

# Decadal Plan for Semiconductors

## FULL REPORT



---

# Table of Contents

|  |     |
|--|-----|
| Introduction . . . . .   | 7   |
| Acronym Definitions . . . . .                                    | 8   |
| <b>Chapter 1</b>   |     |
| New Trajectories for Analog Electronics. . . . .                 | 12  |
| <b>Chapter 2</b>   |     |
| New Trajectories for Memory and Storage . . . . .                | 42  |
| <b>Chapter 3</b>   |     |
| New Trajectories for Communication . . . . .                     | 76  |
| <b>Chapter 4</b>   |     |
| New Trajectories for Hardware Enabled ICT Security. . . . .      | 102 |
| <b>Chapter 5</b>   |     |
| New Compute Trajectories for Energy-Efficient Computing. . . . . | 122 |

---

# Decadal Plan Executive Committee

James Ang, PNNL

Dmytro Apalkov, Samsung

Fari Assaderaghi, Sunrise Memory

Ralph Cavin, Independent Consultant

Ramesh Chauhan, Qualcomm

An Chen, IBM

Richard Chow, Intel

Robert Clark, TEL

Maryam Cope, SIA

Debra Delise, Analog Devices

Carlos Diaz, TSMC

Bob Doering, Texas Instruments

Sean Eilert, Micron

Ken Hansen, Independent Consultant

Baher Haroun, Texas Instruments

Yeon-Cheol Heo, Samsung

Gilbert Herrera, SNL

Kevin Kemp, NXP

Taffy Kingscott, IBM

Stephen Kosonocky, AMD

Matthew Klusas, Amazon

Steve Kramer, Micron

Donny Kwak, Samsung

Lawrence Loh, MediaTek

Rafic Makki, Mubadala

Matthew Marinella, SNL

Seong-Ho Park, SK hynix

David Pellerin, Amazon

Daniel Rasic, SRC

Ben Rathsak, TEL

Wally Rhines, Cornami

Heike Riel, IBM

Kirill Rivkin, Western Digital

Juan Rey, Mentor (Siemens)

Dave Robertson, Analog Devices

Gurtej Sandhu, Micron

Motoyuki Sato, TEL

Ghavam Shahidi, IBM

Steve Son, SK hynix

Mark Somervell, TEL

Gilroy Vandentop, Intel

Jeffrey Vetter, ORNL

Jeffrey Welsler, IBM

Jim Wieser, Texas Instruments

Tomomari Yamamoto, TEL

Ian Young, Intel

Todd Younkin, SRC

David Yeh, Texas Instruments

Victor Zhirnov, SRC

Zoran Zvonar, Analog Devices

The Decadal Plan Workshops that helped drive this report were supported by the U.S. Department of Energy, Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) Research Programs, and the National Nuclear Security Agency, Advanced Simulation and Computing (ASC) Program. SIA and SRC are grateful for their support.

Semiconductors, the tiny and highly advanced chips that power modern electronics, have helped give rise to the greatest period of technological advancement in the history of humankind.

Chip-enabled technology now allows us to analyze DNA sequences to treat disease, model nerve synapses in the brain to help people with mental disorders like Alzheimer's, design and build safer and more reliable cars and passenger jets, improve the energy efficiency of buildings, and perform countless other tasks that improve people's lives.

During the COVID-19 pandemic, the world has come to rely more heavily on semiconductor-enabled technology to work, study, communicate, treat illness, and do innumerable other tasks remotely. And the future holds boundless potential for semiconductor technology, with emerging applications such as artificial intelligence, quantum computing, and advanced wireless technologies like 5G and 6G promising incalculable benefits to society.

Fulfilling that promise, however, will require taking action to address a range of seismic shifts shaping the future of chip technology. These seismic shifts—identified in *The Decadal Plan for Semiconductors* by a broad cross-section of leaders in academia, government, and industry—involve smart sensing, memory and storage, communication, security, and energy efficiency. The federal government, in partnership with private industry, must invest ambitiously in semiconductor research in these areas to sustain the future of chip innovation.

For decades, federal government and private sector investments in semiconductor research and development (R&D) have propelled the rapid pace of innovation in the U.S. semiconductor industry, making it the global leader and spurring tremendous growth throughout the U.S. economy. The U.S. semiconductor industry invests about one-fifth of its revenues each year in R&D, one of the highest shares of any industry. With America facing increasing competition from abroad and mounting costs and challenges associated with maintaining the breakneck pace of innovation, now is the time to maintain and strengthen public-private research partnerships.

As Congress works to refocus America's research ecosystem on maintaining semiconductor innovation and competitiveness, *The Decadal Plan for Semiconductors* outlines semiconductor research priorities across the seismic shifts noted above and recommends an additional federal investment of \$3.4 billion annually across these five areas.

Working together, we can boost semiconductor technology and keep it strong, competitive, and at the tip of the innovation spear.

Sincerely,



John Neuffer  
President & CEO  
Semiconductor Industry Association (SIA)



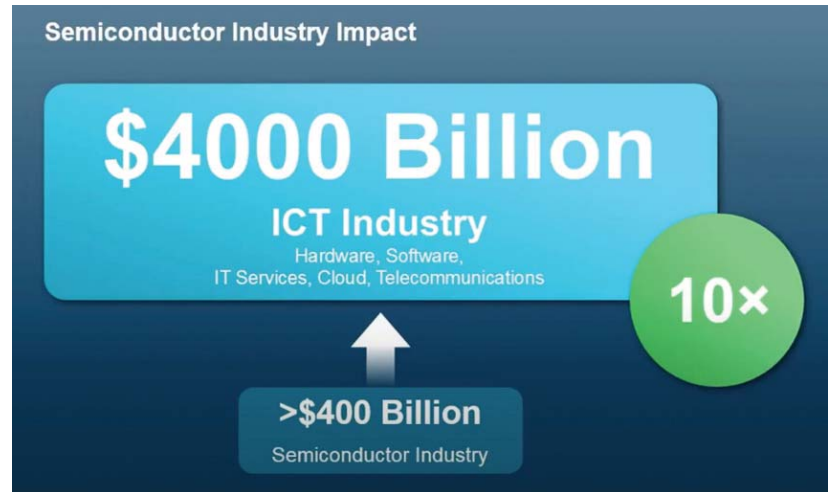
Todd Younkin  
President & CEO  
Semiconductor Research Corporation (SRC)



# Executive Summary

The U.S. semiconductor industry leads the world in innovation, based in large part on aggressive research and development (R&D) spending. The industry invests nearly one-fifth of its annual revenue in R&D each year, second only to the pharmaceuticals sector. In addition, Federal funding of semiconductor R&D serves as the catalyst for private R&D spending. Together, private and Federal semiconductor R&D investments have sustained the pace of innovation in the U.S., enabling it to become the global leader in the semiconductor industry. Those R&D investments have nurtured the development of innovative and commercially viable products, and as a direct result, have led to a significant contribution to the U.S. economy and jobs.

The current hardware-software (HW-SW) paradigm in information and communication technologies (ICT) has made computing ubiquitous through sustained innovation in software and algorithms, systems architecture, circuits, devices, materials, and semiconductor process technologies among others. However, ICT is facing unprecedented technological challenges for maintaining its growth rate levels into the next decade. These challenges arise largely from approaching various fundamental limitations in semiconductor technology that taper the otherwise



necessary generational improvements in the energy-efficiency with which information is processed, communicated, stored, sensed and actuated on. Long term sustainable ICT growth will rely on breakthroughs in semiconductor technology capabilities that enable holistic solutions to tackle information processing efficiency. Disruptive breakthroughs are needed in the areas of software, systems, architectures, circuits, device structure and the related processes and materials that require timely and well-coordinated multidisciplinary research efforts.

This Decadal Plan for Semiconductors outlines research priorities in information processing, sensing, communication, storage, and security seeking to ensure sustainable growth for semiconductor and ICT industries by:

- Informing and supporting the strategic visions of semiconductor companies and government agencies
- Guiding a (r)evolution of cooperative academic, industry and government research programs
- Placing 'a stake in the ground' to challenge the best and brightest researchers, university faculty and students

The Semiconductor Industry Association (SIA) June 2020 report<sup>1</sup> demonstrates that federal investment in semiconductor R&D spurs U.S. economic growth and job creation and presents a case for a 3x increase in semiconductor-specific federal funding. For every dollar spent on federal semiconductor research has resulted in a \$16.50 increase in current GDP.

The Decadal Plan for Semiconductors complements this report and identifies specific goals with quantitative targets. It is expected that the Decadal Plan will have a major impact on the semiconductor industry, similar to the impact of the 1984 10-year SRC Research Goals document that was continued in 1994 as the National Technology Roadmap for Semiconductors, and which later became the International Technology Roadmap for Semiconductors in 1999.

## Trends and drivers

Currently information and communication technologies are facing five major seismic shifts:

### Seismic shift #1

Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces that can sense, perceive and reason

### Seismic shift #2

The growth of memory demands will outstrip global silicon supply presenting opportunities for radically new memory and storage solutions

### Seismic shift #3

Always available communication requires new research directions that address the imbalance of communication capacity vs. data generation rates

### Seismic shift #4

Breakthroughs in hardware research are needed to address emerging security challenges in highly interconnected systems and Artificial Intelligence

### Seismic shift #5

Ever rising energy demands for computing vs. global energy production is creating new risk, and new computing paradigms offer opportunities with dramatically improved energy efficiency

## The Grand Challenge

Information and communication technologies make up over 70% of the semiconductor market share. They continue to grow without bounds dominated by the exponential creation of data that must be moved, stored, computed, communicated, secured and converted to end user information. The recent explosion of artificial intelligence (AI) applications is a clear example, and as an industry we have only begun to scratch the surface.

Having computing systems move into domains with true cognition, i.e., acquiring understanding through experience, reasoning and perception is a new regime. This regime is unachievable with the state-of-the-art semiconductor technologies and traditional gains since the reduction in feature size (i.e., dimensional scaling) to improve performance and reduce costs in semiconductors is reaching its physical limits. As a result, the current paradigm must change to address an information and intelligence-based value proposition with semiconductor technologies as the driver.

<sup>1</sup>Sparkling Innovation: How Federal Investment in Semiconductor R&D Spurs U.S. Economic Growth and Job Creation, SIA Report, June 2020

# Call to Action: Semiconductor Technology Leadership Initiative

Maintaining and strengthening the leadership of the United States in ICT during this new semiconductor era requires a sustained additional \$3.4B federal investment per year throughout this decade (i.e. tripling Federal funding for semiconductor research) to conduct large-scale industry-relevant, fundamental semiconductor research. (The Decadal Plan Executive Committee offered recommendations on allocation of the additional \$3.4B investment per year among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies).

The investments through new public-private partnerships must cover a wide breadth of interdependent technical areas (compute, analog, memory/storage, communications, and security) requiring multi-disciplinary teams to maintain U.S. semiconductor technology leadership.

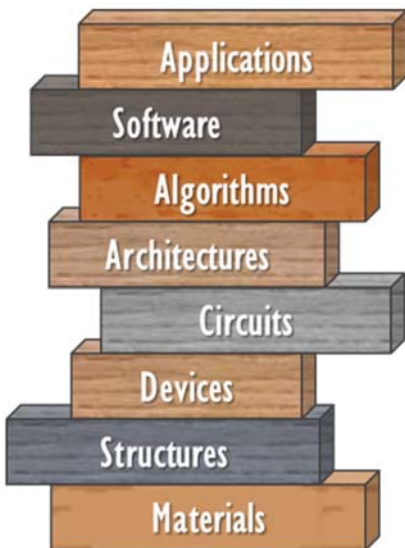
These investments need to be organized and coordinated to support a common set of goals focused on market demand to provide technologies which enable new commercial products and services over the course of the program. The Decadal Plan has identified five seismic paradigm shifts required to accomplish this overarching Grand Challenge.

The Decadal Plan serves as a blueprint for policymakers who recognize this challenge and seek guidance on areas of research emphasis for scientific research agencies and public-private partnerships.

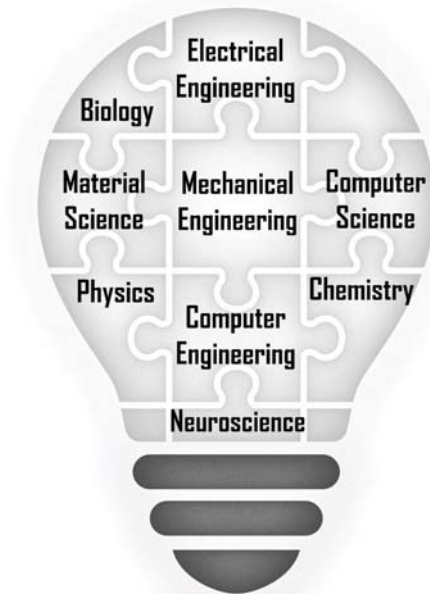
## Semiconductor Technology Breakthroughs Rely On

### Holistic Optimal Solutions

Driven by Hardware/Software Co-Optimization



### Interlocked Multidisciplinary Research





---

# Introduction

Currently information and communication technologies are facing five major seismic shifts:

- Seismic shift #1** Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces that can sense, perceive and reason.
- Seismic shift #2** The growth of memory demands will outstrip global silicon supply presenting opportunities for radically new memory and storage solutions.
- Seismic shift #3** Always available communication requires new research directions that address the imbalance of communication capacity vs. data generation rates.
- Seismic shift #4** Breakthroughs in hardware research are needed to address emerging security challenges in highly interconnected systems and Artificial Intelligence.
- Seismic shift #5** Ever rising energy demands for computing vs. global energy production is creating new risk, and new computing paradigms offer opportunities with dramatically improved energy efficiency.

To support the Decadal Plan development, an international series of five face-to-face workshops was conducted to assess quantitatively each seismic shift, assign targets, and suggest initial research directions. Participants and contributors to these workshops include academic, government and industrial domain experts. The output of these workshops has guided the recommendations in the 2030 Decadal Plan for Semiconductors.

These workshops provided highly interactive forums where key research leaders could evaluate the status of nanoelectronics research and application drivers, discuss key scientific issues, and define promising future research directions. This is instrumental for the published version of the 2030 Decadal Plan for Semiconductors to reflect an informed view on key scientific and technical challenges related to revolutionary information and communication technologies, based on new quantitative analyses and projections.

## The primary objectives of the Decadal Plan include

- Identify significant trends and applications that are driving Information and Communication Technologies and the associated roadblocks/challenges.
- Assess quantitatively the potential and status of the five seismic shifts that will impact future ICT.
- Identify fundamental goals and targets to alter the current trajectory of semiconductor technology.

The Decadal Plan provides an executive overview of the global drivers and constraints for the future ICT industry, rather than to offer/discuss specific solutions: **The document identifies the what, not the how.** In doing so, **it focuses and organizes the best of our energies and skills to the key challenges** in a quantitative manner about which creative solutions can be imagined and their impact measured.

# Acronym Definitions

|                   |  |        |   |
|-------------------|--|--------|---|
| 3GPP              | 3rd Generation Partnership Project   | CMOS   | Complementary Metal-Oxide-Semiconductor             |
| 5G                | fifth generation wireless technology                                       | COTS   | Commercial Off-The-Shelf                            |
| 5GPoA             | 5G Point of Attachment   | CNN    | Convolutional Neural Network                        |
| 6G                | sixth generation wireless technology                                       | CPRI   | Common Public Radio Interface                       |
| 6G                | sixth generation wireless technology                                       | CPU    | Central Processing Unit                             |
| III-V             | compound of type III and type V elements from periodic table               | CS     | Compressive Sampling                                |
| $\Sigma\Delta$    | Sigma Delta—type of analog to digital converter summing difference signals | CTA    | Cross Traffic Alert                                 |
|                   |  | CTE    | Coefficient of Thermal Expansion                    |
|                   |  | CXL    | Compute Express Link                                |
| ACC               | Adaptive Cruise Control  | DAC    | Digital to Analog Converter                         |
| ADAS              | Advanced Driver-Assistance System  | dB     | decibels—logarithm of a ratio                       |
| ADC or A/D        | Analog to Digital Converter  | DDR    | Double Data Rate                                    |
| AEB               | Automatic Emergency Braking  | DARPA  | Defense Advanced Research Projects Agency           |
| AES               | Advanced Encryption System   | DFT    | Design for Test                                     |
| AFE               | Analog Front End   | DNN    | Deep Neural Network                                 |
| AFR               | Annual Failure Rate  | DRAM   | Dynamic Random Access Memory                        |
| AI                | Artificial Intelligence  | DRR    | Data Reduction Ratio                                |
| aJ                | attojoule (10-18 joules)   | DMR    | Digitally Modulated Radar                           |
| AMS               | Analog Mixed Signal  | DNA    | DeoxyriboNucleic Acid                               |
| ANN               | Artificial Neural Network  | DNS    | Domain Name System                                  |
| AR/VR             | Augmented Reality/Virtual Reality  | DOE    | Department of Energy                                |
| ASCR              | Advanced Scientific Computing Research                                     | DPD    | Digital Pre-Distortion                              |
| ASIC              | Application Specific Integrated Circuit                                    | DPPM   | Defective Parts Per Million                         |
|                   |  | DSL    | Domain-Specific Languages                           |
|                   |  | DSP    | Digital Signal Processing                           |
| BaFe              | Barium Ferrite material  | EAMR   | Energy Assisted Magnetic Recording                  |
| BAG2              | Berkeley Analog Generator—2nd generation from UC Berkeley                  | ECID   | Exclusive Chip Identification                       |
| BEOL              | Back End of Line—in semiconductor process                                  | ECRAM  | Electro Chemical Random Access Memory               |
| BER               | Bit Error Rate   | EDA    | Electronic Design Automation                        |
| BIST              | Built in Self Test   | EDFA   | Erbium-doped Fiber Amplifier                        |
| BITS              | Binary Information Throughput (in bits per second)                         | EEPROM | Electrically Erasable Programmable Read Only Memory |
| bits/s, bps, Mbps | Bits per second or Million bits per second                                 | EIRP   | Effective Isotropic Radiated Power                  |
| BS                | Base Station   | ENOB   | Effective Number of Bits                            |
| BSD               | Blind Spot Detection   | eNVM   | embedded Non-Volatile Memory                        |
| BSP               | Bulk Synchronous Programming Model   | ET     | Envelope Tracking                                   |
| BW                | bandwidth  | EVM    | Error Vector Magnitude                              |
| CAD               | Computer Aided Design  | FDM    | Frequency Division Multiplexing                     |
| CBRAM             | Conductive Bridging Random Access Memory                                   | FeFET  | Ferroelectric Field Effect Transistor               |
| CCIX              | Cache Coherent Interconnect For Accelerators                               | fF     | femto Farad—capacitance measure $10^{-15}$ Farad    |
| CFR               | Crest Factor Reduction   | FFT    | Fast Fourier Transform (signal analysis)            |
| CGS               | capacitance from transistor gate to source                                 | FinFET | field effect transistor built as a “fin” vertically |
| Class A           | amplifier circuit with continuous conduction                               | fJ/op  | femto ( $10^{-15}$ ) Joule per operation            |
| Class B           | amplifier circuit with switching conduction—positive and negative          | FLOP   | Floating Point Operation                            |

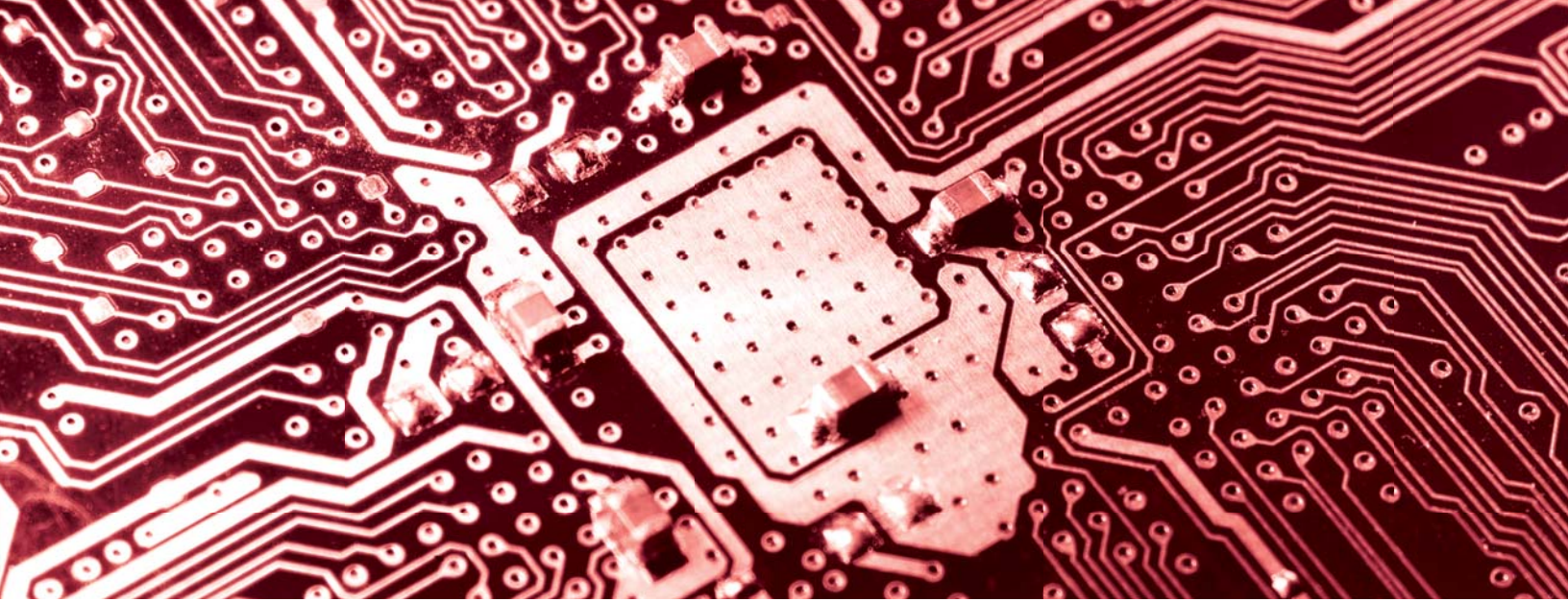
|                  |  |                |  |
|------------------|--|----------------|--|
| $f_{\max}$       | maximum oscillation frequency or power gain frequency  | IDS            | transistor drain to source current                       |
| FMCW             | Frequency Modulated Continuous Wave (radar)            | IEEE           | Instituted of Electrical and Electronics Engineers       |
| FoM              | Figure of Merit  | IIoT           | Industrial IoT   |
| FPAA             | Field Programmable Analog Array                        | IMC            | Integrated Memory Controller                             |
| FPGA             | Field-Programmable Gate Array                          | IMDD           | Intensity Modulation Direct Detection                    |
| FRESCO           | Frequency-stabilized coherent optical                  | InGaAs         | Indium Gallium Arsenide material                         |
| FSO              | Free Space Optical                                     | InP            | Indium Phosphide material                                |
| $f_T$            | transition frequency where current gain is zero        | I/O            | Input / Output   |
| FTE              | Full Time Equivalent (engineering resource personnel)  | IoT            | Internet of Things                                       |
| FTTH             | Fiber to the Home                                      | IP or IP block | semiconductor Intellectual Property core                 |
|                  |  | IPS            | Instructions per Second                                  |
| GAA              | gate all around technology                             |                |  |
| GaN              | Gallium Nitride—transistor material                    | J              | Joule  |
| $g_m$            | transconductance or current dependence on gate voltage | KGD            | Known Good Die   |
| $g_o$            | transistor output conductance                          |                |  |
| GaAsSb           | Gallium Arsenide Antimonide material                   | LCA            | Lane Change Assist                                       |
| GB               | Giga Byte  | LDE            | Layout Dependent Effects                                 |
| Gbps             | Gigabits Per Second                                    | LED            | Light Emitting Diode                                     |
| GDDR             | Graphics Double Data Rate                              | LEO            | Low Earth Orbit  |
| GHz              | giga-hertz ( $10^9$ )                                  | LiFi           | Light Fidelity—wireless optical communication technology |
| GNN              | Graph Neural Network                                   | LNA            | Low Noise Amplifier                                      |
| GPS              | Global Positioning System                              | LO             | Local Oscillator   |
| GPU              | Graphics Processing Unit                               | LOS            | Line of Sight  |
| GS/s             | Giga-Samples per second ( $10^9$ )                     | LPDDR          | Low Power Double Data Rate                               |
| GW               | Giga Watt ( $10^9$ )                                   | LTE            | Long Term Evolution—wireless standard                    |
|                  |  | LTO            | Linear Tape Open   |
| HAMR             | Heat-Assisted Magnetic Recording                       |                |  |
| HBM              | High Bandwidth Memory                                  | M2M            | Machine to Machine                                       |
| HBT              | Hetero-junction Bipolar Transistor                     | MEC            | Multi-access Edge Computing                              |
| HD               | High Dimensional                                       | MEMS           | Micro- electromechanical systems                         |
| HDC              | High Dimensional Computing                             | MESO           | Magnetolectric spin-orbit                                |
| HDD              | Hard Disk Drive  | MID            | Mobile Infrastructure on Demand                          |
| HEMT             | High Electron Mobility Transistor                      | MIEC           | Mixed Ionic-Electronic Current                           |
| HfO <sub>x</sub> | Hafnium Oxide material                                 | MIMO           | Multiple-Input and Multiple-Output                       |
| HIVA             | Hardware-Intensive Virtualization Architecture         | MIST           | Molecular Information Storage                            |
| HPC              | High Performance Computing                             | MIT            | Metal-Insulator Transition                               |
| HTTP             | Hyper Text Transfer Protocol                           | ML             | Machine Learning   |
| HVAC             | Heating, Ventilation and Air Conditioning              | MLC            | Multiple Level Cells                                     |
| HW               | Hardware   | MOS            | Metal Oxide Semiconductor material system                |
|                  |  | mm-Wave        | millimeter wave  |
| IAB              | Integrated Access and Backhaul                         | MRAM           | Magnetic Random Access Memory                            |
| IC               | Integrated Circuit                                     | mMTC           | massive Machine Type Communication                       |
| ICT              | Information and Communication Technologies             | MW             | mega watt power ( $10^6$ watts)                          |
| ICM              | In Compute Memory                                      | mW             | milli watt power ( $10^{-3}$ watts)                      |
| ID               | transistor drain current                               |                |  |

# Acronym Definitions

|            |   |         |  |
|------------|---|---------|--|
| NAND flash | highest-density silicon-based electronic nonvolatile memory | R&D     | Research and Development   |
| NEP        | Noise Equivalent Power                                      | RAID    | Redundant Array Of Inexpensive Disks   |
| NF         | Noise Figure  | RAM     | Random Access Memory   |
| NFC        | Near Field Communications                                   | RAT     | Radio Access Technology  |
| nJ         | nanojoule (10 <sup>-9</sup> joules)                         | RDMA    | Remote Direct Memory Access  |
| NLoS       | Non Line of Sight   | ReRAM,  | Resistive Random Access Memory   |
| nm         | nano meter (10 <sup>-9</sup> meters)                        | RRAM    |  |
| NMOS       | n-channel Metal-Oxide-Semiconductor transistor              | RF      | Radio Frequency  |
| NMR        | Nuclear Magnetic Resonance                                  | RFFE    | Radio Frequency Front End  |
| NVM        | Nonvolatile Memory  | RFIC    | Radio Frequency Integrated Circuit   |
|            |   | RFSOI   | Radio Frequency Silicon on Insulator, specialized semiconductor process for RF chips       |
| OAM        | Optical Angular Momentum                                    | RGB     | Red Green Blue   |
| OFC        | Optical Frequency Comb                                      | RLC     | Resistance, Inductance(L), Capacitance   |
| ORNL       | Oak Ridge National Laboratory                               | RoT     | Root of Trust  |
| OTA        | Over-The-Air  | RRH     | Remote Radio Head  |
| OVS        | Ovonic Threshold Switch                                     | RRM     | Radio Resource Management  |
|            |   | RSA     | (Rivest–Shamir–Adleman) is a cryptosystem that is widely used for secure data transmission |
| P2P        | Pear to Pear  | RTL     | Register-transfer Level is a design abstraction which models digital circuits              |
| PA         | Power Amplifier   |         |  |
| PAE        | Power Added Efficiency                                      | SAR     | Successive Approximation Register type of ADC  |
| PCB        | Printed Circuit Board                                       | SCM     | Software Configuration Management  |
| PCM        | Pulse-Code Modulation                                       | SDR     | Software Defined Radio   |
| PCRAM      | Phase Change Random Access Memory (also PCM)                | SEC     | Statistical Error Compensation   |
| PDM        | Polarization-Division Multiplexing                          | Si      | Silicon material   |
| Petaflops  | 10 <sup>15</sup> floating point operations per second       | SIA     | Semiconductor Industry Association   |
| PHY        | Physical Layer  | SiGe    | Silicon Germanium material for transistor  |
| PKI        | Public Key Infrastructure                                   | SIMO    | Single Input Multiple Output   |
| PLC        | Programmable Logic Control                                  | SiP     | System in Package  |
| PLL        | Phase Locked Loop   | SiOx    | Silicon Oxide material   |
| PPM        | Parts Per Million   | SMF     | Single Mode Fiber  |
| PRBS       | Pseudo-Random Binary Sequence                               | SMR     | Shingled Magnetic Recording  |
| PRD        | Priority Research Direction                                 | SNN     | Spiking Neural Network   |
| PQC        | Post-Quantum Cryptography                                   | SNR     | Signal to Noise Ratio  |
| PUF        | Physical Unclonable Function                                | SoC     | System on Chip   |
| pW/√Hz     | pico Watt(10 <sup>-12</sup> ) per square root of Hertz      | SOT     | Spin Orbit Torque  |
| PZT        | Lead-Zirconia Titanate material                             | SPICE   | transistor / component level simulation program  |
|            |   | SRAM    | Static Random Access Memory  |
| QAM        | Quadrature Amplitude Modulation                             | SrFe    | Strontium Ferrite material   |
| QOS        | Quality Of Service  | SRC     | Semiconductor Research Corporation   |
| QKD        | Quantum Key Distribution                                    | SSD     | Solid State Drive  |
| QLC        | Quad-Level Cell   | STT-RAM | Spin Transfer Torque Random Access Memory  |
| QPSK       | Quadrature Phase Shift Keying                               | SW      | Software   |

|           |   |
|-----------|---|
| Tbps      | Terabit-per-second  |
| TCB       | Trusted Computing Base  |
| TCO       | Total Cost Of Ownership   |
| TEE       | Trusted Execution Environments  |
| TFET      | Tunneling Field Effect Transistor   |
| THz       | Terahertz ( $10^{12}$ )   |
| TLC       | Trust Level Control   |
| TOPS      | Tera Operations per Second ( $10^{12}$ )                                      |
| TPM       | Technological Protective Measures   |
| TPU       | Tensor Processing Unit  |
| TCP       | Transmission Control Protocol   |
| TSV       | Through-Silicon-Vias  |
| TX        | Transmitter   |
| UWB       | Ultra Wideband (communication)  |
| UWBG      | Ultra Wide Band Gap (transistor technology)                                   |
| V2V       | Vehicle to Vehicle  |
| VCO       | Voltage Controlled Oscillator   |
| VCMA      | Voltage Controlled Magnetic Anisotropy  |
| VDD       | Positive supply voltage (CMOS IC)   |
| VDSat     | Transistor drain voltage where output current versus gate voltage flattens ou |
| VGS       | Voltage between gate and source of transistor                                 |
| VLSI      | Very large scale integration  |
| VM        | Virtual Machine   |
| VMM       | Vector Matrix Multiplier  |
| VR        | Virtual Reality   |
| WLAN      | Wireless Local Area Network   |
| WDM       | Wavelength Division Multiplexing  |
| xHaul     | Transport network for 5G  |
| XPoint    | Cross Point   |
| Zettabyte | $10^{21}$ bytes   |
| ZF        | Zero Forcing  |
| ZIF       | Zero Frequency IF (Intermediate Frequency)                                    |
| ZIPS      | $10^{21}$ compute instructions per second                                     |





---

## Chapter 1

# New Trajectories for Analog Electronics

## Seismic shift #1

Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces that can sense, perceive and reason.

### 1.1. Executive Summary

Analog electronics deals with real-world continuously variable signals of multiple shapes (in contrast to digital electronics where signals are usually of standard shape, taking only two levels, ones or zeros). The analog electronics domain encompasses multiple dimensions as shown in **Figure 1.1**. Also, all inputs humans can perceive are analog, which calls for bio-inspired solutions for world-machine interfaces that can sense, perceive, and reason based on ultra-compressed sensing capability and low operation power (**Figure 1.2**). This extends to real-world interfaces such as communication channels (wired or wireless), machine and infrastructure sensing and control, as well as environmental, diagnostics, and converting various sources of nature-produced energy to useable power. *The physical world is inherently analog and*

*the “digital society” drives increasing demand for advanced analog electronics to enable interaction between the physical and “cyber worlds.”*

Sensing the environment around us is fundamental to the next generation of AI, where devices will be capable of perception and reasoning on sensed data that is more stochastic in the presence of noise, as opposed to exact digital precision. In fact, the human brain operates in such a manner, as more of a massive parallel analog computation engine. The world-machine interface lies at the heart of the *current information-centric economy*. **As one example, the next wave of the advanced manufacturing revolution is expected to come from next-generation analog-driven industrial electronics, that includes sensing, robotics, industrial, automotive, medical etc.** For mission-critical applications, the reliability of electronic components is a priority.

The estimated total analog information generated from the physical world across an estimated 45 trillion sensors in 2032 is equivalent to  $\sim 10^{20}$  bit/s. As a reference, the total collective human sensory throughput pales at  $\sim 10^{17}$  bit/s. **Thus, our ability to perceive the physical world is significantly limited and a significant paradigm shift towards extracting key information and applying it in an appropriate way is**

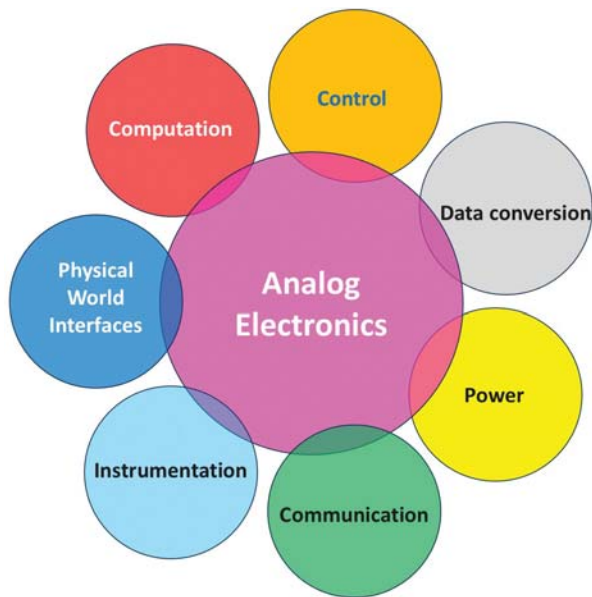


Figure 1.1: The Dimensions of Analog Electronics

**necessary to harvesting the data revolution.** Additionally, **the massive amounts of data from sensors cannot be transmitted to the cloud for processing due to limits of communications capacity, energy and timeliness of information.**

### Call for action

The analog interface bridges the physical and digital worlds. Our collective ability to access the information of the physical world through analog signals is 10,000 trillion times below what is available, and radical advances in analog electronics will be required soon. New approaches to sensing such as sensing to action, analog “artificial intelligence” (AI) platforms, brain inspired/neuromorphic and hierarchical computation, or other solutions will be necessary. Breakthrough advances in information processing technologies, such as developing perception algorithms to enable understanding of the environment from raw sensor data, are a fundamental requirement. New computing models such as analog “approximate computing,” which can trade energy and computing time with accuracy of output (presumably how the brain does) are required. New analog technologies will offer great advancements in communication technologies. The ability to collect, process and communicate the analog data at the input/output boundaries is critical to the future world of IoT and Big Data. Additionally, analog development methodologies require a step increase (10x or greater) in productivity to address the application explosion in a timely manner. Altogether, collaborative research to establish revolutionary paradigms for future energy-efficient

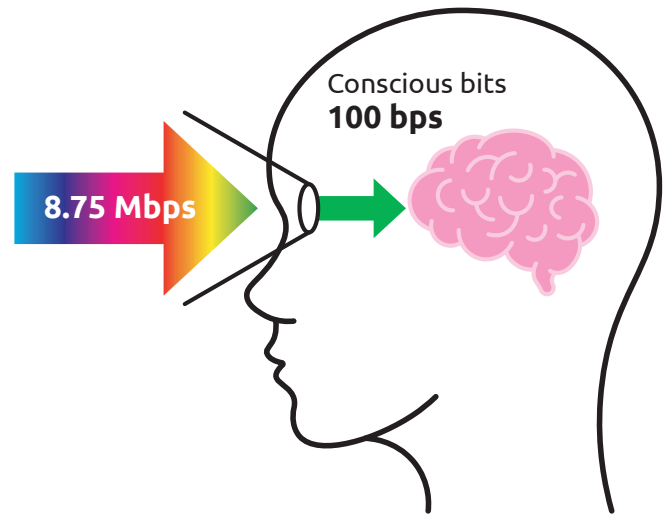


Figure 1.2: The brain’s ability to perceive and reason is based on ultra-compressed sensing capabilities with 100,000 data reduction and a low operation power

analog integrated circuits for the vast range of future data types, workloads and applications is needed.

The **Analog Grand Goal** is for revolutionary technologies to increase actionable information with less energy, enabling efficient and timely (low latency) sensing-to-analog-to-information with a practical reduction ratio of  $10^5:1$ .

**Invest \$600M annually throughout this decade in new trajectories for analog electronics. Selected priority research themes are outlined below.<sup>1</sup>**

Decadal Plan for Semiconductors workshop was held on “New Trajectories for Analog Electronics” to address this Grand Goal. The workshop was organized by and held with experts from academia, industry, and government labs, and it consisted of five sessions:

- Analog ICT Systems Fundamentals, Challenges, and Application Drivers
- Intelligent Sensors: Sensing to Action
- Analog in the THz Regime
- Analog in Machine Learning at the Edge
- Analog Design Productivity and Predictability

The remainder of this chapter on New Trajectories for Analog Electronics covers these sessions in more detail

<sup>1</sup>The Decadal Plan Executive Committee offered recommendations on allocation of the additional \$3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.



and includes resulting research direction recommendations. In the end it becomes obvious that a holistic approach is required to achieve such an aggressive goal and leverage analog technology to better interface with and “make sense” of the real world, as well as make effective use of the information that the real world provides. Co-design of such future microelectronics systems was called out as a key element in the “Basic Research Needs for Microelectronics” report published by the Department of Energy Office of Science Workshop in 2018<sup>1</sup>. This involves evaluating all technology levels, from materials through systems, which all entail analog electronics and microsystems, as well as digital processing and software (Figure 1.3).

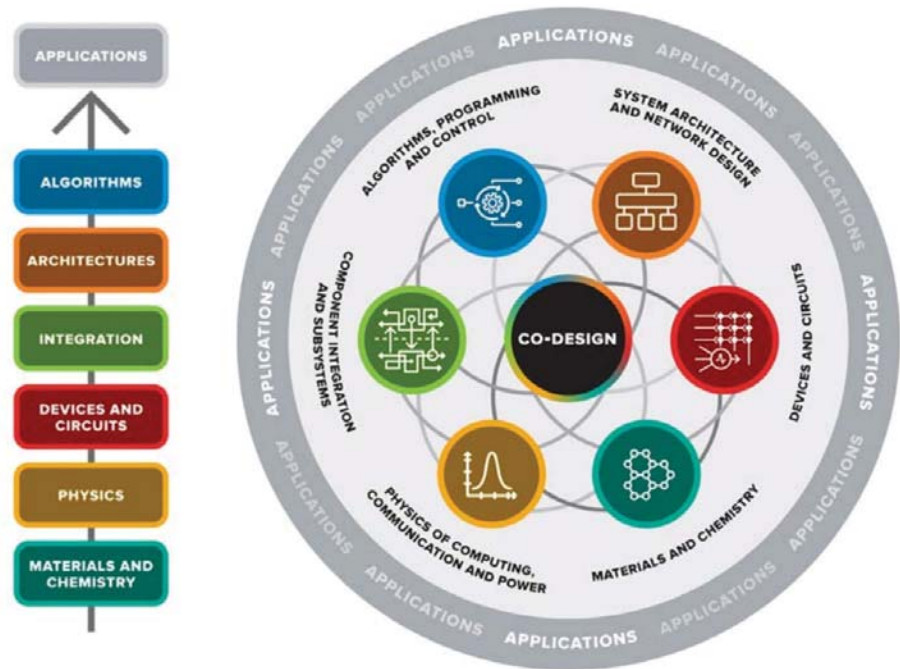


Figure 1.3: Holistic approach required to achieve future analog goals and leverage analog technology for better real-world interface<sup>1</sup> (courtesy of DoE)

## 1.2. Analog ICT Systems: Fundamentals, Challenges, and Application Drivers

### Overview and Needs

While nobody can predict how ICT will develop in the future, *it is worthwhile to explore best-case scenarios that are bounded only by what is technically possible*. Analog information and communication technologies (ICT), which sense and interface to the real world, support daily social life and economic activities. This section identifies fundamental limits and provides an open forum for brainstorming unserved and future applications with their corresponding implications for the semiconductor industry. Furthermore, new emerging analog solutions are discussed in the context of the application space they enable or the trendline in energy, bandwidth, etc., that they dramatically alter. All signals are fundamentally analog waveforms and restricted by analog/physics limitations. In many cases, analog signal processing is more effective and efficient for preparing signals for interpretation by further digital processing.

*For many Analog ICT system drivers, the prime research direction is increasing bandwidth for future 6G wireless networks, data centers, remote medicine, and beamformers for wideband radar, which demand faster Analog-to-Digital Conversion.*

Further downstream, demanding specifications of denser and more power-efficient memory have driven 3D architectures in Very Large System Integration (VLSI), which face their own interesting tests of cost and heat management. Figure 1.4 shows the roadmap for transitions in *FinFET and gate-all-around (GAA) transistor design for storage devices which yield better channel control and, hopefully, eliminate bottleneck design rules*<sup>2</sup>. Other 3D approaches embrace “heterogeneous” integration for more efficient and effective processing of information.

**One fundamental question is how innovation in analog electronics can help with today’s advanced computing and information processing paradigms.** This demands a much deeper appreciation of device scaling, efficient signal processing, and circuit architectures that can be incorporated in high-speed machine interfaces. There is an even greater need for understanding tradeoffs between power consumption and other desirable metrics like high gain, noise, leakage, and reduced supply interference. One set of device-level tradeoffs are visualized in Figure 1.5, considering an ideal NMOS transistor<sup>3</sup>.

>2020: 2.5D/3D fine-pitch assembly + stacking

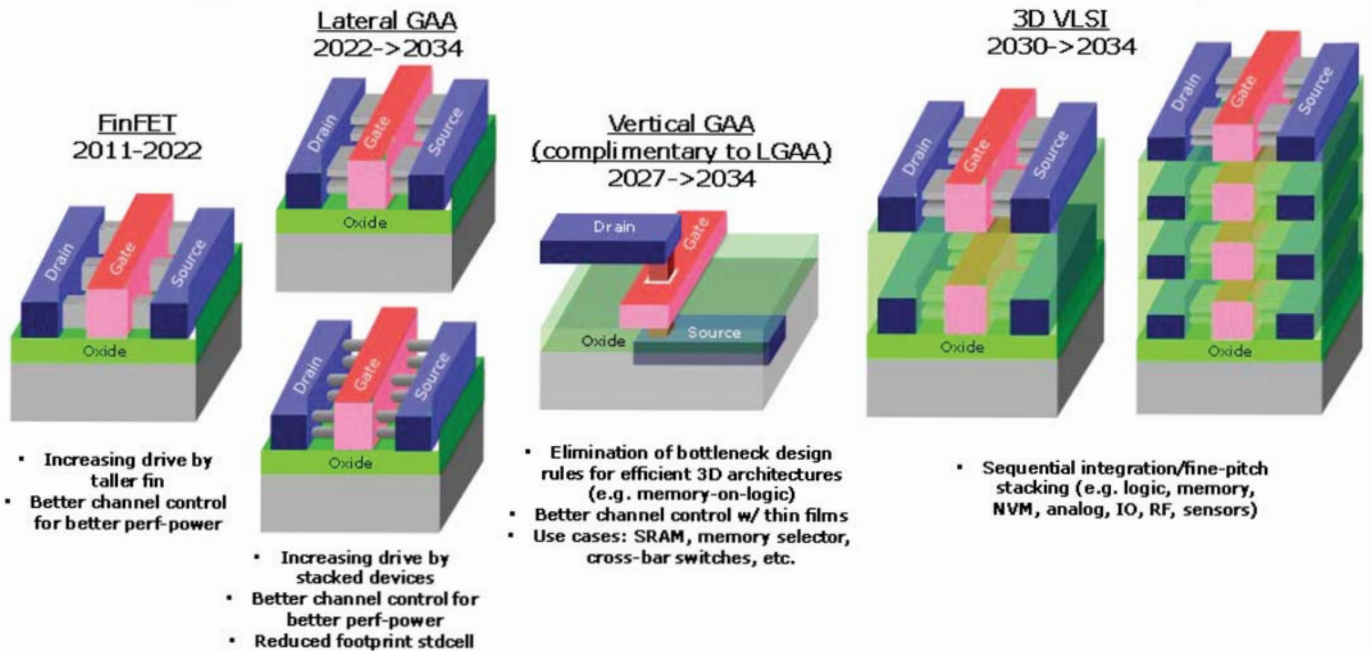


Figure 1.4: Roadmap for evolution of 3D VLSI transistors and architecture<sup>2</sup> (courtesy of Gabriele Manganaro, Analog Devices)

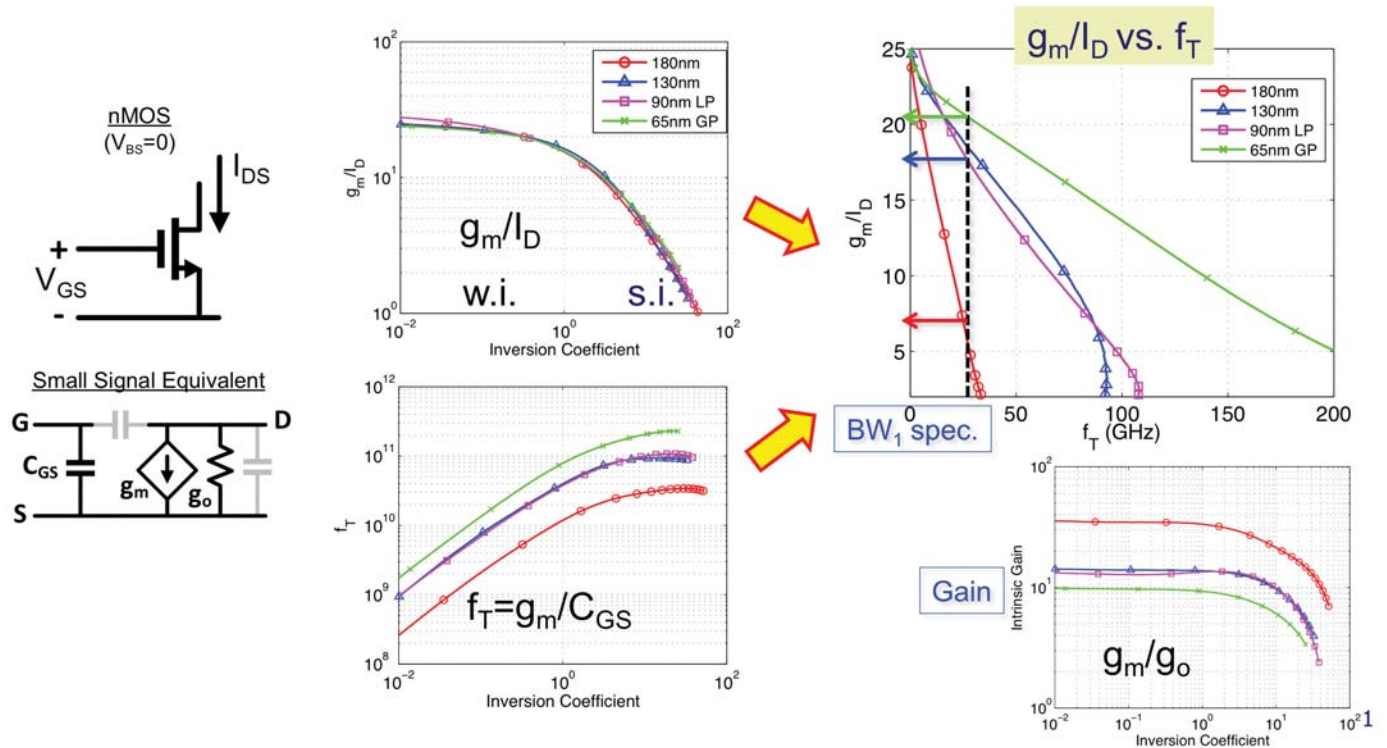


Figure 1.5: An instance of tradeoffs encountered in MOS device scaling<sup>3</sup> (courtesy of Peter Kinget, Columbia University)

*As progression is made toward smaller nodes while power efficiency increases, the improvements in gain and transition frequency have reduced drastically* (limiting use for wireless communication). It should also be noted that power-supply voltage cannot be arbitrarily reduced because of the implications on signal distortion and noise constraints from preferred modes of operation (Class A, Class B, etc.).

Newer circuit architectures will assist in supply-voltage scaling. For example, Operational Amplifiers (Op-Amps), conventional  $V_{DD}$  scaling runs into smaller output swings, higher saturation voltages ( $V_{DSat}$ ), and larger noise consumption in noise-limiting first-stages. Alternative structures, such as switched-mode topologies, may be implemented to achieve rail-to-rail output, low output impedance, and greater bandwidth.

*Future design iterations are underway on three fronts.* The first is better analog-to-digital conversion, which circumvents problems with bandlimited analog signals by efficient oversampling and quantization. The second is conversion of information in sparse analog signals to digital data by compressive sampling (CS). The third is feature extraction from analog signals, aided by machine-learning, when the frequency of the feature is less than the maximum frequency content of the signal. One application of compressive sampling using a single-pixel camera for fast-spectrum scanning is shown in **Figure 1.6**<sup>3</sup>.

## Sensors and Actuators for the next decade

The mobile phone economy is driven largely by cost, size, performance, and bandwidth. Critical to the success of the handset's GPS navigation, optical and electronics image stabilization and fingerprint authentication is robust sensor design. *Sensitivity and accuracy grew tenfold in the past five years while power, cost, and size have reduced to a fifth in that time*<sup>4</sup>. These trends are expected to continue. The fusion of physics and artificial intelligence in on-device computing has made possible better designs in MEMS-based accelerometers, gyroscopes, ultrasonic fingerprint sensors, biometric sensors, and microphones, to name a few. The integration of all these sensors make for seamless execution of activities like navigation dead-reckoning, stability control, impact detection, adaptive lighting, image stabilization, and traction control. **Better sensor performance would imply higher SNR, higher dynamic range and a less-than-1mW power consumption regime. It is also desirable to fabricate these in a <55 nm node and use ultra-small and environment-proof packaging.** Sensors and actuators with associated signal processing are a key focus of Intelligent Sensing discussed in section 1.3.

## High-Yielding and High-Performance ADCs in Sub-16nm Process Technologies

The need for faster Analog-to-Digital Converters (ADC) can be contextualized with an example. In a Digitally Modulated

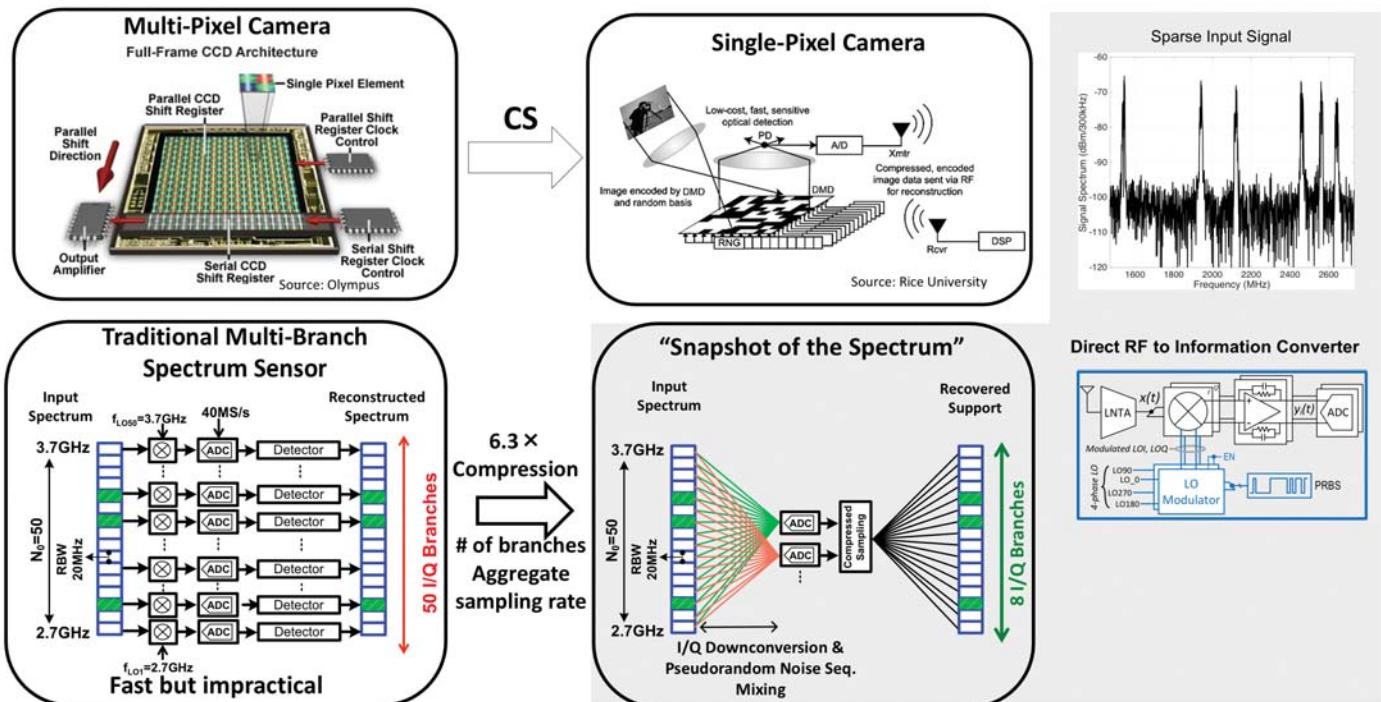


Figure 1.6: Fast-spectrum scanning using compressive sampling<sup>3</sup> (courtesy of Peter Kinget, Columbia University)



Radar-based (DRM-based) automotive radar system, a (digital) Pseudo-Random Binary Sequence (PRBS) modulates the phase of a continuous-wave signal (~79GHz) that is then transmitted as the range-finding signal. The PRBS is designed to enable both unambiguous range (sequence length) and range resolution (pulse-width bandwidth) while also providing excellent immunity to interference. The reflected signal is then received, sampled by an ADC, correlated, and accumulated in real time to determine the target range. Finally, the velocity is determined via an FFT calculation. This scheme is simple but requires extremely fast signal processing of the multi-GHz ADC output data by the correlator and accumulator, in contrast to FMCW-based radar. Computational speed requirements are even greater when MIMO-based beamforming is applied to increase range and angular resolution. The difficulty arises when the received PRBS signals must have very high bandwidth (2-5 GHz) and correspondingly high sampling-rate (4G Sample/second to 10GS/s) ADCs that must also contend with non-idealities of jitter, skew, noise, etc. The situation is further complicated by the need to process (i.e., correlate, accumulate, and do the FFT calculation) the multi-GHz ADC output data in real time on the same die as the ADC. This requirement for GHz signal processing of the ADC output data requires a state-of-the-art CMOS process technology (i.e. 16nm or less) which in turn drives the need for ADCs implemented in the same sub-16nm process. *One example of the architecture of a DMR-based transceiver implemented in a 28nm CMOS process technology and capable of significant processing gain (> 70 dB)<sup>5</sup>.* The performance of this transceiver was indeed limited by the achievable digital signal processing rate of the 28nm technology used, thereby providing direct evidence of the need for ADCs implemented in faster CMOS process technologies.

Given the critical need for GS/s ADCs in sub-16nm process technologies for commercial applications such as the presented DMR-based automotive radar, there is a corresponding critical need for research in this area. At this point, much of the published research on GS/s ADCs being carried out at universities is driven by pressure to ground their ADC research on commonly used figures of merit based solely on power, effective number of bits (ENOB), and sample rate. But these completely ignore critical application-based requirements (e.g., interference, temperature range, cost, etc.) that are essential for real-life use. Therefore, **application-oriented, fit-for-purpose ADCs must be designed to be manufacturable in high-volume production, meeting extreme environmental temperature ranges and reliability requirements, while also maintaining good FoMs.** So, both university-based and industrial research into GS/s ADCs must be redirected with this goal in mind. In turn, this goal will require universities, industry, and funding organizations,

such as SRC, DARPA, NSF, etc., to partner together to enable consistent access to sub-16 nm processes going forward.

## Better Power Electronics with GaN and Beyond

Every electronic system or component requires power that is converted from various sources. The growing pervasiveness of electronics in everything is driving up energy use and necessitates higher efficiency. Additionally, *density is becoming more critical for everything from mobile devices to data centers, with power being challenged by the volume of basic components such as inductors and capacitors.* Focus on **new innovative solutions to address these needs is required to address the size, cost and efficiency of power conversion.** A holistic approach to the power chain provides opportunity for more optimum solutions, from fundamental devices/ components to topologies and overall power-chain architectures.

The Baliga high-frequency figure-of-merit is an important metric for power semiconductor devices operating in high-frequency circuits<sup>6</sup>. This FoM predicts that the power losses incurred in the power device will increase as the square root of the operating frequency, and approximately in proportion to the output power. *This power loss can be reduced by using semiconductors with larger mobility and critical breakdown electric field. Gallium Nitride (GaN) devices have been found to have low gate capacitance and low on-resistance* and, thereby, reduce gate driver losses and improve on this metric. In applications like wireless power transfer and autonomous cars, techniques like Envelope-Tracking, when used efficiently in power converters, reduce wastage of energy by heat dissipation<sup>7</sup>.

Also, *GaN devices are now being integrated at higher levels of hierarchy in Integrated Circuits. These include Monolithic Gate Drivers and Switches, Power System on Chip, Power System in Package (PowerSiP), and Power and Active Interposers with integrated voltage regulators.* Particularly, PowerSoC will enable integration of controller, gate drive, sensing, protection, and inductors (or transformers). Integrated Voltage Regulators can provide significant benefits for VLSI systems in terms of on-chip area, switching frequency, and battery life. A roadmap of current developments in GaN integration is discussed in<sup>8</sup>.

Further improvements in substrate technology will help sustain this trajectory. For example, *200mm GaN-on-silicon and GaN-on-CTE matched substrates look promising.* Also, *Gallium Nitride on Oxide (GaN<sub>o</sub>X)* has superior crystal quality, which makes it suitable for a wide spectrum of applications while offering a more attractive price/performance ratio when compared to SiC (up to 1200V<sup>9</sup>). It is expected that **GaN transistors will replace silicon power MOSFETs with a lower-cost and higher-efficiency solution.** Overall, Ultra-

Wide Band Gap (UWBG) devices for further efficiencies and power density is highlighted in PRD 5 of the “Basic Research Needs for Microelectronics” report published by the Department of Energy Office of Science Workshop in 2018<sup>1</sup>.

## Analog Synaptic Devices for Artificial Intelligence

Bioinspired technology in analog design gives an impetus to realization of neuromorphic learning and in-memory computing that could speed up real-time sensor signal processing. However, **deep-learning algorithms in cloud-based systems are very power-hungry**. Edge systems supporting Internet of Things (IoT) networks perform computing at the sensor node and consume less power. Therefore, they are preferable in realizing adaptive transfer learning. It is here that Emerging Non-Volatile Memories (eNVMs) capable of 100 Tera-Operations/Second/Watt (TOPS/watt) find great applications. These devices include *Filamentary and Non-Filamentary Resistive-change RAMs (RRAMs), Phase-Change Random-Access Memories (PCRAMs), and Ferroelectric RAMs (FeRAMs)*. Their structures are shown in **Figure 1.7**. These eNVMs should be designed better to yield denser In-Compute Memories (ICMs), which would increase cipher texts, consume less energy, and help realize homomorphic encryption.

Also, crossbars used in new architecture should ideally have an efficient analog-to-digital conversion with tunable precision for each column. New architectures may find applications in resolving satisfiability (SAT) problems specific to applications such as crosstalk noise prediction in integrated circuits, model checking, testing of finite-state systems, technology-mapping in logic synthesis, and AI planning and automated reasoning. Here, *analog SAT solvers promise to be more efficient in time, area, and energy*<sup>1</sup>. They provide additional means to trade off time and power by implementing single variable cells or switched variable capacitors<sup>12</sup>. Wider research is now underway for scaling modular parallel circuits and systems for SAT solving and developing programmable ICM unit cells with local SRAM and low-energy charge-domain multiply-accumulate compute elements<sup>13</sup>. Additional *exploration and study of the brain's neural spiking and integration to trigger further spikes, combined with pseudo-analog memory (single variable cells and widely parallel), may provide innovation in processing sensor data for the future*. Further discussion of Analog Machine Learning at the edge is addressed in section 1.5.

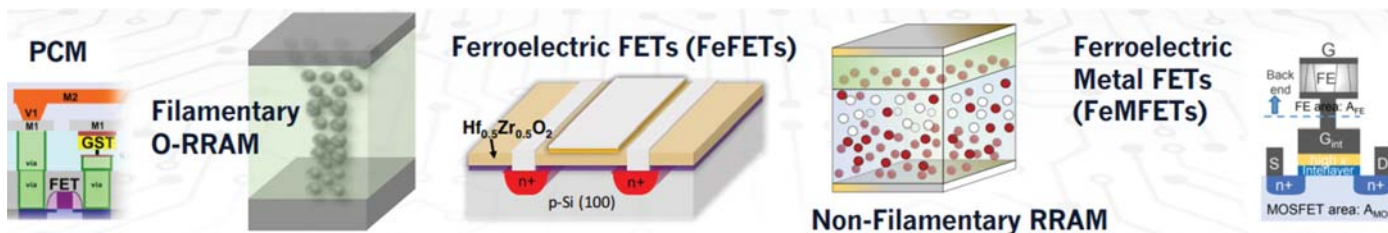


Figure 1.7: Structures of analog synaptic devices used in AI Engines<sup>10</sup> (courtesy of Michael Niemier, University of Notre Dame)

## 1.3. Intelligent Sensing: Sensing to Action

### Overview and Needs

*Over the next decade, a “smart society” including smart factories, smart cities, smart cars and more will become a reality enabled by increased advances in electronic technology.* Key drivers include energy efficiency, safety, productivity, flexibility, and health, as well as entertainment and personalization. Addressing these drivers requires sensing the real world and taking appropriate action in a timely and efficient manner. The vast majority of sensors receive analog inputs from the physical world. *Digitizing these signals creates enormous amounts of raw data, and the data load is predicted to grow at exponential rates given the sheer*

*number of sensors predicted to be deployed.* The questions to be addressed are **how, when, and where to process the data from the growing application of sensors in order to extract information, gather insights, make decisions, and take action.**

With increasing demand for visual information (security cameras, vehicle 360 cams, facial recognition, etc.) and higher resolution, the average **data-acquisition rate per sensor has seen an exponential increase.** Data growth from sensors (**Figure 1.8**) has been estimated to reach 1BB = brontobyte =  $10^{27}$  bytes-per-year by 2032, which corresponds to  $>10^{20}$  bit/s. (See Appendix A3 for details.)

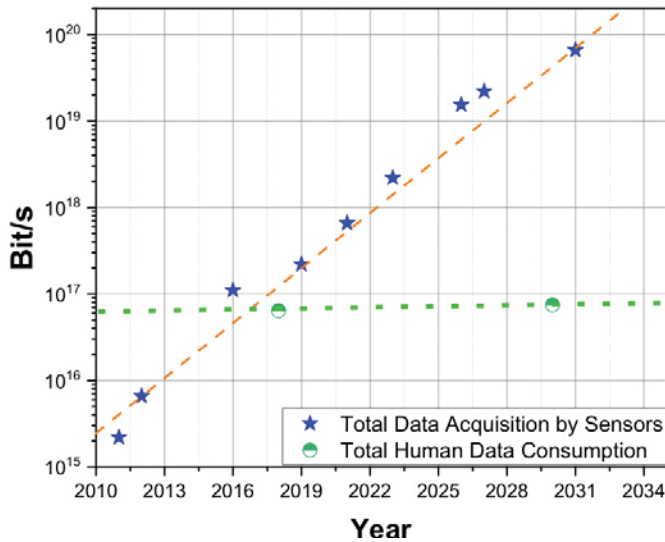


Figure 1.8: World sensor data trend and projection

### Analog Sensor Data-Deluge Problems

There are two key issues with this level of data generation:

- 1) **Digesting or effectively using the data output from sensors toward a smarter society;** and
- 2) **Processing the data efficiently to take appropriate action.**

Digesting the data is well beyond human capability in volume, comprehension, and timeliness. Total human data consumption is estimated to be ~10<sup>17</sup> bit/s (see Appendix A2), while the estimated data generation is >1000X over the next decade. As exponential sensor growth is projected, machine processing is required to make effective use of the sensors being deployed.

This leads to the second problem of processing this data to take appropriate action. As indicated, machine processing is required, which typically streams data over some communications medium for processing and sending the appropriate information back to take action. At the data rates projected (10<sup>20</sup> bit/s), this would require 100MW, assuming only 1pJ/bit. The *most aggressive* communication targets from this Decadal Plan are >100X this level (0.1nJ/bit), resulting in 10GW for communications alone.

Significant change is required if we are to make use of this projected growth in analog sensor data to address the drive for a smarter society. A smarter society can better manage the grid infrastructure in real time and on microgrid scale (including the dynamics of renewable energy), provide flexible/efficient manufacturing with improved yields, and improve building efficiency through lighting and HVAC systems that track need rather than predetermined programs.

**Key to addressing the analog data deluge is increasing the capability of the sensor, signal processing, and subsequent decision-making to take action locally rather than transmitting data any significant distance for processing (i.e., to the cloud as an extreme).** This “Sensing to Action” has the objective of optimizing the system partitioning to manage the amount of data communicated in the system. There will be a balance of what can realistically be processed locally regarding power/energy and cost with global considerations to improve environment and health. This aligns with a model of fast decision action locally and slower integrated learning to improve decision making.

Guidance can be taken from the human sensory/processing system, which generates ~10 Mbits/s via the body’s sensory systems but consciously only processes <50 bit/s for an overall “data to information bits” ratio of about 200,000:1. The brain continues to learn in the background at a slower rate to enhance the “sensing to action” in the foreground. Thus, we initially target a metric of similar magnitude achieving a reduction of “data” to actionable “information bits” of 100,000:1 or Data Reduction Ratio (DRR).

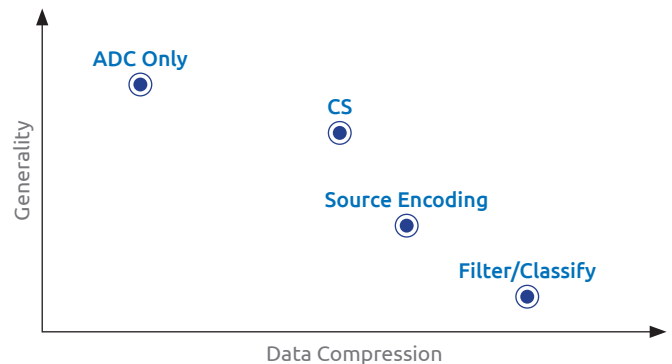


Figure 1.9: Data compression tradeoff (adapted from<sup>14</sup>)

**Data Reduction Ratio:**  $DRR = \frac{DataBits}{InformationBits}$

This is highly aggressive and will not be addressed for every sensing application.

Traditional signal compression is not enough. Compression typically targets reconstruction of the original signal and retains some application generality. For cases where reconstruction is necessary, such as video and audio entertainment or live video for remote medical diagnosis and surgery, this solution works well but has been limited to 10x–200x data reduction (Figure 1.9)<sup>14</sup>.

A paradigm shift in how sensed signal or “information” is processed is required in order to provide an output (analog or few bytes) of detected “actionable information” from the sensed signal. High understanding of the key action objective is needed, as well as the signal and the associated “detection entropy”—and thus certainty or robustness. In classical information theory, Shannon defines the “information entropy” metric as the absolute minimum amount of storage and transmission needed for succinctly capturing any information (as opposed to raw data)<sup>15</sup>. Here this concept is being extended to the minimum actionable output required to take action, which is detected from sensing. This output could be data bits (even a single bit) or an analog output signal controlling driving actuation. **To produce an actionable output, system knowledge will be required, as will consideration of added intelligence to all system components, from the sensor itself to analog signal processing, and possibly neural processing in analog and digital domains.** Therefore, overall co-design will be required, as highlighted in the “Basic Research Needs for Microelectronics” report published by the Department of Energy Office of Science Workshop in 2018<sup>1</sup>.

There is also the possibility to sense more things, especially with the use of more affordable electronics to perform spectroscopy. *The combination of different sensing modalities (sensor fusion) opens the possibility for better system optimization and, potentially, much better sensing capability.* The way humans interact with the world is also an area of increased attention and possibility. *Augmented reality and similar technologies will potentially create different ways for human-machine interaction, and sensing-to-action is one of the key enabling technologies.*

We are on the verge of exploding data from sensors—a data deluge. Currently the data is neither digestible nor practical to transmit any meaningful distance. The need for these sensors for a smart society is established, and there are multiple application trends growing both the number of sensors and the amount of data they produce. The amount of data must be decreased through intelligent reduction by moving toward a model of “Sensing to Action” to transmit minimal information bits. An objective of a Data Reduction Ratio of 100,000:1 on average has been set to address this. There are multiple areas of research needed to tackle such an aggressive target, and an initial list has been recommended.

**Grand Goal:** Analog-to-information compression with a practical compression ratio of 10<sup>5</sup>:1

The second session of the “New Trajectories for Analog Electronic” focused on “Intelligent Sensing: Sensing to

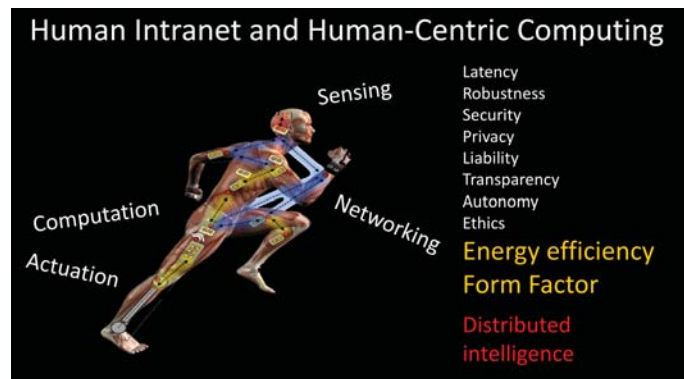


Figure 1.10: Human body internet and human-centric computing (courtesy of Jan Rabaey, UC Berkeley<sup>16</sup>)

Action,” directly aligning with the preceding discussion. Following is a summary of invited speakers’ key points, followed by areas for future research focus.

The keynote from academia<sup>16</sup> highlighted distributed intelligence, including the “Internet of Actions.” *Local and distributed processing enables and provides advantages in latency, energy efficiency, security/privacy, robustness, and autonomy.* This was highlighted via the human body model (Figure 1.10) and other biological systems, which are hierarchical but heavily linked control systems with multiple feedback paths. **The key point was to send only information that is needed, send it as slowly as possible, and process it locally where possible.** Examples were given of early sensing/processing technology, as well as processing with the most appropriate technology—analogue, digital and even chemical. Overall, communication costs are high and processing costs (energy, etc.) are significantly lower<sup>16</sup>.

The first industry panelist<sup>17</sup> discussed a mobile/portable AR/VR application that highlighted a need for local processing at very low power and with multiple sensing modalities to be effective. Latency and frame rate are critical for a natural human interface, which requires local performance. A key requirement to drive technology monolithically is integration for size and weight, in addition to using board, flex, and package integration methods such as a triple-stacked sensor example (pixel+DRAM+logic). **System co-optimization is necessary to satisfy the requirements stressing higher energy cost of data transfer versus processing<sup>17</sup>.**

The second panelist was from industry and focused on “accurate enough” sensing modalities and signal processing, as well as combining multiple sensors for robust decisions and actions<sup>18</sup>. Additionally, *“active sensing” and self-calibration of sensors* could provide additional performance and capability (Figure 1.11). High value was indicated for sensing and action in industrial/robotics, automotive, infrastructure, and health/medical



applications. Also highlighted was the very high “raw data” that sensors produce and necessary information reduction to action for efficiency. A specific example included automotive ADAS/ Autonomous multi-sensor images, where the detected object is the end result and orders of magnitude less “data”<sup>18</sup>.

The next panelist was from academia and discussed *compressive sensing*, where it works and where it does not work<sup>19</sup>. Data can be reduced via sparse/compressive sampling, but post-processing for reconstruction can be costly from an energy perspective. Drawing again on biology, sparse processing locally per the application need has value. The necessity for continued building-block optimal performance innovation for future sensing-to-action applications was stressed<sup>19</sup>.

A fourth panelist, from industry, presented a highly integrated SoC for IoT, which is required for small form factor wearables and low power<sup>20</sup>. *The need for an “always on” processor to*

*offload the main processor*, allowing duty cycling for power savings, was highlighted. Customization of the technology for the application was again stressed, including hardware accelerators for ML that are customized for low power and object detection or other functions. NVM and compute in memory were highlighted as key technology needs<sup>20</sup>.

The final panelist was from academia<sup>21</sup> and highlighted the need for effective and efficient human interfaces to therapeutically address medical conditions. The requirements include local processing, low power, portability, and human body compatibility. Learning and adaptability are key capabilities to make the solution personal and effective. There are very significant challenges to “nerve interface” in the brain but opportunities for in-ear sensors as alternatives to probes while still providing effective “information”. Reference was also made to human biological processing as a model to improve such interfaces and therapies<sup>21</sup>.

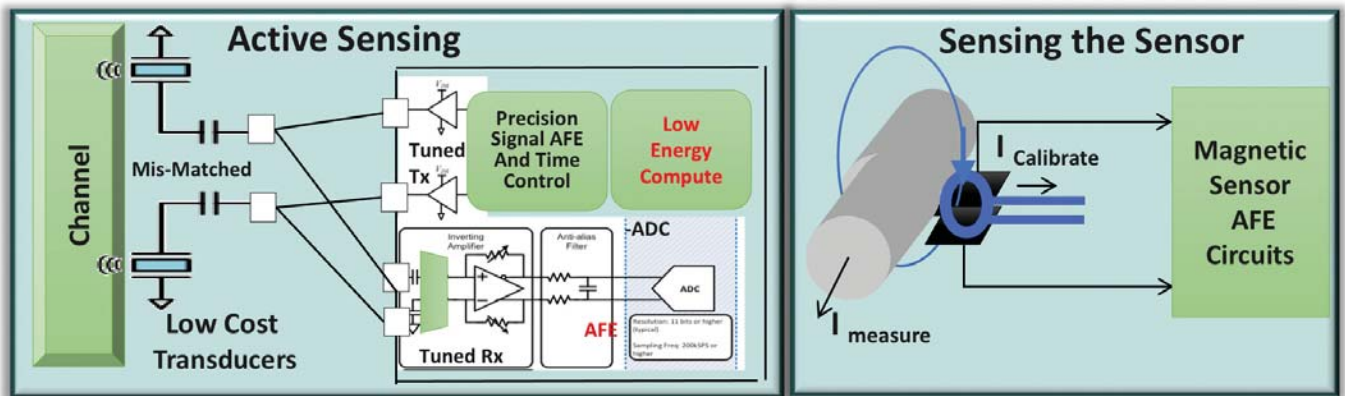


Figure 1.11: Accurate enough active sensing and calibration (courtesy of Baher Haroun, Texas Instruments<sup>18</sup>)

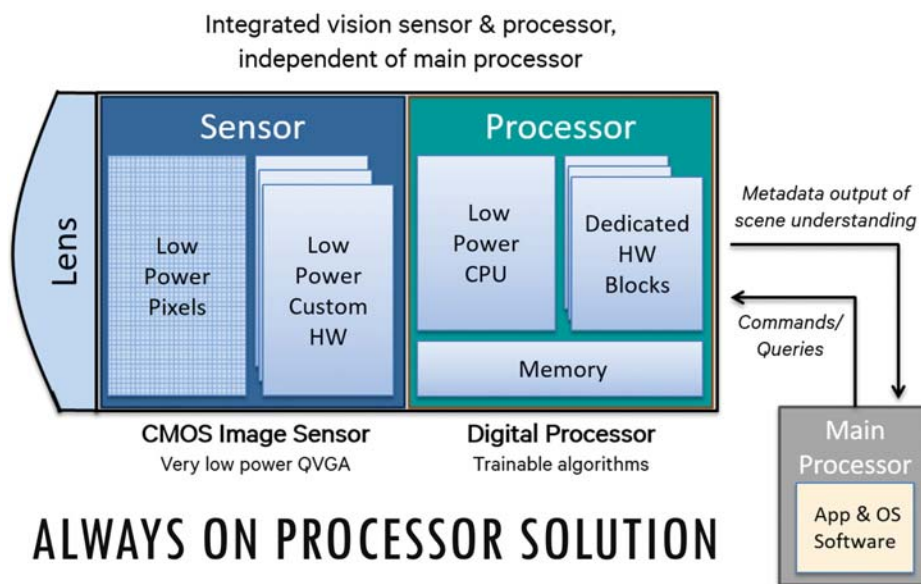


Figure 1.12: Always-on processor (courtesy of Rashid Attar, Qualcomm<sup>20</sup>)



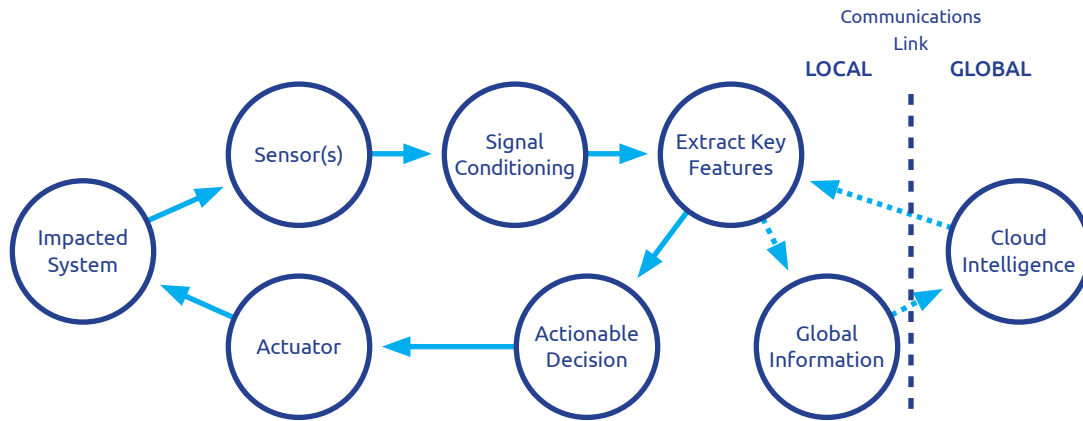


Figure 1.14: Local and hierarchical intelligent sensing

Fig. 1.14 presents an overview of a sensing-to-action system that provides a context to the open questions/challenges outlined below and to the research needs in the following subsection. **The open questions/challenges include:**

- “Trillions” of sensors generate redundant and unused “data.”
- Cloud is not the answer.
  - Communication is a bottleneck.
  - Power to process redundant data is not efficient
  - Latency is too long for local control and action
- Intelligent Sensors are needed to drive local and timely action.

### Key Areas of Focus and Follow-on Research

Systems solution and co-design approach: This holistic approach is recommended for the most robust, compact, energy-efficient, and cost-effective solution. This will require:

- Intelligent sensors and sensor-fusion research—multi-sensor distributed intelligence
- Applications and system knowledge research
- Hierarchical and distributed exploration/optimization
- Collaborative multi-expertise research projects—Moon Shot demonstrator platform
  - Common goal/objective to drive analog technologies—can spin out technologies to other applications
- System approach to optimization that crosses boundaries—sensor, analog processing, digital processing, ML/detection, etc.
- Research most appropriate domain for signal capture—energy and reduced sensor data rates

Leverage human systems as a model: Significant discussion evolved around leveraging of human systems to provide only what is needed with minimal communication via a hierarchical sensing-to-action solution. This will require:

- Local sensing to action for many applications, including “selective” or “detecting” smarter sensors that minimize further signal processing and power
- Learning and adaptive solutions for improving both accuracy and efficiency
- Understanding “minimum” performance needed for robustness—SNR, resolution, number of bits, etc.
- Local feedback (not to cloud) for efficient and timely sensing to action
  - Overlays with ML at the edge and Analog Machine Learning session
  - Heterogeneous sensing—combination and multimodal sensing fusion
- Research how to have always-on “early detection” of anomaly in system for further processing
  - ML may determine “normal” and set thresholds for anomaly detect
- Increasing analog designer’s knowledge and understanding of biology and its interface to our physical world in order to make a substantial and meaningful impact

Flexible, scalable platform and technology: Due to the wide range of sensing-to-action applications, there is a need for a flexible and scalable platform that addresses efficiency (power and cost). This will require:

- Technologies: memory, sensors, domain matched to signal, local, power efficiency
- Advanced building blocks research for performance and/or ultra-low power
  - ADC, PLL, DAC, VCO, high speed links, near zero power
  - New architectures—time-based, ring amps, noise-shaping SAR, Switch Cap PA

- Array-based processing—multi-sensor, multi-modes, beam processing, forming/detecting
  - Can address low-performance sensors with multiple to improve performance (SNR)
- Multi-sensor design and feedback to sensor for improved performance, redundancy, and safety
  - Heterogeneous sensing—combination and multimodal sensing fusion
- New devices for high performance or alternate architectures
  - Analog ML in RRAM, MRAM, or other existing and emerging technologies
  - Digital assist w with new advanced digital devices
- Analog Mixed Signal Machine Learning and/or CMOS neural networks—also highlighted in the Analog in Machine Learning at the Edge session
- Leverage analytics/machine learning for analog circuit design—faster, more optimum, less error prone. This is required to be able to quickly address multiple applications with predictable and reliable design—also highlighted in the Analog Design Productivity and Predictability session

#### Additional topics

- **Secure sensors:** Security and privacy were raised as a need (covered in Chapter 4). Local sensing-to-action improves security simply by reducing the number of attackable interfaces and communications ports.
  - How to detect if sensor is “spoofed”—part of anomaly detection
  - Need for methods to balance privacy and security with the public good—especially when associated with cyber-biological interfaces and sensing
- **Advance human-machine interface:** In order to be more autonomous, local, and mobile, new human interfaces will be needed, especially for on-body and intra-body applications.
  - For VR/AR—not today’s keyboard/mouse/screen
  - For health—on-body or internal sensing and action—including neural interface
- **Moon Shot collaborative projects:** leverage low-cost parts for system solution
- E-nose or E-taste intelligent sensor
- Active biomedical: modulate treatment based on sensing (Parkinson’s, pancreas, cancer, Alzheimer’s)
- Wearable AR/VR all-the-time solutions—FB vision

## 1.4. Analog in the THz & Optical regime

### Overview and Needs

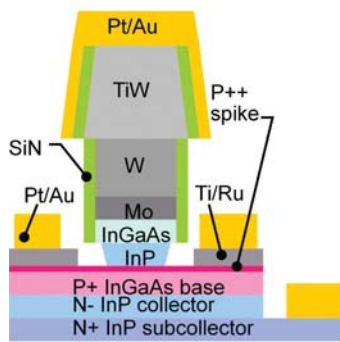
Analog interfaces continue to expand in level of performance to enable new and important communications and sensing modalities. Overall communications challenges are covered in the specific Communications chapter later in this report. ***One vector is toward higher frequencies from 10s of GHz to THz and optical wavelengths for improved sensing applications.*** Electromagnetic waves typically in a range of 300 GHz to 3 THz and above classify as Terahertz (THz) waves. There are many advantages to expanding the use of their shorter wavelength. For communications, it would mean greater spatial multiplexing and parallel channels. More importantly, the massive bandwidth in this portion of the spectrum can be used for high data-rate wireline and wireless communication. For imaging, it would mean finer spatial resolution, enabling applications like through-wall imaging, non-destructive evaluation for detecting manufacturing defects, and highly automated driving/navigation in poor visibility conditions.

New challenges arise to address the deluge of data at such high frequencies, including devices, interconnect, power, linearity, noise, time resolution/noise to packaging, antennas/interface, interference, and signal processing. The Analog in the THz and Optical regime session focused on the key value and challenges for analog device, circuit, and system solutions, addressing analog interface future modality requirements and application needs.

### THz integrated circuits and applications

**Research on transistors that operate in the 100 GHz to 2-3 THz range is gaining momentum. The main challenge is to reduce cost and increase market size.** Bandwidth will increase for the SiGe, InP HBT, InP HEMT, and GaN variety but is likely limited for CMOS, wherein further scaling by shortening gates and use of FinFETs is providing diminishing returns<sup>22</sup>.

It then becomes necessary to consider technologies other than CMOS. *InP Bipolar Transistors, for example, give greater electron transport ( $v_d = 3.5 \times 10^7$  cm/s) than Silicon ( $v_d = 1 \times 10^7$  cm/s).* Also, InP HBTs have wider bandgap and, hence, higher breakdown field. **Figure 1.15** gives better context of the structure of the InP HBT and tabulates strategies to increase bandwidth. **Table 1** gives a scaling roadmap for the InP HBT that leads to ultra-low resistivity contacts<sup>22</sup>. It is prudent to recognize tradeoffs at advanced nodes. For instance, at the



| to double the bandwidth:              | change         |
|---------------------------------------|----------------|
| emitter & collector junction widths   | decrease 4:1   |
| current density (mA/μm <sup>2</sup> ) | increase 4:1   |
| current density (mA/μm)               | constant       |
| collector depletion thickness         | decrease 2:1   |
| base thickness                        | decrease 1.4:1 |
| emitter & base contact resistivities  | decrease 4:1   |

Figure 1.15: InP HBT structure and scaling strategies to double bandwidth (courtesy of Mark Rodwell, UC Santa Barbara<sup>22</sup>)

Table 1: InP bipolar scaling roadmap

| Emitter             |      |   |      |                          |
|---------------------|------|---|------|--------------------------|
| Junction width      | 128  | → | 64   | → 32 nm                  |
| Access resistivity  | 4    | → | 2    | → 1 Ω-cm <sup>2</sup>    |
| Base                |      |   |      |                          |
| Contact width       | 128  | → | 64   | → 32 nm                  |
| Contact resistivity | 5    | → | 2    | → 1.15 Ω-cm <sup>2</sup> |
| Collector           |      |   |      |                          |
| Thickness           | 75   | → | 53   | → 38 Nm                  |
| Current density     | 18   | → | 36   | → 72 mA/μm <sup>2</sup>  |
| Breakdown           | 3.3  | → | 2.75 | → ~2 V                   |
| $f_c$               | 730  | → | 1000 | → 1400 GHz               |
| $f_{max}$           | 1400 | → | 2000 | → 2800 GHz               |
| Digital M/S latch   | 330  | → | 480  | → 660 GHz                |

64 nm/ 2 THz and 32 nm/ 3 THz node, there is a need for higher base contact doping for greater  $\beta$  and for moderate contact penetration. It is possible to employ base regrowth using thin, moderately doped intrinsic base InGaAs or GaAsSb with a carrier concentration of about  $10^{19}/\text{cm}^3$ . By current process runs, it is also seen that GaAsSb intrinsic base is resistant to hydrogen passivation of carbon base dopants.

**A similar prognosis for improving InAs MOS-HEMTs reveals that these transistors will run into scaling limits in terms of gate insulator thickness and source access resistance. These may limit  $f_T$  to about 1200 GHz and  $f_{max}$  to about 3000 GHz.**

## CMOS platforms for THz imaging, sensing, and communication

**There are challenges in realizing high-performance THz circuitry in CMOS technology.** The highest unity current gain frequency,  $f_c$ , and unity maximum available power gain frequency,  $f_{max}$ , of NMOS transistors fabricated are around 280 and 320 GHz, respectively, in 45 nm technology<sup>23</sup>. Interconnects to top metal layers significantly reduce

performance by introducing parasitic capacitance, resistance, and inductance. Also, reducing supply voltage with the technology scaling makes generation of a sufficient power level more difficult. Nevertheless, the nonlinearity of components like Schottky diodes and MOS varactor diodes<sup>24</sup> with cutoff frequencies over 2 THz has been exploited in an attempt to operate in this frequency regime. Further encouraging are findings that efficiency of on-chip patch antennas realized in a 130-nm CMOS process with a 2 mm thick top metal layer and a total dielectric thickness between silicon and the aluminum bond pad layer of 7 mm actually improves to ~80 percent at 1 THz from ~30 percent at 300 GHz<sup>24</sup>. The highest output power level of CMOS circuits is shown to be -1 dBm at 300 GHz. Also, III-V devices generate 5–15 dB higher power over the same frequency range. Interestingly, a cascade of a symmetric MOS varactor frequency quintupler and an asymmetric MOS varactor frequency doubler can be used to generate -23 dBm at 1.3 THz and is only 5 dB less than that of its III-V circuit counterpart<sup>24</sup>. This suggests a possibility to reduce the gap between the output power levels of CMOS and III-V circuits at this frequency regime.

Receivers can be categorized into either incoherent receivers, which detect only the amplitude of input signals, and coherent receivers, which detect both the amplitude and phase. The lowest measured noise figures for CMOS receivers are 9 dB at 200 GHz, rising to 16 dB at 305 GHz, and they are 4.2 dB and 7.7 dB higher than receiver circuits using III-V devices at the respective frequencies. At frequencies above 300 GHz, noise figures of CMOS and SiGe HBT circuits increase due to higher order subharmonic mixing. Also, the signal strength of the integrated frequency multiplied local oscillator at these frequencies is insufficient to properly switch devices, and this again degrades conversion loss and noise figure. *However, this problem can be solved by using a separate high-power-high-efficiency frequency multiplier and by lowering the order of sub-harmonic mixing.*

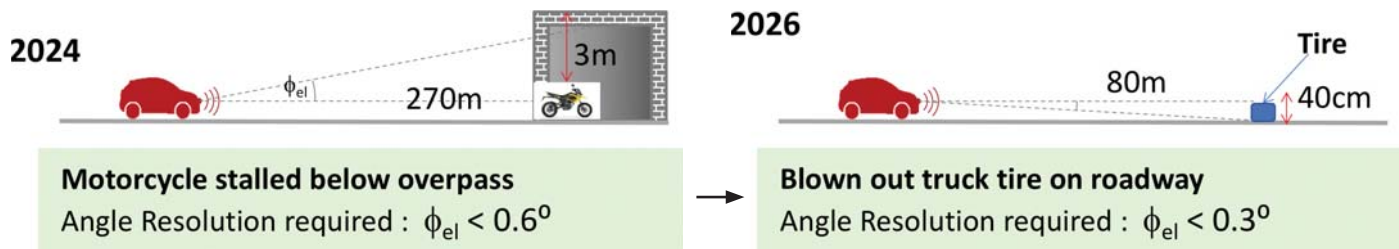


Figure 1.16: Short-term performance targets for High Elevation Angle Resolution in automotive radar (courtesy of Brian Ginsburg, Texas Instruments<sup>30</sup>)

THz transceivers have many everyday applications. A THz imaging system could typically include a pixel array, analog multiplexers, an amplifier bank external to the pixel array, and column and row decoders. A pixel consists of a patch antenna, a diode connected transistor for detection, a double-stub matching network, and an access transistor<sup>25,26</sup>. It was seen that the average responsivity and NEP of the imaging array including the amplifiers was 2600 V/W and 37 pW/√Hz, respectively, with a minimum NEP of 13 pW/√Hz<sup>25</sup>. *This is particularly useful to consider when designing navigation systems in autonomous vehicles that encounter operational challenges in visibly impaired conditions.*

*THz receivers can also be used in gas sensing by rotational spectroscopy<sup>27</sup>.* To ensure that lines of different gases do not overlap, the receiver needs to detect weak energy absorption in the presence of much larger baseline variations, instead of detecting a small signal in the presence of noise. Here, a frequency modulated signal is used to enable detection of small absorption dips. Finally, wireline communication over a dielectric waveguide is being developed to mitigate the complexity of high data-rate communication over copper wires<sup>28</sup>. Frequency-division multiplexing (FDM) and polarization-division multiplexing (PDM) can be simultaneously used to increase the data rate over a given bandwidth<sup>29</sup>.

## Automotive and Industrial Radar

*Increased resolution of automotive radar, now in the 77 GHz regime<sup>30</sup>, has been integrated into algorithms that guide autonomous control of vehicles, including Blind-Spot Detection (BSD), Adaptive Cruise Control (ACC), Lane-Change Assist (LCA), Cross-Traffic Alert (CTA) and Autonomous Emergency Braking (AEB).* Figure 1.16 shows graphically two target milestones in terms of angular resolution for obstacle detection. It is derived that for a 0.3° radar, a 70x70cm array would need as many as 122,500 antennas and requires MIMO or advanced algorithms to have feasible hardware complexity.

**While moving to higher frequencies can achieve the target angular resolution in a smaller form factor, it comes with a**

**penalty in Signal-to-Noise Ratio (SNR)**, which is contingent on output power, antenna gain, wavelength, noise factor, and the target's radar cross-section, among other metrics. Complicating this is the fact that the reflected signal from a target drops as  $1/R^4$  over range R while interfering power from a single radar drops as  $1/R^2$ . It is also important to optimize application-specific data rates for SIMO and MIMO systems. It behooves designers to consider that low-level sensor fusion will require moving 10s-100s of Gbps of radar, camera, and Lidar data to high performance fusion nodes in the near future, and that such high-traffic signal processing should be compatible with passive cooling and IC packaging.

## Analog Beamforming Antennas for 5G

**As 5G is being implemented, we find that each facet of interconnectivity requires its own communication protocol.** To explain further, mobile broadband requires capacity of 10 Tbps/km<sup>2</sup>, Internet of Things requires ultra-high density of 1 million nodes/km<sup>2</sup>, and so on. A 5G analog beamforming antenna array would have to be accurately designed to have the electrical length from the central feed point to each antenna patch be identical in terms of phase and loss. It becomes necessary to evaluate if mm-Wave communication merits the cost of these dense-phased arrays. **There are also associated challenges of IC fabrication, testing, and packaging (particularly, the avoidance of cracks in solder bumps).** In operation, even in the worst case of when 4 channels add coherently, a -60 dB isolation is required with a input-output parasitic capacitance of less than 10 fF<sup>31</sup>.

*To make dense edge-based signal processing arrays, a harder look at electroforming, casting, and 3D printing technologies is needed.* Similarly, for dense broadside arrays, dense control I/O and better thermal design are imperative. It may also help to explore the variations in output power versus the number of antenna elements for different device technologies like GaN, SiGe, SOI/CMOS, etc. All these factors constitute an approach of codesign between antenna design, IC design, and waveform engineering.



## Integrated Silicon Photonics for Communication and Sensing

A key component in integrated photonics is the optical ring resonator. The setup generally consists of an input linear waveguide, the light coupled to a ring waveguide that acts as an optical resonator at resonant wavelengths. The strength of coupling depends on their separation and respective refractive indices. The light of built-up intensity may be coupled to an output bus waveguide, which serves as a detector. The system, therefore, behaves as an optical filter with its appropriate transfer function. Such a transfer function is shifted in and out of resonance by depleting carriers out of the ring resonator. One nice development has been the use of *interleaved junctions in the transmit modulator around the ring*. This structure (**Figure 1.17**), enabled by advanced lithography, helps to fully deplete the junction of carriers at low voltages like 1-1.5 V for the largest frequency shift. The result is highly sensitive structures with quality factors of up to 200,000<sup>32</sup>. This can be driven with CMOS logic inverter (1.2 Vpp) to enable a 5 Gb/s data rate at a 3 pJ/bit optical energy cost<sup>33</sup> and about 3 dB insertion loss.

For communication, if the system is to function as a transceiver, it is necessary to account for insertion loss from input laser to photodetector and associated modulation loss. When accounting for electrical overhead in terms of Energy/bit, design must consider power consumption by the serializer, photodetector capacitance in the receiver, etc. Also, **there is the challenge of increasing voltage requirements for effecting sufficient photons per bit at higher data rates**. Plus, with every new ring added, there is a cost in thermal tuning.

Analog-photonic links are now used to simplify mm-wave node IC architectures and can handle up to 1000 antennas per chip. **They have also greatly increased energy savings in the process and are vastly more energy-efficient than the digital links in the signal chain**. **Figure 1.18** shows a breakdown of comparative power consumption in the constituent (digital and analog/photonic) architectures.

### Key Areas of Focus and Follow-on Research

Device exploration for THz solutions: Key to THz application is transmission and reception of signals in the greater-than 100GHz through low THz range of frequencies. This requires power efficient gain and detection of signals with appropriate linearity and noise via “devices” (transistors typically). This will require:

- Fundamental limits analysis of multiple technologies (i.e., CMOS, GaN SiGe, InP, InGaAs, and beyond) for THz application – receivers/transmitters

- Benchmark comparison of technologies
- System performance implications study

CMOS (or future other) platform integration: Single device performance is part of the solution where additional platform processing and interface would be required. CMOS is predominant today. Integration of specific THz active devices and passives needs exploration and optimization.

- Integration potential and limits of technology
- Optimization function split – platform or other technology for system solution
- Optimization of platform for THz applications (CMOS or future other)

THz array system solutions: It is clear that future THz imaging and communications systems require some type of beam capability for gain and selectivity for SNR and resolution. Architectures that are power efficient, cost competitive, and achieve performance targets will be required, and the optimum split is dependent on technology choice.

- Optimum array per application—frequency, BW size, and signal distribution
- Processing architecture—distributed, aggregated, analog, and digital
- Frequency limits for imaging—range and resolution
- Accurate, low-noise, and tightly synchronized timing solutions/clocks

Silicon photonics integration and application: Silicon photonics holds promise of extending frequency and bandwidth for multiple applications, from sensing to interconnect to communications. Challenges of integration and power/area efficient optical-electrical-optical conversion and interface continue and need to be addressed to make broader application of this technology.

Packaging and heterogeneous integration: At THz frequencies and beyond (optical), interfaces to the outside world are challenged by parasitic effects from traditional resistance, capacitance and inductance to fringing, roughness, and index of refraction at multi-GHz frequencies. The parasitic effects have severe effects on transmission loss and reflection impacting efficiency and SNR. Additionally, integration of multiple technologies will be required and may not be possible on a single chip and/or require very closely couple passives. Architecture optimization covering device through interconnect and packing, plus coupling to the package, is a key area of research need.



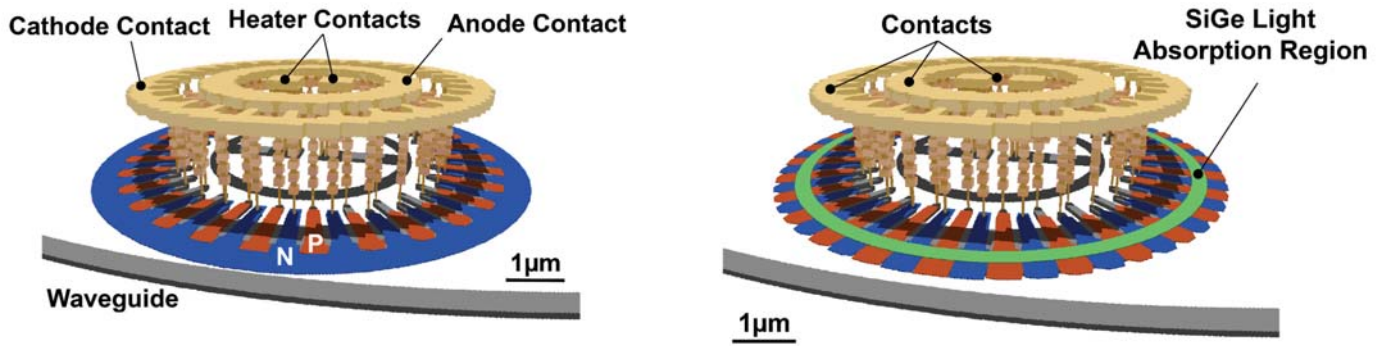


Figure 1.17: Two examples of interleaved junction structures in optical ring resonator transmitters (courtesy of Vladimir Stojanovic, UC Berkeley<sup>32</sup>)

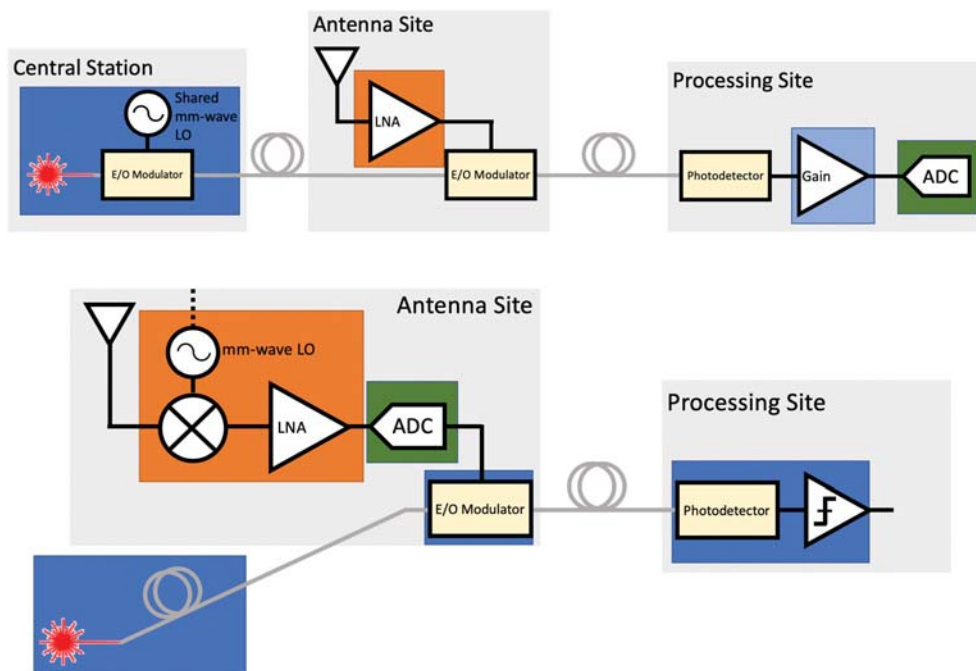
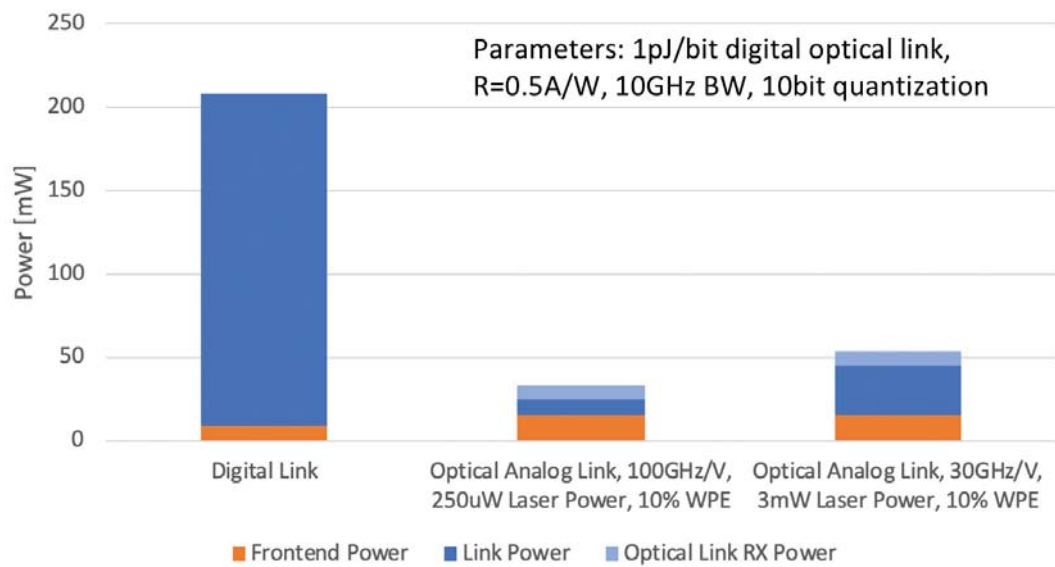


Figure 1.18: Power breakdown between analog/photonic and digital architectures in the signal chain (courtesy of Vladimir Stojanovic, UC Berkeley<sup>32</sup>)

# 1.5. Analog in Machine Learning at the Edge

## Overview and Needs

*Machine Learning (ML) is expected to be one of the next major disruptive technologies.* It will affect the way we access and analyze information, the way we teach, and the way we learn. **However, the current state of ML is characterized by an extensive use of high-performance computational resources, with memory footprints, compute loads, and energy costs that are all quite large.** Unfortunately, all implementation environments—from datacenter to network edge—are subject to significant resource constraints. Likewise, specialized systems to support critical applications, such as autonomous driving, require interaction and communication of multiple components and can approach datacenter complexity. *The use of analog techniques in ML at the edge could offer new solutions to the energy constraints and complexity/footprint challenges that occur where autonomous or local intelligence and decision are necessary.*

**While ML is going to be a huge driver for improving the computational power of systems to be able to run more complex algorithms, it will lead to an increasing complexity of system design and very high power consumption by the machines.** For example, the Sunway TaihuLight supercomputer in China achieves 93 petaflops consuming 15,371 Megawatts, nearly enough to power a small city<sup>34</sup>. In contrast, the brain is one of the most energy-efficient compute systems, which can achieve around 38 petaflops at just 20 Watts operation power<sup>34</sup>. It has a complex architecture

with an intertwined memory and appears to “compute” based on highly simplified analog and digital operations. Training at the edge allows for adaptation to local conditions, which may result in energy, speed, and area benefits. *The main push is for in-memory computing with analog memory.*

Edge computing is a distributed computing topology where information processing is located close to the source of information, bringing computation and data storage closer. Because communication bandwidth is limited and privacy is an utmost concern, all the data cannot be sent to the cloud. Edge computing is beneficial in analyzing data that is collected by sensors at remote locations. Using digital processing devices, the sensors’ data would have to be run through an ADC before use. The use of analog techniques in ML at the edge could offer new solutions to the energy constraints and complexity/footprint challenges that occur where autonomous or local intelligence and decision are necessary. For example, performing inference at the edge would dramatically compress the amount of information that has to be sent downstream. Training is typically conducted at the cloud now (Figure 1.19). *Training at the edge would enable new functionalities, e.g., real-time adaptation to local conditions, resulting in significant energy and speed benefits.*

This section will highlight the challenges and explore possible research directions leveraging analog components, devices, and systems for such ML environments.

|           | Importance<br>(world total in 2020,<br>Source: NVIDIA) | Key differences<br>in HW specs  | Where<br>deployed<br>now                                 | Primary HW<br>metric                                 |
|-----------|--|---|--|--|
| Training  | ~55 ExaFLOP/s  | Medium-to-high<br>(8+32 bit) computing<br>precision   | cloud  | Throughput per<br>chip area                          |
| Inference | ~450 Exa<br>Integer OP/s                               | Low-to-medium<br>(4+8 bit) computing<br>precision<br><br>Persistent<br>(nonvolatile)<br>weights | cloud<br><br>edge<br><i>near term target application</i> | Throughput per<br>chip area<br><br>Energy efficiency |

Figure 1.19: Machine learning/neuromorphic computing applications. (Performance numbers are from<sup>35</sup>)

## Analog In-memory Computing

The critical operation performed by any neuromorphic network and, more generally, many machine learning tasks is a vector-by-matrix multiplication (VMM). *Such operation can be very efficiently implemented with analog circuits utilizing the fundamental Ohm and Kirchhoff laws (Figure 1.20).* This circuit's main component is an analog memory cell with adjustable conductance  $G$ , used at each crosspoint of a crossbar array and mimicking the biological synapse. VMM circuits based on dense emerging analog memories can be remarkably compact, leading to superior speed and energy efficiency. Additionally, dense analog VMM circuits may allow storing all weights locally on a chip, thus dramatically reducing data communication overhead, e.g., moving data in and out from the off-chip memory, which would be typical for digital implementations. This is especially important since many machine learning / neuromorphic computations are both compute- and data-heavy.

Figure 1.20 shows one specific “current-mode” flavor of analog VMM circuit in which inputs and outputs are encoded in instantaneous voltage and current magnitudes, respectively. Other implementations have been suggested, e.g., *time-mode VMM circuits in which inputs and outputs are encoded in time duration of fixed-amplitude voltage pulses, or hybrid VMMs in which bits of input vector elements are applied sequentially, bit by bits, and the outputs are computed by properly accumulating partial results according to the bit significance*<sup>36,37</sup>. Each approach has its cons and pros. For example, variable amplitude encoding may require larger peripheral circuits, while the time-domain approach's main drawback is exponential scaling of latency with the computing precision.

Potential advantages of analog VMM circuits for neuromorphic computing had been recognized several decades ago. However, up until recently, such devices were implemented mostly as “synaptic transistors”, which may be fabricated using the standard CMOS technology<sup>38</sup>. This approach was used to implement many sophisticated, efficient systems—see, e.g.,<sup>39</sup>. However, **these devices have relatively large areas ( $> 10^3 F^2$ , where  $F$  is the minimum feature size), leading to higher interconnect capacitance and hence larger time delays.** The recent advances in memory devices, e.g., based on metal oxide and solid-state electrolyte resistive switching, phase change, magnetic, and ferroelectric materials<sup>49</sup> opened up new opportunities for analog computing. Some emerging memories can be stacked vertically, achieving less than  $4F^2$  effective footprint. *Commercial NOR flash memories are a viable candidate for in-memory analog computing in the near term due to the maturity and accessibility of such technology, while 3D NAND flash memories present intriguing prospects due to their very high density.* For example, recent work shows that inference accelerators based on commercial embedded NOR flash memories, redesigned for analog operation, may dramatically increase the performance and energy efficiency of neuromorphic systems<sup>40</sup>, while modeling shows prospects of reaching fJ/op operation at the chip level<sup>41</sup>.

The Landauer's limit defines minimum energy for computation as  $kT$  times the change in information entropy due to the computation, irrespective of the type of device. (Here  $k$  is a Boltzmann constant, and  $T$  is the temperature in Kelvin)<sup>42</sup>. Landauer's principle is widely understood as an endpoint for digital scaling but hasn't been applied to analog computers.

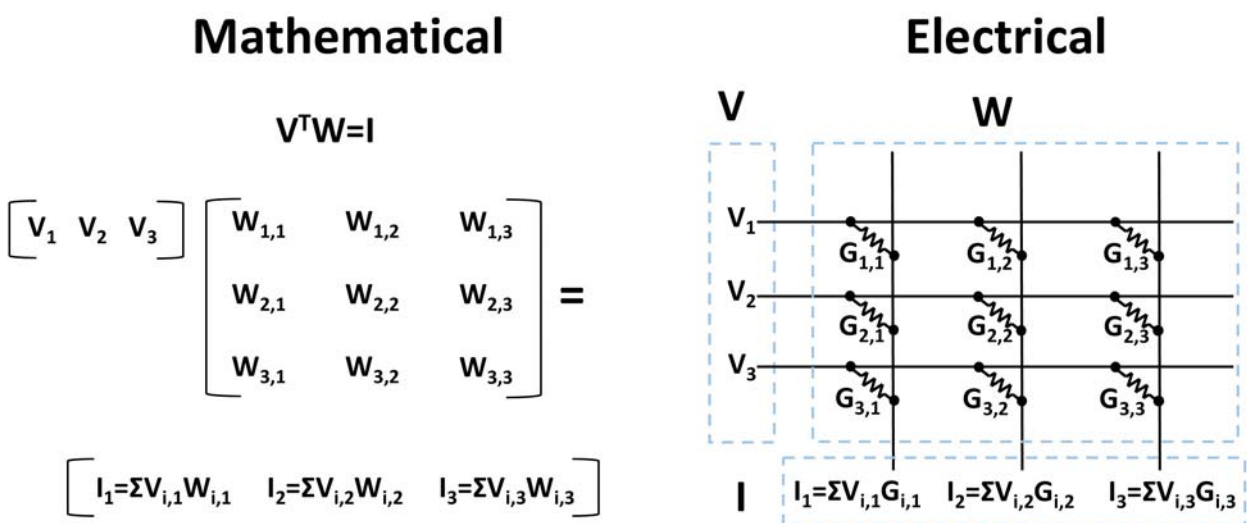


Figure 1.20: Resistive crossbar as an example of computing in memory crossbar computation for electronic vector matrix multiply (courtesy of Matthew Marinella, Sandia National Laboratories<sup>43</sup>)

## Analog and Mixed-Signal Architectures

Purely analog computing is possible for some systems and would be desirable to achieve the best speed and energy efficiency. *One example of such a system is in-sensor computing with image sensors feeding densely interconnected deep multilayer perceptron networks.* However, many neural networks and machine learning models rely on weight reuse. In a weight-stationary (i.e., in-memory) computing, weight reuse implies either a less efficient approach of providing many copies of the same weight or more compact temporal multiplexing of one copy of the weight. The latter requires storing intermediate results after applying the same set of weights to different portions of the input. Because digital circuits are more suitable for implementing intermediate memory, mixed-signal circuits are used for general purpose neuromorphic accelerators. The need for mixed-signal architectures is also driven due to less common tasks in machine learning/neuromorphic computing, which cannot be efficiently implemented in the analog domain.

Several architectures have been developed and optimized for a broad class of ANNs, including CNNs. Examples are the Programmable Ultra-efficient Memristor-based Accelerator (PUMA)<sup>44</sup> and the aCortex<sup>40</sup>. In such architectures, memristor crossbars perform matrix-vector multiplication and are connected via ADC/DAC to digital circuits used to implement other operations, instruction decoder, and instruction memory. The tunability of nonvolatile memories allows reducing the overhead of process variations in analog circuits. ADC/DAC overhead can be avoided with parallel transport of duration vectors in time-domain architectures<sup>36</sup>. An algorithm/HW co-design is used to achieve better physical performance at the same functional performance per system-level task.

**The main challenge for inference and training applications is variations in memory cell I-V characteristics.** However, the requirements are more relaxed for inference due to the utilized feedback tuning algorithms. Additionally, there is a need to reduce write and read currents for memristors (RRAM) and PCRAMs. This would allow lowering the area and energy overhead of peripheral circuits and access transistor overhead for 1T1R arrays. Prior work showed that most neural network inference operations could be performed with 4- to 8-bit weight/activation precision. The other important memory device metrics for inference applications are density, especially for energy efficiency (EE)-optimized designs, multi-level memory, and high retention. For most inference applications, the weights would be changed infrequently and remain stable over an extended period.

**The computing precision requirements are much higher for training. Additionally, a highly linear and gradual conductance update rate and high endurance are required for training applications.** However, the retention characteristics are much more relaxed since the weights are updated frequently. A further challenge is in debugging of issues caused by in-situ training. This may require knowledge of the final value of the weights and full evolution of the system throughout the entire lifetime to fully understand how and why a particular system instance “learned” to do something unexpected or undesired.

Prior work on mixed-signal neuromorphic inference accelerators showed that system-level performance in EE-optimized designs is mainly limited by memory density. The high area overhead of ADC/DAC can be reduced by sharing and time-multiplexing these circuits, resulting in better EE at the cost of lowering computational throughput. This leads to a natural tradeoff between computational throughput

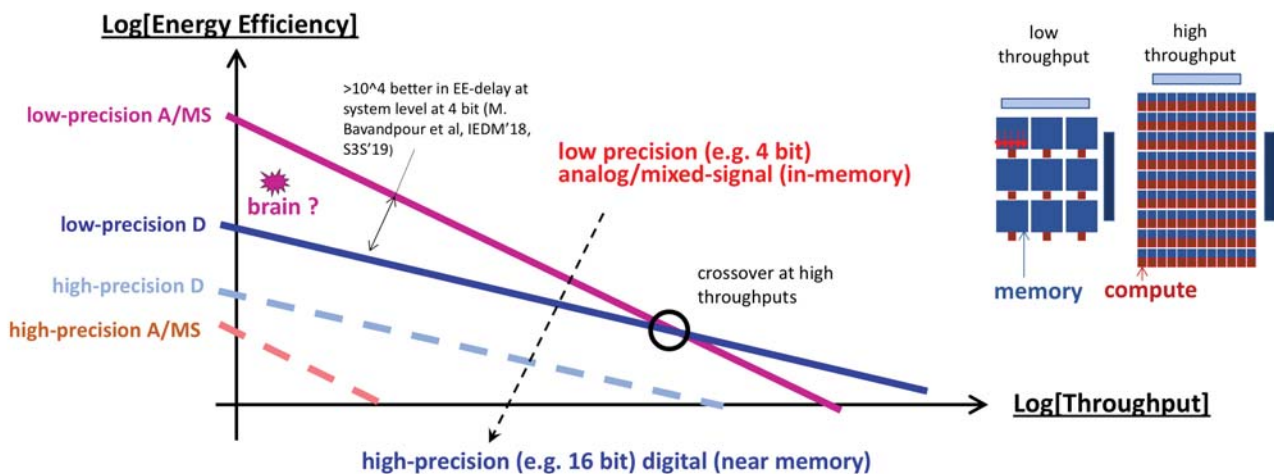


Figure 1.21: Comparison of mixed-signal and digital accelerators (courtesy of Dmitri Strukov, UC Santa Barbara<sup>45</sup>)



and energy efficiency (Figure 1.21). Digital accelerators are better suited for high-precision applications, while mixed-signal accelerators are adequate for low to medium levels of precisions. The gap between mixed-signal and purely digital accelerators is less when there is a significant amount of weight reuse since the high cost of retrieving each weight can be amortized across many computations in digital architecture.

*Neuromorphic inference accelerators could be the first target for analog in-memory computing due to their significance* (see, e.g., expected number of operations performed per second for inference applications in Figure 1.19) and simpler memory device requirements.

### FPAAs, Neuromorphic ADCs, and Stochastic Analog Computing Circuits

Large-scale field-programmable analog arrays (FPAA) consist of programmable and configurable analog and digital components<sup>46</sup> (Figure 1.22). These are made possible by floating-gate transistor circuits<sup>38,47</sup>. *FPAA will be able to aid in the growth and development of neuromorphic computing at the edge*. For example, FPAAs have already been used to implement the ultra-low-power sensor-to-end-result speech classification<sup>50</sup>. Another exciting application of FPAAs is in solving differential equations by constructing analog circuits with matching underlying dynamics<sup>48</sup>.

*Analog to digital converters (ADCs) are an essential component in mixed-signal circuits* (Figure 1.23). ADCs need to be fast,

reliable, and capable of rapid gain-scaling in the edge systems while also being both low power and low cost. The main advantages of making trainable neuromorphic ADC is that they are generic and flexible—logarithmic ADCs are an excellent example. These are suitable for optimization, have an excellent figure of merit, and can self-calibrate based on the application. Additionally, training can compensate for the non-linear characteristic though monotonicity is still required.

One of the defining characteristics of analog computing is noise. While noise is typically detrimental for analog circuit performance, it can be utilized in some applications. *An excellent example is a combinatorial optimization which is used in power grid applications, electronic design automation, logistics, and molecular dynamic simulations*. A promising approach for solving optimization problems is to use generalized Hopfield neural networks. With adequately selected synaptic weights, such network converges (in the ideal case) to its minimum energy state, which corresponds to the solution of the programmed optimization problem. Annealing techniques are commonly used to escape local minima and improve performance, e.g., by probabilistically updating neuron states. Therefore, the efficient hardware for the generalized Hopfield network should efficiently implement not only VMM operation but also stochastic neurons to support metaheuristic techniques. *Mixed-signal circuits based on analog-grade nonvolatile memories that utilize noise have been proposed to implement such stochastic functionality*<sup>49</sup>.

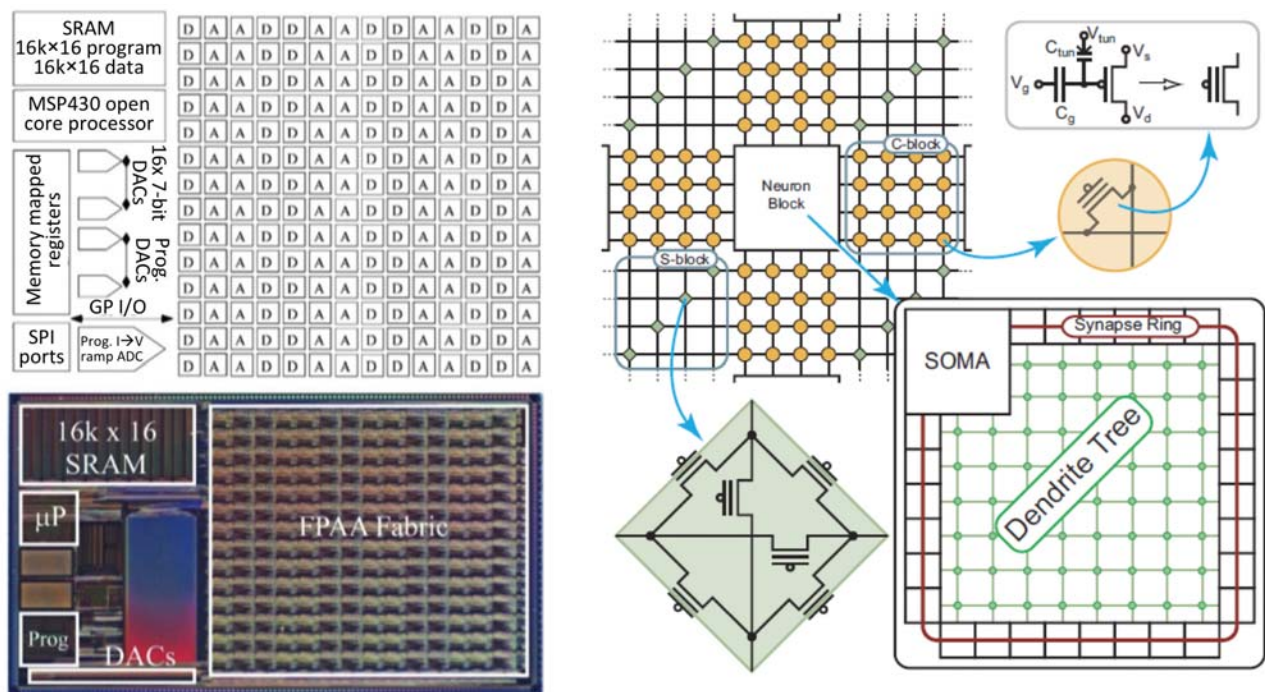


Figure 1.22: Example of FPAA system: General architecture, chip layout in 350-nm process and analog and digital block architecture (adapted from<sup>50</sup>)

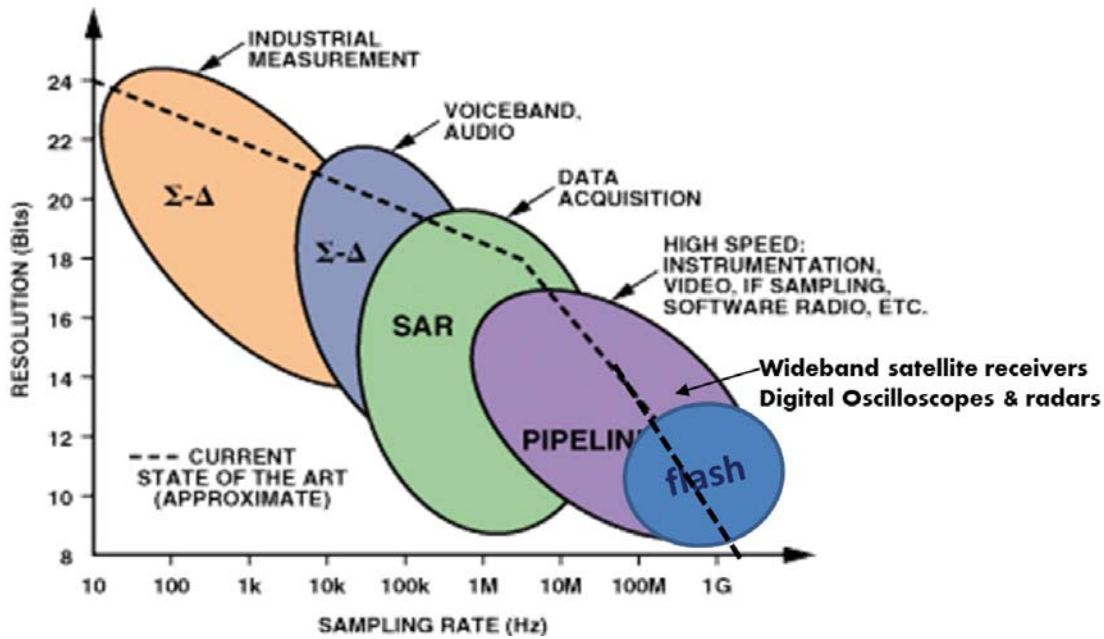


Figure 1.23: Resolution v sampling rate of ADC (courtesy of Shahar Kvatinsky, Technion<sup>51</sup>)

## Key Areas of Focus and Follow-on Research

Analog-based Machine Learning (ML) architectures: There is a tremendous amount of research on AI and ML based on digital processors, including CPU, GPU and TPU (tensor processing unit). The need for more energy-efficient ML and inference at the node and edge requires innovation to facilitate Sensing to Action or true distributed edge intelligence. Analog holds promise to provide parallel matrix processing efficiently, depending on application. Research is needed to explore various analog-based ML architectures that include storage and work collaboratively with digital systems for a total solution. Examples include:

- Compute in memory with ADC/DAC
- Analog summing with comparator sensing
- Spike processing—neural model
- Resistive arrays or capacitive summing
- Understanding optimal partitioning between analog and digital domains for specific applications

Analog element study and optimization: The above-mentioned architectures will require some type of analog device or element for processing. There are many which can be explored, but there is need to identify a “few good solutions” to allow architecture optimization. Elements could include:

- Analog floating gate
- RRAM with multi-level capabilities. The particular goal is lowering variations in I-V characteristics and reducing write and read current of memory cells
- SRAM with multi-level capabilities
- Other NVM with multi-level capabilities which can store, sum, and integrate results

Algorithms for analog AI and ML: Analog ML will bring in non-ideal effects, including noise, non-linearity, gain variations, and process variations. Algorithms will be needed which cannot only operate effectively in this environment but possibly take advantage of them, such as stochastic or “noisy” signals that facilitate optimization or non-linearity that extends dynamic range. Research already indicates that learning systems can tolerate lower-bit resolution.

Co-design across devices, circuits, and algorithms will be needed to address device and circuit nonidealities. It is a recurring theme within this section and is highlighted in the *Basic Research Needs for Microelectronics* report published by the Department of Energy Office of Science Workshop in 2018<sup>1</sup>. To develop better analog compute devices for edge machine learning, we need to look at specific computing workloads and understand their bottlenecks.

# 1.6 Analog Design Productivity and Predictability

## Overview and Needs

*Integrated circuit (IC) design has benefited tremendously over the past half-century from design automation.* Today's multi-billion transistor ICs are possible because of increasingly higher levels of abstraction (transistors to gates to blocks to cores) and the rise of reusable IP (intellectual property) block providers that enable design to be done at the "system" level. Additionally, massive investment in the development of algorithms and computer-aided design (CAD) tools to accelerate and automate the IC design process, as well as a tremendous increase in the available compute power that underlies today's electronic IC design flow, have driven the industry forward. The capability and complexity of CAD tools continue to increase, but so does the complexity of IC manufacturing technologies, especially the restrictions (design rules) on layout. **This leads to a "design gap" between what is possible in a technology and what can be designed in a reasonable time, while still meeting stringent high-volume manufacturing requirements.**

However, the aforementioned improvements in design capability and productivity have overwhelmingly benefitted digital design, not analog/mixed-signal (AMS) design. The time-honored AMS design flow of schematic (circuit) design, then physical (layout) design, then verification, then test is still largely intact today. Yet it's being challenged at 7nm and below as layout-dependent effects (LDEs) and parasitics dominate the inherent device performance<sup>52</sup>. There have been significant improvements in all aspects of that flow. For example, since parasitics and LDEs can have such an impact on transistor performance in modern technologies, coupling schematic and layout design with extracted views is a must. Automating aspects of AMS layout and the increase in available compute power have benefitted AMS just as much as they have digital, but there have been no breakthroughs in abstraction and efficiency comparable to those on the digital side. AMS blocks are almost always the limiting factor in design of complex mixed-signal ICs. With only slight exaggeration, the digital portion is expected to be synthesized and work right the first time, whereas multiple designs for an AMS block may be attempted, knowing some will not meet required performance/cost/power and will have to be discarded. Depending upon the required performance, even the "good" designs can require several iterations in test shuttles to meet requirements. RF/AMS design is inherently an optimization of a multivariable space where the designer is effectively working with  $n$  equations and  $m$  unknowns, where  $m \gg n$  due in part to complexity, model limitations, integration effects, etc.

AMS design time and cost, including test, is therefore a limiting factor in the development of modern mixed-signal ICs and even large mostly digital SOCs, because the latter always include significant amounts of AMS IP. This will be a key limiting factor with the *high growth in applications of semiconductors in IoT and many other areas that may require more unique solutions.* The semiconductor industry would benefit greatly from improved AMS design productivity and predictability. SRC thought leadership is designed to address that need over the next decade.

## Background

AMS block/product development includes the major phases of design (both schematic and physical), verification, and test. Manufacturability constraints, parasitics, LDEs, and variability have all significantly increased the complexity of digital design, as manufacturing technologies have scaled. In addition to those hurdles, AMS design must overcome constraints from lowered supply voltages and less "AMS-friendly" transistor characteristics that can make yesterday's standard AMS techniques obsolete. Digital design has benefitted from abstraction, synthesis, automated regression testing, DFT (Design for Test), BIST (Built-In Self-Test), etc., but is still challenging. AMS has benefitted from improved models, SPICE capabilities, and some layout automation. But it has not fundamentally changed and is still the "long pole in the tent." **Compared to digital design, AMS circuit design disproportionately requires resources, stretches schedules, causes test issues, results in more field failures, and fails to meet specifications.**

AMS circuit analysis and design often use textbook models and simplified conceptions of how transistors operate. This framework can be useful as a basis for conceptualizing new and innovative circuits but is not sufficient to verify that the new ideas will actually work in practice<sup>53</sup>. "Back of the envelope" calculations cannot be trusted, and even the operating point information printed by a model/simulator is only a rough approximation of the terminal-to-terminal metrics for a transistor. Any assistance or automation of the AMS design process must be based on the actual simulation models for the technology. The "actual" simulation must encompass composite devices, and in advanced technologies, long AMS transistors are often not allowed and must be made by "stacking" multiple shorter transistors in series<sup>46</sup>. Getting appropriate design information for such devices is not possible from operating point information but is available from

simulations of the composite stacked transistor. Additionally, as frequencies push into the THz regime discussed in section 1.3, extracted and accurately modeled interconnect and parasitic impact is critical to predict performance. Techniques need to go beyond existing resistance and capacitance toward RLC and transmission line models.

**Verification and test are additional bottlenecks for AMS/mixed-signal (AMS) design<sup>54</sup>.** SPICE-like transient simulations are impossible to use for full SoC system-level verification. As a result, quality/reliability of RF and AMS blocks/circuits is 10X worse than their digital counterparts, and test coverage is poor. This is fundamentally due to the fact that there is no

measurement of a test's defect coverage nor latent defect activation. Efforts are underway in IEEE P2427 to develop standards for defect modeling and coverage<sup>55</sup>. *Behavioral block models and "real number" abstractions of AMS blocks are examples of techniques that can greatly speed up system verification<sup>65</sup>.* Similarly, IEEE P1687.2 is being developed to define AMS test access standards. Automated DFT, BIST, and test generation for AMS have had some success but is nowhere near as advanced as for digital. Substantial improvements in EDA tools and techniques for test coverage are essential, particularly for mission-critical applications like automotive functional safety where failures are not acceptable.

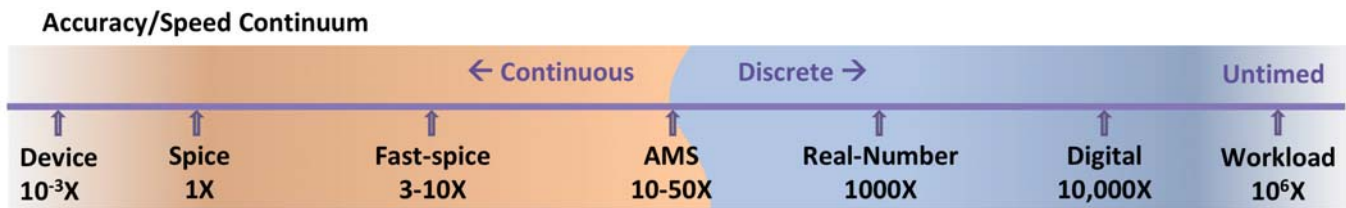


Figure 1.24: Relative simulation/verification time (courtesy of Arturo Salz, Synopsys<sup>65</sup>)

Many factors contribute to the performance and profitability of ICs: design effort, die area, supply current, test, and yield. **A framework for AMS design to maximize profitability based on optimizing those disparate factors could revolutionize the industry.** Automation of design (including AMS BIST) that did not squeeze out every mm<sup>2</sup> of area but got a functional block out with reduced design effort and time, could see faster time to market, increased sales, and ROI for lower volume parts.

Many attempts to automate—or at least improve—AMS design have been proposed over the past three decades. There were initial attempts at “AMS synthesis”, that tried to follow the success of the abstraction of transistors to standard cells (circuit blocks) and the development of languages/algorithms/tools that automatically took a high-level definition of a block and implemented it in those standard cells. Because AMS embodies a large variety of circuits with completely different functions and figures of merit, the synthesis proposals often were limited to one circuit or a small class of circuits, although some were more general. Most were from universities, and some led to spinoff companies to try to commercialize the university research. Unfortunately, none entered mainstream use and moved AMS design up the food chain.

The more modest goal of “AMS optimization” was tried after full-blown synthesis failed to materialize. Often the same

circuit block needs to be “reused” in a different product or “ported” to a different technology. However, there is almost never “exact” reuse: product specifications may be different, so the block has to be tweaked; the supply voltage levels may be different; or the device palette in a new technology is different, so design parameters may need to be adjusted to account for different transistor characteristics. Still, knowing the key design parameters (width and/or length of important transistors, values for specific capacitors or resistors, etc.) and key design performances, it would seem to be a simple task to wrap one of the many available optimization algorithms or codes around this problem and automate reuse and porting of the circuit block. Multiple algorithms/tools have been proposed for this, but again none has entered mainstream use. There are many reasons this has not been successful: an automated tool may find an “optimal” design that violates some unspecified requirement (e.g., noise margin, known historical sensitivity to process or temperature variation, etc.) that a designer just “knows” and automatically takes into account when designing; documenting in detail what the key parameters and performances for a circuit block are can take significant time; and a designer typically gets little or no recognition/reward for this if the block is used by someone else. Furthermore, a different specification/technology may completely preclude a given topology being used, for instance, if supply voltage decrease precludes the use of cascoding.



Parasitics and LDEs are so important for today's devices and technologies that schematic and physical design cannot be decoupled. Improvements in AMS design productivity will have to encompass both aspects. The historical approach to AMS design synthesis/optimization of adjusting transistor geometry and biasing will not work. *This is exemplified by the UC Berkeley BAG2 approach<sup>56</sup>, which encapsulates design methodology/knowledge in customized software scripts that sit "on top" of both schematic and layout views.* As digital design has shown, an additional advantage of this approach is that it is script-based rather than GUI-based, which inherently increases design productivity and enables automated regression testing.

Improving AMS design reuse—or even moving up the food chain to AMS synthesis—by necessity trades simulation time for designer time. Simulation and verification test benches need to be reusable and should link to BIST and production test development and automation. Detailed verification of AMS blocks themselves needs to be done at the SPICE level. But for verification with associated digital circuitry, or at a higher level, it is infeasible to simulate SPICE-level using "AMS-accurate" models. Being able to use much faster "digital-accurate" models in SPICE for digital blocks would be of great benefit. Although it has been proposed in the past, it has never become a reality. **Behavioral models are of significant benefit, but developing them can be a bottleneck.** While there have been attempts in academia and industry, automated creation and verification of behavioral models is in its infancy, yet it is essential for practical simulation time.

**Another possible roadblock to adoption of automation of AMS design is mindset.** Many AMS designers have years or decades of experience and knowledge, look down on coding as an inferior intellectual task (that is not as true with today's graduates), and may even think of automation as a threat to their jobs.

## Targets

AMS design is still an art backed heavily by science. And hardcore AMS designers believe this is extremely unlikely to change. With roughly 60 years of evidence to support their argument, including many failed attempts to prove otherwise, is there a middle ground? At the very least, if the "inspirational" aspect of AMS design cannot be automated, improvements in the non-value-added "grunt" aspects of AMS design would be of great benefit and would free up precious "expert designer" resources to concentrate on the aspects of AMS design that cannot, for now, be automated.

**Defining design productivity has always been an elusive, daunting task. With no clear baseline, it becomes extraordinarily difficult to quantitatively measure productivity.** However, there are improvements that, when made, obviously and undeniably improve productivity. In many cases, they make the impossible possible. Perhaps design predictability is easier to measure, but in the AMS world there are many variables impacting the result. This is primarily dictated by the complexity of the end application and the associated requirements for AMS circuits, and by the ability to model (or not) all parasitics, particularly complex LDEs and substrate coupling in high frequency circuits. While some circuits will have an inherent predictability advantage due to their relative simplicity, predictability can be measured, nonetheless. **Test complexity and the resulting test time also continue to grow, so cost is also increasing.** New paradigms are needed that are more akin to DFT for digital circuits. Here, the target needs to be written around the cost of test as compared to the cost of manufacturing the die. Quality and reliability have standard measures, and AMS circuits significantly trail their digital counterparts in these metrics. But targets can be set, and the AMS community should strive to meet them as more and more applications, such as autonomous vehicles, are becoming mission critical. Lastly, built-in inherent security, methods of measurement, and metrics are beyond the scope of this chapter but are addressed in Chapter 4. The following list defines targets to be met by 2030.

- Simulation speed improvement of > 100x compared to a standard suite of circuits (see suggested set in call for proposals), particularly data converters and PLLs due to a combination of enhanced algorithms, ongoing compute hardware improvements, improvements in statistical analysis, and simplification of the method of simulation
- Model improvements compared to a standard suite of circuits (see suggested set in call for proposals) accurate to within 1% over all process corners and statistical variations, such that when combined with the best simulators ensure all design specifications or parametric yield targets are met
  - Early estimates, perhaps ML-based from schematic entry, accurate to within 10%
- Demonstrated business benefit for reuse/semi-automation (see Appendix B)
  - Component margin dollars exceed development cost by 10X, despite a likely larger die size due to the inefficiencies in semi-automated design approaches
  - Meets comparable performance as verified through simulation and modeling of handcrafted design

- Introduction of a commercial tool that is as commonly used as SPICE that provides reuse through repetition, parameterization, parameter/feature extension, and/or portability, freeing the AMS designer from the mundane elements of design and allowing him/her to concentrate on the creative elements
- Test coverage of defects > 90%; performance correct by design
- Continuous self-test for identification of part wear-out and alert due to TDD, HCI, NBTI, etc.
- Defective parts per million (DPPM) for AMS designs < 0.1
- Detailed study and exploration of AMS design methodology and flow; identify the key points or tasks that can best be automated or addressed with ML algorithms to free designers for creation of new innovative solutions

### Key Areas of Focus and Follow-on Research

Research should address achievement Targets defined above. While not an exclusive list, some proposal topics are:

- Circuit techniques that used to work at 65 nm, where a considerable amount of AMS design has been implemented, no longer apply. Access to small geometry nodes (10 nm and smaller) is a fundamental requirement to validate models and simulation of the latest AMS circuits. Yet, as it stands today, this access is extremely limited to academia. Proposals to resolve this dilemma are needed.
- Definition of a benchmark suite of AMS circuits to measure simulation and modeling improvements against consisting of simple (differential pairs, bandgap regulators, etc.), medium complexity (op amps, filters, LNAs, mixers, etc.) and very complex (continuous time  $\Sigma\Delta$  A/D converters, PLLs, etc.) circuits.
- Definition of a benchmark suite of AMS circuits with estimates of engineering effort in full-time engineer (FTE) equivalents to support analysis/comparison of reuse/semi-automation approaches.
- Reuse/semi-automation tools to improve designer productivity and/or circuit predictability.
- New simulation algorithms providing improvements in accuracy and/or throughput.
- DFT techniques for random AMS circuitry.
- Automated insertion for AMS circuitry.
- Methods to measure AMS defect coverage with acceptable accuracy in acceptable time.
- ML-based approaches to optimize design and/or layout.
- ML-based approaches to identify outliers during test to eliminate defects from escaping to the field.

## 1.7. Summary—New Trajectories for Analog Electronics

### Overview

Analog electronics are key to interfacing and processing real-world conditions and providing means to convert the sensed conditions to real-world actions. Fundamentals of analog impact all electronics, including communications, storage, and computation, which are covered in subsequent chapters of this report. There are opportunities for analog innovation in all these areas, as well as power management and conversion, with focused research addressing key challenges identified above. With the exponentially increasing data from analog sensing, there is need to reduce the raw data to usable and actionable information. This will alleviate the load on storage and communications, as well as provide local, timely, energy-efficient, and secure automation.

The **Analog Grand Goal** is for revolutionary technologies to increase actionable information with less energy, enabling efficient and timely (low latency) sensing-to-analog-to-information with a practical reduction ratio of  $10^5:1$ .

### Research Recommendations Summary

- Study of holistic solutions with key applications knowledge with focus on minimal processing to take action
  - Collaborative multi-expertise research projects demonstrator platform(s)
- Heterogeneous integration to make best use of best technology in an energy-, size-, and cost-efficient manner
  - CMOS platform integration—optimized technologies
  - Package platform integration—multi-technology/multi-die from DC to THz
- Optimum power management (including control) and conversion for efficient and fast response energy control
- Leverage human systems as a model for bioinspired, local “sensing to action” including efficient machine learning and inference at the edge
  - Analog-based ML architectures (compute in memory, synapse, etc.)
  - Architectures and algorithms that leverage analog approach and compensate or take advantage of analog non-idealities
- Flexible, scalable platform and technology, including sensors, memory, and signal representation matched to domain

- Analog building blocks optimized for application and good enough with integrated signal processing
- Efficient array-based signal processing, including multi-sensor/multi-model fusion
  - THz sensing and communications arrays are of significant challenge
  - Accurate, low-noise and tightly synchronized timing solutions/clocks (broadly applies)
- New devices for analog ML, THz operation, and power conversion
  - New analog elements for simultaneous computation and storage
- Silicon Photonics for signal distribution, processing, and sensing
- Methodologies and models to improve analog and mixed-signal simulation >100x without loss of good-enough accuracy for the application
- Design methodology and supporting tools to facilitate reuse of “knowledge” and improve analog design productivity by 10-100X
  - Identify key areas/tasks that can best be automated or addressed via ML algorithms to free designers for creation of new innovative solutions
- Design for test and test methodologies for >90% analog defect coverage for predictable quality
- Models which more accurately predict analog performance prior to silicon covering precision, and THz frequencies over thermal, stress, and aging
  - Include non-ideal effects of physics and parasitics (on-chip and in-package)
  - Include statistical variations to address parametric shifts leading to effective “defect”
- Methodology and infrastructure to provide heterogeneous technology (fab) access to researchers to validate designs in industry current and advanced technologies

## Appendix

### Appendix A: Total analog information from the physical world

Information always requires physical carriers that are material (quasi) particles (e.g., photons, phonons, electrons, etc.) transferring energy rather than thermal noise. Energy is a prerequisite for any information event. In Earth environment, the main source of energy is Sun. The total amount of information available on Earth can be estimated using the number of photons that are reflected from the surface of Earth, and thus collectable by a receiver (e.g. human eye).

The total solar radiation reflected by earth surface  $E_{tot} = 7 \cdot 10^{15}$  J/s (Figure A1). In a first-order approximation, we assume that all this radiation consists of green photons with wavelength  $\lambda=550\text{nm}$ . The corresponding energy of one photon is:

$$E_{ph} = \frac{hc}{\lambda} = 3.6 \cdot 10^{-19} \text{ J} \approx 2.3 \text{ eV} \sim 100k_B T$$

Therefore, the number of equivalent photons per second that can act as information carriers is:

$$N_{ph} = \frac{E_{tot}}{E_{ph}} = \frac{7 \cdot 10^{15}}{3.6 \cdot 10^{-19}} \sim 10^{34} \text{ photons/s}$$

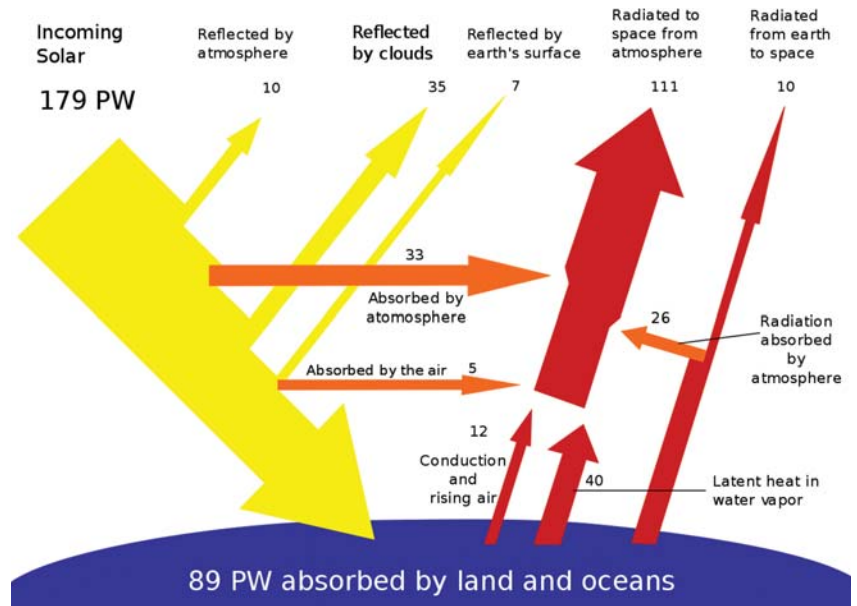


Figure A1: Energy breakdown of incident electromagnetic radiation from the Sun

Table A1: Information acquisition rates of senses and processing<sup>57</sup>:

| Sensory System | Bits per second | Processed by the brain (bits/s) |
|----------------|-----------------|---------------------------------|
| Eyes           | 10,000,000      | 40                              |
| Ears           | 100,000         | 5                               |
| Smell          | 100,000         | 1                               |
| Taste          | 1,000           | 1                               |

Figure A2: Human sensory system of data acquisition via sensors

Based on Table A1 the maximal individual human sensory throughput is  $\sim 10^7$  bit/s. For the total human population of  $\sim 7.5$  billion, the collective human sensory throughput— $\sim 10^{17}$  bit/s. While the individual human sensory throughput is  $\sim 10$  Mbs, the 'conscious bitrate' is  $< 50$  bps, thus the compression ratio of  $> 200\ 000:1$ .

### A3 Global rate of data acquisition via sensors

Janusz Bryzek, CEO of eXo Systems, Inc. and chair of the TSensor Summit, Inc. compiled an inventory of the total number of sensors installed in the world, based on 10 different sources and estimated that this number was approximately 1 trillion ( $10^{12}$ ) in 2018 and will increase to 45 trillion ( $4.5 \times 10^{13}$ ) by 2032<sup>58,59</sup>. The corresponding collective data acquisition rate in 2032 is projected to be  $10^{27}$  bytes-per-year<sup>59</sup>, which is equivalent to  $\sim 10^{20}$  bits/s.

Based on the reported numbers of sensors, one can also estimate the total data acquisition rate by sensors. According to<sup>60</sup>, image sensors constituted  $\sim 30\%$  of the global sensor market in 2018.

The data acquisition rate of image sensors can be estimated based on the following assumptions:

- VGA Standard resolution  $640 \times 480 = 3.07 \times 10^5$  pixels/frame
- 1 frame-per-second acquisition
- 8 bits/pixel (Black and White Video)

This gives acquisition rate of  $2.46 \times 10^6$  bits/s per image sensor. All other common sensors such as flow, level, pressure, temperature, chemical, position, etc., have low-data acquisition rates, e.g., 8–16 bits/s. Since this is much lower than the image sensors, all other sensors can be excluded from this simplified estimate without significant effect on accuracy. Thus, a numerical example for 2018 and 2032 yields:

- Total sensors:  
 $10^{12}$  (2018) and  $4.5 \times 10^{13}$  (2032)<sup>69,70</sup>
- Total sensory throughput:  
 $7.37 \times 10^{17}$  bit/s (2018) and  $1.11 \times 10^{20}$  bit/s (2032)

The data points for total sensory throughput in **Figure 1.10** in this chapter were obtained using projections for the total number of sensors<sup>58,59</sup> (averaged among all sources of data) multiplied by the acquisition rate of image sensors as estimated above.

## Appendix B: Proposed figures of merit

Any tool developed for reuse or semi-automation will reduce development cost and almost certainly come at an increased

die area penalty. Of course, it could provide other benefits for the new AMS designer, such as higher reliability, higher circuit yield in the face of wafer fab parametrics, faster time to market, etc.—but the guru would not likely agree. However, can this tradeoff analysis be simplified to compare the decrease (increase) in die cost versus the increase (decrease) in development cost for the optimized case (partially machine-generated case)?

A good product will produce 10X the development cost in margin dollars. The market sets price, so an increase in die cost will decrease margin dollars. Unless this die cost is offset by some other factor, such as a reduction in test cost, it can only be countered by increased product sales. A reasonable assumption is that end of life for a product, whether hand-optimized or partially machine-generated, is the same. However, it is reasonable to assume that the partially machine-generated product gets to the market sooner and that time difference will result in incremental margin dollars.

Reuse and semi-automated approaches lead to an increase in die area (A). An increase in A will reduce yield (Y), due to defects and to the reduction in Possible Die Per Wafer (PDPW). The change in Y due to a change in A for Seeds' Law is  $\delta Y / \delta A = -D / (1 + AD)^2$  where D = Defect Density. This impact for  $D < .25$  d/cm<sup>2</sup> (achievable with today's manufacturing capability) can be shown to be negligible compared to the reduction in PDPW that reduces linearly with increase in A. For complex AMS circuits and/or processes, die yield is often limited by parametrics, which are not always a straightforward function of defect density (e.g., soft errors shifting parametrics, versus hard functional errors). In this case, for those blocks or elements that would be considered for reuse/semi-automation, the parametric yield is to a first order independent of the design approach (i.e., machine-generated solutions need to be comparable to handcrafted solutions to be effective), and the PDPW is limited simply by the change in area. Thus, for a 1% increase in A, there is a 1% decrease in PDPW and a 1% increase in die cost. Note that  $M\$ = P - C$  where  $M\$$  = Margin dollars, P = Price and C = Cost. Therefore, a 1% increase in cost will be a  $> 1\%$  decrease in Margin dollars.

On the other hand, reuse and semi-automated approaches provide a reduction in development cost and time. To a first order, the development time and cost of the machine-generated circuit/block approaches zero. Development cost of an optimized circuit/block will need to be estimated as the product of the number of Full Time Engineers (FTEs) and an engineer's loaded cost.



To quantify the crossover point for where reuse/semi-automation is a business benefit, consider the following:

$N_0$  = Number of parts sold if a hand-optimized AMS guru completed the design

$N_N$  = Additional units sold if a performance equivalent part made it to the market sooner

$P$  = Price the units are sold at, independent of time

$C_0$  = Cost of hand-optimized AMS-guru-completed design

$C_N$  = Cost of partially reused or semi-automated design

$$\Delta C = C_N - C_0$$

$DC_0$  = Development cost of hand-optimized AMS-guru-completed design

$DC_N$  = Development cost of partially reused or semi-automated design

$$\Delta DC = DC_0 - DC_N$$

In order to satisfy the requirement that the margin dollars exceed 10 times the development cost:

$$(N_0 + N_N) * (P - C_N) > 10 * DC_N \text{ which can be rewritten as}$$

$$N_0 * (P - C_0) - 10 * DC_0 + N_N * (P - C_N) > N_0 * \Delta C - 10 * \Delta DC$$

and for the case where  $N_0 * (P - C_0) = 10 * DC_0$  (i.e., the original design was a good product) then

$$N_N * (P - C_N) + 10 * \Delta DC > N_0 * \Delta C$$

Stating this in words, the margin dollars from the incremental sales plus the savings in development cost must offset the incremental cost applied to all the original unit sales.

This is perhaps an oversimplified analysis due to other considerations such as:

1. In addition to development cost, an amortized amount for creating and maintaining the IP should be included.
2. Time value of money is not considered.
3. Price and cost will vary with time and volume, and that is not considered.
4. The potential time saved, which impacts time to market plus availability of the designer to move on to additional new products though this, is difficult to quantify.

But it is directionally correct, relies on some estimations whose accuracy likely overwhelms the factors not considered above, and gives the developer a strong guideline to ensure his/her approach will be a commercially viable solution rather than technology for technology's sake.

## Contributors

Jim Wieser (Texas Instruments)—*Chair*

Behrooz Abdi (TDK InvenSense)

Elad Alon (UC Berkeley)

Fari Assaderaghi (Sunrise Memory)

Rashid Attar (Qualcomm)

Seyfi Bazarjani (Qualcomm)

Geoffrey W. Burr (IBM Research-Almaden)

Gert Cauwenberghs (UC San Diego)

Michael Flynn (U Michigan)

Brian Ginsburg (Texas Instruments)

Doug Garrity (NXP)

Marcel Geurts (NXP)

Ken Hansen (SRC)

Baher Haroun (Texas Instruments)

Jennifer Hasler (Georgia Tech)

Peter Kinget (Columbia U)

Fritz Kub (Naval Research Lab)

Shahar Kvatinsky (Technion)

Chiao Liu (Facebook)

Gabriele Manganaro (Analog Devices)

Matthew J. Marinella (Sandia National Labs)

Colin McAndrew (NXP)

Michael Niemier (Notre Dame)

Ken O (UT Dallas)

Jan Rabaey (UC Berkeley)

Mark Rodwell (UC Santa Barbara)

Arturo Salz (Synopsis)

Vladimir Stojanovic (UC Berkeley)

John Paul Strachan (Hewlett-Packard Labs)

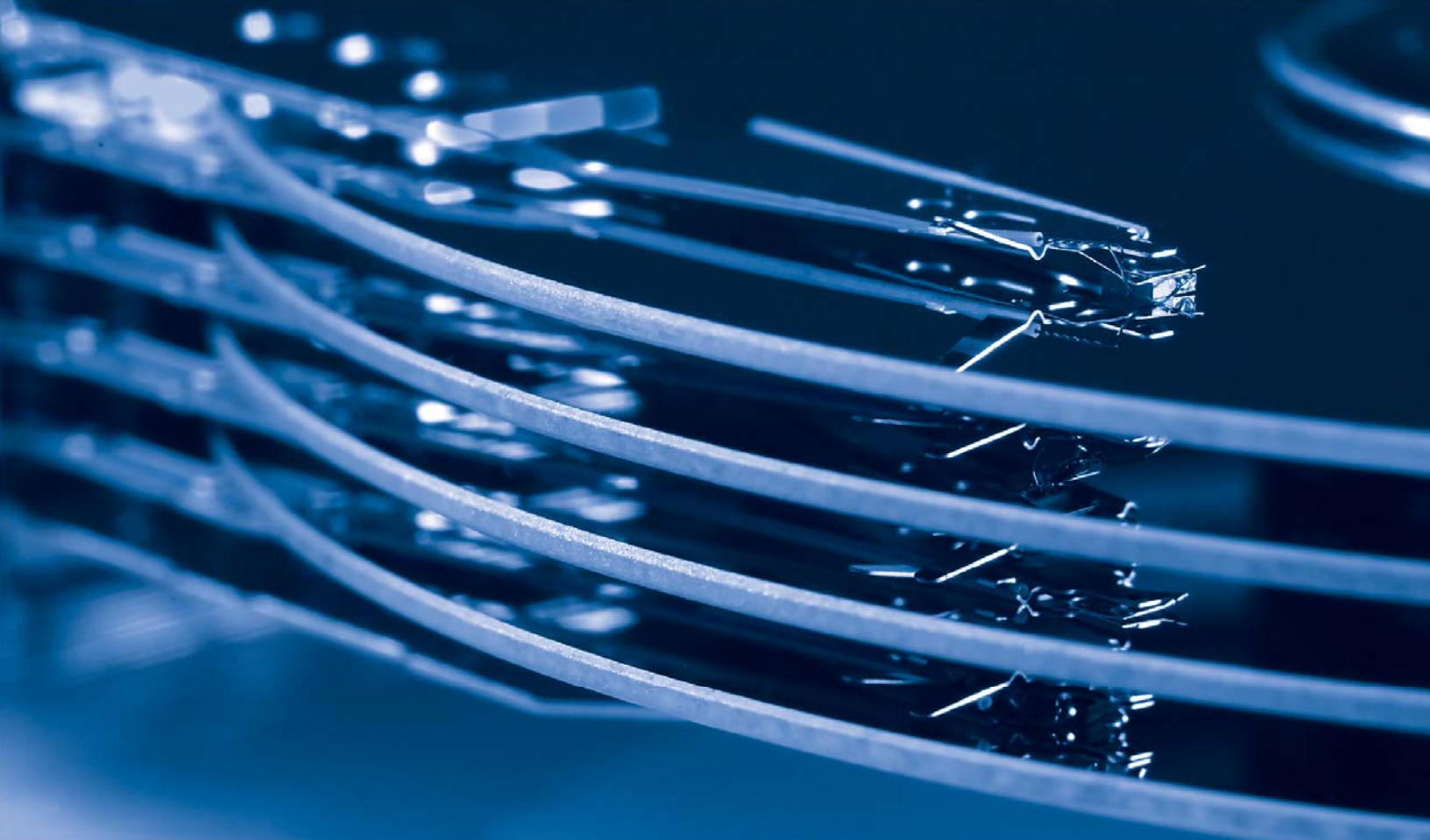
Dmitri Strukov (UC Santa Barbara)

Stephen Sunter (Mentor)

# References to Chapter 1

- <sup>1</sup>Basic Research Needs for Microelectronics. US Department of Energy report 2018, [https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN\\_Microelectronics\\_rpt.pdf](https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf)
- <sup>2</sup>Gabriele Manganaro. "Analog 10 Years from Now" Presented at SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose CA USA Dec. 2019
- <sup>3</sup>Peter Kinget. "The World is Analog - Future Challenges", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>4</sup>Behrooz Abdi. "Sensors and Actuators for the next decade", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>5</sup>D. Guermandi, et al "A 79-GHz 2x2 MIMO PMCW Radar SOC in 28nm CMOS," IEEE Asian Solid-State Circuits Conference Nov. 2016
- <sup>6</sup>B. J. Baliga, "Power semiconductor device figure of merit for high-frequency applications," in IEEE Electron Device Letters, vol. 10, no. 10, pp. 455-457, Oct. 1989.
- <sup>7</sup>A. Lidow et al. "Getting from 48 V to Load Voltage: Improving Low Voltage DC-DC Converter Performance with GaN Transistors" Presented at IEEE Applied Power Electronics Conference and Exposition (APEC) March 2016
- <sup>8</sup>H. Amano et al. "The 2018 GaN power electronics roadmap" J. Phys. D: Appl. Phys. 51 163001
- <sup>9</sup>Fritz Kub. "GaN Power Technology", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>10</sup>Michael Niemier. "Analog Circuits with Beyond CMOS Devices", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>11</sup>Ercsey-Ravasz, M., Toroczkai, Z. "Optimization hardness as transient chaos in an analog approach to constraint satisfaction". Nature Phys 7, 966-970 (2011).
- <sup>12</sup>Yin, et al., "Efficient Analog Circuits for Boolean Satisfiability" IEEE Transactions on VLSI, 26(1), p. 155-167, 2018
- <sup>13</sup>Molnár, B., Molnár, F., Varga, M. et al. "A continuous-time MaxSAT solver with high analog performance". Nature Communications 9, 4864 (2018)
- <sup>14</sup>F.Chen et al., "Compressed Sensing Architecture for Data Compression in Wireless Sensors", IEEE Journal of Solid-State Circuits, Vol.47, No3 March 2012
- <sup>15</sup>Siram Vajapeyam, "Understanding Shannon's Entropy metric" 24-March-2014 – Cornell University - arXiv:1405.2061 [cs.IT]
- <sup>16</sup>Jan Rabaey, "Sensing to Action: The nature of distributed intelligence", Presented at SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose CA USA Dec. 2019
- <sup>17</sup>Chiao Liu, "Intelligent Vision Systems – Bringing Human-Machine Interface to AR/VR", Presented at SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose CA USA Dec. 2019
- <sup>18</sup>Baher Haroun, "Intelligent Sensors: Sensing to Action", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>19</sup>Michael P. Flynn, "Compressed Sensing for Analog Signals and Imaging", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>20</sup>Rashid Attar, "Ultra-Low Power Highly Integrated SoC for IoT", Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>21</sup>Gert Cauwenberghs, "High-Density Neural Interfaces for Pervasive Human-Machine Interaction", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>22</sup>Mark Rodwell. "HBT THz integrated circuits and applications" SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>23</sup>Q. Zhong et al., "CMOS Terahertz Receivers," Proc. IEEE Custom Integrated Circuits Conf., San Diego, CA, Apr. 2018
- <sup>24</sup>Z. Ahmad et al., "Devices and Circuits in CMOS for THz Applications," Proc. IEEE Int'l. Electron Device Meeting, Dec. 2016, San Francisco, CA, Paper 29.8, pp. 734-37.
- <sup>25</sup>Kenneth K. O. "CMOS Platform for THz", Presented at SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>26</sup>D. Y. Kim, S. Park, R. Han and K. K. O, "820-GHz imaging array using diode-connected NMOS transistors in 130-nm CMOS," 2013 Symposium on VLSI Circuits, Kyoto, 2013, pp. C12-C13.
- <sup>27</sup>I.R. Medvedev et al., "Submillimeter Spectroscopy for Chemical Analysis with Absolute Specificity," Optics Letters, vol. 35, issue 10, 2010, pp. 1533-35.
- <sup>28</sup>S. Kang, S. V. Thyagarajan, and A. M. Niknejad, "A 240 GHz Fully Integrated Wideband QPSK Transmitter in 65 nm CMOS," IEEE J. Solid-State Circuits, vol. 50, no. 10, Oct. 2015, pp. 2256-67
- <sup>29</sup>Q. Zhong et al., "300-GHz CMOS QPSK Transmitter for 30-Gb/s Dielectric Waveguide Communication," Proc. IEEE Custom Integrated Circuits Conf., Apr. 2018, San Diego, CA, paper 13-4
- <sup>30</sup>Brian Ginsburg. "Automotive and Industrial Radars", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019

- <sup>31</sup>Marcel Geurts. "Analog Beamforming Antennas", SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>32</sup>Vladimir Stojanovic. "Integrated Silicon Photonics for Communication and Sensing" Presented at SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>33</sup>Sun, C. et al. A monolithically-integrated chip-to-chip optical link in bulk CMOS. *IEEE J. Solid-State Circ.* 50, 828–844 (2015)
- <sup>34</sup>B. Ullmann. "Why algorithms suck and analog computers are the future." <https://blog.degruyter.com/algorithms-suck-analog-computers-future/> (accessed).
- <sup>35</sup>Source: NVidia presenter's day, 2016, <https://investor.nvidia.com/events-and-presentations/presentations/2016/default.aspx>
- <sup>36</sup>Geoffrey W. Burr, "Materials for analog computing," SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>37</sup>M. Bavandpour, S. Sahay, M.R. Mahmoodi, and D.B. Strukov, "Mixed-signal neuromorphic processors: Quo vadis?", in: *Proc. IEEE S3S'19*, San Jose, CA, Oct. 2019, pp. 1-2
- <sup>38</sup>C. Diorio, P. Hasler, A. Minch, and C. A. Mead, "A single-transistor silicon synapse", *Transactions on Electron Devices*, vol. 43, pp.1972-1980. Nov. 1996.
- <sup>39</sup>J. Hasler, "Large-Scale Field Programmable Analog Arrays," *IEEE Proceedings*, 2020.
- <sup>40</sup>M. Bavandpour, M.R. Mahmoodi, H. Nili, F. Merrikh Bayat, M. Prezioso, A. Vincent, K.K. Likharev, and D.B. Strukov, "Mixed-signal neuromorphic inference accelerators: Recent results and future prospects", in: *Proc. IEDM'18*, San Francisco, CA, Dec. 2018, pp. 20.4.1-20.4.4
- <sup>41</sup>M.R. Mahmoodi and D.B. Strukov, "An ultra low energy internally analog, externally digital vector-matrix multiplier circuit based on NOR flash memory technology", in: *Proc. DAC'18*, San Francisco, CA, June 2018, art. 22.
- <sup>42</sup>R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Research and Development*, vol. 5, pp. 183-191, 1961.
- <sup>43</sup>Matthew J. Marinella, "Neuromorphic computing with analog nonvolatile memory," SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>44</sup>Ankit, Aayush, et al. *International Conference on Architectural Support for Programming Languages and Operating Systems*. (2019).
- <sup>45</sup>Dmitri Strukov, "Neuromorphic inference accelerators as the best entry application for analog(mixed-signal) computing," Presented at SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>46</sup>J. Hasler, "Large-Scale Field Programmable Analog Arrays," *IEEE Proceedings*, 2020.
- <sup>47</sup>Jennifer Hasler, "Machine learning at the edge: Analog neural systems ICs," SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>48</sup>J. Hasler, "Starting Framework for Analog Numerical Analysis for Energy Efficient Computing," *Journal of Low Power Electronics Applications*, vol. 7, no. 17, June 2017. pp. 1-22.
- <sup>49</sup>John Paul Strachan, "Perspective on designing and demonstrating hybrid analog-digital hardware," SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>50</sup>S. George S. Kim, S. Shah, J. Hasler, M. Collins, F. Adil, R. Wunderlich, S. Nease, and S. Ramakrishnan, "A programmable and configurable mixed-mode FPAA SoC", *TVLSI*, vol. 24 (6), pp. 2253-2261, 2016.
- <sup>51</sup>Shahar Kvatinsky, "Analog in ML some thoughts," SRC Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>52</sup>Seyfi Bazarjani, "AMS Challenges in 7FF and Beyond," SRC Decadal Plan Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>53</sup>Colin McAndrew, "Better Leveraging Models: Why They Should be Used Backward and Where They Must be Improved," SRC Decadal Plan Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>54</sup>Arturo Salz, "Analog/Mixed-Signal Emulation Technology Innovation," SRC Decadal Plan Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019.
- <sup>55</sup>Stephen Sunter, "Emerging Standards for A/MS Design-For-Test and Test Generation Productivity," SRC Decadal Plan Workshop on New Trajectories for Analog Electronics, IBM Almaden Research Center, San Jose, CA, USA, Dec. 12-13, 2019
- <sup>56</sup>E. Chang et al., "BAG2: A process-portable framework for generator-based AMS circuit design," 2018 IEEE Custom Integrated Circuits Conference (CICC), San Diego, CA, 2018.
- <sup>57</sup>M. Zimmermann, *The Nervous System in the context of Information Technology*, in: Human Physiology, Robert F. Schmidt, and Gerhrad Thews (eds.), Springer-Verlag Berlin Heidelberg 1989/
- <sup>58</sup>Janusz Bryzek, "The Trillion Sensors (TSensors) Foundation for the IoT", <https://www.iot-inc.com/wp-content/uploads/2015/11/2-Janusz.pdf>
- <sup>59</sup>Stephen Whalley, "TSensors and Exponential Abundance", APS Actualization of the Internet of Things Conference, April 17-19, 2017, Monterey, CA, <https://www.aps.org/units/fiap/meetings/conference/upload/2-1-Whalley-Trillion-sensors.pdf>
- <sup>60</sup>BCC Research report [https://www.bccresearch.com/pressroom/ias/global-market-sensors-reach-nearly-\\$154.4-billion-2020](https://www.bccresearch.com/pressroom/ias/global-market-sensors-reach-nearly-$154.4-billion-2020)



---

## Chapter 2

# New Trajectories for Memory and Storage

### Seismic shift #2

The growth of memory demands will outstrip global silicon supply, presenting opportunities for radically new memory and storage solutions.

### 2.1. Executive Summary

Radical new solutions in memory and storage technologies will be needed for future ICT with major innovations in devices, circuits and architectures. By end of this decade, the continuing improvements in ICT energy efficiency and performance may

stall as the underlying memory and storage technologies will reach fundamental scaling limitations. At the same time, training data for AI applications is exploding with no limit in sight. It is becoming increasingly clear that achieving new levels of bit density, energy efficiency, and performance in future information processing applications will require synergistic innovations using unexplored physical principles, from materials and devices to circuits and system-level functions.

Global demand for data storage grows exponentially, and today's storage technologies will not be sustainable in the near future due to the sheer mass of material resources needed to support the ongoing data explosion. Thus, new radical solutions



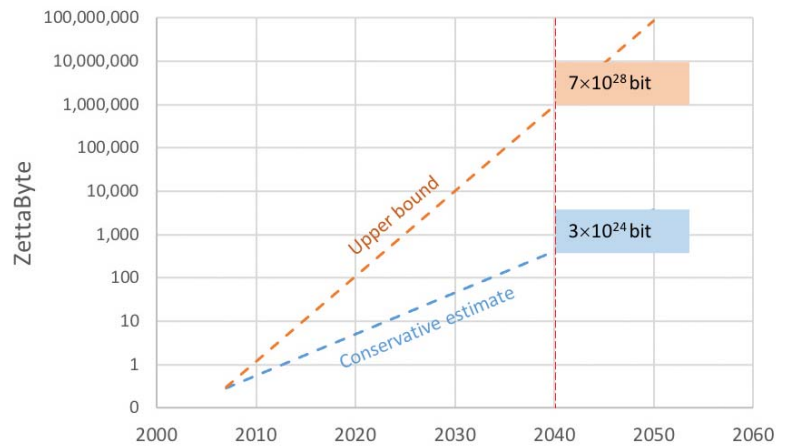
for data/information storage technologies and methods are required. **Figure 2.1** shows the projections of global data storage demand, both a conservative estimate and an upper bound (see Appendix for details). As indicated by **Figure 2.1**, future information and communication technologies are expected to generate enormous amounts of data, far surpassing today's data flows. Currently, the production and use of information has been rising exponentially, and by 2040 the estimates for the worldwide amount of stored data are between  $10^{24}$  and  $10^{28}$  bits, as shown in **Figure 2.1**. Given that the silicon weight associated with one bit of highly scaled NAND flash memory is about 1 picogram ( $10^{-12}$  g)<sup>1</sup>, the *total mass* of silicon wafers required to store  $10^{26}$  bits would be approximately  $10^{10}$  kg—and this would exceed the world's total available silicon supply (**Figure 2.2**).

**Challenge: Global demand for conventional silicon-based memory/storage is growing exponentially (Figure 2.1), while silicon production is growing only linearly (Figure 2.2). This disparity guarantees that silicon-based memory will become prohibitively expensive for future extreme-scale "big data" deployments within two decades.**

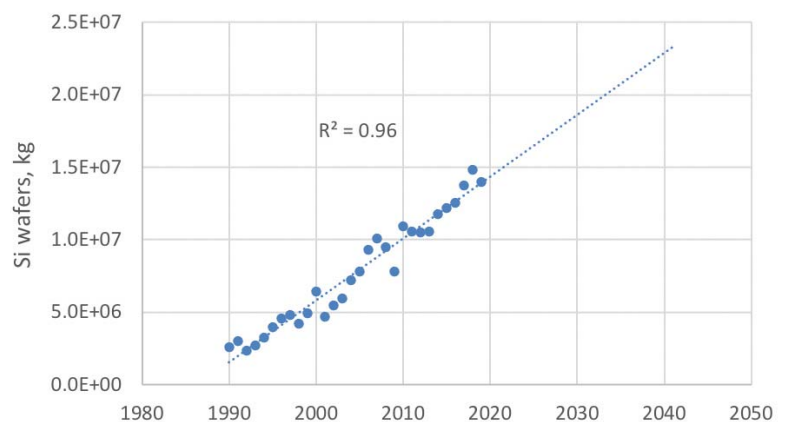
Memory is an essential component of computers, and further advances in computing are impossible without "reinventing" the compute memory system, from memory cells and arrays at various levels of the memory system hierarchy to the memory system architecture. New memory solutions must be able to support multiple emerging applications, such as artificial intelligence, large-scale high-performance heterogeneous computing, and various mobile applications. Novel memory solutions must be able to operate reliably under different application-dependent environmental requirements.

**Memory Grand Goal:** Develop emerging memories and memory fabrics with >10-100X density and energy-efficiency improvement for each level of the memory hierarchy

**Storage Grand Goal:** Discover storage technologies with >100x storage-density capability and new storage systems that can leverage these new technologies



**Figure 2.1: Global demand for memory and storage (utilizing silicon wafers) is projected to exceed the amount of global silicon that can be converted into wafers.**



**Figure 2.2: Global Si wafer supply: 1990-2020 data<sup>2</sup> and future trend**

### Call for action

Radical advances in memory and data storage are required soon. Collaborative research is needed—from materials, devices, and circuits to architecture and processing—for future high-capacity energy-efficient memory and data/information storage solutions serving a vast range of future applications.

**Invest \$750M annually throughout this decade in new trajectories for memory and storage. Selected priority research themes are outlined below.<sup>i</sup>**

<sup>i</sup>The Decadal Plan Executive Committee offered recommendations on allocation of the additional \$3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

## 2.2. Current and Future Application Drivers for Memory & Storage

### Overview and needs

**The amount of data we create as a society is rising exponentially.** Approximately 59 zeta-bytes of data were created and processed in 2020. This was even further increased by the COVID-19 pandemic, which caused an upsurge in work-from-home<sup>3</sup> employees and video communication, as well as the recording and downloading of video data<sup>3</sup>. It is projected that the world will create more than three times the data over the next five years than it did in the previous five<sup>3</sup>. These are enormous numbers, and they are driving a different set of requirements to the ICT systems<sup>4</sup>. Applications fueling these trends include high performance computing (HPC) and data centers, edge computing (including autonomous driving), AR/VR, IoT, and more. There is a strong demand for strategies to improve computing and optimize storage and memory systems to respond to the growing performance needs. Future computing systems must serve a variety of applications with different needs on compute and storage, which requires a flexible approach to adapting compute and storage resources<sup>4</sup>.

To that end, devising new flexible memory and storage system architectures with application-dependent total system performance under as a metric is key.

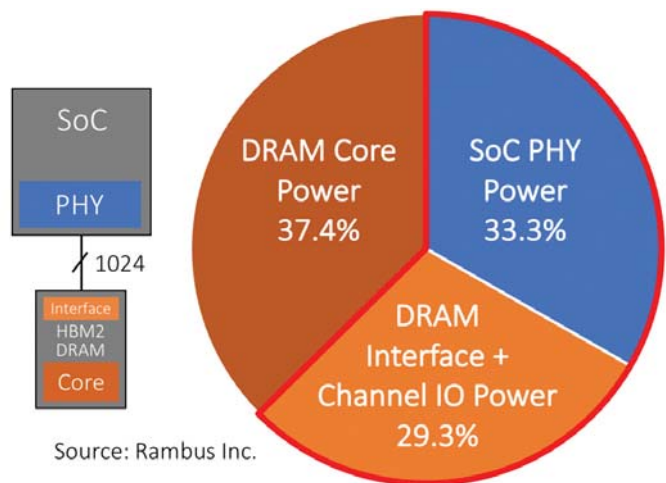
### AI/ML as a key driver for memory and storage

The key drivers for memory and storage are AI and ML, with AI training being most challenging in terms of memory bandwidth and capacity<sup>5,6</sup>. This general trend will continue over the foreseeable future. The AI models and the training data are growing exponentially in size, with more than one trillion parameters being on the horizon. In data centers today, multiple engines are used in parallel for AI training, demanding fast interconnects. Visual and graphics applications drive memory usage with stringent demands in speed, bandwidth, capacity, and energy-efficiency throughout the memory hierarchy.

For these large computer systems, a high utilization is critical for most efficient use and reduction of total cost of ownership (TCO)<sup>5</sup>. Underutilized resources have big negative TCO impact. *Since different tasks require a different system composition for best utilization, the data centers need to be rearchitected in the future using disaggregation and composability. This allows flexible composition and*

*system configuration to optimally serve a particular task.* Considering that the various technical components (CPUs, GPUs, memory, and storage) have different lifecycles, disaggregation additionally improves the system performance and reduces cost, as they can be replaced separately. Common memory systems for AI/ML applications include on-chip memory, high bandwidth memory (HBM), and GDDR—and all have different architectural implications. A universal goal is to realize memory technology with much higher bandwidth and lower latency, while consuming less energy. While HBM DRAMs are already very power-efficient, roughly 2/3 of the power budget is still spent moving data between an SoC and the DRAM (Figure 2.4)<sup>5</sup>. *Reducing the volume of data moved provides an opportunity for large improvement, this requires further research.*

**Different concepts for disaggregation of memory and storage are already proposed, but more research is needed to identify the best way to use disaggregation to achieve TCO benefits at scale and improve latency.** To generate these benefits, a multi-tiered memory approach that includes the use of storage-class memories is needed. The new architectures can pose a challenge but can also provide an opportunity for application development. The impact to legacy code needs to be understood and mitigated.



Source: Rambus Inc.

### Data Movement: 62.6% of Power

Figure 2.3: AI in the data center: HBM2 memory system power (PHY + DRAM power at 2 Gbps, streaming workload; power breakdown for 100% reads or 100% writes<sup>5</sup> (Courtesy of Steven Woo, Rambus Inc.)

## Memory hierarchy in large-scale heterogeneous systems

In recent years, AI/ML workloads have become the major driver for data centers. Similarly, heterogeneous high-performance computing (HPC) systems have fundamentally changed because of the applicability of GPUs for both compute-intensive scientific simulations and AI/ML. For example, the largest U.S. HPC System, named SUMMIT and located at Oak Ridge National Laboratory, is designed for scientific and AI workloads<sup>6</sup>. The system is built from very complex individual heterogeneous nodes that comprise CPUs and GPUs, as well as three memory hierarchies with DRAM, HBM, and NVM and a cache hierarchy within the CPU and GPU. The system is optimized to be good for any type of HPC workload, from traditional modeling and simulation to data analytics and AI, including handling of massive data sets. Such heterogeneous systems possess an incredibly complex memory hierarchy (Figure 2.4) in order to guarantee availability and analysis of huge data sets. **Extraordinary new capabilities are needed for most complex simulations, such as full 3D with sophisticated multi-physics models, multi-time scale simulations, and molecular dynamics models. A complex system may be modeled by a multiplicity of these**

**models which may need to run simultaneously for self-consistent system design space studies. On top of these applications, ML is running and can steer the focus of the simulations to study abnormal activities and behaviors.**

In these complex calculations several individual applications are running in parallel, with a total number of up to 50,000 processes operating simultaneously, interacting with each other and sharing memory. **To program the complex memory hierarchy and to divide the work between the different memory resources, will become even more complex in the future.** This is simply not practical, new ways have to be found to manage the data movement, factoring, and reduction in an automated and hardware-driven manner.

Furthermore, the processes need to be secure and isolated from each other, which requires support in the memory system. The isolation boundaries in the memory system have to be managed with fine granularity to support complexity and security. Another challenge is the persistence of data. The persistence between hierarchies must be automatable for true performance and usability<sup>6</sup>. From today's applications, the fine granularity moving in and out of execution requires very fast switching of state, thus reducing the state-switching latency.

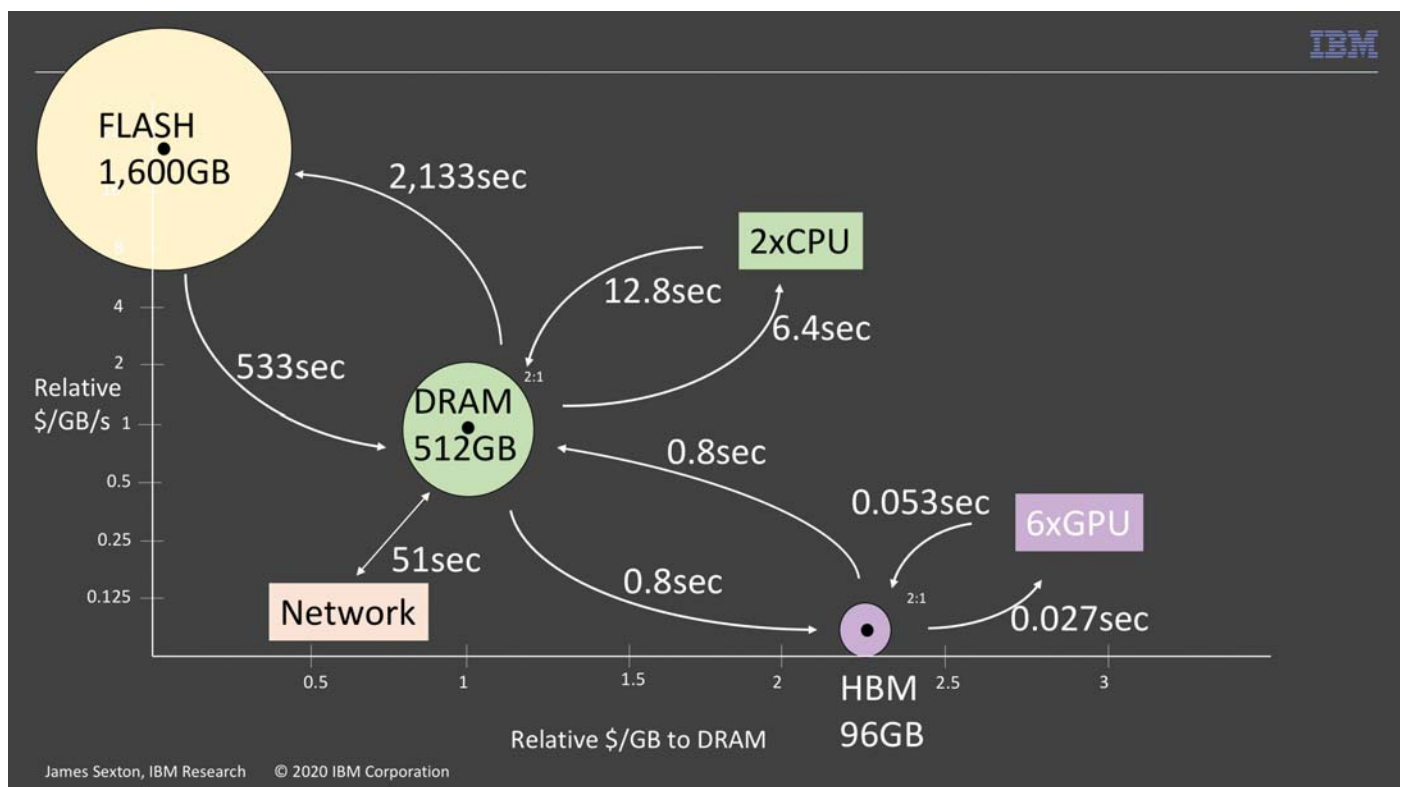


Figure 2.4: Example of complex memory hierarchy in heterogeneous HPC system<sup>6</sup> (Courtesy of James Sexton, IBM Research)

It is recognized now that moving computing near or down into the memory is critical for total system performance and energy efficiency especially for processes where the communication to computation ratio is large. However, ease of programmability needs to be maintained<sup>6</sup>.

## Workload-centric memory organization

Both required memory bandwidth and latency depend on application drivers. Petabytes of memory should be fast, if needed, and should, accordingly, have low latency if needed. Additionally, for HPC, the HBM memory needs to be cheap and energy efficient to enable increases in the memory capacity of accelerators. That memory is too small today and needs a significant increase within the next five years. The hardware should be suitable for the entire range of applications, and heterogeneous systems bring it all together<sup>7</sup>.

**The memory system organization/hardware should allow optimal support of a vast range of applications, heterogeneous systems are the primary pathways in this regard. For example, in the field of biology which is another important application for HPC, in particular, the omics<sup>ii</sup> data is exploding (Figure 2.5<sup>8</sup>), and much bigger datasets need to be stored, accessed, and analyzed. A distributed infrastructure that is linked to a central metadata store needs to be flexible for the evolving data sets, tools, workflows, etc., including expanded capabilities for analysis. In all these huge systems it is critical that the systems are not too complex or that the complexity is taken care of automatically and that they remain programmable.**

Besides HPC and data centers, another key application driver for memory and storage is autonomous driving. Already today, a car makes use of a variety of

different memory types and different memory technology nodes distributed across the system<sup>9</sup>. Examples include *DRAMs for video and graphics processing, NOR/NAND Flash for code, instructions, and data storage, EEPROM for data logging, and embedded memory in CPU like SRAM. Also, emerging NVM memories like PCM and MRAM are used for wake up, personalization, tuning and learning, diagnostics, and data logging.* The major drivers for automotive memory consumption are (i) automotive connectivity, (ii) infotainment and in-vehicle experience, and (iii) advanced driver-assistance systems (ADAS) and autonomous driving<sup>9</sup>.

The timeline for the ADAS and automated driving is shown in Figure 2.6. In fact, “automobiles are becoming high-performance data-processing compute servers on wheels”<sup>9</sup>. A number of sensors like Rada, Vision, and Lidar are combined with monitoring the driving and the driver, as well as the status of the vehicle itself. All these processes that enable simple and complex decisions for fast action are very data- and memory-intensive. **It is expected that moving from level two to level five of autonomous driving within the next 10 years will lead to a 10x increase in memory usage. The level five autonomous drive is estimated to require a 100x processing increase from where we are today in order to process the additional sensing inputs. This drives up NVM usage for code and deep ML data storage, as well as the SRAM/ DRAM usage required for the sensor data processing.**

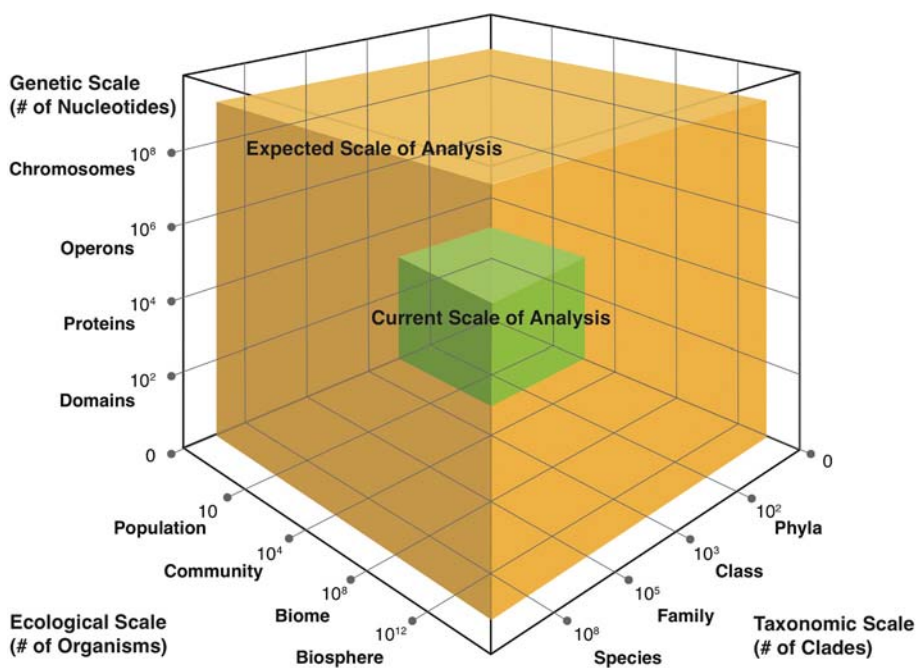


Figure 2.5: Advances in sequencing and omics technologies have far outpaced data infrastructure<sup>8</sup> (Courtesy of Kjersten Fagnan, DOE Joint Genome Institute)

<sup>ii</sup>Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.



In addition, the infotainment and in-vehicle experience will continue to demand increase in memory capacity. Different systems are available with which the occupants interact in different ways via voice commands, gestures, and perhaps even augmented reality. Over-the-air (OTA) updates and security needs require additional non-volatile memory and data logging, so memory usage is increasing (see Table 2.2). Consequently, it is expected that NVM storage needs for autonomous driving are growing by more than 100x in the next decade and the in-vehicle infotainment will demand 10x more NVM over the next decade. Another critical challenge is to make real-time applications work where multiple layers of caching and storage of the same content within a system are needed. This is an area where system optimization is possible.

Further research is needed to develop memory subsystems and optimize the interfaces for improved latency and bandwidth. *Emerging technologies, such as magnetic or resistive memories, are very interesting for meeting the demands of power, density, and high-temperature operation.*

The system form factors of the variety of applications at the edge have similar issues and are in competition with the disaggregation concept. New memory architectures need to be developed that are more efficient about memory management and usage. Therefore, a better understanding of the memory usage profiles for the different applications needs to be acquired to comprehend further memory architecture optimization

Table 2.1: Automotive electronics: applications and functions/features<sup>9</sup>

| Applications Domains           |  | Functions/Features  |
|--------------------------------|--|---|
| Functional Safety and Security | Automotive Connectivity                | <ul style="list-style-type: none"> <li>• Smart Car Access</li> <li>• Vehicle Network Processing/Gateway</li> <li>• V2X Communications</li> </ul>                          |
|                                | Infotainment and In-vehicle Experience | <ul style="list-style-type: none"> <li>• Instrumentation</li> <li>• Infotainment</li> </ul>   |
|                                | ADAS and Autonomous Driving            | <ul style="list-style-type: none"> <li>• Radar, Vision, Lidar, ...</li> <li>• Safe Central Compute for Assisted/Autonomous Drive</li> </ul>                               |
|                                | Powertrain and Vehicle Dynamics        | <ul style="list-style-type: none"> <li>• Power Train Control</li> <li>• Active Suspension Braking/Stability Control</li> <li>• Steering</li> </ul>                        |
|                                | Body and Comfort                       | <ul style="list-style-type: none"> <li>• Interior Exterior Lighting</li> <li>• Tire Pressure Management</li> <li>• Electric pump/motor control</li> <li>• HVAC</li> </ul> |

Major drivers for Automotive memory consumption

Table 2.2: Trends in automotive nonvolatile memory and storage (Source: "A look at Automotive Memory in the E-Mobility Era", SK hynix Newsroom, July 30, 2020)

| Application             | 2020      | 2023       | 2025       | 2030       |
|-------------------------|-----------|------------|------------|------------|
| In-Vehicle Infotainment | ~64/128GB | ~128/256GB | ~256/512GB | ~512GB/1TB |
| Autonomous Drive        | ~8/64GB   | ~128/256GB | ~512GB/1TB | ~1/2TB†    |

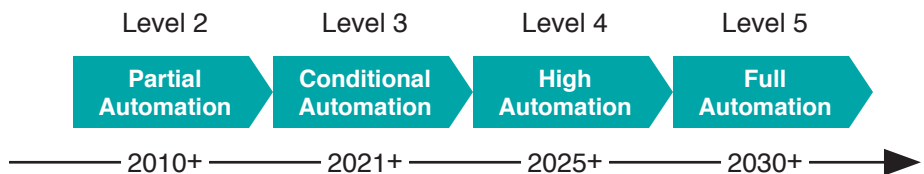


Figure 2.6: Timeline for the ADAS and automated driving<sup>9</sup> (Courtesy of Thomas Jew, NXP)

opportunities. Considering the rapid changes (for example, in AI algorithms), **workload-dependent memory hierarchies and SRAM/DRAM balancing are interesting concepts to be developed to improve flexibility.** Memory is becoming a big part for all

these edge applications, where systems need to be more and more tailored to use cases and integrated through custom packages. Flexibility for different use cases needs to be built into the design. Of course, the memory cost remains essential.

## 2.3. Mobile and IoT Computing Perspectives on Memory Technology

### Overview and needs

Information processing and communication technologies are critical elements to humankind advancements in the knowledge and scientific understanding of the world. In recent decades, the unprecedented improvements in performance, energy efficiency, density, and cost of the computation and communications capabilities have synergistically accelerated those advancements by making them broadly accessible and ubiquitous. The rate of advancement has relied on sustained innovation spanning the entire systems stack at each level and across levels in the hierarchy from software, algorithms, architectures, circuits, devices, interconnects, and materials. The demand for new mobile applications and functionalities, along with the relentless quest for higher levels of energy-efficient computing performance, continue to drive technological changes. By end of this decade, these performance improvements will stall, as the underlying logic, memory,

Annual Size of the Global Datasphere

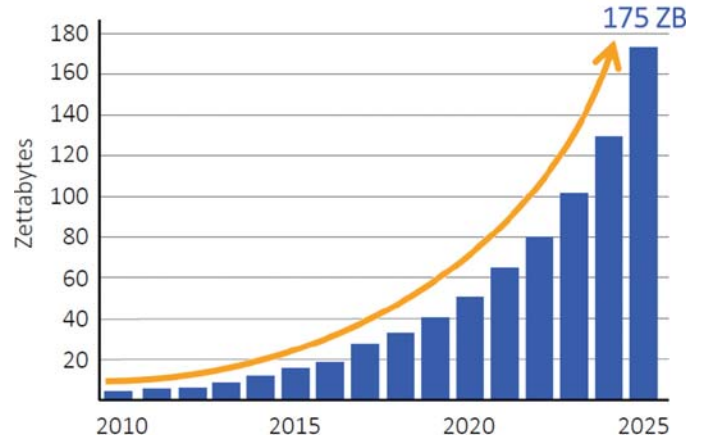
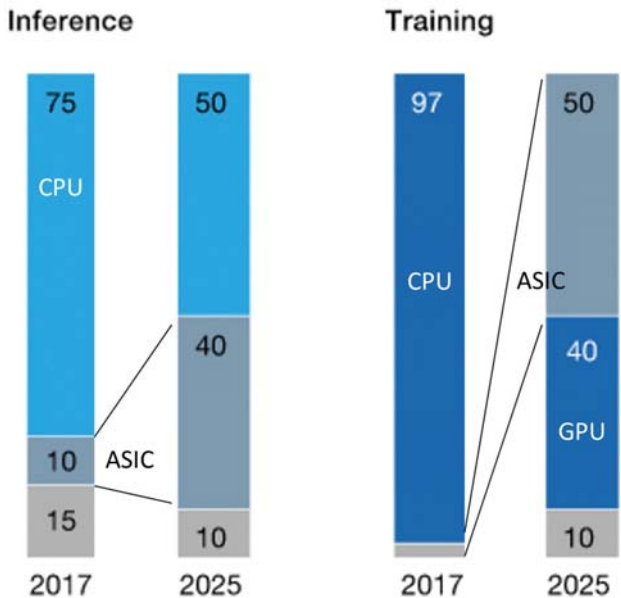


Figure 2.7a: Exploding volumes of information, a proxy for increasing information processing demands both in data centers, at the edge, and in end devices. (Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018<sup>10</sup>; adapted figure courtesy of Gary Bronner, Rambus<sup>11</sup>)

### Data-center architecture, %

ASIC<sup>1</sup> CPU<sup>2</sup> FPGA<sup>3</sup> GPU<sup>4</sup> Other



### Edge architecture, %

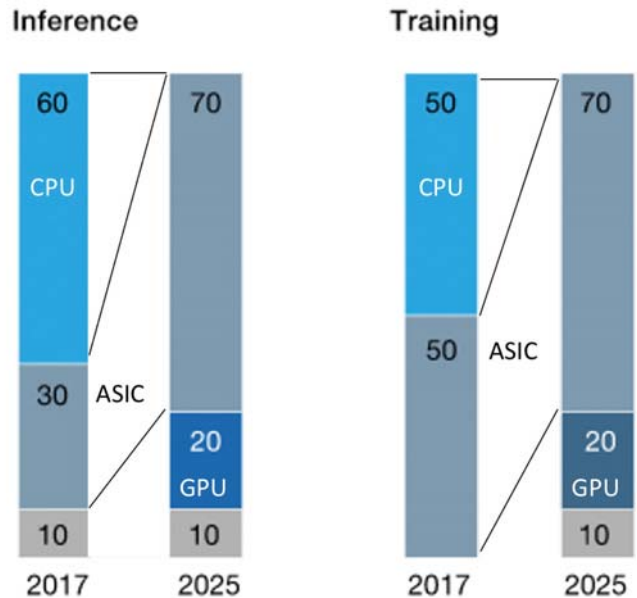


Figure 2.7b: Growth of AI-related computing workloads at the edge is projected to grow at a faster pace than the data centers over this decade. (Source: McKinsey & Company<sup>12</sup>; adapted figure courtesy of Carlos H. Diaz, TSMC)

and storage technologies run into power-scaling limitations associated with the deterministic conventional computing, the current neural networks supporting AI, and the fundamental physical scaling limits of current devices and interconnect technologies. This section seeks to identify memory requirements and solution scenarios that can drive or support dramatic mobile and IoT computing paradigm changes over this decade to maintain the promise of Moore's Law economics.

## Memory systems for AI

*Artificial intelligence operating on abundant data is expected to be a dominant application driver for mobile and IoT throughout all segments, from consumer and industrial to automotive and fast-growing IoT.* The information volume will continue to grow unabated, as exemplified in **Figure 2.7a**. The AI-related workloads will continue to shift to the edge and end devices, and those workloads will increasingly comprise training (as illustrated by **Figure 2.7b**) and expanded cognitive abilities (as shown in **Figure 2.7c**).

*New functionalities and capabilities for sensing, processing, and actuation will add to the compute workloads of mobile and IoT devices.* To that end, disruptive and sustainable performance, as well as energy efficiency improvements and scalable system integration capabilities will be required. Over the next decade, energy-efficiency enhancements of 100X or more, along with corresponding performance improvements, will be needed to enable new cognitive capabilities for the increased computational workload demand that fuels industry growth to positively impact the society at

large. The innovations to effect that level of change span from software to hardware and from architectures to the fundamental building blocks (in the form of design primitives for circuits), and they must include the essential constituents of underlying semiconductor memory, logic and integration technologies. Stepping up research during this decade is key to identify and distill tangible platform-capable alternatives to state-of-the-art logic transistors, memory elements, and integration approaches to meet aggressive energy-efficiency and performance goals stated above. In addition, new system scale-up approaches that utilize the innovations in transistor, memory, and integration technologies are crucial.

## IoT and automotive perspectives on memory

Attaining goals in energy efficiency, performance, and system integration need to be met, while also meeting reliability needs associated with unique environmental requirements specific to the mobile and IoT applications. In consumer applications, semiconductor components (digital logic and memory, analog, et al.) need to have a minimum life of at least three years, with operating ambient temperature of 85°C and early failure rates less than 100ppm. For industrial applications, semiconductor components need to have a minimum life of at least 15 years, with operating ambient temperature of 150°C and early failure rates less than 100ppm. In automotive applications, lifetimes over 20 years, with operating temperature of ~150°C and earlier failure rates less than 1ppb are must. **Consequently, research in semiconductor devices for memory, logic, and integration technologies should include studies on high-temperature reliability, as well as models and methodology to push against intrinsic reliability limits (endurance, safety assurance, software/hardware updates, and security).**

## On-chip memory and its dense 3D-integration for abundant data computing at the edge

Computing engines rely on processing units and a hierarchical system of memory and storage. A representative memory system hierarchy is illustrated in **Figure 2.8**. The memory access frequency and speed are fastest at the top of the hierarchy, whereas both the memory and storage capacity increase away from the processing units, i.e., towards the bottom of the memory system. The exploding volumes of data/information to be processed demand increasing amounts of memory capacity at each level of the memory-system hierarchy. Data movement from the memory system to the processing unit has also become a bottleneck in present-day systems. To tackle these challenges, research efforts need to be stepped up in two principal directions.

### Next Generation(s) AI Challenge

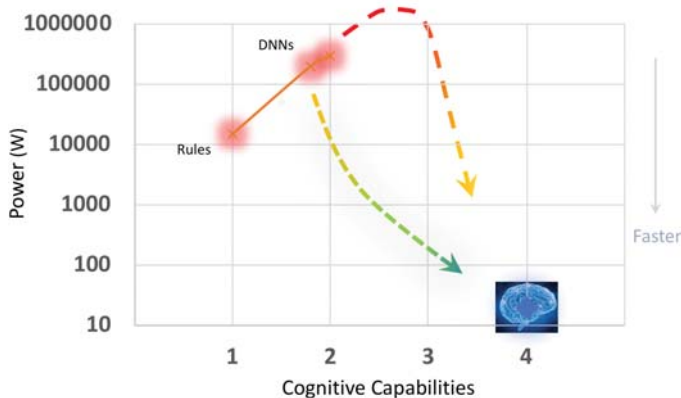


Figure 2.7c: Next-generation AI cognitive capabilities may only become pervasive (ubiquitous) if energy efficiency of the associated algorithms and compute engines improves over two orders of magnitude (> 100x) beyond today's capabilities<sup>13</sup>. (courtesy of Carlos H. Diaz, TSMC)

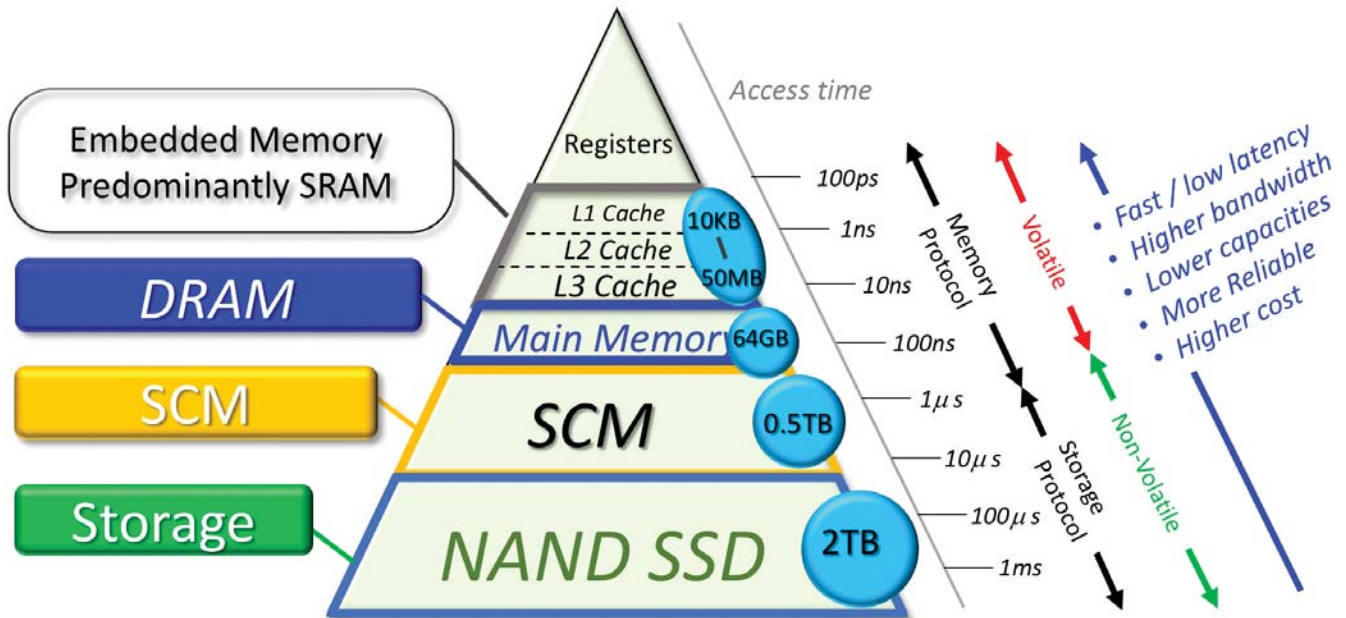
One such research vector relates to the need for denser and more energy-efficient storage (at the memory-cell and array levels) across the memory stack, while maintaining the same or better access speeds of state-of-the-art solutions. Another research vector aims to resolve memory bandwidth and latency bottlenecks through dense 3D integration of the processing units and various memory levels of the memory-system hierarchy. A third research vector must investigate efficient ways of scaling up future systems through an orchestrated approach that combines efficient multi-chip(let) integration with the on-chip integration approaches explored by the above two research vectors. Combined, these three vectors allow the creation of new systems that create the illusion of a single chip with massive on-chip memory and compute.

### Disruptive hardware solutions

While exploring solutions in these spaces, it is also important to ensure disruptive but cohesive hardware and software solutions that can adapt to slow-evolving interface standards, while also accelerating the transition of those emerging concepts. Critical research areas at the cell level include: (i) scalable memory cells capable of storing multiple bits with same or better energy efficiency and controllability/variability than single-bit-cell alternatives at comparable speed; (ii) denser, faster, and more energy-efficient memory

cells for cache and main memory applications than state-of-the-art solutions; (iii) emerging memories having non-volatile capabilities that can be quickly turned ON/OFF and are compatible with leading-edge logic technology nodes; (iv) compute-in-memory accelerators; and (v) data-compression/decompression algorithms. 3D processing and memory-system integration require advances in platform-capable die stacking or monolithic solutions. Optimal solutions may remain application-dependent and will continue to evolve, driven by density, energy efficiency, performance, and cost. Critical challenges are associated with sustainable array-level memory scaling with > 100x denser connectivity and low-cost heat removal capability > 1kW/cm<sup>2</sup>, both adhering to the tight reliability requirements outlined above. Research focus areas include roadmaps for increasingly denser connectivity, thinner stackable active layers, low-thermal budget of high-performance devices, low-cost active layer patterning and integration, and low-cost heat removal. As problem sizes continue to rapidly grow, it is unlikely that on-chip integration alone can keep up. New multi-chip systems that exploit dense on-chip integration and efficient multi-chip(let) integration to scale efficiently and flexibly for large (AI) problem sizes will be crucial moving forward. The goal should be to provide illusory performance and energy as if all memory and compute were densely integrated on a single chip.

## Memory Hierarchy and Emerging Memories



Adapted from E. Wang, IEDM Short Course, 2018

Figure 2.8: Memory system hierarchy (adapted figure courtesy of Carlos H. Diaz, TSMC)



## New information representation

Mobile computing workloads will continue to be diverse, in line with the end applications. Yet, regardless of the workload type, all are subject to tight energy bounds, increased performance, and robustness requirements, as indicated above. Furthermore, AI solutions, including those enabled through AI accelerators, need to remain platform-capable, i.e., able to support multiple application types. Consequently, key opportunities in the quest for intelligent systems with superior energy efficiency and high performance include *mastering new information representation forms that are better suited for enhancing AI's cognitive capabilities, while also facilitating near or in-memory processing*. For example, cognitive models that rely on information representation by high-dimensional vectors are fundamentally one-pass, continuous-learning, and high-level-reasoning capable. High-dimensional computing (HDC) is inherently parallel, local, and error-resilient. These fundamental attributes are well-poised to enable levels of energy efficiency not otherwise attainable by conventional DNN models. These must be leveraged to develop corresponding specialized processing units to augment AI capabilities beyond those otherwise attainable by state-of-the-art deep neural networks or convolutional neural networks. Therefore, research in new information representation and computing paradigms is also necessary, beyond the opportunities in energy efficiency that might arise from revolutionary changes to the essential constituents of the logic and memory technologies and the 3D-integration capabilities. Research is needed in novel types of information representation and processing that: (i) are amenable to near or in-memory processing; (ii) lend themselves to AI models that can scale significantly better with problem size than state-of-the-art DNNs; (iii) can support both continuous learning at the edge and increased levels of cognitive capabilities under tight mobile-application

energy budgets; and (iv) are much more error-resilient and, consequently, capable of opening up the memory and logic operation toward lower power supplies than otherwise possible. *There is great opportunity to create new AI algorithms, or information processing algorithms in general, that are aware of the underlying hardware technologies (e.g., to overcome the emerging memory challenges of write energy, latency, and endurance challenges, as well as ensure error resilience with multi-bit-per-cell storage).*

## Key areas of focus and follow-on research

### Edge- and mobile-application-driven memory systems

- Memory needs: high-power efficiency; total on-chip memory and on-card memory; low power (compute, peak current draw); and encryption
- SRAM: Enhance area and power reduction for SRAM with advanced process nodes
  - Discover 3D-integration technologies to enable more on-die SRAM
- Non-volatile (NV) memories: enable NV memories in cutting-edge process nodes for edge, data-center and auto products
  - Reduce time from technology readiness to availability in cutting-edge process nodes
  - Match SRAM's endurance and write power to enable widespread use
- In-memory compute
  - On-chip – with existing/legacy code
  - Off-chip – on main memory
- External memory: Higher bandwidth
  - ~ GB/sec but at lower pJ/bit such that HBM-like GB/sec fit into a 20W edge card

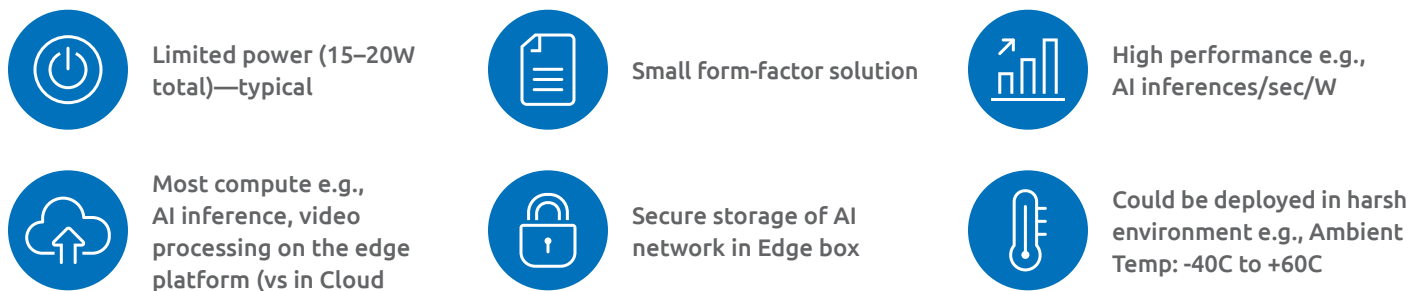


Figure 2.9: Edge computing characteristics and drivers for memory technology<sup>14</sup> (courtesy of Rashid Attar, Qualcomm).

### Memory systems for AI

- Research needs: energy-efficient memory cells and energy-efficient algorithms that minimize bandwidth needs and maximize performance and accuracy
  - Rapidly improve energy efficiency as the data volume to compute increases
  - Expand memory bandwidth for AI applications
  - Discover memory-efficient algorithms (more Ops/byte)

### IoT and automotive perspectives on memory

- Identify and develop alternatives to 6T SRAM for on-chip code/data
  - Smaller area/bit, low leakage, baseline process compatibility, and zero/low process cost adder
- Step up function improvement in performance for emerging memories/cells and scalability
- Cell-level research goal: >10X improvement for key parameters aiming for 100-1000X improvement in power/energy with enhanced reliability and lower cost/area
- Achieve area- power-efficient memories by minimizing IR drop, variations, and stochastic writes
- Identify Flash function replacements @ 28nm and below
  - New capabilities needed with endurance > 1M cycles
- Improve MRAM endurance to attain > 1E10 cycles with enhanced thermal stability, magnetic immunity, and low BER, while also reducing write current

### On-chip memory and its dense (3D) integration for abundant-data computing at the edge

- Address fundamental limits that are posing major challenges in on-chip-memory capacity and connectivity (memory wall), area scalability (miniaturization wall), and the logic and memory power walls of integrated systems
- Ensure “enough” on-chip memory through dense 3D (non-monolithic or monolithic) integration of memory and logic
  - Such technologies are now practically possible in commercial silicon foundries (using silicon and non-silicon technologies)

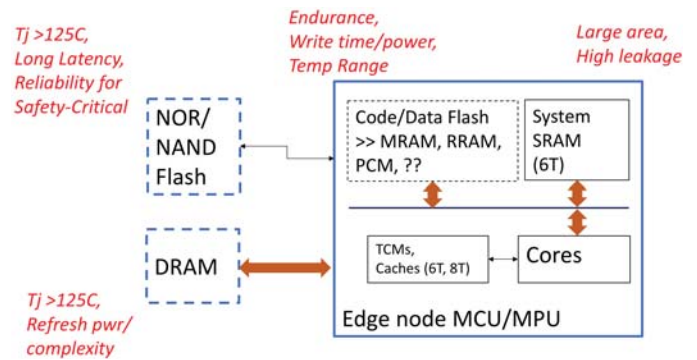


Figure 2.10b: IoT and automotive memory usage and requirements<sup>15</sup> (courtesy of Kelly Baker, NXP)

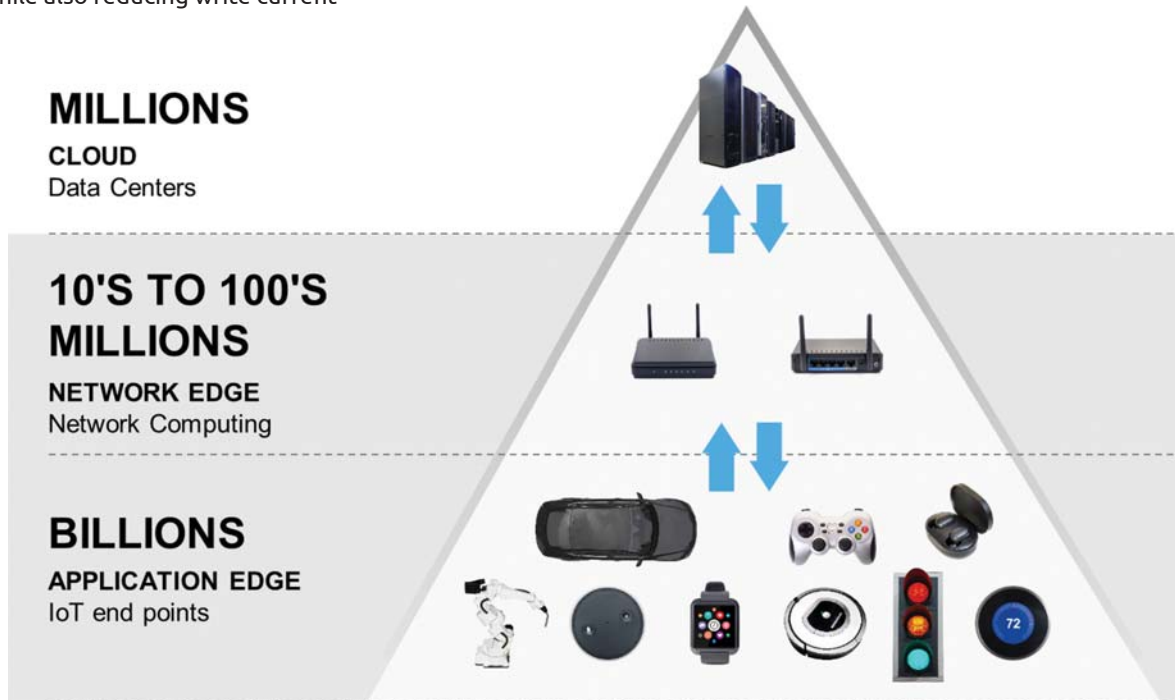


Figure 2.10a: IoT spans “network edge” to “application edge,” demanding high-performance gateways and routers, as well as ultra-low-power sense, compute, actuation, and connectivity<sup>15</sup>. (courtesy of Kelly Baker and Gowrishankar Chindalore, NXP)

- Discover new ways for scale up
  - As problem sizes grow at fast rates, Illusion systems (Figure 2.11a) become crucial. Illusion employs an optimized combination of “enough” on-chip memory (through 3D, multi-bit cells), quick chip wakeup and shutdown, and special multi-chip mapping to scale effectively for a wide range of (AI) workloads.
- Create a new path to scale future (AI) systems through orchestration among dense on-chip integration, efficient multi-chip(let) integration, and Illusion (Figure 2.11)
- Create memory-technology-aware algorithms to overcome write (latency, energy, endurance) challenges and ensure error resilience (e.g., for multi-bit storage)

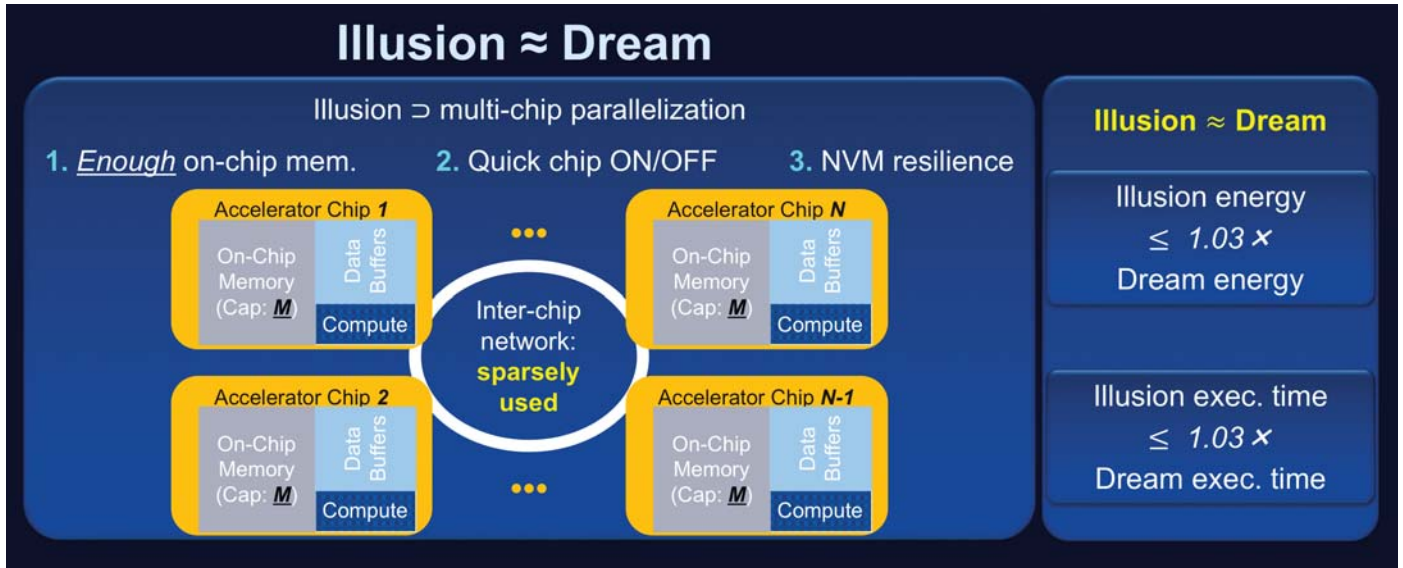


Figure 2.11a: “Illusion” system has the potential of closely approximating the performance, energy, and scalability of a “dream” chip that could hold all memory/compute on-chip<sup>16</sup>. (courtesy of Subhasish Mitra, Stanford University)

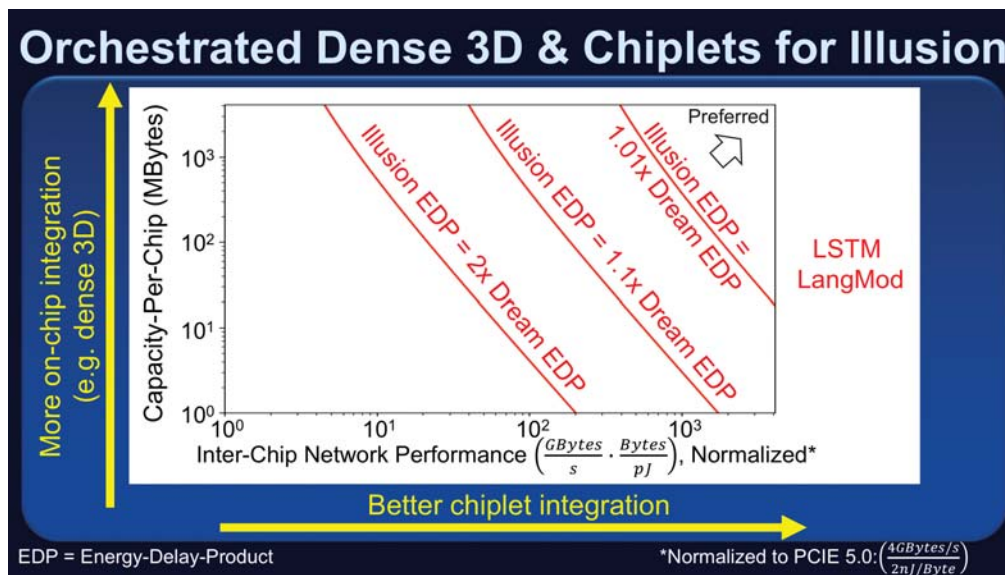


Figure 2.11b: Orchestrated dense on-chip integration, efficient multi-chip(let) integration, and Illusion achieve scaled (AI) systems not possible with the individual approaches alone<sup>16</sup>. (courtesy of Subhasish Mitra, Stanford University)

## In-memory computing across technologies<sup>17</sup>

- Analog IMC trades energy efficiency and throughput for SNR
  - Need high row parallelism (N) for energy efficiency and throughput
  - Level of row parallelism (thus energy/throughput) is limited by SNR
- Establish higher cell SNR and resistance
  - Compute metrics (TOPS/W, TOPS/mm<sup>2</sup>)
  - Bigger cell, if needed
- Address memory-write costs
  - Lower write energy and higher write endurance
  - e-NVM technologies will need to address write energy and endurance
- Create robust architectural and SW abstractions

## Memory in machine vision

- Address exploding volume of video (social media, surveillance)
  - 3D-point clouds and super/hyper-resolution imaging add to volume
- Focus on transformation from archival store-and-play video to interactive video
  - Video content analysis (e.g., content rating)
  - Content-based queries (e.g., surveillance)
  - Synthetic video and video summarization (e.g., digital advertisement, assistive vision)
- Address costly data movement
  - Need compute at various levels of the memory hierarchy, as well as new data representation and in-memory compute primitives
- Discover new memory technologies with more energy-efficient primitives (SRAM, CAM, etc.)
- Screen viable emerging memories using variability as a critical metric

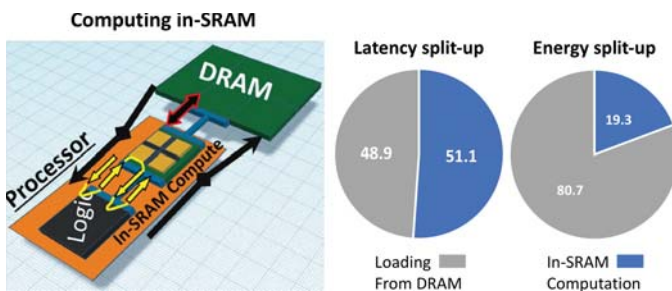


Figure 2.12: Moving data from off-chip memory to processing unit is a computing bottleneck<sup>18</sup>. (courtesy of Vijaykrishnan Narayanan, Penn State)

## 2.4. HPC & Data-Center Computing Perspectives on Memory and Storage

### Overview and needs

The large-scale high-performance-computing community is evolving. *While the traditional focus in scientific computing remains prevalent, e.g. the solution of partial differential equations, the integration of AI/ML methods and various types of data and graph analytics represent the heterogeneous computing workloads of the future.* These new workloads are both data-intensive and compute-intensive, requiring a wider range of data-access patterns than are found in traditional HPC applications. In data-center computing, AI/ML and data analytics are also experiencing tremendous growth, far higher than traditional commercial transaction processing workloads. The integration of commercial workloads with AI/ML and data/graph analytics present significant challenges and opportunities to the memory and storage requirements. Large-scale heterogeneous system architectures are common in both HPC and data-center deployments, but these systems have separate CPUs with DDR/HBM memory and GPUs with GDDR/HBM memories, all integrated with a system-interconnect-network fabric. This system interconnect also extends out to I/O nodes and storage. The current system architecture model is a challenge for future heterogeneous workloads because the separate memory pools, perhaps even within a single compute node, lead to performance, energy, and software inefficiencies. **Future memory solutions are required that can facilitate low-level integration of heterogeneous architectures, rather than continuing to raise the barriers to integration found in current system architectures.**

Memory is one of the main concerns in today's HPC and data-center computing. In fact, "memory is a number-one pain point and limitation to high compute performance and power improvement; storage has been progressing in terms of performance and cost, memory is falling behind" (Nafea Bshara, AWS<sup>19</sup>). Photonics and networking can support fast access to storage, and, with it, the performance/cost trajectory of available storage solutions is headed in the right direction<sup>19</sup>.

Memory bandwidth is a common limitation for both HPC and data-center applications. The second area of concern is capacity scaling. Large-scale system architectures are evolving towards tightly coupled fast memory, supplemented by a much larger shared pool of "far" or "disaggregated"



memory. This high-level perspective indicates that perhaps disaggregated memory solutions that provide very high levels of internal bandwidth for data movement—and some degree of in-memory computing—may be interesting areas for advancement. *Disaggregation of memory and compute has data-center benefits due to flexibility to integrate systems with different development cycles of compute and memory technologies, as well as the need for architectural flexibility for different workloads. However, disaggregation also results in performance constraints due to lower bandwidth and higher latency, which may limit adoption in HPC systems.*

### Workload evolution

The HPC and data-center space is rapidly being consumed by AI and ML applications. As can be seen in Figure 2.13, Deep Learning, which underlies the largest AI applications today (e.g., in autonomous vehicles, fraud detection, digital advertising, and many other domains), is the largest category of AI application in this space. Here, the work, which consists of machine-learning training, and machine-learning inference, is more commonly done with domain-specific accelerators built on custom silicon or FPGAs. In the data center, AI and ML applications have paralleled the growth of graphics applications as the driver for mobile device architectures. *Data-center workloads, most importantly, the ML training workloads, involve largely random access due to multi-core, multi-thread, and multi-tenant scaling of parallelism. In contrast, HPC workloads are designed to have a large degree of locality and thus are very dependent on streaming bandwidth from memory. In both cases, it appears that bandwidth is an important factor, but memory-level parallelism may play a more important role in data-center applications.*

### Architecture evolution

As can be seen in Figure 2.14<sup>20</sup>, DRAM represents the majority of the cost (and even a larger majority of the silicon area) of modern computing platforms in the data center.

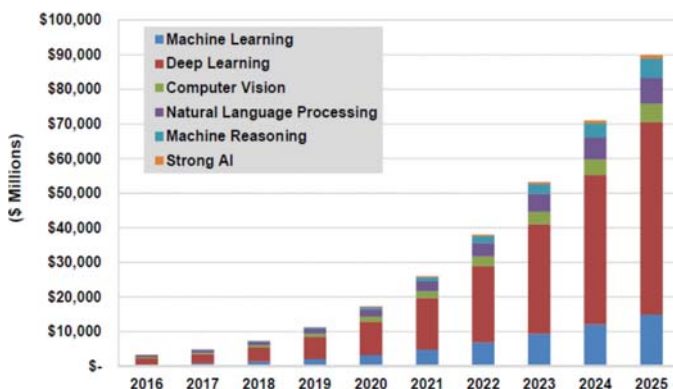


Figure 2.13. Revenue by Artificial Intelligence application category (courtesy of Tractica)

In HPC applications, tightly coupled HBM (High Bandwidth Memory) memories are becoming commonplace, with support for remote direct memory access (RDMA) used for large scale numerical solvers for weather, structural, and fluid modeling. In data centers, to reduce memory cost and to improve memory usage efficiency, far memory and disaggregated memories are being implemented, thereby improving uniformity of memory provisioning, extending capacity expansion range, and reducing node-to-node data-movement requirements in distributed systems.

*Accelerators are becoming much more common, with large cloud and HPC providers generating their own custom silicon (for example Google’s TPU or AWS Inferentia) or accelerating applications with COTS accelerators and FPGA solutions.*

As can be seen in Figure 2.15<sup>19</sup>, logic cores are currently consuming on the order of 500W and are projected to consume on the order of 1KW by the middle to end of the decade.

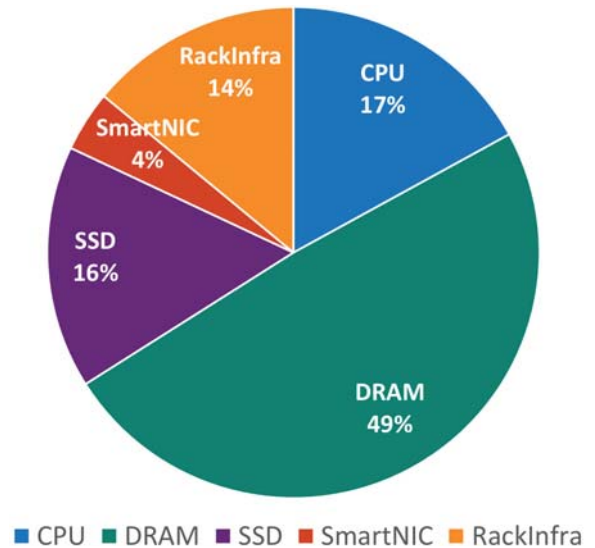


Figure 2.14: Compute node costs by component<sup>20</sup> (courtesy of Sailesh Kottapalli, Intel)

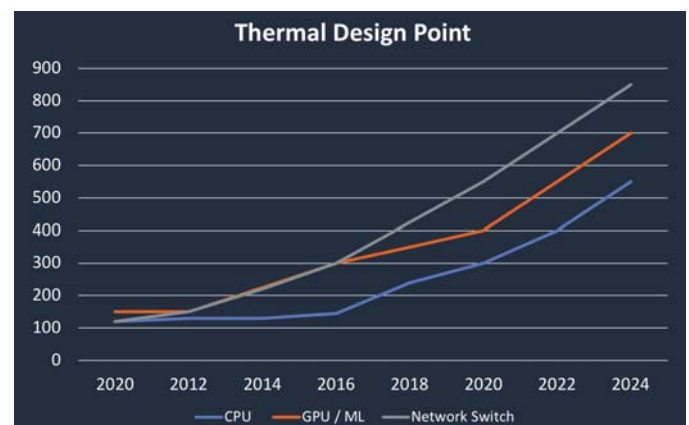


Figure 2.15: Thermally limited operating power by component type<sup>19</sup> (courtesy of Nafea Bshara, Amazon Web Services)

Thus, package-level integration utilizing heat sinks and air cooling is becoming limited by thermal constraints. **While more exotic cooling solutions exist, these tend to be impractical in production environments due to the cost and support requirements. These thermal and power constraints for accelerators are also driving increased demand for DRAM memory in HBM packages and media formats.**

Innovative memory and storage architectures are continuously being proposed by solution suppliers and academia. For these to come to fruition, however, a number of conditions have to be met: **a sufficiently large demand for such solutions must be present for the technology to**

**justify the development cost of the new innovation; the innovation must be sufficiently compatible with existing solutions so it can be seamlessly and timely adopted by the industry, including application developers and data-center operators; and that the hardware and software development cost is not exorbitant.**

An excellent example of innovation that meets all of these criteria can be found in the advent of High Bandwidth Memory (HBM) and illustrated in Figure 2.16<sup>21</sup>. The HBM architecture utilizes relatively standard DRAM memory silicon on top of a higher-performance silicon layer based on a 3D-stacked DRAM process interconnected with a revolutionary Thru-Silicon-Via interconnect technology. This provides exceptionally high memory bandwidth to processors and accelerators through a silicon interposer. In terms of \$/GB, this technology is exceptionally expensive. *For HPC applications, however, the metric of interest is not GB but GB/s—HPC applications favor bandwidth over capacity. From this perspective, HBM has clear advantages over DDR memory in the HPC space on both power and cost*, as can be seen in Figure 2.16 [Daniel Ernst, PhD, Hewlett Packard Enterprise].

*There has recently been a focus on memory architectures targeted towards optimizing bandwidth for a number of disparate platforms (Figure 2.17<sup>22</sup>), including HBM for graphics, AI, and HPC workloads, as well as GDDR for graphics and DDR for general-purpose compute applications.*

Another interesting scaling challenge emerging in the data center and in scientific computing is related to the scaling of memory bandwidth and compute capability per silicon die (Figure 2.18<sup>19</sup>). As it becomes increasingly difficult to scale

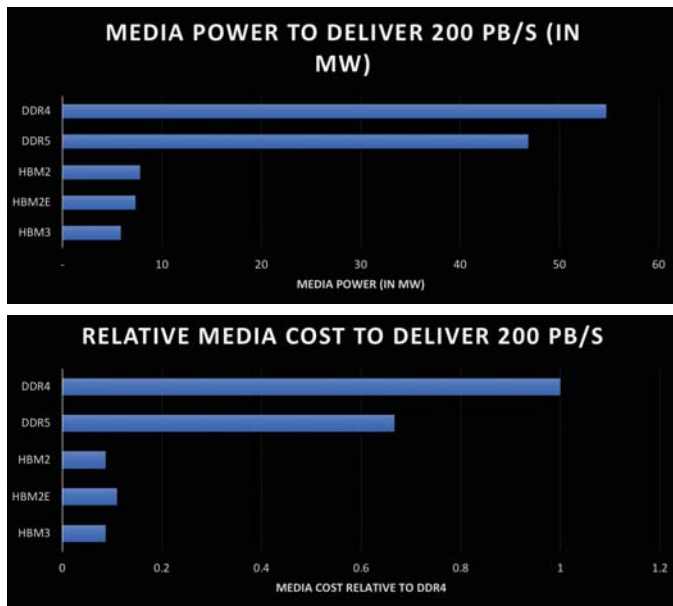


Figure 2.16: Memory power and cost by media type<sup>21</sup> (courtesy of Dan Ernst, Hewlett Packard Enterprise).

Near Term Memory Solutions : Bandwidth ↑, Capacity ↑, ~same Latency

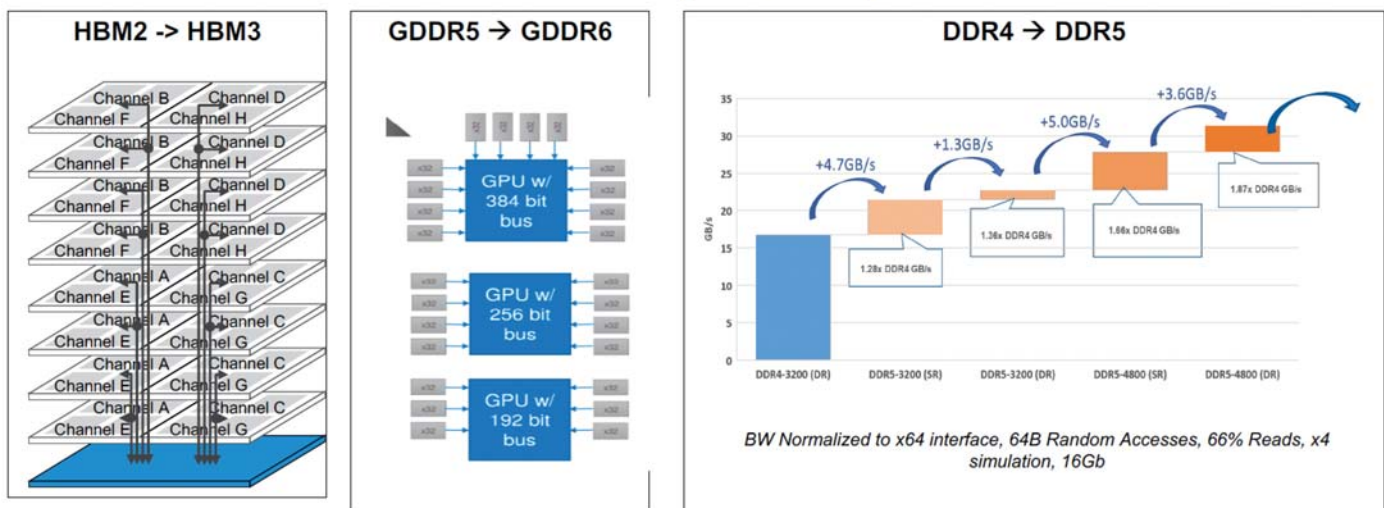


Figure 2.17: Datacenter memory technology options<sup>22</sup> (courtesy of Steve Pawlowski, Micron).

transistor density and operating frequencies, more and more processor vendors and data-center operators are turning to parallelism to address scaling challenges. Examples include the use of clusters of GPUs for ML training and for numerical computing in HPC, as well as the increasing trends for many core architectures. This results in a continued exponential scaling of compute capability per silicon die, especially when we consider accelerators like those used for computation of machine learning and AI workloads. **With increasing memory capacities, memory architectures can, in principle, keep pace with this increasing demand for bandwidth. The limitation, however, is the interfaces to the memory subsystems.** High Bandwidth Memory (HBM) addresses this challenge by providing a very wide interface between the compute accelerator and the memory. On the other hand, GDDR memory addresses this challenge by increasing the frequency of the interface. Both of these solutions, however, have their drawbacks, largely realized in higher cost-per-bit and limited capacities.

### “Far” memory solutions

In many ways system memory in modern computing platforms can be viewed as a cache with the virtual memory system acting as the tags for large cache (Figure 2.19<sup>23</sup>). Thus, future main memories need not be external, provided they are large enough to serve the cache functionality required by the operating system. This opens doors to incorporate numerous emerging memory technologies that can be integrated with the CPU in a single die, providing hundreds of GB in a monolithic fashion.

*The advent of cache-coherent “far” memory protocols like CXL, CCIX, Open-CAPI, and Gen-Z have opened a number of doors with regard to memory subsystems. “CXL is a good start, but needs to be 10X faster” [Nafea Bshara, AWS].* In their simplest form, they provide a means by which cloud and HPC providers can expand memory capacity and bandwidth by

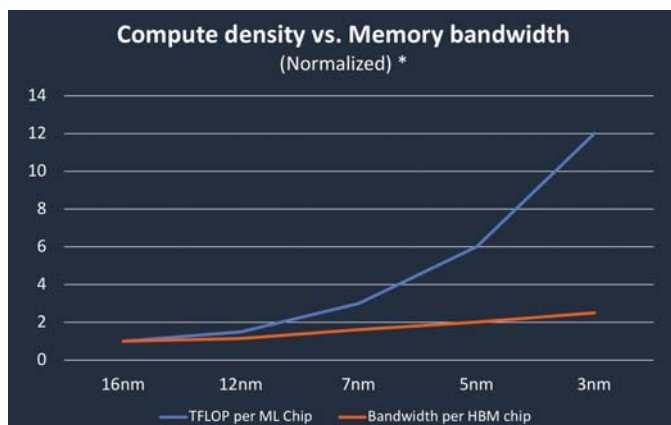


Figure 2.18: Normalized compute density versus memory bandwidth<sup>19</sup> (courtesy of Nafea Bshara, Amazon Web Services).

attaching memory to interfaces that would traditionally be used for storage. With the addition of a switching fabric, these protocols allow service providers to configure these large memory subsystems to be shared among multiple servers in a rack or in a data center. As such, they allow the memory to be disaggregated from the compute infrastructure and provisioned optimally among the many workloads running on the many servers within the distributed computing system. *In addition to adding flexibility for both the data-center architect and the memory-solution provider, these interfaces open the door for subsystem-level acceleration and, with the optional inclusion of both memory and storage behind this interface, allow for the possibility of autonomous data movement within the combined memory/storage subsystem, which in part, alleviates the bandwidth bottleneck into the processor.*

## Main Memory Is a Cache

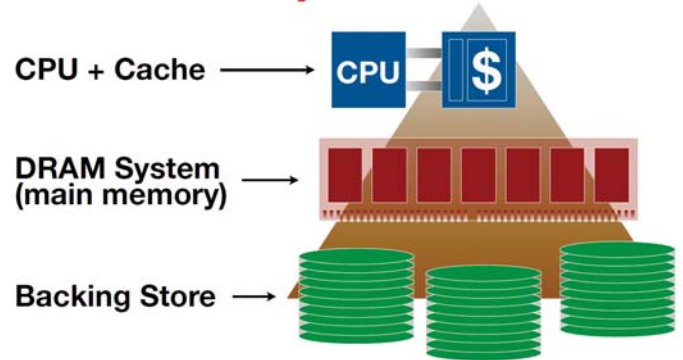


Figure 2.19: Main memory as cache<sup>23</sup> (courtesy of Bruce Jacob, University of Maryland).

### Storage

*Although storage may not be the top pain point in the data center, it is still a very important aspect of the data center. Storage is also exceptionally important for the evolving spectrum of data-intensive HPC applications.* In terms of bit capacity, storage dwarfs all the memory systems combined. In essence, all memory systems in use today are simply caches for the storage systems. In other words, the storage systems (which might include network-attached, block-based storage, shared filesystems, object/blob storage systems, tape archives, and other architectures) are the only subsystems in the data center that have the resilience features required to act as the final resting place for the precious data in a data center. There is no question that there are clear competitive battlegrounds between performance and capacities of NAND or emerging memory SSDs with HDDs and tape to support these differing types of storage. Storage-class memories like 3D XPoint are emerging to provide fast random access, reasonable power efficiency, and nonvolatility. In this role



they may provide an interesting additional tier of memory between the byte-oriented DRAM memory and the block-oriented NAND SSDs. In this growing niche, they provide improved performance over block-oriented NAND SSDs at a higher cost, and they provide a higher capacity, lower performance, and lower-cost solution for relatively fast random-access memory.

### Near-data processing

The model of near-data processing is becoming an age-old concept. Numerous near-data processing proofs of concept can be found in industry and academia. Industry support around a “computational storage” standard is beginning to emerge [www.snia.org]. An advanced concept for compute-near memory is described in Figure 2.20<sup>22</sup>. Some examples of this include the repurposing of dense-memory arrays for compute, as in analog multiply-accumulate on resistive arrays; large-scale state machine, based on a DRAM process like the Automata processor; or in more direct approaches of placing conventional microcontrollers and accelerators for data manipulation onto memory silicon. A similar approach demonstrating the value of tightly coupled memories have represented this strategy using FPGAs with embedded CPUs (as hard macros or as soft IP), with decades of success, especially for DSP and other high-performance embedded applications.

In modern computer architectures, however, there are a number of barriers to adoption of these technologies. The primary barrier to adoption of these technologies is that they are often very difficult to program. **Entirely new programming paradigms need to be adopted. Since it is typically not the case that the entire application**

**can be accelerated on this special-purpose hardware, the new programming paradigm must be seamlessly integrated into more conventional programming models, for example, by adding hardware-specific custom processor instructions, providing higher-level libraries, or by extending existing programming languages and compilers to support alternative programming models, including general-purpose or domain-specific languages (DSLs).** These alternative programming models, libraries, and languages must consider the implications of parallel accesses to memory. Another large barrier to adoption of compute-in-memory on conventional DDR or LPDDR memory busses is that the protocol requires deterministic latency. In such an environment, it is difficult to take segments of memory offline to allow private access for in-memory compute operations to happen. Also, modern computer architectures stripe data structures across memory dice, making compute-in-memory problematic, as the context for computation isn’t normally co-located within a single memory component. Similarly, within storage systems, the abstraction of the filesystem and storage stack—including configurable redundancy features like RAID—make it impossible for the storage devices to understand the underlying structure of the data. *The natural solution to this problem is to build memory- and storage-based near-data compute elements on a specialized accelerator. This closely resembles the model of hosting a graphics accelerator on a graphics card or a TPU processor for matrix multiplication on a specialized subsystem.*

The roofline graph in Figure 2.21<sup>22</sup> shows the relationship between operational intensity (compute operations per byte of data transferred to the compute unit) versus the

DRAM/Memory stacked on computation in logic is a step to drive greater BW/Energy Efficiency.

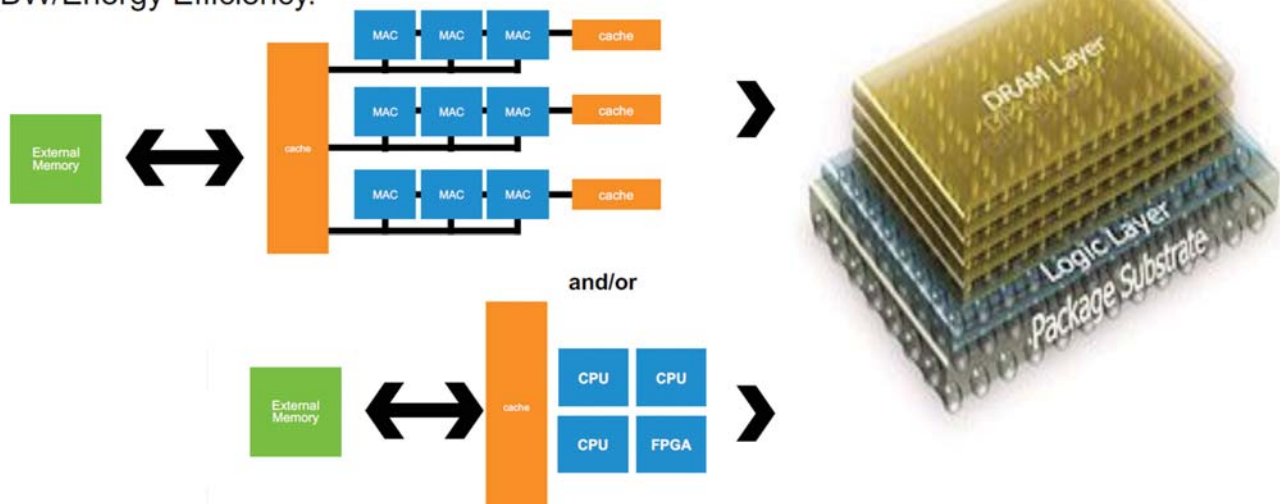


Figure 2.20: Advanced concept for near-data processing<sup>22</sup> (courtesy of Steve Pawlowski, Micron).



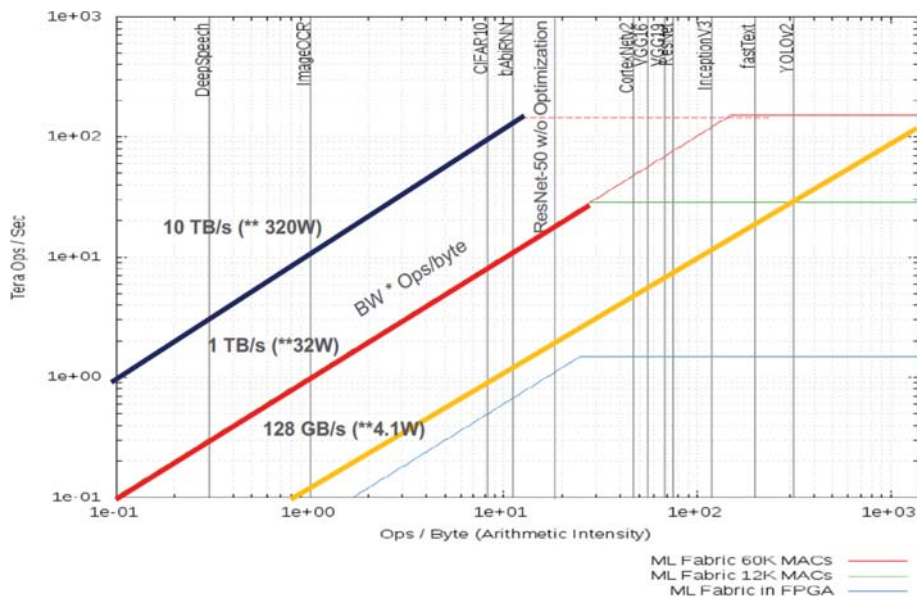


Figure 2.21: Roofline for AI/ML workloads<sup>22</sup> (courtesy of Steve Pawlowski, Micron).

compute rate (TeraOps per second). Applications with lower operational intensity are naturally bound by the rate at which data can be transferred into the processor. Increasing bandwidth for these applications increases performance. Applications with higher operational intensity (on the right of this graph) are compute bound and require a faster or more parallel processor for performance improvements. As was discussed above, memory bandwidth and the energy cost of data movement are extremely important problems in the industry today. Both factors can be improved with the enablement of near-data processing.

The advent of cache-coherent protocols, such as CXL and others, allows for such an accelerator. *To the conventional computing system, this appears as a combined “far” memory, storage, and accelerator. There are clearly large energy advantages to be had in reducing data movement for data-oriented operations, but it is unclear how much data movement must occur between main system memory and this accelerator subsystem.*

This is one area in which relatively modest investments in academia will be pivotal in enabling new technologies in the market. *Development of industry-adopted open-source frameworks that lower the programming barrier for near-data accelerators lowers the barrier to entry of such technologies. Also, supporting the current industry collaboration consortia, standardizing high-bandwidth cache-coherent memory and storage systems enables interfaces in modern computing systems where such accelerators might be installed.* Industry adoption of new technologies requires a market sizeable enough to fund the development of the new technology, as well as technology that is close enough to the extant technology to be brought

to market with reasonable investment. *AI applications and the underlying ML technologies are examples of such technologies, with the scale, breadth of use cases, and future growth needs to justify such investments. Having an interface to plug such a device into, as well as a programming model to make it easy to near-optimally program, significantly lowers the barrier for such technologies.*

### Key areas of focus and follow-on research

*We are transitioning into an age of heterogeneous computing, where computing functions will be as or more likely to be performed on a*

*domain-specific accelerator than on a general-purpose CPU. The advent of compute-centric accelerators like GPUs and AI accelerators has enabled new usages and a wealth of new businesses within the industry.* Because the energy cost of data movement between memory and processing units is high relative to the energy cost of performing the operation, these accelerators—and general purpose CPUs—are notoriously poor at efficiently performing near-data processing.

Imagining a computing system of the future that enabled near-data computing may include one that is heterogeneous in nature, with domain-specific accelerators to perform domain-specific operations. **These would obviously include elements like CPUs, GPUs, AI accelerators and configurable FPGAs. It would also be designed in such a way that data structures are naturally co-located on memory and storage devices, and so the memory and storage devices can reason about the structure of the data. The system would have high peer-to-peer bandwidth among memory, storage, and all of the heterogeneous compute elements. The interconnects enabling direct data movement among memory devices, storage devices, and compute elements need to be extremely high bandwidth and be energy efficient. Near-data processing elements would be tightly coupled with memory dice and storage subsystems, and the software and system design would seamlessly transfer near-data processing subroutines into the memory and storage components such that local processing could be completed. The software framework used to program such a system would need to be ubiquitously adopted and should be designed in such a way that the programmer need not be concerned with the underlying system architecture. The software framework should, however, optimize compute and data movement to ensure that functions with large data**

**Footprints and few compute operations happen in-situ to the memory or storage system, while continuing to transfer operands to accelerators when it is optimal to do so.**

Modern server architectures consume on the order of ~KW, have total memory bandwidths on the order of 100GB/s, and hold total storage bandwidths of 10s of GB/s. These servers typically include one to four processors and 100s to 1000s of DRAM memory dice. The internal bandwidth of *each* DRAM die in a modern DRAM design is on the order of 10s of TB/s, and harnessing this bandwidth local to the DRAM die would consume 10s of W per die. **Reconfiguration of this system architecture with a focus on near-data computing in memory and a software framework to make it easily programmable represents an enormous opportunity. The system that currently traverses 100MB per second per watt might be able to traverse on the order of a TB per second per watt. If this could be achieved, it represents an energy efficiency improvement of well over 1000X.**

For near-data processing to progress, a number of key technologies must be advanced, and significant changes are required in data-center system architectures.

#### *Software infrastructure*

A ubiquitous software framework that lowers the barrier to integration of new near-data processing elements is required. This framework must be capable of optimizing data placement and data movement within the system.

#### *Interconnects*

High-bandwidth, low-energy interfaces are required to move data between memory, storage, CPUs, and accelerators. **These interconnects and interconnect standards should be developed in such a way that they allow data movement directly among elements, rather than moving to a central processor then back out to a different element as is done in systems today. Protocols running on top of these interconnects must ensure coherence and resilience.**

#### *System architecture*

Modern server architectures were not designed with near-data computing in mind and, as such, limit the possibilities of near-data computing in many ways. For example, data structures stored in memory in a modern server always span multiple memory dice. This data layout is determined by the CPU vendor and may be dynamic or configurable. **Because of this, for in-memory compute to be possible, one must first go to great lengths in software to ensure that all the operands for an operation are co-located in the memory die where they will be processed.** Similar complications exist in storage architectures and filesystem abstractions.

For near-data computing to become ubiquitous, new system architectures must be developed. The advent of cache-coherent “far” memory and storage interfaces, such as the emerging CXL standard, is a small step in the right direction.

#### *Near-Data processing elements*

For near-data processing to be realized, near-data processor architectures must be developed. Numerous examples of such architectures can be found in industry today, but upon closer examination, it can be seen that these designs are greatly hindered by having to fit neatly within today’s server and data-center architectures. **Near-data processing-element design must be designed in the context of a full-system design that embraces near-data computing, along with the wealth of other heterogeneous computing operations becoming commonplace today.**

Finally, it should be recognized that the diverse talent required to make this industry-wide revolution happen have yet to enter our higher education institutions. **A focus on developing talent through curricula that explore and teach the concepts of near-data processing will surely bring innovation in both system and software architectures and in the applications that run on such systems. We cannot yet imagine the future applications that will run on an optimized near-data-processing data center.**

## 2.5. Memory Technologies Present and Future

### Overview and needs

The current mainstream memory technologies—SRAM, DRAM, and NAND Flash—were established from advances in semiconductor technology. Each has its respective market dominance because it fills a technological need while simultaneously lowering power and manufacturing costs. These incumbent memory technologies have proved exceedingly hard to supplant with emerging memory technologies—the latest, NAND, was created in the late ‘90s, and the oldest, SRAM, came from research in the early ‘60’s. Over the decades, many new memory-device technologies have been investigated, but none have risen to the forefront as serious challengers to the incumbents.

However, all three of the mainstream memories are now facing scaling challenges in bit density and/or performance. In order to continue performance gains in these technologies, multi-level cell (MLC) and 3D capabilities are being exploited. NAND Flash has managed to go both MLC and 3D monolithically, with vertical strings pushing over 100 layered cells, while DRAM is

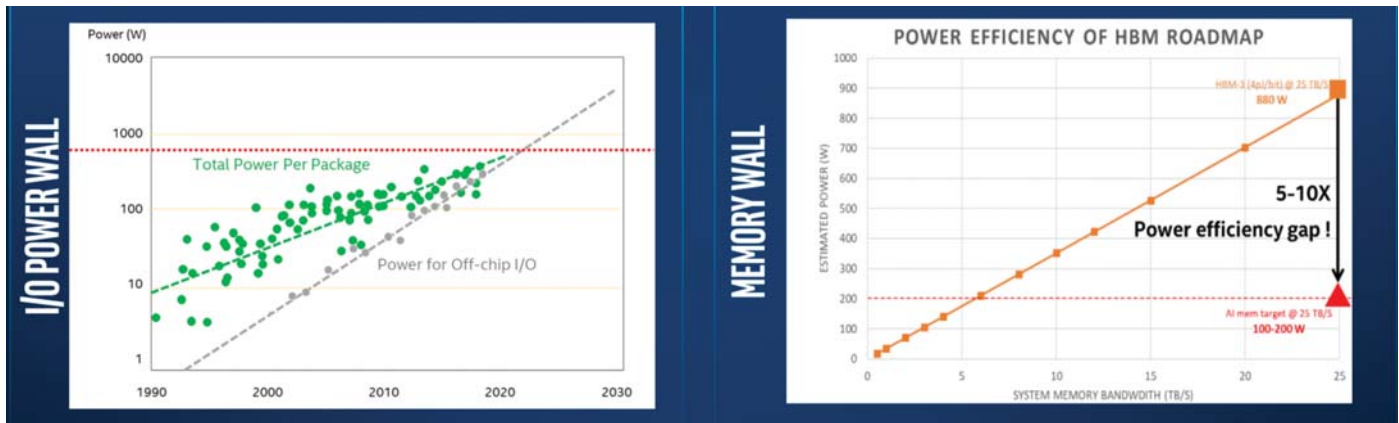


Figure 2.22: Energy from I/O and latency in moving bits back/forth are becoming limiting to compute performance gains<sup>24</sup>. (courtesy of Al Fazio, Intel)

being stacked heterogeneously using through-silicon-vias (TSV) to make high bandwidth memory (HBM) modules.

Multi-function CPU core architectures and parallelism have unleashed more power efficiency in computing and thus performance, but the increased performance has now become limited by the bandwidth to and from the memory.

*Even as memory continues to make great gains in capacity, and to a lesser extent in internal bandwidth, the ability to use that memory has become bottlenecked. Memory resides outside the compute core, consuming power for access and creating latency, which affects scenarios referred to as the “I/O power wall”—where energy is interconnect-dominated—and the “memory wall”—where bandwidth is pinned and locality constrained (Figure 2.22<sup>24</sup>).* Technologies are needed to both enhance memory density and reduce memory power consumption.

The growing interest in new computing paradigms is looking with excitement at the new features that emerging memory devices can provide. Great strides have been made of late in new methodologies for machine learning (ML) and artificial intelligence (AI). *These processing techniques often rely on parallel, nonlinear methodologies, such as neural networks, that lend themselves to weighted memory-cell implementations, like those that can be made with fine-grained, multi-level cell capabilities and analog-like behavior.*

### More efficient memory hierarchy

Quite a few cache-based strategies have been exploited to improve the memory latency and I/O power issues. Taking this further, with the use of extra, more tightly coupled layers of cache in the compute, memory hierarchy can reduce the need to go off-chip for memory access. Some emerging memory technologies have properties that make feasible new layers of cache, such as non-volatility and faster access than NAND,

particularly magnetic RAM (MRAM), phase change (PCRAM), newer ferroelectric devices like the FEFET, and the large variety of resistive RAM (ReRAM). It must be noted, however, that even though non-volatile memory reduces bit movement (and the need for refresh in supplanted DRAM), the energy barrier associated with this persistence generally results in higher write energy that must be considered in the overall design (M. Mayberry, “The Future of Compute”, 2020 IEEE VLSI). Many of the emerging memories can be stacked in 3D XPoint architectures, and such density gains can further improve performance. They are also being considered and implemented as SRAM replacements in mobile devices, largely for reduced standby power due to their non-volatile nature.

There are also performance gaps in memory beyond the compute architecture at the system level, as seen in Figure 2.23. **In designing future systems, computing architects need to be more closely coupled to memory technology, guiding memory device and material-science-level research to fill hierarchy gaps.** One could also envision the insertion of a very tightly compute-coupled layer in the memory hierarchy.

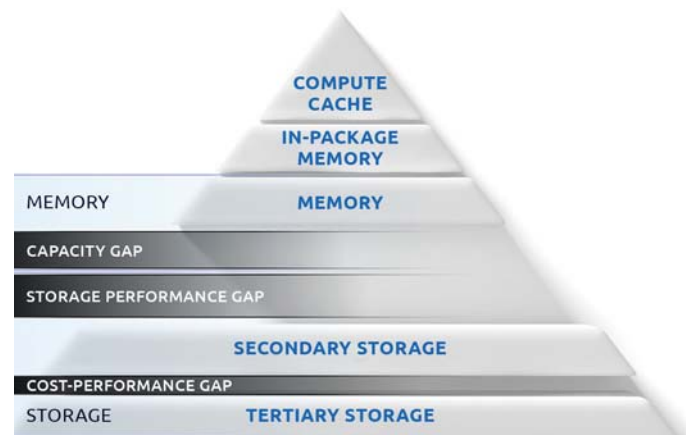


Figure 2.23: Research adjacency: Memory technology research needs pairing tracks with systems research<sup>24</sup> (courtesy of Al Fazio, Intel)



Likewise, concurrent software enablement is critical, as it is software in operating systems and applications that move bits back and forth across the hierarchy.

### Compute in-memory and near-memory

Systems architectures need to keep pace with societal needs. We are now transitioning to an age of what is deemed a "data-driven economy." Growth in memory density has been a great enabler of this transition, and room for further enablement can come by better cache refinement. But as the user demand for instantly accessible information continues to increase, memory access bandwidth still becomes a bottleneck, as illustrated in Figure 2.24<sup>25</sup>. This has spurred some researchers to investigate ways to redefine how memory is used in systems and to innovate ways to incorporate more novel types of memory.

Looking more deeply at data access, the problem is not with the memory itself, but with the way it is being used. The solution could bring compute and memory closer together, creating new compute architectures that can be imagined as a progression of options, with increasing or tighter compute/memory coupling. These options include:

1. Inserting a "tightly" compute-coupled layer in the Memory Hierarchy (e.g. HBM);
2. Moving compute primitives into the memory die;
3. Moving compute primitives into the memory "core" (shown in Figure 2.25); and
4. Merging compute and memory with in-memory computation for Neural Network fabrics' vector matrix multiply acceleration.

These may look to be separate solutions, but compute-in-memory should be considered as a continuum from compute-in-registers to compute-in-memory.

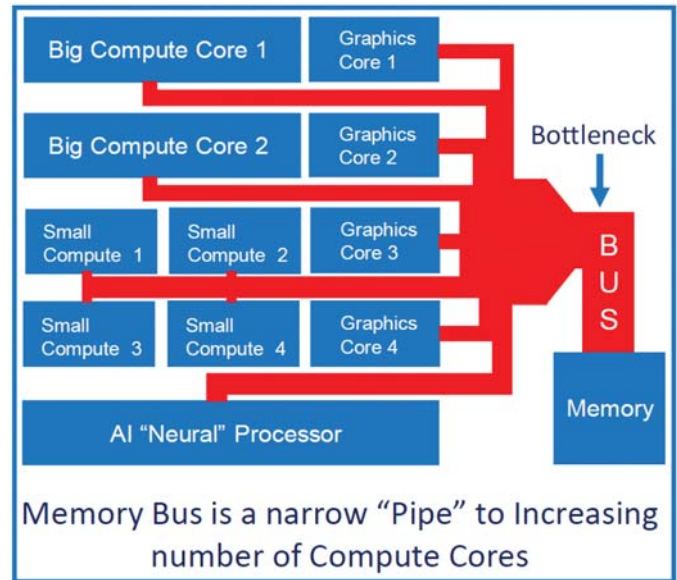


Figure 2.24: How current architectures bottleneck more efficient compute by isolating memory from compute<sup>25</sup> (courtesy of Greg Atwood, Micron Technology)

These new manners to enhance compute take advantage of reduced external (to CPU) bandwidth requirement and employ more internal memory parallelism for compute. Similar to the introduction of GPU, TPU, and FPGA accelerators in data centers, in-memory computation for neural fabrics can take such systems even further by exploiting the unique physics of "emerging memory." Examples of this include resistive, magnetic, and floating-gate technologies for summation (threshold) and sigmoid (triggering) behavior, as well as analog "weight" non-volatile storage. Bridging the gap between digital and analog with more analog-like memories (like ReRAM) may provide synergistic gains.

### Low Off Memory BW ← High On Memory BW

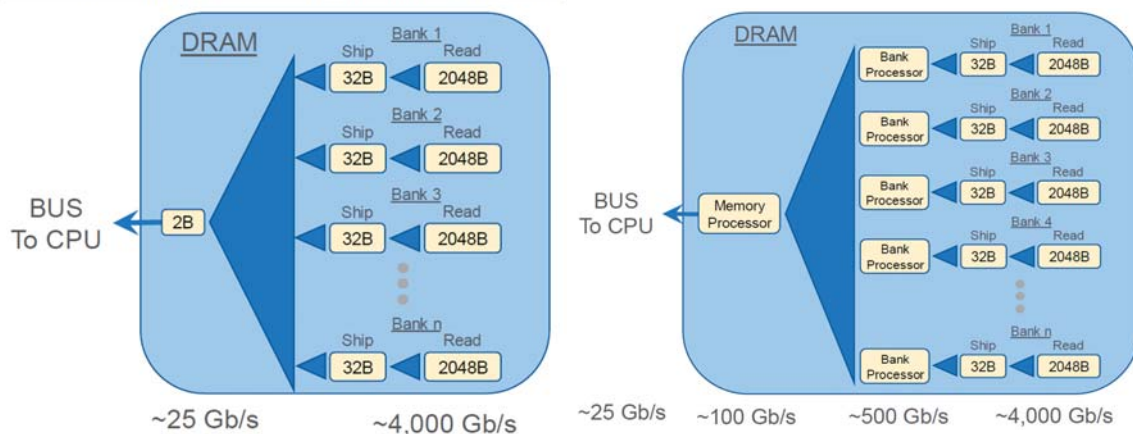


Figure 2.25: a) Memory is currently BUS-bandwidth constrained, and b) moving compute primitives into the memory core could alleviate memory bottleneck through reduced data movement<sup>25</sup>. (courtesy of Greg Atwood, Micron)



Embracing nonvolatile memories (NVMs) for compute comes with significant tradeoffs. When compared to traditional volatile memory like DRAM, emerging NVMs have advantages of low standby power, resilience to unexpected shutdown, and instant-on capability, countered by disadvantages such as high write latency/energy, limited write-cycle endurance, and vulnerability to security attacks. To extract the full value of persistence for computing will require large investments in enabling software development and design for security.

## Emerging memory

In looking to exploit new memory and system architectures, it is beneficial to understand the “emerging memory” contenders and how they might find favorable uses in memory and compute.

### *Spintronic RAM*<sup>26</sup>

More than a decade has passed since giant tunneling magnetoresistance was discovered in Fe/MgO system, imparting a reasonably large switching resistance (>100%) and providing for spin-driven electrical switching. These attributes deliver the basis for a magnetic random-access memory based on spin-transfer torque (STTRAM). Since magnetism is a collective state, magnetic orientation can theoretically be switched with low energy. With these combined attributes, STTRAM has been considered for use as a “universal memory” that rivals the speed of SRAM and the density DRAM. What seemed to be possible, however, has not been achievable to date. The spin torque switching phenomenon has yet to be harnessed in such a way as to be efficient enough to switch at low current, and the small read margin makes sensing difficult. Scaling has also been an issue, limiting devices to around 30nm for existing designs with a perpendicular magnetic orientation. New structures using perpendicular shape anisotropy have recently been proposed and experimentally demonstrated to offer scalability down to sub-10 nm dimensions.

Falling a bit short in its promise of universality, STTRAM has become one of the leading candidates for embedded memory applications, especially in the mobile realm, due to its fast-switching, low-power and non-volatile performance characteristics. Magnetic states are also radiation-tolerant for military and space applicability. To date, using STTRAM as a storage class memory has been limited by the large current densities necessary for switching, sensitivity of the tunneling junction to variation, and the small read margin, all of which serve to hinder high efficiency in large-array implementations.

Some recently discovered magnetic phenomena may help to curtail some of the shortfalls of STTRAM. A current flowing through a heavy metal was found to deflect spins—the spin

Hall effect—creating transverse spin-orbit torque (SOT) in an adjacent magnet. This switching method requires a three-terminal configuration, which limits its density compared to other two-terminal memory elements. However, its fundamental advantages in speed, energy-efficiency, and endurance make a suitable candidate for embedded memories (cache) in leading-edge technologies. Presently, the amount of current needed to impart magnetic switching is not significantly lower than that of spin-transfer torque, but alternative, more exotic materials like topological insulators, 2D magnets, and 2D Weyl Semimetals hold promise for augmenting efficiency in SOT. Another magnetic coupling phenomenon manifests as voltage-controlled magnetic anisotropy (VCMA), which is the ability to harness a voltage to reduce the barrier for magnetic switching. Used in conjunction with SOT, or in a thermal activation regime, faster (< 1 ns) and lower current density magnetic switching may be achieved.

### *Emerging ferroelectrics memories*<sup>27</sup>

Ferroelectric memories (FeRAM) are advantageous in that they can be written at a low voltage and power and at high speed, due to the collective nature of the spontaneous dipole moment. Up until recently, suitable ferroelectric materials for memory applications, typically lead-zirconia titanate (PZT), had dipoles that were only stable in fairly large groups. New momentum in ferroelectric memory development emerged in 2007 when ferroelectric properties were verified in HfO<sub>2</sub>, later confirmed to be a strained, orthorhombic phase, stable at below 10 nm. As such, HfO<sub>2</sub>-based memory cells have the potential to overcome the classical FeRAM scaling issue.

Ferroelectric memories can be used in a variety of forms (**Figure 2.26**), including as traditional FeRAM, where it serves as a capacitive element, in a three-terminal FET (FeFET), and in a switchable resistance-tunnel junction (FTJ). Of these three technologies, the FeFET has garnered the most interest. Unlike the FeRAM, the read is not a destructive process. Moreover, it can be scaled in a planar configuration, and its three-terminal structure lends itself to integration as both a memory and a logic element. This logic compatibility and functionality may serve to lift it above the competition as a suitable compute-in-memory element, especially for AI, which needs dense logic near the memory.

### *Phase change and resistive memory*<sup>28</sup>

Differentiation observed today in market applications makes back end of line (BEOL) resistive, non-volatile memories a valuable option to support or even replace off-chip Flash memory in some new architectures, such as cross-point (XPoint). Two technologies, PCM and ReRAM, are generally

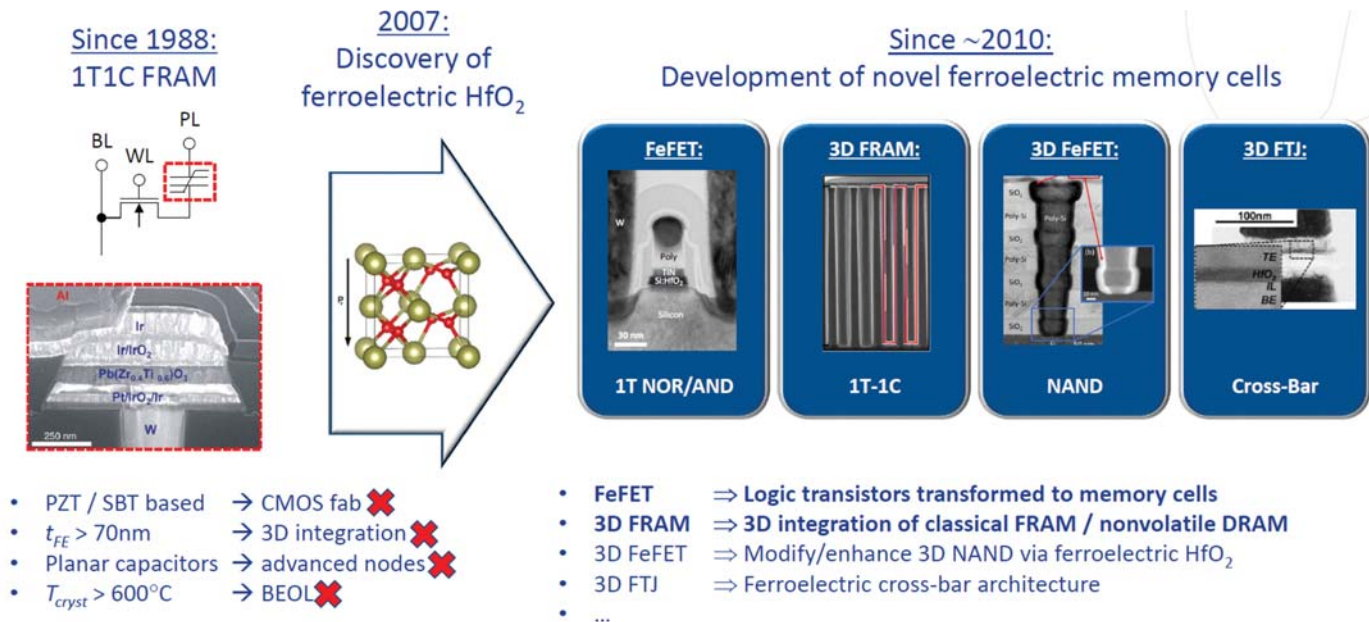


Figure 2.26: How discovery of ferroelectric HfO<sub>2</sub> has led to a resurgence in new ferroelectric devices that may be amenable to VLSI<sup>27</sup> (courtesy of Stefan Müller, Ferroelectric Memory Co)

amenable to such applications. PCM moves between an amorphous high-resistance state and a crystalline low-resistance state, whereas ReRAM moves atoms or atomic vacancies to create or terminate a conductive bridge. Both phenomena allow for intermediate states, so multi-level cells (MLC) or pseudo-analog implementation is possible, either of which would be equally beneficial for ML or AI applications. These two technologies are growing in maturity, thanks to an increasingly consolidated knowledge of the physics behind their functionality and consequential materials tuning for their reliability.

PCM is being used as part of 3D-Xpoint memory with a growing list of applications, primarily for high-end servers. Limited widespread adoption of PCM is related to relatively high current-to-switch and its tendency to have resistance-drift with cycling over time, as well as electrical read (disturb). All this can be attributed to atomic redistribution/segregation under current-driven heating conditions during write and read. ReRAM suffers from intrinsic stochasticity in switched resistance state, largely when switching from low (on) to high (off) resistance states, known as RESET. These negative cell attributes in both PCM and ReRAM can be attributed to uncontrolled atomic motion at the nanoscale. These inherent variabilities in cell properties during use have been met with a variety of compensation strategies, including optimized sensing circuits, write algorithms, and advanced error correction (exemplified for ReRAM in Figure 2.27).

PCM drift, which can be exacerbated with scaling, can manifest in an inability to properly sense the cell state (too-low read

margin) or stuck cells, generally limiting endurance to under 10<sup>6</sup> cycles. Despite the difficulties in achieving high endurance, PCM is being implemented in the storage market and data centers. It's also making inroads in accelerated gaming and graphics, as well as in automotive and military applications, due to better radiation and heat tolerance than Flash memory. ReRAM has yet to achieve any widespread applications

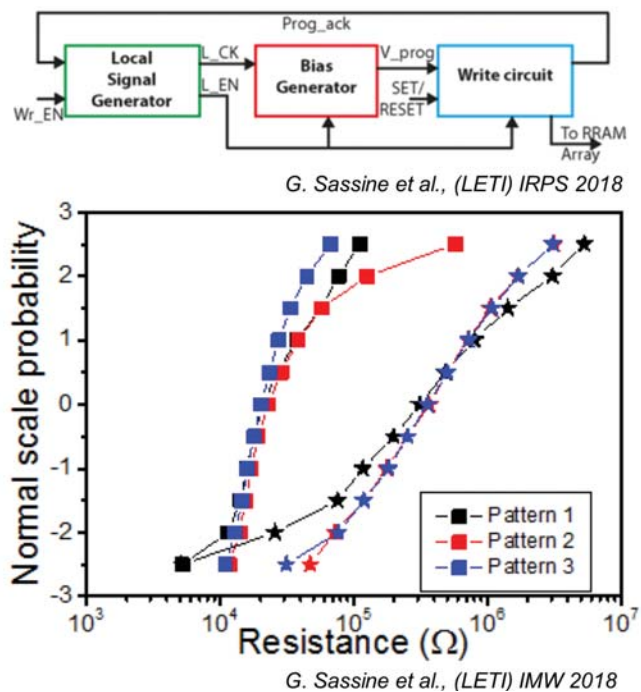


Figure 2.27: Illustration of error correction and/or write verifications schemes necessary to account for inherent stochasticity of ReRAM<sup>28</sup> (Courtesy of Gabriele Navarro, CEA LETI)

despite intensive effort in a very large variety of materials systems. This is largely due to the described stochastic nature of the switching phenomenon in these materials, which leads to high intra- and inter-cell variability. This variability must be compensated for by innovative circuits, architecture, software, or, more hopefully, with materials innovations that greatly dampen the stochastic inconsistency.

### Selector devices for cross-point (X-Point) memories

A selector, or access device, serves to isolate a memory cell in an array and allows for it to be accessed and turned on and/or off, depending on implementation. It improves signal-to-noise ratio, as well as memory-access disturbance immunity. A selector in a cross-point (also known as a cross-bar) memory-cell array is a two-terminal switch serially connected to a matching storage element so the memory resides on top of the selector, integrated at the crossing point of a pair of connecting metal lines, thus allowing for memory arrays with high packing density. An effective selector technology is one that can be realized with a low thermal budget, such that it can be implemented in the backend of an integrated memory device, compatible with mainstream semiconductor technology featuring CMOS under array. This allows for 3D (vertical) deck stacking and relaxing the feature size requirements to achieve an effective end-run on the need for further scaling to get more memory in the same base silicon area. Many emerging memories lend themselves better to this geometry than Flash, a three-terminal device, and DRAM, a very tall device.

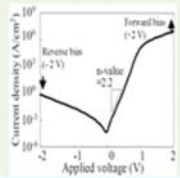
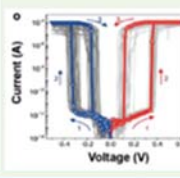
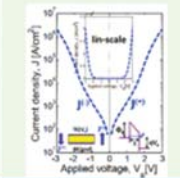
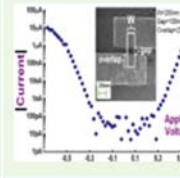
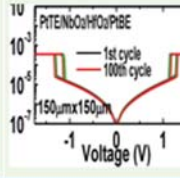
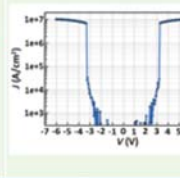
Many different types of selector devices have been studied. Table 2.3<sup>29</sup> lists the attributes of leading select device

candidates for various switching mechanisms. Some memory devices can function in a unidirectional manner, like PCM and some ReRAM, while others require bidirectionality, like STTRAM and other ReRAM. The needs for select device that all emerging memory candidates share are highly nonlinear I-V characteristics, good select-voltage window, fast-access speed, high endurance, excellent device-to-device uniformity, and thermal stability. It has been rather difficult to identify a candidate that fulfills all the necessary criteria, even for a given memory device. More traditional thermionic-based devices are unidirectional in nature, and, while they can have good off current, they are limited in the high-current density regime. Ovonic Threshold (OVS) and Metal-Insulator transition (MIT) selectors both offer high on-current and fast switching speed, but they don't provide suitably low off-current, especially MITs, which limits larger-size array implementations. Filamentary switches provide ultra-low leakage but suffer from low on-current density, small voltage window, and poor reliability. Tunnel devices are difficult to achieve high nonlinearity at reasonable voltages. Mixed ionic-electronic current (MIEC) devices suffer from a rather small voltage window. As one can see, it's difficult to achieve all the necessary attributes in a select device suitable for a large cross-point memory array.

### Key areas for focus and follow-on

Emerging memory technologies reviewed above are holding future promise to supplement or potentially replace some of the existing memory technologies and eliminate the gaps in the memory hierarchy, provided that existing limitations

Table 2.3: Attributes of various identified cross-point select device candidates<sup>29</sup>

| Mechanism              | Thermionic  | Filamentation   | Tunneling   | MIEC   | MIT   | OTS   |
|------------------------|---|---|---|--|---|---|
| Construct.             | P-N or M-S Jx   | Ion (Ag <sup>+</sup> ) in Ox  | MIM   | Cu <sup>+</sup> in SE  | NbO <sub>2</sub>  | Chalcogenide  |
| Switching              | Electronic  | Atomistic   | Electronic  | Atomistic  | Atomistic   | Electronic  |
| Polarity               | Unidirectional  | Bidirectional   | Bidirectional   | Bidirectional  | Bidirectional   | Bidirectional   |
| $\tau_{\text{switch}}$ | sub nsec  | ns ~ 100s ns  | ps or faster  | ns ~ 100s ns   | ns to 10s ns  | sub ns  |
| $J_{\text{MAX}}$       | < 10MA/cm <sup>2</sup>  | 1-10MA/cm <sup>2</sup>  | < 1MA/cm <sup>2</sup>   | ~10MA/cm <sup>2</sup>  | > 10MA/cm <sup>2</sup>  | > 10MA/cm <sup>2</sup>  |
| $J_{\text{Inhibit}}$   | < 1A/cm <sup>2</sup>  | < 1KA/cm <sup>2</sup>   | < 1KA/cm <sup>2</sup>   | < 1KA/cm <sup>2</sup>  | < 1KA/cm <sup>2</sup>   | < 1KA/cm <sup>2</sup>   |
| $V_{\text{Inhibit}}$   | < 3V  | < 1V  | < 3V  | < 1V   | < 3V  | < 3V  |
| I-V                    |  |  |  |  |  |  |
| Reference              | Y. Sasago, et al., VLSI '09   | J. Yang, et al., Adv Func Mtls (2018)   | B. Govoreanu, et al., IEDM'13. P10.2  | K. Gopalakrishnan, et al., VLSI Symposium '10.                                       | X. Liu et al., EDL Oct.'14  | S. Yasuda, et al., VLSI symposium, '17,   |



of each technology are addressed by ongoing research activities. To facilitate entrance of the emerging memories into the market and improve customer confidence in new technologies, niche applications together with novel software and architectures will be important enablers.

Common development directions for all emerging memories include:

- Development of new materials, processes, and structures for enhanced speed, energy-efficiency, reliability, and scalability of binary and multilevel capable memory cells.
- Improved tooling (deposition, etching tools, etc.) to address scalability and device-to-device variations
- Software and architecture development to fully utilize benefits of a specific emerging memory, while mitigating some of the associated risks

In addition to common focus areas for future development, each memory holds its own unique challenges:

- STTMRAM
  - Improve speed to approach cache memory (10-20 ns)
  - Improve endurance (usually limited by MgO barrier breakdown) to  $1e15$  level
  - Improve scalability to  $< 30$  nm dimensions
- SOTMRAM
  - Film and cell-structure stacks for field-free operation and robust magnetic immunity
- PCRAM/ReRAM
  - Address resistance drift, stochasticity of switching and wide cell-to-cell distribution
  - Improve endurance to  $>1e10$
  - Increase speed and reduce power consumption
- FeRAM
  - Reduce write voltage for fast write speed
  - Improve endurance and retention
  - Minimize ferroelectric imprint
  - Solve write/erase disturb

Variability is a universal challenge in emerging memories. Establishing intrinsic variability limits of emerging memory cells along with research on materials and processes to minimize extrinsic variability components are key.

- Research on memory systems for quantum computing

## 2.6. Present and Future Mass Storage Technologies

### Overview and needs

IDC research estimates that corporate data will continue to grow at a 40-50% compound annual growth rate (CAGR), doubling every two to three years. Global memory demand is estimated to exceed 100 zettabytes by 2040. Moreover, the world's demand for storage continues to grow exponentially, yet, evolutionary capacity gains are no longer able to keep up. Mass storage technologies will need to scale dramatically to meet the required capacity, while also continuously improving the price of storage. The ability to store data in an affordable way will allow for expansion of the storage marketplace.

Existing mass storage technologies will continue to scale and provide the foundation for block-level mass storage through the next decade. However, data growth is outpacing the rate of technology advancement.

Recent advancements in solid-state-drive (SSD) technology that use semiconductor cells led to dramatic price declines over the last several years, with the potential for further declines in the next decade. As a result, SSD technology is encroaching into many traditional hard-disk-drive (HDD) segments, such as mobile and gaming. HDD technology is reaching a critical point where the technology roadmap (e.g., energy-assisted magnetic recording) has to emerge soon, or the entire storage roadmap will need to be re-evaluated. Over time, it is possible HDD will migrate to colder storage tiers, while also innovating to preserve performance for warmer storage tiers.

After a decade of consolidation, magnetic tape is potentially poised to inherit significant growth in cold and archival storage, but only if tape technology advances ahead of alternatives. New storage technologies, such as DNA, offer three-dimensional storage that could potentially revolutionize mass storage, but the technology is many logs away from current storage technologies in both capacity and price.

A growing number of use cases need exabyte-scale data sets. It is estimated that 1 exabyte of unreplicated data over five years costs \$100M, with a large footprint and significant power and cooling requirements. Synchronization across multiple exabyte archives is effectively impossible today. As data volumes grow, stakeholders may need to discard an increasing proportion of available data. For use cases such as national security, this could limit key functions.



# Storage Optimization – it's all about tiering!

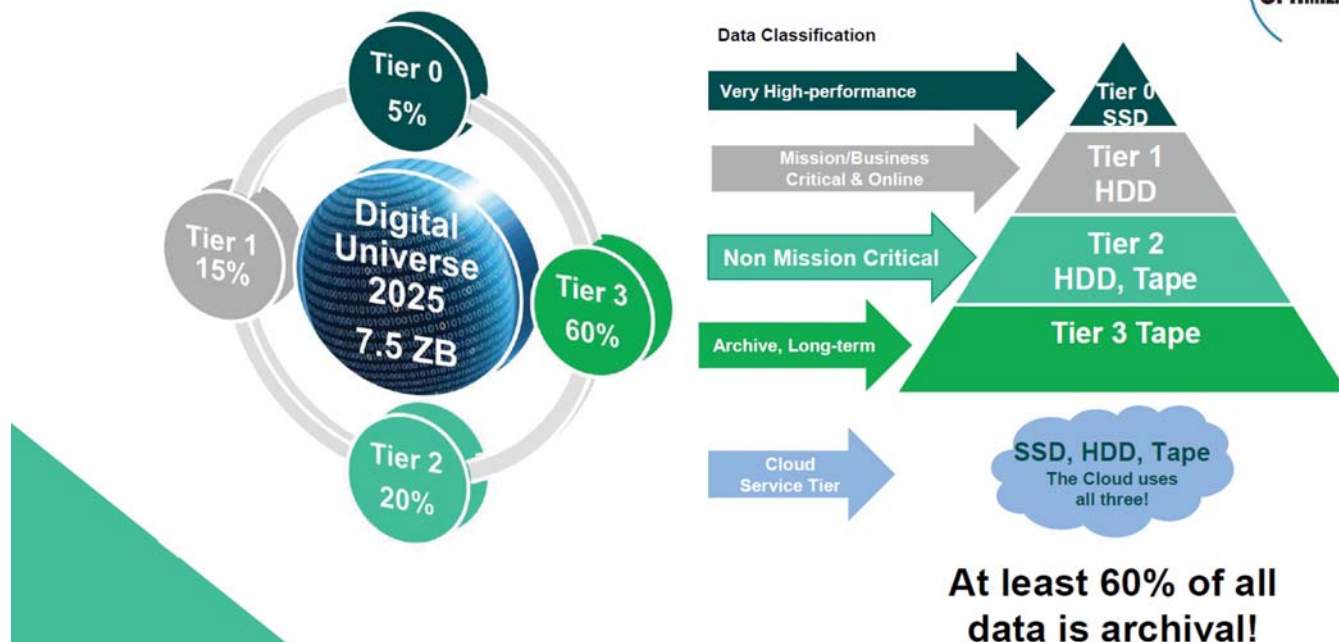


Figure 2.28: Storage tiering<sup>30</sup> (courtesy of Peter Faulhaber, Fujifilm Recording Media U.S.A.)

## Storage optimization with no single-storage solution

A single-storage solution cannot serve all segments. The future of storage is blowing “hot” (data that needs to be accessed right away) and “cold” (data that does not require fast retrieval, e.g., hours or days). At the cold end, DNA could replace or augment tape. Cold-optimized HDDs could also compete with tape. At the other end, there is a whole new category of very fast SSDs—3D Crosspoint is in the market as a storage-class memory. STT-MRAM is now used conventionally as an embedded memory in microprocessors. Each of these has a large enough market as the overall pie grows, and one or more will become mainstream by 2030. We will see the emergence of purpose-built architectures and new semiconductor and magnetic storage devices at the boundaries of existing market segmentations, all tailored to workload, power, and performance requirements. These technologies and solutions will provide opportunities for new growth vectors in the storage marketplace.

To achieve the lowest total cost of ownership, a balance must be reached in terms of the acquisition costs and the quality of service. This is achieved by balancing the various types of storage in the future, including DRAM, Flash, HDD, tape, and novel storage technologies. There is a right place and time in the data lifecycle for all these storage mediums. What is needed are intelligent data-management solutions that classify data and move it automatically by user-defined

policies from expensive tiers of storage to economy tiers of storage. See Figure 2.28 for an illustration of storage tiers<sup>30</sup>.

## Whole-system-level thinking

Some of the biggest opportunities are around whole-system-level thinking to engineer end-to-end in terms of where to take concessions and where to optimize building large-scale storage<sup>31</sup>. For hyperscale cloud use cases, mass storage devices are just one component in terms of the optimization and design of an overall system. Thinking in terms of discrete mass-storage components aims for near perfection regarding reliability of local mechanisms to deal with failures. For the cloud, even if storage devices were perfect, the other sources of risk for a hyperscale cloud use case still require significant countermeasures. Durability is a whole-system-level challenge. As a result, protecting against bit errors above a certain point may have diminishing returns. Instead, consider detecting and responding to those failures across an entire fleet of storage devices. In this way, we can avoid thinking of individual storage devices as discrete components.

This is a different approach to that of storage arrays with a fixed form-factor device with annual failure rates (AFR) and expected risks on the hardware up front. Some cloud systems are designed differently, as a statistical model that is part of an active feedback loop. Component-level failure rates are monitored to actuate proactive steps to avoid durability failures.

If the rates of failure or other environmental factors in the system change in a way that is surprising, the model can adapt by scaling up replication systems to recover more aggressively.



Figure 2.29: Durability concerns for hyperscaler cloud<sup>31</sup> (courtesy of Andy Warfield, Amazon Web Services)

Cloud providers take for granted that things will fail dynamically in a way that will vary over time and components. The systems need to adapt to this style of failure. Meeting a durability promise means reasoning about different durability risks that go far beyond storage component failures (Figure 2.29). This approach allows cloud

providers to achieve 11 9s (or more) of durability, as well as survive the complete loss of an entire facility within a region and remain well ahead of the steady-state failure of mass storage devices or power supplies.

### Flash NAND SSD

NAND Flash’s low cost per bit with non-volatility with moderate performance, power, and reliability characteristics make it an indispensable part of the memory hierarchy<sup>32</sup>. In the past decade, NAND-based storage has successfully transitioned from a 2D to 3D implementation, breaking through capacity barriers and enabling continued cost-per-bit reduction. This new paradigm has simultaneously improved performance and power. Storage devices using NAND Flash are now ubiquitous, offering new opportunities in the storage hierarchy. Flash is and will continue to replace other HDD categories (e.g., performance HDD). Cloud will use both Flash and HDD, not one or the other.

Products based upon 128-layer NAND are becoming mainstream. High capacities in the 10s of TB are available in small form factors. Performance is capable of high bandwidth in the GB/second and read latency in the 10s of microseconds. NAND also has strong endurance with single-digit drive writes per day for mainstream 3-bit-per-cells PLCs.

Cost reduction remains the primary interest, with increased bits stored per memory cell and process technology scaling. Performance improvement is focused on getting the most out of the NAND media and is targeted at what matters most

based on the workload. Whole-density gains are starting to slow down—TLC (3 bits) to QLC (4 bits) was a 25% scaling benefit, whereas QLC (4 bits) to PLC (5 bits) will be a 10% actual scaling benefit. Fundamental improvements in processes, tools, and materials will be required to continue delivering this level of process scaling into the next decade. Figure 2.30 illustrates that increased bits per cell will double the number of states for every additional bit per cell added, but the cost benefit is diminishing.

This creates tradeoffs in performance, energy, and reliability. The only reason Flash continues to improve is because we multiple the small gains by very large (400) layers on top. The industry needs to rationalize this technology scaling capability with the capital required to ensure business viability<sup>33</sup>. For every 1% of bit growth, the capital required is growing 26% for Flash and 22% for HDD from 2014-2019 (Figure 2.31). All the gains we get because of technology could be lost in the capital cost if we are not careful. Capital intensity could potentially ruin the technology advancements in both HDD and NAND Flash.

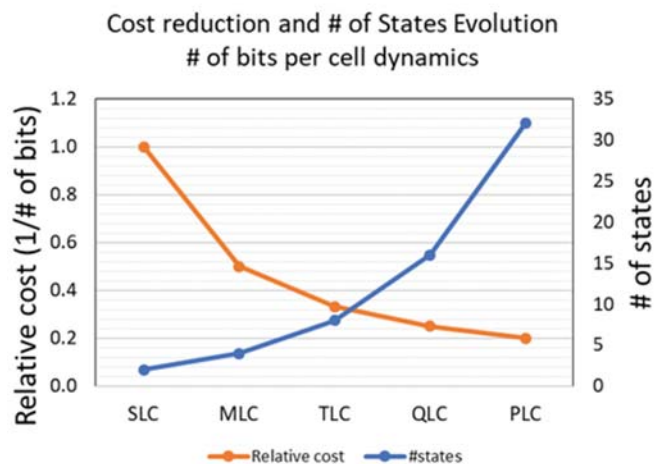
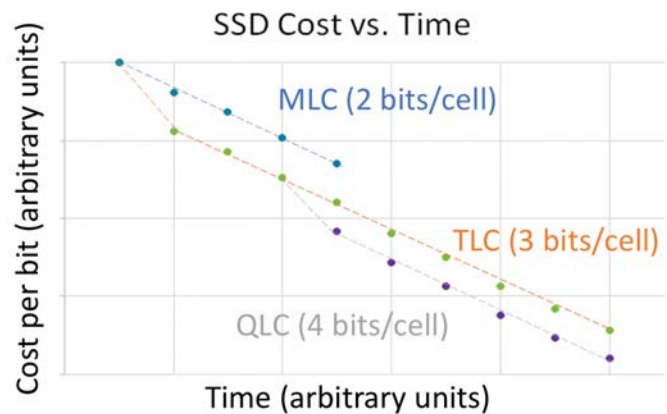


Figure 2.30: Cost per bit analysis for SSD NAND<sup>32</sup> (courtesy of Mark Helm, Micron Technology)

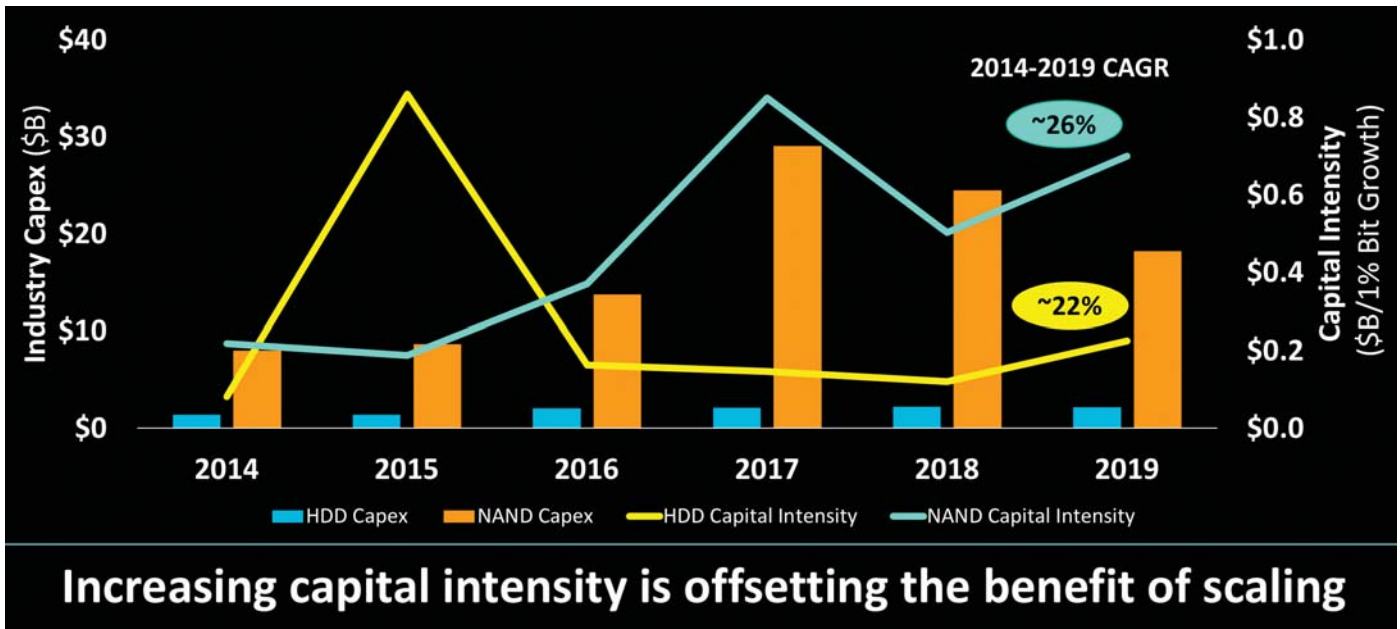


Figure 2.31: Capital Requirement for every 1% of bit growth for SSD NAND and HDD<sup>33</sup> (courtesy of Siva Sivaram, Western Digital)

### Hard Disk Drives (HDD)

Roughly every 10-20 years there is a major technology introduction that creates an inflection point in HDD areal-density growth<sup>34</sup>. Why is this important? These inflection points significantly reduce the cost of storage and play a big role in enabling all the data-hungry applications in our future. For example, 20 years ago, magnetoresistive readers fueled 10 years of areal-density growth before it started to slow. Then, around 2005, perpendicular recording fueled the past 15 years of growth with a 10x improvement in areal density. Perpendicular recording is also starting to slow down. More recently, shingled magnetic recording (SMR) technology

has been introduced, which enables up to 20% density improvement leveraging existing perpendicular recording technology under tailored workload conditions. Limiters to HDD scaling include linear density (sub-10nm scale), track density (nm scale of servo positioning), and grains per bit (nm scale). HDD technology developed many unique wafer capabilities, and lithography dimensions are on par with leading semiconductor processing.

We are approaching one of these inflection points, and this will unlock 10-15 years of robust areal-density growth (see Figure 2.32). Over the next decade heat-assisted magnetic recording (HAMR) will enable advances in magnetics for

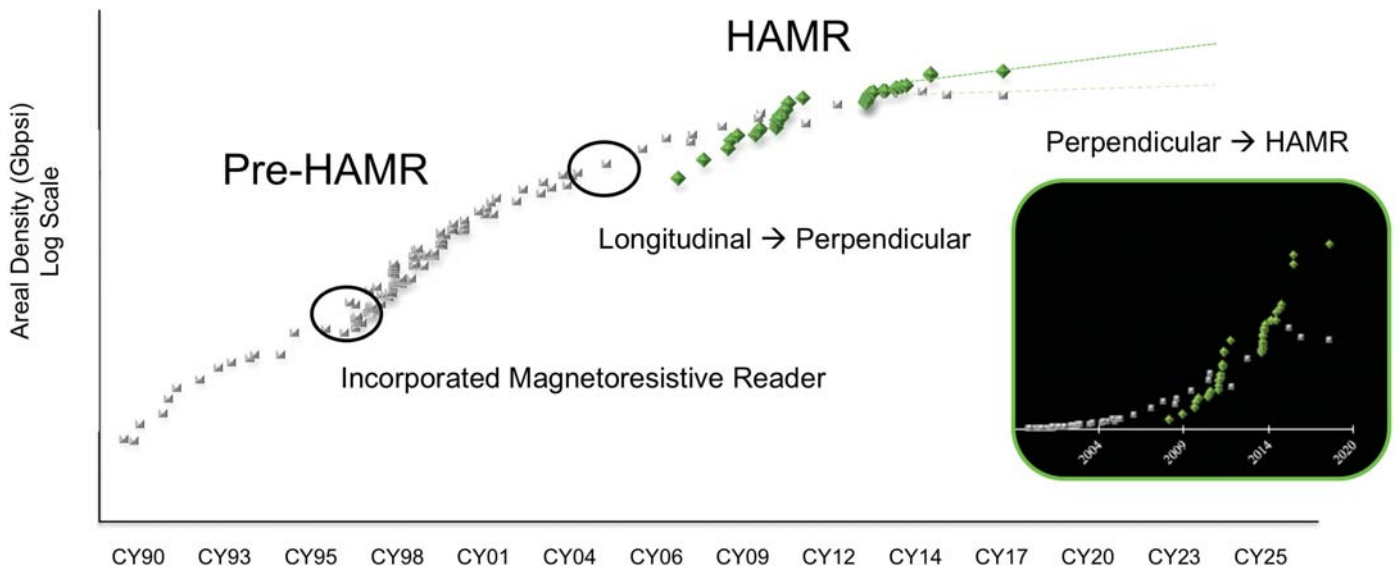


Figure 2.32: HDD areal-density gains over time, 1990-2025<sup>34</sup> (courtesy of John Morris, Seagate)



large and important markets<sup>34</sup>. HAMR enables HDD technology to write smaller bits on higher coercivity media on a glass substrate. After cooling, the grains are very stable and not switchable by the head field alone. This whole process occurs in under two nanoseconds.

HAMR will enable HDD to continue to supply >90% of the demand for mass-capacity storage. It is unlikely that business-critical HDD will be outpaced by Flash. For many storage classes, a number of functions are handled above the physical device's storing data, such as compression, de-duplication, and erasure encoding. This allows physical-mass storage devices to focus on raw-storage efficiency in terms of dollars per terabyte. Cooling and power costs of storage tiers are minimal compared to compute nodes.

### Magnetic tape

Magnetic tape is one of the oldest storage technologies used today at scale<sup>30</sup>. It offers the lowest total cost of ownership with the longest (coldest) retrieval times. The current LTO Roadmap specifies Generation 12 at 144 TB native capacity and 360 TB compressed. LTO-12 is expected around 2027 and certainly before the end of the decade<sup>30</sup>.

Achieving higher capacity and better TCO in magnetic recording is all about areal density. IBM and Fujifilm demonstrated the ability to get to 123 billion bits per square inch in 2015, which would equate to a tape cartridge capacity of 220TB using Barium Ferrite (BaFe). Beyond Barium Ferrite, Fujifilm is working to commercialize a new magnetic particle that can store up to 400 terabytes, or 22 times more than LTO-9 capacity. This new magnetic particle is called "Strontium Ferrite" (SrFe) and has magnetic properties that are even better than Barium Ferrite.

Fujifilm believes this new particle technology can be applied beyond LTO-10 for cartridge capacities of 400TB or more by 2029<sup>30</sup>.

The evolution of magnetic tape particles started in 1994 with legacy metal particles, followed in 2006 by Barium Ferrite; Strontium Ferrite is expected for 2025<sup>30</sup>. This progressive reduction in particle size enables higher areal density and higher-capacity cartridges. Fujifilm is also working on the next magnetic particle beyond Strontium Ferrite that the company calls Epsilon Ferrite ( $\epsilon\text{-Fe}_2\text{O}_3$ ). This technology is enabled by Focused Millimeter Wave-Assisted Magnetic Recording (F-MIMR). Fujifilm expects Epsilon Ferrite to deliver up to a 1 PB tape media cartridge by 2035. **Figure 2.33** compares Barium Ferrite, Strontium Ferrite, and Epsilon Ferrite.

### DNA storage

Molecular information storage (MIST) is an emerging paradigm that uses polymers like DNA to encode information with higher bit density and greater stability than conventional storage media. Several efforts to develop scalable MIST technologies are currently underway, including a large public-private partnership launched by IARPA in 2018<sup>35</sup>.

## Epsilon Ferrite Beyond BaFe and SrFe

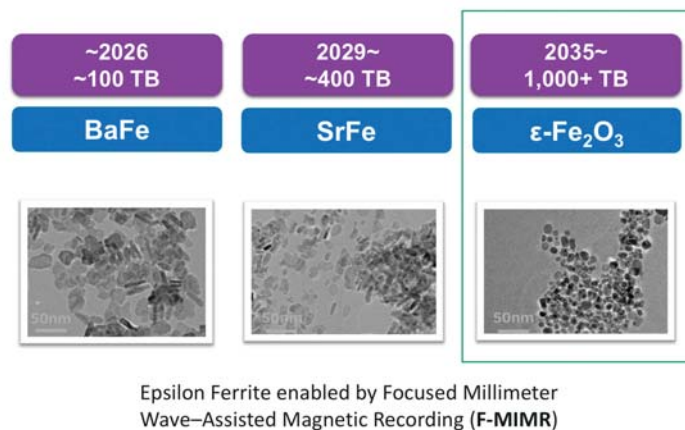


Figure 2.33: Comparison of Barium Ferrite, Strontium Ferrite, and Epsilon Ferrite<sup>30</sup> (courtesy of Peter Faulhaber, Fujifilm Recording Media U.S.A.)

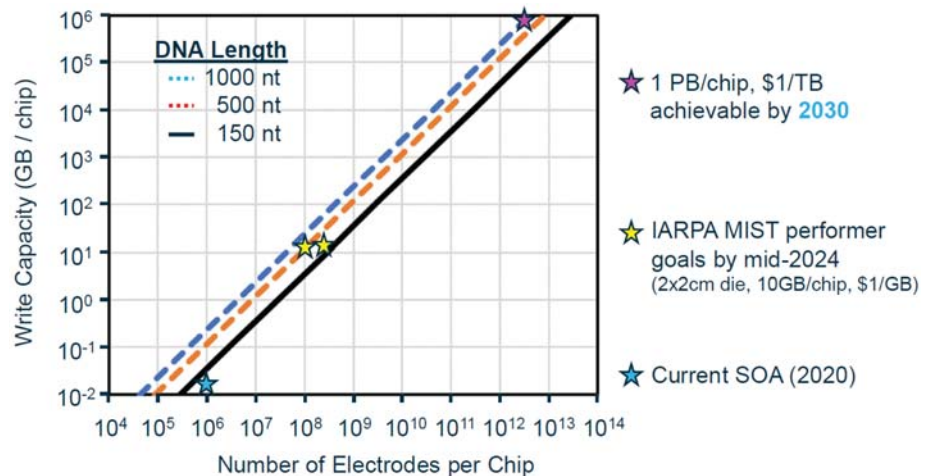


Figure 2.34: Scaling Potential for DNA Synthesis Chips<sup>33</sup> (courtesy of David Markowitz, IARPA)



DNA storage should be viewed as a “right tool for the right job,” not a panacea that will replace other storage technologies like tape or HDD. The IARPA roadmap for DNA storage aims for tabletop DNA storage devices that can achieve a TB-per-day write and read workflows for \$1,000 in resource cost (see **Figure 2.34**). This achievement will fundamentally de-risk this technology stack for future scaling and commercial product development.

Commercialization of DNA storage will be possible in six years to satisfy use cases that require extreme media longevity. There are several use cases in the U.S. national security community that have this requirement, so it is reasonable to expect the U.S. government will be among the first customers of these first-generation devices. Within 10 years, IARPA sees second-generation devices that satisfy extreme scale, resource efficiency, and fast-parallel search. IARPA is focusing on getting the field on this path. Executing on the roadmap will require public and private partnerships<sup>35</sup>.

### Key areas of focus and follow-on research

#### SSD NAND

- Increasing bits stored per memory cell will require improvements in processes, tools, and materials.
- Addressing the need to keep scaling the technology with business viability, given the increasing cost of adding bit growth.
- Balancing performance, energy, and reliability in SSD NAND.

#### HDD

- Energy-Assisted Magnetic Recording (EAMR) is required to write smaller bits on higher coercivity media (requires heating the bits to high temperatures followed by a rapid cooling).
- Balancing areal-density gains with IO performance through solutions like multi-actuators.
- Adding platters in the same near-line form factors will be a challenge.

#### Magnetic tape

- New materials such as Strontium Ferrite and Epsilon Ferrite are considerations for future magnetic tape cartridges to increase areal density.
- Magnetic tape drives require read/write heads that are more complicated than hard disk drives (HDDs).

#### DNA storage

- Logarithmic improvements in DNA synthesis costs and capacity are required for hyperscale applications.

- Solutions are needed for the volume of liquid consumables required for DNA storage.
- DNA sequencing capacity needs to improve dramatically, including lower capital costs.

## 2.7. Summary—New Trajectories for Memory and Storage

### Overview

Radical new solutions in memory and storage technologies will be needed for future ICT. It is becoming increasingly clear that in future information-processing applications, synergistic innovations from materials and devices to circuits and system-level functions will be key to achieving new levels of bit density, energy efficiency and performance. Those innovations likely rely on yet-to-be-explored physical principles and structures (materials and interfaces).

Memory is an essential component of ICT, and further advances in computing are impossible without ‘reinventing’ the compute-memory system, including information representation, device physics, memory hierarchy architecture, and physical implementation. Traditional planar Flash nonvolatile memory can no longer be scaled below 28nm, and alternatives must be able to support the rugged environment of the automotive market. Also, global demand for data storage continues to grow exponentially, and today’s storage technologies will not be sustainable in the near future due to the excessive material resources needed to support the ongoing data explosion. Thus, new radical solutions for data-storage technologies are needed.

In summary, revolutionary changes to ICT memory and storage will be required soon. This will necessitate a cross-disciplinary, cross-functional approach to realize a solution space with multi-decade longevity to replace the current solutions.

**Memory Grand Goal:** Develop emerging memories and memory fabrics with >10-100X density and improve energy efficiency for each level of the memory hierarchy.

**Storage Grand Goal:** Discover storage technologies with >100x storage density capability and find new storage systems that can leverage these new technologies.

## Research recommendations summary

### Memory technology for edge, mobile, and IoT computing

- Cell-level research goal: >10X improvement for key parameters aiming for 100-1000X improvement in power/energy with enhanced reliability and lower cost/area.
- SRAM: Disruptive area scaling solutions for SRAM in advanced nodes.
  - Identify new alternatives to 6T SRAM for on-chip code/data: smaller area/bit, low leakage, baseline-process compatibility, and zero/low process cost adder.
- Non-volatile memories: Enable NVM in cutting-edge process nodes for edge, data center, and auto products.
- Flash function replacements @ 28nm and below
  - New capabilities needed with endurance > 1M.
- Research on scalable multi-level emerging non-volatile memory cell solutions and related high-efficiency memory arrays.
- Sustainable denser 3D (non-monolithic or monolithic) integration of memory and logic: high memory capacity and bandwidth.
- Memory technology-aware algorithms to overcome write (latency, energy, endurance) challenges and to ensure error resilience.

### Memory technology for HPC and data centers

- *Software infrastructure*: A ubiquitous software framework that lowers the barrier to integration of new near-data processing elements.
  - This framework must be capable of optimizing data placement and data movement within the system.
- *Interconnects*: High-bandwidth, low-energy interfaces are required to move data between memory, storage, CPUs, and accelerators.
  - These interconnects and interconnect standards should be developed in such a way that they allow data movement directly among elements, rather than to a central processor then back out to a different element as is done in systems today.
  - Protocols running on top of these interconnects must ensure coherence and resilience.
- *System architecture*: Modern server architectures were not designed with near-data computing in mind and, as such, limit the possibilities of near-data computing in many ways.
  - For in-memory compute to be possible, one must first

go to great lengths in software to ensure that all the operands for an operation are co-located in the memory die where they will be processed.

- For near-data computing to become ubiquitous, new system architectures must be developed.
- *Near-data processing elements and near-data processor architectures must be developed.*
  - Near-data processing element design must be designed in the context of a full-system design that embraces near-data computing, along with the wealth of other heterogeneous computing operations becoming commonplace today.

### Emerging memory technologies

- There is room in the compute-memory hierarchy for additional cache layers that allow for implementation of emerging memory technologies, such as MRAM, PCRAM, and FeFETs.
  - Emerging memory implementations must consider methods/paths for scaling and 3D capability, including the select device for X-point implementations.
- MRAM needs endurance to attain > 1E10 with enhanced thermal stability, magnetic immunity, and low BER, while also reducing write current.

Variability is a universal challenge in emerging memories. Establishing intrinsic variability limits of emerging memory cells along with research on materials and processes to minimize extrinsic variability components are key.

- Research on memory systems for quantum computing.

### Mass storage technologies

#### SSD NAND

- Increasing bits stored per memory cell will require improvements in processes, tools, and materials.
- Addressing the need to keep scaling the technology with business viability, given the increasing cost of adding bit growth.
- Balancing performance, energy, and reliability in SSD NAND.

#### HDD

- Energy-Assisted Magnetic Recording (EAMR) is required to write smaller bits on higher coercivity media (requires heating the bits to high temperatures followed by a rapid cooling).
- Balancing areal-density gains with IO performance through solutions like multi-actuators.
- Adding platters in the same near-line form factors will be a challenge.

### Magnetic tape

- New materials such as Strontium Ferrite and Epsilon Ferrite are considerations for future magnetic tape cartridges to increase areal density.
- Magnetic tape drives require read/write heads that are more complicated than hard disk drives (HDDs).

### DNA storage

- Logarithmic improvements in DNA synthesis costs and capacity are required for hyperscale applications.
- Solutions are needed for the volume of liquid consumables required for DNA storage.
- DNA sequencing capacity needs to improve dramatically, including lower capital costs.

## Appendix: Global Data Storage Trends

Global trends in information storage are based on research by Hilbert and Lopez<sup>36</sup>, where a detailed inventory of all storage media was created (their data-storage inventory includes, among others, paper books and newspapers, audio and video tapes, photo negatives and prints, all types of digital storage etc.). A summary of their findings is shown in **Figure A2.1**, from which several observations can be made and include: (i) the majority of data was analog before 2002; (ii) the analog data reached a maximum around 2000 and steadily decreased afterwards; (iii) the amounts of stored analog and digital data became equal around 2002; and (iv) after 2007 the vast majority of information became digital—a trend that continues today.

Extrapolation of the digital line in **Figure A2.1** provides projections for required global data storage. An important caveat is that the growth rate of digital storage (red line) is considerably higher than the total growth rate (blue line). This reflects analog data dominating the total storage capacity for the majority of the measurement period. Extrapolation of the digital line in **Figure A2.1** is treated as an upper bound for storage, while the extrapolation of the total storage capacity is used as a conservative estimate.

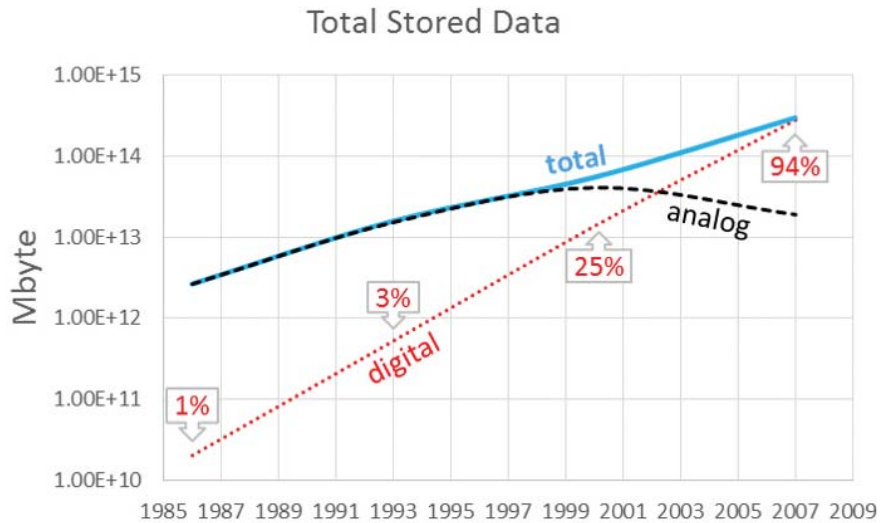


Figure A2.1: Timeline of analog and digital data storage including percentages of digital data with time<sup>36</sup>

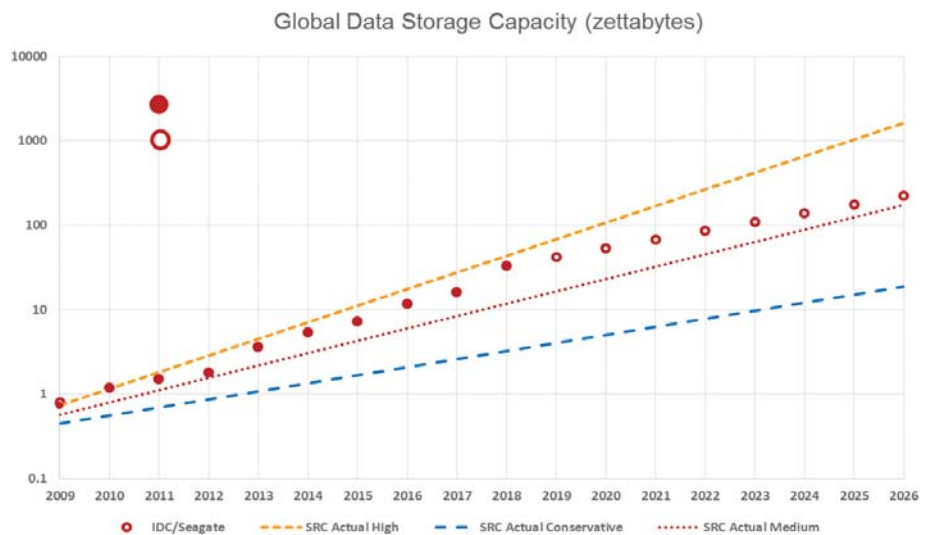


Figure A2.2: Estimated and projected storage data/storage demand in 2009-2026, including a conservative estimate and an upper bound, and comparison with independent estimates of the data storage<sup>37</sup> (solid and open red dots)

Figures A2.2 and A2.3 show the projections of global storage demand, both a conservative estimate and an upper bound. These extrapolations are compared to independent estimates of the data storage in 2009-2018 (solid red dots) and projected storage needs in 2020-2030 (open red dots) in Figures A2.2 and A2.3. All estimages and projections are within the defined boundaries (formed by the 'conservative' and 'upper bound' lines).

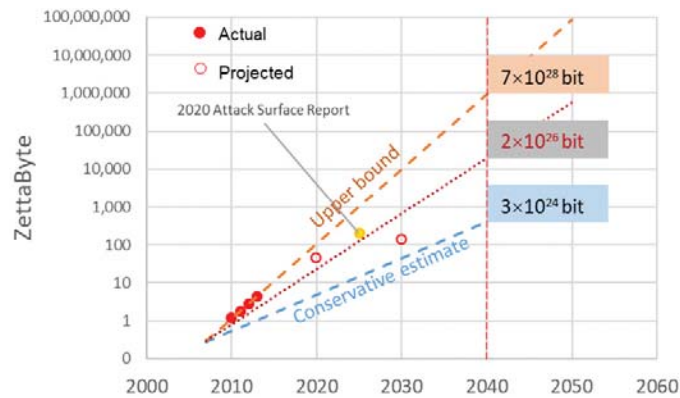


Figure A2.3: Estimated and projected storage data/storage demand in 2000-2050, including a conservative estimate and an upper bound

## Contributors

James Ang (Pacific Northwest National Lab)  
 Dmytro Apalkov (Samsung)  
 Rashid Attar (Qualcomm)  
 Kelly Baker (NXP)  
 Shekhar Borkar (Qualcomm)  
 Nafea Bshara (Amazon Web Services)  
 Gary Bronner (Rambus)  
 Carlos Diaz (TSMC)  
 Sean Eilert (Micron)  
 Dan Ernst (Hewlett Packard Enterprise)  
 Kjiersten Fagnan (DOE-JGI)  
 Peter Faulhaber (Fujifilm Recording Media)  
 Simon Hammond (Sandia National Lab)

Ken Hansen (SRC)  
 Mark Helm (Micron)  
 Ron Ho (Facebook)  
 Bruce Jacob (U Maryland)  
 Thomas Jew (NXP)  
 Matthew Klusas (Amazon Web Services)  
 Sailesh Kottapalli (Intel)  
 Steve Kramer (Micron)  
 Rafic Makki (Mubadala)  
 Matthew Marinella (Sandia National Lab)  
 Subhasish Mitra (Stanford)  
 John Morris (Seagate)  
 Vijay Narayan (Penn State)

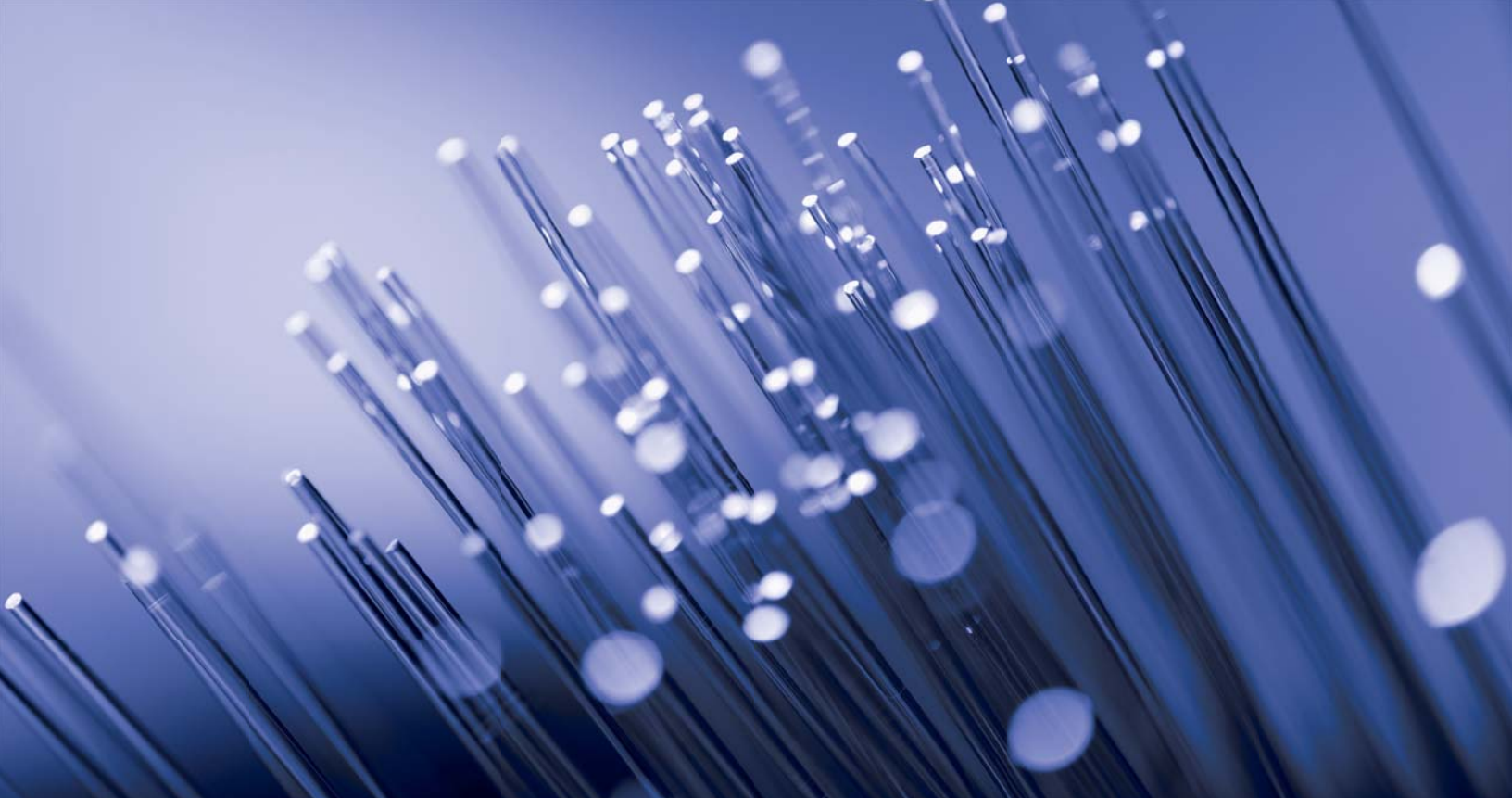
Stephen Pawlowski (Micron)  
 David Pellerin (Amazon Web Services)  
 Heike Riel (IBM)  
 Kirill Rivkin (Western Digital)  
 Gurtej Sandhu (Micron)  
 James Sexton (IBM)  
 Siva Sivaram (Western Digital)  
 Naveen Verma (Princeton)  
 Andy Warfield (Amazon Web Services)  
 Steven Woo (Rambus)  
 Ian Young (Intel)

## References to Chapter 2

- <sup>1</sup>V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, W. L. Hughes, "Nucleic acid memory", Nature Mat. 15 (2016) 366-370
- <sup>2</sup>Silicon Wafer Shipment Statistics, <https://www.semi.org/en/products-services/market-data/materials/si-shipment-statistics>
- <sup>3</sup>IDC Global Data Sphere Report, 2020, <https://www.idc.com/getdoc.jsp?containerId=prUS46286020>
- <sup>4</sup>Ron Ho, "Memory and Future of Computing", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>5</sup>Steven Woo, "Memory Solutions for Datacenter Workloads", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>6</sup>James Sexton, "Compute Memory Trends: from Application Requirements to Architectural Needs", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>7</sup>Simon Hammond, "Memory and Storage Demands in Biological Data Analysis", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>8</sup>Kjiersten Fagnan, "Memory and Storage Demands in Biological Data Analysis", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).



- <sup>9</sup>Thomas Jew, "Automotive and Embedded NVM", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>10</sup>D. Reinsel, J. Gantz and J. Rydning, "Data age 2025: The digitization of the world from Edge to Core", An IDC White Paper #US44413318, Nov 2018, <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>
- <sup>11</sup>Gary Bronner, "Memory Trends and Outlook - Mobile, Wearable, and IoT Application Driven", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>12</sup>G. Batra, Z. Jacobson, S. Madhav, A. Queirolo, and N. Santhanam, "Artificial-intelligence hardware: New opportunities for semiconductor companies", McKinsey & Company, Jan. 2, 2019 <https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies>
- <sup>13</sup>Carlos H. Diaz, "Next Gen AI and Technology Trends", Invited Talk at the SRC – C-BRIC Annual Review, Oct. 6-8, 2020
- <sup>14</sup>Rashid Attar, "Memory Trends and Outlook - Mobile, Wearable, and IoT Application Driven", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>15</sup>Kelly Baker, "IoT and Automotive Perspective on Memory", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>16</sup>Subhasish Mitra, "On-Chip Memory and its Dense 3D Integration for Abundant-Data Computing at the Edge", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>17</sup>Naveen Verma, "In-memory Computing Across Technologies", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event)
- <sup>18</sup>Vijaykrishnan Narayanan, "Memory in Machine Vision", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>19</sup>Nafea Bshara, "The Future of Clouds", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>20</sup>Sailesh Kottapalli, "Memory Hierarchy in Datacenter Processors", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>21</sup>Dan Ernst, "Application-optimized Architecture Trends for Memory and Storage Systems", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>22</sup>Stephen Pawlowski, "Emerging In-memory Compute Architectures", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>23</sup>Bruce Jacob, "Whither External Memory?", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>24</sup>Al Fazio, "Memory Technologies: The Long Gestation from Research to Mainstream", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>25</sup>Greg Atwood, "Memory Technology Enablement of Future Computing Systems", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>26</sup>Jaewoo Jeong, "Spintronic RAM: Present and Future", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>27</sup>Stefan Müller, "Emerging Ferroelectrics Memories", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>28</sup>Gabriele Navarro, "PCM and RRAM", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>29</sup>Derchang Kau, "Selector Devices for Cross-bar Memories", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>30</sup>Peter Faulhaber, "Is the LTO Roadmap Enough to Fend Off Competitive Technologies like HDD?", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>31</sup>Andy Warfield, "Durability and Storage from a Systems Perspective and How Much Work is Involved as a Consumer of that Media", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>32</sup>Mark Helm, "NAND SSD - Is the sky the limit and what will it take to fuel the rocket?", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event).
- <sup>33</sup>Siva Sivaram, "Emerging Trends in HDD, SSD and Moonshot Technologies", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event)
- <sup>34</sup>John Morris, "Innovating HDD", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event)
- <sup>35</sup>David Markowitz, "DNA Storage: How Could it Potentially Intercept Existing Storage Mediums?", SRC workshop on New Trajectories for Memory and Storage Technologies, Oct. 25-29, 2020 (Virtual Event)
- <sup>36</sup>M. Hilbert and P. Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information", *Science* 332 (2011) 60-65
- <sup>37</sup>IDC FutureScape: Worldwide IT Industry 2019 Predictions, <https://www.idc.com/getdoc.jsp?containerId=US44403818>



---

## Chapter 3

# New Trajectories for Communication

### Seismic shift #3

Always available communication requires new research directions that address the imbalance of communication capacity vs. data generation rates.

### 3.1. Executive Summary

The current state of the developed world is characterized by (almost) *always-available* communication and connectivity, which has a tremendous impact on all aspects of life. A manifestation of this is Cloud Storage and Computing. The ability to get data from anywhere and send it to everywhere has transformed both the way we do business and our personal habits and lifestyle. Social networks are a prime example.

However, the main concept of the cloud is based on the assumption of constant connectivity, which is not guaranteed. Furthermore, the demand grows daily for communication to become more ubiquitous as we become more connected. An alarming trend is a growing gap between the world's technological information storage need, as just discussed, and communication capacities shown in **Figure 3.1**. For example, while it is currently possible to transmit all world's stored data in less than one year, it is predicted it will require at least 20 years for the transmission in 2040. A global storage-communication crossover is expected to happen around 2022, which may have a tremendous impact on ICT. Besides the cloud being used as storage for mass information, it is also heavily used for compute, specifically for AI application. Although edge computing for AI applications is a fast-growing trend, there are numerous applications that rely on cloud compute

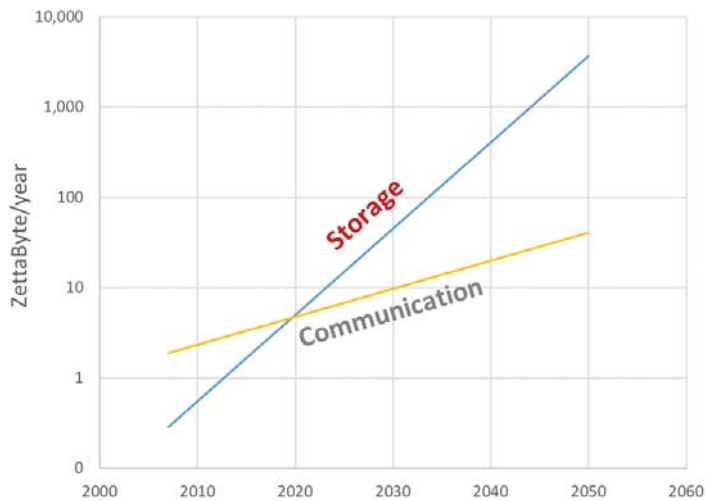


Figure 3.1: The Global Communication Data Generation Crossover occurs when the data generated exceeds the world's technological information storage (see Chapter 2) and communication capacities (see Appendix to this chapter), creating limitations to transmission of data.

capabilities. The explosion of information generated on the edge and processed and/or stored in the cloud will require tremendous growth of communications infrastructure.

### Call for action

Radical advances in communication will be required to address growing demand. For example, the cloud technologies may undergo substantial changes, with emphasis shifting toward edge computing and local data storage. Broadband communications will expand beyond smart phones to immersive augmented reality, virtual meetings, and smart office settings. New capabilities will enrich user experiences through new use cases and new vertical markets. This requires collaborative research spanning a broad agenda aiming at establishing revolutionary paradigms to support future high-capacity, energy-efficient communication for the vast range of future applications. The DOE Office of Science published a report in March 2020 to identify the potential opportunities and explore the scientific challenges of advanced wireless technologies<sup>1</sup>.

**Challenges would include wireless communication techniques expanding to sub-THz region, wireless and wireline technologies interplay, new approaches to network densification, increasing importance of security, new architectures for mmWave, and device technology to sustain bandwidth and power requirements, packaging, and thermal control.**

<sup>1</sup><https://www.osti.gov/servlets/purl/1606539>

<sup>ii</sup>The Decadal Plan Executive Committee offered recommendations on allocation of the additional \$3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

### Communication Grand Goals:

- Advance communication technologies to enable moving around all stored data of 100-1000 zettabyte/year at the peak rate of 1Tbps@<0.1nJ/bit
- Develop intelligent and agile networks that effectively utilize bandwidth to maximize network capacity

**Invest \$700M annually throughout this decade in new trajectories for communication. The priority research themes are outlined in this Chapter<sup>ii</sup>.**

## 3.2. Future Communication Technologies: Fundamentals, Challenges, and Application Drivers

### Overview and needs

Over the past several decades, communication technologies have supported daily social life and economic activities and continue to increase in importance. While it is not easy to predict how communication technologies will develop in the future, it is worthwhile to explore best-case scenarios that are bounded only by what is technically possible. This chapter identifies fundamental limits and provides an open forum for brainstorming unserved and future applications, as well as their corresponding implications for the semiconductor industry. Furthermore, new emerging communication solutions are discussed in the context of the application space they enable and the technology trends they may establish in energy, spectral efficiency, etc.

In many ways, technology is driving the future of human and machine communication. The amount of data that is being generated continues to rise exponentially, which means that we will have to come up with better ways of transmitting large amounts of data securely and without loss. The future will be driven by intelligent edge nodes based on high-speed devices, local processing, and local analytics. *Edge computing* is a very promising area of research because pulling *intelligence to the edge* could significantly improve latency to the required level (i.e., below 1ms). There is a large window of the electromagnetic spectrum that remains

untapped for communication<sup>1</sup>. Technology innovations are needed to leverage a vast, untapped spectrum. Innovative semiconductor technologies (e.g., RFSOI, FinFET, and SOI/SiGe-based photonics) will provide process platforms for the future. To establish true ubiquitous coverage at lower costs, significant technology advances are required in areas such as power amplifiers at higher frequencies, channel estimation techniques, low-cost security for IoT devices, etc. An important topic is the application of AI/ML and exploring parallels between communication networks and neural networks. Security needs to be made a priority, with a focus on hardware, analog, and overall network security. Better mitigation strategies are also necessary for dealing with a network that is under attack. Finally, business model innovations are needed that adapt to the changing requirements and demands of the end user.

### mmWave trends

With the introduction of 5G and the increase in the number of devices that communicate with each other, *mmWaves are being explored and implemented as a key technology for two reasons*. First, the spectrum is large and capable of accommodating the sizable number of devices requiring a communication channel. Second, most of these applications typically require communication among devices in close proximity, and mmWaves have proved effective in LOS communication. Cellular networks operate mostly in relatively narrow licensed bands below 4 GHz, where signals propagate reasonably far through free space, but where available spectrum is somewhat limited. Unfortunately, the bandwidth of available spectrum has a direct impact on the maximum data rate of the transmissions in these bands. According to the well-known Shannon-Hartley Theorem, capacity is a linear function of bandwidth. Early deployments of 5G mmWave are underway, but coverage is intermittent for now.

For mmWave systems, LOS directionality requirements are addressed through narrow beams with the need to track

receivers. The mmWaves signals are absorbed by many common materials which limits their range and increases blockage probability. The device power consumption is very high due to the high sampling rate and large number of antennas. As network density increases, we need fiber access of mmWave backhaul with improvements in edge services. To get power-efficient directional search, a conventional analog phased array can be used, which performs beamforming with RF phase shifters. It consumes low power because there is only one mixer+ADC per stream, but it can only search in one direction at a time. The most promising architecture is a fully *digital phased array* where all streams are digitized, and it can search in all directions simultaneously. It has low power consumption when utilizing low resolution ADC, and the dominant power is due to the local oscillator. Fully digital architectures have fast directional searches that are robust to blockages and have improved rate enabling multiple streams.

THz and sub-THz are a largely unexplored spectra that provide massive bandwidth with a LOS link band greater than 100 meters with active antenna arrays and modest power consumption. While 140GHz is prohibitive for smartphones, it may be a strong candidate for applications of robots, drones, and point-to-point links.

In summary, 5G offers low latency and high peak rates for mobile edge computing and real-time control systems, but there are challenges, including MEC deployment, intermittent links, TCP adaptation, and multi-path routing<sup>2</sup>.

### Power aspects

Initial estimates are that 5G networks will consume over 3.5 times more power than 4G networks and may need around 2 to 3 times smaller cells to obtain full coverage at higher frequencies. Electricity already makes up about a third of a carriers' average operational costs. In the past, to reduce power consumption, efforts were made to improve the electronics, especially power amplifiers, DSPs, and displays.

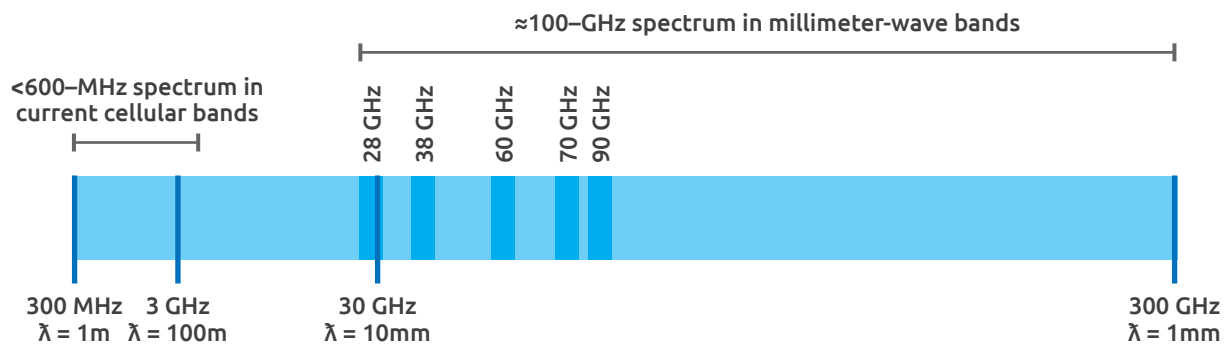


Figure 3.2: mmWave spectrum breakdown (adapted from<sup>2</sup>)



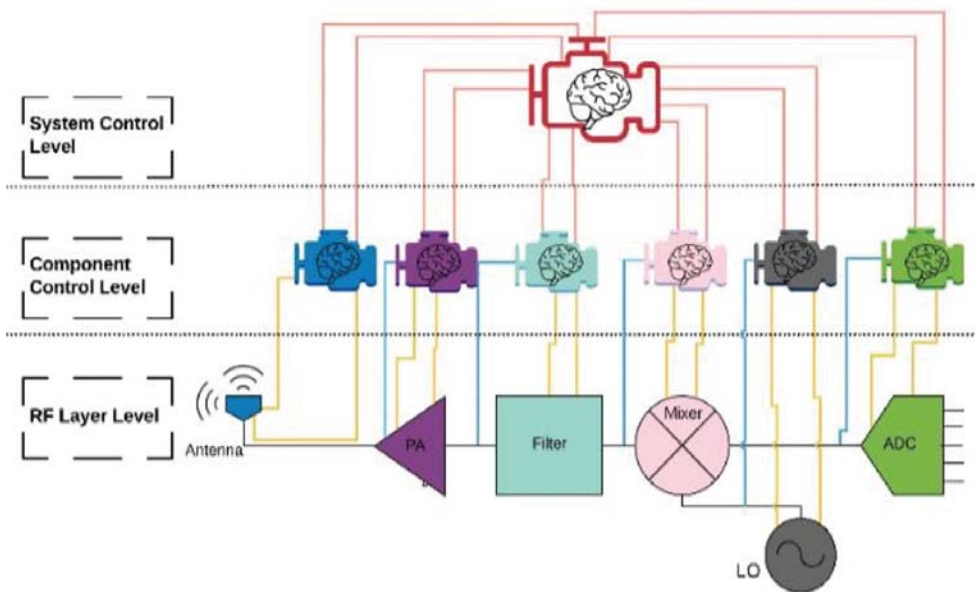


Figure 3.3: Network enhancements (courtesy of Jeffrey H. Reed, Virginia Tech<sup>3</sup>)

More efficient antennas and renewable energy sources have been considered as well. Improvements in the efficiency of Radio Resource Management (RRM) and planning can also lead to a reduction in power consumption. These approaches should continue to be refined, but there are more innovative and potentially more effective ways that could be explored. Some innovations that have come up in RF include analog FFT, *cognitive RF-AI circuit control*, neuromorphic training with analog circuits, and avoiding cooling through distributed radio systems. Neural network techniques have been used to obtain near-optimal solutions to complex optimization problems in Radio Resource Enhancements.

Multi-objective optimization across networks can be used to optimize power while maintaining a target QoS. Some other avenues that have been explored in radio resource enhancement include power modulation, coding, and context-based optimization of antenna resources and bands. Power could be reduced by viewing the whole *system as a neural network*, and employing adaptations that use multi-objective metrics to balance power while meeting QoS, as well as trading excessive degrees of freedom provided by the array. New research has shown that a neural network<sup>3</sup> can be utilized to design/optimize a complete radio system.

Future networks will be driven by mobile, P2P, V2V, M2M, and embedded applications, to name a few. For example, *it is anticipated that by 2030 M2M will constitute 90% of the total communication traffic*<sup>1</sup>. This is typically information-centric/named-data networking, where data is named instead of using data containers. This directly secures data and fully

utilizes wireless broadcast media. These will give rise to new communication needs like wireless, infrastructure-free communication, scalable communication, and secure connection, which are particularly important for communications with random encounters.

### Security aspects

The challenge with this exploding communication technology is that networking needs to be done with a potentially unlimited number of computing devices that would need autoconfiguration

and auto-updates, basically working autonomously, with security as the key underpinning factor<sup>4</sup>. With the rise in IoT technology, **security becomes even more critical**. From a security perspective, IoT devices are much more heterogeneous as compared with traditional devices, and they are mostly neglected. Automatic updates, that are already canonical in the desktop and mobile operating system space, require cryptographic primitives for resource-constrained devices, as well as building a PKI infrastructure to support trusted updates. Over time, the risk these devices pose to the Internet commons will only grow unless they are taken offline. The recent Mirai attacks were primarily enabled due to the absence of security best practices in the IoT space, which resulted in a fragile environment that was ripe for abuse. This is a combination of device and communication problems, inseparable in cyberspace that is made of trillions of interconnected devices.

One of potentially disruptive approaches is to apply **security measures that are bio-inspired**, like from our immune system<sup>4</sup>. Any invader that breaches the physical barrier of skin or mucosa is countered by the immune system, and this system is capable of learning and adapting to protect us from multiple invaders. As a parallel, in information-centric networks, all entities should have: a semantically meaningful name that represents the context; produce keys; and a test anchor that issues certificates and installs security policies. Encryption at the data producer can prove useful, so that all data exchanges are authenticatable. *Measures like a hardware pseudo-random number generator, TPM to safely keep private*

keys, and hardware crypto-accelerators can help in improving security. These trends also impact societal trends, as they empower end users by placing more control with them.

### Emerging quantum communication technologies

Quantum technology is being explored because of its unique properties, and there are plans for a global fiber network distributing quantum entanglement (Figure 3.4). Some of the devices that will be used are quantum repeaters (QR), switches (QS), interconnects, routers, and memories. Qubits are already being used for creating secure keys—random strings of 0s and 1s—that can then be used to encode classical information, an application called quantum key distribution (QKD). The fundamental issue is how to devise quantum memories, quantum repeaters, and quantum interconnects that carry entanglement/preserve coherence over secure coast-to-coast networks. There are limits to extending entanglement and coherence that have not yet been quantified. We need to determine the origins and limits on the control of decoherence and extend quantum state storage times to permit efficient interrogation, minimal error, and teleporting the quantum state to another link. Scalable quantum error correction and quantum algorithms are needed to exploit the quantum computing advantage and demonstrate quantum superiority in security, capacity, and machine learning<sup>5</sup>.

### Key areas of focus and follow-on research

- Development of intelligent edge nodes with a focus on always-on devices, high bandwidth devices, and new modalities for security
- Technology innovation to enhance transmit power, especially in the untapped spectrum (100GHz-1THz)
- Innovative semiconductor process platforms to include RFSOI, FinFET, and SOI/SiGe-based photonics

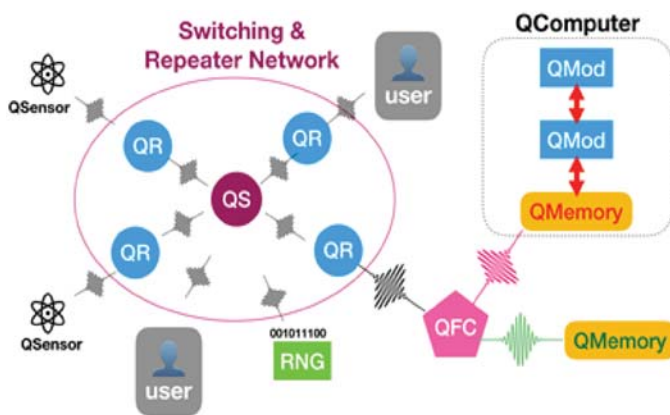


Figure 3.4: Quantum Internet flow (courtesy of Fil Bartoli, NSF and Lehigh University<sup>5</sup>)

- Solutions for cost-effective ubiquitous coverage
- Security for communication designed into the hardware
- mm-Wave applications that overcome issues of blockage and power consumption, with massive data-rate support and ultra-low latency
- Exploring parallelism between communication networks and neural networks, using multi-objective metrics to balance power and quality of service, as well as radio resource enhancements
- Learning from biology to shape future communication systems, finding inspiration from immunity systems
- Merging innovations from quantum technologies into communication networks

## 3.3. Life After 5G

### Overview and needs

As 5G is now entering into the massive commercialization phase, it is important to identify trends, challenges, and research goals to support future developments, including 6G and beyond. This section identifies key elements of future mobile communication and explores cellular networks, infrastructure, and subscriber alternatives to maximize system-spectral efficiency and minimize energy consumption. This section also includes analysis of transceiver components from antenna/modulation to demodulation/antenna against anticipated requirements for latency, data rate, and overall quality of service.

There is significant interest on the differentiation in the portfolios of 5G and 6G, considering some of the challenges presented to 5G deployments. Expectedly, in 6G, there will be an even greater increase in area traffic capacity, connection density, peak data rate, and spectrum efficiency, and there will be a lowered end-to-end latency in service<sup>6</sup>. Stronger roadmap emphasis would be on industry verticals including, but not limited to, automotive, e-health services, energy, media and entertainment, and industrial automation<sup>7</sup>. The standardization partnership (3GPP) that produces the specification for 5G has completed release 16 (Rel-16), which expands the features for low-latency and time-sensitive IoT communication. This has improved on Release 15 (Rel-15), which featured limited bandwidth, laid stress on fixed wireless connection and smart phones, and lacked native support data-driven learning. It is speculated that 6G could open up the 300 GHz to 3 THz range. The classification of frequency bands of interest is highlighted in Figure 3.5<sup>8</sup>.

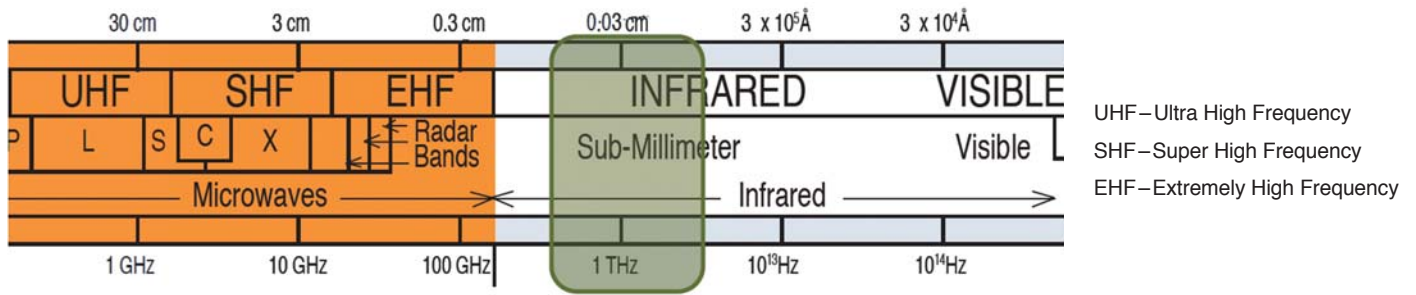


Figure 3.5: New spectrum from 100 GHz to 3 THz presents opportunities for communication and imaging<sup>8</sup>

## 6G trends

It is necessary to foresee the hardest challenges in extending to terahertz operation. The **first is the heightened path loss beyond 1 THz. Secondly, for the 1-10 THz band, only a few channel measurements are available thus far<sup>9,10</sup>. Thirdly, THz frequency is beyond the  $f_{max}$  of current CMOS and SiGe transistors, and this makes low-cost, high-power and low-phase-noise design difficult<sup>11</sup>.** As it becomes imperative to reduce power consumption and have effective link budgets, multiple antennas will become important to provide larger and more effective aperture. Also, as mixed-signal devices like ADCs and DACs dominate the total power consumption at higher frequencies, greater efforts will be made to develop **architectures that will utilize ultra-low resolutions** of AD converters.

A rethinking of overall system architecture is also required. A possible approach is that the Base-Station (BS) can work in synergy with User Equipment (UE) to synchronize, encode/decode, and perform channel estimation<sup>6</sup>. In this approach, system design prioritizes good correlation properties of synchronization signals that are typically distorted by ADC and DAC quantization noise<sup>12</sup>. The new system must also better handle the channel estimation compromised by the low-resolution quantization so beam misalignment and error can be avoided.

Here, the convergence of EM wave theory, circuit theory, and information theory will help design physically consistent antenna and RF models, so a closer spacing in an array can provide more gain and broader bandwidth. Of course, implementation may be application-specific. For aerial vehicles, for example, this would mean better propagation modelling, MIMO modes, channel-tracking, and joint communication, while simultaneously managing interference and employing efficient sensor fusion and distributed computing. In medicine, it would mean better patient-data processing, replacing wires with cellular links and enabling secure links. Also, better situational awareness can be

achieved through machine learning. For radar and mmWave applications, capturing reflections from surfaces and running classification problems for single-beam selection would be interesting<sup>13</sup>. In the process, this would also collect sufficient data at spatial, frequency, and temporal scales to enable ray-tracing for better channel consistency and elegantly incorporate mobility of scatterers and transceivers<sup>14</sup>. **The challenge will be to obtain meaningful data sets at the appropriate spatial/frequency/temporal scales with enough detail but without a computational burden.**

## 5G as stepping stone to future systems

To outline future requirements, it is important to revisit 5G as the communication innovation platform for this decade. It is estimated that the deployment of 5G wireless technology will yield an estimated \$13.2 trillion in goods and services by 2035<sup>15</sup>. One of the main features of 5G is its distributed functionality in terms of latency, i.e., low latency would be assigned for delivery of local value content and storage and AI-processing, while longer-latency service would be assigned to Big Data and aggregated value processing. As the roadmap in **Figure 3.6** shows, the first step will be delivery of mobile broadband using existing mobile frequencies and mmWave frequencies for increased bandwidth. Currently, while the Sub-6 service (i.e., 5G deployed on frequencies under 6GHz) may have better coverage than mmWave communication and is likely the most valuable spectrum, it isn't easily available, resulting in lower data rates. Release 16 (Rel-16) also seeks to enable deployment in currently unlicensed spectrums. Rel-16 will help anticipate further requirements in Release 17 (Rel-17), such as highly synchronized support for industrial IoT, which will provide access to network edge computing and automation. It is expected that Rel-17 will also expand operating frequencies to the 50-150 GHz range.

**Satellite-based broadband internet** and mobile communication show promise to bridge the world's digital divide. Photonics is another field that will provide massive improvements to communication infrastructure because of its ability to transfer

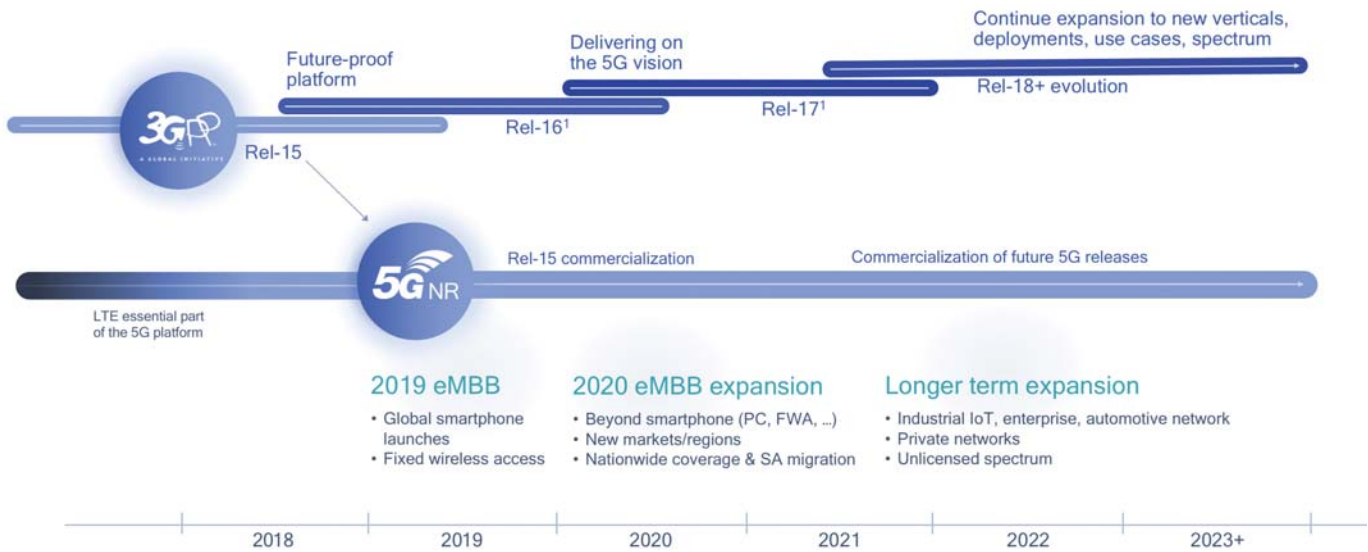


Figure 3.6: Release versions' timeline for the deployment of 5G services (courtesy of John Smees, Qualcomm<sup>15</sup>)

data over long distances at the speed of light. 5G delivers the ability to expand network capacity and data rates while driving new use cases and application requirements, such as wide-area and IoT applications, enhanced mobile broadband, and ultra-reliable, mission critical applications. 5G internet goes beyond just smartphones to competitive gaming, VR, and smart offices. Several factors of 5G implementation must be considered to address both user and operator challenges.

**Device complexity and cost, power consumption and battery life, network densification, and network security and reliability are some of the integral challenges.** 5G mmWave requires *cell densification* due to high propagation path loss. Base station cell coverage hinges on two key characteristics: (1) high effective isotropic radiated power (EIRP), which gives greater TX coverage per cell, and (2) lower noise figure (NF), which gives greater RX coverage per cell. When the same coverage is maintained and operating power is lowered by using a smaller array, cooling requirements are lowered and the system has less weight in addition to lower operating cost. At higher EIRP and lower NF for base stations, we can provide greater coverage at the same array size. Fewer base stations are needed for equivalent coverage<sup>1</sup>.

### Antenna arrays considerations

Emerging applications, such as augmented reality, holographic video, and wireless cognition, may pose further challenges to physical layer design. This places stress on existing Massive MIMO techniques that could provide solutions. Consider augmented reality, for example, which requires data rates of up to 3 GB/s/user. One of them is Ubiquitous cell-free Massive MIMO<sup>16,17</sup>, which refers to a distributed system that implements coherent user-centric transmission to circumvent

inter-cell interference limitations and provides increased macro-diversity. Additionally, large intelligent surfaces can be employed to surround users with arrays for effective near-field operation. This can be coupled with Holographic MIMO with continuous transmitting and receiving surfaces.

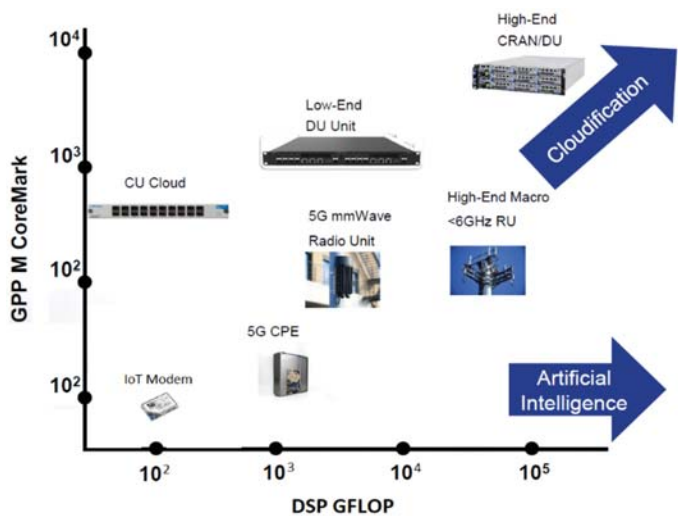
Systems in THz band will revisit the concept of spatial multiplexing versus providing each user with his/her own spectral band. Different approaches may be required to maintain the power aperture product as carrier frequency increases. Consideration must also be given to **meta materials** and new types of adaptive arrays.

### Baseband computation challenges

There are a number of issues in relation to baseband signal processing that must be effectively addressed for successful 5G and (prospective) 6G deployment. **The increased bandwidth would mean a higher sampling rate. This would place stress on data converter design, which is coupled to the power requirements of the ADCs/DACs.** There are also a large number of use cases in terms of performance, latency, form-factors, and cost considerations. All this is compounded by the need for multi-band RF design and interfacing. **Figure 3.7** shows the shift in general purpose computing that is required to make this possible, including processing requirements that must increase by over three orders of magnitude.

It is hoped that, when these are resolved, 6G will realize data rates in excess of 100 Gbps with sub -1 ms end-to-end latency. The expectation is that 6G will be a *Multiple Radio Access Technology (Multi-RAT) that uses less power and has greater coverage on land, sea and, probably, in space.* Developing higher-fidelity Software Defined Radio (SDR) then becomes





CU—Centralized Unit  
 DU—Distributed Unit  
 RU—Resource Units  
 CPE—Customer Premises Equipment  
 CRAN—Centralized Radio Access Network

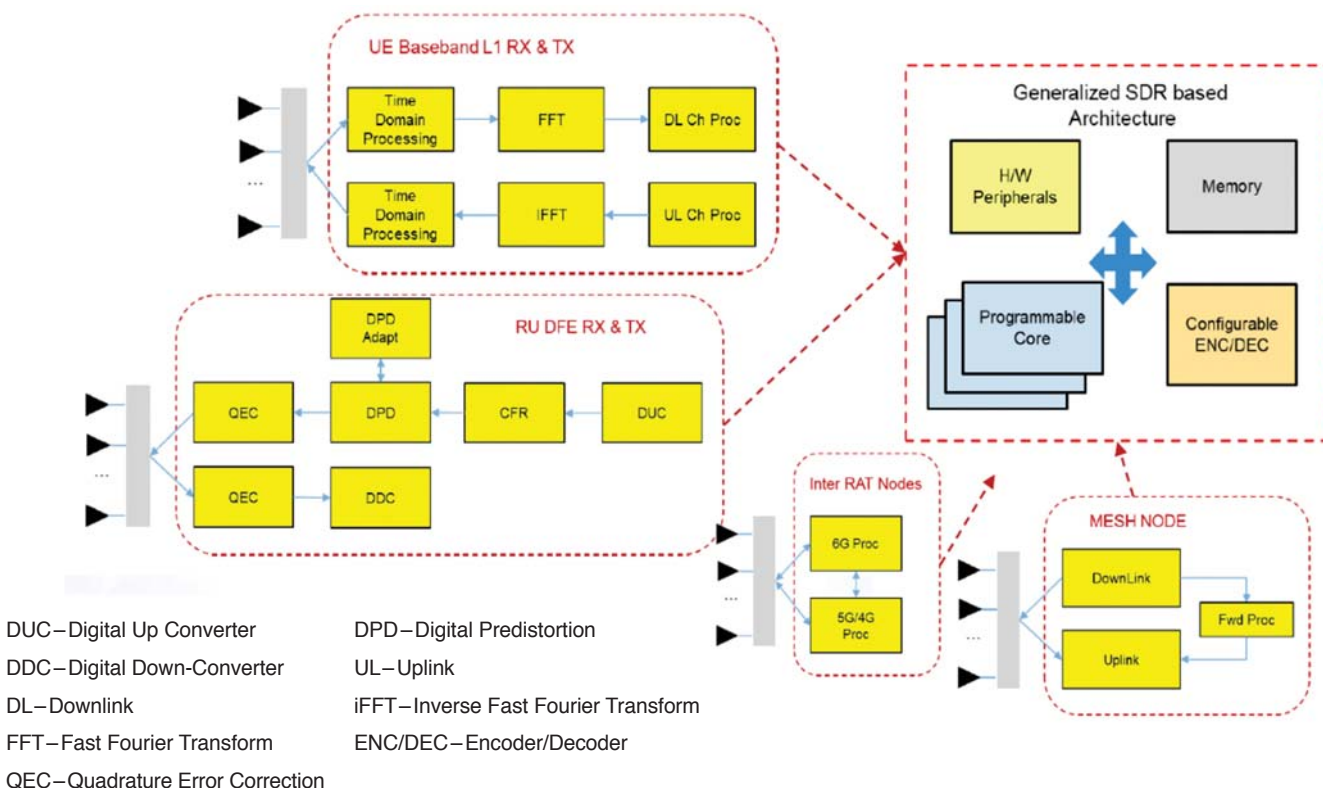
Figure 3.7: Evolution of general-purpose computing and DSP scaling for 5G and 6G (courtesy of Jayesh Kotecha, NXP<sup>18</sup>)

necessary, as pre-silicon design and verification efforts are becoming too costly for state-machine-based modems. Since scalability will be a key feature of future architectures, important trade-offs will be required in platform optimization between programmable section and hardware accelerators.

The skeletal **5G/6G-oriented computing architecture** is shown in **Figure 3.8**. Faster computing will further transform analog functionality into purely digital computing. Standards should also evolve and facilitate multi-layer densification topologies. They should also establish multiple Power Efficiency vs Coverage vs Spectral Efficiency trade-off points and support cross-layer optimization in terms of AI-based scheduling and interference management. Lastly, Integrated Access and Backhaul (IAB) should be evaluated more thoroughly as a means to control costs of ultra-dense 5G mmWave networks<sup>19</sup>.

### Management of client-level data generation

Cloud computing may not sustainably support the ever-increasing generation of digital data. Local data, such as user-generated image, video, and speech, is best handled near the source to alleviate processing-power concerns. **Figure 3.9** shows the progression over the past two decades of contributions towards structured productivity data, unstructured broadcast media data, and locally processed data, with the highest growth being in the last category. This complicates the frequency planning and the number of band allocation that a wireless system has to contend with. Also, as higher frequency carriers are used, cell density increases and this diversifies the types of links, traffic services, and interference scenarios.



DUC—Digital Up Converter  
 DDC—Digital Down-Converter  
 DL—Downlink  
 FFT—Fast Fourier Transform  
 QEC—Quadrature Error Correction  
 DPD—Digital Predistortion  
 UL—Uplink  
 iFFT—Inverse Fast Fourier Transform  
 ENC/DEC—Encoder/Decoder

Figure 3.8: 5G and 6G oriented architectures for supporting an increased number of device types (courtesy of Jayesh Kotecha, NXP<sup>18</sup>)

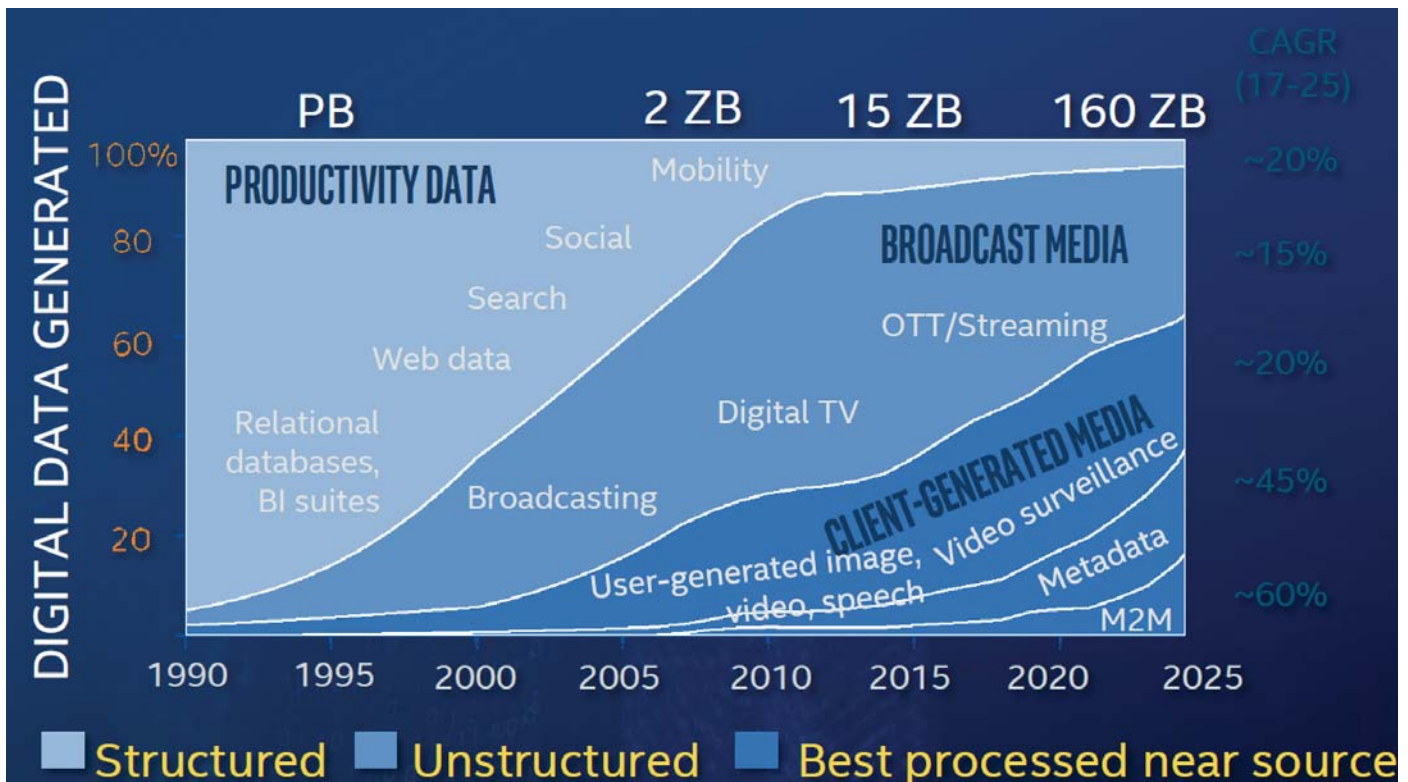


Figure 3.9: Progression in contributions of structured, unstructured, and user-generated local data (courtesy of Ariela Zeira and May Wu, Intel<sup>20</sup>)

**A systems view is necessary to set new objectives given the foreseeable complexity.** Network designs should be tailored to harvest and process data conducive to learning algorithms. Also, the structure of the networks should be tailored to be dynamic and self-aggregating, with opportunistic connections. More concretely, it is possible to transform signal processing in communication by combining the merits of different AI-based approaches. This preserves decades of domain knowledge with appreciation for compensating model inaccuracies. Neural networks can assist a probabilistically directed graphical model, such as the Bayesian Network. Neural networks are inherently good for complicated unknown scenarios and exhibit robustness. But they *suffer drawbacks in terms of the need for larger data sets and are, in fact, inefficient for small-scale models. These flaws can be remedied by a complementary Bayesian Network, wherein the variables and their dependencies explicitly represent model features and efficiently handle small data sets.*

### Key areas of focus and follow-on research

- Develop strong roadmap with emphasis on industry verticals (e.g., automotive, e-health, energy, media and entertainment, and industry automation)
- Consider systems with high bandwidth utilizing large arrays in THz range; focus on changing aspects of PHY; establish wideband models for arrays; extend operations for THz, including new models for signal processing, new RF architectures, and new devices
- Expand utilization of AI/ML in wireless systems, data-driven models, integration into network; self-healing and self-learning approaches
- Impact of network densification and solutions for backhaul
- Cross-layer optimization, utilizing AI for scheduling and interference management
- New design principles for wireless networks, from data insight to action; security as key component of design; new performance metrics; expanded role of ad-hoc and mesh networks

### 3.4. Fixed-line Communication

#### Overview and needs

The exponential growth in data volumes that need to be communicated cannot be supported entirely through wireless technologies. In fact, the core of a mobile network is represented by fixed-line optical fiber communication, and the advent of 5G+ calls further increases optical fiber-line capacity. It is necessary to identify the key value of and challenges for the fixed/wireless convergence, including device, circuit, and system solutions. Improving network capacity would mean more spectrum (100 Gb/s per cell site), spatial reuse, and better spectral efficiency. To achieve this, *a new digital fabric will be implemented that integrates existing access networks, newer cloud computing paradigms, and optical and fixed-line data transfer infrastructure, among others* (Figure 3.10).

#### Trends and challenges of the fixed/mobile convergence in the 5G and 6G era

*Intense R&D is underway in newer technologies like low-cost, easy-to-install small cells<sup>22</sup>, spectral reuse with grids of beams, zero-forcing algorithms, and distributed Massive MIMO<sup>23</sup>.*

An interesting approach is to split the workload between edge computing and the radio site for all functionalities like analog signal processing, data converters, Common Public Radio Interface (CPRI), and distributed systems, as shown in Figure 3.11. In particular, for each shared functionality, there should be a good trade-off between flexibility and efficiency. On one hand, at a processing-hardware level, cloud-computing hardware consists of high-power CPUs, GPUs, and possibly other processor units capable of multi-functional pooling for

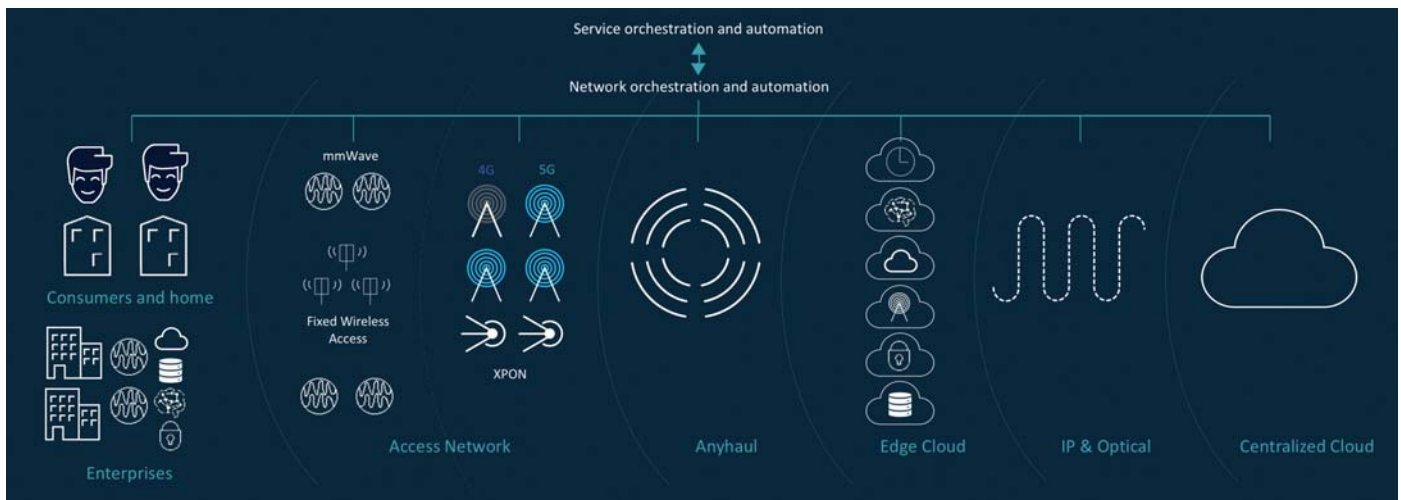


Figure 3.10: Proximal digital network for productivity and knowledge creation (courtesy of Peter Vetter, Nokia Bell Labs<sup>21</sup>)

|            | Analog                 | $\Sigma\Delta$    | CPRI              | eCPRI<br>Low layer split | High layer split | Classical distributed |
|------------|------------------------|-------------------|-------------------|--------------------------|------------------|-----------------------|
| Edge cloud | Layer 3                | Layer 3           | Layer 3           | Layer 3                  | Layer 3          |                       |
|            | Layer 2                | Layer 2           | Layer 2           | Layer 2                  | Layer 2 high     |                       |
|            | Layer 1                | Layer 1           | Layer 1           | Layer 1 high             |                  | 10 Gbps<br>>5 ms      |
|            | ADC / DAC              |                   |                   |                          | 10 Gbps<br>5 ms  | Layer 3               |
|            | 3x64x100MHz<br><0.1 ms | 1 Tbps<br><0.1 ms | 1 Tbps<br><0.1 ms | 100 Gbps<br><1 ms        |                  | Layer 2 high          |
|            |                        | ADC / DAC         |                   |                          | Layer 2 low      | Layer 2 low           |
| Radio site |                        |                   | ADC / DAC         | Layer 1 low              | Layer 1          | Layer 1               |
|            | RF                     | RF                | RF                | RF                       | RF               | RF                    |

Figure 3.11: Options for a functional split in architectures for a case study of 64 TX/RX Massive MIMO 100 MHz using 3-sectors and 16 Layers (courtesy of Peter Vetter, Nokia Bell Labs<sup>21</sup>)



graphical rendering, AI, Machine Learning, and digital signal process. This allows for flexible programming and dynamic instantiation of new functions, supporting so-called slicing-of-network subsystems for multiple tenants. On the other hand, at the radio site, RF Integrated Circuits and System-on-Chip circuits have specialized programming and are low-power systems.

A **flexible xHaul architecture** is emerging that supports the most optimal functional split between remote radio units and central functions in the edge cloud<sup>24</sup>. It essentially consists of high-capacity switches and fiber or wireless optical links interconnecting remote radio sites, 5G points of attachment (5G PoAs), and centralized-processing units.

In order to meet wireless capacity demands of users with smart devices, service providers are deploying “small cells” to benefit from spectral reuse or allow for line-of-sight operation of mmWave frequencies. These can be “femtocells” or “picocells,” which are suitable for small businesses and are generally of low-power design. These are connected back to the network via fiber-based options (a process known as backhaul). This allows for Gb/s connectivity as **small cell sites are aggregated back to the core when mmWave technology is employed in combination with highly directive antennas**. **Figure 3.12** shows the trends and strategies in replacing direct wireline connections via distributed architectures for varying distances and data speeds, including Fiber to the Home (FTTH) and wireless small cells. The fixed wireline infrastructure will increasingly be used to support the backhaul or fronthaul of a high density of small-cell future 5G/6G networks.

## Scope for optical fiber in 5G

*The main drivers for higher throughput all-optical fiber communication have been better Internet service and links between datacenters. The main challenges are the cost and bandwidth of components and the affordability of photonic integrated circuits.* Fiber has been preferred for long-haul systems of greater 1000 km (submarine access), while RF options have been preferred for short mobile access of less than 1000 km. A typical coherent optical communication system consists of a transmitter and a receiver that convert digital bits to optical signals using electro-optical modulators (25-100 Gigabaud). The channel adds noise, so the signal requires amplification at spatial intervals along the chain of propagation. Similar to RF transceivers, the optical receiver demodulates the encoded information in amplitude, phase, and polarization, albeit by mixing with laser, which acts as a local oscillator (LO). Higher-order QAM is transmitted with higher symbol rate and a greater number of channels and bits/symbol, as shown in **Figure 3.13a**. **Figure 3.13b** shows a generic scheme for multiplexing modulated laser subcarriers and subsequent demodulation with coherent receivers. Encouraging results using **wavelength-division multiplexing (WDM)** in single-mode fiber show that capacities have been achieved recently of up to 115 Tb/s over 100 km using 250 channels<sup>25</sup> and 50.5 Tb/s over 17000 km using 295 channels<sup>26</sup>. While the standard transceiver rate today is 30 Gigabaud with QPSK modulation at 100 Gbit/s, it is hoped that faster rates like 80 Gigabaud with 256 QAM modulation at 800 Gbit/s

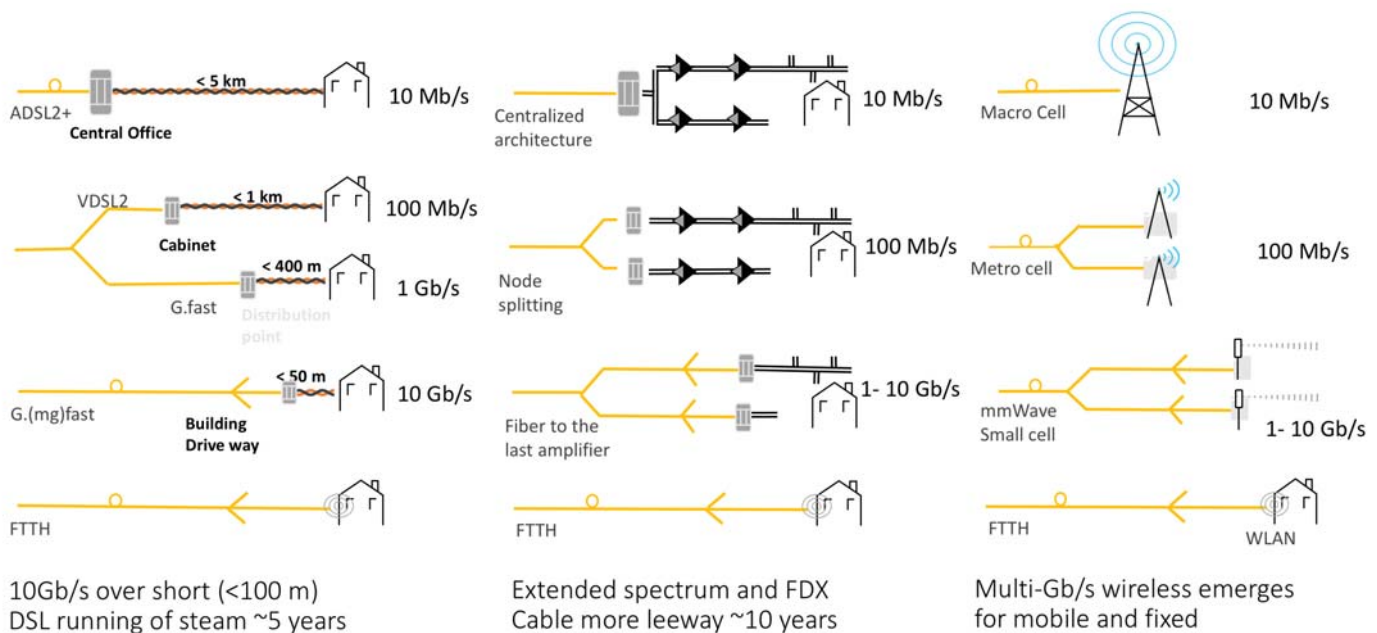


Figure 3.12: Various schemes for substitution of long-distance wireline access with Fiber to the Home (FTTH) and small cells (courtesy of Peter Vetter, Nokia Bell Labs<sup>21</sup>)



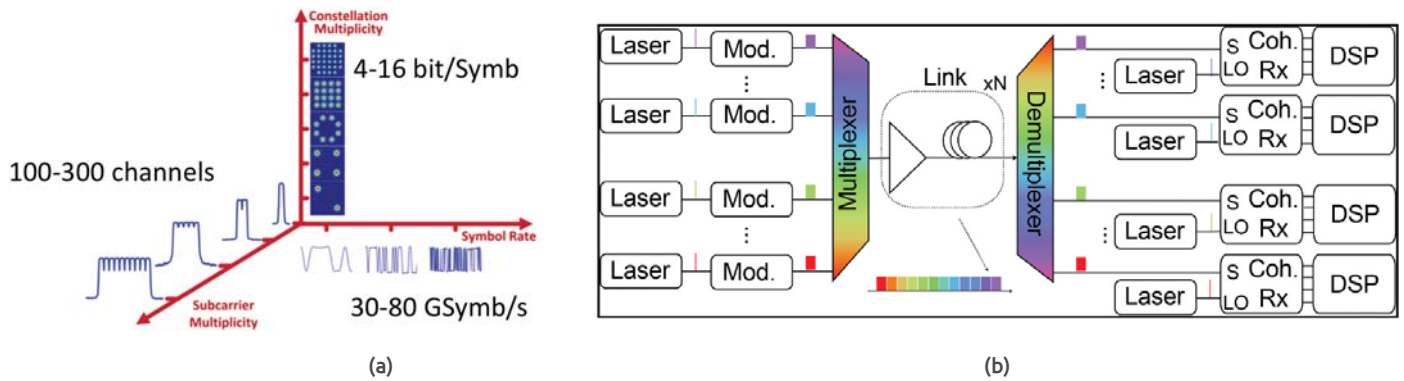


Figure 3.13: (a) Higher-order QAM signals are transmitted by larger total capacity fiber. (b) Lasers provide local oscillator frequencies that can be subsequently modulated and multiplexed for transmission (courtesy of Magnus Karlsson, Chalmers University<sup>27</sup>).

gain more traction<sup>27</sup>. *In the next five years, it is estimated that even higher symbol rates of greater than 100 Gigabaud at 1.6 Terabit/s channel rate will be materialized in integrated photonics.* Further, nonlinear equalization algorithms for higher order QAM signals<sup>28</sup> coupled with constellation shaping will improve optical signal integrity. This will be aided by faster ADCs/ DACs in 10 nm ASICs with bandwidths exceeding 100 GHz. Collectively, if these targets are met, **co-integrated CMOS and optics** on the same Silicon chip may be more inexpensively realized.

### Fiber links for low-energy frequency-stabilized coherent WDM

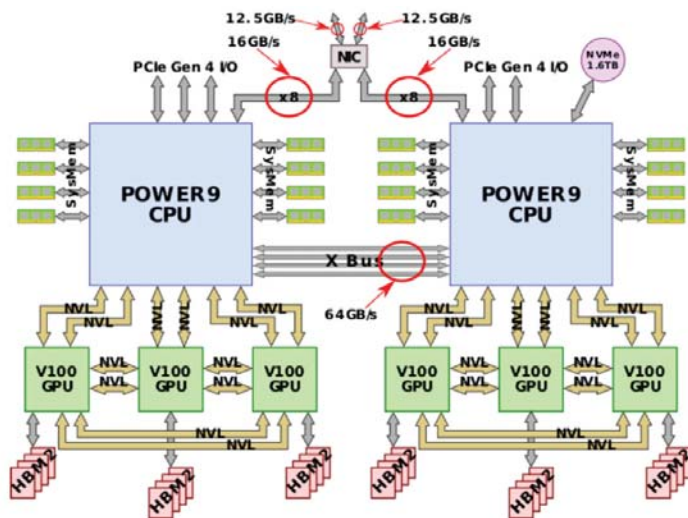
*Datacenter ethernet switching speed, currently at 25 Tb/s<sup>29</sup>, is on a steady upwards march towards 100 Tb/s.* However, fiber links that keep datacenter interconnect capacity manageable will soon face **fundamental scaling barriers in terms of power envelope, power consumption, and heat dissipation.** There will also be difficulties encountered by optical I/O modules in tracking 100 Tb/s ASIC switching signals. An elegant *Frequency-stabilized coherent optical (FRESCO) architecture<sup>30</sup>* is suggested to circumvent some of these issues. Here, narrow linewidth and laser stabilization technology developed for frequency standards and atomic clocks<sup>31</sup> is adopted for coherent fiber interconnects. Moreover, the laser used has continuous emission of a near-perfect single frequency of light with nearly zero frequency fluctuations and very long coherence time. In this system, a shared ultra-stable, spectrally pure laser that drives an ultra-stable, spectrally pure, shared optical comb (TX + LO) is employed. This laser is modulated with an integrated silicon photonic coherent transceiver. The shared WDM optical source consists of a silicon photonic tunable laser, a micro-scale ultra-stable optical reference cavity, a Silicon Nitride Brillouin laser, and a nonlinear optical frequency comb (OFC) source. This

arrangement facilitates coherent QAM modulation operating at 64 and 72 Gigabaud and under 64- and 256-QAM schemes and supports up to 1.6 Tb/s per wavelength on its frequency stabilized link<sup>29</sup>. The advantage of this new design is that it has no high-speed digital logic, no high bandwidth optical phase-locked loops, no Costas loops, and no out-of-band carriers.

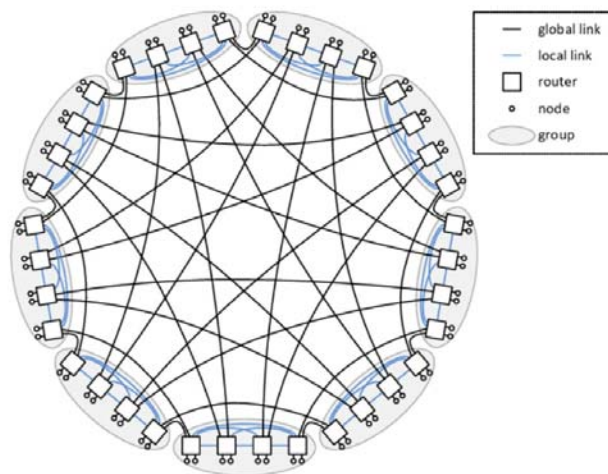
### HPC for fixed-line interconnection network fabrics

Traditional High-performance Computing (HPC) workloads are being written using the Bulk Synchronous Programming model (BSP)<sup>32</sup>. This model encourages message aggregation in order to maximize communication performance. **However, this model faces challenges in data partitioning, system partitioning, network congestion, and job isolation. With the advent of data-centric computing, emerging workloads will now integrate numerical simulation with data analytic capabilities, e.g., graph analytics and machine learning that are asynchronous and require fine-grained communication.**

From a computing fabric perspective, it is recognized that the primary cost in terms of performance and energy is data movement. In current designs, about 1 nJ of energy per 64-bit word is required for local interconnect traffic<sup>33</sup>. Current heterogeneous system architectures like ORNL's Summit are challenged to implement traditional message aggregation strategies, because the data structures for the GP-GPUs are stored in a HBM2 memory pool, which is distinct from the data structures for the CPUs and their DDR4 memory pool. This leads to the need for higher message-frequency requirements for the system interconnection network fabric. The large memory capacity on each Summit heterogeneous compute node also leads to the need to support larger messages. To hide current latencies of a microsecond caused by message aggregation, a node-level architecture blending overlapping functionalities of communication and



(a)



(b)

Figure 3.14: (a) ORNL Summit Compute Node node-level architecture to hide latencies (b) Dragonfly Network Topology systems-level architecture to keep network diameter low (courtesy of Kevin Barker, Pacific Northwest National Lab<sup>32</sup>)

computation would try to increase message frequency. This may actually increase per-node memory, which could lead to larger messages, as in the schematic in Figure 3.14a. *Another interesting challenge from a systems perspective is the desire to keep the network diameter low.* As Figure 3.14b shows, a “Dragonfly” topology may be the best option at present. This must also employ high-radix switches and maintain link bandwidth to avoid impairing multi-hop communications. This wired network currently makes for high density communication between nodes.

Looking forward to the next generation, there is a need for high bandwidth in the range of **400 Gb/s/link**.

Longer term, the priority research directions for HPC Interconnection fabrics include:

- Tighter integration between network, processing, and memory, in order to reduce latency, increase BW, and reduce energy
- Reconciliation of intra-node and inter-node network disparity to answer the question of being able to have a single network both inside and outside the network
- Technologies to develop richer network topologies (high-radix, high BW) that will be more robust to task/job placement and network traffic patterns

### Single and multi-mode free-space optical communication links

Free-space optical (FSO) communication has the advantage of large bandwidth, unregulated spectrum, and high directionality<sup>34</sup>. However, it is encumbered with line-of-sight

requirements and link outage from environmental interference. Classical pointing and tracking techniques and adaptive optics have helped ameliorate some of these problems.

*An exciting novelty in the non-line-of-sight propagation (NLoS) communication paradigm is Light Fidelity (Li-Fi), illustrated by Figure 3.15.* Its mode of operation of transmission and handoff is similar to Wi-Fi. It has the advantage of operation in areas susceptible to electromagnetic interference and achieves speeds of up to 10 Gbit/s using RGB LED at 1.5 m. Further development lies in more efficient optical TDMA / CDMA scheme realizations and better spatial frequency reuse methods.

Wavelength Division Multiplexing (WDM) has been making greater strides. In essence, WDM multiplexes multiple carrier signals (lasers of different wavelength) onto an optical fiber and enables bidirectional communications on this channel.

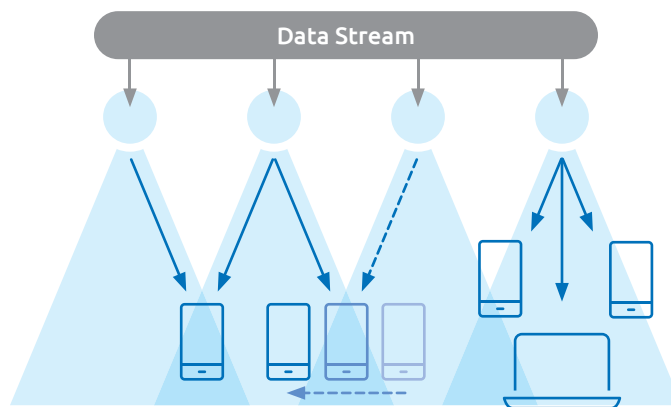


Figure 3.15: Simplified scheme for multiple access and handoff using Li-Fi<sup>34</sup>

Larger data capacity can be achieved by increasing the number of wavelengths with the use of Erbium-doped Fiber Amplifier (EDFA).

*Optical Angular Momentum (OAM) beams are also becoming popular for multiplexing.* Unlike plane waves that propagate with constant phase along the wavefront, OAM propagates twisting phase-fronts corresponding to different modes. As these modes are strictly orthogonal, data bits can be effectively twisted clockwise or anticlockwise and multiplexed or demultiplexed over a fiber or free space. This concept shows promise in application of back-haul, datacenter and building-to-building communications, as the single aperture pair structure used eliminates the need for DSP. The caveat is that it requires careful beam structuring and link design.

Other optical methods are also being pursued, including satellite-to-ground quantum key distribution<sup>35</sup> and laser communication for data relay services or optical backbone network (by the European Data Relay System).

### Key areas of focus and follow-on research

- Co-integrated CMOS and optics solutions for next-generation, affordable integrated photonics

- Very low-cost, low-power optical transceivers for line rates 100 Gb/s or even 1 Tb/s; more affordable fronthaul allowing for more centralized Flexible signal processing and higher density of small cells
- Low power consumption (mainly for thermal management reasons) for links in datacenters and HPC (IM/DD links today are more energy efficient and lower cost than coherent single mode (SMF)-based solutions but will have problems to go beyond 0.1 km and 100-200 Gb/s per fiber.)
- Spatial division multiplexing (SDM) (Parallel fiber links may have an advantage in ultra-long haul (submarine) over spatial multiplexing solutions.)
- New wavelengths and new amplifiers critical for metro-long-haul links (>100 km)
- High-capacity (> 1Tb/s /wavelength) links involving development of new optoelectronics with high bandwidths, new ADC/DACS for high-throughput DSP (ASIC-based) solutions, and algorithms to fit those
- Tighter integration between network, processing, and memory, in order to reduce latency, increase BW, and reduce energy
- Technologies to develop richer network topologies (high-radix, high BW) that will be more robust to task/job placement and network traffic patterns

## 3.5. IoT and M2M Communication Concepts

### Overview and needs

*The current human-centric communication is evolving into machine-centric communication.* This evolution is seeing exponential growth, which will be manifested by a dramatic increase in the number of Internet-of-Things (IoT) and Machine-to-Machine (M2M) communicating devices. While many technologies are characterized by the specific nature of applications or architectures, a wide variety of devices, networks, and applications defines IoT. This rich diversity consists of a variety of applications, from health- and wellness-related wearable devices to autonomous and complex industrial control systems, and from a localized network of interconnected machines to a large slew of sensors connected over wireless. IoT applications and devices will be central to advances in computation (AI-ML) and communication technologies (5G/6G) in the years and decades ahead, both in terms of driving requirements for them and enabling faster adoption. A number of challenges exist in efficient development and adoption of IoT systems in

terms of device development, coordination of communication protocols, reliable and high bandwidth networks, and security. This section highlights these challenges and sheds some light on approaches to address them.

Exponential population growth, a rise in urbanization, the need and demand for transportation and connectivity—all powered by limited natural resources—are placing a lot of stress on our infrastructure. Technology will enable us to optimize our resources and make us more efficient, safer, and healthier. *A new era is approaching with distributed intelligence at the edge.* Rather than being connected to a single device, all devices are untethered, giving them wide coverage and making them reconfigurable and flexible. This enables new applications like personalized electronics, wellness/health, manufacturing, transport, retail, and finance. These characteristics are also essential for truly autonomous vehicles and robots. Data analytics, inference, and modelling will drive large amounts of data transfer, and next generation connectivity will unlock the full potential of artificial

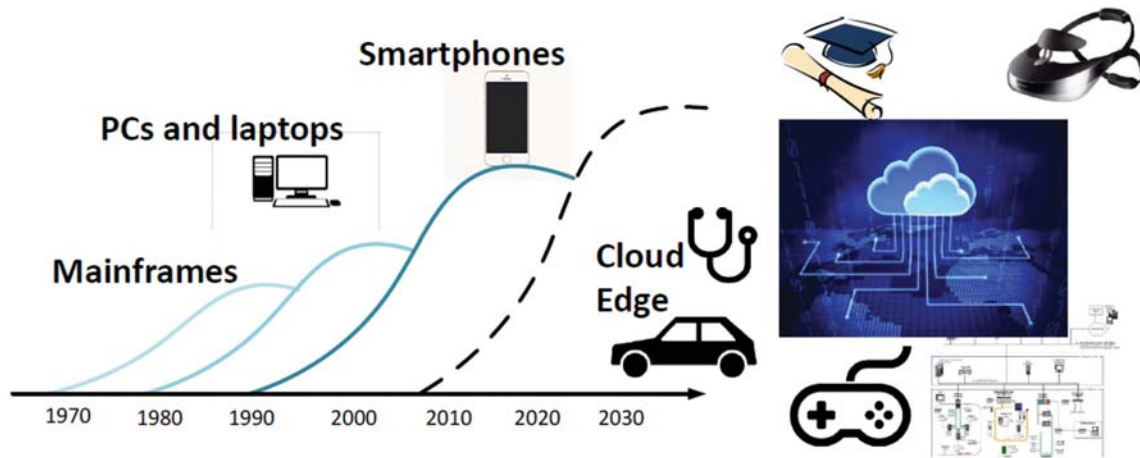


Figure 3.16: Application drivers (courtesy of Ajith Amerasekera, UC Berkeley<sup>36</sup>)

intelligence and machine learning. **Driver applications are diversifying, and custom silicon is needed for optimized performance across a diverse set of applications with various demands.** This calls for new design techniques, EDA methodologies, and reuse. Semiconductors are being used in applications that they have never been used before, giving rise to new challenges across the industry.

The current focus is on enhancing mobile, broadband, and fixed wireless access, with defining network characteristics that include higher data rate, smaller cells, new physical layer, and so on. The most important characteristics that enable **IoT include low power and latency control.** There is greater value when we include expanded application opportunities. Some applications will be addressed in 5G+ and the rest will drive new connectivity techniques<sup>36</sup>.

Mobile connectivity is increasingly becoming important and has multiple application drivers and user needs. Today's capabilities are up to 100 Mbps, which is good enough for driver augmentation. Typically, 5G is now used by consumers, smart cities, and in industry verticals. **Some of the performance targets are coverage (expected to be 20dB better than smartphones), ultra-low device cost, a battery life greater than 10 years, very high connection density (~1M connected devices/km<sup>2</sup>), reliable network, and bounded low latency.** There is also commercial growth in massive IoT devices like wearables and broadband IoT devices where high performance is critical. However, the major application drivers for innovation are critical IoT devices like those in autonomous vehicles, traffic control, and smart grids. Another category of devices is used in industrial automation, such as collaborative robotics, advanced automation, and control.

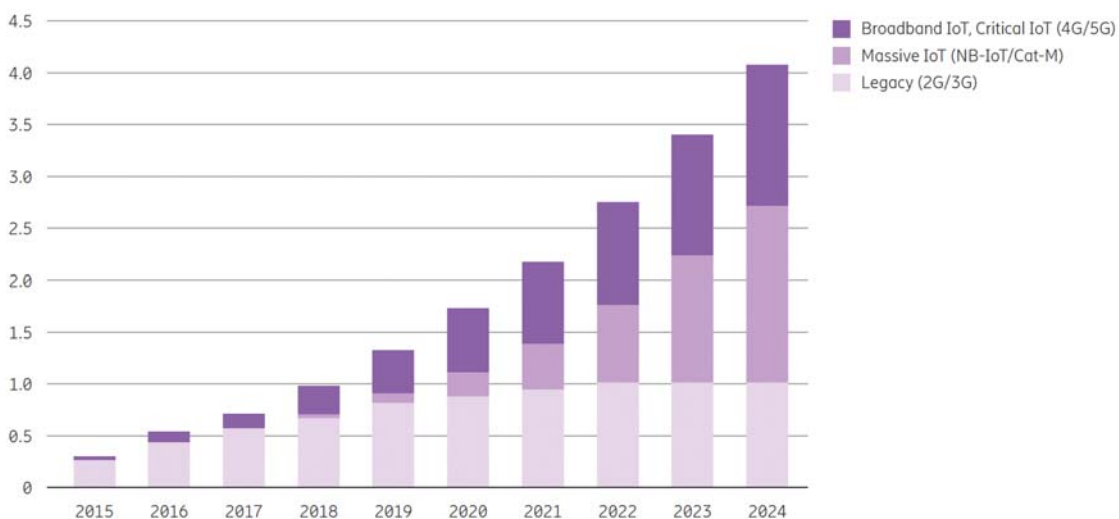


Figure 3.17: Cellular IoT connections by segment and technology (courtesy of Y.-P. Eric Wang (Ericsson)<sup>37</sup>)



The vision of Industry 4.0 is a factory where all devices and end products are fully connected using flexible and open standards. Cellular networks provide these capabilities and are the key enabler. The two major technologies that are being used are *LTE-machine-type communication* and *Narrowband IoT devices*. The future network migration involves spectrum re-farming and random-access network software upgrades. This requires narrowband built in forward compatibility and dynamic spectrum sharing. A few important considerations are small form factor, new frequency bands, power/energy management, device cost, network deployment and management, and reliability<sup>37</sup>.

## IoT for autonomous systems

Wireless networks will be the enabling technology of future autonomous systems that will be deployed in factories, rescue missions, patrolling, etc.<sup>38</sup> **These autonomous systems depend on communication between machines that are more sensitive than humans to delays and errors.** The traditional viewpoint is that ultra-high reliability and ultra-low latency are needed, but an alternative view would be to look at the *joint design of communications and control*<sup>39</sup>. The control loops are designed to tolerate errors, and the communication system takes advantage of this tolerance to improve resource allocation. However, an autonomous system operating in a non-autonomous wireless network is not truly autonomous. The network itself must be autonomous by learning to communicate and learning to allocate wireless resources. Such allocation of resources is a complex optimization problem.

The optimization challenge can be circumvented by learning optimal resource allocation using machine learning models to generate simulated data<sup>40</sup>. There is some success in using fully connected neural networks in small-scale systems, but *graph neural networks (GNNs) are the only solution that scales to large networks*<sup>41</sup>. *GNNs outperform known model-based heuristics and transfer within graph families*. They can be trained offline but executed online on a different network, able to transfer to large graphs without retraining. The combination of joint design and learning techniques results in integrated wireless autonomous systems where the wireless network is itself autonomous, and the distinction between the autonomous system and the wireless communication network gets blurred.

In the context of industrial control, wireless autonomous systems take the form of several plants that share communication links, while an attempt is made to close as many control loops as possible within a target delay<sup>42</sup>. A

joint design approach allows adaptation of communication reliability to the state of the plant. Plants that are close to instability get more bandwidth to achieve more reliable communication. A learning approach allows for reduced deployment costs and adaptability to varying plant configurations. In the context of communication infrastructure, wireless autonomous systems enable the deployment of *mobile infrastructure on demand (MID), either to augment capacity in urban setting or provide coverage in rural settings where it may be unavailable*<sup>43</sup>. *Autonomous joint design of MID allows for joint reconfiguration of communication routes and physical positions. Autonomous learning of MID allows provision of rate guarantees with high probability, despite meager knowledge of channel models, as well as adaptation to avoid the typical rate dips of fixed infrastructure.*

As autonomous vehicles become more prevalent, we need better communication technology to handle the massive amounts of data that will be transferred between vehicles and over the network. A good candidate for this is mmWave because it has a large spectrum that can be used by multiple devices. In addition, it works well with line-of-sight communication. The vision is to have a combination of sensing, learning, communication, and *multiband connectivity that supports V2X*. The vehicles exchange sensor data with the sensing built into the infrastructure. It should also enable high-data-rate infotainment applications and joint communication and radar. Both beam training and conventional channel estimation/tracking have high overhead with large arrays. The solutions are to exploit channel structure by making them as sparse as possible, exploiting channel statistics and spatial consistency, and using out-of-band information. We can also leverage mmWave base stations, which may aid both automation and communication. For example, it provides augmented sensing for vehicles, safety for vulnerable road users, and data for the government. This builds into a smart city infrastructure for the future<sup>44</sup>.

## IoT security considerations

As the world becomes more connected, security is a significant concern. The number of attack points increases, and it is necessary to create new security measures and standards. It is expected that by 2022, the number of connected IoT devices will reach \$29 billion, with a market value of \$1.2 trillion. Different companies develop their IoT platforms for third-party developers to build apps to realize service and provision automation. Security and privacy measures are not keeping pace with rapid IoT growth. The

Mirai botnet attacks, which exploited a vulnerability in IoT devices that could have been easily prevented, compromised multiple devices and brought down a DNS provider. Basically, it caused an internet outage. The most critical security and privacy threats come from IoT platforms and their affiliated applications. Individually, IoT devices may be secured, but they open up vulnerabilities while connected as a system for various applications. *Incremental expansion of the systems or the applications expose new vulnerabilities. Peer interactions between connected devices also add to vulnerabilities. Unintentional insider threats rise with the scaled versions of IoTs, and unconventional usages of IoT devices will also introduce new vulnerabilities.*

Many vulnerabilities can be resolved by implementing good authentication practices like individual authentication, though it suffers from scalability issues. The same is true of secure onboarding. Proximity-based authentication might prove effective, but it is susceptible to side-channel attacks. Another solution might be continuous authentication. *Network-based detection can be used to detect suspicious behavior and incorporate user intention to improve accuracy.* The main advantage of such a system is low overhead at host, as it is easy to deploy. Plus, a large number of devices can be monitored without introducing overhead and signatures not revealed by system-level approaches. Most suspicious traffic is transmitted with simple unencrypted HTTP requests, and the number of malware families is not huge. Variants of the same malware exhibit similar behaviors. The challenges lie in the components' interaction, which leads to conflict, repetition, and unforeseeable outcomes. There are many attacks that could happen at device, network, or application level, and these attacks escalate due to chains of interaction. To mitigate these attacks, a multi-layer IoT hypothesis graph must be created that characterizes system states at each layer and attacks (Figure 3.18)<sup>45</sup>.

### IoT network challenges

Humans are increasingly becoming connected to the internet in the form of human computing, wearable computing, and symbiotic computing. For example, a mesh network of hubs forms robust communication skin around body, and hubs communicate locally with energy-frugal sensor nodes. The physical layer for last hop is optimized for location and channel, and the protocol stack is adaptively tuned to adjust to changing conditions and optimize robustness and energy consumption. There is a push to solve interference problems

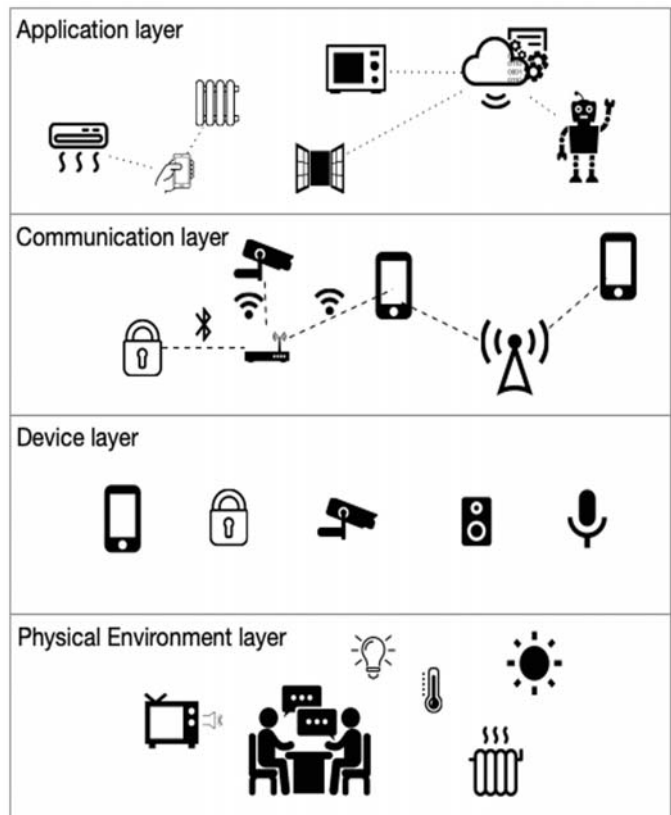


Figure 3.18: Multi-layer system graph (courtesy of Prasant Mohapatra, UC Davis<sup>45</sup>)

caused by high user density with many different types of connectivity. Today it is addressed by a control of spectrum, techniques for avoidance, and filtering. This will need to be adjusted by *active interference cancellation, learning collaboration, etc.* Spectrum congestion is another issue that needs to be tackled by developing intelligent radio networks that collaborate to manage and optimize the spectrum in a heterogenous environment, and to provide autonomous and spontaneous collaborative intelligent radio networks that are driven by learning. These use techniques that include human-robot interaction like game-theory, in combination with machine learning.

Network reliability can also be improved by using relays that will faithfully transmit the data by overcoming fading with spatial diversity. Wired networks will also be of huge importance, with the use of more copper and the push to use optical fiber networks with high data rates. Some key technology needs are in signal processing, where dynamic and static techniques for interference mitigation, management, and cancellation are needed and to produce low SNR signals.

Mesh networks require channel prediction and modeling in order to use machine learning and adaptive techniques to make the system more efficient. System design will require a power optimization perspective. Semiconductors will also undergo a major shift, requiring customized compute for machine learning, control and signal processing, and data converters for high speed with low SNR requirements. They will also play a role in high-bandwidth modulation in free space and plastic waveguides. Reconfigurability and reusability will be major characteristics to look for. **Silicon photonics for new architectures and scaling to 1000s of antennas will be necessary, in addition to new materials to filter signals above 10GHz. Also needed are new packaging methods for above 100GHz signals like that in wide bandwidth configurable antennas**<sup>36</sup>.

There is a push to understand how to make a “world without wires,” action that requires high bandwidth, highly reliable and ultra-low cost, and power for IoT. The projected IoT growth will stress the infrastructure, and multiple classes of IoT devices have very different requirements. **Ad-hoc/**

**heterogenous connectivity is needed for local networks, in addition to self-organizing networks.** The information age needs to be brought to the physical world, and system-level thinking is necessary. Future IoT may utilize LEO satellite communication, and better energy harvesting research will benefit smart phones and many other devices. New ways are needed for network and devices to meet so many different requirements for IoT device classes and cellular devices.

### Key areas of focus and follow-on research

- High-bandwidth, high-reliability solutions for autonomous systems
- System solutions for classes of IoTs: Massive IoT, Broadband IoT, Critical IoT, Industry IoT Capacity, High BW, reliability (autonomy), and low latency (autonomy)
- Security topics for IoT, including authentication of individual devices, proximity-based authentication, use of side channels for proximity estimation, and secure onboarding (scalability issues)
- Role of IoT in future wireless systems beyond 5G

## 3.6. Communication ICs in the Nanoscale Era

### Overview and needs

Over the past half century, communications has benefitted from a repeated cycle of a confluence of market needs, technology capabilities, and industry organizational structure. Examples include:

- *Insatiable demand to communicate voice, video, and data from anywhere at any time for convenience, personal use, or as a tool for work;*
- *Moore’s Law dimensional scaling, which has provided increased levels of computation throughput, faster and more accurate signal processing, increasing frequency response, enhanced energy efficiency, and higher levels of integration at each successive node; and*
- *Standardization of communication protocols, enabling open competition and creating scale that has resulted in new features, greater performance, and lower cost.*

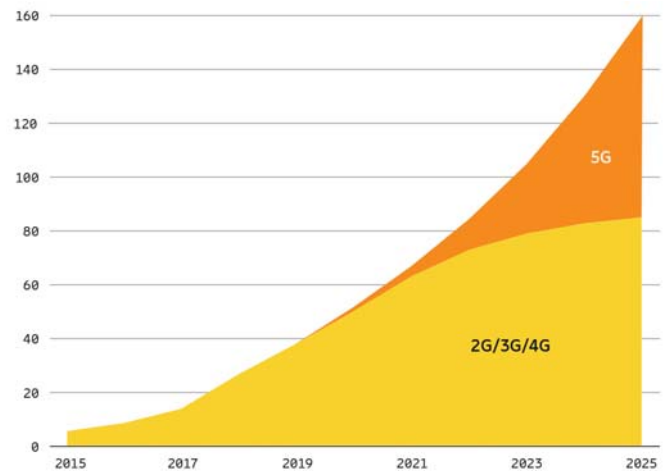


Figure 3.19: Source: Ericsson Mobility Report, Nov. 2019

While there are many communications systems such as Zigbee, UWB, NFC, Bluetooth, WLAN, and other ad hoc IoT protocols, the focus here is on cellular 5G and 6G. Both base stations and handsets are considered, with the emphasis on RF and associated supporting computation. Each cellular generation has been introduced roughly at the beginning of each decade and has generally provided increased data rate

(see Figures 3.19 and 3.20), improved spectral efficiency, and enhanced connectivity. Bringing us a step closer to ubiquitous coverage, these enhancements have driven higher frequency, wider BW, and improved linearity requirements for RF stages. Furthermore, the trend continues in complexity, as shown by the growing number of RF bands required to support cellular systems around the world (see Figure 3.21).

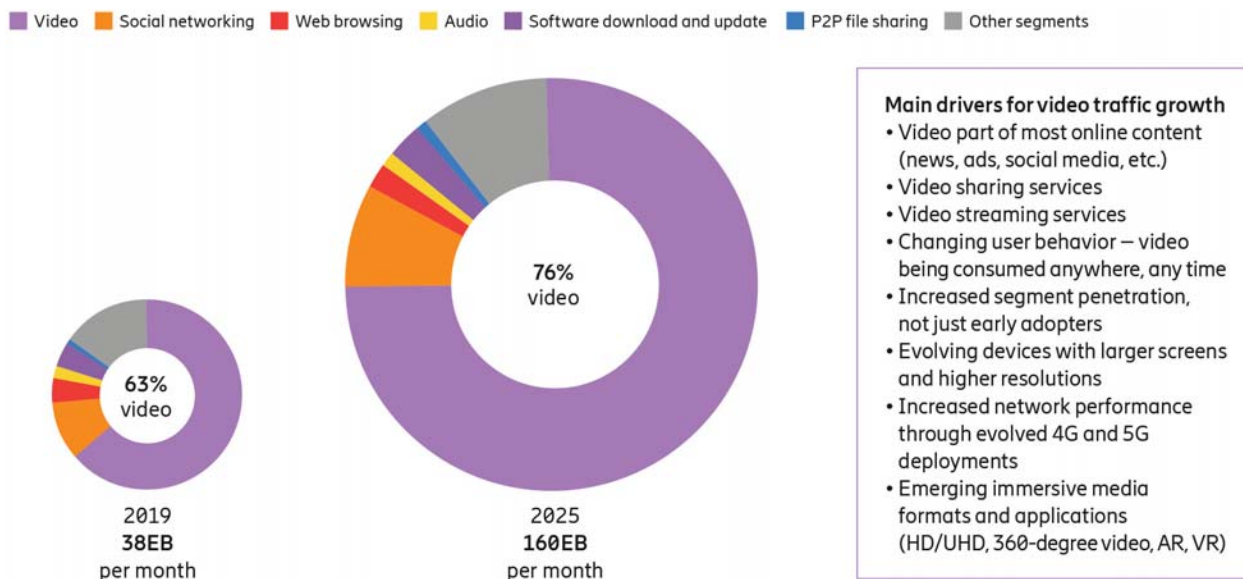


Figure 3.20. Source: Ericsson Mobility Report, Nov. 2019

|             | <1GHz             | 3GHz                 | 4GHz                                 | 5GHz       | 24-28GHz                                      | 37-40GHz                             | 57-71GHz           | >71GHz                     |
|-------------|-------------------|----------------------|--------------------------------------|------------|---|--------------------------------------|--------------------|----------------------------|
| USA         | 600MHz (2x35MHz)  | 2.5/2.6GHz (B41/n41) | 3.45-3.55GHz, 3.7GHz, 4.2GHz         | 5.9-7.1GHz | 24.25-24.45GHz, 24.75-25.25GHz, 27.5-28.35GHz | 37-37.6GHz, 37.6-40GHz, 47.2-48.2GHz | 57-64GHz, 64-71GHz | 71-76GHz, 81-86GHz, >95GHz |
| Canada      | 600MHz (2x35MHz)  |                      | 3.55-3.7 GHz                         |            | 26.5-27.5GHz, 27.5-28.35GHz                   | 37-37.6GHz, 37.6-40GHz               | 57-64GHz, 64-71GHz | 71-76GHz, 81-86GHz         |
| EU          | 700MHz (2x30 MHz) |                      | 3.4-3.8GHz                           | 5.9-6.4GHz | 24.5-27.5GHz                                  |                                      | 57-66GHz           | 71-76GHz, 81-86GHz         |
| UK          | 700MHz (2x30 MHz) |                      | 3.4-3.8GHz                           |            | 26GHz   |                                      |                    |                            |
| Germany     | 700MHz (2x30 MHz) |                      | 3.4-3.8GHz                           |            | 26GHz   |                                      |                    |                            |
| France      | 700MHz (2x30 MHz) |                      | 3.46-3.8GHz                          |            | 26GHz   |                                      |                    |                            |
| Italy       | 700MHz (2x30 MHz) |                      | 3.6-3.8GHz                           |            | 26.5-27.5GHz                                  |                                      |                    |                            |
| China       | 700MHz            | 2.5/2.6GHz (B41/n41) | 3.3-3.6GHz                           | 4.8-5GHz   | 24.75-27.5GHz                                 | 37-42.5GHz                           | 59-64GHz           |                            |
| South Korea | 700/800MHz        | 2.3-2.39GHz          | 3.4-3.42GHz, 3.42-3.7GHz, 3.7-4.0GHz | 5.9-7.1GHz | 25.7-26.5GHz, 26.5-28.9GHz, 28.9-29.5GHz      | 37.5-38.7GHz                         | 57-66GHz           | 71-76GHz, 81-86GHz         |
| Japan       |                   |                      | 3.6-4.1GHz                           | 4.5-4.9GHz | 26.6-27GHz, 27-29.5GHz                        | 39-43.5GHz                           | 57-66GHz           | 71-76GHz, 81-86GHz         |
| India       | 700MHz            |                      | 3.3-3.6GHz                           |            | 24.25-27.5GHz, 27.5-29.5GHz                   | 37-43.5GHz                           |                    |                            |
| Australia   |                   |                      | 3.4-3.7GHz                           |            | 24.25-27.5GHz                                 | 39GHz                                | 57-66GHz           |                            |

**Legend:**  
 - Existing band: Blue line  
 - Unlicensed/shared: Red line  
 - Licensed: Green line  
 - New 5G band: Yellow box

Figure 3.21. Global snapshot of allocated/targeted 5G spectrum (courtesy Jeremy Dunworth, Qualcomm<sup>46</sup>)



## State-of-the-art cellular communications

While 5G has successfully launched, the mmWave (24GHz and above) bands are largely unused and require further development. As part of Release 17 of 3GPP, it is supporting bands up to 114GHz. Additionally, unlike previous generations of cellular, the technology underpinnings for 6G are undefined. **One key technical challenge proposed/suggested for 6G is to extend beyond 100Gbps to 1Tbps, which places substantial demands on the baseband processing, RF devices, and the overall tradeoffs between communication-system planning and link budget.** Although there is still much to be accomplished below 6GHz, this report will focus on mmWave, where substantial research is needed. Some of the key challenges over the next decade will be: advancing mmWave communications system architecture and cell planning; optimizing integrated RF and antenna designs; enhancing active devices for efficiency and gain; and new materials for innovating passive components, such as tunable filters and circulators for adequate performance at these frequencies and reconfigurable antennas.

Understanding RF propagation and modelling of the channel in the mmWave frequency band is a critical element for system planning and communication component/subsystem specification setting. Data collected by Nokia (see Figure 3.22) shows that, with a 28GHz carrier, a path loss of approximately 30dB @250m that needs to be accounted for. Further, outdoor-to-indoor path loss is on the order of 25-40dB.

Extending the frequency above 28GHz creates even greater challenges due to higher path losses. Link demonstrations have been conducted at SRC's ComSenTer (a collaboration of 12 global commercial semiconductor companies and DARPA) utilizing chips capable of operating at 120GHz and delivering 80 Gbps at 10 cm<sup>47</sup>. *Further, the research team is targeting a module consisting of an array of elements both on the Rx and Tx side. This is capable of communicating over 10m, supporting 8 users at 100 Gbps, and leveraging two polarizations, resulting in a total throughput of 1.6 Tbps.* Even with this outstanding performance, massive densification of the base station will be required, creating a business-case dilemma. It is likely that the only viable solution is to have the customer share in the cost

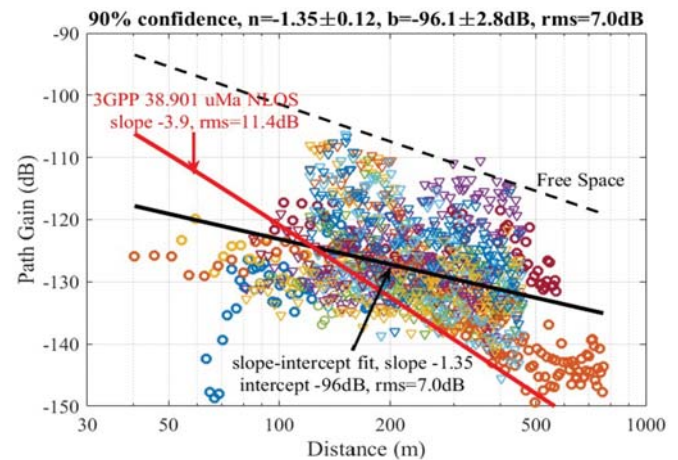


Figure 3.22: Source: Nokia 2020

of the infrastructure, much as is done today in the customer purchase of a WiFi Access Point.

Several key semiconductor device elements for mmWave operation include: a high-frequency technology; low-loss BEOL for optimized on-chip RF passive components and **advanced modelling and simulation capability** to account for layout-dependent effects; and transmission-line-based devices and cross-coupling. It is generally accepted that  $1/3$  of  $f_{\text{max}}, f_t$  is an upper frequency bound of operation that can tolerate process, voltage, and temperature variations, while still maintaining adequate gain. Scaling CMOS devices beyond 20nm no longer improves these devices, as other limits due to prevailing gate and interconnect resistance result in a peak performance for  $f_{\text{max}}$  of approximately 450GHz<sup>48</sup>. For small signal RF operation, CMOS circuits have been demonstrated to support **mmWave operation into the 100GHz range**<sup>49</sup>. For silicon technologies, *Si-Ge offers higher frequency performance, higher breakdown voltage, and higher power out compared to CMOS, particularly at high temperatures. IHP has demonstrated a 700 GHz  $f_{\text{max}}$  DOT7 device<sup>50</sup> that extends performance well above CMOS. Of all the PA devices, GaN stands out as offering the highest Tx power, highest PAE (see Figure 3.23), widest BW, greatest power density, and highest reliability<sup>51</sup>.* The primary challenge with these devices is a degradation in EVM due to the temperature-dependent, time-constant differences between the fill and release of

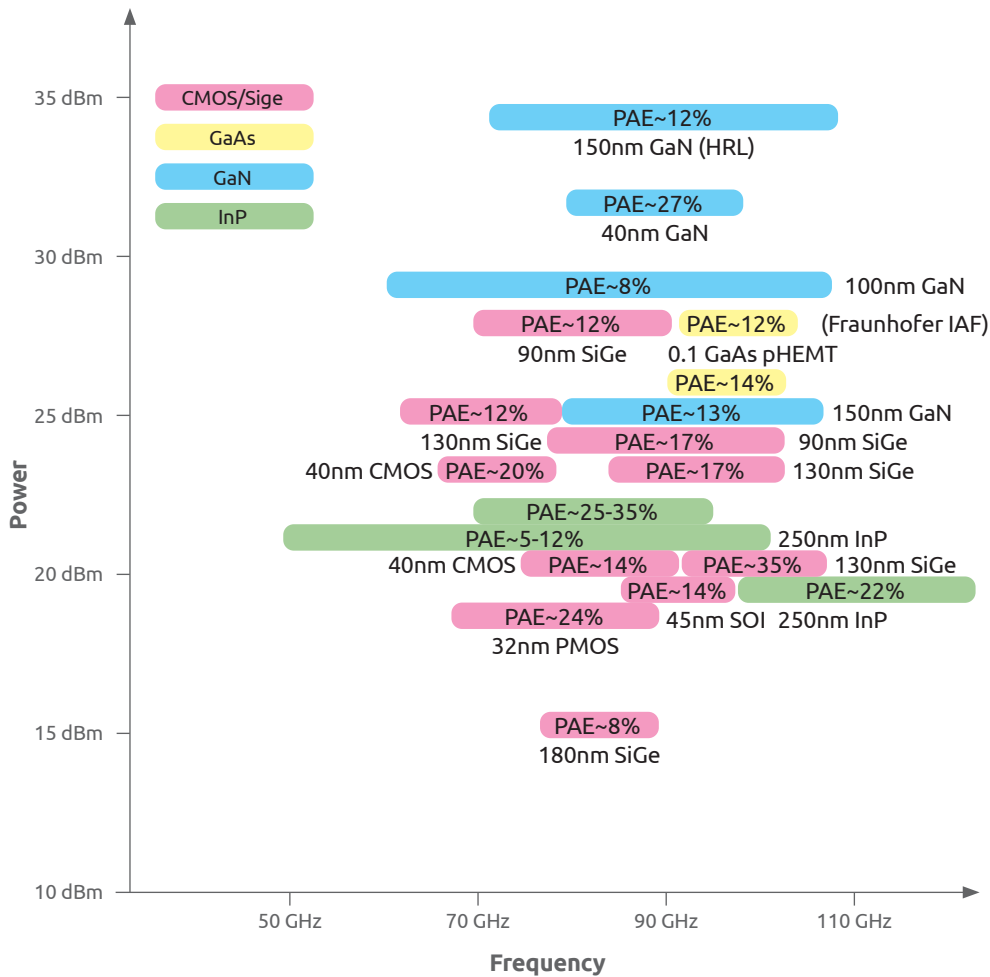


Figure 3.23: A comparison Power Added Efficiency of several technologies operating at mmWave frequencies

charge in traps caused by a lattice mismatch. Many process improvements have been explored in an attempt to reduce the lattice mismatch at the GaN/substrate interface, but a solution has yet to be provided. Moreover, the devices have not been shown to operate above 120GHz.

Building a highly efficient transmitter, particularly in the mmWave band, goes well beyond the device. It requires an optimization of the device, circuit, topology, architecture, and integration/signal processing. Well-defined models are fundamental to making these trade-offs. Digital Predistortion (DPD), Crest Factor Reduction (CFR), or Envelope Tracking (ET) remain challenges at mmWave for the handset. A scalable four-way asymmetric G conjugate matched reconfigurable transmitter design has been shown operating across the range of 37-73 GHz, with power out varying between 16.3 and 19.3dBm, with a peak drain efficiency of 40% that degrades (worst case) by 50% under backoff conditions up to 10 dB<sup>52</sup>.

Both Tx and Rx blocks must be translated to or from the digital domain through a data converter. **Wider channel BWs and increased linearity will place substantial demands on these blocks that already are power hungry.**

Much like the radio was moved to the antenna to create a Remote Radio Head (RRH) to optimize power consumption of a base station, integrated antennas and antenna modules are being co-designed with the RFIC for the handset to minimize losses and optimize system performance in the handset at the mmWave frequencies<sup>46</sup>. Taking into account the system trade-offs, design constraints, and the need for increased levels of integration, many partitioning decisions need to be made. **The challenge is to determine what is monolithically integrated versus “package” or heterogeneously integrated for a complete RFFE module for each of the handset and the base station<sup>53</sup>.**

## Key challenges for communication ICs

Looking toward 6G, the challenges are many. As thought of today, mmWave covers at least two octaves—that's a lot to handle. The increasing demand for high-speed data pushes operating frequency higher. At the same time, propagation losses increase, requiring a densification unheard of today and challenging traditional business-case models. We must continue to search for new non-traditional paths forward. For example, multipath used to be considered a problem, but it's now a solution when utilized constructively with computationally powerful engines. Phased arrays are becoming practical to implement and are providing a means for capturing more information. Applications may emerge, driving IC requirements in the future. On the other hand, there is need to advance enabling technologies. A balance must be struck within the coming decade.

In order to take advantage of mmWave technology, an organized and ambitious direction must be set. **Key challenges identified include:**

- Improve channel models to accuracies comparable to those below 1GHz, as link performance is critical to system planning
- Data converters for infrastructure ZIF receivers supporting a multi-band 600 MHz half BW @ 50 mW
- Si-Ge based mmWave transmitters operating beyond 1 W/mm<sup>2</sup>
- Handset implementable DPD, CFR, or ET for 28GHz and above, 400 MHz BW 5G NR
- For GaN devices, trap algorithmic linearization techniques to offset the impact due to dislocations at the GaN/substrate interface, eliminating their dominance on Error Vector Magnitude (EVM) specifications—reduce from 2.5-3% to below 1%

**GaN devices and PAs extended to 200 GHz**, operating at 40dBm @ 30% PAE

- Reconfigurable mmWave transmitters 20dBm/40% PAE
- Programmable PA matching networks
- Multiband beamforming

## Key areas of focus and follow-on research

Broad technology areas have been identified to address challenges outlined in the previous section. Proposals should support achievement of the Targets defined above. While not an exclusive list, some proposal topics are:

- New approaches and improved devices for mmWave regime (Circuits now operate much closer to  $f_{\max}$ ,  $f_c$ , and many circuit techniques available to RFIC/RFFE designers in the sub-GHz regime do not scale to >24GHz.)
- Efficient digital beamforming approaches to improve network capacity
- Enhancements in mmWave channel models
- Approaches to optimize SiP-based package approaches at mmWave
  - EDA tools to include high frequency and thermal effects and high speed interconnects between devices
  - Test methods to identify Known Good Die (KGD) in the mmWave regime
  - Test and repair strategies
- ML techniques applied to transmitters to improve linearity and offer simultaneous spectral agility and back-off efficiency
- New materials and passive devices in the mmWave band for tunable filters, circulators, etc.
- Reconfigurable transmitters
- Re-examination of underlying physics of communications

## 3.7. Summary— New Trajectories for Communication

Radical advances in communication will be required to address growing demand. For example, the cloud technologies may undergo substantial changes, with emphasis shifting toward edge computing and local data storage. Broadband communications will expand beyond smart phones to immersive augmented reality, virtual meetings, and smart office settings. New capabilities will enrich user experiences through new use cases and new vertical markets. This requires collaborative research spanning a broad agenda, aiming at establishing revolutionary paradigms to support future high-capacity, energy-efficient communication for the vast range of future applications.

### The communication grand goals

- **Advance communication technologies to enable moving around all stored data of 100-1000 zettabyte/year at the peak rate of 1Tbps@<0.1nJ/bit**
- **Develop intelligent and agile networks that effectively utilize bandwidth to maximize network capacity**

## Research recommendations summary

- Re-examination of underlying physics of communications
  - Explore parallelism between communication networks and neural networks; use multi-objective metrics to balance power and quality of service; determine radio resource enhancements
  - Learn from biology to shape future communication systems; find inspiration from immunity systems
  - Merge innovations from quantum technologies into communication networks
- mmWave and beyond technologies—going from hype to universal deployment
  - Massive MIMO systems with 1000 of antennas
  - New materials and passive devices in the mmWave band for tunable filters, circulators, etc.
  - mmWave applications with massive data rate support, ultra-low latency; overcome issues of blockage and power consumption
  - Consider systems with high bandwidth, utilizing large arrays in THz range; focus on changing aspects of PHY; establish wideband models for arrays; extend operations for THz, including new models for signal processing, new RF architectures, and new devices
  - Technology innovation to enhance transmit power, especially in the untapped spectrum (100GHz-1THz)
  - New approaches and improved devices for mmWave regime (Circuits now operate much closer to  $f_{\max}$ ,  $f_t$ , and many circuit techniques available to RFIC/RFFE designers in the sub-GHz regime do not scale to >24GHz.)
  - Efficient digital beamforming approaches to improve network capacity
  - Enhancements in mmWave channel models
  - Approaches to optimize SiP-based package approaches at mmWave
  - EDA tools to include high frequency and thermal effects; high speed interconnects between devices
  - Test methods to identify Known Good Die (KGD) in the mmWave regime
- Photonics
  - Innovative semiconductor process platforms to include RFSOI, FinFET, SOI/SiGe-based photonics
  - Co-integrated CMOS and optics solutions for next generation, affordable integrated photonics
- Very-low-cost, low-power optical transceivers for line rates 100 Gb/s or even 1 Tb/s (More affordable fronthaul allows for more centralized Flexible signal processing and higher density of small cells.)
- Low-power consumption (mainly for thermal management reasons) for links in datacenters and HPC (IM/DD links today are more energy efficient and lower cost than coherent single mode (SMF)-based solutions, but will have problems to go beyond 0.1 km and 100-200 Gb/s per fiber.)
- Spatial division multiplexing (SDM) (Parallel fiber links may have an advantage in ultra-long haul (submarine) over spatial multiplexing solutions.)
- New wavelengths and new amplifier critical for metro-long-haul links (>100 km)
- High-capacity (> 1Tb/s /wavelength) links involving development of new optoelectronics with high bandwidths; new ADC/DACS for high-throughput DSP (ASIC-based) solutions and algorithms to fit those
- Secure communications
  - Security for communication designed into the hardware
  - Security topics for IoT, including authentication of individual devices, proximity-based authentication, use of side channels for proximity estimation, and secure onboarding (scalability issues)
- System-level optimization from the network to the edge
  - New design principles for wireless networks, from data insight to action; security as key component of design; new performance metrics; expanded role of ad-hoc and mesh networks
  - Technologies to develop richer network topologies (high-radix, high BW) that will be more robust to task/job placement and network traffic patterns
  - High-bandwidth, high-reliability solutions for autonomous systems
  - Impact of network densification; solutions for backhaul
  - Reconfigurable transmitters
  - System solutions for classes of IoTs, including Massive IoT, Broadband IoT, Critical IoT, Industry IoT Capacity, High BW, reliability (autonomy), and low latency (autonomy)
  - Development of intelligent edge nodes with focus on always-on devices, high-bandwidth devices, and new modalities for security
  - Role of IoT in future wireless systems beyond 5G



# Appendix

## Global trends in communication

Global trends in communication are based on research by Hilbert and Lopez<sup>54</sup>, where a detailed inventory of all communication media<sup>iii</sup> was created (the communication inventory includes, among others, postal service and newspapers, radio and TV broadcast, fixed-line telephone and internet, mobile telephone and internet, GPS, etc.). Extrapolation of the communication inventory in<sup>54</sup> provides *conservative* projections for global communication trends (Figure A1).

Another scenario is based on the assumption of the prevalence of wireless communication with aggressive growth trend sustainable over long time. This model is compared to the Cisco data based on the past and projected communication capacity in 2012-2022 (orange dots) in Figure A2. At this point it is still uncertain which of the two scenarios will be realized in the next 20 years. In the current Decadal Plan, working documents the conservative trend is tentatively used.

Figure A1: Estimated and projected (conservative trend) global communication capacity in 2010-2050<sup>54</sup>

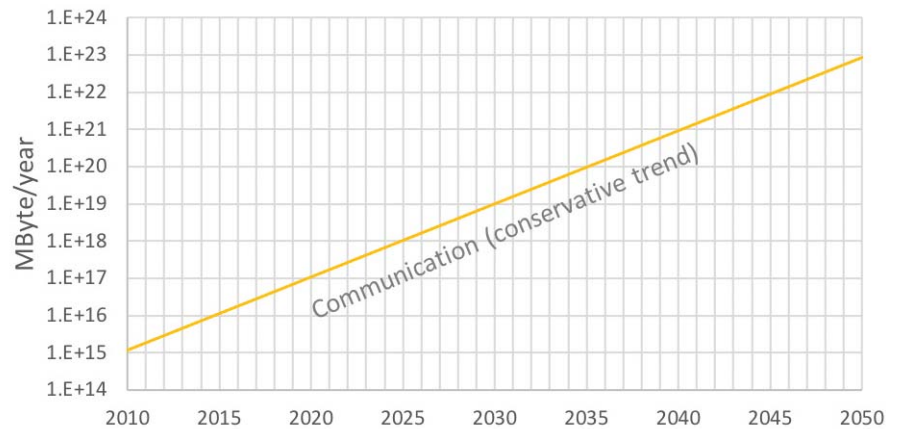
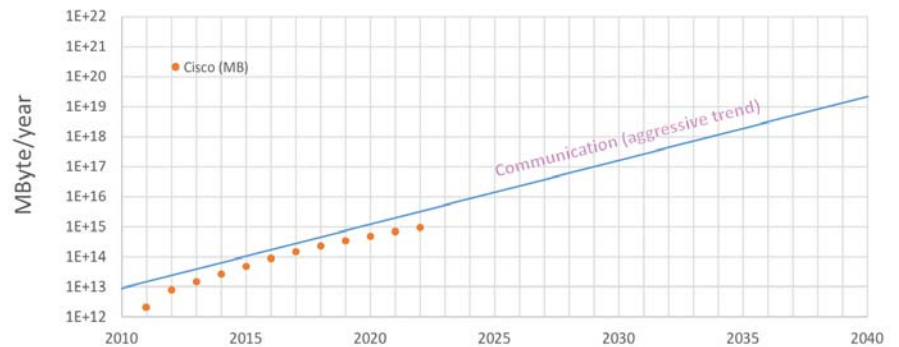


Figure A2: Estimated and projected (aggressive trend) global communication capacity in 2010-2050



## Contributors

Tamer Ali (MediaTek)  
 Ajith Amerasekera (UC Berkeley)  
 James Ang (Pacific Northwest National Lab)  
 Kevin Barker (Pacific Northwest National Lab)  
 Filbert Bartoli (NSF)  
 Bami Bastani (GLOBALFOUNDRIES)  
 Pete Beckman (Argonne National Lab)  
 Daniel Blumenthal (UC Santa Barbara)  
 Ramesh Chauhan (Qualcomm)  
 Jeremy Dunworth (Qualcomm)  
 Peter Gammel (GLOBALFOUNDRIES)  
 Nuria Gonzalez Prelcic (U Vigo)

Ken Hansen (SRC)  
 Robert Heath (UT Austin)  
 Tingfang Ji (Qualcomm)  
 Magnus Karlsson (Chalmers University)  
 Nikolaus Klemmer (Qorvo)  
 Jayesh Kotecha (NXP)  
 Ali Niknejad (UC Berkeley)  
 Rafic Makki (Mubadala)  
 Thomas Marzetta (NYU)  
 Prasant Mohapatra (UC Davis)  
 Tony Montalvo (Analog Devices)  
 Sundeep Rangan (NYU)

Jeffrey Reed (Virginia Tech)  
 Alejandro Ribeiro (U Pennsylvania)  
 Kaushik Sengupta (Princeton U)  
 John Smee (Qualcomm)  
 Peter Vetter (Nokia Bell Labs)  
 Y.-P. Eric Wang (Ericsson)  
 Alan Willner (USC)  
 May Wu (Intel)  
 Ariela Zeira (Intel)  
 Lixia Zhang (UCLA)  
 Zoran Zvonar (Analog Devices)

<sup>iii</sup>In the communication process, a medium is a channel or system of communication—the means by which information is transmitted between the sender and the receiver.

# References to Chapter 3

- <sup>1</sup>Bami Bastani, "The future of communication", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>2</sup>Sundeep Rangan, "Millimeter Wave: Challenges and new technologies", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>3</sup>Jeffrey H. Reed, "Towards energy efficient 5G systems", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>4</sup>Lixia Zhang, "Designing security into cyberspace", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>5</sup>Fil Bartoli, "Internet of the Future: Quantum and Secure?" SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>6</sup>Robert W. Heath. "Going towards 6G and beyond", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>7</sup>"5G empowering vertical industries", The 5G Infrastructure Public Private Partnership (5G PPP) White Paper, Feb. 2016
- <sup>8</sup>United State Frequency Allocation. National Telecommunications and Information Administration. <https://www.ntia.doc.gov/files/ntia/publications/2003-allochrt.pdf>.
- <sup>9</sup>C. Han and Y. Chen, "Propagation modeling for wireless communications in the terahertz band", IEEE Communications Magazine, 2018.
- <sup>10</sup>G. A. Siles, J. M. Riera and P. Garcia-del-Pino, "Atmospheric attenuation in wireless communication systems at millimeter and THz frequencies [Wireless Corner]", IEEE Antennas and Propagation Magazine, vol. 57, no. 1, pp. 48-61, Feb. 2015.
- <sup>11</sup>H. Khatibi and E. Afshari, "Towards efficient high power mmWave and terahertz sources in silicon: One decade of progress," 2017 IEEE 17th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF), Phoenix, AZ, 2017, pp. 4-8.
- <sup>12</sup>Peng Sun, P. Schniter, R. W. Heath, Jr., Zhongyong Wang, "Joint channel-estimation/decoding with frequency-selective channels and low-precision ADCs," Proc. of Asilomar, 2017.
- <sup>13</sup>Y. Wang, A. Klautau, M. Ribero, M. Narasimha, and R. W. Heath, Jr., "MmWave vehicular beam training with situational awareness by Machine Learning," Proc. of GLOBECOM Workshops, Dec. 2018
- <sup>14</sup>A. Klautau, P. Batista, N. González Prelcic, Yuyang Wang, and R. W. Heath, Jr., "5G MIMO data for Machine Learning: Application to beam-selection using Deep Learning", Proc. of the Information Theory and Applications, San Diego, California, February 11-16, 2018.
- <sup>15</sup>John Smee. "Wireless network densification", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>16</sup>Thomas L. Marzetta. "Research directions for antenna arrays", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>17</sup>Giovanni Interdonato et al. "Ubiquitous cell-free massive MIMO communications". EURASIP Journal on Wireless Communications and Networking (Aug. 2019)
- <sup>18</sup>Jayesh H. Kotecha. "Baseband computation challenges and directions for 5G/6G mmWave", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>19</sup>O. Teyeb, A. Muhammad, G. Mildh, E. Dahlman, F. Barac and B. Makki, "Integrated access backhauled networks," 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 2019, pp. 1-5.
- <sup>20</sup>Ariela Zeira and May Wu. "Rethinking wireless communication design principles in a Smart World: Transforming wireless signal processing". SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>21</sup>Peter Vetter. "Trends and challenges of the Fixed/Mobile convergence in the era of 5G and 6G", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>22</sup>M. De Ree, G. Mantas, A. Radwan, S. Mumtaz, J. Rodriguez and I. E. Otung, "Key management for beyond 5G mobile small cells: A survey", IEEE Access, vol. 7, pp. 59200-59236, 2019
- <sup>23</sup>S. Buzzi, C. D'Andrea, A. Zappone and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach", IEEE Trans. Wireless Communications, vol. 19, no. 2, pp. 1250-1264, Feb. 2020.
- <sup>24</sup>D. Camps-Mur et al., "5G-XHaul: A novel wireless-optical SDN transport network to support joint 5G backhaul and fronthaul services", IEEE Communications Magazine, vol. 57, no. 7, pp. 99-105, July 2019.
- <sup>25</sup>K. Schuh et al. "Single carrier 1.2 Tbit/s transmission over 300 km with PM-64 QAM at 100 GBaud", Optical Fiber Communication Conference Postdeadline Papers, OSA Technical Digest (online) (Optical Society of America, 2017), paper Th5B.5
- <sup>26</sup>Oleg V. Sinkin et al. "SDM for power-efficient undersea transmission", J. Lightwave Technol. 36, 361-371 (2018)
- <sup>27</sup>Magnus Karlsson. "Roadmap of optical communications" SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>28</sup>P. Lei et al., "Functional-link neural network based nonlinear equalizer", 2019 18th International Conference on Optical Communications and Networks (ICOON), Huangshan, China, 2019, pp. 1-3.

- <sup>29</sup>Daniel J. Blumenthal, "Low-energy frequency-stabilized coherent WDM Fiber Links", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>30</sup>D. J. Blumenthal, "Ultra-stable integrated lasers and low-cost, low-energy coherent data center interconnect", in OSA Advanced Photonics Congress (AP) 2019 (IPR, Networks, NOMA, SPPCom, PVLED), OSA Technical Digest (Optical Society of America, 2019), paper NeM4D.1.
- <sup>31</sup>Ludlow, A.D., et al., "Optical atomic clocks". *Reviews of Modern Physics*, 2015. 87(2): p. 637-701.
- <sup>32</sup>Kevin J. Barker. "High performance computing needs for fixed-line interconnection network fabrics", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>33</sup>Kogge, P. and Shalf, J. "Exascale computing trends: Adjusting to the new normal for computer architecture", *J. Computing in Science and Engineering*, vol. 15, Nov. 2013.
- <sup>34</sup>Alan E. Willner. "High-Capacity Free Space Optical Communication Links using Single and Multi-mode" SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>35</sup>S. Liao et al., "Satellite-to-ground quantum key distribution". *Nature* 549, p. 7670, 2017.
- <sup>36</sup>Ajith Amerasekera, "Distributed Edge Intelligence," SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>37</sup>Y.-P. Eric Wang, "Cellular IoT", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>38</sup>Alejandro Ribeiro, "Wireless Autonomous Systems", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>39</sup>K. Gatsis, A. Ribeiro and G..J Pappas, "Optimal Power Management in Wireless Control Systems", *IEEE Transactions on Automatic Control* 59 (2013), 1495 - 1510.
- <sup>40</sup>M. Eisen, C. Zhang, L.F.O. Chamon, D.D. Lee, and A. Ribeiro, "Learning optimal resource allocations in wireless systems", *IEEE Transactions on Signal Processing* 67 (2019), 2775-2790.
- <sup>41</sup>M Eisen and A Ribeiro, "Optimal wireless resource allocation with random edge graph neural networks", arXiv preprint arXiv:1909.01865.
- <sup>42</sup>M. Eisen, M.M. Rashid, K. Gatsis, D. Cavalcanti, N. Himayat and A. Ribeiro, "Control aware radio resource allocation in low latency wireless control systems". *IEEE Internet of Things Journal* 6 (2019), 7878-7890.
- <sup>43</sup>D Mox, M Calvo-Fullana, J Fink, V Kumar and A Ribeiro, Mobile Wireless Network Infrastructure on Demand. arXiv preprint arXiv:2002.03026.
- <sup>44</sup>Nuria González Prelcic, "mmWave for vehicular networks", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>45</sup>Prasant Mohapatra, "IoT communications security," SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>46</sup>Jeremy Dunworth, "The Future of RF Semiconductors: RF Front-end beyond 5G," SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>47</sup>Ali Niknejad, "What to Expect from "6G": Wireless Links from 100 Gbps to 1Tbps," SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>48</sup>Peter Gammel, "Silicon Technology for sub-THz Applications", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>49</sup>D. Daly, et al., "Through the Looking Glass – 2020 Edition", *IEEE Solid-State Circuits Magazine* 12 (2020) 8-24
- <sup>50</sup>H. Rücker & B. Heinemann, "High-performance SiGe HBTs for next generation BiCMOS technology", *Semicond. Sci. Technol.* 33 (2018) 114003
- <sup>51</sup>Nikolaus Klemmer, "GaN for Communications", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>52</sup>Kaushik Sengupta, "Integrated Transmitters and Power Generation in the mmWave and THz: Challenges and Approaches," SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>53</sup>Tony Montalvo, "The future of RF semiconductors", SRC Decadal Plan Workshop on New Trajectories for Communication, Qualcomm, San Diego, CA, USA, February 11-12, 2020.
- <sup>54</sup>M. Hilbert and P. Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information", *Science* 332 (2011) 60-65





---

## Chapter 4

# New Trajectories for Hardware Enabled ICT Security

### Seismic shift #4

Breakthroughs in hardware research are needed to address emerging security challenges in highly interconnected systems and artificial intelligence.

### 4.1. Executive Summary

Today's highly interconnected systems and applications require security and privacy for proper operation (Figure 4.1). Corporate networks, social-networking, and autonomous systems are all built on the assumption of reliable and secure communication but are exposed to various threats and attacks, ranging from exposure of sensitive data to denial of service. The field of security and privacy is undergoing rapid flux these days as new use cases, new threats, and new platforms emerge. For instance, new threat vectors

through the emergence of quantum computing will create vulnerabilities in current cryptographic methods. Thus, new encryption standards resistant to quantum attack must be developed, with consideration given to the impact of these standards on system performance. Also, privacy has emerged as a major policy issue drawing increased attention by consumers and policymakers across the globe. Technical approaches to enhancing privacy include obfuscating or encrypting data at the time of collection or release.

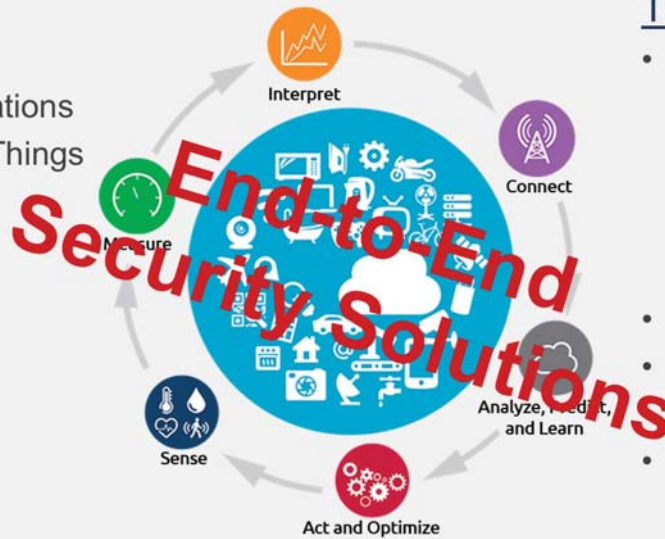
*In another direction, devices have permeated the physical world, and thus trust in these devices becomes a matter of safety.* Security has therefore never been more important. Safety and reliability of systems need to consider malicious attacks, in addition to the traditional concerns of random failures and degradation of physical-world systems. Security of cyber-physical systems needs to consider how to continue to function or fail gracefully while under or after attacks. **To do secure sensor fusion over time, intelligent algorithms**



# Security: A System-Level Property

## Domains:

- Consumer
- Communications
- Internet of Things
- Automotive
- Industrial
- Healthcare
- Aerospace
- Military
- Financial
- ...



## Technologies:

- Hardware
  - Sensors
  - Processors
  - Actuators
  - Radios
  - Power
- Software
- Network / Cloud
- Cryptography
- ...

Constraints: Performance, Cost, Power, Form Factor, Criticality, ...

Hierarchical Approach: From sensors/actuators to cloud, each level should support a hierarchical security monitoring/reacting protocol

Figure 4.1: Systems view of security<sup>1</sup> (courtesy of Yiorgos Makris / University of Texas at Dallas)

are needed that sift through contextual data to evaluate trust. This is a difficult problem as contextual data has tremendous variety and quantity—the systems of the future are actually systems of systems with limitless possibilities for communication and signaling. For instance, cars can communicate with each other and with roadside infrastructure. Like humans, we need to augment systems with the intelligence to trust or not trust all they perceive.

Our hardware is also changing. Complexity is the enemy of security, and today's hardware platforms are highly complex due to the drivers of performance and energy efficiency. Modern system-on-chip designs incorporate an array of special-purpose accelerators and intellectual property (IP) blocks. The security architecture of these systems is complex, as these systems are now tiny distributed systems where we must build distributed security models with different trust assumptions for each component. Furthermore, these components are often sourced from third parties, implying the need for trust in the hardware supply chain. The pursuit of performance has also led to subtle issues in microarchitecture. For instance, many existing hardware platforms are vulnerable to speculative execution side-channel issues, famously

exposed by Spectre and Meltdown. Driven by these problems and others, the future requires fundamentally new hardware designs with innovative security approaches.

The major workload of today is artificial intelligence (AI). Many security systems, for instance, use anomaly detection to identify attacks or employ feature analysis for contextual authentication. AI capabilities continue to increase, and applications for these trusted systems continue to grow. **However, the trustworthiness of the AI for these systems is unclear. This is a problem not just for security systems but also for general systems with implicit trust assumptions, for instance, visual object detection in autonomous vehicles.** Researchers have shown that small perturbations to an image can sway neural network models into the wrong conclusion. A well-placed small sticker on a stop sign can make a model classify it as a Speed Limit 45 sign<sup>1</sup>. Other applications of deep learning systems have similar trust issues: the output of speech recognition might be manipulated with imperceptible audio changes, or malware might go undetected with small changes to the binary. The brittleness of deep learning models is related to their famous inscrutability. Neural networks are black boxes with no

<sup>1</sup>"What is adversarial machine learning?" by Ben Dickson—July 15, 2020 <https://bdtechtalks.com/2020/07/15/machine-learning-adversarial-examples/>

explanation for their decisions. **Other important problems with neural networks are algorithm bias and fairness.** Approaches are needed to make deep learning systems more trusted, explainable, and fair.

*Finally, in the last decade, the systems that we must secure have become immeasurably more complex. The cloud has become the standard for outsourcing computation and storage, while maintaining control. We are still grappling with security challenges arising from the cloud—multi-tenancy, provider assurance, and privacy—while cloud offerings continue to increase in complexity.* Clouds now offer trusted execution environments and specialized, shared hardware and software. At the same time, interest in edge computing is growing as we realize clouds lack the performance and privacy guarantees of nearby compute infrastructure. The heterogeneous nature of the edge implies that trust in edge service providers is a major issue and, of course, the security of IoT devices has plagued us for years. Developing security must be made easier for resource-constrained, often low-cost devices. Even if care is taken in the security design, difficulties arise from extreme environments, such as medical implants. **To compound the problem, systems have become more complicated at every level. Modern system-on-chip designs incorporate an array of special-purpose accelerators and IP blocks—basically tiny distributed systems where distributed security models must be built with different trust assumptions for each component.**

**Security Grand Goal:** Develop security and privacy advances that keep pace with technology, new threats, and new use cases. Examples include trustworthy and safe autonomous and intelligent systems, secure future hardware platforms, and emerging post-quantum and distributed cryptographic algorithms.

## Call for action

The pace at which today's systems are increasing in intelligence and ubiquity is astounding. At the same time, the increased scale and complexity of these systems have forced hardware specialization and optimization to address performance challenges. All these advances in capability must go hand-in-hand with advances in the security and privacy. *Examples include securing weaknesses in the machine-learning or conventional cryptography, protecting privacy of personal data, and addressing vulnerabilities in the supply chain or hardware.*

**Invest \$600M annually throughout this decade in new trajectories for ICT security. Selected priority research themes are outlined below.<sup>ii</sup>**

## 4.2. ICT Security: Fundamentals and Applications

### Overview and needs

Standard cryptographic methods today offer security that meets the anticipated threats. **In the future, with the emergence of quantum computing, standard cryptographic methods will be vulnerable to quantum attack.** Additionally, applications requiring more sophisticated security features, such as distributed consensus, are growing in demand. **At the same time, privacy has emerged as a major policy issue that is drawing increased attention by policymakers across the globe.** *Technical approaches to enhancing privacy include obfuscating or encrypting data at the time of collection or release, and more work is needed on algorithms that can gain insight from obfuscated or encrypted data while preserving individual privacy. Also, new encryption standards resistant to quantum attack must be developed, with consideration given to the impact of these standards on system performance.*

### Securing systems for future defense applications

Secure microelectronics is one of the highest R&D priorities for the U.S. government, and for the Department of Defense in particular<sup>2</sup>. For example, the Air Force works to out-innovate adversaries and, to do so, must rely upon the commercial technology base. **This creates challenges, as defense applications often sit outside the normal mode of operation for commercial products and have security challenges well beyond most commercial applications. Additionally, defense applications typically require the enhanced use of commercial technologies while integrating with non-commercial technologies. The dependency upon commercial industry for key components provides a special vulnerability for defense systems<sup>2</sup>.**

The military applications must take advantage of the benefits of commercial technologies, while also managing the security risks by collaborating with industry to establish risk-based frameworks for security across the entire lifecycle of a microelectronics system. These frameworks must analyze

<sup>ii</sup>The Decadal Plan Executive Committee offered recommendations on allocation of the additional \$3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

all risks and vulnerabilities within the system, manage the risks with appropriate mitigations, prevent migration of the vulnerabilities between layers and domains, and implement the appropriate security posture for the application. In doing this, the United States Armed Forces will be able to accelerate the adoption of advanced technology from the microelectronics industry and out-innovate its adversaries (Figure 4.2)<sup>2</sup>.



### Future Air Force Battlespace

- Combination of Manned, Unmanned, and Autonomous platforms working cooperatively to achieve mission objectives
- Quick decision on-board processing necessary at the edge for optimal decision making
- High bandwidth data exchanges for cloud and remote processing
- Resilient systems capable of operating under duress in contested environments

Figure 4.2: Future Air Force battlespace<sup>2</sup>  
(courtesy of Len Orlando, AFRL)

### Computing for threat intelligence

Cybersecurity has involved a continuous arms race throughout the history and evolution of computing, with adversaries taking advantage of security vulnerabilities, and defenders in a state of continuous catchup. *This happened during the mainframe, PC, and web eras, and is now taking place in the cloud and mobile eras. The advent of artificial intelligence is*

*creating even greater vulnerabilities, where AI will be pitted against AI. We must prepare for this future<sup>3</sup>.*

There are two dynamics where AI will be pitted against AI (Figure 4.3). The first is where AI will be used to automatically employ multiple tools to attack defenses. Defenders will employ AI techniques to counter the rapidly developing AI attacks. The second dynamic is adversarial AI, where attackers will exploit vulnerabilities in AI models to fool or exploit AI systems.

AI-powered attacks will be more evasive, pervasive, and adaptive than any prior era of computing. For example, researchers at Endgame Systems have shown how malware could be evolved to evade security defenses. *Recognizing the potential of AI-powered cyberattacks, DARPA created a Cyber Grand Challenge in 2014 to pit AI attackers against AI defenders with no human intervention. AI systems attacked other AI cyber systems while simultaneously defending against the attack of other AI systems. During the 2018 Black Hat conference, IBM demonstrated DeepLocker, a proof-of-concept to demonstrate the integration of known malware and AI methods to create a sophisticated attack. DeepLocker evaded almost all security methods deployed today.* To guard against such attacks, defenders must employ a combination of AI and continuous machine learning to extract features and patterns, improve decision making, and detect unknown threats. Natural language processing must be employed to help security analysts consolidate threat intelligence. *Reasoning can be facilitated by highlighting evidence of breaches, assisting threat remediation planning, and helping to anticipate new threats.* Automation must be employed to reduce the burden on human analysts, thereby decreasing reaction time.

Adversarial AI is used to evade detection by fooling models, poisoning training data, and stealing training data and trained models. Multiple examples demonstrate that only simple alterations may be needed for this, including one instance where placing a sticky note on a stop sign can make AI algorithms mistake it for a speed limit sign. Protecting AI services requires the integration of Data Security, Model Security, and Application Security. **At the data level, it is critical to protect the integrity, provenance, and quality of the data. Model security is achieved by baking security and privacy guarantees into the entire process of model development and construction.** Application security is accomplished by end-to-end management of AI applications, including operations modeling and application testing<sup>3</sup>.



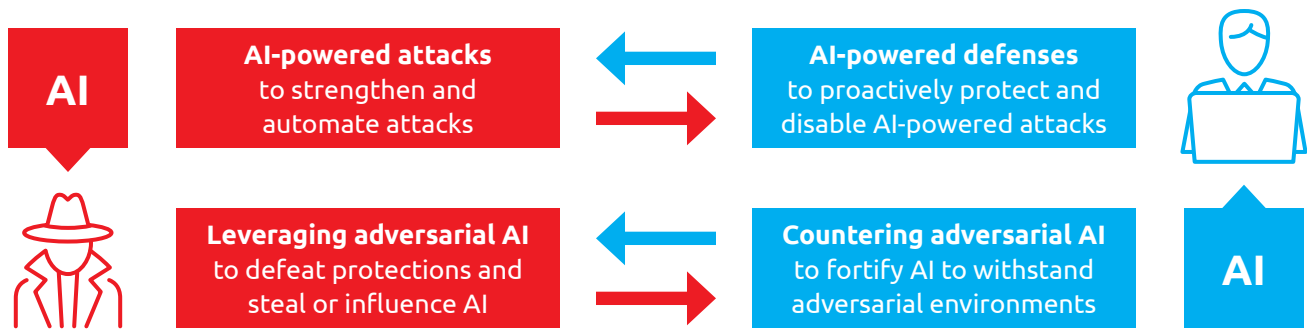


Figure 4.3: The two duals in Cybersecurity and AI<sup>3</sup> (courtesy of Josyula Rao, IBM)

## Embedded security

Interfacing with the analog world creates unique security challenges that requires creating embedded systems resistant to attack<sup>4</sup>. Medical devices, autonomous vehicles, and the Internet of Things depend crucially on embedded security. In fact, the serious security risk represented by analog electronics and sensors represent is often overlooked. With simple side-channel electromagnetic, acoustic, or optical attacks, sensors can be spoofed, and security systems overcome<sup>4</sup>.

Data from sensors are often trusted by the systems that depend upon them for operation, without the type of scrutiny given to digital data. **Therefore, by manipulating the physical phenomena the sensor is designed to interpret, an adversary can cause an undesired course of action.** For example, by altering the voltage interrogated by a thermal sensor, a system can be forced to interpret a temperature as being well below absolute zero<sup>4</sup>. The MEMS accelerometer on a smartphone can be tricked to measure steps by simply playing a YouTube video with embedded sounds that cannot be easily detected. Information can even be communicated using this method. Using a laser, commands can be injected into voice-controlled systems from a distance, even through windows<sup>4</sup>.

When designing systems, it is critical to consider the physics of analog sensors in order to address security. Microprocessors should not blindly trust sensors and should include algorithms to determine the reasonableness of sensor data and detect sudden or unusual changes. Sensors and the systems they interface with should consider potential side-channel attacks and design both physical and software measures to mitigate these attacks. Additionally, there should be a focus on designing trustworthy systems and remembering physics when writing control systems, rather than focusing only on trustworthy components.

## Cryptography in the quantum era

While we are not certain when it will arrive, **quantum computing represents a critical threat to public key cryptosystems and the information they protect**, and the Cryptographic Technology Group at the National Institute for Standards and Technology (NIST) is responsible for developing quantum-resistant encryption standards<sup>5</sup>. The security of well-deployed public key cryptosystems is based on the hardness of factorization (e.g., RSA signature and RSA public key encryption) or upon discrete logarithm problem (e.g., Diffie-Hellman Key Agreement over finite fields and elliptic curves). Quantum computing changes what we believed about the hardness of discrete log and factorization problems, such as when they can be built to a size that can execute Shor's algorithm. Both factorization and the discrete logarithm problem can be solved in polynomial time, thereby providing exponential speedup over classical factoring and discrete logarithm algorithms. Quantum computing can also impact the security of symmetric key-based cryptography algorithms by using Grover's algorithm to search the Advanced Encryption System (AES) keys. However, the speedup using a quantum computer is quadratic, unlike the exponential speedup offered by Shor, and can be mitigated by increasing the key size.

***NIST is developing post-quantum cryptography standards for deployment in 2024. This process commenced in 2016 with the development of post-quantum cryptography criteria and requirements and a call for proposals.*** The transition and migration to the new standards will not be trivial, so early preparation is essential. Plans must be made for algorithm changes for existing systems and for next-generation hardware cryptographic libraries and accelerators (Figure 4.4). Firsthand experience through prototyping is important, as the new algorithms will clearly impact areas



such as power consumption, computing resources, and implementation costs, among other impacts. The transition and migration will be a journey, but the result will be security protection in a post-quantum world<sup>5</sup>.

*Quantum-safe solutions are currently being explored both in government and in the private sector. For example, Amazon is currently assessing the strategy for protecting the overall network architecture for cloud computing of its Amazon Web Service (AWS) against quantum attack<sup>6</sup>. AWS employs multiple protocols (TLS, SSH, IPSec, MACSec, and DWDM) to meet the various needs of the network, and all the protocols require end-to-end security.* In addition, their customers have different requirements for the length of time information must be protected. Information must be protected for as long as the information is sensitive. Some information has limited temporal value, e.g., temporary security credentials, while some must be protected for years, e.g., credit card information. Other information must be protected for decades, e.g., trade secrets and classified information, and there is some that must be protected over a lifetime, e.g., personal information or DNA.

Protecting information on the cloud requires end-to-end protection, from customer to the cloud, and throughout the period the information has value<sup>6</sup>.

There is a range of opinions regarding the time of first availability of a Shor-capable quantum computer, ranging from a decade to many decades. **While there is a range of opinions, and because sensitive information can have value for a very long time, it is critical to prepare for a post-quantum future today.** The protection schemes must assume that an adversary may collect and store encrypted information during transport for subsequent analysis—hence the urgency. AWS’s approach is to use a hybrid key exchange method that combines a classical and a post-quantum key to assure the security and confidentiality is as strong as the combined key security.

AWS is an active participant in the NIST post-quantum standardization activity and works with other international standards bodies on hybrid key exchange standards. AWS has already deployed post-quantum cryptography for customer evaluation and to protect the AWS Key Management Service. This will provide an excellent experience base to support the standardization and rapid deployment of post-quantum cryptography<sup>6</sup>.

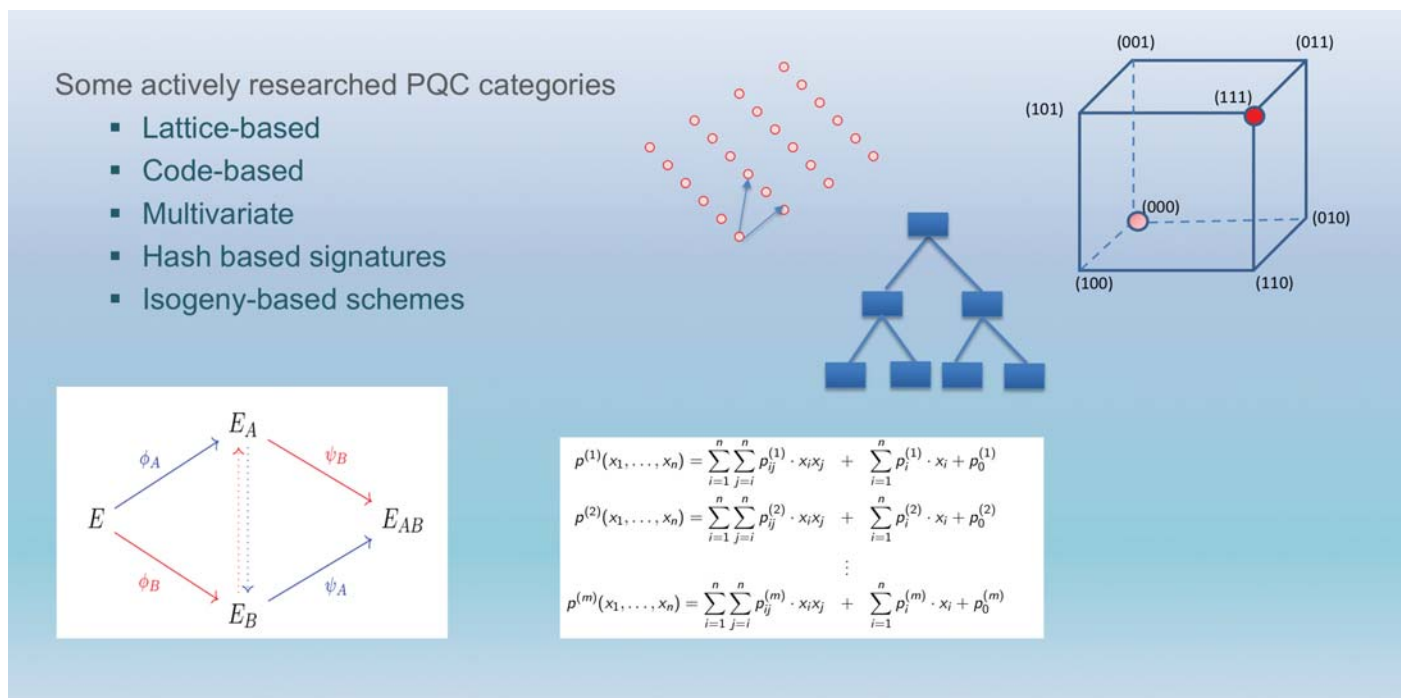


Figure 4.4: Post-quantum cryptography (PQC)<sup>5</sup> (courtesy of Lily Chen, NIST)

### Additional themes and considerations

- Meeting the security needs of the U.S. Department of Defense (DoD) will require a partnership with industry, so the DoD can take advances of industry innovation. It also requires the government to operate at the speed of industry.
- **The cost and complexity of implementing quantum-resistant encryption is significantly larger and more difficult than past transitions.** This is due to both the complexity of the quantum-resistant algorithms and the diversity of applications requiring protection.
- Protecting IoT devices is more than encrypting the digital channel. IoT systems must treat sensors as untrustworthy and should validate sensor data via the software stack and through the overall system design.
- AI and other applications that access large amounts of stored information may be negatively impacted by the overhead of quantum-resistant decryption. However, this may be offset by the positive impact of quantum computers on these applications.
- Protecting information systems requires attention at all levels of system hardware, from design through manufacturing and deployment. It is truly a system-engineering problem.

## 4.3. Security and Safety of Autonomous Systems

### Overview and needs

In the IoT era, a multitude of devices are connected to formulate end-to-end autonomous systems. A tremendous amount of information is generated on the edge using variety of sensors, which is passed through multiple stages of processing and interconnect to form end-to-end autonomous systems. The safety and security of these autonomous systems, therefore, is heavily dependent on how security is handled in various stages of the end-to-end systems. This section focuses on specific challenges currently faced and recommendations on addressing them for security of autonomous systems.

***As devices permeate the physical world, trust in these devices becomes a matter of safety. Thus, security has never been more important.*** System safety and reliability

need to consider malicious attacks, in addition to the traditional concerns of random failures and degradation of physical-world systems. Security of cyber-physical systems needs to consider how to continue to function or fail gracefully, even after attacks. Intelligent algorithms are needed that sift through contextual data to evaluate trust and do secure sensor fusion over time. Contextual data has tremendous variety and quantity. The systems of the future are actually systems of systems with limitless possibilities for communication and signaling. *For instance, cars can communicate with each other and with roadside infrastructure. Finally, as with human processing, systems must be augmented with the intelligence to trust or not trust all they perceive.*

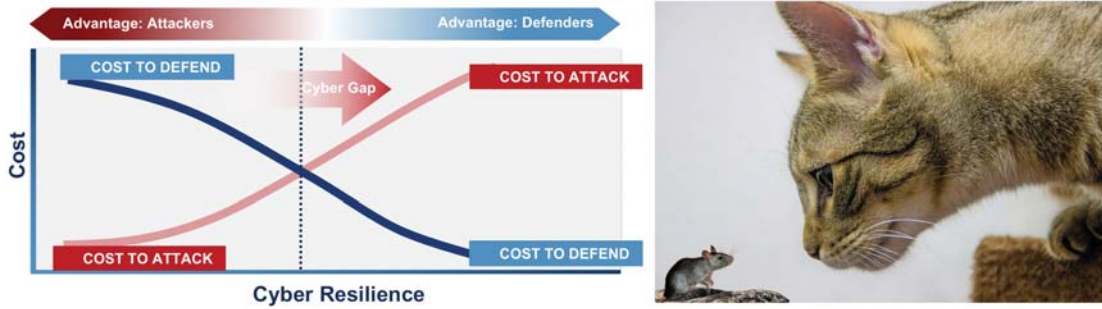
### Endpoint security in hyperconnected future

The ICT value chain ranges from the sensors generating information to different processing nodes to users, and it involves a variety of devices and HW/SW platforms coming from multiple suppliers and stakeholders. The hyperconnected world is built on smart devices, systems, and systems of systems that are moving toward autonomous control using AI. This autonomous control necessitates trusted data, which involves identifying data source and integrity<sup>7</sup>.

**Lack of adequate data security (data identity and data integrity) and the current security attack defense strategies based on known threats in cyber systems favor attackers, with cost of attack being low against the cost to defend the systems.** Research needs to focus on building fundamental solutions to eliminate root causes and tip the balance in favor of defenders by lowering the cost to defend and raising the cost to attack (Figure 4.5)<sup>7</sup>.

Since security is directly impacted by the complexity of processing, enabling the safety and security of autonomous systems requires building HW-primitive secure functions at the device level and then addressing security hierarchically at each level of the value chain. Hardware Root of Trust (data identity and integrity) is required at each level of the value chain to ensure the system's security, from data creation to processing at different stages. The systems also need to be able to adopt the HW RoT to account for network failures or network upgrades. *AI can play an important role in building secure and safe autonomous systems but will not address all security issues unless HW RoT is deployed throughout the connected systems* (Figure 4.6)<sup>7</sup>.

## Required: Stop Cat and Mouse Games!



- Must fundamentally change cost to attack verses cost to defend
- Research & develop fundamental solutions that eliminate root source – not prevent based on today's technology level

Figure 4.5: The current security attack defense strategies based on known threats in cyber systems favor attackers with cost of attack being low against the cost to defend the systems. Research needs to focus on building fundamental solutions to eliminate root causes and tip the balance in favor of defenders by lowering the cost to defend and raising the cost to attack<sup>7</sup>. (courtesy of Doug Gardner, Analog Devices)

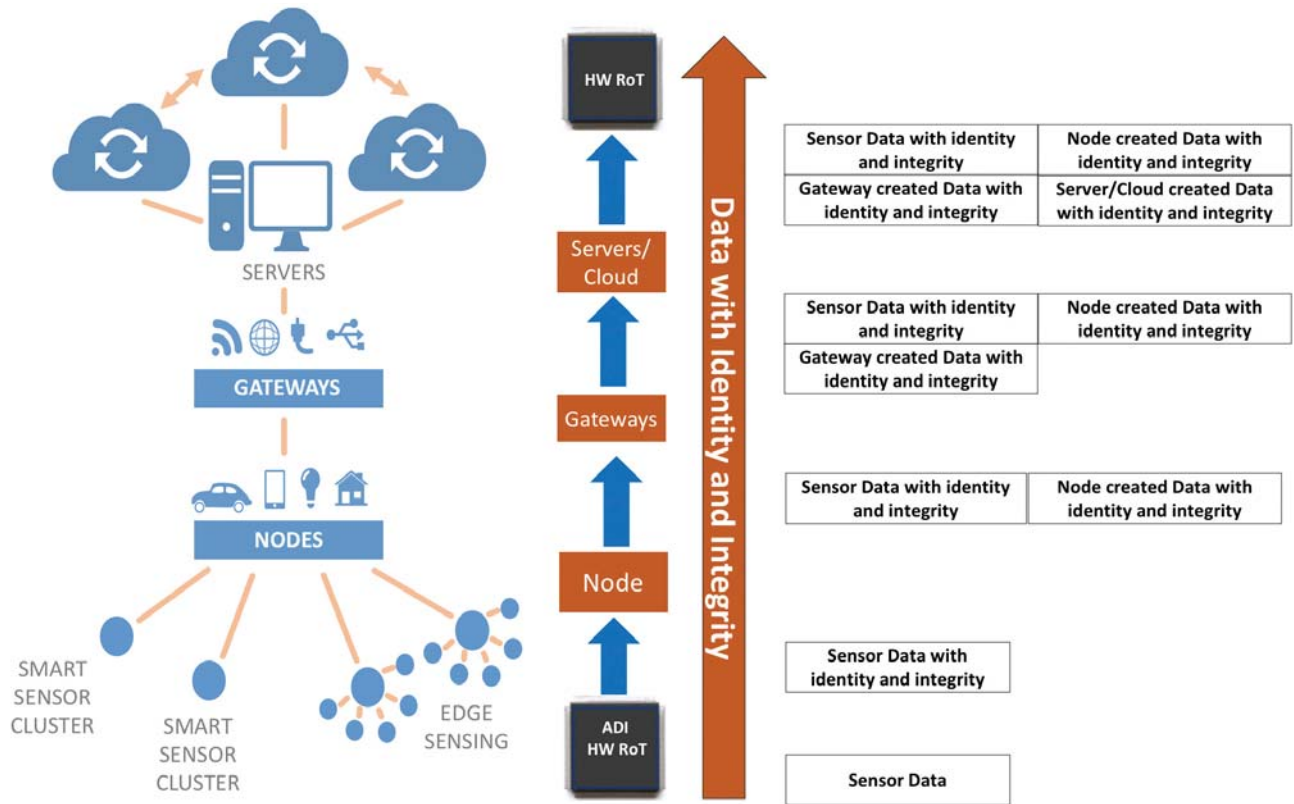


Figure 4.6: Security and safety framework for autonomous systems<sup>7</sup> (courtesy of Doug Gardner, Analog Devices)

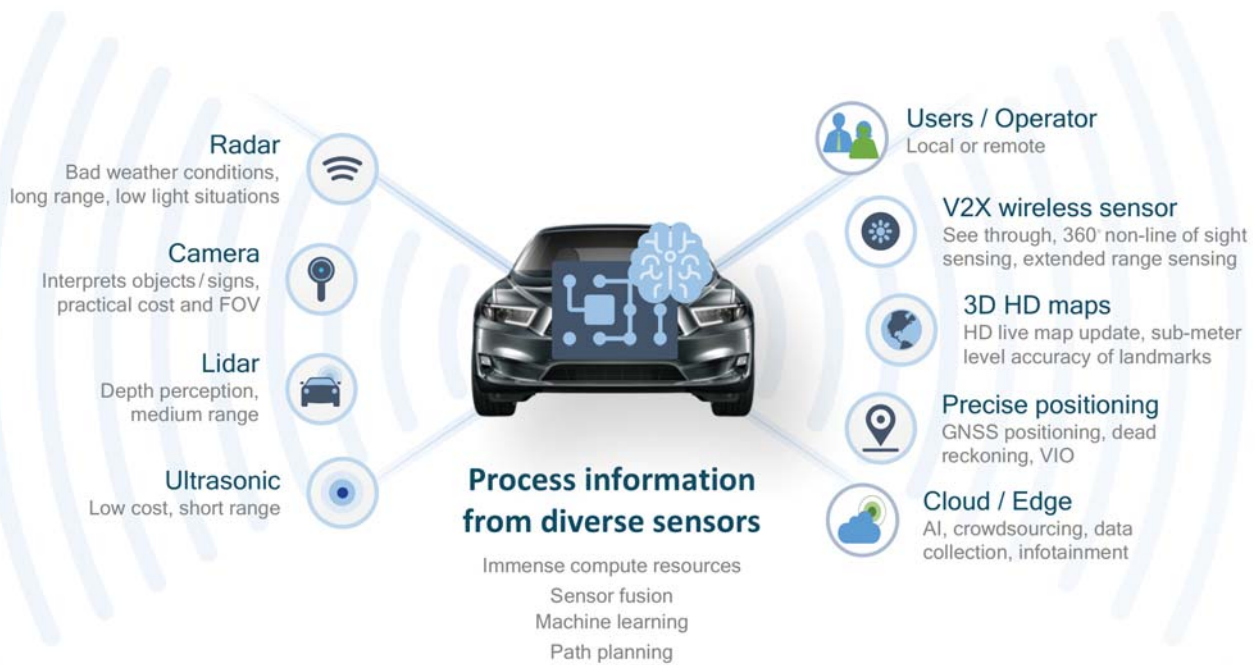


Figure 4.7: The nervous system of the autonomous vehicle<sup>8</sup> (courtesy of William White, Qualcomm)

## Security in autonomous vehicles

Autonomous vehicles are built using a large number of integrated sensors, thereby increasing their susceptibility to adversarial attacks (Figure 4.7). As an example, an attacker may flash high-speed-limit images around a corner and cause vehicle to lose control. There is a need to build security for each of the sensor systems that generate critical data, as well as building robust AI networks that together provide the autonomy of vehicle operation. **The secure systems must be built at a more robust HW level with cryptographic solutions, as current public-key cryptography is seriously threatened by the huge processing power of quantum computing<sup>8</sup>.**

*The growing complexity of systems like automotive- and 5G-IoT is more vulnerable to security threats, as increasing code size generally leads to higher security risk. A Hardware-Intensive Virtualization Architecture (HIVA) approach was proposed based on decomposing applications to fit into small VMs (Virtual Machines) secure execution environments<sup>9</sup>. Different types of vulnerabilities, e.g., buffer/resource management, quantum cryptography, AI-enabled attack, etc., can be mapped against specific HW building blocks to tackle the individual vulnerabilities (Table 4.1). This process is just beginning, and solutions for many types of vulnerabilities still need to be picked up by research and development.*

## Smart grid security

The smart grid is essentially an evolution of the existing grid, using internet connectivity and consisting of embedded

computing systems distributed over telemetry and remote control (SCADA) toward Industrial IoT (IIoT). While smart-grid deployments benefit from the flexibility and cost of telecommunication and network-based devices, the autonomous systems using these smart grids are also susceptible to a widening collection of known and unknown cyberthreats<sup>10</sup>.

Addressing the security risks for smart grids requires a continuous analysis of threats and adaptation of system architecture via a full-stack “DevSecOps” posture (Figure 4.8). *Using Machine Learning, sustained telemetry, digital twins, and “Analysis by Synthesis” for devices, adaptation of compute architecture can mitigate new threats. By also including an evaluation of ROI for upgrades and synthesis, the system components can be upgraded accordingly<sup>10</sup>.*

Table 4.1: Vulnerabilities defended by hardware building blocks in HIVA<sup>9</sup>

| Vulnerability             | Hardware Building Blocks                                     |
|---------------------------|--|
| Buffer overflow           | Tagged memory  |
| Resource management       | MMU (Memory Management Unit)<br>MPU (Memory Protection Unit) |
| Post Quantum Cryptography | Lattice engine and more                                      |
| Information leakage       | to be explored   |
| Permission                | to be explored   |
| Code injection            | to be explored   |
| Numeric error             | to be explored   |
| Video/sensor data path    | to be explored   |
| AI-enabled attacks        | to be explored   |
| Digital M/S latch         | to be explored   |



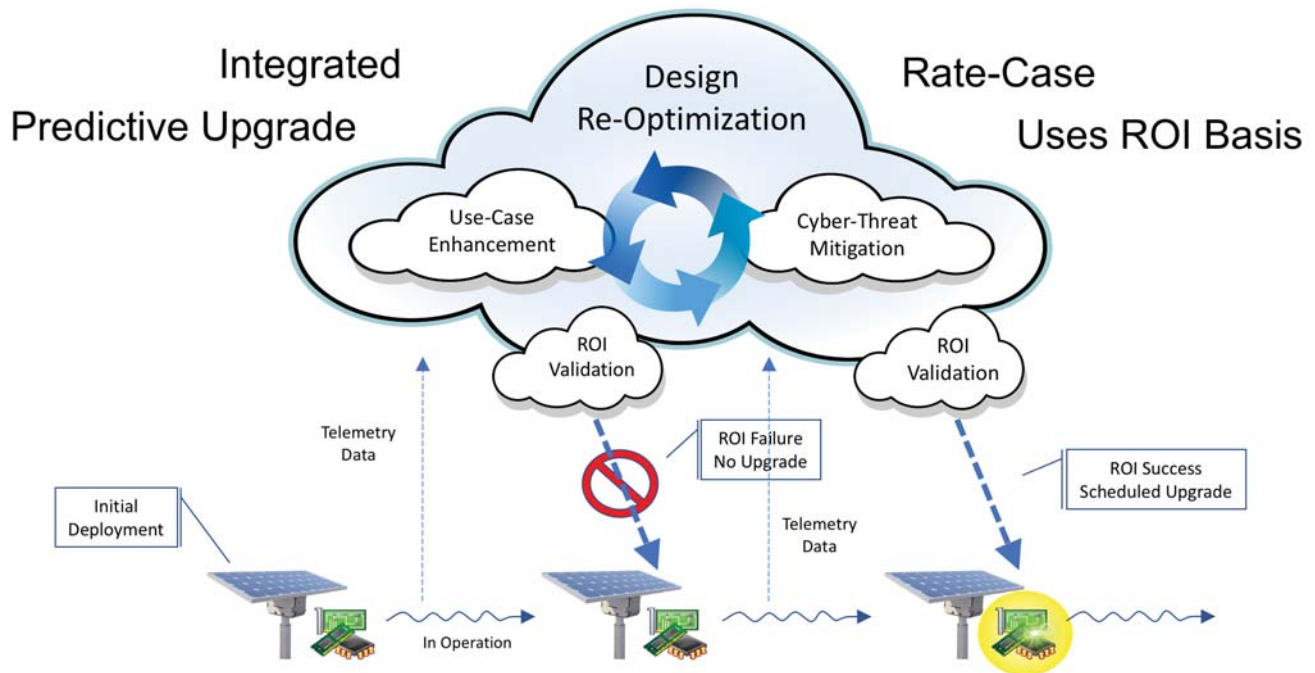


Figure 4.8: Integration of the DevSecOps engineering practice with Machine Learning and telemetry will allow for a joint and continuous optimization of the CPU architecture, including critical use-case analysis, metrics for evolving/simulated threat vectors, and cost of upgrade vs. liability of compromise.<sup>10</sup> (courtesy of Stan McClellan, Texas State University)

### Autonomous system security

There is a need to derive threat models of complex autonomous systems with various layers of control to obtain a security-aware autonomy architecture that will enable tradeoffs between security-related overhead and the overall quality of control in the presence of attacks. Machine learning (ML) networks have to model adversarial networks to evaluate different threat scenarios. Stealthy attack scenarios for automotive systems, for instance, must employ adversarial learning approaches and HW-SW co-design of cyber-physical security components<sup>11</sup>.

#### Key areas of focus and follow-on research

- Building end-to-end security for highly connected autonomous systems, HW RoT and security primitives, and trusted data
- Minimizing the security threats at the root, which requires building secure smaller (small code size) components and constructing complex systems from these secure components
- Lengthening system durability and security through careful dynamic assessment and adaptation against failures and threats
- Using AI and ML to find appropriate solutions to detect, analyze and adopt security threats
- Finding alternatives to current cryptography techniques, which are under threat due to quantum computing and research is required to find alternatives

## 4.4. Secure Hardware Design

### Overview and needs

Complexity is the enemy of security, and today's hardware platforms are highly complex due to the drivers of functionality, performance, and energy efficiency. Modern SoC designs incorporate an array of special-purpose accelerators and IP blocks (Figure 4.9). **The security architecture of these systems is complex, as these are now tiny distributed systems where security models with different trust assumptions for each component must be built. Further, these components are often sourced from third parties, implying the need for trust in the hardware supply chain.** The pursuit of performance has also led to subtle issues in microarchitecture. For instance, many existing hardware platforms are vulnerable to side-channel attacks. Finally, for many IoT devices, primary design considerations are cost, time to market, energy efficiency, and consumer usability. **Security is very low on the priority list (if it's there at all), and available HW resources are severely constrained. Driven by these and other problems, the future requires fundamentally new hardware designs.**

The hardware security topics include, among others, HW attacks from HW Trojans (used for Denial-of-Service attacks and information leakage), untrusted foundries, counterfeit ICs, physical attacks, side-channel attacks, fault injection (used for violation of Integrity and Confidentiality), reverse engineering, and fake parts<sup>12</sup>.

It should be emphasized that PCB supply chain has received little attention in HW security, all the while being easier to attack. **SoCs are also becoming new attack points, as they are connected to the outside world through WiFi and other communication means.** In designing secure SoCs, security rule checks can be recommended with the objective to provide automated security assessment and possible countermeasures of a given design for target vulnerabilities<sup>12</sup>.

The entire lifecycle needs to be considered in the context of HW security: design, fabrication, assembly, distribution, lifetime, end of life/recycling. In an SoC lifecycle, the integration phase (going from RTL to layout) is a particularly critical step. Overall, three guidelines can be recommended: protect the IP, protect the assets, and protect the supply chain. For example, steps in protecting the supply chain should include IC authentication (using ECID, PUF chip ID), PCB authentication, subsystem authentication, and HW/FW self-authentication.

## Secure computing

**A critical question in the context of hardware security is how to trust “remote” computation.**<sup>13</sup> One approach would be to minimize the Trusted Computing Base (TCB). Multi-tenancy requires architectural isolation of processes, which is fundamental to maintaining correctness and privacy. However, performance optimization at the microarchitectural level makes this a big challenge. Isolation breaks because of the

shared microarchitectural state. Side-channel attacks exploit this, as was shown in famous *Meltdown* and *Spectre* cases.

*A stronger form of isolation can be achieved by using enclaves, as they strengthen the process abstraction. Processes guarantee only isolation of memory, whereas enclaves provide a stronger guarantee.* In fact, no other program can infer anything private from the enclave program through its use of shared resources or shared microarchitectural state. Enclaves also decouple performance considerations from security. Finally, enclaves provide security guarantee under chosen threat models.

Three strategies for building enclaves can be suggested<sup>13</sup>: spatial isolation, temporal isolation, and cryptography.

An important topic in secure computing is mitigation of microarchitectural vulnerabilities<sup>14</sup>. **Currently, the HW/SW landscape is evolving towards heterogenous computing, where utilization is driving increased disaggregation and sharing of resources.** In the cloud SW landscape, third parties are delivering programming to the cloud as VMs, applications, or even function levels. Cloud, on the other hand, supports Confidential Computing via Trusted Execution Environments (TEEs), VMs, processes, or even functions. Cloud is also making many HW primitives SW controlled. With all these changes, different attack surfaces emerge that HW must comprehend and defend. **Figure 4.10** provides an evolutionary view of cache side-channel attacks and mitigations for them. Several relevant questions arise in this environment: **How can a HW developer go about creating new features in microarchitecture? Would it create new vulnerabilities? How would the mitigation stand the test of time and defend against new attacks? These challenges apply not only to cache side-channel attacks, but also to all optimization problems across the platform.**

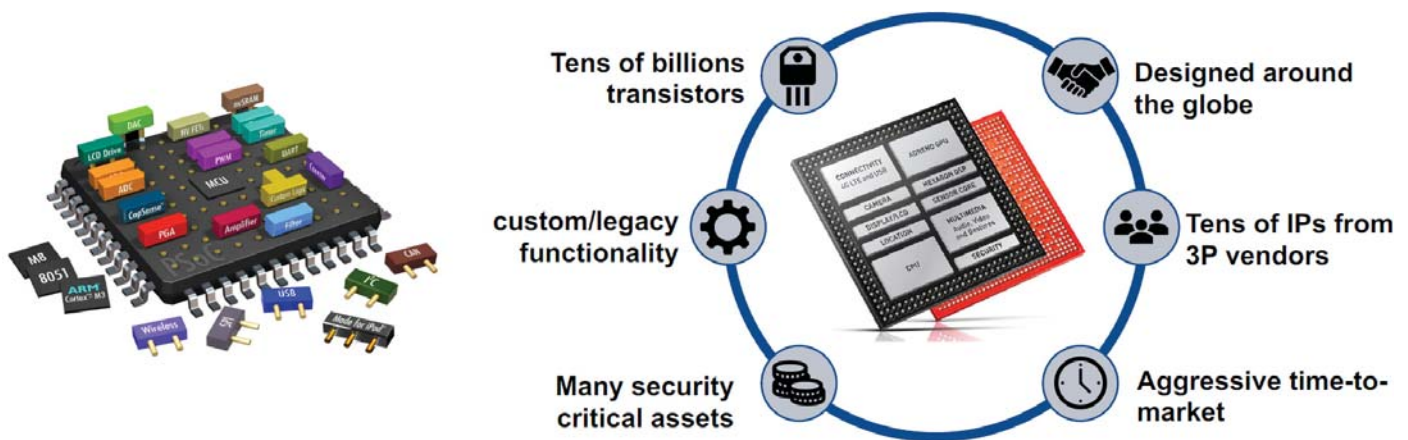


Figure 4.9: SOC security is a challenge<sup>12</sup> (courtesy of Mark Tehranipoor, University of Florida)

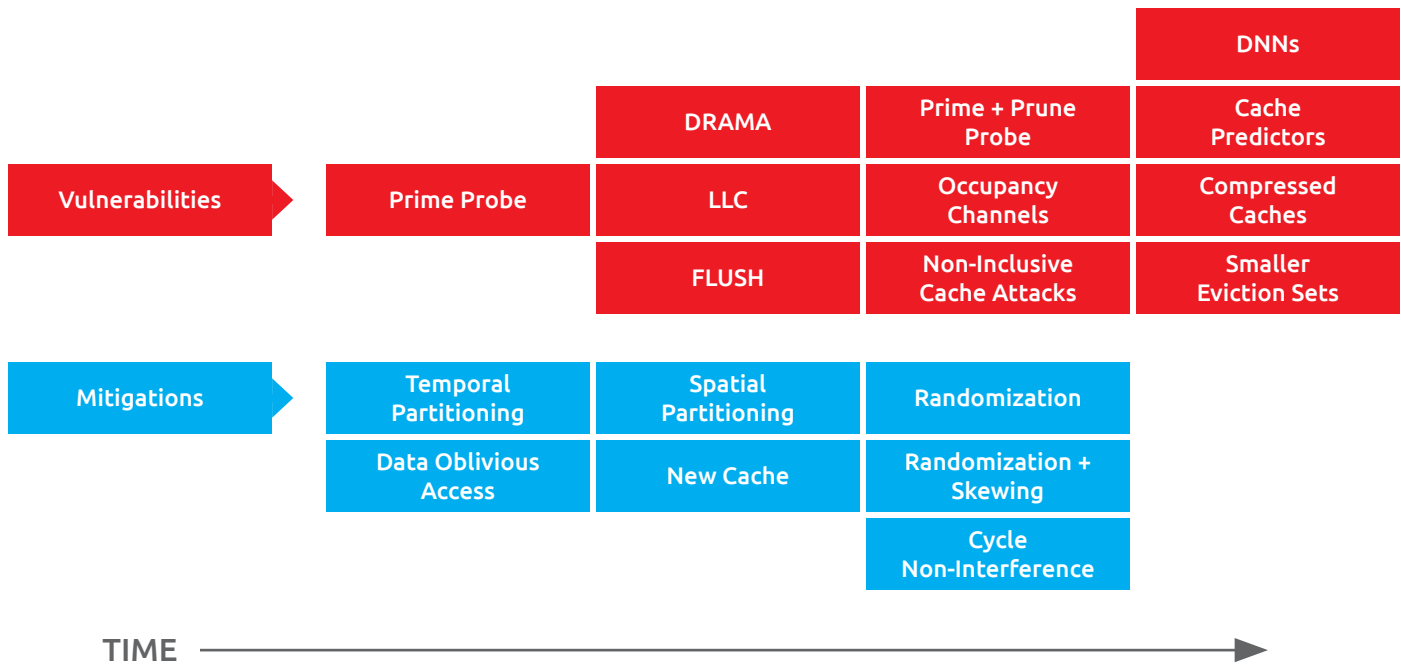


Figure 4.10: An evolutionary view of cache side channels<sup>14</sup> (courtesy of Carlos Rozas, Intel)

## Hardware for cryptography

All hardware, from cloud to edge to IoT, needs security and cryptography<sup>15</sup>. In the past, we assumed that security attacks take place on the channel between communicating parties. Protection was provided through strong mathematical algorithms and protocols. Therefore, the focus was on efficient implementations of the cryptographic algorithms, such as DES and 3DES. Currently, the attack models include the channel and the end points. Not only are strong algorithms and protocols needed, but so are the secure implementation of them. The present need is for efficient and side-channel/attack-resistant implementations<sup>15</sup>.

## Security and trust in the analog/mixed-signal/RF domain

Typical IC lifecycle involves numerous entities around the globe<sup>1</sup>. This distributed nature justifiably raises security concerns, and because of it many threat models and solutions have been created. However, they are mostly in the digital domain. Given that analog components play a huge role in many systems, from sensing and measuring to interpreting, acting, and optimizing, their security should not be overlooked. Not only do analog security and trust lag behind their digital counterpart, a 2017 survey<sup>16</sup> showed that analog efforts were simply mirroring digital domain approaches, such as HW Trojans, reverse engineering, and counterfeiting. The picture is slowly changing though, as a 2019 follow-up survey<sup>17</sup> showed emergence of analog-specific solutions to security challenges.

In the RF communication area, integrity and privacy are well-studied through communication and information theory, and mature solutions exist (jamming, interference, covert channels, etc.). But physical-layer security is an area of opportunity for exploration and development<sup>1</sup>.

Given that the world is analog, and security must be an end-to-end solution, the analog domain should not be the weakest link. However, the analog/RF IC security and trust topics have yet to receive the attention and support needed to take off. **One of the key challenges is that analog/RF designs still have limited automation, which hampers security analysis and solutions, and formal methods are extremely challenging for analog domain<sup>1</sup>.**

## Key areas of focus and follow-on research

### General

- Improved security of billions of very-low-cost and easy-to-implement devices
  - While the design of mission-critical systems carefully considers security and attempts to address it while absorbing necessary costs, many IoT devices are driven by cost and quick/easy consumer adoption. They do not have even the most rudimentary security defenses.
- Hardware futureproofing
  - Hardware is generally inflexible. Some systems built on a given HW may remain in the field for over a decade. Security attacks, however, continue to evolve.

- "Security maturity" signoff of HW design before its tape out
  - HW design has advanced to have systematic methods to evaluate it for manufacturability, reliability, testability, etc.
- Vulnerability of the NVM systems to hardware attacks
- Methods to prevent/mitigate IP piracy at foundries
- Ways to synthesize RTL for guaranteed security against known attacks

### Secure computing

- Building single-chip secure processors
  - Open source, formally verified TCB
  - Secure against all practical SW attacks
  - Secure against physical attacks of memory
  - Enhanced physical security against invasive attacks
- Minimizing performance overhead
  - Interplay of security, performance, usability, and complexity creates a very challenging problem.

- Developing threat models
  - What actual threats need to be mitigated? How effective are they in practical scenarios?
- Documenting HW/SW 'contract' for microarchitectural properties around security
  - How can SW reason and express the security property it desires?
  - How can HW innovate on microarchitectural design to improve performance and power?

### Hardware for Cryptography

- Crypto-diversity, lightweight crypto with side-channel security, and low-latency crypto
- More than Moore: Post-quantum computing cryptos, including secure implementation as through masking
- Hardware-entangled crypto, which has minimum Root of Trust through PUF (Physically Unclonable Function) and TRNG (True Random Number Generator)
- Homomorphic encryption

## 4.5. AI Security and Privacy

### Overview and needs

*AI is now integral to many security systems, for instance, anomaly detection to identify attacks or feature analysis for contextual authentication. AI's capabilities continue to increase, and applications for these trusted systems continue to grow.* However, the trustworthiness of the AI for these systems is unclear. This is a problem not just for security systems but even for general systems with implicit trust assumptions, for instance, visual object detection in autonomous vehicles. Researchers have shown that small perturbations to an image can sway neural networks models into the wrong decision. A well-placed small sticker on a stop sign can make a model classify it as a Speed Limit 45 sign. Other applications of deep learning systems have similar trust issues: the output of speech recognition might be manipulated with imperceptible audio changes, or malware might go undetected with small changes to the binary. **Why deep learning models are so brittle is unclear as deep neural networks are famously inscrutable. A related problem is algorithm bias and fairness. Approaches are needed to make deep learning systems more trusted, transparent, and fair.**

### Adversarial machine learning

Adversarial machine learning is in its early days<sup>18</sup>. Similar to the early days of crypto, defenses against attacks are proposed and then quickly broken.

As illustrated in **Figure 4.11**, all steps of the machine-learning pipeline can be attacked. During the training phase, the attacker can augment or replace some small fraction of the training set. The goal of the attacker is to maximize the error of the resulting model (untargeted) or cause misclassification of a specific sample (targeted). The challenge problems for training-set poisoning include addressing targeted attacks and assessing if system-level considerations or hardware can help.

During the test output phase, model theft uses queries to steal the machine-learning model. For example, an attacker may want to replicate a cloud service that provides a classification service. Challenge problems related to model

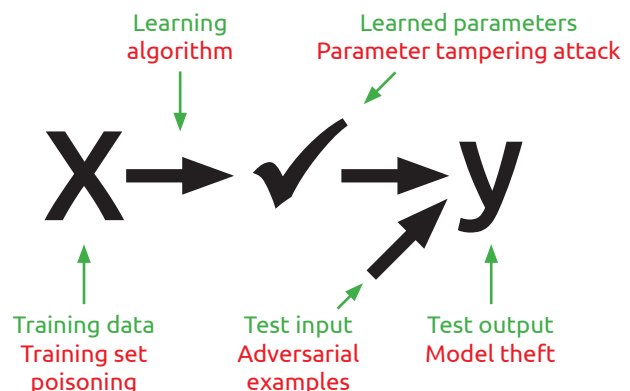


Figure 4.11: Attacks on the machine-learning pipeline<sup>18</sup> (courtesy of Somesh Jha, UW-Madison)



theft include finding robust defenses and investigating trusted hardware or new economic models.

During the test input phase, adversarial examples are the most studied attack on machine learning. Adversarial examples are perturbations created by attackers to cause misclassification. Robust defenses have proved elusive. These attacks are usually white box in the sense that the models are known to the attacker. Attacks can also be black box, for example when the attacker replicates the model and does a white-box attack on the replicated model. It turns out that adversarial examples transfer well, so that adversarial examples on the replica will likely be adversarial examples for the original. **Challenge problems in adversarial examples include determining if adversarial training can be accelerated and understanding adversarial models and what makes them robust.**

Deepfakes are not part of the machine-learning pipeline but this is a related important problem with deep societal implications. The essential technical problem is to determine if content is produced by a generative model, as opposed to being natural. This is a difficult problem, a sort of Turing test.

Note that the problems facing security in machine learning are not necessarily new, and we should take lessons from past work that may apply, including:

- Training-set poisoning is similar to what is done in robust statistics.
- Model theft is similar to what is done in active learning.
- Adversarial examples are similar to what is done in robust optimization.

## Secure computation for AI

*We are now seeing more entities interested in or concerned about sharing personal and proprietary data. The entities here can be roughly described as institutions, customers, and regulators. Institutions are looking at new services that require collaboration on shared data. Customers are gaining understanding of the value and risks of sharing their data. Regulators understand the excitement around sharing data, but also are concerned about potential abuse<sup>19</sup>.*

Encryption methods are one of the main technologies that enable data sharing and yet maintain privacy. In the usual model of inference, a user submits data to a cloud service provider that uses a machine-learning model to provide an inference based on the user’s data. Homomorphic encryption is one way for a user to avoid submitting sensitive data. The user submits homomorphically encrypted data, and the model is evaluated over the encrypted data. The encrypted result is returned to the user, who can then decrypt the result. Currently, open-source tools are available that allow anyone to integrate homomorphic encryption into ML applications<sup>20</sup>.

Figure 4.12 outlines how encryption scope is gradually shifting in two phases toward (a) stronger encryption techniques and (b) techniques to compute on encrypted data. The first phase includes the shift from pre-quantum to post-quantum cryptography. The second phase includes the shift from cryptographic methods to protect the confidentiality of data in transit and at rest to cryptographic methods to protect the confidentiality of data in use with the ability to compute on encrypted data.

| Security Mechanism   | Building Block  | Data Types                   | Functions  | Evaluate/ Produces   | Operations  | Complexity  |   |
|--|---|------------------------------|--|--|---|---|---|
| Bulk Encryption/ Message Integrity                                 | Block Cyphers/ one-way Hash Functions                     | Bytes                        | Key generation/ Encryption/ Decryption/ Digest   | Cipher Text/ Message Authentication Code   | Finite Field Arithmetic/ Boolean Arithmetic                                     | $O(\#rounds)$<br>"AES128 is 10 rounds"              | Secure storage/ communication<br>pre-quantum  |
| Key Management/ Digital Signature                                  | RSA/ ECC  | Big Numbers                  | Key generation/ Encryption/ Decryption/ Protocol | Transfer Key/ Agree Key/ Digital Signature   | Modular Arithmetic  | $O(\#bit\ key)$<br>">2048 for RSA"                  |   |
| Security Mechanism   | Building Block  | Data Types                   | Functions  | Evaluate/ Produces   | More Operations   | Increased Complexity                                | Secure storage/ communication<br>post-quantum |
| Bulk Encryption/ Message Integrity                                 | Block Cyphers/ Open-way Hash Functions                    | Bytes                        | Key generation/ Encryption/ Decryption/ Digest   | Cipher Text/ Message Authentication Code   | Finite Field Arithmetic/ Boolean Arithmetic                                     | $O(\#rounds)$<br>"AES256 is 14 rounds"              |   |
| Post-quantum Key Management/ Digital Signature (based on lattices) | Lattice-based Cryptography (R-LWE)                        | Polynomials of big numbers   | Key generation/ Encryption/ Decryption/ Protocol | Transfer Key/ Agree Key/ Digital Signature   | Modular Arithmetic/ Polynomial Arithmetic/ Noise distribution                   | $O(\text{polynomial degree})$<br>"degree 512 – 2K"  |   |
| Secure Computing Mechanism   | Building Block  | Data Types                   | Functions  | Evaluate/ Produces   | More Operations   | Increased Complexity                                | Compute on Encrypted Data                     |
| Homomorphic Encryption (BGV, BFV, CKKS)                            | Lattice-based Cryptography (R-LWE, other algebraic rings) | Polynomials of big numbers   | Key generation/ Encryption/ Decryption/ Protocol | Arithmetic or Boolean Circuits/ Arithmetic or Boolean Operations on Encrypted Inputs | Modular Arithmetic/ Polynomial Arithmetic/ Noise Distribution/ Noise Management | $O(\text{polynomial degree})$<br>"degrees 8K – 32K" |   |
| Secure Multi-Party Computation (Yao's Garbled Circuit, GMW)        | Pseudo-Random Function (AES based), Oblivious Transfers   | Garbled Bits/ Garbled Tables | Key generation/ Encryption/ Decryption/ Protocol | Boolean circuits/ Key Agreement/ Oblivious Transfer                                  | Finite Field Arithmetic/ Boolean Arithmetic                                     | $O(\#inputs/\#critical\ path\ length)$              |   |

Figure 4.12: Trends in applied cryptography<sup>19</sup> (courtesy of Rosario Cammarota / Intel)

**The methods of the second phase bring a paradigm shift in both cryptography and computing.** In traditional cryptography, cryptographic techniques are decoupled from the applications. In contrast, to protect data in use, computing must meet cryptography, and the corresponding cryptographic methods become entangled with the applications.

Examples of the techniques in the second phase are homomorphic encryption and secure multi-party computation. **These changes bring new fundamental data types, additional challenges with data movement, and an increase in the complexity of the cryptographic algorithms. Current implementations of these schemes in software are still impractical in storage, communication, and computational overhead, despite progress in making this software more efficient. In addition to the overheads, applications need to be rewritten to instruct existing hardware to process encrypted data.**

In spite of these challenges, the benefits of introducing these cryptographic methods to data privacy and economic growth would be unparalleled. Hence, there is a clear need for new hardware for cryptography to compute on encrypted data. Besides technologies, standardization and educational paths are required.

### Secure collaboration for AI

Many organizations need to learn from cross-organizational data but cannot share their data. For example, banks would like to share data for anti-money laundering. Open-source platforms are currently being developed to enable secure collaboration<sup>21</sup>. One example is MC2, as illustrated in **Figure 4.13**. Two approaches used in these platforms are secure, multi-party computation and hardware enclaves. There is an inherent tradeoff between these approaches: multi-party computation is slow and expensive, while hardware enclaves require trusted hardware (with trust assumptions) and setup.

Differential privacy, in addition to cryptography and hardware enclaves, provides yet another complementary approach.

While the traditional approach is to have different research thrusts of the algorithm and the hardware, with security and privacy coming later, an important task is co-optimization of algorithms, hardware, and security<sup>22</sup>. Federated learning involves training a machine-learning algorithm on multiple local datasets contained in local nodes. This is done without exchanging data but, instead, by exchanging model parameters, as with neural network weights. *One solution to the problem of machine learning on encrypted data is optimization of deep-learning computations represented as Boolean circuits. But this approach has been only partially successful so far. Hybrid and co-optimization of the design appears to be a more promising approach<sup>22</sup>.* To fully address federated learning, co-optimization of the algorithm, protocol, software, and hardware is needed. We also need end-to-end automated solutions to remove non-recurring engineering costs.

### Key areas of focus and follow-on research

- Building defenses for all parts of the machine-learning pipeline (e.g., model-stealing, training-set poisoning, and adversarial), as they can all be attacked
- Solving deepfakes, an important unsolved problem with deep societal implications
- Merging of cryptography and artificial intelligence critical for user privacy
  - Algorithms such as Fully Homomorphic Encryption require advances in hardware, storage, and communication to be fully practical.
- Learning from cross-organizational data that cannot be shared
  - Algorithms and co-optimization are needed across the hardware, software, and network.

### Secure multi-party computation (MPC)

Efficient protocols:

- Delphi** [UsenixSecurity'20] - CNN inference
- Helen** [IEEEESP'19] - linear models training
- Senate** [UsenixSecurity'21] - general analytics/ML
- Bost et al.** [NDSS'15] - classical ML inference

### Hardware enclaves

Oblivious & efficient protocols, and automatic partitioning:

- Visor** [UsenixSecurity'20] - ML inference
- Civet** [UsenixSecurity'20] - general programs
- Membuster** [UsenixSecurity'20] - general programs
- OCQ** [Eurosys'20] - distributed analytics
- Oblix** [IEEEESP'18] - multi-user search
- Opaque** [NSDI'17] - analytics

Figure 4.13: Projects in MC2<sup>21</sup> (courtesy of Raluca Ada Popa/UC Berkeley)

## 4.6. System-level Security

### Overview and needs

The proliferation of hyper-connectivity, the continuation of Moore's Law, and the implications of system-of-systems has created unprecedented solution complexity—spanning from the cloud to system-on-chip designs—that impact our ability to assess and attest to system-level security. The challenge of understanding and ensuring system-level security creates an unprecedented level of complexity, especially given the integrated nature of such designs and their dependence on hardware, software, design flows, and lifecycle considerations. Here, system-level security challenges are explored, along with proposed solutions and recommendations.

In the last decade, the cloud has become the standard for outsourcing computation and storage, while maintaining control. **We are still grappling with security challenges arising from the cloud (multi-tenancy, provider assurance, and privacy), while cloud offerings continue to increase in complexity.** Cloud offerings include trusted execution environments and specialized, shared hardware and software. **At the same time, interest in edge computing is growing, as cloud lacks the performance and privacy guarantees of nearby compute infrastructure.** The heterogeneous nature of the edge implies that trust in edge service providers is a major challenge. Similarly, the security of IoT devices has been a plaguing challenge for years. Developing security must be made easier for small, often low-cost devices. Even if care is taken in the security design, difficulties arise from resource constraints or extreme environments, such as medical implants.

*Finally, systems have become more complicated at every level.*

Modern system-on-chip designs incorporate an array of special-purpose accelerators and IP blocks, which are basically tiny distributed systems where distributed security models must be built with different trust assumptions for each component.

### Securing platform integrity and trust with hardware security

There are many hardware security challenges today, such as **counterfeiting, malicious insertions in microelectronics, backdoors, side-channel attacks, extraction and bypass of roots of trust, and supply-chain assurance challenges**<sup>23</sup>.

Considerable security risks are associated with the involvement of third parties, so it's vital to develop risk-mitigation strategies as the use of third-party IP, software, and solutions continues to grow with system complexity.

*The mitigation strategies include early-stage analysis, industry-driven research and standards, and emerging technologies.*

Special attention must be given to overall system integrity with supply chain, as well as security design. The implications for system security in the emerging era of heterogeneity must be addressed. For example, the PNNL's Data Model Convergence (DMC) initiative explores purpose-built architectures for both data analysis and scientific simulation and has a long background in applying mathematics, program analysis, and machine-learning techniques to problems of national security (Figure 4.15)<sup>24</sup>. The complexities of converged workloads in high-performance computing involving modeling and simulation, ML/AI, and graph analytics call for more robust

and scalable heterogeneous program analysis techniques to properly evaluate software and system-level security. Correctness and attestation of the system must be proven, as we move from unit to component to system, and finally to system of systems.

**The current challenges can be characterized as the “blurring of lines” on programming language, compiler infrastructure, and architecture. More research is needed to understand how to leverage system heterogeneity to prove a solid foundation for system-level security.**



Figure 4.14: Third-party security risks<sup>23</sup> (courtesy of Yousef Iskander / Microsoft)



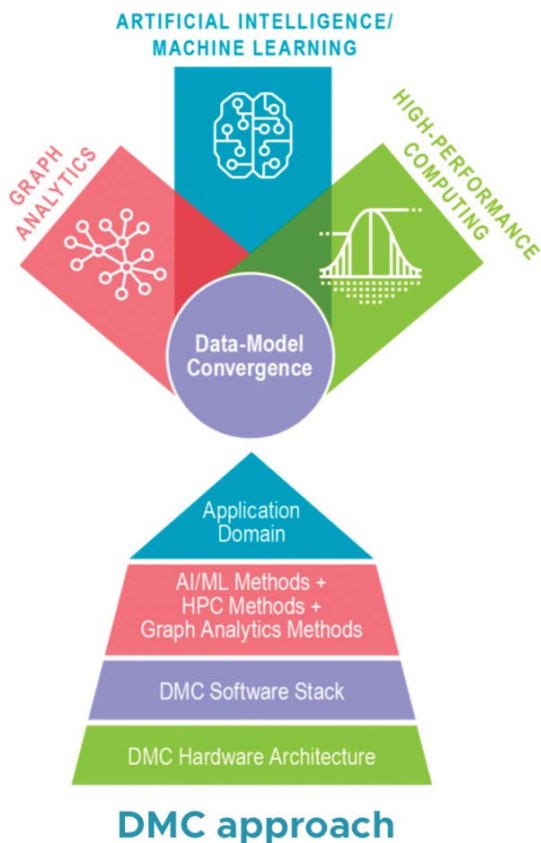


Figure 4.15: Data Model Convergence (DMC) approach<sup>24</sup>  
(courtesy of Mark Raugas / PNNL)

## Confidential computing

Confidential computing is an important area of exploration, and various technology solutions need to be tested for delivery of secure computing environments<sup>25</sup>. **Current megatrends that are driving the importance of confidential computing include the proliferation of cloud computing, the growth of AI and analytics, and the emergence of the network and edge.** Cryptography is an essential component in providing security for data at rest and data in motion, and homomorphic encryption is being discussed for data in use. How can data be protected when it is actively in use in computing environments that are not in our direct control? *Trusted Execution Environments (TEEs) based on separation (i.e., physical, temporal, logical, cryptographic), is a promising pathway to confidential computing, which requires significant new research efforts<sup>25</sup>.*

## The future of secure systems

There are new security challenges associated with emerging computing technologies. For example, FPGAs have been growing in use for cloud computing. This calls for better understanding of their vulnerabilities, especially

those associated with multi-tenant FPGAs, i.e., where the optimization of compute resources has driven the use of FPGAs by multiple users at the same time<sup>26</sup>.

Possible types of the cloud FPGA attacks include on-chip voltage sensors to extract encryption key, temperature residue attacks, on-chip voltage supply attacks, and crosstalk-based attacks (information extraction from adjacent wires). *Remediation approaches to address them include a regular rectangular grid of voltage sensors, bitstream and netlist scanners, power throttling, and others.*

Overall, the system security vulnerability increases together with the compute performance as shown in Figure 4.16. **This figure also shows an alarming trend that, in recent years, the growth rate of the security vulnerabilities has become greater than the performance<sup>27</sup>.** The “patch and pray” model, commonly used today, is unsustainable given the growing landscape and complexity of system-level security. One of the approaches to identifying and mitigating entire classes of vulnerabilities through hardware architectures is the ongoing DARPA System Security through Integrated Hardware and Firmware (SSITH) program<sup>27</sup>.

## Key areas of focus and follow-on research

**Addressing system-level security brings an even greater challenge to the semiconductor industry, as all levels of system hardware and all lifecycle stages (from sourcing and design to manufacturing and deployment) need to be considered.** Protecting both the infrastructure/enterprise and the data in today’s complex edge-to-cloud environment will require research in a number of key focus areas:

- The contribution (and timing) of technologies, such as homomorphic encryption and blockchain to address system level security challenges
- Innovative techniques to leverage system heterogeneity to improve system-level security
- Progression of tools and methodologies to provide system-level attestation and trust
- Specific challenges in securing IoT devices
  - Implementing a systematic approach to update devices amid an ever-changing threat landscape
  - Innovative methods and technologies specific to IoT/ constrained device protection that go beyond digital data encryption to validate sensor data via the software stack and through the overall system design
- The compatibility and relevance of academic research with industry focus





Figure 4.16: Performance and security trends<sup>27</sup> (Courtesy of Keith Rebelló / DARPA)

## 4.7. Summary—New Trajectories for Hardware-enabled ICT Security

The pace at which today’s systems are increasing in intelligence and ubiquity is astounding. At the same time, the increased scale and complexity of these systems have forced hardware specialization and optimization to address performance challenges. All these advances in capability must go hand in hand with advances in security and privacy. Examples include securing weaknesses in the machine-learning or conventional cryptography, protecting privacy of personal data, and addressing vulnerabilities in the supply chain or hardware.

**Security Grand Goal:** Develop security and privacy advances that keep pace with technology, new threats, and new use cases. Examples include trustworthy and safe autonomous and intelligent systems, secure future hardware platforms, and emerging post-quantum and distributed cryptographic algorithms.

### Research recommendations summary

#### Security and safety of autonomous systems

- Building end-to-end security for highly connected autonomous systems, HW RoT and security primitives, and trusted data
- Minimizing security threats at the root by building secure smaller (small code size) components and constructing complex systems from these secure components
- Dynamic assessment and adaptation against failures and threats for longer durability and ongoing security of these systems
- Desired level of security in a system is highly dependent on the expected lifetime of a system.
- Using AI and ML to find appropriate solutions to detect, analyze, and adopt security threats
- Finding alternatives to current cryptography techniques that are under threat due to quantum computing

## Secure hardware design

- Improved security of billions of very-low-cost and easy-to-implement devices
  - While the design of mission-critical systems carefully considers security and attempts to address it while absorbing necessary costs, many IoT devices are driven by cost and quick/easy consumer adoption. They do not have even the most rudimentary security defenses.
- Hardware futureproofing
  - Hardware is generally inflexible. Some systems built on a given HW may remain in the field for over a decade. Security attacks, however, continue to evolve.
- "Security maturity" signoff of HW design before its tape out
  - HW design has advanced to have systematic methods to evaluate it for manufacturability, reliability, testability, etc.
- Vulnerability of the NVM systems to hardware attacks
- Methods to prevent/mitigate IP piracy at foundries
- Ways to synthesize RTL for guaranteed security against known attacks

## Secure computing

- Building single-chip secure processors
  - Open source, formally verified TCB
  - Secure against all practical SW attacks
  - Secure against physical attacks of memory
  - Enhanced physical security against invasive attacks
- Minimizing performance overhead
  - Interplay of security, performance, usability, and complexity creates a very challenging problem.
- Developing threat models
  - What actual threats need to be mitigated? How effective are they in practical scenarios?
- Documenting HW/SW 'contract' for microarchitectural properties around security
  - How can SW reason and express the security property it desires?
  - How can HW innovate on microarchitectural design to improve performance and power?

## Hardware for cryptography

- Crypto-diversity, lightweight crypto with side-channel security, low-latency crypto
- More than Moore: Post-quantum computing cryptos, including secure implementation as through masking
- Hardware-entangled crypto, which has minimum Root of Trust through PUF (Physically Unclonable Function), and TRNG (True Random Number Generator)
- Homomorphic encryption

## AI security and privacy research

- Building defenses for all parts of the machine-learning pipeline (e.g., model-stealing, training-set poisoning, and adversarial), as they can all be attacked
- Solving deepfakes, an important unsolved problem with deep societal implications
- Merging of cryptography and artificial intelligence critical for user privacy
- Algorithms such as Fully Homomorphic Encryption require advances in hardware, storage, and communication to be fully practical.
- Learning from cross-organizational data that cannot be shared
- Algorithms and co-optimization are needed across the hardware, software, and network.

## System-level security

- The contribution (and timing) of technologies, such as homomorphic encryption and blockchain to address system-level security challenges
- Innovative techniques to leverage system heterogeneity to improve system-level security
- Progression of tools and methodologies to provide system-level attestation and trust
- Specific challenges in securing IoT devices
  - Implementing a systematic approach to update devices amid an ever-changing threat landscape
  - Innovative methods and technologies specific to IoT/constrained device protection that go beyond digital data encryption to validate sensor data via the software stack and through the overall system design
- The compatibility and relevance of academic research with industry focus

## Contributors

|   |   |                                |                                     |
|---|---|--------------------------------|-------------------------------------|
| Fari Assaderaghi<br>(NXP, Sunrise Memory) | Kevin Fu (U. Michigan)                    | Hungwen Li (MediaTek)          | Mark Raugas (PNNL)                  |
| Rosario Cammarota (Intel)                 | Doug Gardner (Analog Devices)             | Yiorgos Makris (UT Dallas)     | Keith Rebelló (DARPA)               |
| Matthew Campagna (Amazon)                 | Gilbert Herrera<br>(Sandia National Labs) | Stan McClellan (Texas State U) | Carlos Rozas (Intel)                |
| Ramesh Chauhan (Qualcomm)                 | Yousef Iskander (Microsoft)               | Len Orlando (AFRL)             | Mark Tehranipoor (U. Florida)       |
| Lily Chen (NIST)                          | Somesh Jha (UW Madison)                   | Miroslav Pajic (Duke U)        | Russel Tessier<br>(U Massachusetts) |
| Richard Chow (Intel)                      | Marc Joye (Zama)                          | Ron Perez (Intel)              | Ingrid Verbauwhede<br>(KU Leuven)   |
| Debra Delise (Analog Devices)             | Farinaz Koushanfar<br>(UC San Diego)      | Raluca Popa (UC Berkeley)      | William Whyte (Qualcomm)            |
| Srini Devadas (MIT)                       |   | Josyula Rao (IBM)              |                                     |

## References to Chapter 4

- <sup>1</sup>Yiorgos Makris, "Security and Trust in the Analog/Mixed-signal/RF Domain", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>2</sup>P. Len Orlando, "Securing systems for future defense applications", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>3</sup>Josyula Rao, "Computing for Threat Intelligence", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>4</sup>Kevin Fu, "Embedded Security: Protecting Analog Sensor Cybersecurity", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>5</sup>Lily Chen, "Cryptography in the Quantum Era", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>6</sup>Matthew Campagna, "Quantum-safe Solutions for Cloud Computing", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>7</sup>Doug Gardner, "End Point Security in the Hyperconnected Future", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>8</sup>William Whyte, "Security in Autonomous Vehicles—Future Trends", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>9</sup>Hungwen Li, "Securing Automobile System by Hardware-Intensive Virtualization", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>10</sup>Stan McClellan, "Smart Grid Security", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>11</sup>Miroslav Pajic, "Autonomous System Security", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>12</sup>Mark Tehranipoor, "Hardware Security", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>13</sup>Srini Devadas, "High-performance Secure Processors", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>14</sup>Carlos Rozas, "Secure Computation for AI", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>15</sup>Ingrid Verbauwhede, "Cryptography Hardware", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>16</sup>A. Antonopoulos, et al, "Trusted Analog/Mixed-Signal/RF ICs: A Survey and a Perspective," IEEE Design & Test of Computers (D&T), vol. 34, no. 6, pp. 63-76, 2017
- <sup>17</sup>K. Subramani, et al, "Trusted and Secure Design of Analog/RF ICs: Recent Developments," International On-Line Test Symposium (IOLTS), pp. 125-128, 2019
- <sup>18</sup>Somesh Jha, "Machine Learning and Security", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>19</sup>Rosario Cammarota, "Secure Computation for AI", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event)
- <sup>20</sup>Marc Joye, "End-to-end encryption", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).
- <sup>21</sup>Raluca Popa, "Secure Collaborative Learning", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).
- <sup>22</sup>Farinaz Koushanfar, "Distributed and Federated Learning on Encrypted Data", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).
- <sup>23</sup>Yousef Iskander, "Securing Platform Integrity and Trust with Hardware Security", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).
- <sup>24</sup>Mark Raugas, "System Security in an Era of Heterogeneity", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).
- <sup>25</sup>Ron Perez, "Confidential Computing", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).
- <sup>26</sup>Russell Tessier, "Challenges and Solutions with Emerging Computing Technologies", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).
- <sup>27</sup>Keith Rebelló, "The future of secure systems", SRC workshop on ICT Hardware Enabled Security, Aug. 25-27, 2020 (Virtual Event).





---

## Chapter 5

# New Compute Trajectories for Energy-Efficient Computing

### Seismic shift #5

Ever-rising energy demands for computing versus global energy production are creating new risk, but new computing paradigms offer opportunities with dramatically improved energy efficiency.

### 5.1. Executive Summary

Rapid advances in computing have provided increased performance and enhanced features in each new generation of products in nearly every market segment, whether it be servers, PCs, communications, mobile, automotive, entertainment, among others. These advances have been enabled with decades of R&D investments by both the private sector and the government, yielding exponential growth in compute speed, energy efficiency, circuit density, and cost-effective production capability. Sustained innovation

in software and algorithms, systems architecture, circuits, devices, and semiconductor process technologies have been foundational to that growth pace. Although this trend has persisted for decades by successfully overcoming many technological challenges, **it is now recognized that conventional computing is approaching fundamental limits in energy efficiency and, therefore, presents challenges that are much harder to surmount.** Consequently, disruptive innovations in information representation, information processing, communication, and information storage are all pressing and critical to sustainable economic growth and United States technological leadership.

*The number of information bits being processed and the number of computations per year continues to increase unabated, and it is projected that in 2050 we will be dealing with  $10^{42}$ – $10^{46}$  bits (see Appendix).* As shown in Figure 5.1a, the total energy consumption by general-purpose computing continues to grow exponentially and is doubling



approximately every three years, while the world's energy production is growing only linearly, by approximately 2% a year. The rising global compute energy is driven by ever-growing demands for computation (Figure 5.1b), and this is in spite of the fact that the chip-level energy per one-bit transition in compute processor units (e.g. CPU, GPU, FPGA) has been decreasing over the last 40 years (as manifested by Moore's law), and is ~10 attojoules or  $10^{-17}$  J in current processors. However, **the demand for computation growth is outpacing the progress realized by Moore's law.** In addition, Moore's law is currently slowing down as device scaling is approaching fundamental physical limits. If the exponential growth in compute energy is left unchecked, market dynamics will limit the growth of the computational capacity, which would cause a flattening of the energy curve (the "market-dynamics-limited" scenario in Figure 5.1a). Thus, a radical improvement in energy efficiency of computing is required to avoid the limiting scenario.

**The underlying technical challenge is bit-utilization efficiency in computation**, i.e., the number of single bit transitions needed to implement a compute instruction. The current CPU compute trajectory is described by a power formula (shown as inset in Figure 5.2) with an exponent  $p \sim \frac{2}{3}$ . The theoretical basis for the observed trajectory and for the value of the exponent is not clearly understood, so **the theoretical basis for computation needs further research.** As an observation, if it is possible to increase the exponent in the formula by only ~30%, the compute efficiency, and thus energy consumption, would have a 1,000,000x improvement. This is illustrated as "new trajectories" in Figure 5.2.

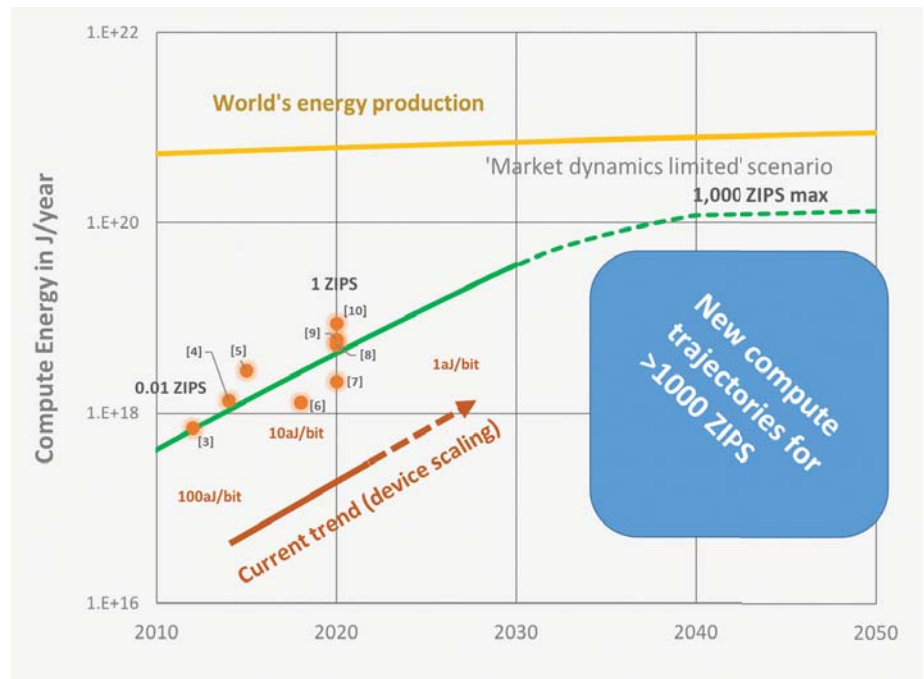


Figure 5.1a: Total energy of computing (see Appendix for details): The solid green line indicates continuing the current computing trajectory while improving the device energy performance. The dashed green line indicates a "market-dynamics-limited" scenario, stopping further increase in the world's computing capacity and resulting in a flattening of the energy curve. The blue box indicates a scenario where a radically new computing trajectory is discovered. The Decadal Plan model (green line) is compared to independent data by different groups (circled dots).

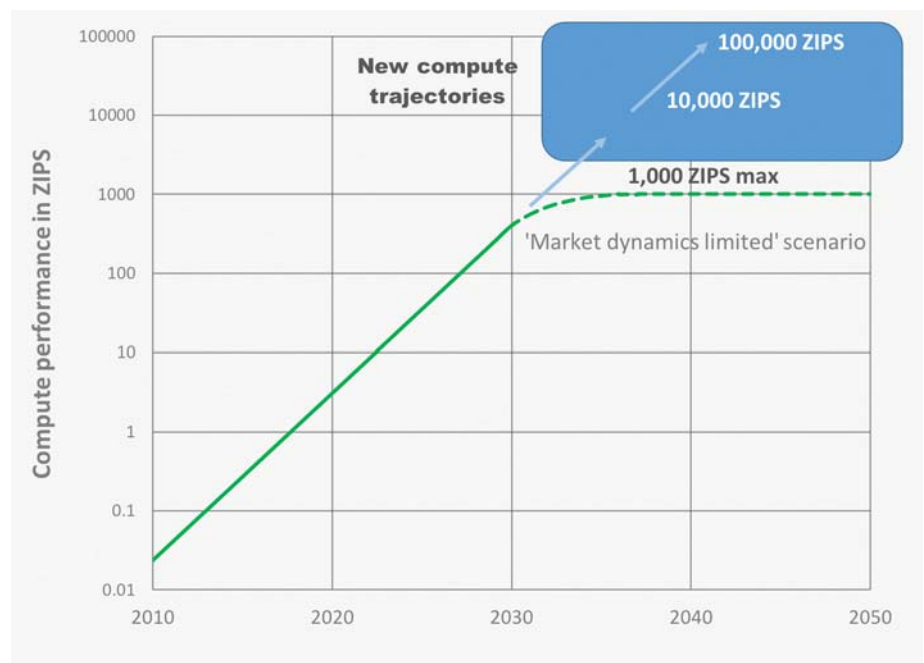


Figure 5.1b: World's technological installed capacity to compute information, in ZIPS, for 2010-2050. The solid green line indicates the current trend (see Appendix). The dashed green line indicates a "market-dynamics-limited" scenario, stopping further increase in the world's computing capacity due to limited energy envelope. The blue box indicates a scenario where a radically new computing trajectory is discovered.

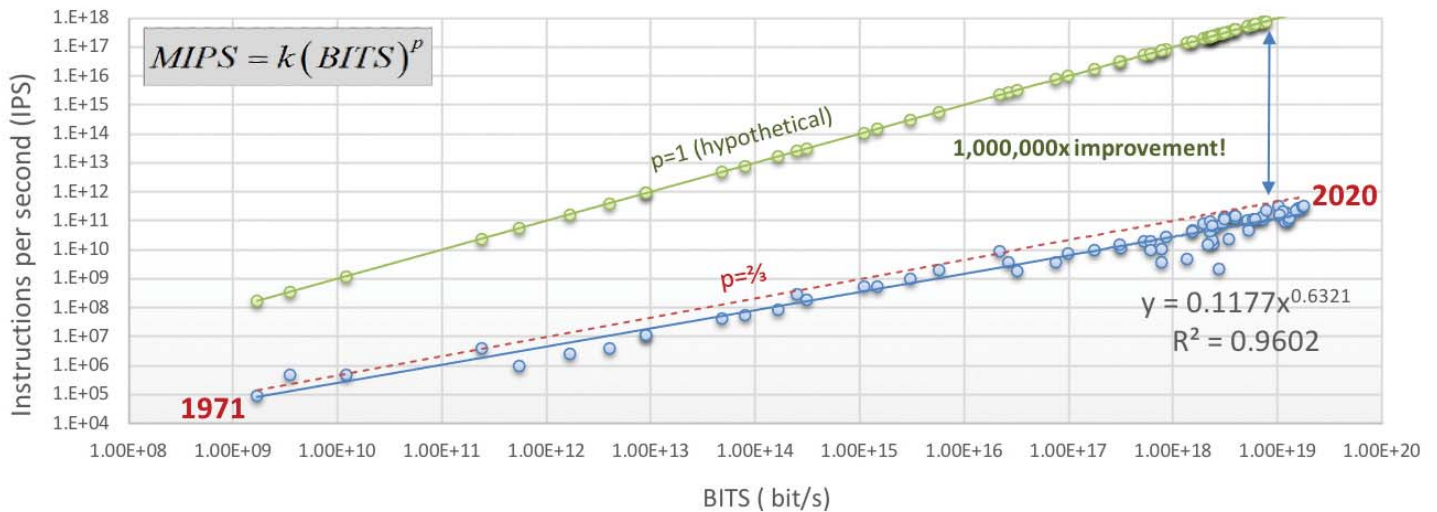


Figure 5.2: The current CPU compute trajectory is described by a power formula (shown as inset) with an exponent bounded by  $p \sim \frac{2}{3}$ . The theoretical basis for the observed trajectory and for the value of the exponent is not clearly understood, so the theoretical basis for computation needs further research. As an observation, if it is possible to increase the exponent in the formula by only  $\sim 30\%$ , the compute efficiency, and thus energy consumption, would have a 1,000,000x improvement. This is illustrated as “new trajectories”.

## Call for action

Revolutionary changes to computing will be required soon.

**Computational loads continue to grow exponentially, as evidenced by the growth in artificial intelligence (AI) applications and training demands.** New approaches to computing, such as *in-memory compute, special purpose compute engines, different AI platforms, brain inspired/ neuromorphic computation, quantum computing, or other solutions, will be necessary and will need to be combined in a heterogeneous manner.* The scope of potential heterogeneous computing architectures is described in a recent National Science and Technology Council (NSTC) report on the Future of Computing<sup>1</sup>. This research will require a cross-disciplinary, cross-functional approach to realize commercially viable and manufacturable solutions with multi-decade longevity to replace the mainstream digital approach. This document is intended to stimulate multilateral collaborative research “from materials to architecture and algorithms” to establish revolutionary paradigms that support future energy-efficient computing for the vast range of future data types, workloads, and applications. For additional background, see the DOE Office of Science, Basic Research Needs for Microelectronics workshop report<sup>2</sup>.

The **Computing Grand Goal** is to discover computing paradigms/architectures with a radically new computing trajectory, demonstrating  $>1,000,000x$  improvement in energy efficiency. Changing the trajectory not only provides immediate improvements but also provides many decades of growth potential (as shown in Figure 5.1). This would be much more cost-effective than attempting to dramatically increase the world’s energy supply.

**Invest \$750M annually throughout this decade to alter the compute trajectory. Selected priority research themes are outlined below.<sup>1</sup>**

This chapter addresses fundamental ICT capabilities and limits, and it describes outcomes from our open forum for brainstorming new computing applications and their corresponding implications for the semiconductor industry. Furthermore, emerging architectures were discussed in the context of the application space they enable and the corresponding trend lines in energy, storage, or communication that they can dramatically alter.

<sup>1</sup>The Decadal Plan Executive Committee offered recommendations on allocation of the additional \$3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market-share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

## 5.2. Energy-Efficient Computing: Fundamentals, Challenges, and Application Drivers

### Overview and needs

ICT systems support daily social life, scientific discoveries, advances in healthcare, engineering innovations, and economic activities. ICT has and will continue to contribute greatly to the global economy. Applications are only limited by the computational power that can be delivered by today's technology. While we may not be able to predict how ICT will evolve, we know that market dynamics and energy constraints will require a shift to a new compute trajectory enabled by the research and development strategies that are the goal of this SRC Decadal Plan.

These strategies are informed and shaped by an understanding of fundamental limits in energy, storage, and communication based on trends from widely exploited general-purpose computing platforms, such as CPUs, GPUs, FPGAs, etc. Emerging nonconventional computing architectures will also need to be evaluated against established trends to identify their potential impact.

**Zettascale computing ( $10^{21}$  FLOP/s) may be the first milestone that we are not able to reach in a reasonable and predictable timeframe with existing technologies, as compared to prior mega-, giga-, tera-, peta-, and exa-scale milestones<sup>11</sup>.** Thus, attempting to build a general-purpose zettascale computing system by 2030 to adeptly address the needs of a broad collection of the world's most challenging problems is not currently possible. One example of the

future computational challenges is the accurate simulation of the X-51A recoverable hypersonic vehicle over a one-hour flight path (Figure 5.3). This challenge could generate as much as an exabyte of data and requires tightly integrated artificial intelligence, data analytics, and high-end physics capabilities due to the complex aerodynamic structure, challenging subsonic, supersonic, and hypersonic combustion effects (from rest to full speed), and extreme material property variations attributable to non-linear heat injection throughout the flight path.

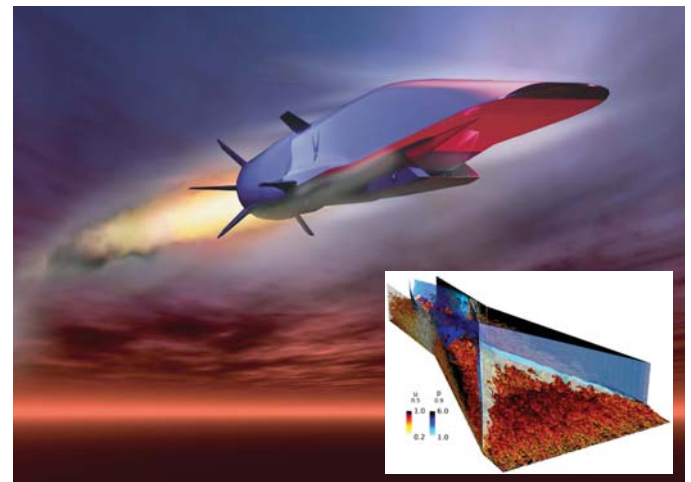


Figure 5.3: Artist rendering of the X-51A recoverable hypersonic vehicle with a companion computational fluid dynamics simulation (by Michael C. Adler and Datta V. Gaitonde from The Ohio State University, using DoD HPCMP resources)<sup>12</sup>

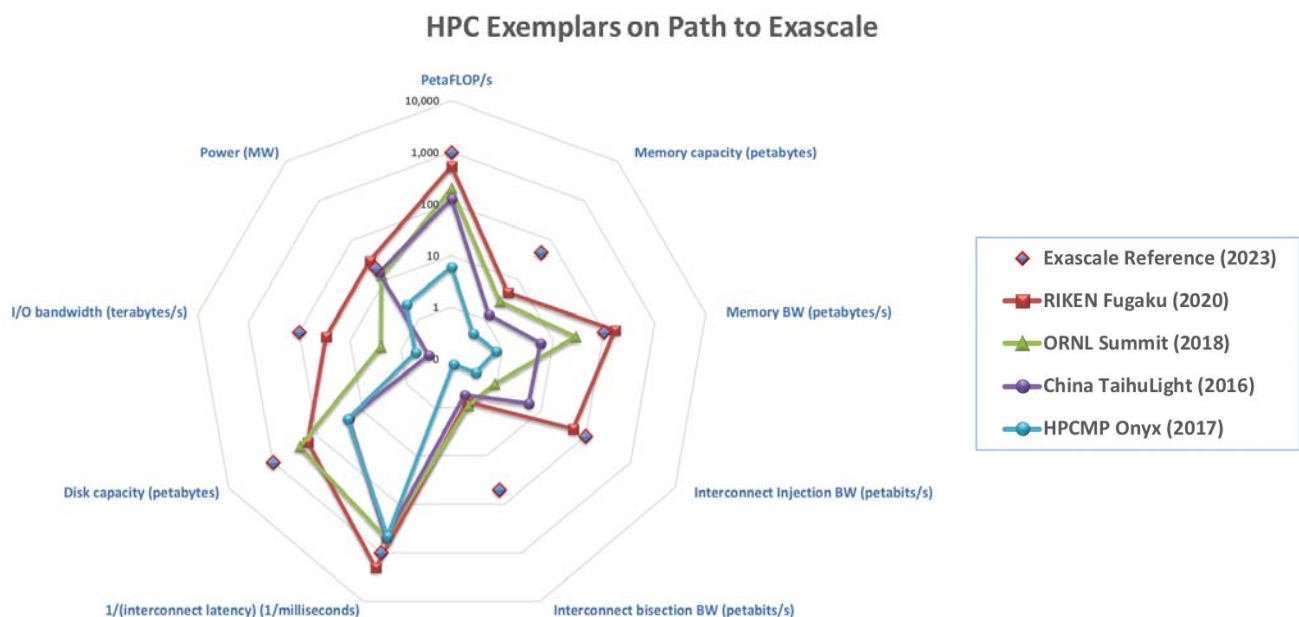


Figure 5.4: Dimensions of high-end systems performance<sup>11</sup> (courtesy of Roy Campbell, DoD and James Ang, PNNL)

As shown in **Figure 5.4**, **some dimensions of high-end systems performance noticeably lag behind the needs, in particular interconnect and I/O bandwidth, as well as memory capacity.** In addition, while we have greatly benefited from shrinking the feature size of CMOS transistors (a dominant driver for growth in the FLOP/s axis), *we anticipate exhausting this trend in less than 10 years, as smaller CMOS feature sizes approach the fundamental physics limits of scaling/shrinking devices.* We will, therefore, be forced to consider concepts other than CMOS shrinking (e.g., 3D stacking) to realize sustainable technological improvements. These concepts, however, will only provide temporary relief. As the size and complexity of our problems increase, *we will not only have to reconsider our use of broadly applicable systems but will also have to consider alternatives to general-purpose von Neumann processors at the core of these systems.* By 2030, we anticipate an era of specialized computing (perhaps one system type per problem class), systems with reconfigurable, heterogeneous computing processors, or processors that employ dataflow constructs in lieu of historical control-flow methods. Many of these advanced architecture concepts represent a substantial departure from today's high-end computers, and thus will be developed through an application-driven hardware/software co-design approach.

The commercial world embraced data-centric computing to develop and apply a broad range of machine-learning and deep-learning methods to detect, recognize, classify, and train on large data sets collected from IoT devices, including sensors, smart-home devices, cameras, etc. This shift toward data-centric needs rather than compute-centric needs is also recognized by the DOE scientific research community for both computational and experimental data sources.

What will computers look like beyond Moore's Law? This is still an open question<sup>13</sup> that has many facets. **The theoretical basis for computing performance is less solid than the theoretical basis for information storage and communication, like the Shannon limit and others.** New metrics for measuring performance are needed that would account for widespread use and the fundamental limits of accelerators<sup>13</sup>. On the elemental level, new device structures and aggressive introduction of new materials for the development of chips beyond silicon can be expected<sup>14</sup>. **Improvements in hardware alone will not be enough to handle future needs. Sustained innovation in software and algorithms is one of the critical future targets<sup>15</sup>.** The DOE Advanced Scientific Computing Research (ASCR) program hosted a scientific-community workshop on Productive Computational Science in the Era of Extreme Heterogeneity that defines basic computer-science research needs and

opportunities to develop operating systems, runtime systems, programming environments, and tools to support heterogeneous computing<sup>16</sup>.

## 5.3. Impact of Emerging Device Technologies

### Overview and needs

Radically new solutions are needed for future ICT, with major innovations in devices, circuits, and architectures. Circuits could be digital, analog, or hybrid. *New device physics needs to be explored that includes spintronics, nanophotonics, nanoionics, superconductivity, and other phenomena for emerging computing technologies.* Also, a new conceptual abstraction is needed for emerging information representations and associated novel processor architectures. Possible alternate computing models include, for example, *high-dimensional computing, analog "approximate computing," and others. A juxtaposition or hybridization of computing models may drive the development of neuromorphic or brain-like architectures, which could be a significant component of future compute systems.* New architectural research congruent with novel semiconductor technology research will be increasingly important to achieve platform-capable, self-consistent, and optimal system solutions having enhanced performance and energy-efficiency with minimum added complexity. For example, the role of future transistors, interconnects, and memories is critical for emerging architectures. However, in addition to assessing the potential benefits of those emerging technology concepts on conventional architectures, we also need to look at the reverse: *new architectures driving semiconductor technology requirements, e.g., extending the operating space not otherwise attainable under conventional architectures through error-resilient computing architectures.* We must discover how new architectures can take full advantage of emerging technology components or interconnect fabrics to create disruptive, platform-capable solutions that might permeate multiple application segments with no or minimal added complexity.

### Shannon models of computing

As traditional deterministic solutions to computing are reaching their limits, *new nondeterministic, "accurate-enough" methods are being considered to build systems that can cope with and/or exploit inherent variability or device stochastic characteristics for performance-power-area benefits<sup>17</sup>.* Nanoscale logic circuit fabrics can be treated as noisy communication channels on which inference type machines are built (statistical information processing), so that information is processed reliably (**Figure 5.5**).



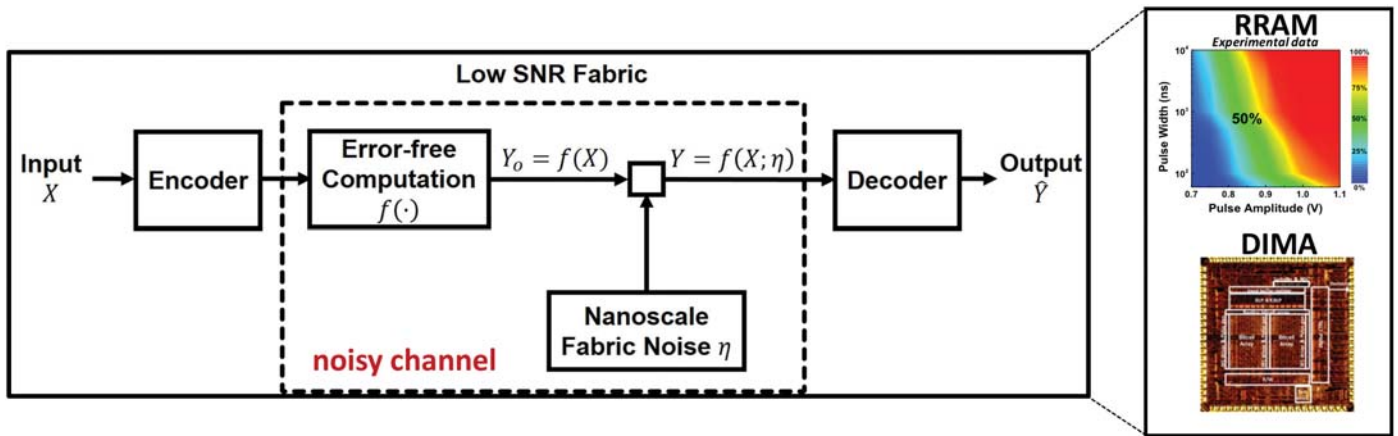


Figure 5.5: Shannon-inspired model of computing<sup>17</sup> (courtesy of Naresh Shanbhag, University of Illinois at Urbana-Champaign)

*The Shannon-inspired probabilistic computational models* may enable not only increased levels of network reliability despite poor logic gate-level reliability, but might also offer opportunities for the design of energy-efficient computing architectures employing system-level considerations<sup>17,18</sup>. This approach would be similar to the design of communication links, which considers an overall minimization of bit-errors by employing error control coding and channel estimation. Just as these links are able to operate at their fundamental limits on channel capacity, the Shannon model of computing might enable the design of computing systems that operate at their fundamental limits on energy efficiency.

A Shannon-based computing model (Figure 5.5) comprises the use information-based metrics, plus the design of low Signal-to-Noise Ratio (SNR) circuit fabrics (such as in-memory architectures, spintronic logic, voltage overscaled circuits) and development of Statistical Error-Compensation (SEC) techniques (such as Algorithmic Noise Tolerance, Stochastic sensor NOC, Soft NMR, Likelihood processing)<sup>18</sup>. This model strives to shape the error statistics of nanoscale devices seen

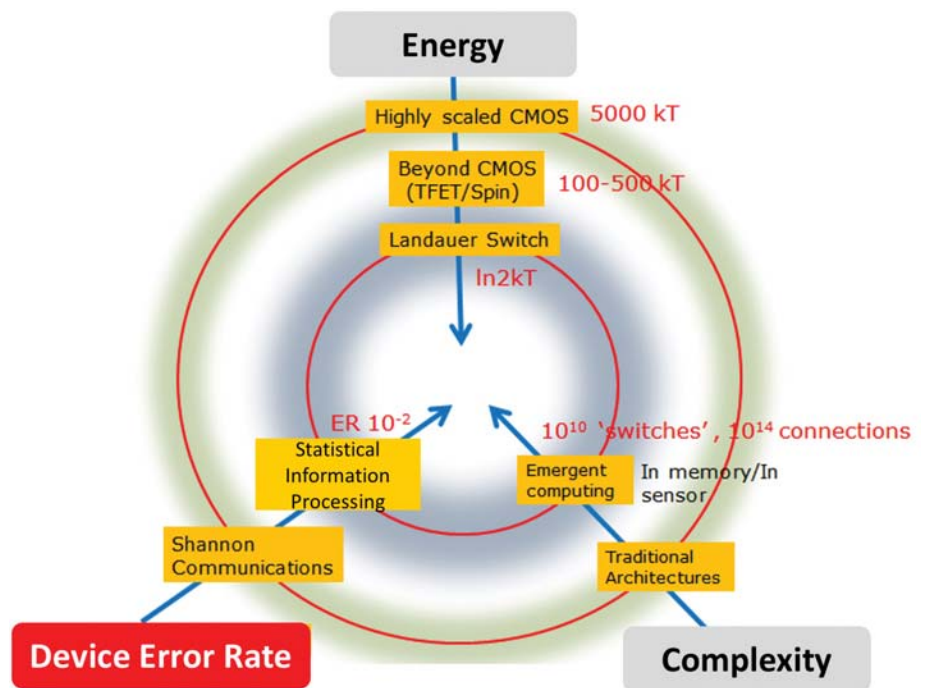


Figure 5.6: An integrated approach for reducing energy and device error rate and increasing architectural complexity, employing both Shannon-inspired and traditional computing models<sup>17</sup> (courtesy of Naresh Shanbhag, University of Illinois at Urbana-Champaign, S. Manipatruni, D. Nikonov and I. Young, Intel).

at the system level via circuit, logic, and architectural methods in order to realize error compensation with minimal overhead. This can be applied, for example, to spintronics-based logic to enhance energy efficiency<sup>17</sup>. One envisions an integrated approach that

combines traditional and Shannon-based computing with emerging devices in order to design computing systems of the future that operate at the limits of energy efficiency and reliability (Figure 5.6).

## High-dimensional computing

*Information representation by high-dimensional (HD) vectors* having identical and independently distributed random components has been shown to be inherently one-shot, continuous learning, and capable of high-level reasoning with unsurpassed representation error tolerance. High-dimensional vector spaces enable the memory-efficient representation of very large numbers of distinct or similar objects or sets of them. The associated computing engine relies essentially on an associative memory and an arithmetic unit supporting three component-wise arithmetic operations, which are intrinsically parallel and local, thus suitable for tightly integrated logic and memory. The fundamental error tolerance of HD computing, inherent parallelism and locality, and one-pass/continuous learning capabilities may also allow for simultaneous breakthroughs in energy efficiency.

## Selected examples of emerging devices

*Photonics* presents significant opportunities to address bandwidth bottlenecks in data movement for high-performance computing systems. Novel high-speed, low-power transceivers, light sources, waveguides/modulators, and photodetectors represent some of the essential building blocks for integrated photonics—a field where sustainable dimensional scalability remains a key challenge. Beyond photonics as a potential interconnect fabric, photonic devices have also been shown to enable certain classes of mathematical operations like matrix vector multiplication as illustrated in **Figure 5.7**. Overall system density and energy efficiency represent a key research challenge to overcome, compared, for example, with alternative digital or mixed-signal systems for AI applications.

## Cryogenic computing

*Cryogenic electronics* holds promise for high-performance computing, as CMOS devices at a very low temperature (e.g., 6K) have Energy-Delay characteristics that are six times better than at 300K. Moreover, the transistors subthreshold slope gets three times steeper, the carrier mobility within the devices improves, and the resistivity of interconnects is reduced<sup>20</sup>. To take advantage of the CMOS performance boost at cryogenic temperatures, new gate-stack and interconnect technologies need to be deployed<sup>21</sup>.

## Spin logic

*Spin logic devices*, one of the alternatives to CMOS transistors, use magnetic field orientation to represent information. It is suggested that total power consumption for spintronic devices can be less than CMOS, while retaining high throughput (**Figure 5.8**)<sup>22</sup>. *Magnetolectric spin-orbit (MESO) devices*<sup>23</sup> can have multiple input current levels, which make them a potential candidate for analog circuits. Low-voltage spintronics can demonstrate stochastic switching activity and can be used to implement Shannon-based computing. Also, spintronic devices can mimic the functionality of spiking neurons.

## Improving materials to sustain the compute trajectory

The driving factors for ICT design are power, performance, area, and cost. As critical dimensions approach atomic length scales, there are not only lithographic challenges, but also other fundamental ones associated with the patterned features fidelity and mechanical integrity, as well as the related enhanced physical and electrical variability. A compendium of current areas of design thrust and the material processes being explored are shown in **Figure 5.9**<sup>24</sup>.

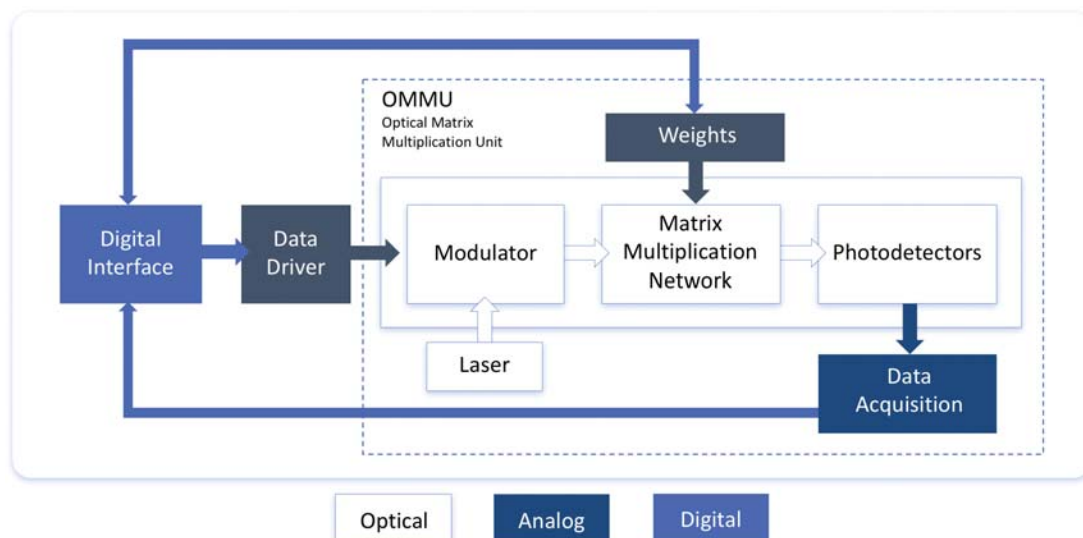
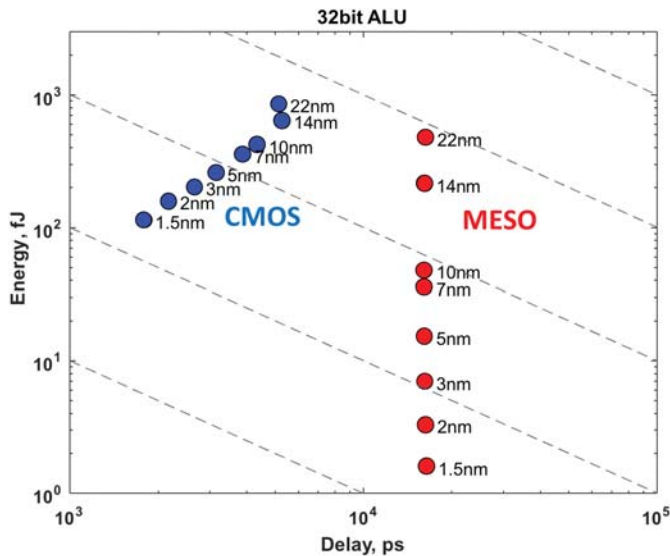


Figure 5.7: Schematic of optical computing system<sup>19</sup> (courtesy of Yichen Chen, Lightelligence)

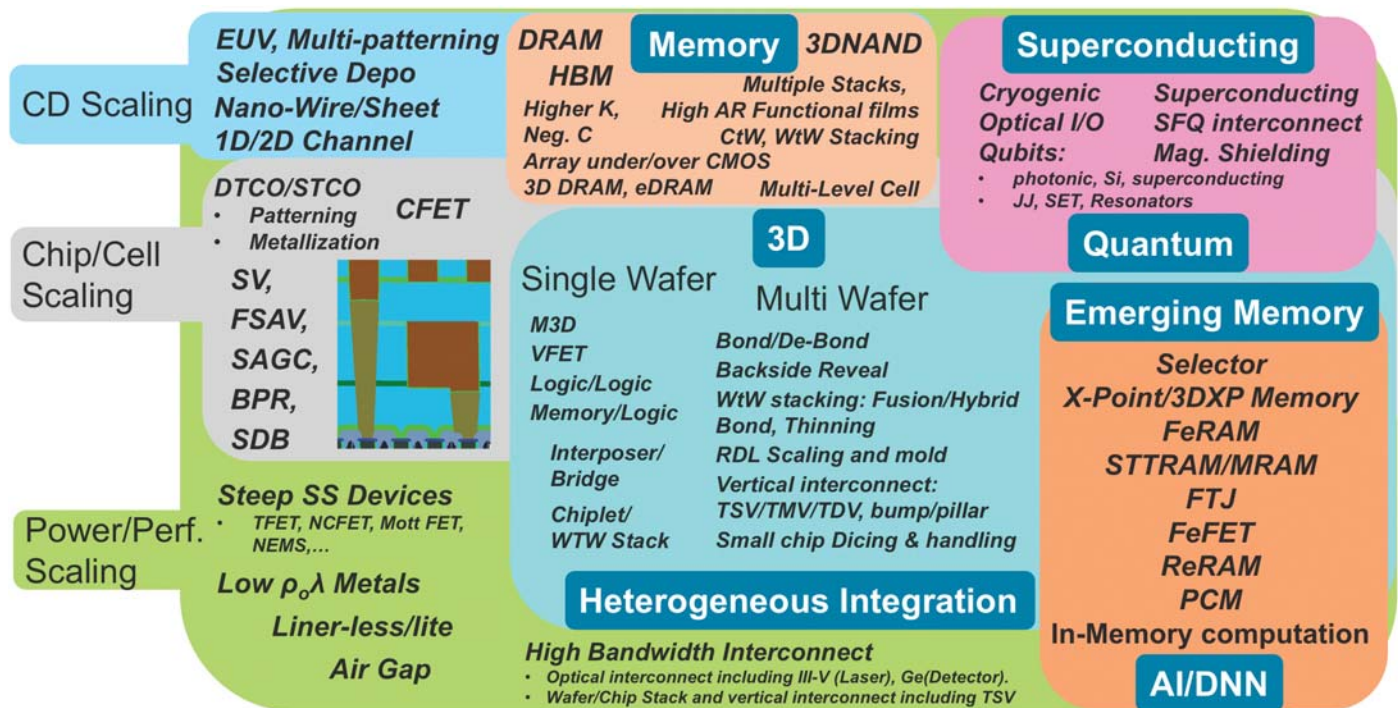


CMOS possesses higher switching speed that continues to improve with scaling.  
MESO devices energy scales better, but speed is limited by the magnet.

Figure 5.8: Energy-delay characteristics of logic circuits implemented with magnetolectric spin-orbit (MESO) devices as compared with CMOS<sup>22</sup> (courtesy of Dmitri Nikonov and Ian Young, Intel. Corp.)

### Key areas of focus and follow-on research

Reaching new highs for deterministic high-performance computing capabilities and AI cognitive capabilities, along with orders of magnitude improvements in energy efficiency on such workloads, will require significant research investments on multiple but synergistically interlocked areas. Needed research areas include new types of information representation and processing, new compute paradigms like high-dimensional computing or Shannon-inspired statistical computing, and related research on generalized processor engines. Those areas explore new building blocks and interconnect fabrics that press against the fundamental limits of advanced logic and memory technologies and their operating conditions. Device (transistor and memory) research must focus on enabling added technology functionalities while minimizing complexity. This would include *charge-based and non-charge-based devices (e.g., spintronics, photonics, and others), as well as new materials and processes to related logic-and-memory building blocks and dense, high-speed, and low-energy interconnect fabrics.*



R. Clark / TTCA TFPT / October 8, 2019

Source: TEL



Figure 5.9: Drivers and technologies for better power, performance, area, and cost Scaling<sup>24</sup> (courtesy of Robert Clark, Tokyo Electron)



## 5.4. Artificial Intelligence and Brain-inspired Computing

### Overview and needs

*Artificial Intelligence (AI) is expected to be one of the next major disruptive technologies.* It will affect the way we access and analyze information, the way we teach, and the way we learn or enhance our knowledge. However, **the current state of AI is characterized by an extensive use of high-performance computational resources and by large memory footprint, compute load, and energy cost.** For example, today's high-performance GPU-based object-recognition engines require 1-10 J/frame, which results in system power consumption of tens to hundreds of watts. Also, today's inference machines often require tens of GB of local memory. As a result, **many AI applications are bounded to datacenter environments, that are less resource-constrained, than network edge computing resources.** Likewise, specialized systems that support critical applications, such as autonomous driving, require interaction and communication of multiple components, which approach datacenter complexity. As we proceed with more advanced AI engines and applications, **resources and energy consumption may become prohibitive at both the datacenter and at the network edge.**

Many corporations in the communications and sensors space have now invested in the seamless integration of AI into ubiquitous computing. Applications ranging from Internet of Things (IoT) to massive Machine Type Communication (mMTC) in 5th generation wireless communication (5G) suite are immensely demanding in terms of computation and networking resources (see Chapter 3). To this end, the physical design of High-Performance Computing (HPC) systems is evolving to enable high data-transfer rates to serve researchers' and organizations' with requirements to train models that combine simulations with the vast influx of digital data. The fusion of AI and HPC is made possible by availability of ever-increasing data. This is summed up in **Figure 5.10.**

*The brain provides a wide range of complex cognitive capabilities that would be invaluable to implement in a computing system,* and it does so rapidly and at extremely low power, despite operating at the relatively slow timescales of organic material. The brain achieves these benefits by using a combination of unique neural algorithms, a configurable and adaption-compatible architecture that relies on event-based communication, and device "technology" (neurons and synapses) that is analog in behavior, configurable, and three-dimensional.

A convergence or integration of computing models may drive the development of *neuromorphic or brain-like architectures,*

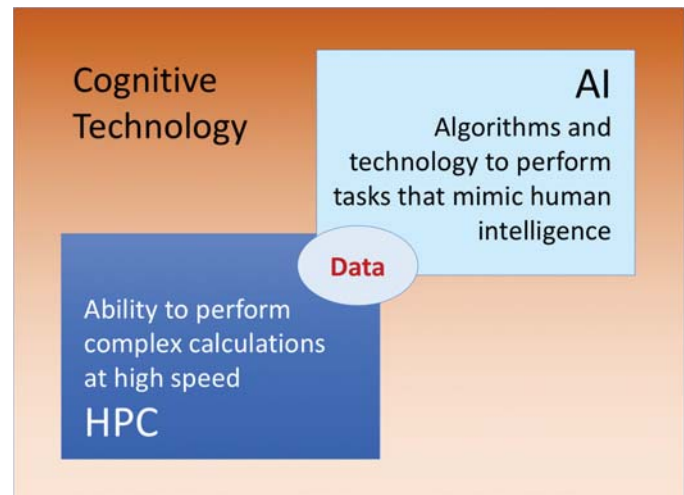


Figure 5.10: The fusion of Artificial Intelligence and High-Performance Computing is made possible by an increased production of data<sup>25</sup> (courtesy of Fred Streitz, DOE/AITO).

which are expected to be a significant component of future compute systems. *Living systems like the human brain can be viewed as an information processor that is extraordinarily efficient in the execution of its functions, consequently it is a good reference for new information-processing technologies.*

**Figure 5.11** lays out a convenient roadmap<sup>26</sup> that points to the current AI development and projects ultimately bridging the gap between artificial intelligence and natural intelligence, in terms of energy efficiency.

### High-dimensional (HD) representation for brain-inspired computing

HD information representation as described in Section 5.3 above is the basis of a corresponding computing architecture that may enable performance and energy efficient attainments of brain-like cognitive capabilities relying still on von Neumann like computing architectures<sup>27</sup>. Computing acceleration accuracy can both be enhanced, as shown by a classification experiment comparing HD computing versus support vector machines<sup>28</sup>. (See **Table 5.1**)

Table 5.1: Comparison of HD computing versus support-vector machine (adapted from<sup>28</sup>)

| ARM Cortex M4              |            |              |
|----------------------------|------------|--------------|
| Kernel                     | Cycles (k) | Accuracy (%) |
| High-dimensional computing | 12.35      | 90.70        |
| Support-vector machine     | 25.10      | 89.60        |



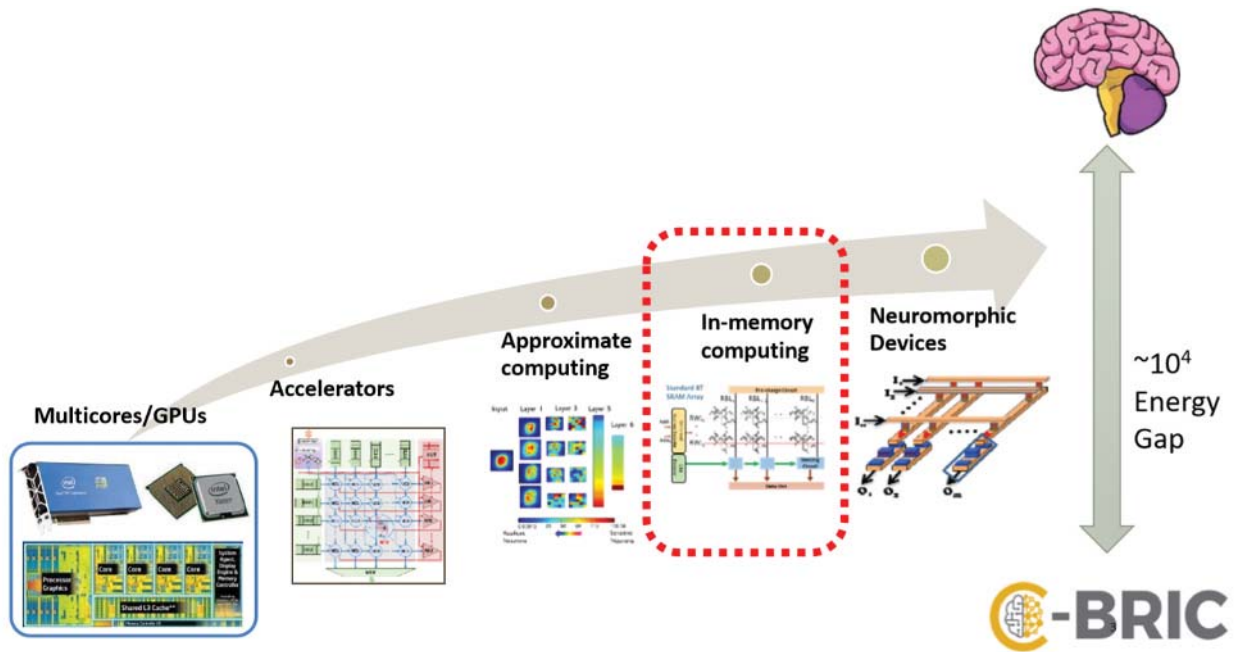


Figure 5.11: Roadmap for the development of hardware to bridge capabilities of Artificial Intelligence (AI) and Natural Intelligence (NI)<sup>26</sup> (courtesy of Anand Raghunathan and Kaushik Roy, Purdue University)

## Neurons and neuromorphic computing

Neurons are excitable in nature. This means that they produce electrical events called action potentials, which are also known as nerve impulses or spikes. Spikes are the basic currency of the brain. They allow neurons to communicate with each other, computations to be performed, and information to be stored. Any time a neuron spikes, neurotransmitters are released from hundreds of its synapses, resulting in communication with hundreds of other neurons<sup>29</sup>.

*There is a push to emulate this behavior in computing.* Because it is event-driven, it is triggered only when the threshold voltage is achieved and not at each propagation cycle, which means that it expends energy only when the neuron crosses the threshold and is reliable. It is robust to noise and does not accidentally spike due to interference. The most important advantage is that it is efficient over long distances, as neurons often project across the brain or whole body<sup>30</sup>.

*Neuromorphic hardware* is an electronic circuitry that mimics the natural biological structure of the nervous system. In contrast to a traditional von Neumann architecture, which has a powerful logic core and operates sequentially on data fetched from memory, neuromorphic computing distributes both computation and memory among a number of relatively primitive neurons that communicate with other neurons<sup>31</sup>. Depending on the application, it can either be brain-inspired

hardware to support spiking, or deep-neural-networks hardware that uses back-propagation algorithms. Von Neumann hardware is fast, serial, power hungry, and dense in time. Neuromorphic hardware provides improvements because it is real time, power efficient, parallel, and sparse in time<sup>32</sup>. Because there is less communication, less computation, and fewer memory lookups, using spikes in neural networks is more efficient.

**A grand challenge in the realization of neuromorphic computing is the ability of a compute engine to learn from unstructured stimuli, with energy efficiency comparable to the human brain.** This involves the realization of Spiking Neural Networks (SNN)<sup>33</sup> in which a neuron can fire independently of other neurons and, thereby, send pulsed signals that directly change the electrical states of those other neurons. This could then elegantly encode information within these signals and their timing, and in the process, as well as simulate natural learning mechanisms by dynamically remapping the synapses between these artificial neurons. The advances in such frameworks may be also bolstered by increased development of less-power-hungry quantum and analog computing and system architectures, which could make their transition from research to production more appealing. Current examples of brain-inspired computing chips include the IBM TrueNorth<sup>34</sup> and Intel Loihi<sup>35</sup>.

## Understanding the spike-based information processing

The current activation level is modeled as a differential equation and is normally considered to be the neuron's state, with incoming spikes pushing this value higher, eventually either firing or decaying. Various coding methods exist for interpreting the outgoing spike train as a real-value number, relying on either the frequency of spikes or the interval between spikes to encode information. One of the concerns about spiking is that it often requires paying a time penalty. However, this is not always the case. There are several coding schemes that are time advantageous by implementing fast threshold gate circuit algorithms on spike hardware. *An example is training deep neural networks for binary communication using the Whetstone method<sup>28</sup>*, which allows the use of spiking communication with no time penalty and minimal accuracy reduction. This was used to train deep neural networks for binary communication<sup>36</sup>. During the training process, the activation function at each layer is slowly refined to the threshold activation. Many other methods for training spiking neural networks exist, such as *the temporal dithering method<sup>37</sup>*, which can effectively interpolate between spiking and non-spiking coding schemes, has no time penalty beyond the simulation time-step, has mathematically proven guarantees, and is interoperable with networks trained using deep learning and/or the Neural Engineering Framework.

## Neural algorithms

The capabilities of spiking can be leveraged by developing algorithms that use time in computing. While spiking lowers the precision of communication, this precision is sometimes unnecessary or can be compensated for in other ways.

**Neuroscience-constrained algorithms that incorporate a broad range of neural plasticity and dynamics are still unexplored from an algorithms' perspective (Table 5.2<sup>38</sup>)**. This suggests that brain-inspired hardware may enable new algorithmic capabilities that extend far beyond the deep learning technologies in use today<sup>39</sup>. Further, it is increasingly recognized that neurons can be treated as powerful logic gates. Since algorithms are circuits, they become a model of parallel computation, energy efficient because of event-driven communication and high fan-in logic. For example, *stochastic differential equations using Monte Carlo simulations and many classes of graph analytics are solvable by spiking circuits<sup>40</sup>*.

Brain-inspired algorithms provide unsupervised learning capabilities and novel unseen classes. However, the unsupervised algorithms are still not as good at classification as back propagation-based algorithms. Spike algorithms use less energy, but the energy to convert information to spikes must be included to the total energy count in cases when the information representation is non-spike based<sup>41</sup>.

Table 5.2: Neural algorithms<sup>38</sup>

| Algorithm Class                         | Current Algorithms   | Inspiration   | Application   |
|---|--|---|---|
| Deep Vision Processing                  | Deep Convolutional Networks (VGG, AlexNet, GoogleNet), HMax, Neocognitron    | Hierarchy of sensory nuclei and early sensory cortices  | Static feature extraction (e.g., images) and pattern classification                 |
| Temporal Neural Networks                | Deep Recurrent Networks (e.g., long short-term memory), Hopfield Networks    | Local recurrence of most biological neural circuits, especially higher sensory cortices                                       | Dynamic feature extraction (e.g., videos, audio) and classification                 |
| Bayesian Neural Algorithms              | Predictive Coding, Hierarchical Temporal Memory, Recursive Cortical Networks | Substantial reciprocal feedback between "higher" and "lower" sensory cortices   | Inference across spatial and temporal scales  |
| Dynamical Memory and Control Algorithms | Liquid State Machines, Echo State Networks, Neural Engineering Framework     | Continual dynamics of hippocampus, cerebellum, and prefrontal and motor cortices  | Online learning content-addressable memory and adaptive motor control               |
| Cognitive Inference Algorithms          | Reinforcement learning (e.g., Deep Q-learning) Neural Turing Machines        | Integration of multiple modalities and memory into prefrontal cortex, which provides top-down influence on sensory processing | Context and experience dependent information processing and decision making         |
| Self-organizing Algorithms              | Neurogenesis Deep Learning   | Initial development and continuous refinement of neural circuits to specific input and outputs                                | Automated neural algorithm development for unknown input and output transformations |

## Remarks on the human brain versus computers

The final goal for brain-inspired computing is the ability to mimic human behavior, which goes beyond computation to recognition, reasoning, and expression of feelings<sup>42</sup>. Modeling the brain's representations with holistic hypervectors has been justified on several grounds, including the size of neural circuits, the brain's tolerance for variation and noise in the input signal, and robustness against component failure.

**Current systems can't function like the human brain.** A field where mimicking the human brain might be important is in autonomous vehicles, as it is emotion that creates a faster reaction to danger. Common sense would be one of the most difficult things to implement, because it requires a large amount of information. Overall, to be able to mimic the human brain, computers need to "understand" others' goals and intentions and be able to adapt based on circumstances. It should also be able to develop a conceptual model of task and a mental model of others<sup>42</sup>.

## AI engines

The industry has managed to progress past the production of general-purpose processors, advanced GPUs, and AI accelerator chips to now give way to approximate computing hardware. *Approximate computing*<sup>43</sup> refers to the tradeoff

in effort expended with computation quality. It has become pervasive in newer CPUs, GPUs, FPGAs, and memory. **A key issue to be addressed is the memory access energy, which is about three orders of magnitude in excess of compute energy.** This constraint underscores the inconvenience of moving data to the CPU for computation and makes near-memory computing<sup>44</sup> (NMC) imperative.

That said, the need for precision-scaling neutralizes the gains from current accelerators' near-memory compute. Research is underway to design memory arrays that exploit parallelism to lower data-movement cost. Critical evaluations of the performance of binary, ternary, and super-ternary (analog) in-memory computing with resistive switching<sup>45</sup> are being done. This in-memory computing requires better design of crossbar-based programmable architectures and peripheral circuits<sup>46</sup>. It can be realized by storing data in RAM and processing it in parallel across a cluster of computers. Obviously, this development in hardware should also be complemented with greater development in AI programming frameworks like graph networks and Causal/Explainable AI<sup>47</sup> to yield 'co-design' development that could achieve greater workload capabilities. The expected parallel developments in both hardware and algorithm streams are outlined in **Figure 5.12** and are exemplified by Cerebras Systems' CS-1 'wafer-scale-system'<sup>48</sup>.

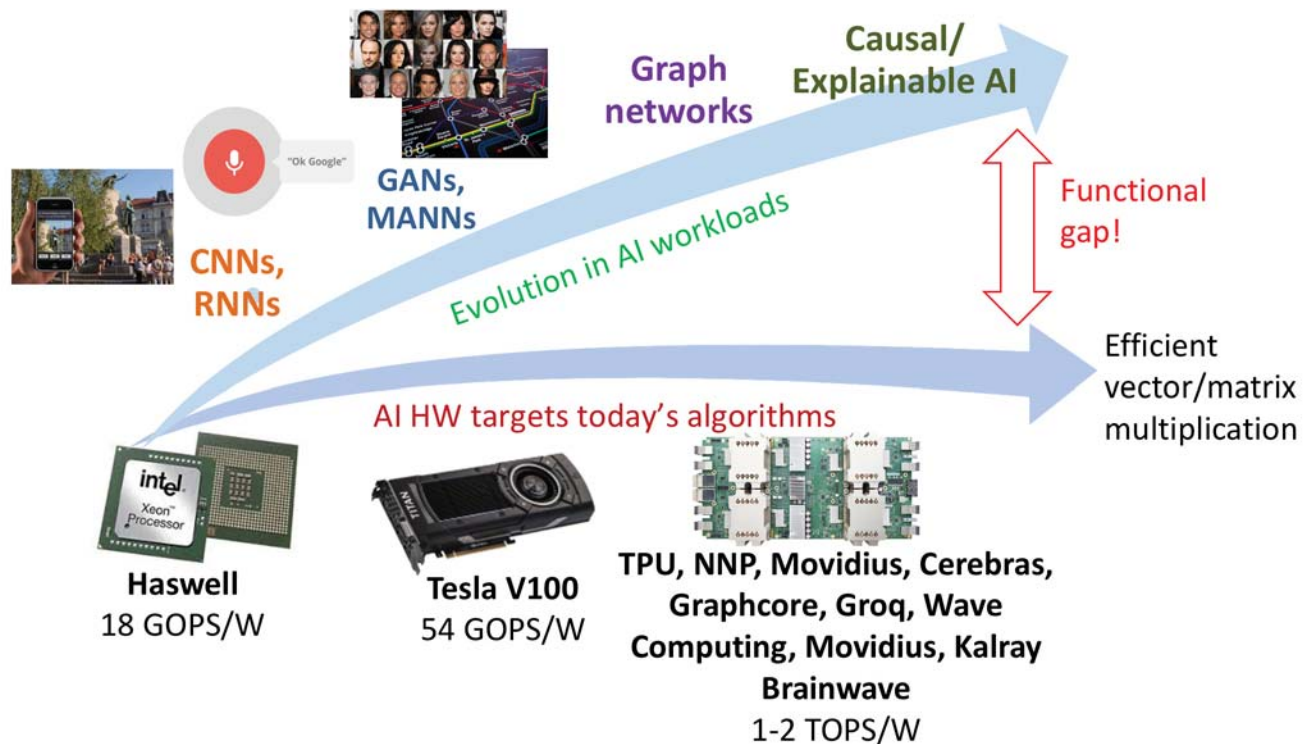


Figure 5.12: Evolution of newer generations of AI hardware and algorithms to enable co-design opportunities<sup>26</sup> (courtesy of Anand Raghunathan, Purdue University)



## Emerging non-volatile memory

It is seen that deep learning algorithms in cloud-based systems are very power hungry. In this regard, *edge systems*<sup>49</sup> that support *Internet of Things (IoT) networks* could perform local computing at the sensor node and consume less area and power, while realizing adaptive transfer learning and processing of continuous data. It is hoped that *analog computing realized by Emerging Non-Volatile Memories (eNVMs) capable of 100 Tera-Operations/Second/Watt (TOPS/watt)* could help in this regard. This would involve better design of devices like *Filamentary and Non-Filamentary Resistive-change RAMs (RRAMs)*, *crystalline/amorphous Phase-Change Random-Access Memories (PCRAMs)*, and *Ferroelectric Transistors (FeFETs)*, some structures of which are shown in **Figure 5.13**<sup>50</sup>. Currently, **eNVMs are still wanting in terms of energy/latency metrics and cycling endurance when compared with SRAMs in advanced technology nodes.**

*Better circuit-level simulators* are being developed to benchmark the area, latency, power consumption, and leakage power from these building blocks for AI Inference Engines for various technology nodes. These would help reduce, for instance, the dynamic energy expended in transitions in these eNVM memory cells and help understand the various factors affecting throughput, latency, and inference accuracy. CAD tools would also increase the appreciation for the impact of factors like device-to-device variation, on-state resistance, programming voltage, write endurance and its degradation during online training, and retention failure on inference algorithms. These would also address the greater difficulties in write capability of eNVMs, in comparison with read capability.

## Key areas of focus and follow-on research

In the US, the Department of Energy (DOE) is focused on the research, development, and use of AI for accelerating scientific discovery<sup>51</sup>. *AI can also be used to address power grid disruptions, fraud and anomaly detection, nuclear deterrent assurance, and other challenging problems.* The DOE plans to continue partnerships with academia, industry, and other government agencies to develop the tools and hardware to compete effectively against growing leadership in adversarial countries and help reduce the AI expertise gap in government. Overall, there is a strong need for foundational research in Scientific Machine Learning and AI (**Figure 5.14**)<sup>52</sup>. Strong foundations lay the groundwork for greater AI-enabled capabilities in massive scientific data analysis, machine learning-enhanced modeling and simulations, and intelligent automation and decision-support for complex processes and systems.

In 2019, the DOE and National Labs conducted four 'AI for Science' townhalls with a final objective of obtaining community consensus to guide the strategic planning for scientific AI for the next five to 10 years. These townhall meetings on AI for Science were documented in a DOE report<sup>16</sup>. Since **algorithms like deep learning, as data-fitted functions between inputs and outputs, may have reached a stagnation point in their potential**, it is necessary for the National Labs to drive greater collaboration with industry and academia to co-design heterogeneous computing solutions that integrate AI, data analysis, and scientific computing hardware designs. Algorithms and computer architectures for AI are evolving quickly and growing more diverse (e.g., neuromorphic, quantum, brain-inspired computing). Emerging trends indicate a need for specialized device

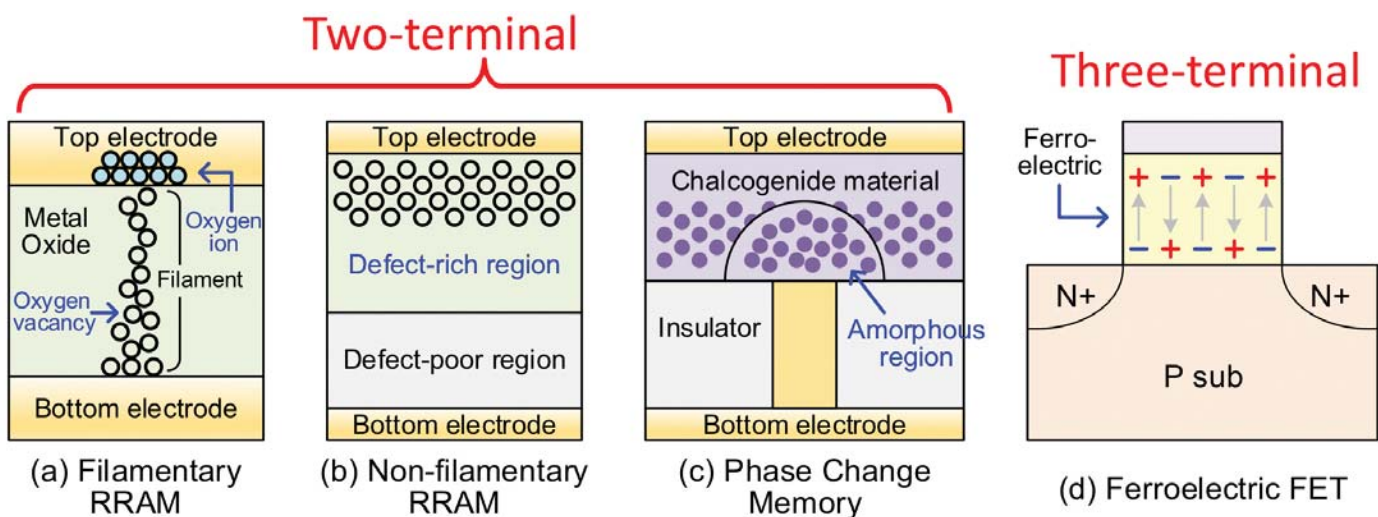


Figure 5.13: Structures of analog devices used in AI Engines<sup>50</sup> (courtesy of Shimeng Yu, Georgia Tech)



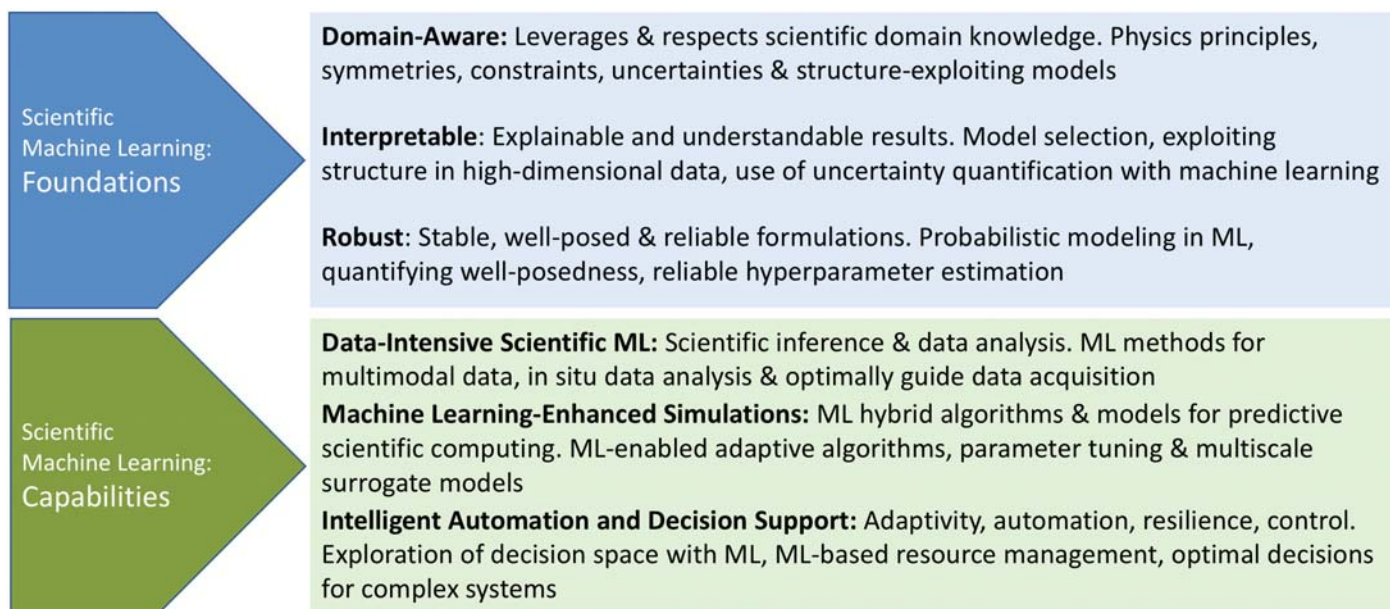


Figure 5.14: Advances in 6 Priority Research Directions (PRDs) are needed to develop the next generation of machine learning methods and artificial intelligence capabilities<sup>51</sup> (courtesy of Steven Lee, DOE/ASCR).

architectures at different scales for five major use cases: (1) AI research and development prototyping; (2) offline training of AI models in production; (3) inference on deployed servers; (4) inference deployed at the edge of a network; and (5) online learning on servers and at the edge. **An important key to success is a strategy that leverages community and industry investments so that AI algorithms and hardware serve the DOE science mission.**

Importantly, neuromorphic computing is still a maturing technology. Because it represents paradigm shifts in math, algorithms, architecture, and hardware, it is important to consider the cross-technology stack implications of modifying each technology in isolation from others. A Roadmap for reaching the potential of brain-derived computing has been recently discussed<sup>53</sup>. Significant technological opportunities exist in the following areas:

- **Learning from devices to algorithms**—How can the properties of an online adaptive device be realized within a brain-inspired learning algorithm?
- **Achieving brain-like connectivity**—Neurons have very high fan-in/fan-out, and they connect to thousands of other neurons using point-to-point connections in three dimensions. In contrast, all neuromorphic hardware today relies on some form of conventional routing technology of spiking events.
- **What is the most effective hardware architecture for leveraging both analog devices and event-based spiking communication?**

## 5.5. Large-scale Quantum Computing

### Overview and needs

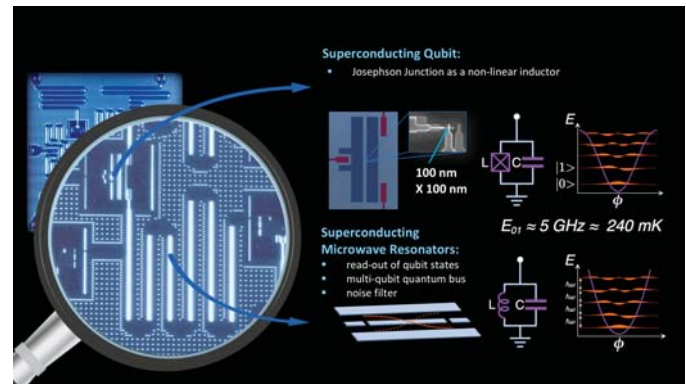
Modern “classical” semiconductor technology is based on quantum mechanics. For example, the quantum phenomenon of energy bands is a foundational concept of all semiconductor devices, from transistors to light-emitting diodes (LED). Quantum mechanical tunneling is the basic operation mechanism of flash memory, while the effect of quantum confinement is used in high-electron-mobility transistors (HEMPT).

Other more peculiar quantum effects, namely quantum superposition and quantum entanglement, have been proposed as a foundation for the paradigm of quantum information processing that includes quantum computing. There are theoretical predictions that quantum computing using quantum bits (“qubits”) would result in a significant speedup over the current computing technologies for certain problems. And several small-scale quantum computers (QCs) have been demonstrated.

Large-scale QCs will be needed for the realization of complex practical tasks. This section addresses the topics of hardware realizations of large-scale QCs and the practical potential of QCs to define new computing trajectory for broad-purpose information processing, with orders-of-magnitude improvement in energy efficiency.



a) Trapped ion-based qubits with optical control<sup>56</sup> (courtesy of ionQ)



b) Superconducting qubits<sup>55</sup> (courtesy of IBM)

Figure.5.15. Examples of types of qubits

## Quantum computing promises and challenges

*As an emerging technology, quantum computing has been gaining momentum, and there has been significant progress that includes recent efforts towards the goal of quantum supremacy<sup>54</sup>, the point at which QCs can solve problems not feasible by classical means.*

Similar to classical computers, as the number of computing elements (bits or qubits) increases, more possibilities can be calculated with more accuracy. Quantum superposition and quantum entanglement are two of the major concepts that contribute to the enormous potential of QCs. Just 300 logical qubits would allow for the exploration of more combinations than the number of particles in the universe. To realize a logical qubit, multiple physical qubits are needed to insure computational robustness. **Thus a challenge in QC is the realization of as many as possible logical qubits each comprised of the minimum number of physical qubits.**

## Qubit technologies

There are different types of qubit technologies, such as trapped atomic ions, neutral atoms, superconducting circuits, topological qubits, photonic

qubits, and semiconductor spins. The important metric is the combination of qubit coherence and gate quality<sup>55</sup>. Because of their properties, *trapped ions<sup>56</sup> and superconducting qubits<sup>56</sup> appear to be among the most promising and prevalent thus far (Figure 5.15).* While superconducting circuits can exploit chip-fabrication processes common to conventional VLSI, ion traps feature nearly perfect qubit quality and replicability. Both systems have been shown to satisfy DiVincenzo's criteria for a physical system to implement a logical qubit<sup>57</sup>. Gate operations, initializing and reading qubit states, and storing operations can all be designed

to keep the computer functional for the execution of quantum gate sequences. It is this sense of control that provides a measure of confidence for scaling the number of qubits.

## Challenges

There are two general challenges in the full realization of QCs. First, **as the QC is scaled, it typically becomes more difficult to isolate the system from noise in the environment and prevent errors (Figure 5.16).** Moreover, some technologies, especially in the solid state, cannot be easily replicated. The minute differences between qubits must be cataloged, and these

NOISE IS THE MAJOR CHALLENGE IN QC  
HOW LONG BEFORE A HARDWARE ERROR?



Q-CTRL's quantum firmware addresses this issue using *control engineering*

Figure 5.16: The considerable effect of noise in quantum computers<sup>58</sup> (courtesy of Q-CTRL) The hybrid quantum-classical algorithms (see Figure 5.17) have shown good performance in that they provide high accuracy by reducing approximations needed to render problems computable and at higher speed<sup>61</sup>. A potential key advantage is that the quantum computing portion can virtually decouple the compute power from energy consumption. This provides the capability to encode and manipulate data in exponentially large state spaces at a lower cost.

differences can also drift in time, making it very difficult to scale. Such noise in the environment disrupts quantum systems and causes errors in the computation through dephasing and relaxation<sup>58,59</sup>, where analog errors get turned into digital errors. The second challenge involves **imperfections in the control operations of the quantum gates**, which are usually exacerbated through unwanted crosstalk of the control systems and qubits as the system is scaled.

One way to address these challenges is to *exploit error correction*. However, quantum error correction adds significant overhead in terms of additional qubits and gate operations, in some cases requiring a multiplicative factor of 10,000 or more qubits, straining the ability to scale. Consequently, in the short term, there has been a focus on NISQ (noisy intermediate-scale quantum) systems that can be combined with classical computers<sup>60</sup>.

Superconducting qubit QCs operate at 20 mK, requiring significant cooling using a dilution refrigerator, and the electronics readouts are done at room temperature, or 300 K (see **Figures 5.18 and 5.19**)<sup>60</sup>. To improve scaling, *architectures have been proposed that include readouts at lower temperatures as depicted in Figure 5.19*. Ion trap QC

systems can be run at room temperature but require a vacuum environment for the floating atoms in space. These two technologies also have complementary challenges to scaling. First, superconductors enjoy the ability to fabricate many qubits on-chip but have finite coherence time owing to their noisy solid-state environment. Next, the complex wiring of large numbers of qubits and their crosstalk, all in a cryogenic environment, cause errors that generally grow as the system is made larger. **These challenges may be fundamental in nature, requiring new approaches to the basic qubit materials and their substrates, their interconnects, and their refrigeration.** Ion trap qubits have negligible qubit idle/memory errors, and the qubit technology of isolated atoms will never be improved, as they are perfectly replicable atomic clocks. Here the challenge is the scaling of the classical controllers, typically laser beams that must be engineered to address the individual atoms in the vacuum chamber.

### Comparing QCs

Given the differences in types of qubits and control requirements, it is often difficult to compare QCs. One gauge sometimes used is the number of qubits. However, comparing QCs by counting qubits is not accurate because it

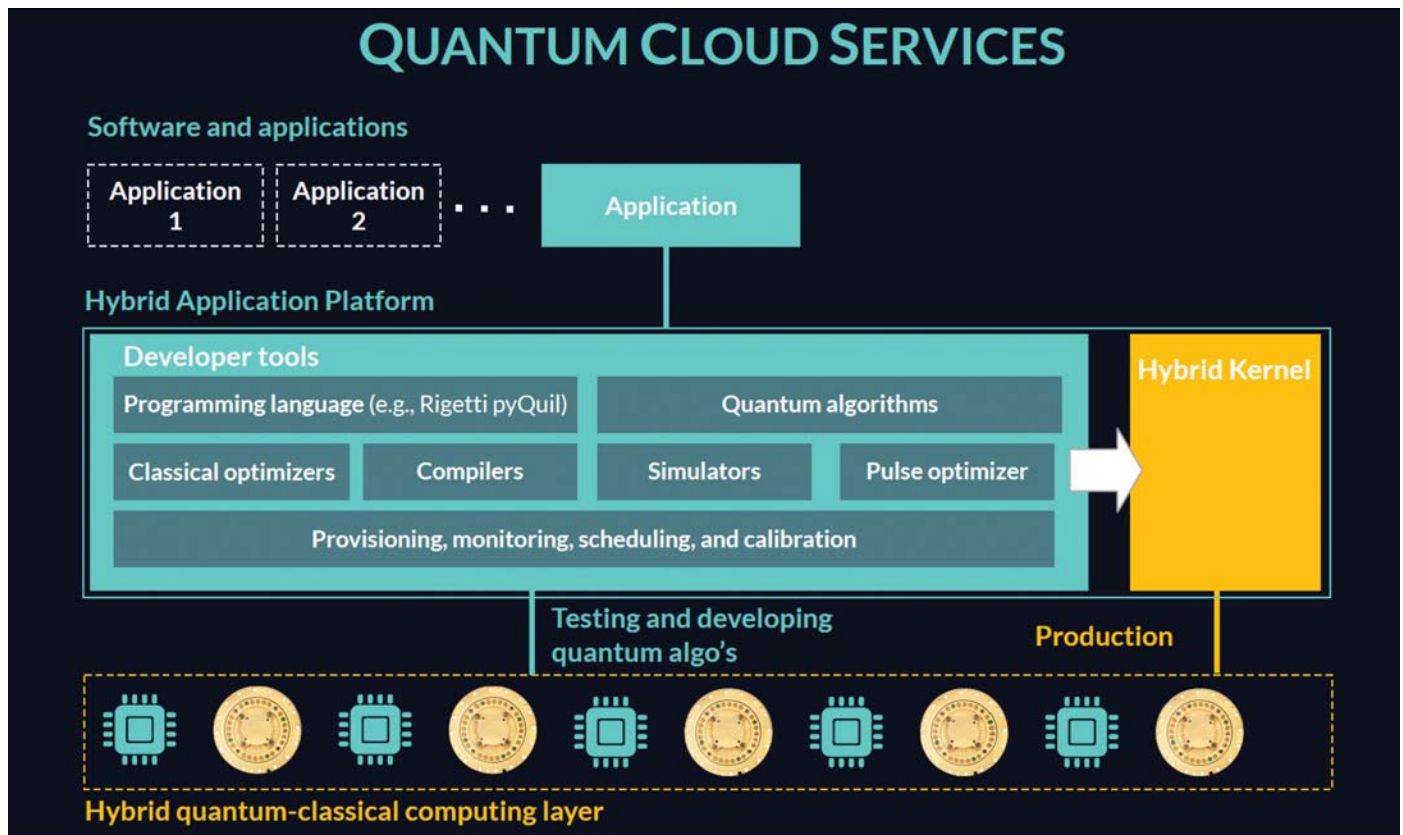


Figure 5.17: Quantum cloud services with hybrid quantum/classical algorithms<sup>61</sup> (courtesy of Rigetti)



is the quality of the qubits and control mechanisms that truly count. It is more helpful to think of the number of equivalent “logical qubits,” defined to be qubits that have a coherence time long enough to be usable by a quantum gate. Another metric that can be useful to assess a QC is Quantum Volume, which is a metric that depends on three dimensions: number of qubits, error rate, and qubit connectivity.

### Applications

QCs are not expected to completely replace classical computers, but rather they will be used for a class of applications that optimize functions by simultaneously sampling all inputs. Also, due to the fundamental nature of quantum systems, a natural area of study will be molecular simulations, with applications ranging from design of materials (e.g., batteries) to target proteins in drug discovery. This is in addition to factorization (Shor’s algorithm) with applications in several domains, including decoding of encrypted data.

### Key areas of focus and follow-on research

In summary, QC is a paradigm that is slowly being realized, with a number of difficult engineering and science challenges that remain unfulfilled. The next big leap in quantum computing will require much higher levels of investment in QC education and R&D. Colleges and universities

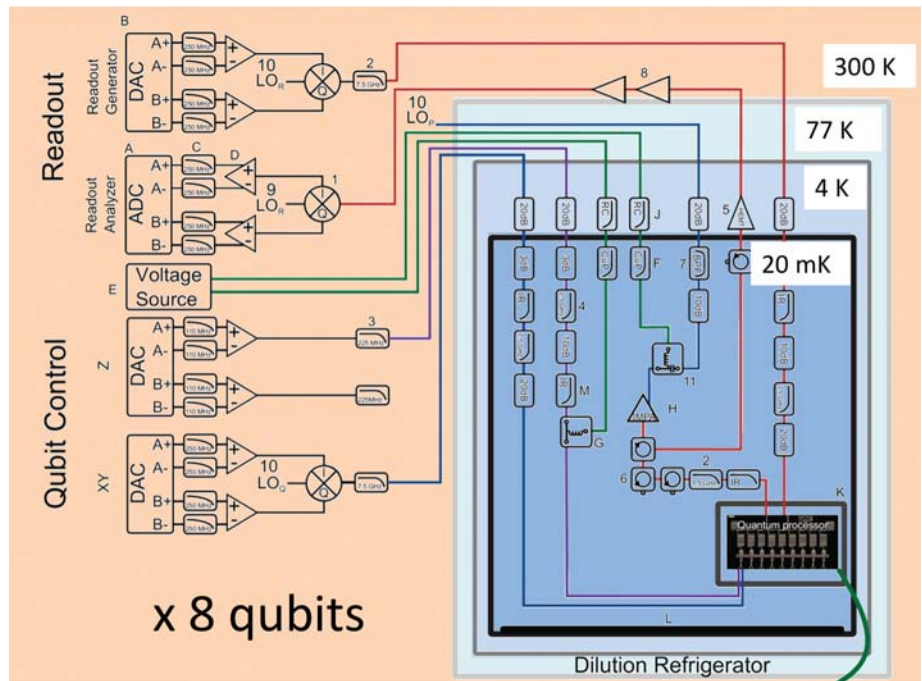


Figure 5.18: Quantum computer setup<sup>59</sup> (courtesy of Edoardo Charbon)

should consider developing additional undergraduate courses on the theory and practice of QC in the same vein that logic systems and computer programming courses are offered as required or electives depending on the major. There is also a need for additional R&D funding programs that attract the brightest minds to this important field of computing.

With applications ranging from combinatorial optimization to molecule ground state calculations, the future of QC looks promising. **To make QC practical, we need to make them much more scalable.** Indeed, scalability remains the key challenge given the very

high error rates that we are seeing in today’s modest QCs. Qubit and system control, as well as robustness and related error tolerance are among the key drivers. Scalability considerations include parameters such as: noise budget, power budget per qubit, physical dimensions, replicability, and bandwidth for multiplexing<sup>57</sup>. These parameters are not necessarily separate components of the puzzle but very much interrelated. **As of the date of writing of this Chapter, there is no single preferred pathway identified for scaling as it very much depends on the type of qubit implementation technology.** It remains to be seen which technology or set of technologies will win out.

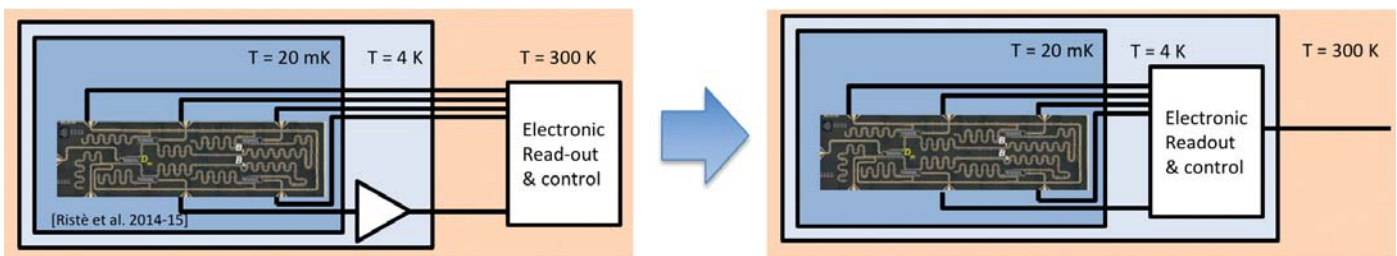


Figure 5.19: Proposed architecture for scalability<sup>59</sup> (courtesy of Edoardo Charbon)



## 5.6. Summary: New Compute Trajectories for Energy-efficient Computing

### Overview

*The total energy consumption by computing continues to grow exponentially, which may become unsustainable even before 2030.* Thus, radical improvement in the energy efficiency of computing is needed. The Grand Goal of computing research should be the discovery of computing paradigms/architectures with new computing trajectory, providing >1,000,000x improvement in energy efficiency. It is becoming increasingly clear that in future information-processing applications, synergistic innovations—from materials and devices to circuits and system-level functions using unexplored physical principles—will be a key to achieving new levels of energy efficiency and performance. Paradigm-shifting new solutions will be needed for future computers, with major innovations in devices, circuits, and architectures. *New approaches to computing will be necessary, such as in-memory compute, special purpose compute engines combined in a heterogeneous manner, different AI platforms, brain-inspired/neuromorphic computation, quantum computing, and other solutions.*

In summary, revolutionary changes to computing are needed very soon. This necessitates a completely new way of thinking to realize a solution space with multi-decade longevity replacing the von Neumann/CMOS-based approach that has served the ICT community well for over half a century. This report is intended to stimulate collaborative research, from materials to architecture, aiming at establishing revolutionary paradigms to support future energy-efficient computing for the vast range of future data types and heterogeneous workloads.

The **Computing Grand Goal** is to discover computing paradigms/architectures with a radically new computing trajectory, demonstrating >1,000,000x improvement in energy efficiency. Changing the trajectory provides immediate improvements and establishes a buffer of many decades (as shown in **Figure 5.1**). This would be much more cost effective than attempting to dramatically increase the world's energy supply.

### Research recommendations summary

Reaching new highs for deterministic high-performance computing capabilities and AI cognitive capabilities, along with orders-of-magnitude improvements in energy efficiency on such workloads, will require significant research investments on multiple but synergistically interlocked areas.

These include:

- Advancing the theoretical basis for computing performance
- New types of information representation and processing, as well as new compute paradigms
  - High-Dimensional computing
  - Shannon-inspired statistical computing
  - Heterogeneous computing that integrates general purpose processors with specialized accelerators for: AI/ML, Data Analytics, Neuromorphic Computing, and Quantum Computing
- Exploring new building blocks and interconnect fabrics that press against the fundamental limits of advanced logic and memory technologies
  - device (transistor and memory) research enabling added technology functionalities while minimizing complexity
  - charge-based and non-charge-based devices
    - spintronics, photonics, and others
    - dense, high-speed, and low-energy interconnect fabrics
    - new materials and processes
- Establishing partnerships among academia, industry, and government agencies to help reduce the AI expertise gap and develop HPC, AI/ML and data analytics tools and hardware to compete effectively with adversarial countries
- Foundational research in Scientific AI/ML, data analytics and HPC
  - optimization of fundamental principles of domain-awareness
  - effective use of structured high-dimensional data
  - learning-assisted uncertainty quantification
- Driving greater collaboration with academia, DOE National Labs, and other US Government agencies in AI hardware design and algorithmic co-optimization, since, as optimization functions between inputs and outputs
- Applying machine-learning to multimodal data and optimally guiding data acquisition
  - intelligent automation to enhance exploration of the decision-space and learning-based resource management capabilities

- Understanding factors contributing to analysis of ML, such as model reduction and multi-fidelity modeling, computational complexity and optimization, and statistics and uncertainty quantification
  - ethical, legal, and societal implications of AI
  - develop better shared public datasets and environments for AI training
  - strengthen public-private R&D partnerships
- Neuromorphic computing, representing paradigm shifts in math, algorithms, architecture, and hardware
  - consider the cross-technology stack implications of modifying one technology in isolation from others
- Significant technological opportunities
  - learning from devices to algorithms
    - realize the properties of an online adaptive device within a brain-inspired learning algorithm
  - achieving brain-like connectivity
    - neurons have very high fan-in/fan-out, connecting to thousands of other neurons using point-to-point connections in three dimensions
    - all neuromorphic hardware today relies on some form of conventional routing technology of spiking events
  - finding the most effective hardware architecture for leveraging both analog devices and event-based spiking communication
- Scalable quantum computing
  - scalability considerations include noise budget, power budget, physical dimensions, and bandwidth
- Creating a full stack that includes the chip, system, software, and cloud access
- Developing hybrid quantum-classical algorithms in both superconducting and ion trap systems

## Appendix: Global Compute Inventory

The theoretical basis for computing performance is less solid than the theoretical basis for information storage and communication, like the Shannon limit and others. It is not easy to quantify the computational power of compute engines. There are multiple characteristics like MIPS, MOPS, and FLOPS, and there is no straightforward formula for the conversion between them<sup>62</sup>.

General-purpose computing has advanced to an extent where we are able to perform many classes of complex calculations and simulations that are mapped with heterogeneous programming models to systems with CPUs and GPUs integrated with a communication fabric. The overall computational performance of CPU is often measured in (millions of) instructions per second (IPS or MIPS) that can be executed across a standard set of benchmarks. In turn, GPU and supercomputer performance is typically measured in FLOPS (Floating Points Operations Per Second). While there is no straightforward formula for the conversion between FLOPS and MIPS, and the two types of metrics are regarded as complementary, for the purposes of this study, we utilized one single metric for computing performance. Following Hilbert and Lopez<sup>63</sup>, we choose the Dhrystone VAX MIPS as the unit of measurement because of the availability of relevant and consistent statistics. According to<sup>62</sup>:

$$1 \text{ FLOPS} \approx 3 \text{ IPS (A1)}$$

As can be seen in **Figure A1**, the world's technological effective capacity to compute information is  $\sim 3 \times 10^{15}$  MIPS in 2020, and is projected to rise to  $\sim 10^{22}$  MIPS in 2040 (based on research by Hilbert and Lopez<sup>62</sup>).

Another indicator of the ultimate performance of an information processor, realized as an interconnected system of binary switches, is the maximum binary throughput (in bit/s or BITS), that is, the maximum number of on-chip binary transitions per unit time. It is proportional to the product of the number of devices (transistors)  $N_{tr}$  with the clock frequency of the microprocessor  $f_{clk}$ :

$$BITS = a \cdot \frac{N_{sw}}{t_{sw}} = a \cdot N_{sw} \cdot f_{clk} \quad (\text{A2})$$

Where  $a$  is the activity factor of a digital circuit (here we assume  $a=0.01$ )

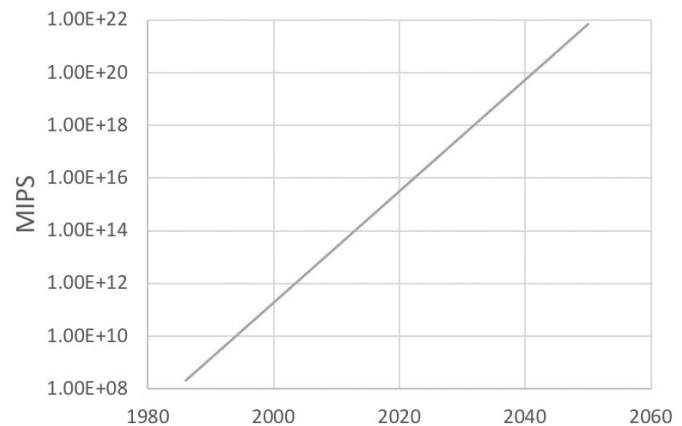


Figure A1: Trend in world's technological information processing capacity (Based on research by Hilbert and Lopez<sup>62</sup>)

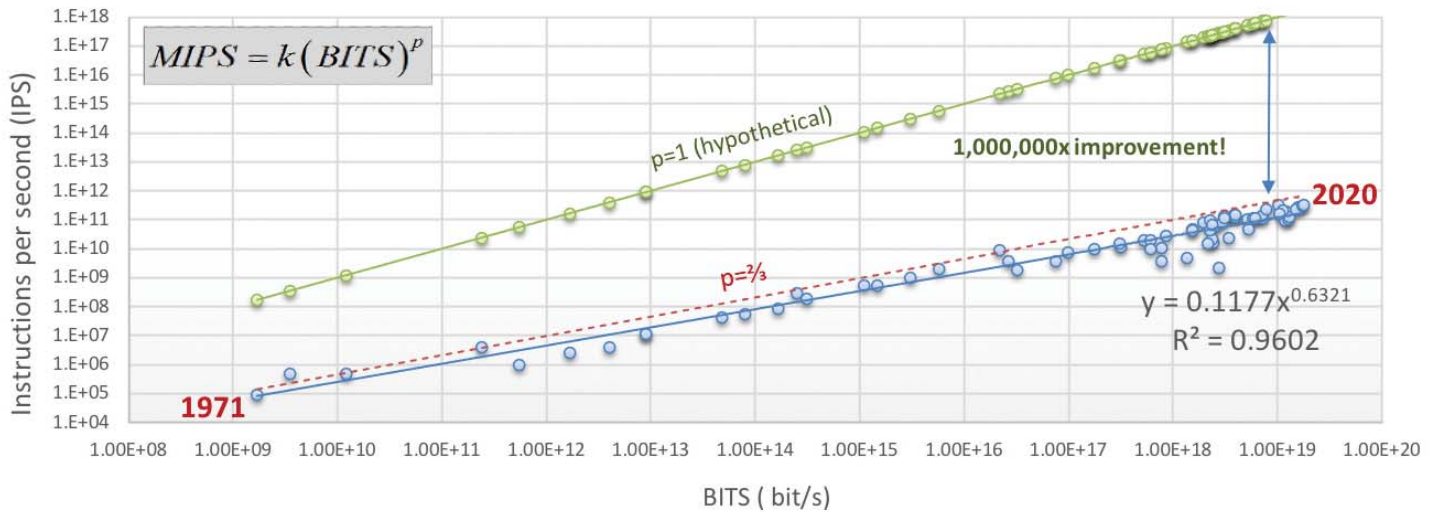


Figure A2: A 1971-2019 trend for computational performance in IPS as a function of binary information throughput in bit/s for CPU

### General-purpose CPUs

For general-purpose CPUs, there is a strong correlation between the overall computational performance measured in (M)IPS and the system’s binary throughput (BITS), measured in bit/s, which represents a characteristic number of “raw” binary transitions in the system needed to implement an instruction, as can be seen in **Figure A2**, and to a good approximation:

$$IPS = k(BITS)^p \quad (A3)$$

For a variety of CPU chips produced in 1971-2019 by different companies,  $k \sim 0.1$  and  $p \sim 2/3$  with a high degree of accuracy (the determination coefficient  $R^2=0.96$ ). For details, see **Table A1**. This strong correlation suggests a possible fundamental law behind the empirical observation. It is also instrumental for speculations about future developments.

### Graphics processing units (GPU)

The current trend is toward the increasing use of GPU for different computational tasks, including general-purpose computations and machine learning/artificial intelligence. **Figure A3** presents the percentage of GPU in the world’s technological installed compute capacity, as estimated by this group based on<sup>62</sup>.

For a variety of GPU chips produced in 1995-2018 (see **Table A2**), there is also a strong correlation between the overall computational performance (measured in FLOPS or MIPS) and the system’s binary throughput (BITS), measured in bit/s (see **Figure A4**). And to a good approximation, it is also described

by (A3) with  $k \sim 10^{-3}$  (when converted to the IPS metrics using (A1)) and  $p \sim 0.85$  with a high degree of accuracy (the determination coefficient  $R^2=0.99$ ).

To account for both CPU and GPU in the world’s technological installed capacity to compute information, we define “effective” values for the parameters  $k$  and  $p$  in (A2) as:

$$\begin{aligned} k &= n_{GPU} k_{GPU} + n_{CPU} k_{CPU} \\ p &= n_{GPU} p_{GPU} + n_{CPU} p_{CPU} \end{aligned} \quad (A3)$$

where  $n_{GPU}$  and  $n_{CPU}$  are, respectively, the relative proportions of GPUs and CPUs in the world’s installed compute infrastructure ( $n_{GPU} + n_{CPU} = 1$ ).

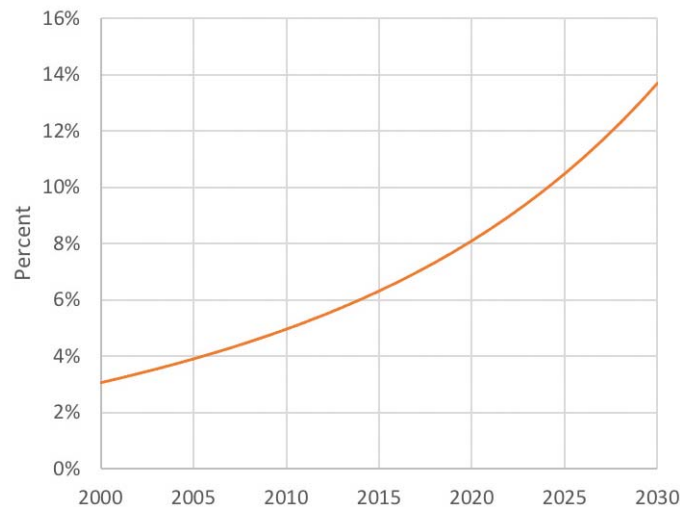


Figure A3: Percentage of GPU in the world’s technological installed compute capacity

## Global “raw” bit transitions per second:

From (A2):

$$BITS = \frac{a}{k} \cdot (IPS)^{\frac{1}{p}} \quad (A4)$$

Global “raw” bit transitions per year:

$$BITY = y \cdot BITS \cdot 3600 \cdot 24 \cdot 365 \quad (A5)$$

(y is computing engine utilization factor, here we assume  $y=0.005$ )

The resulting total number of binary transitions required for computing per year is shown in **Figure A5**. For example, for the scenario, where all computations are performed by CPUs, the total number of “raw” binary transitions required for computing is  $\sim 10^{37}$  bits/y in 2020 and could reach  $\sim 10^{46}$  bits/y by 2050. A scenario with an increasing use of GPU for different computational tasks, along with CPU, suggests  $\sim 10^{36}$  bits/y in 2020 and  $\sim 10^{42}$  bits/y in 2050. The dashed middle line on **Figure A5** is the geometric mean of the two scenarios, which can be regarded as a “nominal trend” that projects  $\sim 10^{44}$  bits/y in 2050.

## Total energy of computing

The total energy of computing per year,  $E_{tot}$ , can be obtained by multiplying the global “raw” bit transitions per year (BITY) and the energy per one bit transition  $E_{bit}$ :

$$E_{tot} = BITY \cdot E_{bit} \quad (A6)$$

The energy per one-bit transition in compute processor units (e.g., CPU, GPU, and FPGA) has been decreasing over last 40 years (as manifested by Moore’s law), and is  $\sim 10$  attojoules or  $10^{-17}$  J in current processors. However, the demand for computation growth is outpacing the progress realized by Moore’s law. In addition, Moore’s law is currently slowing down as device scaling is approaching fundamental physical limits. The physics-based theoretical lower limit, known as the Landauer limit for binary switching, is  $3 \times 10^{-21}$  J/bit.

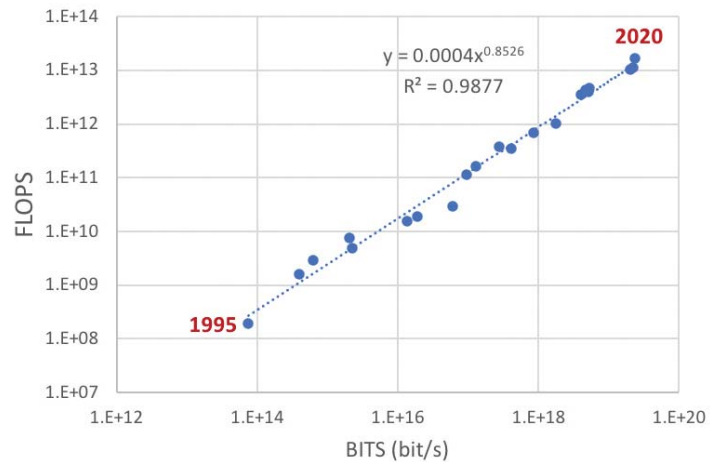


Figure A4: A 1995-2018 trend for computational performance in FLOPS as a function of binary information throughput in bit/s for GPU

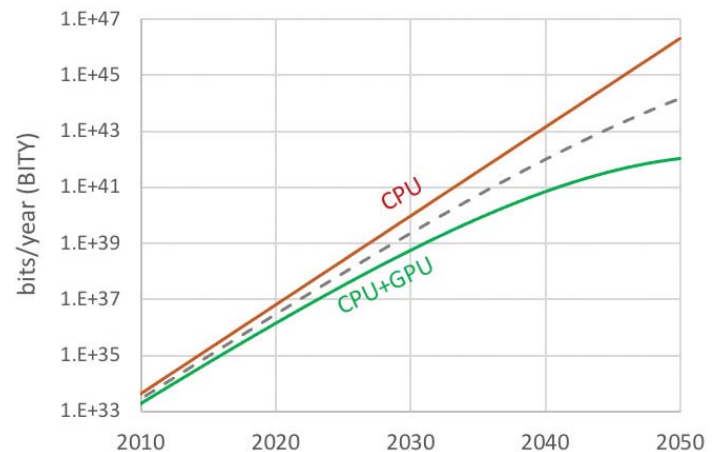


Figure A5: The total number of “raw” binary transitions required for computing (the top solid line represents a hypothetical scenario, where all computations are performed by CPU; the bottom solid line represents a scenario with an increasing use of GPU for different computational tasks, along with CPU; the dashed middle line is the geometric mean of the two scenarios)

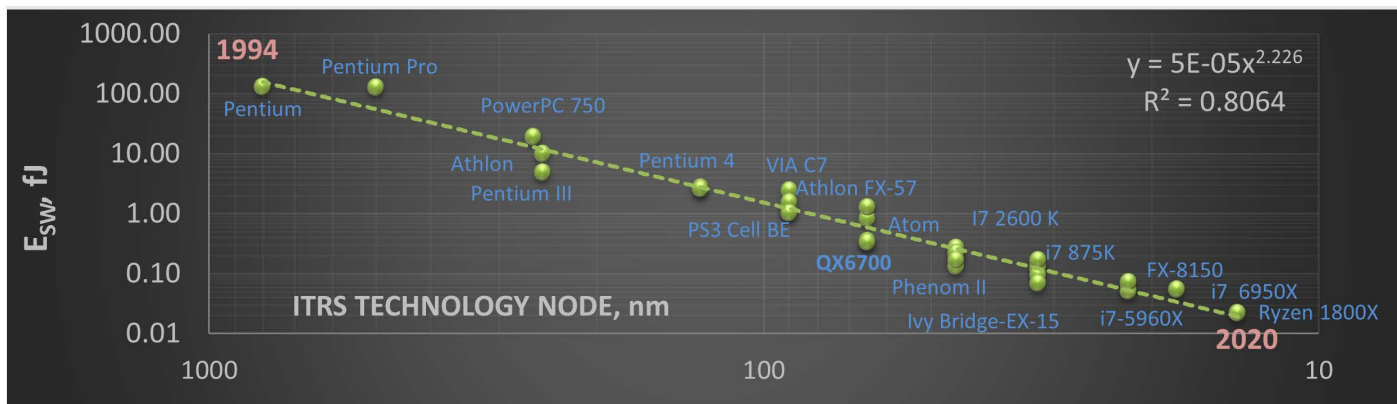


Figure A6: Energy per bit transition in compute processor units



**Table A1. A 1971-2019 CPU IC summary**

| Processor                               | Year | MIPS    | Clock frequency, MHz | Transistor count |
|---|------|---------|----------------------|------------------|
| Intel 4004                              | 1971 | 0.092   | 0.74                 | 2,300            |
| Intel 8080                              | 1974 | 0.5     | 2                    | 6,000            |
| MOS Technology 6502                     | 1975 | 0.5     | 1                    | 3,510            |
| Motorola 68000                          | 1979 | 1.00    | 8                    | 68,000           |
| Intel 286                               | 1982 | 2.66    | 12.5                 | 134,000          |
| Motorola 68020                          | 1984 | 4.00    | 20                   | 200,000          |
| Intel 386DX                             | 1985 | 11.4    | 33                   | 275,000          |
| Motorola 68030                          | 1987 | 11      | 33                   | 273,000          |
| Motorola 68040                          | 1990 | 44      | 40                   | 1.20E+06         |
| DEC Alpha 21064 EV4                     | 1992 | 300     | 100                  | 3.10E+06         |
| Intel Pentium Pro                       | 1996 | 541     | 200                  | 5.50E+06         |
| IBM PowerPC 750                         | 1997 | 525     | 233                  | 6.35E+06         |
| Intel Pentium III                       | 1999 | 2050    | 600                  | 9.50E+06         |
| AMD Athlon                              | 2000 | 3560    | 1200                 | 2.20E+07         |
| AMD Athlon XP 2500+                     | 2003 | 7530    | 1830                 | 5.43E+07         |
| Pentium 4 Extreme Edition               | 2003 | 9730    | 3200                 | 5.50E+07         |
| VIA C7                                  | 2005 | 1800    | 1300                 | 2.50E+07         |
| AMD Athlon FX-57                        | 2005 | 12000   | 2800                 | 1.14E+08         |
| AMD Athlon 64 3800+ X2 (Dual core)      | 2005 | 1460    | 2000                 | 1.54E+08         |
| Xbox360 IBM "Xenon" (Triple core)       | 2005 | 1920    | 3200                 | 1.65E+08         |
| PS3 Cell BE (PPE only)                  | 2006 | 10200   | 3200                 | 2.41E+08         |
| AMD Athlon FX-60 (Dual core)            | 2006 | 18900   | 2600                 | 2.33E+08         |
| Intel Core 2 Extreme X6800 (Dual core)  | 2006 | 27100   | 2930                 | 2.91E+08         |
| Intel Core 2 Extreme QX6700 (Quad core) | 2006 | 49200   | 2660                 | 5.82E+08         |
| P.A. Semi PA6T-1682M                    | 2007 | 88000   | 2000                 | 1.10E+07         |
| Intel Core 2 Extreme QX9770             | 2008 | 59500   | 3200                 | 8.00E+08         |
| Intel Core i7 920                       | 2008 | 82300   | 2660                 | 7.31E+08         |
| Intel Atom N270                         | 2008 | 3850    | 1600                 | 4.70E+07         |
| AMD Phenom II X4 940                    | 2009 | 42800   | 3000                 | 7.58E+08         |
| AMD Phenom II X6 1100T Thuban           | 2010 | 78400   | 3300                 | 9.04E+08         |
| Intel Core i7 Extreme Edition 980X      | 2010 | 148,000 | 3330                 | 1.17E+09         |
| Intel Core i7 2600K                     | 2011 | 128,000 | 3400                 | 1.16E+09         |
| AMD E-350                               | 2011 | 10000   | 1600                 | 3.80E+08         |
| Intel Core i7 875K                      | 2011 | 92100   | 2930                 | 7.74E+08         |
| AMD FX-8150                             | 2011 | 109,000 | 3600                 | 2.00E+09         |
| Xeon E3-1290 v2                         | 2012 | 100,000 | 3700                 | 1.40E+09         |
| Ivy Bridge-EX-15                        | 2013 | 200,000 | 2800                 | 4.30E+09         |
| i7-5960X                                | 2014 | 238,000 | 3000                 | 2.60E+09         |
| Intel core i7 6950X                     | 2016 | 334,000 | 3000                 | 3.40E+09         |
| AMD Ryzen 1800X                         | 2017 | 305,000 | 3,600                | 4.80E+09         |
| AMD Ryzen2 2700X Pinnacle Ridge         | 2019 | 334,000 | 3,700                | 4.80E+09         |

**Table A2. A 1995-2018 GPU IC summary<sup>63</sup>**

| Processor                                | Year | FLOPS    | Clock frequency, MHz | Transistor count |
|--|------|----------|----------------------|------------------|
| STG-2000                                 | 1995 | 1.92E+08 | 75                   | 1.00E+06         |
| Riva 128                                 | 1997 | 1.60E+09 | 100                  | 4.00E+06         |
| Riva TNT                                 | 1998 | 2.88E+09 | 90                   | 7.00E+06         |
| Riva TNT2 Ultra                          | 1999 | 4.80E+09 | 150                  | 1.50E+07         |
| GeForce 256 DDR                          | 2000 | 7.68E+09 | 120                  | 1.70E+07         |
| GeForce3 Ti500                           | 2001 | 1.54E+10 | 240                  | 5.70E+07         |
| GeForce4 Ti4600                          | 2002 | 1.92E+10 | 300                  | 6.30E+07         |
| GeForce FX 5900 Ultra                    | 2003 | 2.88E+10 | 450                  | 1.35E+08         |
| GeForce 6800 Ultra Extreme               | 2004 | 1.15E+11 | 425                  | 2.22E+08         |
| GeForce 7800 GTX                         | 2005 | 1.65E+11 | 430                  | 3.02E+08         |
| GeForce 7950 GX2                         | 2006 | 3.84E+11 | 500                  | 5.56E+08         |
| S870 GPU Computing Server (4xG80)        | 2007 | 3.46E+11 | 600                  | 6.81E+08         |
| S1070 Server 500 Configuration (4xGT200) | 2008 | 6.91E+11 | 602                  | 1.4E+09          |
| S2050 GPU computing server (4xGF100)     | 2011 | 1.03E+12 | 575                  | 3.10E+09         |
| K20X GPU Accelerator (1xGK110)           | 2012 | 3.94E+12 | 732                  | 7.08E+09         |
| K40 GPU Accelerator (1xGK110B)           | 2013 | 4.67E+12 | 745                  | 7.08E+09         |
| K80 GPU Accelerator (2xGK210)            | 2014 | 3.58E+12 | 562                  | 7.10E+09         |
| M60 GPU Accelerator (2xGM204)            | 2015 | 4.25E+12 | 899                  | 5.20E+09         |
| P100 GPU Accelerator (1xGP100)           | 2016 | 1.01E+13 | 1,300                | 1.53E+10         |
| V100 GPU Accelerator (1xGV100)           | 2017 | 1.67E+13 | 1,130                | 2.11E+10         |
| T4 GPU Accelerator (1xTU104)             | 2018 | 1.12E+13 | 1,620                | 1.36E+10         |

## Contributors

|  |                                 |  |
|--|---------------------------------|--|
| J. Bradley Aimone (Sandia National Labs)   | Tayfun Gokmen (IBM)             | Chad Rigetti (Rigetti Computing)       |
| Rob Aitken (ARM)                           | Stephen Kosonocky (AMD)         | Kaushik Roy (Purdue U)                 |
| Stefano Ambrogio (IBM)                     | Steven Lee (DOE/ASCR)           | Naresh Shanbhag (UIUC)                 |
| James Ang (Pacific Northwest National Lab) | Rafic Makki (Mubadala)          | Yichen Shen (Lightelligence)           |
| Michael Biercuk (U Sydney)                 | Christopher Monroe (U Maryland) | Fred Streitz (DOE/AITO)                |
| Roy Campbell (DOD/HPCMP)                   | Dmitri Nikonov (Intel)          | Stephen Trimberger (U Maryland)        |
| Edoardo Charbon (EPFL)                     | Don Norman (UCSD)               | Aaron Voelker (Applied Brain Research) |
| Jacqueline Chen (Sandia National Labs)     | Bruno Olshausen (UC Berkeley)   | David Wentzlaff (Princeton U)          |
| Robert Clark (TEL)                         | John Owens (UC Davis)           | Ian Young (Intel)                      |
| Oliver Dial (IBM)                          | Anand Raghunathan (Purdue U)    | Shimeng Yu (Georgia Tech)              |
| Carlos Diaz (TSMC)                         | Titash Rakshit (Samsung)        |  |
| Chris Eliasmith (Applied Brain Research)   | Heike Riel (IBM)                |  |

# References to Chapter 5

- <sup>1</sup>National Strategic Computing Initiative Update: Pioneering the Future of Computing, <https://www.nitrd.gov/pubs/National-Strategic-Computing-Initiative-Update-2019.pdf>
- <sup>2</sup>Basic Research Needs for Microelectronics, <https://doi.org/10.2172/1616249>
- <sup>3</sup>W. Van Heddeghem et al, "Trends in worldwide ICT electricity consumption from 2007 to 2012", *Computer Communications* 50 (2014) 64
- <sup>4</sup>IEA (2017), *Digitalization and energy*, IEA Publications, Cedex, Paris
- <sup>5</sup>European Commission (2015), *Eco-design preparatory study on enterprise servers and data equipment*. Luxembourg: Publications Office of the European Union
- <sup>6</sup>E. Masanet, et al, "Recalibrating global data center energy-use estimates", *Science* 367 (2020) 984
- <sup>7</sup>H. Fuchs et al, "Comparing datasets of volume servers to illuminate their energy use in data centers", *Energy Efficiency* 13 (2020) 379
- <sup>8</sup>Malmodin et al., "The future carbon footprint of the ICT and E&M sectors Proceedings of the 1st International Conference on Information and Communication Technologies for Sustainability ETH Zurich, February 14-16, 2013 pp. 12-20
- <sup>9</sup>A. S. G. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030", *Challenges* 6 (2015) 117-157
- <sup>10</sup>L. Belkhir and A. Elmelig, "Assessing ICT global emissions footprint: Trends to 2040 & recommendations", *J. Cleaner Production* 177 (2018) 448
- <sup>11</sup>Roy Campbell, "Towards zettascale computing", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>12</sup><https://www.af.mil/About-Us/Fact-Sheets/Display/Article/104467/x-51a-waverider/>
- <sup>13</sup>David Wentzlauff, "Limits of Architectural Innovation", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>14</sup>R. Aitken. "Moore's Law Ending? No Problem", *EETimes*, March 27, 2019, <https://www.eetimes.com/moores-law-ending-no-problem/#>.
- <sup>15</sup>John Owens, "Future Hardware for Computation", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>16</sup>Rick Stevens, Valerie Taylor, Jeff Nichols, Arthur Barney Maccabe, Katherine Yelick, and David Brown, "AI for Science", February 1, 2020, <https://doi.org/10.2172/1604756>
- <sup>17</sup>Naresh Shanbhag. "Computing for the Nanoscale Era", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>18</sup>N. Shanbhag et al. "Shannon-Inspired Statistical Computing for the Nanoscale Era" *Proc. IEEE* 107 (2019) 90-107
- <sup>19</sup>Yichen Shen. "Reinventing computing for AI using integrated photonics". *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>20</sup>Stephen Trimberger, "Cryogenic Computing", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>21</sup>H. L. Chiang et al., "Cold CMOS as a Power-Performance-Reliability Booster for Advanced FinFETs", *IEEE 2020 VLSI Technology Symposium*.
- <sup>22</sup>Dmitri Nikonov and Ian Young. "Computing in 2050: What does physics have to say?" *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019
- <sup>23</sup>Manipatruni, S. et al. (2019). Scalable energy-efficient magnetoelectric spin-orbit logic". *Nature*, 565(7737), 35-42,42A
- <sup>24</sup>Robert D. Clark. "Materials and Processes to Support New Compute Trajectories". *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019
- <sup>25</sup>Fred Streit, "Artificial Intelligence: View from the DOE", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019
- <sup>26</sup>Anand Raghunathan. "In-memory computing for next-generation AI Hardware." *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019
- <sup>27</sup>P. Kanerva, "Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors," ed: *Cognitive Computing*, 2009, pp. 139-159.
- <sup>28</sup>F. Montagna, A. Rahimi, s. Benatti, D. Rossi, and L. Benini, "PULP-HD: Accelerating Brain-Inspired High-Dimensional Computing on a Parallel Ultra-Low Power Platform, <https://arxiv.org/pdf/1804.09123.pdf>," ed, 2018.
- <sup>29</sup>Queensland Brain Institute. <https://qbi.uq.edu.au/brain-basics/brain/brain-physiology/how-do-neurons-work> (accessed 2020).
- <sup>30</sup>Brad Aimone, "Computing with spikes," *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>31</sup>D. Monroe, "Neuromorphic computing gets ready for the (really) big time", *Communications of the ACM* 57 (2014) 13-15
- <sup>32</sup>Aaron Voelker, "The need for neuromorphics," *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.

- <sup>33</sup>S Ghosh-Dastidar, H Adeli. "Spiking neural networks". *International Journal of Neural Systems* 19 (2009), 295-308
- <sup>34</sup>D. S. Modha, "Introducing a Brain-inspired Computer: TrueNorth's neurons to revolutionize system architecture", <https://www.research.ibm.com/articles/brain-chip.shtml>
- <sup>35</sup>M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro* 38 (2018) 82-99.
- <sup>36</sup>W. Severa, C. M. Vineyard, R. Dellana, S. J. Verzi, and J. B. Aimone, "Training deep neural networks for binary communication with the Whetstone method," *Nature Machine Intelligence*, vol. 1, no. 2, pp. 86-94, 2019/02/01 2019, doi: 10.1038/s42256-018-0015-y.
- <sup>37</sup>A. R. Voelker, D. Rasmussen, C. Eliasmith, "A Spike in Performance: Training Hybrid-Spiking Neural Networks with Quantized Activation Functions" *arXiv:2002.03553*, 2019.
- <sup>38</sup>J. B. Aimone, "Neural Algorithms and Computing Beyond Moore's Law," *Communications of the ACM*, vol. 62, p. 110, 2019.
- <sup>39</sup>Aimone, J. B., Hamilton, K. E., Mniszewski, S., Reeder, L., Schuman, C. D., & Severa, W. M. (2018). Non-neural network applications for spiking neuromorphic hardware. In *Proceedings of the Third International Workshop on Post Moore's Era Supercomputing* (pp. 24-26).
- <sup>40</sup>Aimone, J. B., Hamilton, K. E., Mniszewski, S., Reeder, L., Schuman, C. D., & Severa, W. M. (2018). Non-neural network applications for spiking neuromorphic hardware. In *Proceedings of the Third International Workshop on Post Moores Era Supercomputing* (pp. 24-26).
- <sup>41</sup>Stefano Ambrogio, "Brain-inspired computing using phase-change memory devices," *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>42</sup>Don Norman, "Human-technology collaboration", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>43</sup>Sparsh Mittal. "A Survey of Techniques for Approximate Computing" *ACM Computing Surveys*, Vol. 48, No. 4, Article 62 (2016)
- <sup>44</sup>S. F. Yitbarek, T. Yang, R. Das and T. Austin, "Exploring specialized near-memory processing for data intensive operations" 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, 2016, pp. 1449-1452.
- <sup>45</sup>Ielmini, D., Wong, H.P. "In-memory computing with resistive switching devices" *Nature Electronics* 1 (2018) 333-343.
- <sup>46</sup>W. Rao, A. Orailoglu and R. Karri, "Logic Mapping in Crossbar-Based Nanoarchitectures" *IEEE Design & Test of Computers*, vol. 26, no. 1, pp. 68-77, Jan.-Feb. 2009.
- <sup>47</sup>F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2018, pp. 0210-0215.
- <sup>48</sup>S. K. Moore, "Huge chip smashes deep learning's speed barrier", *Spectrum* 57 (2020), pp. 24-27.
- <sup>49</sup>H. El-Sayed et al., "Edge of Things: The Big Picture on the Integration of Edge, IoT and the Cloud in a Distributed Computing Environment", *IEEE Access* 6 (2018) 1706-1717.
- <sup>50</sup>Shimeng Yu. "Devices and Materials Requirements for Artificial Intelligence", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>51</sup>Steven Lee, "Scientific Machine Learning and AI", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019
- <sup>52</sup>Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence, <https://doi.org/10.2172/1478744>
- <sup>53</sup>J. B. Aimone, "A Roadmap for Reaching the Potential of Brain-Derived Computing"
- <sup>54</sup>Arute, F., Arya, K., Babbush, R. et al. Quantum supremacy using a programmable superconducting processor. *Nature* 574, 505-510 (2019).
- <sup>55</sup>Oliver Dial, "Superconducting qubits", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, United States of America, October 15-16, 2019.
- <sup>56</sup>Christopher Monroe, "The State of Quantum Computing and Future Outlook", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019
- <sup>57</sup>D. P. DiVincenzo, "The Physical Implementation of Quantum Computation," ed. *arXiv:quant-ph/0002077*, 2000.
- <sup>58</sup>Michael Biercuk, "Mitigating Noise and Error in Quantum Processors", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019
- <sup>59</sup>Edoardo Charbon, "Electronic Interface for Quantum Processors", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, Oct. 15-16, 2019.
- <sup>60</sup>M. Brooks, "Beyond quantum supremacy: the hunt for useful quantum computers", *Nature* 574 (2019) 19-21.
- <sup>61</sup>Chad Rigetti, "Full-stack, hybrid quantum-classical computing with superconducting qubit", *Workshop on the New Compute Trajectories for Energy-Efficient Computing*, Sandia National Laboratory, Livermore, California, United States of America, October 15-16, 2019.
- <sup>62</sup>M. Hilbert and P. Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information", *Science* 332 (2011) 60-65
- <sup>63</sup>GPU Statistics: <https://owensgroup.github.io/gpustats/>







Semiconductor  
Research  
Corporation

4819 Emperor Blvd  
Suite 300  
Durham, NC 27703

919.941.9400  
[www.src.org](http://www.src.org)