U.S. DEPARTMENT OF ENERGY

Office of ENERGY EFFICIENCY & RENEWABLE ENERGY

Advanced Manufacturing Office

Workshop on Manufacturing and Integration Challenges for Analog and Neuromorphic Computing

Workshop Report

August 11-13, 2021

Within the DOE Office of Energy Efficiency and Renewable Energy (EERE), the Advanced Manufacturing Office (AMO) collaborates with industry, small business, universities, national laboratories, state and local governments, and other stakeholders on emerging manufacturing technologies to drive U.S. industrial decarbonization, economic competitiveness, and energy productivity. AMO has a mission to develop technologies that reduce manufacturing energy intensity and industrial carbon emissions; increase the competitiveness of the U.S. manufacturing sector, with a focus on clean energy manufacturing; and reduce the life cycle energy and carbon impact of manufactured goods in the industry, buildings, transport, power, and agricultural sectors.

This document was prepared for DOE/EERE's AMO collaborative effort by DOE AMO and Energetics Incorporated.

Disclaimer

List of Acronyms

5G	Fifth Generation (of tele-communications technology)
ADC	Analog-to-digital converter
AI	Artificial intelligence
AMO	Advanced Manufacturing Office
ASIC	Application specific integrated circuit
CAD	Computer aided design
CHIPS	Creating Helpful Incentives to Produce Semiconductors
CMOS	Complementary metal oxide semiconductor
CNN	Convolutional neural network
CNT	Carbon nanotube
COMPETES	Creating Opportunities for Manufacturing Pre-Eminence in Technology and Economic Strength
CRS	Congressional Research Service
DOE	Department of Energy
ECRAM	Electrochemical random-access memory
EDA	Engineering design automation
EERE	Office of Energy Efficiency and Renewable Energy
FPAA	Field programmable analog arrays
FPGA	Field programmable gate arrays
GaN	Gallium nitride
GHG	Greenhouse gas
GHz	Gigahertz
GAO	General Accounting Office

GPU	Graphics processing unit
I/O	Input-output
ML	Machine learning
mmWave	Millimeter wave
NDAA	National Defense Authorization Act
NLP	Natural language processing
NIST	National Institute of Standards and Technology
PVT	Process-voltage-temperature
RDD&D	Research, development, demonstration, and deployment
ReRAM	Resistive random-access memory
RF	Radio frequency
SC	Department of Energy Office of Science
SIA	Semiconductor Industry Association
SiC	Silicon carbide
SRC	Semiconductor Research Corporation
TOPS	Tera operations per second
USICA	United States Innovation and Competition Act

Executive Summary

The U.S. Department of Energy's (DOE's) Advanced Manufacturing Office (AMO) held its third virtual workshop on Semiconductor R&D for Energy Efficiency on August 11-13, 2021. This public workshop – titled "Manufacturing and Integration Challenges for Analog and Neuromorphic Computing" – brought together more than 180 leading scientific and technical experts to identify challenges and opportunities to improve manufacturing capabilities for analog and neuromorphic computing architectures to significantly increase energy efficiency of microelectronic devices. The workshop featured perspectives of researchers from national laboratories, universities, government agencies, and industry with expertise in novel and emerging devices, processes, and metrology and characterization. This workshop is intended to inform an AMO research, development, demonstration, and deployment (RDD&D) plan to significantly reduce semiconductor energy use by 2030.[1] It was co-sponsored by the Semiconductor Research Corporation (SRC) and the DOE Office of Science's Advanced Scientific Computing Research (ASCR).

Discussions at the workshop explored four major topics: 1) applications and impacts of analog and neuromorphic computing; 2) analog hardware for communications; 3) analog hardware for sensors; and 4) neuromorphic architecture and devices.

The key takeaways from the workshop discussions can be grouped in three categories: grand challenges and solutions to enable energy efficient microelectronics, technology problems that must be overcome to address the grand challenges, and the technology pathways that must be explored to solve the necessary technology problems, Figure ES-1.

Grand challenges: Two grand challenges in energy efficiency emerged throughout workshop discussions.

Data deluge: Over the next decade, the amount of generated data is projected to be >1000x greater than what is possible for human consumption (SRC 2021). The deployment of sensors and internet-of-things (IoT) devices in all sectors have caused an explosion in data generation, and methods in which to extract useful information must drastically improve. Sending this data to the cloud or other "off-node" hardware for analysis and storage is driving unsustainable energy consumption. Moreover, most data from these nodes are sparse and in differing formats, often incompatible with one another, leading to further computational (and energy) intensity. Solutions include methods for edge or near-edge computation that only communicate pertinent data and hence save time and enormous amounts of energy. In the case of industrial automation, the subset of pertinent information is much smaller and less energy intensive to filter than in less constrained applications such as autonomous vehicles (AV).

Energy consumption of artificial intelligence/machine learning (AI/ML): AI/ML has enabled advances in many disparate fields, including manufacturing, communications, medicine, and transportation, through rapid analysis of large data sets. While these advances benefit society, the energy used to train and operate these models is increasing orders of magnitude faster than efficiency can be increased. A recent study shows that cutting-edge AI doubles its energy consumption for training every 2-3 months (Mehonic and Kenyon 2021). Furthermore, current neuromorphic computing approaches typically rely on traditional complementary metal oxide semiconductor (CMOS) devices, limiting computational performance and energy efficiency compared

[1] This workshop report summarizes the presentations, panel discussions, and facilitated discussions that took place at this event. More detailed summaries are available in the Appendix. Note that the results presented here are a snapshot of the viewpoints expressed by the experts who attended the workshop and do not necessarily reflect those of the broader semiconductor research and manufacturing communities.

with devices optimized for neuromorphic computing. Solutions include novel devices, architectures, and circuitry such as application specific integrated circuits (ASICs) designed for neuromorphic computing that minimize energy consumption while maximizing performance. Solutions presented, such as IBM's neuromorphic hardware approach that would more than double (2.5x) energy efficiency annually, are promising, but still insufficient to counter the rapid and accelerating doubling times of energy use by AI. The highest leverage energy efficiency solutions may lie in application-specific ML software and algorithms. Just as ASICs can be customized to reduce hardware energy use 1000-fold, perhaps physics-driven application-specific AI/ML algorithms can do the same to reduce software energy use.

Technology Problems: To address the grand challenges, a number of technological problems were identified; three are discussed below.

Local processing: Local processing is needed to improve energy efficiency of sensors and IoT, as well as address the data deluge, by only transmitting necessary data, while discarding the rest. Processing analog signals, prior to digitization, by using field programmable analog arrays (FPAAs) or other devices and circuitry in conjunction with traditional digital computation is a promising approach. When such tools use AI/ML, they must include application-specific training to minimize energy use. Challenges include integration of local processing approaches or devices with existing circuitry and the lack of computational tools for design and modeling/simulation.

Hardware optimized for neuromorphic computing: Hardware needs to be optimized for specific applications in order to achieve performance and energy efficiency metrics (e.g., up to 1000x improved energy efficiency) currently not possible using conventional devices and architectures. Neuromorphic-optimized photonics, organic semiconductors, and electrochemical devices are promising approaches that are being explored. Challenges include material compatibility and integration with existing CMOS, long-term device stability, and device scaling.

Advanced packaging: Advanced packaging is needed to enable continued improvement in performance, energy efficiency, and cost through functional diversification (i.e., heterogenous integration), increases in device density on the package level, and reduction in interconnect distances with 3D approaches. As advanced packaging shrinks to the micro-level, new energy efficiency problems may emerge that must be overcome. Challenges include I/O routing, thermal management, and metrology.

Technology Pathways: To enable advances in the technology problems, four technology pathways of primary importance are highlighted below.

Material development and process integration: Advanced materials and manufacturing processes will enable next generation devices for analog and neuromorphic computing. Emerging devices such as electrochemical random-access memory (ECRAM) and organic semiconductors, optics, and atomically precise manufacturing are promising approaches driving more efficient computation and communication. The primary challenge will be process integration with existing processes and tools. Thermal budget, strict contamination compliance, and the sequence of process flow must be considered carefully.

Co-design: Energy efficiency must be considered at all levels of the stack (i.e., device, circuit, architecture, and algorithms) to ensure efficiency improvements at one level are realized on the system level. Due to the increased complexity of neuromorphic systems, compared with conventional devices and architectures, co-design is necessary. A wide range of experts, from device physicists and circuit designers to biologists and

neuroscientists, must collaborate to design and manufacture energy-efficient, high performance neuromorphic systems.

Electronic Design Automation (EDA)/Computer Aided Design (CAD): EDA/CAD tools are essential to designing, developing, and deploying novel semiconductor technologies. They allow for iterative development to optimize designs to achieve key metrics before actual devices are fabricated. Existing tools must integrate novel materials, processes, and device physics to enable commercialization of next generation devices and help validate results. Information sharing between academic researchers and industry can accelerate this development. With advanced packaging/heterogenous integration becoming the go-to approach to improve performance, AI/ML guided design tools that can holistically evaluate multi-scale phenomena (circuit to systems) is needed.

Workforce Development: Workforce development is a central issue in the semiconductor industry. Throughout the workshop series, participants have highlighted the difficulty in attracting top students to hardware-related electronics curricula and jobs. The industry is struggling to compete against "hot" fields such as AI/ML and software development. Even within AI/ML, the focus is more on mathematics than on real-world physics-based training. A fundamental change in the curricula, recruitment, and messaging is needed to ensure the domestic workforce has sufficient expertise to lead the industry in all key areas of innovation.

Figure ES-1: Key topics from the workshop discussions are summarized in three categories: Grand Challenges, Technology Problems, and Technology Pathways.

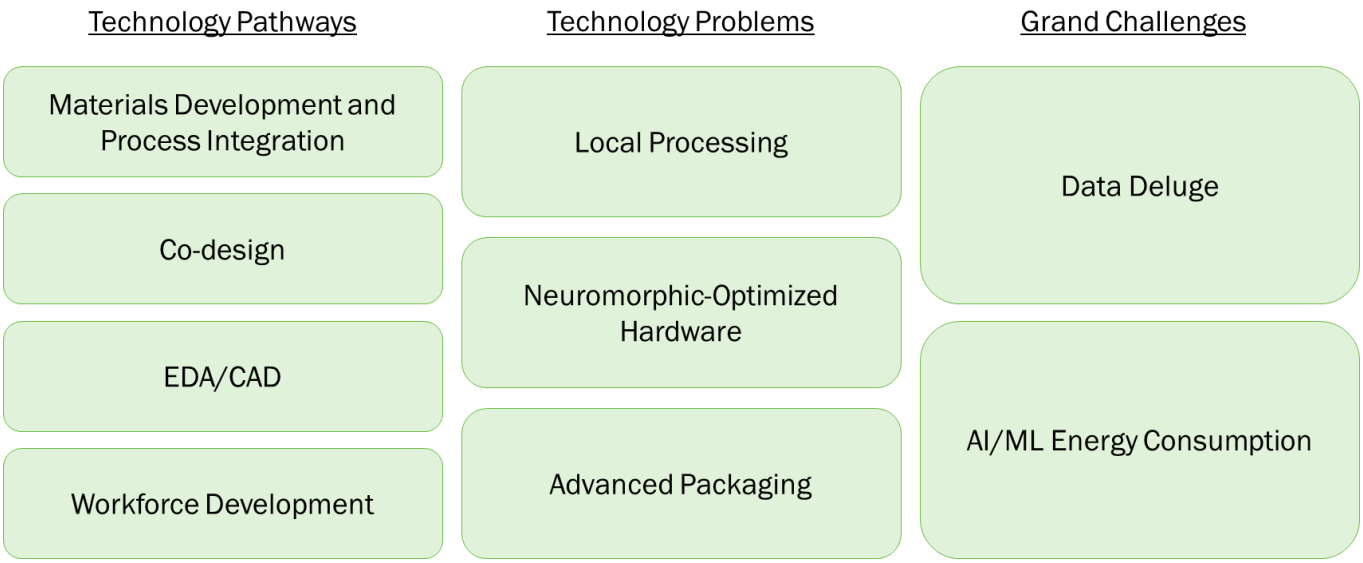

Table of Contents

List of Figures

List of Tables

Background

On August 11-13, the Department of Energy's Advanced Manufacturing Office (AMO) within the office of Energy Efficiency and Renewable Energy (EERE), with co-sponsors DOE Office of Science and the Semiconductor Research Corporation (SRC), held the third in its ongoing series of workshops on different topics related to semiconductor research and development (R&D) to increase energy efficiency. This workshop focused on manufacturing and integration challenges for analog and neuromorphic computing. In addition to the industry needs and RDD&D opportunities, AMO's goals on environmental protection and a new goal on industrial greenhouse gas (GHG) reductions by 2050 were also addressed. The output of this workshop will inform AMO's future R&D portfolio investments; provide perspectives on trends, drivers, and challenges for analog and neuromorphic computing; and help the stakeholder community understand the opportunities on the horizon.

Workshop Series

Semiconductors power key products that are rapidly growing in importance in all sectors of the economy including consumer goods, finance, transportation, and manufacturing. Advances in semiconductor technology are critical for global competitiveness as well as economic, national, and climate security. According to the 2021 World Semiconductor Trade Statistics report, the global semiconductor industry, in 2021, had just under $553 billion in sales, up from $440 billion in 2020, with the Americas accounting for roughly 20%. In 2020, according to Semiconductor Industry Association (SIA), semiconductors were the fourth largest U.S. export ($49 billion) behind aircraft, refined oil, and crude oil. An October 2020 Congressional Research Service (CRS) report divides the industry into four segments, shown in Figure 1. AMO's first workshop focused on integrated sensor systems (20%). AMO's second workshop focused on logic, communications (42%) and memory (25%) devices. This third workshop focuses on both analog devices (13%) and an emerging class of brain-inspired (neuromorphic) hardware that can incorporate both digital and analog elements.

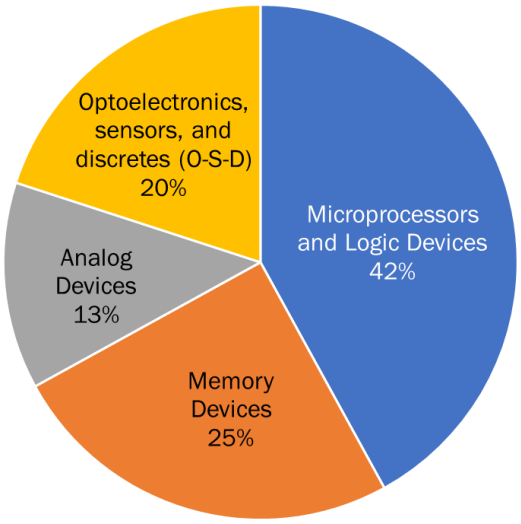

Figure 1: Market segments of the semiconductor industry and the breakdown of sales, in percentages, data year: 2020 (CRS 2020).

Because of its economic importance and the potential for its products to improve quality of life and reduce energy use and GHG emissions in other sectors, growth of the semiconductor industry is desired; but innovation is needed to ensure this growth is accompanied by major improvements in energy efficiency of its products. In the past, the energy use of semiconductor industry products was flat or declining due to efficiency increases related to miniaturization. However, since 2010, as shown in Figure 2, it has begun to dramatically increase—doubling every three years (SIA 2019).

Multiple trends in semiconductor-related energy use are combining to make increased energy efficiency of semiconductor products a top priority for the industry and the federal government. These trends include the exploding electricity use of new applications ranging from Bitcoin to AI/ML algorithms; the decrease in improvements in energy consumption per chip; and the acceleration of the use of microelectronics for

decarbonization—especially electrification (DOE AMO 2022)—which, together have led to the exponential growth in energy use of semiconductor industry products.

Separate from the growing computational energy use, the recent shortage in semiconductor products that has disproportionately affected multiple supply chains – in particular, the auto-industry – has prompted increased government scrutiny of semiconductor manufacturing and its supply chain. The U.S. Congress passed into law in its fiscal year 2021 National Defense Authorization Act (NDAA) unfunded authorizations for significant activities in its version of the "Creating Helpful Incentives to Produce Semiconductors (CHIPS) for America Act", to increase domestic semiconductor manufacturing by authorizing massive increases in Federal R&D and supply chain and manufacturing support. In June 2021, the Senate passed the United States Innovation and Competition Act of 2021 (USICA), that establishes a Directorate for Technology and Innovation at the NSF aimed at specific critical technologies, including semiconductors, and authorizes specific funding amounts for its version of the CHIPS Act (S.1260 2021). In February 2022, the House passed its own version of this competitiveness legislation Creating Opportunities for Manufacturing Pre-Eminence in Technology and Economic Strength (COMPETES Act) that also included authorization of appropriations for its variation on the CHIPS Act as well as a new Microelectronics R&D activity within the DOE Office of Science. In his State of the Union Address, President Biden said that the two legislative branches should reconcile the legislation so that he could promptly sign it. On May 12, 2022, the House and Senate held their first conference committee meeting to reconcile differences in the two bills including the CHIPS provisions—the conference should result in a Bipartisan Competitiveness bill that can be passed in both Houses of Congress by the end of Summer 2022.

An additional driver for AMO efforts with respect to semiconductor energy efficiency is the Biden Administration's goal of cutting GHG emissions by 50% by 2030 through aggressive industrial decarbonization and electrification. This has driven an increased EERE interest in developing more energy efficient semiconductor devices. U.S. leadership in manufacturing and deployment of these devices can lead to a re-shoring of semiconductor manufacturing foundries spurring domestic job creation, increase productivity and competitiveness of the U.S. manufacturing industry, and combat the climate crisis through reduced energy consumption across all sectors that utilize semiconductor technology. Other Biden Administration carbon goals that drive semiconductor efficiency R&D include goals for a zero-carbon grid by 2035 and the overall goal of a net zero carbon economy by 2050. These goals all increase the urgency of deploying decarbonization technologies such as ultra-energy-efficiency, massive electrification, and increased digitalization that increase use of semiconductors.

Workshop Motivation

Since 2010, semiconductor energy use has doubled every 3 years. By 2030, semiconductors could consume nearly 25% of planetary energy production, Figure 2 (SRC 2021). Innovation in semiconductor device and architecture energy efficiency is essential not just for the economy, but also to address the climate crisis. In particular, advances in modeling, simulation, artificial intelligence, machine learning, and other analysis and discovery techniques are driving innovations in manufacturing processes and approaches, that have the potential to reduce energy consumption, and thereby limit GHG emissions. While computing approaches now underpin the success of nearly every modern industry, significant slowing in efficiency improvements and exponentially increasing electricity use is emerging in some important applications. To permit continued U.S. participation and manufacturing growth in these application areas, the underlying microelectronic systems need to continue to improve as well.

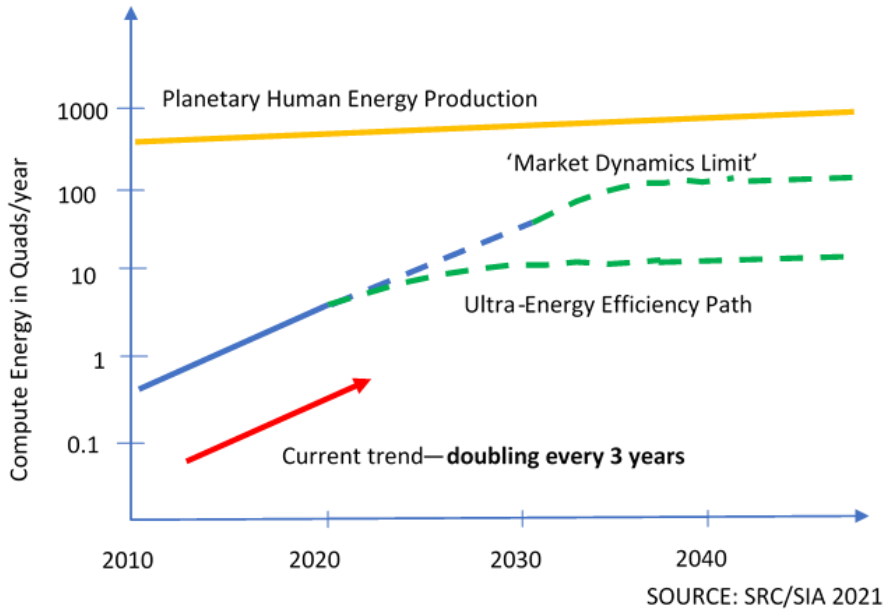

Figure 2: SRC projects that absent significant increase in federal contributions to semiconductor R&D for energy efficiency, the "Market Dynamics Limit" will be reached by 2035, limiting the world's computing capacity and economic growth. Alternatively, prioritizing ultra-energy-efficiency in semiconductor products can achieve a trajectory in which computing/economic growth and energy use are decoupled.

Unlike Workshops 1 and 2 that were purely focused on hardware, this workshop discussed software and algorithms, in cooperation with Office of Science (SC), and the new hardware required to take advantage of such algorithms. Leveraging both novel algorithms/software and hardware is the only way to reach the 1000-fold increase in overall energy efficiency needed to avoid the "market dynamics limit." In particular, how to leverage these alternative computing architectures, including analog and neuromorphic computing, was the focus of this workshop. While the SC-led session focused on non-traditional computational architectures, AMO-led sessions focused on associated analog hardware. Together with the inherently more efficient hardware addressed in earlier workshops, these neuromorphic devices and architectures provide pathways to reduce computational energy use by 1000-fold in key applications.

Though analog computing has a long history, new system opportunities enabled by new hardware and software exist. Communication and sensing are two applications where advanced analog hardware can drastically improve energy efficiency of microelectronic systems. Due to the significant energy penalty, reducing data generation and transmission for integrated sensor systems was one of the major conclusions of the first workshop. In addition, analog devices that utilize novel materials or architectures (such as neuromorphic) have been shown to provide as much as 1000-fold energy and performance improvements in certain applications, compared with digital devices. Performing computation on analog signals (e.g., sensor input and RF[2] signals) in the analog domain (before conversion to digital memory) maintains signal fidelity and in many cases can optimize computing efficiency in ways digital computation cannot. Similarly, analog signal processing can greatly improve the bandwidth and efficiency of wireless communication.

AI/Machine-learning (ML) has emerged as a major driver of neuromorphic architectural approaches for applications including advanced process control/optimization, object recognition, and speech recognition. Unfortunately, the use of traditional von-Neuman hardware architectures for these AI/ML software

[2] Radio frequency (RF) refers to electromagnetic waves in the frequency or band of frequencies in the range 100MHz to THz, suitable for use in telecommunications.

applications consumes an enormous amount of electricity. Advanced hardware for neuromorphic computing is far less (up to 1000x less) energy intensive than using a software only approach with conventional digital CMOS-based technologies.

Identifying the emerging technologies and key challenges and R&D pathways for analog and neuromorphic hardware can illuminate RDD&D pathways to drastically reduce semiconductor computational energy use and accelerate the transition toward a sustainable path that avoids planetary energy impacts, while revitalizing a key domestic industry that offers high-paying jobs. By partnering with U.S. industry to further develop these technologies based on non-traditional architectures, AMO hopes to increase the competitiveness of domestic device and chip manufacturing, spur domestic job creation in this growing field, and combat the climate crisis by flattening the curve of semiconductor energy consumption across all sectors that use semiconductors by 2030.

Workshop Overview

To better understand the challenges and opportunities in improving device and manufacturing capabilities for analog and neuromorphic computing architectures and AMO's role in this area, the U.S. Department of Energy AMO, co-sponsored by DOE Office of Science and SRC, held the Workshop on Manufacturing and Integration Challenges for Analog and Neuromorphic Computing on August 11-13, 2021. Representatives from industry, academia, the DOE national laboratories, and non-governmental organizations gathered virtually to hear presentations and participate in panel discussions with subject matter experts, as well as contribute to topical facilitated discussions/brainstorming sessions. The workshop was divided into four technical topics: 1) applications, benefits, and metrics of analog and neuromorphic computing; 2) analog hardware for communications; 3) analog hardware for sensors; and 4) neuromorphic architecture and devices. This workshop report summarizes the promising technologies and RDD&D challenges and opportunities identified for the four technical workshop topics from the presentations, panels, and facilitated discussions.

Below is a brief overview of the workshop agenda. More detailed summaries of all of the talks are included in Appendix B.

On the first day of the workshop, participants learned about efforts of the federal government in coordinating semiconductor R&D activities, the unsustainable energy consumption of leading-edge AI models, and the applications and impacts of deploying analog and neuromorphic computing. Plenary talks featured speakers from DOE, the National Science and Technology Council, and Stanford University. The day concluded with a panel discussion and facilitated session on applications, benefits, and metrics of analog and neuromorphic computing.

The second day began with opening remarks from the SRC, a co-sponsor of the workshop, followed by two technical sessions. The first comprised a panel discussion on analog hardware for communications and included a talk from one of AMO's current projects followed by an extensive Q&A. The panel discussed key opportunities and challenges in developing next generation communications technologies. The second featured two technical talks, discussing FPAA for high efficiency sensor systems, and an overview of intelligent sensing. The talks were followed by combined Q&A session. Both technical sessions were followed by a facilitated discussion.

The third and final day opened with an introduction from DOE Office of Science (SC), a co-sponsor of the workshop. The talk gave a general overview of SC and the importance of semiconductors to its basic science research. Following the opening remarks, the workshop moved on to the final technical session, comprising two panel sessions and a final facilitated discussion. The first panel discussed brain-inspired computational approaches and included perspectives from biology, electrical engineering, and software engineering. The

second panel discussed neuromorphic hardware, including promising technologies such as electrochemical non-volatile memory, organic semiconductors, and co-packaged optics. Both panel sessions were followed by extensive Q&A and the workshop ended after a closing discussion.

Applications, Benefits, and Metrics of Analog and Neuromorphic Computing

To explore the breadth of applications and benefits derived from analog and neuromorphic hardware and set the stage for subsequent discussions, the workshop opened with the Analog and Neuromorphic Applications and Impacts panel. It included talks from Gina Adam (George Washington University), Brian Calvert (AI startup), Gabriella Carini (Brookhaven National Laboratory), and Steve Spurgeon (Pacific Northwest National Laboratory).

Gina Adam highlighted the growing performance gap between processors and memory, with processors now having superior speed compared to memory. She described how analog technologies that have higher speed, smaller size, and greater energy efficiency, can be deployed for data-driven memory applications that are more tolerant to analog's lower precision computing to close the performance gap. Brian Calvert described how custom analog architectures could provide fast, high-volume data processing demanded by cutting-edge applications such as high-energy physics, astronomy, genomics, and autonomous vehicles. Gabriella Carini discussed how high-energy physics detectors can benefit from analog and neuromorphic computing architectures that mimic the brain's ability to reduce incoming data from the senses. Steve Spurgeon discussed how neuromorphic computing can enable scientists' to better interpret and act on large volumes of data using domain-grounded data reduction and inference from prior research.

A complete list of applications, impacts, benefits, and metrics of analog and neuromorphic computing, identified by workshop participants, can be found in Appendix C, Table C-1 to Table C-5.

Table 1. Participant Input on Applications, Metrics, and Benefits of Analog and Neuromorphic Computing	
Applications	
High Impact Application Areas	**Emerging Applications**
• Edge computing for autonomous systems • In-situ data analysis for systems control • Advanced, data-intensive scientific research • Advanced electric grid control • Machine learning • High performance computing.	• Non-machine learning (ML) use cases of neuromorphic computing, such as graph and optimization algorithms • Extremely small form factor wearables • Very low power multi-sensor fusion • Physical systems modeling, such as chemical reactions.
Metrics to measure performance	**Benefits of deploying analog and neuromorphic computing**
• Time to solution for a specific error rate • Bit/joule • Bit/m^2 • Tera operations per second (TOPS)/watt • Latency • Energy/unit information • AI model training time and energy.	• Privacy of data (due to locality) • Less reliance on cloud infrastructure and thus less communication energy used • New optimization techniques and methods • More efficient processing and reduced latency for inference and data filtering • Improved computational performance at lower power.

Analog Hardware

Analog hardware encompasses a wide range of technologies that process continuously variable signals of many shapes and sizes (versus digital electronics that process only 1's and 0's). According to the Congressional Research Service, analog devices account for roughly 13% or $54 billion of the global semiconductor market, Figure 1 (Platzer, Sargent, and Sutter 2020).

Analog hardware is used in all semiconductor applications, Figure 3. This workshop focused on two specific application areas: physical world interfaces (i.e., sensors) and communication where energy efficiency opportunities abound.

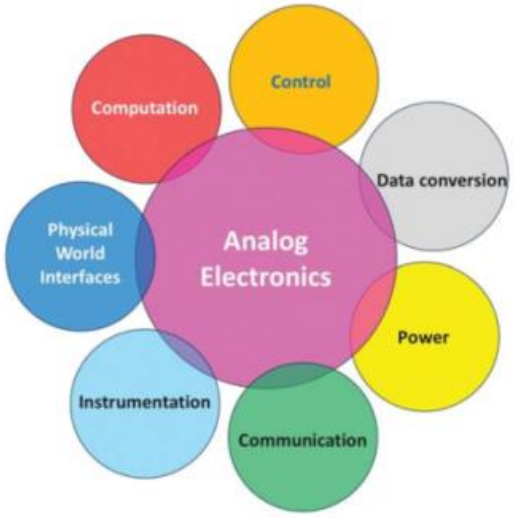

Figure 3: The domains in which analog electronics are used (SRC 2021)

As consumer use of high-tech goods continues to explode, partly from the pandemic and high levels of virtual work, communication technologies have not kept pace with the exponentially increasing data that result from use of these goods. Although the transition to fifth generation (5G) telecommunications technology is far from complete, higher frequencies are increasingly being utilized, thus advanced analog electronics that can accommodate higher frequencies, better than digital electronics, will play a central role in enabling devices to use terahertz (THz)[3] frequencies likely required for 5G+. The transition to 5G and beyond (which could include an order of magnitude increase in base station energy use), coupled with the overall semiconductor energy consumption crisis, requires more energy efficient technologies to maximize system-spectral efficiency and thus minimize energy consumption (SRC 2021).

Analog Hardware for Communication

The exponential growth of IoT devices coupled with ever increasing data consumption, and manufacturers generating significantly larger amounts of data to monitor and control their processes, is part of an overall trend causing what SRC refers to as a "data deluge." In the first workshop, data communication was highlighted as the most energy intensive process in the operation of microelectronic devices—thus the data deluge also is an energy consumption problem. As manufacturers look to further automate their plant operations, maintaining low levels of energy use while transmitting larger amounts of data will be critical. Novel ways to transmit data, efficiently and without loss, whether through intelligent edge nodes, higher frequency regimes, or advanced device technology, are needed to meet the growing demand for sustainable data transmission (SRC 2021).

The Analog Hardware for Communications session included a panel session and a facilitated discussion. The panel consisted of Thomas Cho (Samsung), Jim Booth (NIST), and Steffen McKernan (Carbon Technology Inc.).

Thomas Cho outlined the history of CMOS in mobile devices and the technical challenges facing the mobile industry, including slow adoption of new processes for RF, crowded spectrum, and the development of new radio architectures to combat propagation loss at higher frequencies. He closed by highlighting key considerations and potential R&D pathways for developing next generation energy-efficient analog communications technologies such as beamforming, chiplet based systems, and carrier aggregation. Jim Booth gave an overview of 5G communications, discussed approaches for analog electronics that can improve spectral and energy efficiency. These include opportunities for adaptive materials and devices for communication technology, including nonlinear dielectric materials, liquid crystals, and phase change materials. Steffen McKernan reviewed the material and electronic properties of carbon nanotube (CNT) based electronics that make it a promising technology for energy efficient RF communications, such as high linearity

[3] THz refers to electromagnetic waves in the superhigh 300 GHz to 3 THz frequency range.

which can lead to 75% improvement in energy efficiency. A more detailed summary of each panel speaker's remarks can be found in Appendix B.

Table 2, on pages 10-11 summarizes participant input from the facilitated session following the panel that was focused on promising technology and R&D challenges and opportunities in the panel subject area. Table C-6 to Table C-10 in Appendix C include the full participant input.

Promising Technologies

Below is a summary of the promising technologies for communications that were discussed at the workshop.

CNT: CNTs' one-dimensional geometry has inherent linearity, which leads to a high dynamic range. This makes CNT-based devices more energy-efficient and also well suited to electromagnetically noisy environments – such noisy environments were identified as a major challenge for use of sensors in manufacturing facilities during our first workshop. In addition, CNT's high carrier mobility, high saturation velocity, and ballistic transport allow for extremely low power operation and high energy efficiency. Recent research advances supported by AMO include overcoming challenges of growing or depositing pristine CNTs with tight diameter control and semiconducting-to-metallic CNT selectivity. Overcoming these challenges will result in a pathway for CNT devices to be a viable alternative to conventional silicon-based technologies. The ability to fabricate at manufacturing-relevant CNT selectivity, density and uniformity, results in higher performance and more stable CNT devices that can out-compete less energy efficient conventional devices.

Wide bandgap materials: Gallium arsenide- and silicon germanium-based devices have been used for RF systems for decades. Next generation wireless communications operating at extremely high frequencies, over 100 gigahertz (GHz), will require amplification unachievable with conventional Si-based technologies due to their increased signal attenuation of high frequency signals. Wide bandgap semiconductors are ideal candidates for >100 GHz communication given their higher output power and operating temperature and voltage. In particular, gallium nitride (GaN) boasts faster switching speeds, higher thermal conductivity, and lower on-resistance, translating to improved energy efficiency over conventional Si-based devices. Silicon carbide (SiC) is also a well-characterized substrate for RF applications and many products that incorporate it are coming onto the market. However, both GaN and SiC, as well as other emerging wide bandgap materials such as diamond, aluminum nitride, and gallium oxide, face challenges due to high frequencies' higher sensitivity to impurities and other defects in crystal growth which results in lower yields that raise costs, and thus limit deployment. Manufacturing R&D approaches focused on ultra-precision control, such as those explored in our second workshop, may be well suited to overcoming such challenges.

Co-packaged optics: The need to reduce energy consumption by shortening interconnects and increase bandwidth by using the speed of light rather than electrons is increasing interest in (1) co-packaging and (2) optics as an emerging approach for short-to-medium distance connections, such as chip-to-chip communications in data centers and backhaul networks (IRDS 2021). A combination of these two solutions, co-packaged optics – where the optical module is packaged with the silicon switch – would replace an incumbent technology comprising copper interconnects and pluggable optical modules. Co-packaged optics are still in an early developmental stage and tight collaboration between a wide range of stakeholders, including photonics vendors, co-packaged optics developers, and data center operators, will be needed to accelerate their commercialization (Chopra 2021; Minkenberg, Krishnaswamy, Zilkie, and Nelson 2020). The projected growth of cloud computing, AI engines, and other data-intensive applications, would drastically increase energy consumption, power dissipation, and power density—all of which limit capacity and hence performance of incumbent technology.

Hybrid (analog-digital) beamforming devices: Emerging wireless communications networks (5G and 6G) use higher frequencies that will require large antenna arrays to maintain performance (e.g., latency, bandwidth)

in part because higher frequency signals are prone to environmental attenuation. Cost effective and energy efficient beamforming therefore will be a central consideration in 5G+ wireless designs. Beamforming (either analog or digital) is used for directional signal transmission and reception. Analog beamforming uses a single RF chain to control all antennas, while digital beamforming has a dedicated RF chain for each antenna. Analog beamforming, which has fewer circuits than digital, has lower energy consumption, but at the cost of data rate. Hybrid beamforming optimizes aspects of both analog and digital beamforming to limit hardware complexity and thus energy consumption (and cost) while maintaining data rates similar to purely digital beamforming (National Instruments 2019; Han, Yan, and Yuan 2021). To date, however, hybrid beamforming's increased computational complexity degrades other performance and research is needed to overcome this. (Yu and Zhang 2021).

Challenges and R&D Pathways

Summarized below are the challenges and R&D opportunity areas that were discussed throughout the session.

Signal attenuation: As wireless networks use higher frequencies (mmWave and THz) for future generations of telecommunications (5G+, 6G), environmental signal attenuation increases. THz waves are especially prone to atmospheric attenuation due to absorption by water vapor above 1 THz (Tamosiunaite, Tasmosiunas, Zilinskas, Valusis 2017). Even though >100 THz waves have larger bandwidth and other advantages, their increased signal attenuation will require vastly more base stations and relay points. If implemented in the same way as previous generations (3G and 4G), much higher cell tower and overall energy use will be required to cover the same ranges.

Material quality: Although emerging semiconducting materials, such as wide bandgap semiconductors and CNTs, exhibit improved performance and energy efficiency compared with conventional devices, they haven't been used at large enough scale for long enough to overcome material quality issues in comparison with incumbent ultra-pure ($10^{-9\text{-}12}$purity) Si. Although WBG (i.e., SiC and GaN) have been used longer than CNTs and are in some commercialized products, WBG substrate defects are still a significant source of yield loss and cost. Similarly, controlling CNT chirality, and thereby electronic properties, has been the primary barrier to its wider development and commercialization. Atomic precision including self-assembly and other novel approaches to material growth, defect characterization, and in-situ metrology overviewed in our Workshop 2 may be needed to overcome material quality issues.

Spectral management: More efficient use of the available spectrum is also important for overall energy efficiency, in addition to developing devices for higher spectral frequencies. While some unused frequencies within purchased bands serve a purpose (i.e., to limit interference), there are still opportunities to use these bands more efficiently – compared with the current practice of selling spectrum to providers that get exclusive access to that band. In 2004, the General Accounting Office (GAO) noted that "[t]he current structure and management of spectrum use in the United States does not encourage the development and use of some spectrum efficient technologies" (GAO 2004). Spectrum efficient technologies may include devices that can sense which parts of the spectrum are unused and adjust their operating frequencies accordingly. Recently, there has been increased R&D in tunable, chip-scale and/or integrated RF devices (Hagerstrom et al. 2018).

Cost: The cost of developing and commercializing new, more energy efficient technologies was noted as a primary barrier. Currently, CMOS technology is very low cost, and the industry is largely optimized for silicon/CMOS production. Workshop participants identified a short-term (~5year) need for technology roadmapping for new CMOS-compatible technologies and a longer-term (>10 year) need to develop a R&D roadmap to commercialize non-silicon-based technologies for communications, such as wide bandgap semiconductors and CNTs. Increasing production capacity of non-silicon-based devices will require significant cost, time, and industry backing. U.S. industry currently spends ~$60B per year on RD&D and the proposed

CHIPs Act funding, $52B over 5 years, nearly an order of magnitude increase from historical Federal semiconductor R&D spending, was the scale deemed necessary to commercialize Beyond CMOS technologies (SIA 2022).

Electronic design automation (EDA) tools: It was noted several times throughout the session that more comprehensive and updated EDA tools were needed to design, develop, and deploy novel analog devices for communications. The available simulation packages do not include more energy efficient materials and atomically precise processes. Moreover, most system simulation packages are relatively inflexible. For example, they do not take into account the underlying physics of novel materials, so validating components and circuits requires specialized tools and expertise that few manufacturers can access. There is also currently no holistic design tool that is able to capture multi-scale phenomena (from circuit to systems). Even more, given the tight developmental turn-around time expected from consumers (e.g., new mobile phone models every couple of years), industry does not have time to model and integrate results from each hierarchical level into the next design. As part of efforts to broaden and accelerate simulation of emerging designs, robust information sharing between academia/researchers and industry was recommended.

Thermal management: The trends towards smaller chip sizes, increased bandwidth, increased operating frequency, and 3D chip structures have made thermal management a priority. The area of silicon on a chip that is not useable due to thermal load consideration – also known as dark silicon – has been increasing for years, reducing operational and energy efficiency. Reliable, scalable, cost-effective solutions are urgently needed for transferring heat—including engineered materials that transfer heat faster – from inner layers of 3D-stacked chips to their exteriors. Furthermore, for emerging advanced packaging techniques, the increased heat can cause cracking and interface separation of the many different materials bonded and packaged together on smaller chips.

Minimizing Undesirable Tradeoffs: Innovation is needed to avoid diminished performance and reliability that otherwise might be caused by the introduction of novel materials, device architectures, and integration schemes that increase energy efficiency. For example, CNT-based electronics have been shown to greatly improve energy efficiency while maintaining performance, compared with conventional silicon technologies, but reliability is key concern due to the lack of a technique to deposit and/or grow aligned, pristine semiconductor CNTs at manufacturing relevant scales. Large scale growth of CNTs from catalysts have been demonstrated but the inability to control chirality and diameter result in devices that are leaky, hysteretic, and unstable. Some participants stated that markets will not accept the tradeoff of reducing transistor switching speed for improving energy efficiency; historically that has not been true, as evidenced by the transition from the faster bipolar junction transistor (BJT) to more energy efficient metal oxide semiconductor field effect transistor (MOSFET) in the 1990s.

Community: Participants noted that novel technology development for communications could benefit from a strong community; GaAs was used as an example. With federal and private support, there was a concerted effort to develop GaAs-based devices, creating a strong sense of community that accelerated development. A community around other materials and devices/systems, such as GaN, CNTs, and RF front end—perhaps catalyzed by a ManufacturingUSA consortia – such as AMO's PowerAmerica Institute that accelerated SiC commercialization for power electronics – could have a similar effect.

Table 2. Participant Input on Analog Hardware for Communication

Most Promising analog devices for energy-efficient, next generation communications	
• III-V and III-N semiconductors (e.g., GaN, GaAs, InP, and AlN)	• RF-Photonics integration • Redox transistors

• Carbon Nanotubes (CNT) • Hybrid analog-digital beamforming • Neuromorphic approaches for RF signal processing	• Passives made from metamaterials • Energy harvesting approaches (thermoelectrics, piezoelectrics, and triboelectrics).
Manufacturing and Integration Challenges	
• Achieving acceptable material quality, including defect density and surface cleanliness, for novel semiconductor materials, such as CNTs, GaN, and other wide bandgap materials. • Ensuring process (e.g., thermal budget) and material (e.g., contamination) compatibility with existing production lines when introducing new processes and materials. Thermal budget becomes increasingly important as chips get smaller. • Lack of design tools capable of modeling and simulating non-traditional materials and processes. • Engineering interfaces between materials to maximize reliability (e.g., accounting for differences in coefficient of thermal expansion) and performance. • Developing test protocols for mmWave and THz communications systems.	
R&D Pathways	
• System level co-design, simulation, and characterization to maximize performance and energy efficiency with more comprehensive EDA tools that incorporate multi-scale phenomena (device to system)—e.g., evaluating how novel materials affect performance and reliability of existing components. • Leveraging DOE's high performance computing capabilities to explore many designs in parallel. • Building a community to collaboratively develop solutions. This can include creating of a consortium, similar to SEMATECH, to develop standardized modules. • Identifying new application drivers, outside of mobile communication (e.g., industrial IoT), to accelerate R&D for higher frequency communications.	

Analog Hardware for Sensors

As we move towards an ever "smarter" society, the quantity and types of data from electronic sensors has exploded. Manufacturers and researchers, in particular, are increasingly relying on a larger pool of sensors to monitor, communicate, and control their processes. From AMO's first workshop, key applications of sensor technology in manufacturing and research included in-situ monitoring and control of manufacturing processes, real-time material property measurements, and detectors measuring nanoscale reactions at particle accelerators.

As the number of sensors and data exponentially increases, the conventional approach of transmitting all the data off-node for processing and sending back commands for action from a central controller will become unsustainable. The urgent need to quickly and efficiently generate only actionable information from the growing data deluge was a key conclusion from the first workshop on Integrated Sensors Systems for Manufacturing Applications. Advances in analog hardware and advanced software can drastically reduce data transmission load and improve system energy efficiency. Participants deemed sensor data reduction of >100,000x, one of the goals of SRC's Decadal Plan for Semiconductors, to be within reach with a more analog focus.

The Analog Hardware for Sensors session included two technical talks, from Jennifer Hasler (Georgia Tech) and Jim Wieser (Texas Instruments), and a facilitated discussion. Jennifer Hasler discussed the potential for programmable and configurable analog techniques, in particular FPAAs, to enable ultra-low power computing for sensors and other IoT applications. Jim Wieser gave an overview of intelligent sensing and related human sensing as well as processing approaches to guide intelligent sensor development. He also discussed key challenges such as the enormous amount of raw data production and energy and power constraints. A more detailed summary of the technical talks can be found in Appendix B.

Table 3, on page 13, summarizes participant input from the facilitated session identifying the promising technology and R&D challenges and opportunity areas. Appendix Table C-11 to Table C-13 include the full participant input.

Promising Technology/Approaches

Below is a summary of the promising technology for sensors that were prominently discussed at the workshop.

FPAA: FPAAs include some of the same elements as the more commonly used field programmable gate arrays (FPGA), namely, routing and logic (i.e., digital), but also include analog blocks based on floating gate devices, enabling analog signal processing which can improve computational efficiency by 1000x for certain sensing applications. FPAA can be combined with sensors and energy harvesting approaches to create low-power computational systems for IoT and sensor applications that completely eliminate the need for batteries. The lack of EDA tools and infrastructure to model and simulate FPAAs are primary challenges.

THz sensing: THz waves exhibit desirable properties, such as the ability to penetrate optically opaque materials, non-ionizing photon energy, and unique spectral signatures for chemicals, that make them attractive for industrial sensing applications. There have been significant efforts to miniaturize THz systems to chip-scale for use in portable, battery-operated systems. Primary challenges for THz sensors have been their limited spectral range, angle of incidence, and polarization which limits them to a single sensing modality (Wu, Lu, and Sengupta 2019).

Analog signal processing: Processing incoming signals, before digitization by analog-to-digital (ADC) converters, can drastically reduce the computational load by filtering out unnecessary data. By placing an analog signal processor *before* the ADC, only relevant and actionable data is digitized and transmitted for further processing. Less data reduces energy consumption, size, and cost of the sensor node. In one example, a 100x reduction in data digitization and processing was observed (Rumberg 2016). However, noise, linearity, programmability, and flexibility are challenges to overcome to develop advanced analog signal processors that improve energy efficiency of the sensor system.

Challenges and R&D Pathways

Summarized below are the challenges and R&D opportunity areas that were discussed throughout the workshop.

Local processing: For applications such as manufacturing that don't require much autonomous decision making, local processing can readily reduce the amount of data transmitted off the sensor node, thereby greatly reducing the energy consumption and addressing the "data deluge." Transmitting only actionable information to the cloud or local networks for analysis will require specialized hardware, including analog electronics. Analog signal processing (e.g., FPAAs), in conjunction with more traditional digital computation, provide a promising path forward.

Changing the sensing paradigm: Currently, sensor systems use a "digital-first" paradigm where all incoming signals are immediately digitized, before any processing occurs. This method creates a huge amount of digital data, whether useful or not, and dominates system power consumption. By shifting the digital-first paradigm to analog-first, energy efficiency of the system can be improved by reducing the computational and data transmission load. Advances in programmable analog circuits have enabled flexibility to enable more general-purpose analog devices for sensing applications but more development is needed (Rumberg 2016).

Energy and size constraints: Energy and size constraints will be a primary consideration for sensors deployed in IoT systems. It is estimated that by 2025, the IoT data output will be four times higher than 2019 (Jovanovic 2022). With limited energy supply (e.g., batteries) and increasing computational load of IoT devices, sensors must minimize the amount of energy consumption while maintaining acceptable performance. At the same

time, sensors must also have a small form-factor for easy integration with the other components on the device. With the range of applications and environments where IoT devices are used is expected to greatly increase, the specific size, energy, and other requirements for sensors and supporting electronics systems will range widely and be highly application specific. Advanced analog hardware leveraging analog signal processing at the sensor node will reduce data generation and computation in the cloud, increase processing speed, and ultimately reduce energy consumption. Building out more heterogeneous integration options and modular design structures can greatly accelerate the ability to create systems optimized for key manufacturing applications and industrial markets.

Sensor fusion: Sensor fusion combines multiple types of data from different sensors that are monitoring the same environment or process. Advantages of deploying sensor fusion hardware and software approaches include increased spatial and temporal coverage, reduced uncertainty, robustness against interference, and increased resolution. Sensor fusion is increasingly being used for mmWave, THz, infra-red, and optical frequencies for environmental sensing for autonomous vehicles and systems (Wu, Lu, and Sengupta 2019). A primary challenge of sensor fusion is the required increase in computation to translate and/or combine the multiple sources of data to generate accurate, useful information. Sensor fusion through software implementations in the cloud produces significant data transmission loads, computational burden, and energy consumption that edge computing approaches can avoid.

Process, voltage, temperature (PVT) sensitivity: Analog electronics – because they respond to the details of the signal rather than just a threshold – are inherently more sensitive to PVT variations that can introduce noise and degrade the signal. Devices (e.g., memristors) and circuit techniques have been used to mitigate the impacts of PVT but require additional chip area, increasing energy consumption and costs. New analog architectures that can efficiently mitigate PVT effects are needed.

Table 3. Participant Input on Analog Hardware for Sensors

Most promising analog approaches for improving sensor energy efficiency	
• Floating-gate devices • FPAA • Analog signal processing	• Analog architecture • >100 GHz sensing modalities • Printed organic electrochemical transistors.
Challenges	
• Establishing programmable standard cell libraries for analog devices to reduce development cost for a wide variety of application specific requirements. • Designing sensors that adhere to strict energy and size requirements dictated by form factor and application. • Mitigating sensitivity to and effects of PVT variations through energy efficient components or novel architectures.	
R&D Pathways	
• Using local, analog signal processing to greatly reduce data processing and communication and avoiding the energy intensive “digital-first” paradigm. • Developing a machine learning framework that can process various types and multiples of signals (i.e., sensor fusion) and use the knowledge to guide process control and optimization. • Establishing an open dialog between designers and manufacturers to ensure the greatest impact and effectiveness of sensor development. • Developing self-calibrating sensors that can operate in-situ under different or changing conditions (e.g., high temperature, high pressure, vibration, etc.).	

Neuromorphic Architecture and Devices

Significant acceleration in analysis of massive amounts of data have been achieved in recent years by computing approaches known as machine learning (ML). A subset of ML approaches, that are based on the function of neurons accepting input signals and propagating them to other neurons (i.e., neuromorphic) has proven to be optimal for applications that are particularly relevant to manufacturing such as object recognition and process optimization. While such neuromorphic computing, which involves many computations done in parallel, have been revolutionary for data analysis, it is typically implemented on traditional CMOS hardware. Neuromorphic computing improves computational performance (in some applications) compared with traditional von-Neumann computing, but it still falls short of the even greater performance and energy efficiency achievable from coupling neuromorphic computing with neuromorphic-optimized hardware.

An interim hardware solution for ML users is to leverage existing more parallel computing hardware such as graphics processing units (GPU) and FPGAs to circumvent traditional serial hardware limits on energy efficiency and speed. Experts recognize, however, that new devices that are specifically designed for neuromorphic computing approaches would provide many more orders of magnitude improvements in energy efficiency and speed over even GPU and FPGA approaches (Zhu, Zhang, Yang, and Huang 2020).

Furthermore, while the term "neuromorphic" refers to the operation of neurons in the human brain, the actual operation of today's neuromorphic computing approaches in silicon is still far from how experts now understand the brain to operate. In particular, today's so-called neuromorphic computing still uses orders of magnitude more energy per operation than the human brain. Our recently improved understanding of how the brain works – at many levels below that of the neuron – suggests that many more methods for storing and processing data (particularly sensory data) in extremely energy efficient ways remains to be developed. When adapted to grow from today's silicon systems the most modern understanding of the brain can provide a roadmap for realizing additional gains in performance and efficiency. One recent estimate of the potential energy efficiency improvements possible with such brain-inspired hardware and software is a factor of a million (1,000,000X) (Shankar 2021).

Brain-inspired Computing

The human brain requires only 12 Watts to perform complex tasks that silicon-based serial architecture computers require up to a million times more energy to perform or cannot perform at all. While neuromorphic inspired computing approaches and circuit architectures have unlocked some of these efficiency gains, the way in which neurons in the brain operate is still vastly different and more energy efficient. These differences include a) the number of connections between individual switching devices – a neuron has tens of thousands of connections whereas a transistor only has two and b) the storage of data – traditional memory is stored in a separate location in silicon from where processing is done, whereas neurons in many areas of the brain perform both computation and storage functions. While not every detail of how the brain operates needs to be mimicked in silicon, there are likely more aspects of neural function that electronics developers can leverage to realize additional gains in energy efficiency.

The Brain Inspired Computing panel session included Lawrence Spracklen (Numenta), Bruno Olshausen (UC – Berkeley), Dhireesha Kudithipudi (UT – San Antonio), and Bobby Kasthuri (Argonne National Lab and University of Chicago).

Lawrence Spracklen discussed Numenta's software-based approach, inspired by the neocortex, to reduce the costs of AI by one million-fold. Bruno Olshausen introduced vector symbolic architecture as a holistic representative framework that combine key ideas from artificial intelligence and cybernetics/neural networks. Dhireesha Kudithipudi discussed the importance of choosing the appropriate model to build algorithms or

topologies to maximize energy efficiency and highlighted three potential innovations inspired by human brain models: neurogenesis, neuromodulation, and metaplasticity. Narayanan Kasthuri discussed the ongoing effort to produce a "connectome," a map of every neuronal connection in a biological brain, and its importance to AI and neuromorphic computing. A more detailed summary of each panel speaker's remarks can be found in Appendix B.

Based on the panel talks and subsequent Q&A session, the following key themes emerged:

Emulating the brain: The brain is a highly developed, energy efficient organ that can detect signals that are below the noise floor for electronic components. It is also capable of drastically (>100-10,000X) reducing the megabits of data absorbed by the senses each second, passing along fewer than 100 bits of data per second to the conscious mind for processing. Finally, it accomplishes all of this on an energy budget much smaller than that of modern computing hardware. While neural networks derive their name from a major aspect of organic brains, there are still significant differences in both structure and capability between brains and the most advanced neural networks. There is still much to learn from how the brain stores, processes, and communicates information. Participants discussed how the study of brains, including the brains of animals like mice and jumping spiders, still remains many steps ahead of their implementation in neuromorphic computing.

Plasticity: The brain's ability to continuously break, re-form, and grow new connections may be the key to how and why it's so efficient. As new information comes into a brain, new connections can be made. By contrast, most modern neural networks are extremely rigid – once trained, any attempt to train them on new tasks or update them with new information risks erasing the training they previously received (known as catastrophic forgetting). Innovative hardware implementations of neurogenesis, neuromodulation, and metaplasticity should bring increased flexibility and plasticity to neuromorphic computing. Since brain structure changes radically during growth from infancy to adulthood, neuromorphic computing might be improved by incorporating a similar transformation in a neural network as it is trained.

Architectures: Alternative computing hardware architectures can leverage existing technology to drastically reduce energy consumption and compute/memory resources. AI with architectures that more closely mimic the human brain, with a sparse neural network and active dendrites, can reduce the amount of energy, cost, and data required for AI training. For example, vector symbolic architecture provides a robust, low energy method of performing symbolic computation in a highly distributed network, combining the strengths of Boolean computing systems and neural networks. Pyragrid and MetaPlasticNet are hardware implementations of advanced neuromorphic computing using standard semiconductor devices, providing an avenue to realize major improvements in hardware energy efficiency in the short-term, as more radical emerging devices are developed over the long-term.

Parallel development of hardware and software: Neuromorphic architecture can either be built into newly designed hardware or simulated in traditional von-Neumann architecture computers. The latter – software implementations of neuromorphic computing in traditional architecture – is a near-term solution that utilizes existing digital and analog hardware. For example, Numenta's roadmap outlines how understanding and applying principles from the neocortex in software alone can result in a **one-million-fold reduction in AI costs.** Hardware implementations of neuromorphic computing, while having a longer time horizon for development and deployment, likely will have even more drastically improved energy efficiency compared with software approaches using conventional technology (i.e., CPU, GPU, FPGA). The development of these two approaches, software and hardware, do not need to be mutually exclusive and can complement one another. To realize the most efficiency and performance gains, neuromorphic software and hardware developers should work collaboratively in a co-design approach.

Neuromorphic Hardware

As mentioned previously, hardware researchers are realizing orders of magnitude efficiency and processing speed improvements in neuromorphic approaches by parallelizing and otherwise optimizing hardware for energy intensive applications such as machine learning. These innovations and adaptations to date are occurring at larger sub-system and block levels. The basic devices that make up the systems are still the same, CMOS transistors combined in von-Neumann-based structures. Instead, building hardware where the individual devices and their connections to each other are optimized for neuromorphic computing can result in even more orders of magnitude improvements in operational efficiency. Developing such neuromorphic hardware devices are ongoing avenues of research and multiple candidate technologies exist at the lab-scale. These candidate technologies, as described below, vary widely in how much they diverge from existing, commercially available technologies and thus each has their own barriers to scaling up and integrating with traditional systems. Because they are optimal only for a subset of all computing, neuromorphic devices will not completely replace CMOS and von-Neumann-based architectures, but can be complementary to them, providing orders of magnitude improvements in computational power and efficiency in certain functions.

The Neuromorphic Hardware session included a panel session and a facilitated discussion. The panel consisted of Vijay Narayanan (IBM), Sean Shaheen (University of Colorado – Boulder), and Bhavin Shastri (Queens University).

Vijay Narayanan discussed analog synaptic non-volatile memory devices, including phase change memory, resistive random-access memory (ReRAM), and electrochemical random-access memory (ECRAM). He also discussed device requirements to implement these technologies, and key considerations for analog AI, such as heterogenous integration needed to meet IBM's goal to increase energy efficiency by 2.5x per year from 2020 to 2025. Sean Shaheen highlighted the advantages of organic semiconductors for neuromorphic computing and their most promising applications. Bhavin Shastri discussed the emergence of photonics as a promising approach for AI and neuromorphic computing. He also discussed the wide array of architectures and applications in which it can be used. A more detailed summary of each panel speaker's remarks can be found in Appendix B.

Table 4, on pages 18-19, summarizes participant input from the facilitated session identifying the promising technology and R&D challenges and opportunity areas. Appendix Table C-14 to Table C-18 includes the full participant input.

Promising Technology

Below is a summary of the promising technologies for neuromorphic hardware that were prominently discussed at the workshop.

Memristors: Memristors are circuit elements with variable, programmable resistance: once a memristor is programmed with a resistance value, it will hold that value until it is programmed with a new one. This property makes memristors useful as elements of analog and neuromorphic computing devices. Memristors can compensate for PVT variations in analog circuits and form the basis of artificial synapses for neuromorphic computers. They have faster switching speeds and require less energy than other solid-state memory technologies, and can theoretically be scaled down to a smaller size than other logic components to enable higher-density data storage. However, there are still challenges inherent in scaling memristors to small sizes while maintaining their functionality. Performance issues at scale include endurance, memory retention, device variations, and analog on/off ratio with linear conductance. (Ang, Zhou, Yew, and Berco 2019).

Electrochemical devices: Electrochemical computing devices such as organic electrochemical transistors, electrochemical random-access memory (ECRAM), and electrochemical memristors mimic the

electrochemical processes the brain uses for computation. In general, these devices rely on ion diffusion through a polymer to retain their states. They feature low switching energy and high linearity, making them desirable for energy-efficient neuromorphic architectures. However, they tend to have slower switching speeds than comparable semiconductor devices. Long-term operational stability can also be a challenge for some electrochemical devices (van de Burgt 2017).

Photonics: Optical communication systems over long distances have become increasingly dominant since the 1970s, and optical information processing systems have been considered since 1985. Unlike transistors, optical data processing systems are highly linear, which makes them unsuitable for digital computing. However, the linearity of optical systems makes them highly desirable for neuromorphic and analog computing. In addition, photonics can compute in parallel with multiple wavelengths of light. Photonic components are still much larger than transistors, but are currently shrinking exponentially, following a "photonic Moore's law." Because they do not suffer from capacitive effects, photonics may ultimately be scalable to much smaller sizes than electronic computing components (Hamerly 2021). Photonics' integration with electronics requires very high – even atomic – precision.

Organic semiconductors: Some carbon-based polymers or other small organic molecules possess semiconducting properties or can be doped into a semi-conductive state. These materials may offer advantages over traditional metallic semiconductors. Organic semiconductors can be processed at very low temperatures, fabricated into complex structures via 3D printing, and tuned by varying the chemistry used in their construction. They also feature extremely low energy use per computation. In the future, organic materials may complement silicon in some applications, such as memristor crossbar arrays made of polymers, organic circuits that exhibit neuron-like behavior, and integrated sensing and computing devices. For most computing applications, organic components will need to be integrated with CMOS digital circuits such as selectors and driver transistors; scalable methods of integrating these components are currently being investigated (Tuchman et al. 2020).

In-memory computing: One area of innovation aimed at realizing near-term impacts is adapting existing CMOS devices to be more optimized for neuromorphic computing approaches. This approach differs from specialized neuromorphic chips available in systems today – which leverage higher-level organization of logic blocks – in that it aims to use the fundamental device technology (e.g., random access memory) directly to store data and perform logic operations. This would allow IC designers to further optimize the hardware for highly parallel, brain-like spiking operations by integrating the functionality of logic and memory devices, eliminating significant data transfer operations that are a requirement and drawback of today's von-Neumann-based hardware. An advantage of using existing memory technologies for in-memory computing is that the manufacturing infrastructure is already mature and therefore successful developments can be rapidly scaled and commercialized. A disadvantage of using conventional charge-based memory devices, however, is that they have a much larger footprint than memristors or other advanced memory devices and are not as energy efficient, limiting the ultimate energy efficiency and scaling potential of compute-in-memory approaches on conventional hardware (Sebastian, Le Gallo, Khaddam-Aljameh, and Eleftheriou 2020). Thus, they have limited application for the goal of countering exponential energy growth or use of Si for memory.

Challenges and R&D Pathways

Summarized below are the challenges and R&D opportunity areas that were discussed throughout the session.

Manufacturing at scale: Although devices for neuromorphic circuits, like ECRAM and organic semiconductors, have been developed in the lab, scaling to manufacturing-relevant throughputs is a challenge. High-volume manufacturing requires significantly larger substrates and process chambers as well as faster processing times. As a result, the physics and chemistry for material growth, etching, and deposition may be

different at the manufacturing scale than at the lab scale. More extensive metrology and a deeper understanding of physics for key materials will be necessary to transition from lab to fab, notwithstanding design and process integration concerns.

Material and process integration: The introduction of novel materials and processes into an existing process flow introduces contamination concerns and may introduce process integration challenges, such as different thermal budgets for existing structures as well as downstream processes. Contamination concerns include electromigration, inadvertent doping, and contamination of process equipment. Such risks are especially high for general purpose fabs (in a model where they serve multiple fabless manufacturers) that use the same tool for multiple product lines. In addition, the integration of new processes or process modules may, depending on the materials and process parameters, require adjustment to accommodate thermal budget and potentially additional steps (e.g., surface prep and cleans) for proper integration.

Device scaling: Today's most promising neuromorphic devices have much larger linewidths than leading edge logic devices. Absent innovation to shrink them further, accommodating larger form-factors may affect device operation. In addition, for devices that rely on non-traditional operation, such as organic semiconductors, performance at smaller scales is unknown. In most cases – except in-memory computing – advanced packaging techniques will be needed to integrate new neuromorphic hardware with traditional CMOS hardware in order address challenges associated with such linewidth differences.

Higher dimensional devices: A neuromorphic circuit that uses the third dimension (i.e., stacked crossbar arrays) can increase compute capacity while lowering energy use compared with a 2D system. Because the third dimension can no longer be used for heat dissipation, however, such 3D configurations have more challenging thermal management requirements (especially for compute chips). One effective 3D strategy is to pack more bits into each signal and signaling less often, moving to a system that mimics neurons. This lowers the need for thermal management because this allows for "sparse signaling," reducing heat generation while maintaining computational capacity. This neuron-inspired approach to increase both the energy efficiency and compute capacity is still nascent and will require significant mathematical development before it can be commercialized; but it is an example of the type of transformational innovation needed to increase energy efficiency exponentially to counter the current trajectory of unsustainable energy use.

Co-design: Co-design of energy efficiency across the device, circuit, architecture, and algorithm (the stack) is essential as evidenced throughout the plenary talk and the Brain-inspired Computing and Neuromorphic Hardware panels. The interconnectedness across the stack makes co-design a necessity to design and manufacture vastly more energy-efficient, high performing neuromorphic systems. Experts from across the stack as well as those in fields not typically associated with semiconductors, such as biology and neuroscience, must collaborate closely to develop effective solutions. In particular, learning from other technologies/approaches in semiconductors as well as those outside the field can provide inspiration for novel solutions. Two examples of co-design learnings from other fields were provided during the workshop. First, principles and codes from quantum computing can be leveraged. Second, cross-cutting lessons learned from memory chip developers on their path to develop and commercialize 3D memory chips, can be useful for co-designing 3D compute chips.

Table 4. Participant input on Neuromorphic Hardware

Promising more energy efficient non-Von-Neumann approaches to computing	
• Redox transistors • Electrochemical devices (ECRAM) • Continuous memristors	• Photonics • Room temperature quantum materials

• Organic semiconductors	• Silicon-oxide-nitride-oxide-silicon (SONOS) memory

Challenges

Design

- Maintaining a small size to meet form factor requirements.
- Developing a multi-scale and multi-modal co-design capability for holistic modeling and simulation.
- Simultaneously co-designing and understanding interactions between device, circuit, architecture, and algorithm (stack) interactions in order to maximize inference and training accuracy.
- Developing robust design verification frameworks and benchmarking platforms for consistent evaluation of device performance, including energy efficiency.

Manufacturing and integration

- Non-availability of embedded non-volatile memory technology in advanced CMOS nodes (where high-speed IO IP is available).
- Small manufacturing/process changes can affect the accuracy and energy efficiency of an analog system, such as inference accuracy.
- Managing stresses induced by the presence of large fields as the devices are scaled down
- Understanding sources of device variability and whether devices can be fault tolerant to them. Even small changes in processes can affect the accuracy of an analog system.
- Integrating current arithmetic and numerical representations between neuromorphic and existing products/devices/components (e.g., IEEE 754 floating-point vs. neuromorphic "arithmetic").

R&D pathways

- Using a co-design approach for holistic development of neuromorphic systems that considers considerations from devices to algorithms.
- Leveraging insights and solutions from a biology and quantum computing to inspire more energy efficient solutions.
- Pushing devices to higher dimensions for increased energy efficiency will increase computational capacity and improve energy efficiency.
- Standardizing data representation, devices, and processes can help accelerate development and deployment of neuromorphic systems.

Cross Cutting Issues

Summarized below are two cross-cutting issues that were brought up in every session; addressing these issues will have a broad impact on the semiconductor R&D and fabrication community.

Workforce development: There was consensus among participants that workforce development was a central issue in the semiconductor industry. Participants highlighted that the comparatively larger starting salaries (and hype) in sectors like software (e.g., artificial intelligence, machine learning) and finance has made it increasingly difficult to attract incoming college students into the hardware-aspect of electronics, especially analog. Compounding this, electrical engineering curricula typically don't have a separate analog electronics track – students are exposed only briefly to analog electronics in multi-course modules that only provide basic concepts of analog fundamentals. A fundamental change in curricula, recruitment, and messaging may be needed to attract the brightest minds back into semiconductor hardware manufacturing fields. Some participants stated that outreach should begin with K-12, similar to robotics, to garner interest early.

The importance of the workforce having multidisciplinary training was also emphasized. For example, neuromorphic computing research benefits from those that are trained in both neuroscience and semiconductor device research – but this requires additional time and funding. In order for co-design to fully emerge as a promising R&D pathway, a multidisciplinary trained workforce will be needed for the most challenging problems. While NSF has recently initiated more efforts relating to workforce[4], participants agreed that such efforts also need to be integrated with efforts such as R&D consortia.

Strengthening the lab-to-fab pipeline: The need for a prototyping facility or a general fab capability for academic researchers and small businesses was discussed in all topics – and discussed in all workshops in the series. Securing fab runs at commercial fabs is difficult due to a number of reasons including, cost, contamination concerns, and manufacturing schedule. A prototyping service or facility may provide a low-cost, high flexibility option for testing designs, materials, and processes. However, challenges exist, including the need for sufficiently advanced node toolset to ensure products coming out of the prototyping facility are industry relevant, and the substantial cost of building, operating, staffing, and maintaining a prototyping facility. A public-private R&D consortium, similar to AMO's PowerAmerica could help solve these issues by bringing together academic researchers and companies to create a collaborative community with a shared goal.

[4] NSF and Intel announced a new $100M joint effort to address semiconductor challenges and workforce shortages (Intel, 2022).

References

Ang, D. S., Y. Zhou, K. S. Yew, and D. Berco. 2019. "On the area scalability of valence-change memristors for neuromorphic computing." *Applied Physics Letters*. 115(17). https://aip.scitation.org/doi/10.1063/1.5116270.

Chopra, Rakesh. 2021. "Co-Packaged Optics and an Open Ecosystem." Cisco Blogs. https://blogs.cisco.com/sp/co-packaged-optics-and-an-open-ecosystem.

Department of Energy (DOE) Advanced Manufacturing Office (AMO). 2022. "Semiconductor: Supply Chain Deep Dive Assessment." https://www.energy.gov/sites/default/files/2022-02/Semiconductor%20Supply%20Chain%20Report%20-%20Final.pdf.

Hagerstrom, A. M., Lu, X., Dawley, N. M., Nair, H. P., Mateu, J., Horansky, R. D., Little, C. A. E., Booth, J. C., Long, C. J., Schlom, D. G., & Orloff, N. D. 2018. "Sub-Nanosecond Tuning of Microwave Resonators Fabricated on Ruddlesden–Popper Dielectric Thin Films." *Advanced Materials Technologies*, 3(8), 1800090. https://doi.org/10.1002/admt.201800090

Hammerly, Ryan. 2021. "The Future of Deep Learning is Photonic." IEEE Spectrum. https://spectrum.ieee.org/the-future-of-deep-learning-is-photonic.

Han, Chong, Longfei Yan, and Jinhong Yuan. 2021. "Hybrid Beamforming for Terahertz Wireless Communications: Challenges, Architectures, and Open Problems." Preprint. https://arxiv.org/pdf/2101.08469.pdf.

Hennessy, John L., and David A. Patterson. 2012. *Computer Architecture: A Quantitative Approach*. Waltham: Elsevier, Inc.

HEP Software Foundation. 2019. "A Roadmap for HEP Software and Computing R&D for the 2020s." *Computing and Software for Big Science*. 3. https://link.springer.com/article/10.1007/s41781-018-0018-8.

Intel. 2022. "Intel Invests $100M in Ohio and National Semiconductor Education and Research." Intel, March 17,2022. https://www.intel.com/content/www/us/en/newsroom/news/intel-invests-100m-ohio-national-education.html.

The International Roadmap for Devices and Systems (IRDS). 2021. *More Moore*. https://irds.ieee.org/images/files/pdf/2021/2021IRDS_MM.pdf.

Jovanovic, Bojan. "Internet of Things statistics for 2022 – Taking Things Apart." DataProt, March 8, 2022. https://dataprot.net/statistics/iot-statistics/.

Margalit, Near, Chao Xiang, and Steven M. Bowers et al. 2021. "Perspective on the future of silicon photonics and electronics." *Applied Physics Letters*, 118, 22051. https://aip.scitation.org/doi/pdf/10.1063/5.0050117.

Marr, Bo, Brian Degnan, Paul Hasler, and David Anderson. 2013. "Scaling Energy Per Operation via an Asynchronous Pipeline." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 21(1). https://ieeexplore.ieee.org/document/6133317.

Mehonic, Adnan and Anthony J. Kenyon. (2021). "Brain-inspired computing: We need a master plan." arXiv. https://arxiv.org/abs/2104.14517.

Minkenberg, Cyriel, Rajagopal Krishnaswamy, Aaron Zilkie, and David Nelson. 2020. "Co-packaged datacenter optics: Opportunities and challenges." *IET Optoelectronics*. https://doi.org/10.1049/ote2.12020.

National Instruments. 2019. "The Case for Hybrid Beamforming in 5G mmWave Prototypes." IEEE Spectrum. https://spectrum.ieee.org/the-case-for-hybrid-beamforming-in-5g-mmwave-prototypes.

Platzer, M., Sargent, J., Sutter, Karen. 2020. *Semiconductors: U.S. Industry, Global Competition, and Federal Policy*. Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R46581.

Rumberg, Brandon. 2016. "Analog Processing Improves Battery Life and Sensor Intelligence In the IoT." Fierce Electronics. https://www.fierceelectronics.com/components/analog-processing-improves-battery-life-and-sensor-intelligence-iot.

Sebastian, Abu, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. 2020. "Memory devices and applications for in-memory computing." *Nature Nanotechnology*. 15, 529-544. https://www.nature.com/articles/s41565-020-0655-z.

Semiconductor Industry Association (SIA). 2022. *Funding for the CHIPS for America Act & Enacting a FABS Act Investment Tax Credit*. Washington D.C.: Semiconductor Industry Association. https://www.semiconductors.org/wp-content/uploads/2022/03/CHIPS-FABS-summary-3.23.2022_1.pdf

Semiconductor Industry Association (SIA). 2021. *2021 State of the U.S. Semiconductor Industry.* Washington D.C.: Semiconductor Industry Association. https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf.

Shankar, Sadasivan Sadas. 2021. "Lessons from Nature for Computing: Looking beyond Moore's Law with Special Purpose Computing and Co-design." *2021 IEEE High Performance Extreme Computing Conference*. https://ieeexplore.ieee.org/document/9622865.

SIA (Semiconductor Industry Association). 2019. Response to DOE/BES RFI. https://www.semiconductors.org/wp-content/uploads/2019/09/Semiconductor-Industry-Association-Response-to-DOE-RFI-Basic-Research-Microelectronics.pdf.

Tamosiunaite, Milda, Stasys Tamosiunas, Mindaugas Zilinskas, and Gintaras Valusis. 2017. "Atmospheric Attenuation of Terahertz Wireless Networks" In Broadband Communications Networks. https://doi.org/10.5772/intechopen.69590.

The Semiconductor Research Corporation (SRC) and the Semiconductor Industry Association (SIA). 2021 *The Decadal Plan for Semiconductors.* https://www.src.org/about/decadal-plan/.

Tuchman, Yaakov, Tanyaradzwa N. Mangoma, Paschalis Gkoupidenis, Yoeri van de Burgt, Rohit Abraham John, Nripan Mathews, Sean E. Shaheen, Ronan Daly, George G. Malliaras, and Alberto Salleo. 2020. "Organic neuromorphic devices: Past, present, and future challenges." *MRS Bulletin*. 45. https://www.cambridge.org/core/journals/mrs-bulletin/article/organic-neuromorphic-devices-past-present-and-future-challenges/A641A09A984E4D18AED39AD90B70CA3F.

United States General Accounting Office. 2004. *Spectrum Management: Better Knowledge Needed to Take Advantage of Technologies That May Improve Spectrum Efficiency.* https://www.gao.gov/assets/gao-04-666.pdf.

van de Burgt, Yoeri, Ewout Lubberman, Elliot J. Fuller, Scott T. Keene, Gregorio C. Faria, Sapan Agarwal, Matthew J. Marinella, A. Alec Talin, Alberto, Salleo. 2017. "Á non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing." *Nature Materials*. 16(4). https://doi.org/10.1038/nmat4856.

World Semiconductor Trade Statistics (WSTS). 2021. *WSTS Semiconductor Market Forecast Fall 2021*. San Jose, CA: WSTS. https://www.wsts.org/esraCMS/extension/media/f/WST/5263/WSTS_nr-2021_11.pdf.

Wu, Xue, Huaixi Lu, and Kaushik Sengupta. 2019. "Programmable terahertz chip-scale sensing interface with direct digital reconfiguration at sub-wavelength scales." *Nature Communications*. 10, 2722. https://www.nature.com/articles/s41467-019-09868-6.

Yu, Xianghao and Jun Zhang. 2021. "Hybrid Beamforming for 5G Millimeter-Wave Systems." IEEE Signal Processing Society. https://signalprocessingsociety.org/publications-resources/blog/hybrid-beamforming-5g-millimeter-wave-systems.

Zhu, Jiadi, Teng Zhang, Yuchao Yang, and Ru Huang. 2020. "A comprehensive review of emerging artificial neuromorphic devices." *Applied Physics Reviews*. 7, 011312. https://aip.scitation.org/doi/abs/10.1063/1.5118217

Appendix A: Agenda

DAY 1: August 11, 12:30 PM – 4:25 PM EDT	
Time	**Activity**
12:30 – 1:50	**Opening Plenary**
12:30 – 12:45	Opening Remarks from Organizer (Paul Syers, DOE AMO, Workshop Chair)
12:45 – 12:55	Welcome from DOE AMO (Diana Bauer, Acting Deputy Director, DOE AMO)
12:55 – 1:10	Semiconductor R&D for Energy Efficiency Workshop Series Overview and Q&A (Tina Kaarsberg, DOE AMO)
1:10 – 1:25	NSTC Interagency Working Group Update and Q&A (Lisa Friedersdorf, Co-chair of the Subcommittee for Microelectronics Leadership, OSTP)
1:25 – 1:55	Keynote Talk (Kwabena Boahen, Professor, Stanford University)
1:55 – 2:10	**BREAK**
2:10 – 3:10	**Panel on Analog and Neuromorphic Applications and Impacts (Pete Tseronis, Dots and Bridges, Moderator)**
2:10 – 2:40	- Gina Adam, George Washington University - Gabriella Carini, BNL - Steven Spurgeon, PNNL - Brian Calvert, AI Startup
2:40 – 3:10	Panel Q&A on Analog and Neuromorphic Applications and Impacts
3:10 – 3:25	**BREAK**
3:25 – 4:25	Facilitated Discussion on Analog and Neuromorphic Applications, Global Impacts, and Benefits (Emmanuel Taylor, Energetics, Facilitator)
4:25 – 4:30	**Day 1 Concluding Remarks (Paul Syers, DOE AMO, Workshop Chair)**

DAY 2: August 12, 12:30 PM – 5:30 PM EDT	
Time	**Activity**
12:30 – 12:55	**Opening Plenary**
12:30 – 12:40	Welcome Back (Paul Syers, DOE AMO)
12:40 – 12:55	Opening Remarks (Dave Henshall, SRC)
12:55 – 1:55	**Panel on Advanced Analog Hardware for Communications (Tina Kaarsberg, DOE AMO, Moderator)**
12:55 – 1:25	- Thomas Cho, Samsung - Jim Booth, NIST - Steffen McKernan, Carbon Technology Inc.
1:25 – 1:55	Panel Q&A on Advanced Analog Hardware for Communications
1:55 – 2:05	**BREAK**
2:05 – 3:05	Facilitated Discussion on Advanced Analog Hardware for Communications (Emmanuel Taylor, Energetics, Facilitator)
3:05 – 3:15	**BREAK**
3:15 – 4:15	**Advanced Analog Hardware for Sensors (Paul Syers, DOE AMO, Moderator)**
3:15 – 3:45	Technical Talk and Q&A (Jim Wieser, Texas Instruments)
3:45 – 4:15	Technical Talk and Q&A (Jennifer Hasler, Georgia Tech)
4:15 – 4:25	**BREAK**
4:25 – 5:25	Facilitated Discussion on Advanced Analog Hardware for Sensors (Emmanuel Taylor, Energetics, Facilitator)
5:25 – 5:30	**Day 2 Concluding Remarks (Paul Syers, DOE AMO, Workshop Chair)**

DAY 3: August 13 12:30 – 5:30 EDT	
Time	**Activity**
12:30 – 12:55	**Opening Plenary**
12:30 – 12:40	Welcome Back (Paul Syers, DOE AMO)
12:40 – 12:55	Opening Remarks (Robinson Pino, DOE SC)
12:55 – 1:55	**Panel on Brain-Inspired Computational Approaches (Sadas Shankar, SLAC/Stanford University, Moderator)**
12:55 – 1:25	- Lawrence Spracklen, Numenta - Bruno Olshausen, UC – Berkeley - Dhireesha Kudithipudi, University of Texas – San Antonio - Bobby Kasthuri, ANL/University of Chicago
1:25 – 1:55	Panel Q&A on Brain-Inspired Computational Approaches
1:55 – 2:10	**BREAK**
2:10 – 3:10	**Panel on Neuromorphic Hardware (Robinson Pino, DOE SC, Moderator)**
2:10 – 2:40	- Vijay Narayanan, IBM - Sean Shaheen, University of Colorado – Boulder - Bhavin Shastri, Queens University
2:40 – 3:10	Panel Q&A on Neuromorphic Hardware
3:10 – 3:20	**BREAK**
3:20 – 4:35	Facilitated Discussion on Neuromorphic Hardware (Emmanuel Taylor, Energetics, Facilitator)
4:35 – 4:55	**BREAK**
4:55 – 5:25	**Closing Discussion**
5:25 – 5:30	**Day 3 Concluding Remarks (Paul Syers, DOE AMO, Workshop Chair)**

Appendix B: Plenary and Panel Talk Summaries

Day 1

Opening Remarks from the Organizer – Paul Syers, Technology Manager, DOE AMO

As digital technology advances, the volume of data being transmitted wirelessly increases. Hence, even small increases in wireless communication efficiency can lead to significant energy savings. Improving the energy efficiency of signal transmission and signal processing will require new hardware, such as analog and neuromorphic devices. The objective of this workshop is to identify opportunities, needs, and barriers to manufacturing and integrating analog and neuromorphic computing devices and designs with existing semiconductor systems to achieve significant improvements in operational energy efficiency. During this workshop, participants will gain insight into how analog and neuromorphic computing systems are used and how hardware is and can be leveraged, identify manufacturing-related hurdles to incorporating analog and neuromorphic computing hardware into advanced electronics systems, and prioritize R&D needs to advance semiconductor R&D for analog and neuromorphic designs and devices within the AMO mission space.

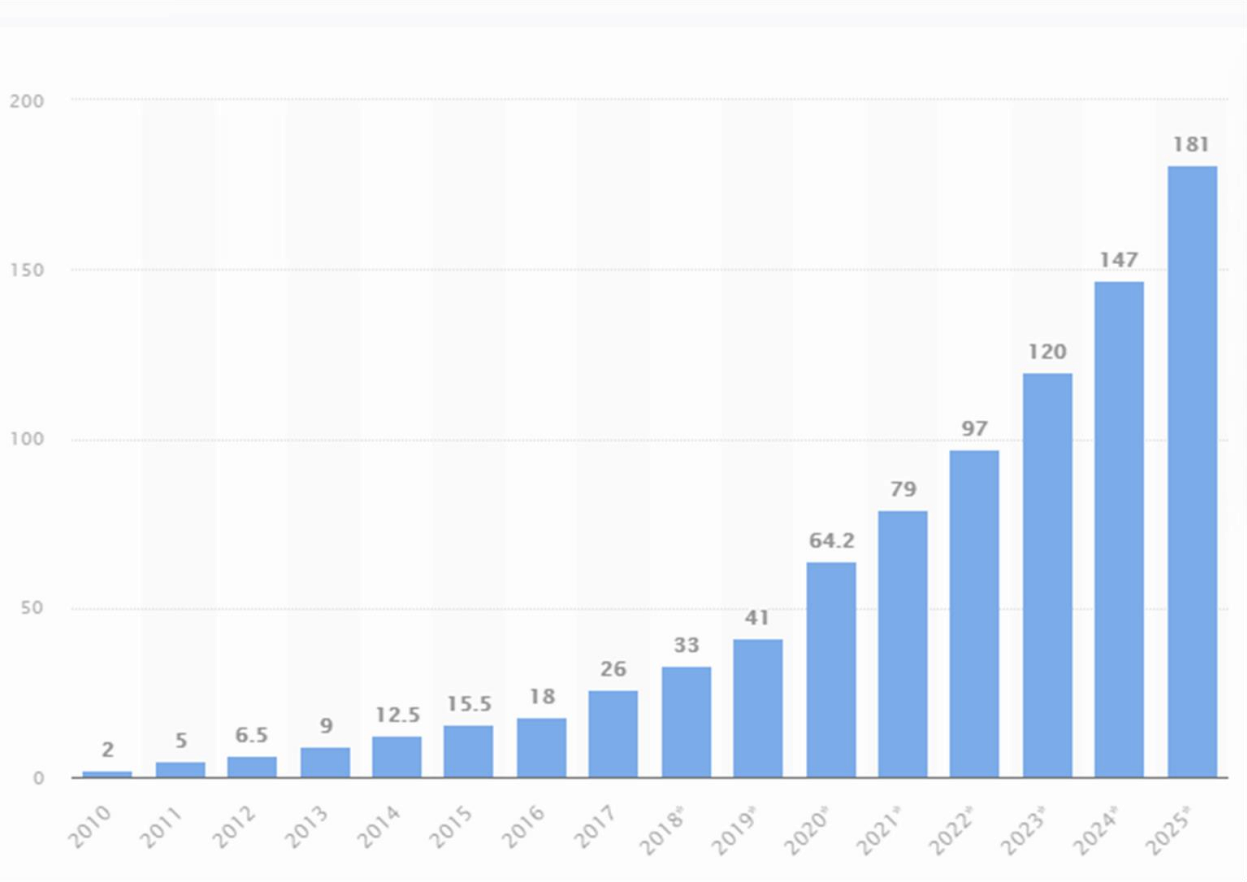

Figure 4: Data volume in zetabytes, with projections to 2025 (Statista).

Welcome from DOE AMO – Diana Bauer, Acting Deputy Director, DOE AMO

The Advanced Manufacturing Office organized this workshop in cooperation with the DOE Office of Science and the Semiconductor Research Corporation. The DOE Office of Science has been at the leading edge of microelectronics, both as a consumer and as a center of research and development that has enabled many technological breakthroughs. The DOE Office of Energy Efficiency and Renewable Energy (EERE) engages

in semiconductor RDD&D as part of its energy efficiency mission. EERE programs involved in semiconductor RDD&D include Better Plants, Solid State Lighting, and Photovoltaics. AMO's Manufacturing USA Institute, PowerAmerica, sponsors wide bandgap semiconductor RDD&D and workforce development to produce semiconductors suitable for manufacturing applications. AMO's atomic precision manufacturing for microelectronics portfolio could leverage both DOE Office of Science and National Nuclear Security Administration resources to bring innovations from early-stage research to applied research and deployment. The improved energy efficiency of the advanced microelectronics and power electronics developed by DOE and its collaborators will be necessary to enable exascale and larger computers and a smart electricity grid.

The Biden Administration's climate agenda calls for a 50% reduction in carbon emissions by 2030 and a carbon emissions-free power sector by 2035, while encouraging good-paying jobs, environmental justice, and robust collaboration between the federal government, states, and the private sector. AMO supports these goals by increasing energy and material efficiency in manufacturing, driving energy productivity, economic growth, and decarbonization. AMO has funded semiconductor R&D projects, including a range of analog hardware related projects, through the Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs. The Creating Helpful Incentives to Produce Semiconductors (CHIPS) for America Act of 2020 included in the FY 2021 National Defense Authorization Act, Title 99, authorizes DOE to coordinate microelectronics research, development, manufacturing, and supply chain security activities, expanding the opportunities for stakeholders in the microelectronics sector to engage with AMO.

Semiconductor R&D for Energy Efficiency Workshop Series Overview – Tina Kaarsberg, Technology Manager, DOE AMO

The U.S. semiconductor industry is responsible for more than half of global semiconductor sales and spends more on semiconductor R&D as a fraction of its sales (16.4%) than any other country. Semiconductors are the U.S. industry's fourth largest export after oil, vehicles, and aerospace technology. The U.S. semiconductor industry is responsible for 1.25 million direct and indirect high-paying jobs. Semiconductors are thus important to achieving the Biden Administration's goal of net zero greenhouse gas emissions by 2050. Semiconductors help accelerate R&D on decarbonization technologies, indirectly, through machine learning approaches, and directly by enabling energy efficient equipment and energy management systems, and lowering the net cost of electrification. Semiconductors can also help adapt to climate change by enabling advanced energy efficient cooling technologies and computers, sensors, and communications for extreme weather prediction.

However, the rapidly increasing energy use of semiconductors threatens to make the climate problem worse. Current trends in semiconductor use and energy consumption would lead semiconductors to consume up to 20% of the planet's energy by 2040. There is an urgent need for innovation to reduce semiconductor energy consumption while improving their performance. Consistent with the Energy Act of 2020 and the CHIPS Act of 2021, AMO has launched a series of four workshops to discuss opportunities for innovation in semiconductor energy efficiency.

Keynote: The Future of AI: A 3D Silicon Brain – Kwabena Boahen, Professor, Stanford University

GPT-3, which represents the state of the art in machine learning and AI, is an AI model that uses an array of programs inspired by human neuron function to respond to text prompts with clear and sensible output text comparable to text written by humans. But training GPT-3 requires 4.6 million dollars, 355 GPU-years, and the equivalent of 50 car-years' carbon emission. As AI training shifted from CPUs to GPUs, the doubling time for AI training has increased: the number of operations needed to train AI models on CPUs doubled every 24 months, but the number of operations on GPUs is doubling every 3.4 to 2 months, leading to an unsustainable exponential increase in energy consumption. AI that can run on personal devices rather than the cloud has the potential for improvements in response time, customization, energy savings, and security, but running a

program like GPT-3 on a single device will require 15 times more processing speed and 180 times more battery capacity than is currently available.

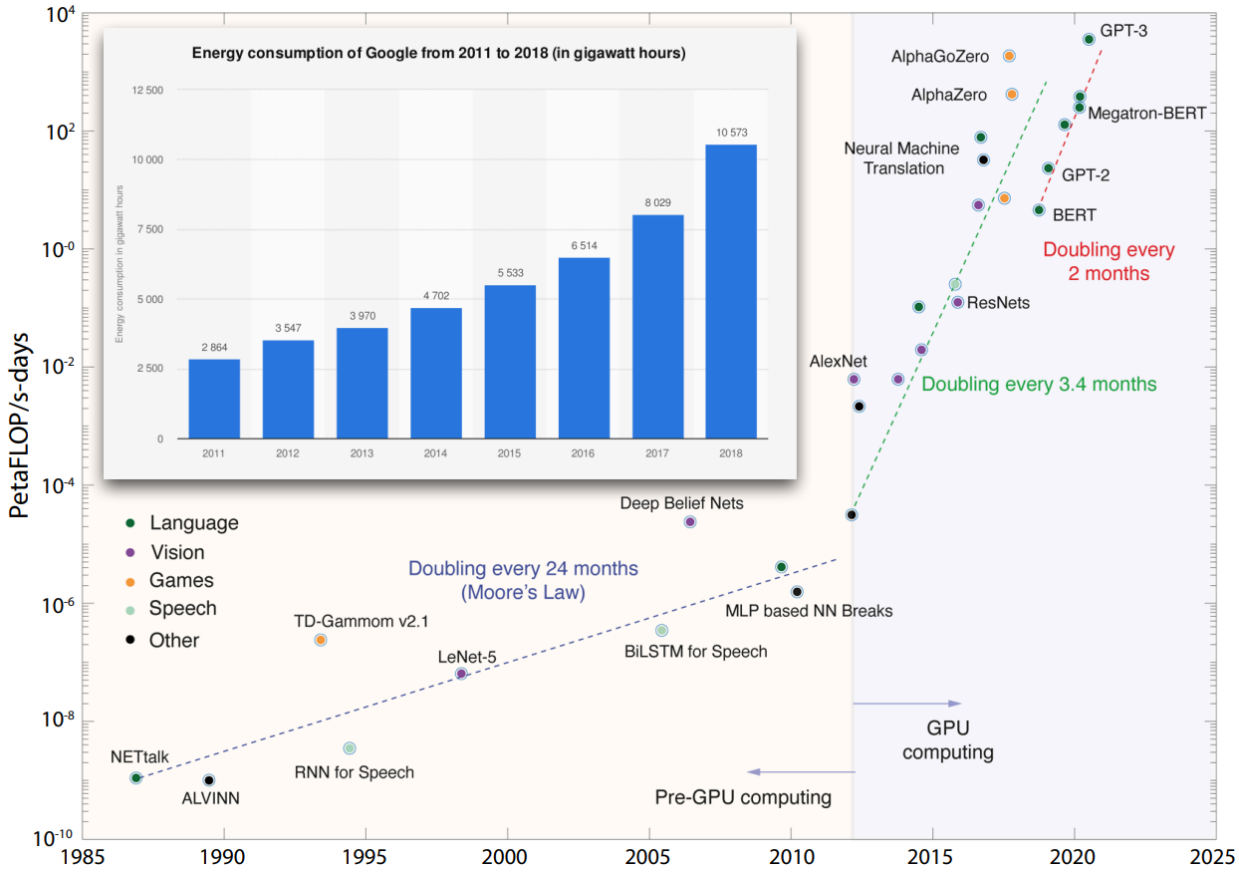

Figure 5: Training data required (PetaFLOPS/s-days) for AI models. Subplot: Energy consumption of Google from 2011 to 2018 (Boahen 2021).

Neuromorphic computing could improve the performance and energy efficiency of computers by basing computer architecture on the organizing principles of organic brains, which are several orders of magnitude more energy efficient than conventional computers. Organic brains are extremely complex, functioning at six spatial and temporal scales; finding out which organizing principles are the most useful to adopt is a major challenge. The scaling of computing capacity with problem size is determined by the codes used to represent information and the primitives that operate on those codes; this is why the energy use of a quantum computer scales polynomially rather than exponentially with problem size and room temperature.

Dr. Boahen's research group investigated how neuron energy use scales with problem size by constructing a virtual model from first principles. The model shows that in a 2D array of neurons, energy use scales as the square of the number of neurons. By transitioning to a 3D array of neurons, the distance that signals travel is reduced, and energy consumption is reduced scaled to the power of 1.5. However, a 3D array of neurons has less surface area per neuron than a 2D array and thus similar to circuits, is less capable of dissipating heat, so the number of signals it can process will be limited by its thermal capacity. To solve this problem, designers can use arrays of multiple neurons that use decimal (1-10) instead of binary (0/1) coding, allowing each signal to transmit more information. In addition, the data processed by a neural network exists on a manifold – so it doesn't have as many degrees of freedom as the input, meaning a data point's position can be encoded with fewer signals. By switching from binary to decimal coding and exploiting the manifold nature of neural

networks, his group reduced energy use until it was simply proportional to the number of neurons. This produces processing power and energy use savings of 2 to 3 orders of magnitude – enough savings to run GPT-3 on a phone. Industry has recently built the first 128-layer 3D decimally encoded memory chips for 1-terabyte phones. Dr. Boahen concluded that computing would be much more energy efficient if it follows phone chip manufacturers' lead and moved to 3D chips.

Panel on Analog and Neuromorphic Applications and Impacts

Gina Adam, George Washington University: While some computing applications such as modeling, simulation, and graphics are limited primarily by processing power, other applications such as sensor networks and IoT are limited by memory and bandwidth, with most of their energy spent on data access and transmission. Because processor speed has historically increased faster than memory access speed, data-driven applications have encountered a bottleneck in capacity. Although analog computing was emphasized less in the past, because it is less precise than digital computing and can be difficult to design and operate, it is beginning to enjoy a resurgence because it is fast, energy efficient, and smaller compared with digital computing making it an ideal solution for small, bottleneck-limited applications such as sensor networks and IoT.

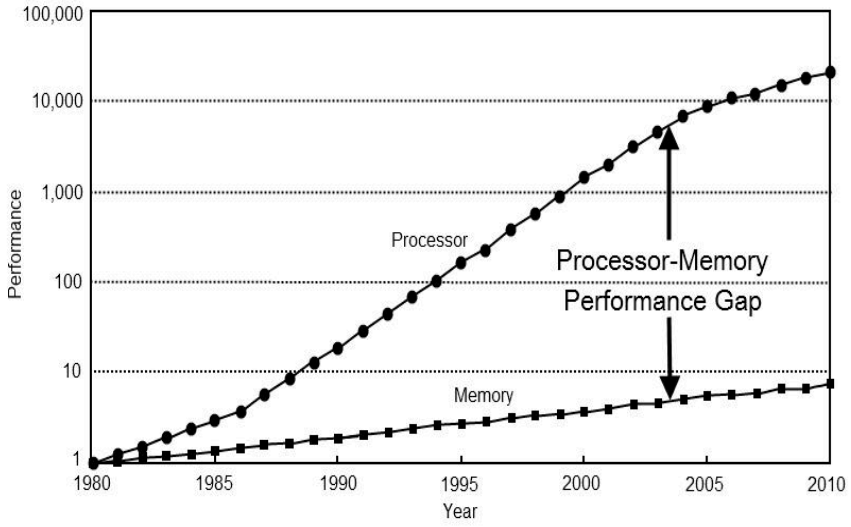

Figure 6: The growing gap between processor and memory performance. Data-driven applications are constrained by memory access speed (adapted from Hennessy and Patterson 2012).

A major challenge to analog computing is errors caused by process-voltage-temperature (PVT) variations. While digital circuits are insensitive to PVT variations because of the discrete nature of digital signals, analog circuits can be highly sensitive to PVT variations. Clever arrangements of analog transistors can reduce noise due to PVT variations, but can't eliminate it entirely. In addition, analog circuits can generate more heat, compared with digital circuits, creating more PV variation. In the short term, solutions that could provide robustness to PVT variations include 1) tunable elements such as memristors that can compensate for these variations, 2) optimizing blocks within analog computing systems independently of one another to reduce error propagation, and 3) inserting optimized analog blocks into digital systems. In the long term, neuromorphic computing is a potential solution, since the brain is apparently already able to compensate for variability in its signaling processes. Designers are looking to adapt techniques from neuroscience to enable analog circuits to better compensate for variability.

Gabriella Carini, Brookhaven National Laboratory: Modern scientific computing must process an overwhelming amount of data from sources such as particle accelerators, observatories, and distributed sensor

networks. For instance, CERN's Large Hadron Collider produces 0.5 petabytes (10^{21} bytes) per second (PBps), and the upcoming Packed Ultrawideband Mapping Array is anticipated to produce 40 GBps on each of its 32,000 channels. Distributed sensing is even more complex, since it involves heterogeneous inputs. The traditional approach to large amounts of data is to stream, store, and analyze it; however, as such researchers near fundamental limits in computing power and energy use, a new approach is needed to analyze more information with less power. The human brain can analyze large amounts of information despite using relatively little energy. The brain also greatly reduces incoming data; neuromorphic computing may similarly allow scientists to generate more information from less data. However, this approach faces both technological and sociological challenges: neuromorphic computing is not a mature technology yet, and even when it is, scientists may not accept an approach that uses less data. Hardware must be co-designed with software to optimize such data processing.

Steven Spurgeon, Pacific Northwest National Laboratory: Answers to major technological challenges such as transformative manufacturing, quantum information science, and energy storage will depend on harnessing large volumes of data produced by advanced experimental instrumentation. Each generation of experimental instrumentation has produced data at exponentially higher rates; analyzing this data using existing computer techniques will soon become practically impossible. To continue to investigate pressing scientific problems, new ways to quickly interpret and act on high bandwidth, heterogeneous data from multiple sensors, modalities, and scales must be developed. Interpreting this data will require inferences from prior knowledge and knowledge of heterogeneous systems. As continued scaling of devices becomes unsustainable, domain-grounded reduction and inference will be essential for obtaining useful information from large-scale experiments.

Brian Calvert, AI Startup: Large amounts of computing power and data are required for both academic applications like high energy physics and industry applications such as self-driving vehicles, and Dr. Calvert has experience with both. High energy physics experiments like those used at the Large Hadron Collider (LHC) often have on the order of 100 million data channels operating at 40 MHz, producing around 100 TB/s of raw data. Advanced, multi-stage data reduction systems make microsecond decisions on which data to keep or throw away, resulting in a final data rate of around 300 MB/s. High energy physics already uses some machine learning in offline analysis, online/offline particle reconstruction, and simulation of accelerator components. The high-energy physics community recently authored a white paper on projections for future computing needs, in which they noted that data requirements are increasing so drastically that current computing infrastructure, which is dependent on off-the-shelf products, will not be usable. Transitioning to machine learning for nearly all applications will become a practical necessity, especially for real-time data reduction and large-scale detector physics simulations. Engineering expertise will be needed for custom compute architecture. Multiple other fields, including astronomy and genomics, are also facing such computing challenges.

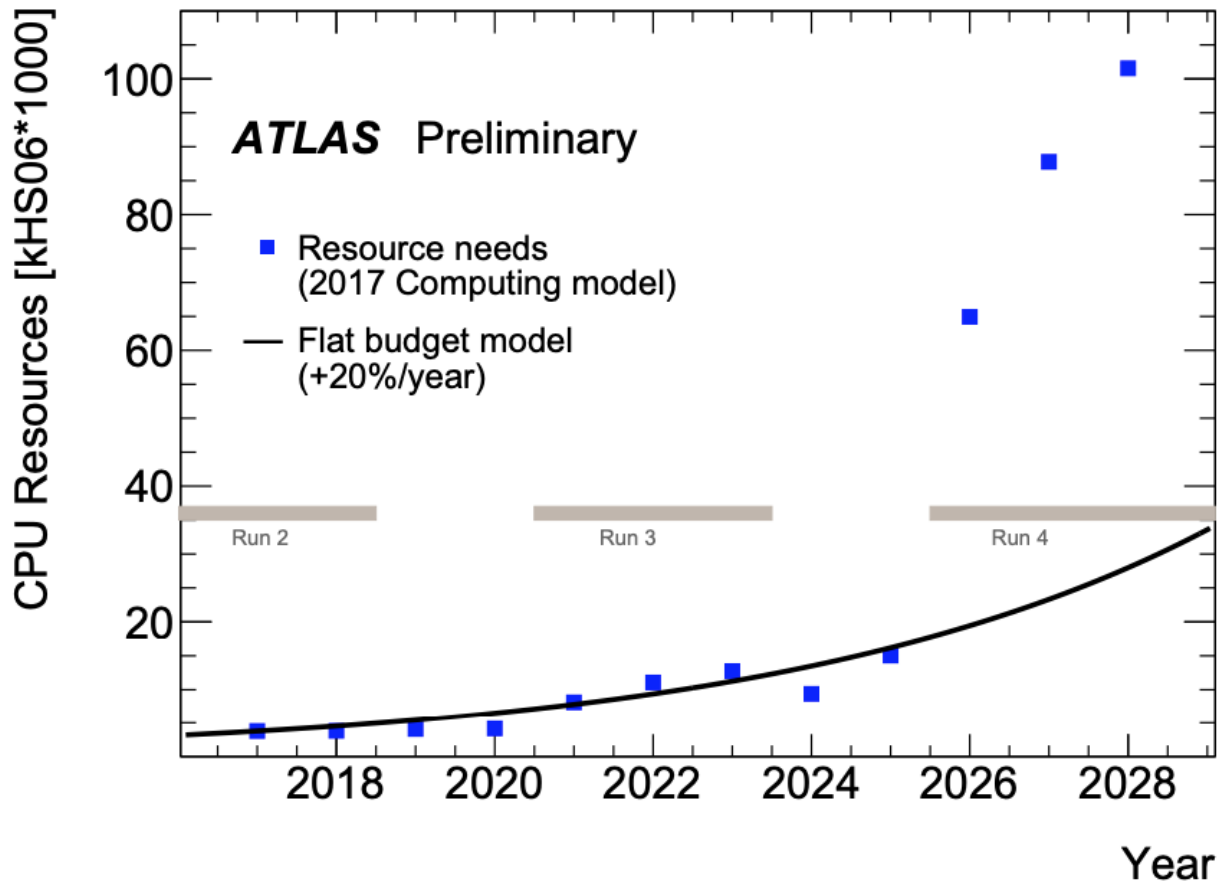

Figure 7: Projected computing resource requirements for the ATLAS experiment at the LHC (HEP Software Foundation 2017).

Autonomous vehicles must integrate multiple high-volume data streams, including radar, LIDAR, cameras, and ultrasonic sensors, to operate safely. Their systems must also be secure, redundant, and low latency, which rules out traditional horizontal scaling. The demand for large-scale, efficient computing for autonomous vehicles has led to a growing interest in custom computing architecture—including for low power as current approaches can require up to 15kW for computing alone. This also implies a corresponding demand for a new workforce and more training programs.

Day 2

Opening Remarks from SRC – Dave Henshall, Director of Business Development and Government Relations, SRC

Microelectronics has the potential to enable a wide variety of groundbreaking future technologies, including sustainable computing and communications; "Industry 4.0" featuring robotics, automation, and advanced manufacturing; 5G+ enabling smart cities and autonomous vehicles; new human-machine interfaces such as augmented and virtual reality; personalized, targeted healthcare and therapies; and quantum computing, information, and positioning systems. The SRC manages collaborative research by recruiting collaborators, running solicitations, managing performance, and ensuring technology transfer to collaborators. In the past, SRC has worked with NIST, the National Science Foundation, and DARPA to design a 3D, monolithically integrated chip for artificial intelligence applications. SRC also provides thought leadership; its Decadal Plan

for Semiconductors, created with input from industry and academia, outlines a series of goals that will enable efficient, high-performance computing into the next decade. These goals include smart sensing that can effectively leverage exponentially growing amounts of analog data; semiconductors several orders of magnitude more energy efficient than present-day technology; drastically increased memory, storage, and communication capacity; and new techniques to meet security challenges.

Analog data is growing exponentially, with an anticipated 45 trillion sensors in place by 2032. Managing this data energy efficiently will require "smart" sensors and new techniques for interpreting data. Meanwhile, semiconductor energy use is also growing rapidly, necessitating orders-of-magnitude increases in efficiency.

Panel on Advanced Analog Hardware for Communications

Thomas Cho, Samsung: The use of CMOS transistors for radio frequency communications is rapidly increasing as transistor sizes decrease, with the total number of connected devices estimated to grow from 27 billion to 39 billion between 2020 and 2025. These devices will enable a wide variety of applications including self-driving cars, smart buildings, industry automation, drones, and wearable devices. However, they face several technical challenges. The end of Moore's Law scaling will impact RF communications as well as transistors; newer RF processes don't always guarantee cost savings, reducing the incentives for manufacturers to adopt them; and the spectrum below 6 GHz is becoming crowded, forcing manufacturers to consider higher frequencies that don't propagate well. New radio architecture will be required to compensate for the loss in propagation. These and other challenges will need to be overcome to continue improving RF power, performance, and area.

CMOS will remain the standard for RF and digital into the near future despite Moore's Law slowing, since no replacement technology is sufficiently mature. However, CMOS is still unsuitable for some applications such as RF front ends below 6 GHz. In new applications such as millimeter waves, which need an array of antennas with power amplifiers and low noise amplifiers, CMOS can be a low-cost solution; however, this is an incremental solution rather than a breakthrough. This workshop provides an opportunity for stakeholders to connect and collaborate on innovative RF technologies.

Jim Booth, NIST: 5G wireless systems have high data rates, shorter latency times of less than 1 millisecond, and the potential to enable massive machine-to-machine communication, enabling a wide range of advanced technologies. However, implementing 5G poses challenges in spectrum management, energy efficiency, and security. The spectrum is a valuable and finite resource, so 5G must use the spectrum efficiently and allow multiple users to coexist. Energy efficiency is necessary to enable massive machine-to-machine communications and preserve the battery life of 5G devices. In addition, as data rates increase, mobile devices will need increased processing power to process the data they receive; this in turn will require more energy unless energy efficiency is increased. Analog components such as active antennas, low-loss switches, tunable phase shifters and filters, matching networks, and energy harvesting can help address these energy challenges. Adaptive materials and devices present opportunities to resolve tradeoffs between energy efficiency and spectrum usage, but these must be balanced against the need to reduce losses at millimeter-wave frequencies. Multiple antennas that can form and direct beams can help overcome losses; a hybrid beamforming system that combines analog and digital components has the potential to do this more efficiently than a digital system.

NIST's Communication Technology Laboratory promotes the development and deployment of advanced communications technologies by disseminating high-quality measurements, data, and research supporting U.S. innovation, industrial competitiveness, and public safety. Its research areas related to 5G include public safety, spectrum sharing and allocation, cybersecurity, and supply chain security. NIST has participated in several studies on microelectronics for 5G communications and has published benchmark industry best practices for low loss measurements.

Steffen McKernan, CarbonTech Inc: Carbon Nanotubes (CNTs) can transform semiconductors by acting as "a switchable metal." Signal amplifiers require high linearity to transmit large quantities of data over a narrow spectrum, which usually demands more power; however, CNTs are highly linear, drastically reducing the power needed for amplification and enabling more data transmission for less spectrum and power. Nanotubes are also highly rugged; they require less cooling and packaging than silicon and GaN transistors. They also carry current at a cross section less than 1% that of copper, dramatically reducing the power consumption of transistors. CNTs also operate at frequencies into the terahertz range with low noise. In addition, they can be constructed more quickly and in fewer steps than present day CMOS transistors. If CNTs are commercialized, the semiconductor industry will benefit from an unprecedented leap forward in performance coupled with a reduction in cost. Nanotubes can be used to make complementary transistors, supplant conventional transistors in RF front ends to improve millimeter wave performance, and create advanced sensors.

However, it can take up to four decades to develop semiconductors made of a new material from prototypes and niche applications to use in mainstream commercial products. To commercialize CNTs this decade to meet AMO's 2030 or bust goal, CarbonTech will need to move at twice that pace. While some companies are already investing in CNT transistors, they are still firmly in the "valley of death," a promising technology that needs substantially more demonstration to be implemented. Government needs to build a community of stakeholders who can participate in this effort.

Technical Talks: Advanced Analog Hardware for Sensors

Jennifer Hasler, Georgia Institute of Technology: Advanced chips made up of a grid of computational logic blocks (CLBs) enable reconfigurable devices, such as Field Programmable Gate Arrays (FPGAs), However, FPGAs have high energy requirements – in any given application, an FPGA will use 30 to 100 times more energy than a chip that was purpose-built for the same function. This large energy use is part of a general trend in which digital computing is nearing limits of power efficiency.

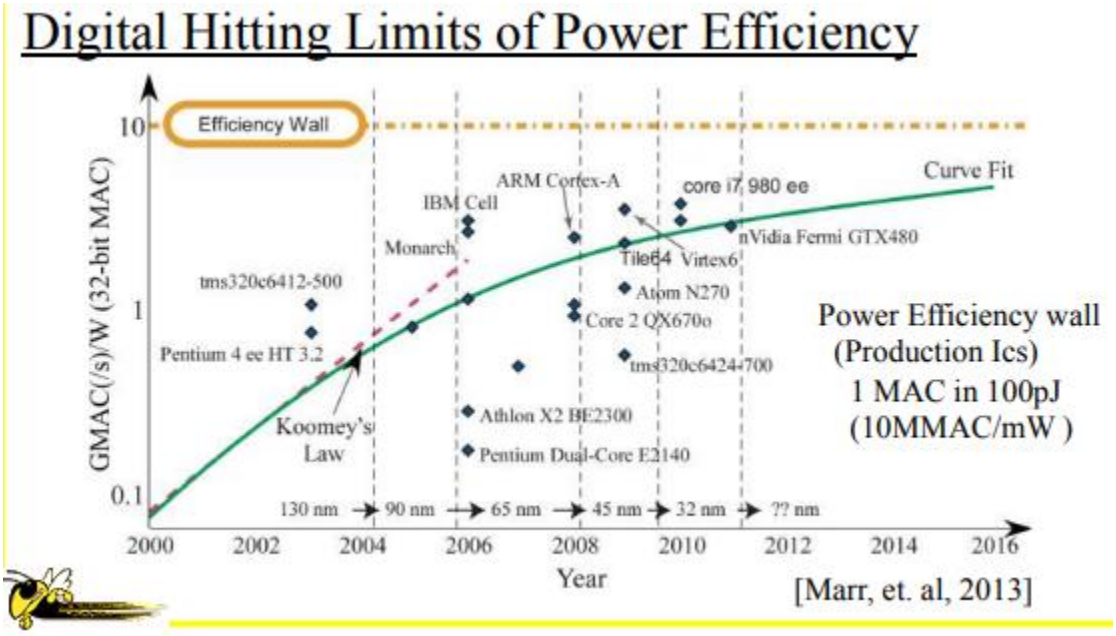

Figure 8: Energy consumption of digital devices since 2000 (Hasler 2021).

Field Programmable Analog Arrays (FPAAs) replace the CLBs found in FPGAs with computational analog blocks (CABs). An individual CAB can perform much more complex calculations than a single CLB. In some applications, such as multiplication, analog systems comprising one or two analog transistors can be used to

replace large digital transistor arrays. Because FPAAs can perform computation quickly and output the results in digital code, potential efficiency improvements of up to 1000x over FPGAs are possible. These energy efficiency improvements are comparable to improvements in efficiency due to CMOS scaling from 1979 to the present day. FPAA-based devices can overcome the energy use drawbacks of both cloud-based and local digital processing. In one experiment, an FPAA was able to perform speech recognition with only 23 microwatts of power. FPAAs have been demonstrated in many other applications, including embedded machine learning, vector-matrix multiplication, and spatiotemporal beamforming.

FPAA chips have the potential to bring laptop-like computing capabilities to devices with the power supply of a cell phone, operate individual sensors off of energy harvested from the environment, and even add functionality to radio frequency identification (RFID) devices. They can also simplify the process of designing complex analog computing devices, potentially paving the way for neuromorphic architectures.

Jim Wieser, Texas Instruments: Analog hardware serves as the interface between computers and the real world by enabling sensors and communications systems. The Decadal Plan for Semiconductors calls for fundamental breakthroughs in analog hardware to generate more capable sensors and communications interfaces for applications such as self-driving vehicles, robots, IoT, climate and environmental sensing, industrial emissions and process control, and wearable sensors for health and lifestyle applications. Wearable sensors, in particular, need to be made more compact and energy efficient. When designing analog sensors, how sensor data is used to make decisions must be considered. Sensors produce a large volume of raw data which is first converted to digital data, then operated on by a digital computer system. Processing a large volume of data this way requires large amounts of computing power and memory, and hence is an energy-intensive process – even moving the data from the sensor to the processor imposes a large energy cost. When local processors implement machine learning algorithms to interpret data, those algorithms are generally static – after the initial training, they are rarely updated.

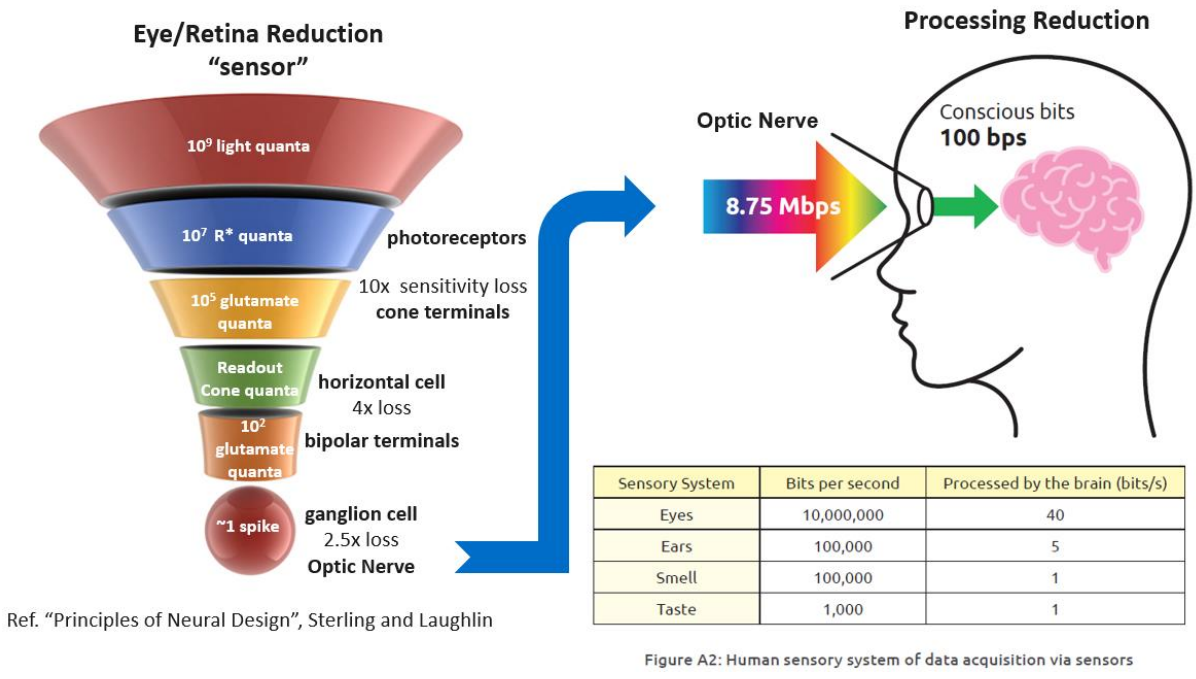

Figure 9: Data reduction in the human optical system (SRC 2021).

Biological "sensors" like our eyes don't function this way – they pre-process the vast majority of the data they receive in order to reduce the amount of data that must be processed by the conscious mind. Analog circuits

are more energy-efficient for applications that require an effective resolution of less than 8 bits. Bio-inspired design can produce analog circuits that pre-process sensor data. Some prototype analog pre-processing classifiers have been demonstrated to reduce the data transmitted to a digital system by up to 1000x and perform complex calculations with 3-10 times less power than an equivalent digital system. These intelligent sensors can make actionable decisions locally without transmitting data to the cloud, thus avoiding drawbacks in latency, security, and energy use.

Day 3

Welcome from DOE Office of Science – Robinson Pino, Program Manager, Office of Science

DOE's missions include securing U.S. leadership in energy technology, science, and innovation, and maintaining U.S. nuclear safety and security. DOE's Office of Science (SC) includes a total of 10 laboratories and 27 user facilities, which include supercomputers, particle accelerators, fusion testbeds, and other advanced experimental equipment, and are free to use for non-proprietary work published in open literature. High performance computing and simulation are vital to all of DOE's missions, and future computing technologies, including quantum, neuromorphic, and probabilistic computing, hold promise for next-generation DOE mission applications. In particular, new directions for applied mathematics and computer science could enable new scientific discoveries across SC, and advanced energy efficient microelectronics and power electronics will be needed to enable exascale computers and a smart electricity grid.

Present-day supercomputers take up about the same area as a basketball court, have 300,000 to 500,000 leading-edge processor and memory chips, and use more than 10 MW of power. They have a useful lifetime of 4-5 years. Emerging energy-related applications, including autonomous vehicles, smart grid, edge computing, AI, and machine learning will require new computing paradigms.

SC's programmatic activities provide foundational technology for autonomous operations, sense-making, and reconfigurable *in-situ* analysis. SC aims to improve not just information processing, but information understanding. Neuro-inspired architecture in software-hardware can save large amounts of energy and offer a path for much needed technological evolution; success in this field will require a cross-disciplinary effort in physics, chemistry, biology, mathematics, engineering, computer science, and neuroscience.

SC's microelectronics priorities are to define innovative material, device, and architecture requirements driven by applications, algorithms, and software; revolutionize memory and data storage; reimagine information flow unconstrained by interconnects; redefine computing by leveraging unexploited physical phenomena; and reinvent the electricity grid through new materials, devices, and architectures. The DOE 5G for Science Initiative's goal is to revolutionize wireless communication in extreme environments through advances in materials science and physics, reinvent scientific instrumentation and critical national infrastructure with wireless technology to provide rapid, AI-driven adaptation, reinvent the digital continuum linking the wireless edge to advanced scientific user facilities, data centers, and high-performance computing, revolutionize AI-enabled edge computing for advanced wireless, and accelerate innovation using community testbeds. The 5G for Science Initiative also will reinvent the digital continuum linking the wireless edge to advanced scientific user facilities, data analysis, and high-performance computing.

Panel on Brain-Inspired Computational Approaches

Lawrence Spracklen, Numenta: AI today has made significant progress, to the point where it can outperform humans at a variety of different tasks. However, this progress was achieved largely through brute force – increasing the size of models to encompass trillions of parameters and investing massive computing, power, and data resources into training. The much-publicized GPT-3 model was estimated to cost $10 million to train. Once trained, the models are static and can't be updated; attempts to update models with new data may impede

their ability to process the data they have. Finally, models are highly fragile and can produce wildly different outputs with only small changes in input. The current state of the art is a long way from anything that can be termed artificial general intelligence (AGI).

Numenta examined the neocortex to see how its features could inform AI design. The neocortex has a large number of neurons, but neuron interconnections and activations are sparse. Also, biological neurons are much more complex than the point neuron abstraction learned by neural networks. As a result, biological brains can identify images based on a single training example, not the hundreds of examples needed by AI. Numenta plans to take some of these concepts and apply them to AI. Because it takes a long time to build new hardware and write software for new hardware, Numenta is trying to apply neocortex-inspired optimizations to AI today by switching from dense to sparse neural network, creating active dendrites, and refining AI training to reduce the amount of data required. Each step in this roadmap can achieve cost reductions of 10x to 100x, resulting in a 1-million-fold reduction in AI costs compared with today's hardware.

Bruno Olshausen, UC – Berkeley: An advanced supercomputer like LBNL's NERSC consumes five megawatts of power and occupies a large building. Insects like the jumping spider have similar impressive capabilities for pattern recognition and geometric reasoning on a miniscule energy budget. Most neural network experiments aim to get the neural network working on a large, powerful system, and then try to make it more efficient afterwards. This approach is too limited to get the massive improvements in energy efficiency needed to offset exploding energy use. The principles that biological computation systems use to run efficiently must be studied and systems built at the appropriate scale from those principles.

High performance computing in the present day has two different modes: AI and neural networks. AI relies on symbolic computation, which is powerful but brittle: a single misplaced bit can corrupt an entire AI process. In addition, ensuring that every bit is correct imposes a large energy cost. The neural network model is more robust but cannot perform symbolic computation. Vector symbolic architecture (VSA) is a way to unite these two modes because it does symbolic computation in a highly parallel distributed system. The basic primitive of VSA computation is a high-dimensional vector, not a single neuron, allowing it to perform symbolic computation in a decentralized manner and making it more robust to failure and more energy efficient.

Dhireesha Kudithipudi, University of Texas at San Antonio: The computational requirements for training conventional AI and machine learning are doubling every 3 to 4 months. Depending on the neuron/synapse model chosen, neuromorphic AI can yield energy savings of up to 60x. An interdisciplinary design flow approach can translate design cues from the nervous system to a neuromorphic model that can solve computing problems.

While biological systems can learn over their entire lifespan: every experience leads to behavioral adaptations that improve performance, few AI have this ability – in fact, learning new tasks often renders them unable to perform previously known tasks. There are several processes that can be adapted from biological systems to solve this problem. Processes that mimic neurogenesis and synaptogenesis would allow a network to overcome fixed capacity limitations and scale the network dynamically by adding and pruning neurons throughout the network's lifetime. Neuromodulation can identify and influence neuronal activity patterns to dynamically adapt and modulate the internal state of the system in response to changes in context. Metaplasticity is a process that governs plasticity depending on the activity being performed to reduce catastrophic forgetting. The University of Texas at San Antonio is currently testing neural networks built to use these types of processes to overcome traditional neural network limitations.

Narayanan Kasthuri, University of Chicago and Argonne National Laboratory: Argonne National Laboratory is participating in the first large-scale effort to understand the hardware of a living brain. This effort will inform AI and neuromorphic computing. Santiago Ramón y Cajal, one of the founders of modern

neuroscience, responsible for discovering the existence of neurons and observing that they were specialized cells was only able to see .01 percent of neurons in the small volumes he examined under his optical microscope. In the 20th century, scientists were able to map the human genome and gain insights into the physical basis of life; in the 21st century, he said we should aim to produce maps of neural connections (also known as connectomes) to understand the physical basis of thought.

In a human brain, there are approximately 100 billion neurons, each of which makes up to 10,000 connections. That's about 10 times more than the number of stars in the Milky Way galaxy. The individual cells and connections exist at the nanoscale, so mapping them will require extremely advanced techniques in microscopy and in machine learning to interpret the resulting images. Scientists are now preparing to produce a complete connectome of a mouse brain, which has about 100 million neurons with 100 billion connections. It's estimated that the complete mouse brain connectome will occupy about 2 exabytes of data – seven orders of magnitude more than the entire human genome. The mouse brain mapping project is a collaboration between NIH and DOE national labs, leveraging synchrotron sources and big data. The resulting data could enable the design of neural networks. Brain development is also of interest to project participants, and they have discovered that many of the rules for emulating adult brains – such as minimizing the number of connections – do not apply in babies' developing brains. This higher interconnectedness is what gives young brains their plasticity and ability to learn rapidly. Incorporating a more connected and plastic brain in early stages of training to a more static one may improve neuromorphic computing and AI.

Panel on Neuromorphic Hardware

Vijay Narayanan, IBM: The amount of computing power used by AI training has increased 750x in the past two years and is continuing to rise, with a corresponding increase in energy use. Today, the carbon footprint of training a single natural language processing (NLP) model is equivalent to the lifetime CO_2 emissions from five cars. Innovations across the stack are required to vastly increase energy efficiency including in algorithms, accelerators, materials, and software. IBM developed and has been following a technology roadmap to improve energy efficiency of AI hardware by 2.5x per year through 2025. IBM's near-term approach is to apply approximate computing principles to digital AI cores (i.e., traditional technology), using algorithms that can accommodate reduced precision with no loss in accuracy. Longer term, analog AI cores will be developed to overcome the von-Neumann bottleneck. These cores can compute in memory with 100x the energy efficiency of traditional architecture.

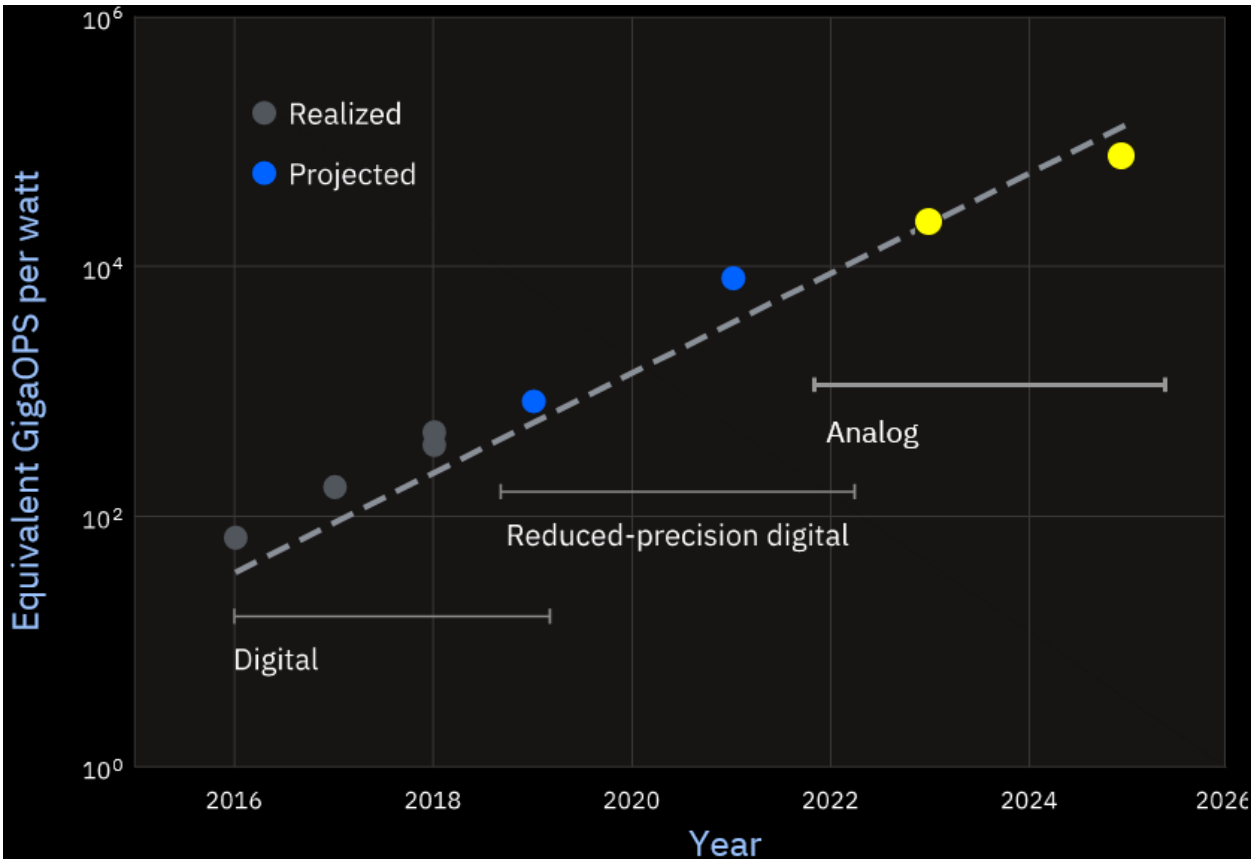

Figure 10: Energy efficiency in GigaOPS per watt of digital, reduced-precision digital, and analog devices (Narayanan 2021).

Hardware implementation of approximate computing uses an array of tunable analog resistive elements that map onto a neural network. By varying conductivity and current in each of these elements, the neural network can perform calculations in parallel with significant improvements in time and power efficiency, compared to serial calculations. Because the computing is analog, each element can encode multiple weights. This approximate analog architecture builds on designs already in use for memory, such as phase change memory and resistive or electrochemical RAM, which use variable resistivity to encode multiple values. Once these weights have been encoded, they can be used for different applications. For example, they can be applied to forward inference which requires long-term retention with resistance to drift, but not endurance and programming speed. But such methods would not be successful for training which requires high programming speeds and endurance, but only modest retention. Currently, IBM's architecture is more efficient for large fully connected matrices such as those used in NLP, speech, and recommendations, but less efficient for convolutional neural networks (CNNs). Scaling approximate computing technology to the data center level will require high-speed IO IP in advanced CMOS nodes. Mapping it to larger models and networks will also require a scalable, chiplet-based design paradigm.

Sean Shaheen, University of Colorado – Boulder: Traditional von Neumann architecture is tailored for computational tasks with bound, deterministic input-output mapping. Von Neumann systems are unable to "learn" like brains, and data transfer between the CPU and storage is energy intensive. Neuromorphic computers are designed to perform real-world computation and classification with unbound and poorly behaved input data, co-allocate data computation and storage to reduce energy use, learn in real time, and resist degradation. Such computers would emulate the brain's sensory system which demonstrates modular and

hierarchical organization. The ability to model these brain functions in computing would open opportunities for improved designs that can be tailored for specific applications.

The use of organic semiconductors in neuromorphic computing can drastically reduce energy per computation and enable a wide range of new capabilities including 3D fabrication and integration, combined sensing and computation, and unconventional computing approaches like reservoir computing. They are not direct competitors with silicon but could supplement silicon in applications such as memristive devices for crossbar array architectures, organic electrochemical transistors for emulating neurons, and semiconducting polymer-based neuronal circuits. The University of Colorado Boulder (CU) is studying organic semiconductor devices to assemble larger multi-gate organic electrochemical "neuro-Boolean" circuits that have the potential to execute Boolean logic with sub-femtojoules of energy consumption per operation. In related work, models of reservoir computers based on realistic organic electrochemical device transfer functions have shown improved performance for time-series classification tasks as compared to standard neuronal activation functions. Other neuromorphic approaches based on metal oxide semiconductors for memristors, and quantum materials with precise and ultra-low energy conductivity switching mechanisms are also being pursued at CU Boulder.

Bhavin Shastri, Queen's University: Deep neural networks are made of many layers with many neurons in each layer. Data is represented into the neural network as a vector and is multiplied by a matrix each time it moves from one layer to the next. The matrix multiplication process is time intensive while data movement is resource intensive. As neural networks become more accurate, demand for computational power increases; currently, the computational power used by neural networks is doubling every 3.4 months. In addition, each layer of a neural network needs to be trained, an energy-intensive process that can emit as much carbon as several cars over their entire lifetimes. Most neural networks today are software networks running on traditional digital hardware. Such hardware is limited in how quickly it can respond. Today, many companies and researchers are trying to build neuromorphic hardware that emulate the operation of neurons; however, even analog electronics face limits in how much they can be miniaturized without suffering from capacitive effects, and so face tradeoffs between bandwidth and complexity. Another option is optical neural networks, which could enable new applications with photons that cannot be achieved with electrons.

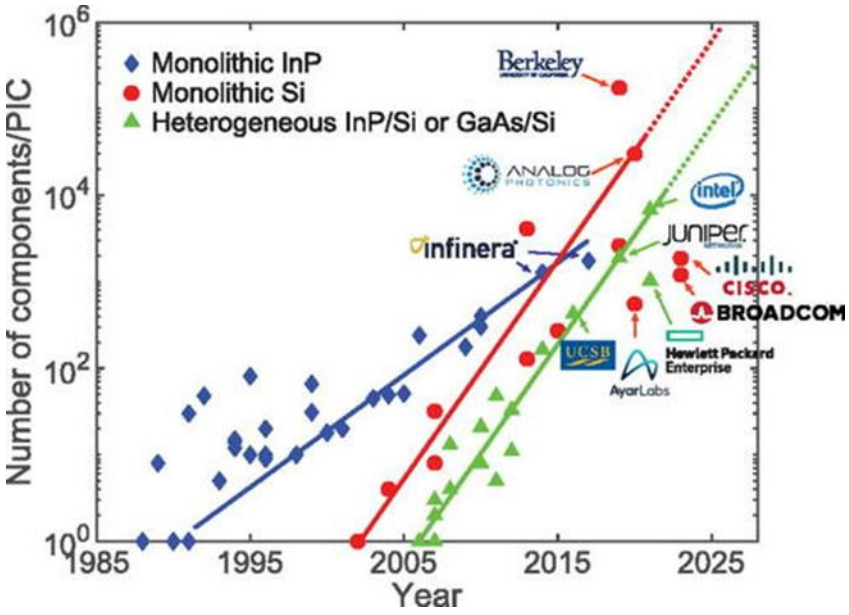

Figure 11: The number of photonic components on a single waveguide for three photonic integrated platforms (Margalit, Xiang, Bowers, et al. 2021).

Optical information processing-systems have been considered as early as 1985. However, these systems were extremely large and thus struggled with nonlinear programming. In the decades since, photonic neural networks have shrunk radically in size, following a “Photonic Moore’s Law”. Neuromorphic photonic architectures can include reservoir computing, multiwavelength networks, coherent networks, spiking networks, diffractive optics, and superconducting networks. Their applications now can include nonlinear programming for robotics, predictive control, and autonomous vehicles; high-performance computing and machine learning; intelligent signal processing for communications; and use in scientific experiments such as particle accelerators. Several startups have begun to release photonic neural network products.

Appendix C: Full Workshop Facilitation Tables

Table C-1: Application areas for advanced analog and neuromorphic deployment
• Data analytics on the edge (e.g., small- and large-scale scientific instruments, remote sensor, satellite, etc.). • Autonomous vehicles. • Autonomous robots for science experiments and manufacturing. • Real-time data analysis (e.g., scientific facilities, extreme weather, etc.). • High throughput materials research. • AI-guided in-situ manufacturing process control. • On-sensor filtering of data. • Applications that make heavy use of (or could benefit from) partial differential equations. • Low-energy communication. • High-noise, low signal computations in IoT. • High-signal, low-energy detectors. • Ultra-low power, edge AI for inference and limited online learning. Applications: Cognitive agents, customized conversational agents that ensure data locality and privacy, sensor fusion. • Aggregated electricity demand response in large buildings and communities. • Hardware for machine learning. • Extremely power constrained nano-robots. • Optimization applications (e.g., convex quadratic optimization). • Biomedical data analytics and smart wearable devices. • Integration of renewables (e.g., grid, industrial processes, etc.). • Analog simulators for quantum computers. • In Vivo imaging (human and machines).

Table C-2: Application areas of neuromorphic computing that is currently not getting enough or any attention
• Non-machine learning use cases of neuromorphic computing, including graph and optimization algorithms. • Miniaturized, flexible, wearable devices (such as personal health monitoring devices, edge sensors) that require small form factor, low weight, and extended operating time but do not have space for bulky battery pack. • Multisensory fusion (vision, RF, audio, text, etc.) near the sensor at very low power for various applications. • Neuromorphic computing-enabled micro drones for environmental monitoring (wildfires, micro weather, etc.). • Neural network architectures using neuromorphic architectures and brain scanning. • New MEMS sensors for simultaneous sensing and computing. • Modeling chemical reactions. • Inspection and boroscopy.

Table C-3: The impact of analog or neuromorphic hardware on energy efficiency for specific applications	
Potential Impact	**Application**
Direct integration with sensors and communication will allow closed control loops at much lower energy and higher precision by remaining in the analog domain (instead of paying costly conversion to digital).	Sensing in large scale DOE experimental instruments.
Analog Edge-AI and Neuromorphic computing ICs are the only promising pathway for realizing 100X higher performance in terms of throughput put unit energy i.e., TOPS/W.	Analog Edge-AI and Neuromorphic computing ICs should be targeted to replace general purpose Edge AI processors. The Digital edge AI processors are typically in the Watts range in power consumption for reasonable datasets. This is not feasible for embedded neural network computing near the sensor.
3-6 orders of magnitude savings should be observed, if implemented efficiently in principle.	Distributed sensing and analysis.
Certain computations at the edge are not currently possible with conventional computing because of the power constraints. Extremely low power neuromorphic implementations will enable these computations, allowing for more intelligent use of the data that is collected there.	On-sensor data analysis.
It seems like until now the energy efficiency was only discussed in the context of global energy consumption. There is an important issue regarding local energy dissipation which will limit computational power.	
Ultra-low power IoT devices capable with simultaneous sensing/computing capabilities.	Various: energy, health, transportation, security, etc. Allows to deploy a high number of IoT devices with a minimum (energy) impact.

Table C-4: Anticipated benefits from the deployment of advanced analog devices.	
Benefits	**Metrics**
Advanced Analog Mixed-Signal (AMS) integrated circuits are promising for on-chip deep learning applications near the sensor or the end user. They will ensure (i) locality and privacy of data, (ii) less reliance on cloud infrastructure, (iii) significant reduction in total energy overheads in communication between edge and server. The overarching metric would be energy efficiency, classification accuracy, size, and latency.	Energy-efficiency: TOPS/W and TOPS/bits/W. The bits would capture the fidelity or signal to noise ratio (SNR) of the operations performed. Latency: time (nanosecond to microsecond) Size: Volume (mm^3)

	Can combine TOPS/W and Classification accuracy for a consolidated metric.
New solvers for PDEs (partial differential equations), DAEs (differential-algebraic equations), ODEs (ordinary differential equations); new optimization techniques and methods.	Convergence time, error, accuracy, processing speed.
Energy efficiency.	Energy processed/unit of information processing.
Low power, real estate, reduced processing and faster processing times.	bit/Joule, bit/m^2.
Local energy consumption as measured by temperature fluctuations, stability of systems, interactions between devices.	A standard test can be implemented to measure energy consumption for a certain task.

Table C-5: Anticipated benefits from the deployment of neuromorphic devices

Benefits	Metrics
Enormous reduction in latency for inference and data filtering.	Higher "experimental" data throughput (images/sensed data/etc.).
New machine learning algorithm development (i.e., more brain-inspired learning approaches).	Training time, amount of data required to train.
For neuromorphic, the advantages come from at least three aspects: 1) energy efficiency or unit system-level information processed; 2) high-noise signal processing; 3) storage energy efficiency.	Energy/unit information.
Low power. Novel computing paradigms inspired by the emerging understanding of the brain.	Same as energy efficiency metrics.
Lighter and faster devices, low power/carbon efficient devices.	Performance and carbon released in pounds.
Parallel computing, computing closer to the signal/process.	Operations per second, overall performance, energy, data movement.

Table C-6: What impact does wireless communications have on the energy consumption of electronic systems?

- Lowers energy consumption per bit compared with electronic wireline. This is even more true if you take into account the build costs of wireline. But the lower cost likely increases use so much that the overall energy consumption is higher.
- Wireless will make communications ubiquitous. Overall energy consumption is expected to go up.
- Any wireless system would need to spend the necessary overhead for communication. Even though the wireless transceivers are fairly advanced and optimized, several blocks such as power amplifiers, phase-locked loops, time-interleaved ADCs, etc. can be still further optimized.

- In the case of mmWave, this is one of the challenges because of its energy footprint on both base stations (NB) and user equipment (UE).
- Power efficiency decreases with frequency. This in turns increases the heat load that electronic systems working at high frequencies need to dissipate for the same amount of power.
- Energy consumption should follow the **IoT proliferation trend typically 2x increase in 1.5yr.** However, security will create electronic power dissipation significantly higher than that.
- As we move up in frequency for a given amplifier technology (GaAs, Si, CNTs), the energy per bit will increase.
- Quadrupling of the number of towers quadruples energy use.
- Communication applications like energy maps appear to be energy hogs.
- mmWave communication is a beam-based "directed energy" challenge; as such, for its implementation, a large number of small cells will be needed, multiplying its energy footprint.
- Atmospheric absorption at higher frequency can reduce range, particularly close to the resonant frequencies of oxygen and water molecules. There will be a tradeoff between these losses and gains achieved from more accurate beamforming achievable at higher frequencies. This may be an issue for frequency range 2 of 5G and higher mmWave frequencies.
- Use of wireless experiences "rebound effect" as costs get lower, it gets more use.
- As we move to higher frequencies (sub-THz, THz), e.g., 6G and beyond, the energy profile of wireless communication systems will increasingly impact its energy profile/footprint.
- At higher operating frequencies, the devices need signal amplification, which consumes more power.

Table C-7: What are the most promising analog devices or approaches for energy-efficient, next-generation communications technologies?

- Wideband gap materials/III-V (e.g., GaN, SiC, etc.).
- CNTFETs.
- Hybrid analog-digital approaches for beamforming and full duplexing.
- Fully integrated frequency agile transceivers for spanning a wider range of spectrum.
- Energy efficient devices in the backhaul infrastructure.
- Neuromorphic approaches for online RF signal processing.
- Metamaterial passives.
- Chiplet-based architecture (DARPA-driven).
- Ultra-wideband RFICs.
- Integrated phased-array antennas with PAs/LNAs and MMICs.
- Free space optical systems (can achieve 100Gb/s for ~1W).
- RF/Photonics integration.
- Energy harvesting approaches that utilize thermoelectrics, pyroelectrics, piezoelectrics, triboelectrics.
- Simultaneous sensing/processing (role of analog computing, neuromorphic), quantum sensing (wide RF spectrum with one device) and networking.
- Analog edge-computing.
- Devices based on redox gating materials.

Table C-8: For the devices or approaches identified, what are the most significant manufacturing barriers?

Material quality

- Avoiding contamination of CNTs due to sensitivity with surface adsorbants.
- Controlling diameter, chirality, and alignment of CNTs.
- Reducing defect density and cost of GaN.

EDA tools

- Lack of EDA tools that can model and simulate unconventional materials, processes, multi-component packaging, and heterogeneously integrated packages.
- Understanding the limitations of current modeling capabilities (i.e., accuracy), particularly for nonlinear, anisotropic, and inhomogeneous materials.

Cost

- Achieving competitive cost or cost parity with traditional CMOS technology.

Process-related

- Achieving atomic level control of deposition and growth processes.
- Micromachining high-precision passive devices for THz and mmWave components and devices.
- Dissipating heat in heterogeneously integrated III-V circuits.
- Forming large-scale arrays of piezoelectric energy harvesting devices.
- Incentivizing lab-to-fab transition to accelerate the incorporation of new materials into the market.

Other manufacturing challenges

- Lack of domestic semiconductor manufacturing capabilities.
- Leveraging current fab infrastructure and technologies for new designs and materials.

Table C-9: What are the primary challenges in integrating these devices or approaches with existing semiconductor processes or products (e.g., sensors)?

Process integration

- Ensuring process compatibility when integrating new processes and materials, including thermal budget/management, chemical compatibility, and contamination concerns.
- Ensuring dimensional compatibility of layers and subcomponents, particularly for high-frequency mmWave circuits.
- Engineering interfaces between materials to maintain intended electronic properties.

Other integration challenges

- Distributed nature of knowledge within the semiconductor industry. Individual or organization with materials expertise may not have system level understanding to integrate the components in a way that showcases their strength.
- Developing new testing and characterization techniques such as THz spectroscopy and 3D mapping of signal propagation, as well as leveraging existing techniques such as over-the-air testing.
- Developing system level design tools to evaluate integration.

Table C-10: What research pathways can be pursued to overcome the challenges identified?

Modeling and simulation

- Leveraging DOE Lab's high performance computing capabilities to give designers and industry the ability to explore many designs in a massively parallel way.
- Enabling system-level design with new EDA design tools, including design-for-manufacturing and test capabilities.
- Developing multi-scale (atomic scale to system scale) design, modeling, and validation capabilities.
- Evaluating process-material-performance interactions.
- Developing open-source design tools to lower the cost of exploration of novel material designs
- Developing new design tools for metamaterials-based antennas.
- Creating a design library for FPAAs (using analog components) and its interface with digital systems (for hybrid approaches).

Materials, devices, and processes

- Researching new materials and evaluating their processing and reliability for next generation devices.

- Better understanding of how the integration of novel materials affects the performance of existing components - cost of integration vs benefits.
- Evaluating thermal management solutions.
- Field Programmable Analog Arrays (FPAAs).
- Electrically steerable optics without mechanical motion.
- RF/Photonics integration (RF to photons).

Other R&D pathways

- Identifying a strong application driver.
- Evaluating energy efficiency at the system level for novel materials and devices.
- Creating community to accelerate development of novel technologies through a competitive-collaborative way.
- Create programs that support material-device co-development.

Cross cutting

- Significant applied research funding >10M/project.
- Create a consortium, similar to SEMATECH, to collectively create standardized modules.
- Given the cost and limited capable fab-owning companies, government has to step in to shoulder the burden until these technologies are mature enough to generate revenue through customers.

Table C-11: In what ways can analog devices or approaches improve sensor operations (e.g., sensing capabilities, speed of data transformation/analysis, efficiency of operation)?

- Provide higher energy efficiency and facilitate local sensing to action.
- Enable low power online edge processing to reduce energy consumption and unnecessary data communication.
- Natively analyze analog sensor data (rather than performing an analog to digital conversion), and potentially do the analysis more efficiently, reducing overall cost of computation.
- Enable multi-sensing capabilities with sensor fusion.
- Enable on chip learning.
- Offer better architectural visibility up front regarding what performance metrics are achievable and allowing migratability from node to node, without having to recreate the wheel each time.
- Reduce reaction to stimuli which improves data security and interference mitigation.
- Leverage backscattering of existing RF spectrum (BT, LTE, 5G, Wi-Fi) for near zero energy.
- Enable sensing in harsh environments by leveraging different sensing modalities such as high-resolution radar, impedance sensing etc.
- Integrate memory and sensor to improve speed and efficiency.

Table C-12: What are the most promising analog devices or approaches for improving sensor performance (e.g., energy efficiency, signal processing, signal fidelity, etc.)?

- Floating-Gate Devices (CMOS).
- Backscatter energy harvesting from new spectrum (e.g., mmWave 5G), approximate computing.
- Holistic development analog sensor with associated analog processing.
- Separation of complex computation from simple reaction, simpler computation in hostile environments.
- Analog architectures with enhanced dynamic range.
- >100GHz sensing modalities (including THz) for higher resolution manufacturing "inspection" capability.
- Silicon Technologies, Inc.'s ADONIS platform.
- Printed sensors that use organic electrochemical transistors.

Table C-13: What R&D is needed to accelerate deployment of sensors for in-situ process and/or quality control?
• Developing a machine learning framework that can process various types of sensing signals and use the knowledge to guide process control and optimization. • Understanding the key in-situ process areas that need to be sensed for improved process control. For example, what type of sensing modalities are most important and what type of "information" is required? • Connecting designers with opportunities in the manufacturing space. • Developing over-the-air (OTA) testing for 5G IoT sensors. • Establishing an overall system-of-sensor approach with multiple sensing modalities (i.e., sensor fusion), required connectivity, and detection of key parameters to control locally but report globally. • Developing a program to bring analog designers into manufacturing spaces, similar to the Technologist in Residence program. • Developing self-calibrating in-situ sensors to operate under harsh conditions (e.g., high temperature, high humidity, vibration, dust, etc.). Known uncertainty in measurements (i.e., standards) are needed.

Table C-14: In which application areas are hardware implementations of neuromorphic computing anticipated to have the greatest impact on energy efficiency?
• Autonomous systems. • Sensing and control of building systems. • Small footprint devices deployed in the field or wearables (e.g., IoT, medical/health, federated AI /computing, human augmentation devices, warfighters/soldiers, airborne sensors/drones, etc.). • Local processing of data at the sensor location. • Sensors in dangerous environments. • IoT and high-performance computing at scale. • Applications that use neural networks (e.g., classification, learning, NLP). • Optimization problems, including those for industrial plants and agriculture. • Deep neural network for smart sensor processing at the edge. • Predictive maintenance/Prognostics in machinery and critical equipment. • Control of electric motors and battery storage systems. • Industrial automation (using ML and in-situ sensors) for reactive maintenance etc. • Human-Machine Symbiosis, including brain-computer/brain-machine interfaces and haptics devices • Approximate computing. • Closed-loop regulation of morphology and function in solid state batteries. • Printed technology interfaces with OLEDs and compute.

Table C-15: What are the most promising devices geared towards non-Von-Neumann approaches to computing? Why?
• Superconducting optoelectronic devices for extremely high-speed neuromorphic computing applications such as processing data collected in a high energy physics experiment. • Redox molecules. • Phase transition oxides. • Electrochemical devices engineered to capture key neuronal functions. • Devices that use the four, fundamental degrees of freedom of quantum materials (lattice, charge, orbit, and spin) at room temperature. This is because they have multiple levels, are controllable, they may combine multiple functionalities. They have been used to emulate neurons, dendrites, axons, synapses. • Spin torque oscillators that can do neuromorphic computation in the time domain. It can be used for auditory recognition.

- Devices with a lot of nuances in their state (continuous memristors) and tunability have the potential to target multiple applications, from online learning to ML accelerators.
- For near-term energy efficient inference: flash: floating gate, SONOS, and related devices.
- For longer-term energy efficient training: electrochemical RAM (ECRAM), and possibly non-filamentary or bulk versions of RRAM.
- Organic semiconductors.
- Photonics.
- Super low bandgap diodes.
- DNA-based devices in the far future.
- Single molecule chem sensors.
- Quantum coherent neural networks / devices.

Table C-16: What design challenges are most significant in the further development of neuromorphic or similar non-Von-Neumann devices?

Modeling, simulation, and other software approaches

- Lack of modeling tools at all levels: materials, devices, circuits, systems, including the functionality for co-design.
- Lack of multi-scale/multi-modal design capability.
- Developing software for non-von Neumann machines.
- Lack of robust design verification frameworks, benchmarking platforms for easy use and device add-on by academia.
- Developing algorithmic innovations to achieve 'Iso-Accuracy' for a wide-variety of tasks. This will ensure wider adoption over niche applications (where some loss of accuracy is acceptable).

Holistic design

- We need to think of the large picture: Algorithm + device +architecture solution.
- Understanding the device/circuit/architecture/algorithm interaction - device technology cannot be separated from architecture and software, like CMOS. Device physics directly affect the final performance of the neural network in terms of inference accuracy, training accuracy.

Circuits

- Evaluating the energy/conversion of peripheral circuitry (DAC/ADC), which has direct impact on the TOPS/W metric.
- Minimizing the area of peripheral components/circuitry.

Other

- Developing chiplet based scalable design to span wide range of model sizes, applications & power envelopes.
- Understanding the brain; in biological systems there is a correlation between temperature range in which animals' function and cognition. Perhaps it may be fruitful to spend some effort on providing more precise thermal control instead of designing devices and systems that work in a very broad temperature range.
- Balancing analog and/or digital design approaches with novel devices, and the challenges of integration of the architecture with a software simulator.
- Lack of standards for floating point/fixed point representation (e.g., equivalent to IEEE 754).

Table C-17: What are the most significant manufacturing challenges for neuromorphic or similar non-von Neumann devices?

- Non-availability of NVM memory technology in advanced CMOS nodes (where high-speed IO IP is available). Need to take advantage of advanced packaging/Heterogeneous Integration Technologies for system level performance.

- Controlling processes at very small scale.
- Developing fault tolerant systems.
- Managing variability/reproducibility from device-to-device and compatibility between different technologies.
- Scaling to small sizes while retaining the intended neuromorphic behavior (conductive switching, etc.).
- Proper routing of devices. Some applications require all-to-all connections between layers, scaling quadratically with device count. Our 2D fabrication technology might not be suited for routing at this level.
- Sensitivity of neuromorphic devices to defects, impurities, and small changes to manufacturing processes.
- Identifying large applications drivers to increase end-user demand.
- Standardizing these devices. Right now, we don't know the winning solution; CMOS only, or CMOS + NV memory.
- Managing the variety of materials used. Some designs require a high precision filament formation; others may require fast switching materials. Some require multi-scale manufacturing or use ionic diffusion. Flexibility of manufacturing processes may be key.
- Managing stress induced by the presence of large fields as the devices ae scaled down.
- Navigating the valley-of-death for emerging technologies.
- Testing following chip manufacturing will become significantly more complicated, as the pass/fail, speed binning of digital systems will not be sufficient. There may be complex versions of analog trimming required.

Table C-18: What are the primary challenges in integrating such devices with existing semiconductor processes or products?

- Leveraging heterogenous integration/packaging approaches.
- Addressing the vast compatibility issues that may crop up in manufacturing, including materials compatibilities, chemical compatibilities, small range ionic diffusion, electromigration, thermal diffusion.
- Integrating vastly different materials in a fab imposes challenges in process temperature, material handling, substrate preparation, waste stream handling, etc.
- Integrating novel devices/structures with CMOS, including spintronic neuromorphic and magnetic tunnel junctions.
- Integrating current arithmetic and numerical representations between neuromorphic and existing products/devices/components (e.g., IEEE 754 floating-point vs. neuromorphic "arithmetic").
- The lack of high-performance analog devices/sensors. There may also be issues with how to sum their inputs to deliver high fidelity bits to our remarkably efficient digital processors.

Appendix D: Workshop Attendees

Name	Organization
Gina Adam	George Washington University
Praneet Adusumilli	IBM Research
Khurram Afridi	Cornell University
Moinuddin Ahmed	Argonne National Laboratory
John Aidun	Sandia National Laboratories
Igor Alvarado	National Instruments
Vinay Amatya	Pacific Northwest National Laboratory
Emad Andarawis	GE Research
Allison Arabelo	Texas A&M University
Hanu Arava	Northwestern University
Jayasimha Atulasimha	Virginia Commonwealth University
Prasanna Balaprakash	Argonne National Laboratory
John Baniecki	SLAC National Accelerator Laboratory
Joseph Bates	Singular Computing
Diana Bauer	DOE AMO
Christopher Bennett	Sandia National Labs
David Bergsman	University of Washington
Getnet Betrie	Argonne National Laboratory
Harish Bhandari	RMD
Kshitij Bhardwaj	Lawrence Livermore National Lab
Kwabena Boahen	Stanford University
Jim Booth	NIST

Name	Organization
Frank Borris	U.S. Department of Energy
Mary Breton	IBM
Brian Calvert	AI Start-up
James Cameron	DuPont Electronics & Industrial
Suma Cardwell	Sandia National Laboratories
Gabriella Carini	Brookhaven National Laboratory
Krishnendu Chakrabarty	Duke University
Frances Chance	Sandia National Laboratories
Abhijit Chatterjee	Georgia Tech
Degang Chen	Iowa State University
Lizhong Chen	Oregon State University
Yiran Chen	Duke University
Zhihong Chen	Purdue University
Mark Cheng	University of Alabama
Ramesh Chettuvetty	Infineon Technologies
Jay Chittooran	Samsung
Kyeongjae Cho	University of Texas – Dallas
Thomas Cho	Samsung
Eugene Chow	PARC, a Xerox Company
Eric Church	DOE HEP
Bennett Cromer	Cornell University
Jeremiah Croshaw	University of Alberta
Abby Davis	MITRE Engenuity

Name	Organization
Grzegorz Deptuch	Brookhaven National Laboratory
Jia Di	University of Arkansas
Alexander Edwards	The University of Texas at Dallas
Dan Ewing	Dept of Energy Kansas City National Security Campus
Mark Feng	Polykala Technologies LLC
Domingo Ferrer	GlobalFoundries
Martin Frank	IBM Research
Lisa Friedersdorf	White House Office of Science and Technology Policy
Joseph Friedman	University of Texas at Dallas
Robert Galli	Infineon
Maya Gokhale	Lawrence Livermore National Laboratory
Ramesh Harjani	University of Minnesota
Jennifer Hasler	Georgia Institute of Technology
David Henshall	SRC
Robert Hershey	Robert L. Hershey, P.E.
Quang Anh Hoang	Drexel University
Qiang Huang	University of Alabama
Jon Ihlefeld	University of Virginia
Subramanian Iyer	UCLA
Conrad James	Sandia National laboratories
Rajiv Joshi	IBM
Tina Kaarsberg	US Department of Energy
Vedant Karia	University of Texas – San Antonio

Name	Organization
Narayanan Kasthuri	University of Chicago/ Argonne National Laboratory
Jiyoung Kim	The University of Texas at Dallas
Kyungtae Kim	Los Alamos National Laboratory
Wiley Kirk	3D Epitaxial Technologies
Scott Koziol	Baylor University
Dhireesha Kudithipudi	University of Texas San Antonio
Shruti Kulkarni	Oak Ridge National Laboratory
Santosh Kurinec	Rochester Institute of Technology
Nick Lalena	Advanced Manufacturing Office
Min-Ha Lee	KITECH North America
Wai Lee	Texas Instruments
Vincent Leung	Baylor University
Elizabeth Lewis	Department of Energy Office of Science
Xiuling Li	University of Texas, Austin
Zhiyong Li	Sandia National Laboratories
Sam Lilak	UCLA
Ankur Limaye	Pacific Northwest National Laboratory
Cosmi Lin	UC Berkeley
JengPing Lu	PARC, a Xerox Company
Rob MacCurdy	CU Boulder
Atif Mahmood	Binghamton University
Christian Mailhiot	Sandia National Laboratories
Rafic Makki	Mubadala Capital

Name	Organization
Anil Mane	Argonne National Laboratory
Vinayak Manmadkar	Innovation Laboratory Energy (ILAB-E)
Matthew Marinella	Sandia National Laboratories
JW McCamy	Vitro Architectural Glass
Steffen McKernan	Carbon Technology, Inc.
Elspeth McSweeney	Brookhaven National Lab
Apurva Mehta	SLAC National Accelerator Lab
Jeremy Mehta	DOE AMO
Abdeljalil Mekkaoui	DHS
Haichao Miao	Lawrence Livermore National Laboratory
Todd Miller	GE Research
William Miller	DOE Office of Science ASCR
Sandeep Miryala	Brookhaven National Laboratory
AhmedElmutasim Mohamed	Eltesmanians Int for Sustainable Development Co. Ltd
Wouter Mortelmans	MIT
Joseph Moses	Akanu Ibiam Federal Polytechnic Unwana
Staci Moulton	Forge Nano
Karthikeyan Nagarajan	The Pennsylvania State University
Hari Nair	Cornell University
Chang-Yong Nam	Brookhaven National Laboratory
Vijay Narayanan	IBM
Kai Ni	Rochester Institute of Technology
John Oakley	SRC

Name	Organization
Yaw Obeng	NIST
Bruno Olshausen	UC Berkeley
Christopher Oshman	DOE
Samuel Palermo	Texas A&M University
Satyavolu Papa Rao	NY CREATES & CNSE/ SUNY Polytechnic Institute
Dishit Parekh	IBM Research
Gregory Parsons	North Carolina State University
Amanda Petford-Long	Argonne National Laboratory
Robinson Pino	DOE Office of Science
Kevin Pintong	PARC
Manuel Quevedo	UT-Dallas
Priyanka Raina	Stanford University
Lavanya Ramakrishnan	Lawrence Berkeley National Laboratory
Mark Raymond	The Research Foundation for SUNY
Juan Rey	Siemens EDA
Ron Rohrer	Southern Methodist University
Ridah Sabouni	Energetics
Sourabh Saha	Georgia Institute of Technology
Fernando Salcedo	ORAU
Andrés Sarmiento	UFSC
Vishal Saxena	University of Delaware
Sina Sayyah Ensan	Pennsylvania State University
Kristine Schroeder	Silicon Technologies, Inc.

Name	Organization
Ivan Schuller	University of California San Diego
Catherine Schuman	Oak Ridge National Laboratory
Tapan Shah	General Electric
Sean Shaheen	University of Colorado Boulder
Sadasivan Shankar	SLAC National Laboratory (Research Technology Manager) and Stanford University (Materials Science)
Bhavin Shastri	Queen's University
Yeuan-Ming Sheu	DARPA
Sam Shichijo	University of Texas at Dallas
Kenta Shimizu	Energetics
Nikhil Shukla	University of Virginia
Miguel Singh	MLS Consulting Engineering Company
Matthew Skaggs	Hydro Dynamic Power Systems
Jaqueline Soriano	Victory13
Laura Spinella	National Renewable Energy Laboratory
Lawrence Spracklen	Numenta
Steven Spurgeon	Pacific Northwest National Laboratory
Narayan Srinivasa	Intel Corporation
Michele Steinbach	North Carolina State University
Adam Stieg	UCLA / California NanoSystems Institute
Changwon Suh	DOE AMO
Jonathan Sun	IBM Research
Paul Syers	DOE AMO
Emmanuel Taylor	Energetics

Name	Organization
Praveen Thallapally	Pacific Northwest National Laboratory
Jeremy Theil	Xperi
Shiva Shankar Thiagarajan	The University of Texas at Dallas
Zhiting Tian	Cornell University
Cliff Tsay	Quick's Net Consulting
Pete Tseronis	Dots and Bridges LLC
Antonino Tumeo	Pacific Northwest National Laboratory
Jeffrey Vetter	Oak Ridge National Laboratory
Aaron Vigil-Martinez	MAD Ventures
Nguyen Vu	University of Michigan
Chunguang Wang	Purdue University
Dawei Wang	Carbon Technology, Inc
Jian-Ping Wang	University of Minnesota
Xinjun Wang	University of Maryland
Jim Wieser	Texas Instruments
H.-S. Philip Wong	Stanford University
Tianyao Xiao	Sandia National Laboratories
J. Joshua Yang	University of Southern California
Angel Yanguas-Gil	Argonne National Laboratory
Jinkyoung Yoo	Los Alamos National Laboratory
Shimeng Yu	Georgia Institute of Technology
Andriy Zakutayev	National Renewable Energy Laboratory
Ramtin Zand	University of South Carolina

Name	Organization
Yuping Zeng	University of Delaware
Yuepeng Zhang	Argonne National Lab
Jingzhou Zhao	Western New England University
Victor Zhirnov	Semiconductor Research Corporation
Peng Zhou	The University of Texas at Dallas

U.S. DEPARTMENT OF ENERGY

Office of ENERGY EFFICIENCY & RENEWABLE ENERGY

For more information, visit:
energy.gov/eere/amo

DOE/EE-2632 • August 2022