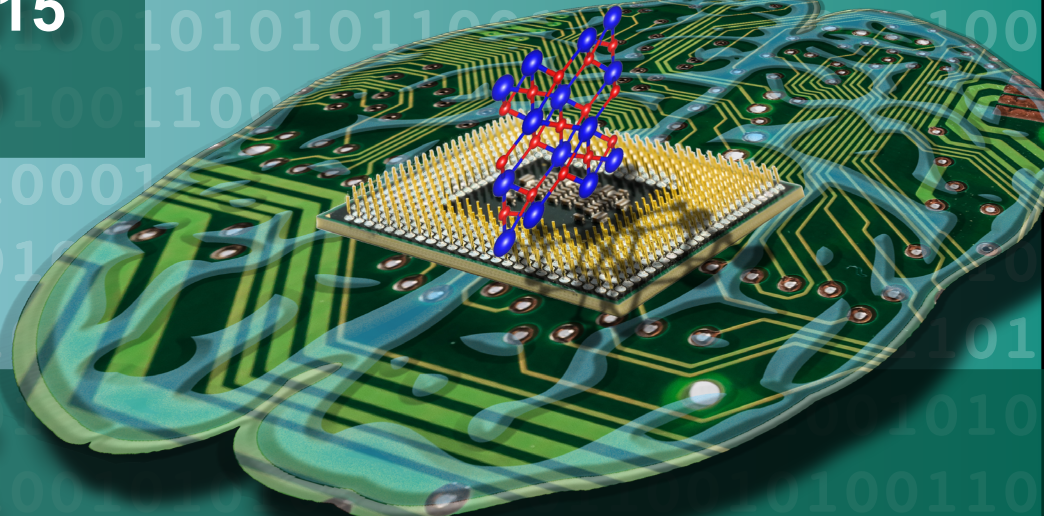
Neuromorphic Computing: From Materials to Systems Architecture

Report of a Roundtable Convened to Consider Neuromorphic Computing Basic Research Needs

October 29-30, 2015
Gaithersburg, MD



Organizing Committee
Ivan K. Schuller (Chair),
University of California, San Diego
Rick Stevens (Chair),
Argonne National Laboratory and University of Chicago

U.S. DEPARTMENT OF
ENERGY
Office of Science

Neuromorphic Computing: From Materials to Systems Architecture

Report of a Roundtable Convened to Consider Neuromorphic Computing Basic Research Needs

October 29-30, 2015
Gaithersburg, MD

Organizing Committee

Ivan K. Schuller (Chair), University of California, San Diego

Rick Stevens (Chair), Argonne National Laboratory and University of Chicago

DOE Contacts

Robinson Pino, Advanced Scientific Computing Research

Michael Pechan, Basic Energy Sciences

Contents

EXECUTIVE SUMMARY

Computation in its many forms is the engine that fuels our modern civilization. Modern computation—based on the von Neumann architecture—has allowed, until now, the development of continuous improvements, as predicted by Moore's law. However, computation using current architectures and materials will inevitably—within the next 10 years—reach a limit because of fundamental scientific reasons.

DOE convened a roundtable of experts in neuromorphic computing systems, materials science, and computer science in Washington on October 29-30, 2015 to address the following basic questions:

Can brain-like ("neuromorphic") computing devices based on new material concepts and systems be developed to dramatically outperform conventional CMOS based technology? If so, what are the basic research challenges for materials sicence and computing?

The overarching answer that emerged was:

The development of novel functional materials and devices incorporated into unique architectures will allow a revolutionary technological leap toward the implementation of a fully "neuromorphic" computer.

To address this challenge, the following issues were considered:

The main differences between neuromorphic and conventional computing as related to: signaling models, timing/clock, non-volatile memory, architecture, fault tolerance, integrated memory and compute, noise tolerance, analog vs. digital, and *in situ* learning

New neuromorphic architectures needed to: produce lower energy consumption, potential novel nanostructured materials, and enhanced computation

Device and materials properties needed to implement functions such as: hysteresis, stability, and fault tolerance

Comparisons of different implementations: spin torque, memristors, resistive switching, phase change, and optical schemes for enhanced breakthroughs in performance, cost, fault tolerance, and/or manufacturability

The conclusions of the roundtable, highlighting the basic research challenges for materials science and computing, are:

1. Creating the architectural design for neuromorphic computing requires an integrative, interdisciplinary approach between computer scientists, engineers, physicists, and materials scientists

2. Creating a new computational system will require developing new system architectures to accommodate all needed functionalities
3. One or more reference architectures should be used to enable comparisons of alternative devices and materials
4. The basis for the devices to be used in these new computational systems require the development of novel nano and meso structured materials; this will be accomplished by unlocking the properties of quantum materials based on new materials physics
5. The most promising materials require fundamental understanding of strongly correlated materials, understanding formation and migration of ions, defects and clusters, developing novel spin based devices, and/or discovering new quantum functional materials
6. To fully realize open opportunities requires designing systems and materials that exhibit self- and external-healing, three-dimensional reconstruction, distributed power delivery, fault tolerance, co-location of memory and processors, multistate—i.e., systems in which the present response depends on past history and multiple interacting state variables that define the present state
7. The development of a new brain-like computational system will not evolve in a single step; it is important to implement well-defined intermediate steps that give useful scientific and technological information

Successfully addressing these challenges will lead to a new class of computers and systems architectures. These new systems will exploit massive, fine-grain computation; enable the near real-time analysis of large-scale data; learn from examples; and compute with the power efficiency approaching that of the human brain. Future computing systems with these capabilities will offer considerable scientific, economic, and social benefits.

This DOE activity aligns with the recent White House “A Nanotechnology-Inspired Grand Challenge for Future Computing” issued on October 20th, 2015 with the goal to “Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain”. This grand challenge addresses three Administration priorities: the National Nanotechnology Initiative (NNI), the National Strategic Computing Initiative (NSCI), and the BRAIN initiative.

WHY NEUROMORPHIC COMPUTING?

Computers have become essential to all aspects of modern life—from process controls, engineering, and science to entertainment and communications—and are omnipresent all over the globe. Currently, about 5–15% of the world's energy is spent in some form of data manipulation, transmission, or processing.

In the early 1990s, researchers began to investigate the idea of "neuromorphic" computing. Nervous system-inspired analog computing devices were envisioned to be a million times more power efficient than devices being developed at that time. While conventional computational devices had achieved notable feats, they failed in some of the most basic tasks that biological systems have mastered, such as speech and image recognition. Hence the idea that taking cues from biology might lead to fundamental improvements in computational capabilities.

Since that time, we have witnessed unprecedented progress in CMOS technology that has resulted in systems that are significantly more power efficient than imagined. Systems have been mass-produced with over 5 billion transistors per die, and feature sizes are now approaching 10 nm. These advances made possible a revolution in parallel computing. Today, parallel computing is commonplace with hundreds of millions of cell phones and personal computers containing multiple processors, and the largest supercomputers having CPU counts in the millions.

"Machine learning" software is used to tackle problems with complex and noisy datasets that cannot be solved with conventional "non-learning" algorithms. Considerable progress has been made recently in this area using parallel processors. These methods are proving so effective that all major Internet and computing companies now have "deep learning"—the branch of machine learning that builds tools based on deep (multilayer) neural networks—research groups. Moreover, most major research universities have machine learning groups in computer science, mathematics, or statistics. Machine learning is such a rapidly growing field that it was recently called the "infrastructure for everything."

Over the years, a number of groups have been working on direct hardware implementations of deep neural networks. These designs vary from specialized but conventional processors optimized for machine learning "kernels" to systems that attempt to directly simulate an ensemble of "silicon" neurons, better known as neuromorphic computing. While the former approaches can achieve dramatic results, e.g., 120 times lower power compared with that of general-purpose processors, they are not fundamentally different from existing CPUs. The latter neuromorphic systems are more in line with what researchers began working on in the 1980s with the development of analog CMOS-based devices with an architecture that is modeled after biological neurons. One of the more recent accomplishments in neuromorphic computing has come from IBM research, namely, a biologically inspired chip ("TrueNorth") that implements one million spiking neurons and 256 million synapses on a chip with 5.5 billion transistors with a typical power draw of 70 milliwatts. As impressive as this system is, if scaled up to the size of the human brain, it is still about 10,000 times too power intensive.

Clearly, progress on improvements in CMOS and in computer hardware more generally will not be self-sustaining forever. Well-supported predictions, based on solid scientific and engineering data, indicate that conventional approaches to computation will hit a wall in the next 10 years. Principally, this situation is due to three major factors: (1) fundamental (atomic) limits exist beyond which devices cannot be miniaturized, (2) the local energy dissipation limits the device packing density, and (3) the increase and lack of foreseeable limit in overall energy consumption are becoming prohibitive. Novel approaches and new concepts are needed in order to achieve the goals of developing increasingly capable computers that consume decreasing amounts of power.

The Need for Enhanced Computing

The DOE has charted a path to Exascale computing by early in the next decade. Exascale machines will be orders of magnitude faster than the most powerful machines today. Even though they will be incredibly powerful, these machines will consume between 20 and 30 megawatts of power and will not have intrinsic capabilities to learn or deal with complex and unstructured data. It has become clear that the mission areas of DOE in national security, energy sciences, and fundamental science will need even more computing capabilities than what can be delivered by Exascale class systems. Some of these needed capabilities will require revolutionary approaches for data analysis and data understanding.

Neuromorphic computing systems are aimed at addressing these needs. They will have much lower power consumption than conventional processors and they are explicitly designed to support dynamic learning in the context of complex and unstructured data. Early signs of this need show up in the Office of Science portfolio with the emergence of machine learning based methods applied to problems where traditional approaches are inadequate. These methods have been used to analyze the data produced from climate models, in search of complex patterns not obvious to humans. They have been used to recognize features in large-scale cosmology data, where the data volumes are too large for human inspection. They have been used to predict maintenance needs for accelerator magnets—so they can be replaced before they fail—to search for rare events in high-energy physics experiments and to predict plasma instabilities that might develop in fusion reactors. These novel approaches are also being used in biological research from searching for novel features in genomes to predicting which microbes are likely to be in a given environment at a given time. Machine learning methods are also gaining traction in designing materials and predicting faults in computer systems, especially in the so-called "materials genome" initiative. Nearly every major research area in the DOE mission was affected by machine learning in the last decade. Today these applications run on existing parallel computers; however, as problems scale and dataset sizes increase, there will be huge opportunities for deep learning on neuromorphic hardware to make a serious impact in science and technology.

Neuromorphic computing may even play a role in replacing existing numerical methods where lower power functional approximations are used and could directly augment planned Exascale architectures. Important questions for the future are which areas of science are most likely to be impacted by neuromorphic computing and what are the requirements for those deep neural networks. Although this roundtable did not focus on an application driven agenda, it is increasingly important to identify these areas and to further understand how neuromorphic hardware might address them.

VON NEUMANN vs. NEUROMORPHIC

System Level

Traditional computational architectures and their parallel derivatives are based on a core concept known as the von Neumann architecture (see Figure 1). The system is divided into several major, physically separated, rigid functional units such as memory (MU), control processing (CPU), arithmetic/logic (ALU), and data paths. This separation produces a temporal and energetic bottleneck because information has to be shuttled repeatedly between the different parts of the system. This “von Neumann” bottleneck limits the future development of revolutionary computational systems. Traditional parallel computers introduce thousands or millions of conventional processors each connected to others. Aggregate computing performance is increased, but the basic computing element is fundamentally the same as that in a serial computer and is similarly limited by this bottleneck.

In contrast, the brain is a working system that has major advantages in these aspects. The energy efficiency is markedly—many orders of magnitude—superior. In addition, the memory and processors in the brain are collocated because the constituents can have different roles depending on a learning process. Moreover, the brain is a flexible system able to adapt to complex environments, self-programming, and capable of complex processing. While the design, development, and implementation of a computational system similar to the brain is beyond the scope of today’s science and engineering, some important steps in this direction can be taken by imitating nature.

Clearly a new disruptive technology is needed which must be based on revolutionary scientific developments. In this “neuromorphic” architecture (see Figure 1), the various computational elements are mixed together and the system is dynamic, based on a “learning” process by which the various elements of the system change and readjust depending on the type of stimuli they receive.

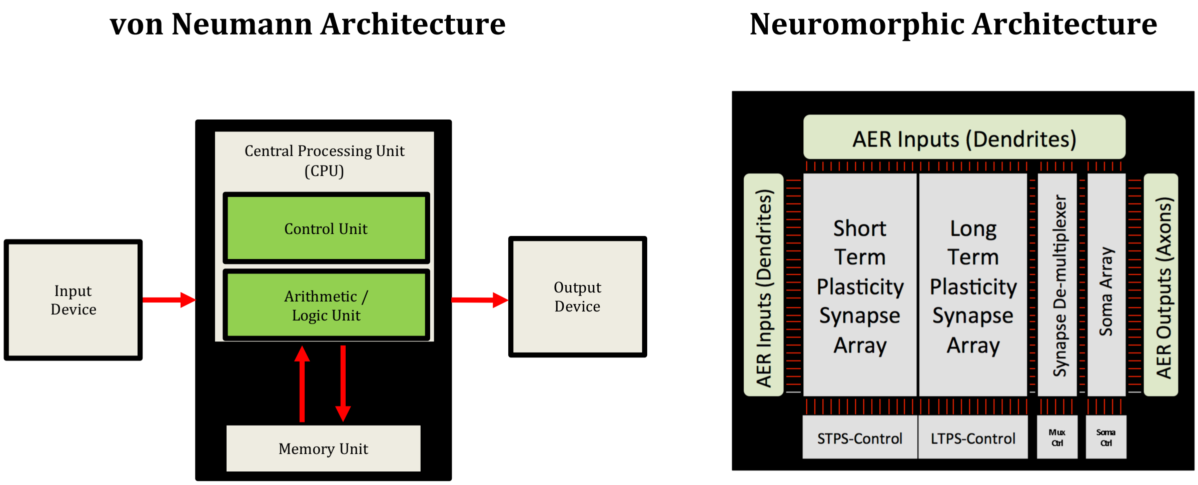


Figure 1. Comparison of high-level conventional and neuromorphic computer architectures. The so-called "von Neumann bottleneck" is the data path between the CPU and the memory unit. In contrast, a neural network based architecture combines synapses and neurons into a fine grain distributed structure that scales both memory (synapse) and compute (soma) elements as the systems increase in scale and capability, thus avoiding the bottleneck between computing and memory.

Device Level

A major difference is also present at the device level (see Figure 2). Classical von Neumann computing is based on transistors, resistors, capacitors, inductors and communication connections as the basic devices. While these conventional devices have some unique characteristics (e.g., speed, size, operation range), they are limited in other crucial aspects (e.g., energy consumption, rigid design and functionality, inability to tolerate faults, and limited connectivity). In contrast, the brain is based on large collections of neurons, each of which has a body (soma), synapses, axon, and dendrites that are adaptable and fault tolerant. Also, the connectivity between the various elements in the brain is much more complex than in a conventional computational circuit (see Figure 2).

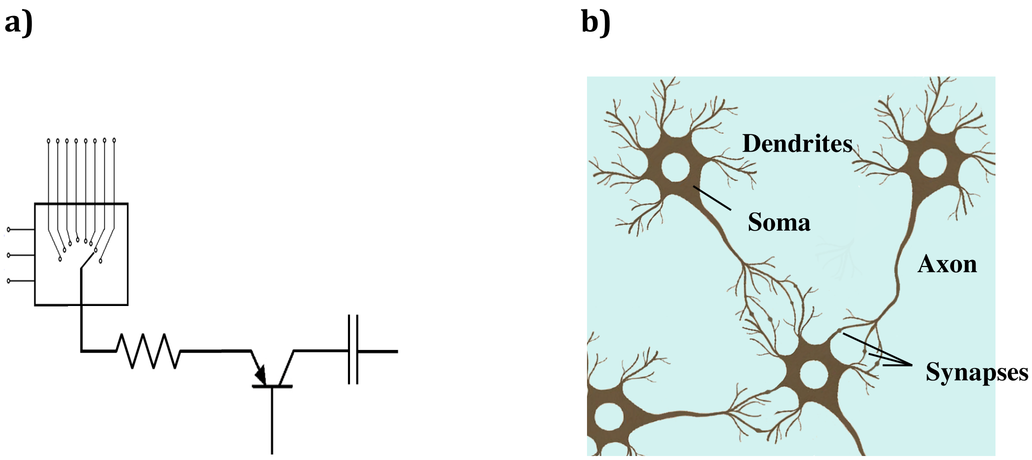


Figure 2. Interconnectivity in a) conventional and b) neuronal circuits.

Performance

A comparison of biological with technological systems is revealing in almost all aspects. Although the individual constituents of silicon systems naively seem to exhibit an enhanced performance in many respects, the system as a whole exhibits a much-reduced functionality. Even under reasonable extrapolations of the various parameters, it appears that the improvement in computational capacity of conventional silicon based systems will never be able to approach the density and power efficiency of a human brain. New conceptual development is needed.

Inspection of the delay time versus power dissipation for devices in many competing technologies is quite revealing (see Figure 3). The neurons and synapses exist in a region in this phase diagram (upper left corner) where no other technologies are available. The energy dissipation in a synapse is orders of magnitude smaller although the speed is much slower.

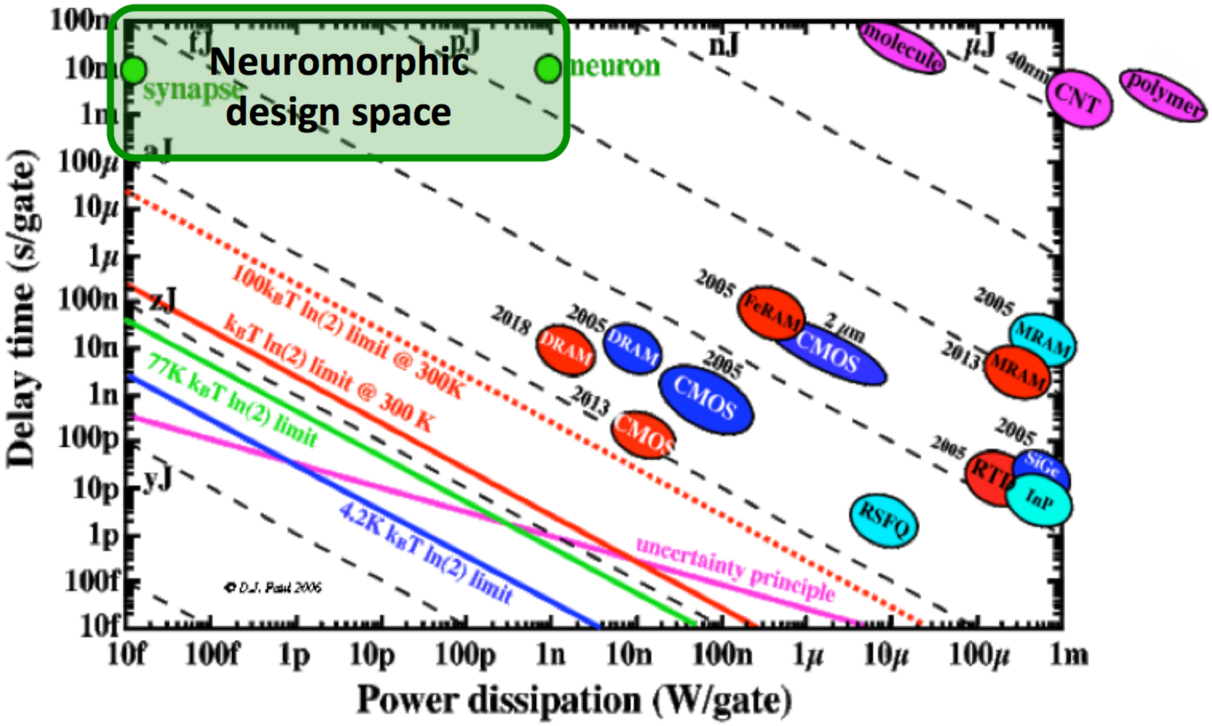


Figure 3. Delay time per transistor versus the power dissipation. The operating regime for neuromorphic devices is in the upper left corner indicating the extreme low power dissipation of biological synapses and the corresponding delay time. Systems built in this region would be more "brain-like" in their power and cycle times (after userweb.eng.gla.ac.uk).

Table 1 compares the performance of biological neurons and neuron equivalents built from typical CMOS transistors currently used in computers. While the values listed in this table may not be exact and are debatable, the differences are so large that it is clear that new scientific concepts are needed, including possibly to the system architecture itself. While silicon devices may exhibit certain advantages, the overall system computational capabilities—its fault tolerance, energy consumption, and ability to deal with large data sets—is considerably superior in biological brains. Moreover, there are whole classes of problems that conventional computational systems will never be able to address even under the most optimistic scenarios.

	Biology	**Silicon**	**Advantage**
Speed	1 msec	1 nsec	1,000,000x
Size	1µm - 10µm	10nm - 100nm	1,000x
Voltage	~ 0.1V	V_{dd} ~1.0V	10x
Neuron Density	100K/mm^2	5k/mm^2	20x
Reliability	80%	< 99.9999%	1,000,000x
Synaptic Error Rate	75%	~ 0%	>10^9
Fan-out (-in)	10^3-10^4	3-4	10,000x
Dimensions	Pseudo 3D	Pseudo 3D	Similar
Synaptic Op Energy	~ 2 fJ	~10pJ	5000x
Total Energy	10 Watt	>>10^3 Watt	100,000x
Temperature	36C - 38C	5C - 60C	Wider Op Range
Noise effect	Stochastic Resonance	Bad	
Criticality	Edge	Far	

Table 1. Comparison of biological and silicon based systems. This table shows a comparison of neurons built with biology to equivalent structures built with silicon. Red is where biology is ahead; black is where silicon is ahead. The opportunity lies in combining the best of biology and silicon.

Technology that has the advantages of both biological and engineered materials, with the downsides of neither, is needed. Thus, major changes are required in nanoscale device designs, new functional materials, and novel software implementations. The general philosophy of building a conventional computational system relies on the ability to produce billions of highly controlled devices that respond the same way to a well-defined stimulus or signal. In contrast, neuromorphic circuits elements (especially synapses) intrinsically are expected to respond differently depending on their past history. Consequently, the design as well as the implementation of new architectures must be dramatically modified.

EXISTING NEUROMORPHIC SYSTEMS

Architectures

A range of computing architectures can support some form of neuromorphic computing (see Table 3 in the appendix for a partial list of historical efforts to build chips that directly implement neural network abstractions). At one end of the spectrum are variations of general-purpose CPU architectures with data paths that are optimized for execution of mathematical approximations of neural networks. These artificial neural network accelerators support the fast matrix oriented operations that are the heart of neural network methods. This architectural approach can beat general-purpose CPUs in power efficiency by a large margin by discarding those features that are not needed by neural network algorithms and may be well suited for integration into existing systems as neural network accelerators (see Figure 4).

At the other end of the spectrum are direct digital or analog implementations of networks of relatively simple neurons, typically based on an abstract version of a leaky integrate-and-fire (LIF) neuron. These have discrete implementations of synapses (simple storage), soma, and axons (integrating and transmitting signals to other neurons). Implementations can be in analog circuits or digital logic. Analog implementations of spike timing dependent plastic (STDP) synapses—a form of learning synapse—have been demonstrated with a few transistors, and an analog leaky integrate-and-fire model of the soma required around 20 transistors. Axons in this case are implemented as conventional wires for local connections within a chip.

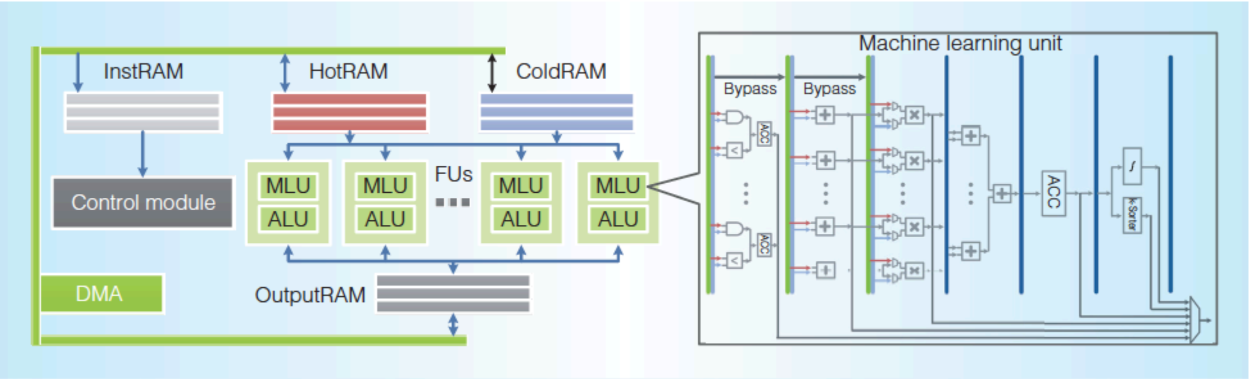


Figure 4. Variation of CPU data path optimized for execution of a mathematical approximation of a neural network. This variation shows dramatic power reduction compared to general purpose CPU when executing key machine learning kernels including neural network execution. It uses a relatively simple data path, and many fixed-width multiple/add units enable this architecture to perform well on the dense floating-point workload characteristic of neural network training. This type of architecture is aimed at a low-power accelerator add-on to existing CPUs. *Figure: Chen, Tianshi, et al.* IEEE Micro *(2015): 24-32.*

For long-distance (off-chip) signaling, an (electronic or optical) analog to digital to analog conversion can be used. The digital equivalent implementation can take considerably more transistors depending on the degree of programmability. The TrueNorth system, for example, is estimated to require about 10 times as many transistors as does an analog equivalent; many of these transistors are used to provide a highly programmable soma, but do not currently support on-chip learning. Power differences between analog and digital implementations are also significant, with analog being several orders of magnitude more efficient. Even the best current analog chips, however, are four to five orders of magnitude less power efficient than biological neurons.

In this report, we focus on the abstract hardware implementation of an abstraction of a neural type network. Clearly novel devices based on new functional materials will have a major impact for such an implementation. Therefore, collaborative research in the synthesis of new nanostructured materials, design and engineering of nanoscaled devices and implementation of creative architectures is needed. This is precisely the approach necessary to dramatically improve power and density needed to reach "brain-like" performance and "brain-like" power levels.

Demonstrations

Recent neuromorphic processor test chips have typically implemented 256 neurons and 65K–256K synapses, whereas IBM's TrueNorth chip, announced in 2015, contained one million neurons and 256 million synapses. Some groups are experimenting with test chips including novel synapse devices based on memristors. Full-system-scale neuromorphic prototypes under development include the one billion-neuron SpiNNaker hybrid CPU/Neuron project at the University of Manchester and the wafer-scale neuromorphic hardware system "FACETS" being developed at the University of Heidelberg that contains 180K neurons and 4 x 10^7 synapses per wafer. The completed system should contain many such wafers. The FACETS system is unique in that it supports more than 10,000 synapses per neuron, making it potentially able to simulate systems that are more biologically plausible and perhaps more powerful.

Project Name	Programmable Structure	Component Complexity (Neuron/Synapse)	On-Chip Learning	Materials/Devices
Desired	**Neurons and synapses**	**< 5 / < 5**	**Yes**	**Novel Materials?**
Darwin[6]	Neurons and synapses	> 5 / > 5	Yes	Fabbed with existing CMOS processes
DANNA[1]	Neurons and synapses	2 / 2	Yes	FPGA, ASIC
TrueNorth[2]	Fixed (Synapses on/off)	10 / 3	No	Fabbed with existing CMOS processes
Neurogrid[3]	Fixed (Synapses on/off)	79 / 8	No	Fabbed with existing CMOS processes
BrainScaleS[4]	Neurons and synapses	Variable	Yes	Wafer-Scale ASIC
SpiNNaker[5]	Neurons and synapses	Variable	Yes	ARM Boards, Custom Interconnection

Table 2. Comparison of some recent neuromorphic device implementations.

A common architectural pattern in neuromorphic hardware is to arrange the synapses in a dense crossbar configuration with separate blocks of logic to implement learning in the synapse and integrate-and-fire for the soma (see Figure 5). A challenge for such systems is the fan-in fan-out ratios for the crossbars. Real biological neurons can have up to 20,000 synapse connections per neuron. Existing neuromorphic chips tend to limit the number of synapses to 256 per neuron.

We currently do not understand precisely when or if artificial neural networks will need to have the number of connections that approach those in biological systems. However, some deep learning networks in production have layers with all-to-all connections between many thousand neurons, but these approaches also use methods to "regularize" the networks by dropping some connections. It might be possible to get by with limitations in the number of synapses per neuron in the thousands. It has also recently been determined that dendrites are not just passive channels of communication from the synapse to the soma, but that they might also play a role in pattern recognition by filtering or recognizing patterns of synaptic inputs and transmitting potentials only in some cases. This might require that we revise the ideas of simple synapses connected by wires.

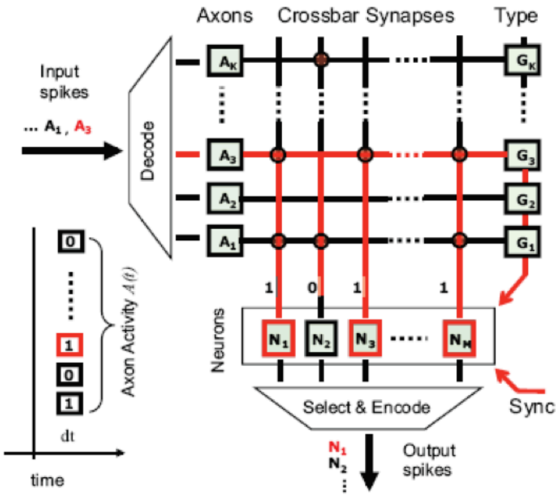


Figure 5. Example of a simple neuromorphic architecture. This diagram illustrates the dominance of the synapse crossbar in most neuromorphic architectures implementations and explains in part why most groups are focused on implementation of synapses as the key scalable component. Since synapse area will dominate most designs, it is imperative that designs minimize synapse area.

For historical comparison, a list of early neuromorphic chips and their scale is available in Table 3 in the Appendix. While many of these implementations have produced considerable advances, none are based on developing completely novel approaches nor based on new neuromorphic materials/devices nor approach the performance expected from a brain-like device.

IMPLEMENTATION NEEDS

In this section, we will outline the important concepts and building blocks that will most likely be used to implement a neuromorphic computer.

Neuromorphic Concepts

The following concepts play an important role in the operation of a system, which imitates the brain. It should be mentioned that sometimes the definitions listed below are used in slightly different ways by different investigators.

Spiking. Signals are communicated between neurons through voltage or current spikes. This communication is different from that used in current digital systems, in which the signals are binary, or an analogue implementation, which relies on the manipulation of continuous signals. Spiking signaling systems are time encoded and transmitted via "action potentials".

Plasticity. A conventional device has a unique response to a particular stimulus or input. In contrast, the typical neuromorphic architecture relies on changing the properties of an element or device depending on the past history. Plasticity is a key property that allows the complex neuromorphic circuits to be modified ("learn") as they are exposed to different signals.

Fan-in/fan-out. In conventional computational circuits, the different elements generally are interconnected by a few connections between the individual devices. In the brain, however, the number of dendrites is several orders of magnitude larger (e.g., 10,000). Further research is needed to determine how essential this is to the fundamental computing model of neuromorphic systems.

Hebbian learning/dynamical resistance change. Long-term changes in the synapse resistance after repeated spiking by the presynaptic neuron. This is also sometimes referred to as spike time-dependent plasticity (STDP). An alternative characterization in Hebbian learning is "devices that fire together, wire together".

Adaptability. Biological brains generally start with multiple connections out of which, through a selection or learning process, some are chosen and others abandoned. This process may be important for improving the fault tolerance of individual devices as well as for selecting the most efficient computational path. In contrast, in conventional computing the system architecture is rigid and fixed from the beginning.

Criticality. The brain typically must operate close to a critical point at which the system is plastic enough that it can be switched from one state to another, neither extremely stable nor very volatile. At the same time, it may be important for the system to be able to explore many closely lying states. In terms of materials science, for example, the system may be close to some critical state such as a phase transition.

Accelerators. The ultimate construction of a neuromorphic–based thinking machine requires intermediate steps, working toward small-scale applications based on neuromorphic ideas. Some of these types of applications require combining sensors with some limited computation.

Building Blocks

In functional terms, the simplest, most naïve properties of the various devices and their function in the brain areas include the following.

1. **Somata** (also known as **neuron bodies**), which function as integrators and threshold spiking devices
2. **Synapses**, which provide dynamical interconnections between neurons
3. **Axons**, which provide long-distance output connection between a presynaptic to a postsynaptic neuron
4. **Dendrites**, which provide multiple, distributed inputs into the neurons

To implement a neuromorphic system that mimics the functioning of the brain requires collaboration of materials scientists, condensed matter scientists, physicists, systems architects, and device designers in order to advance the science and engineering of the various steps in such a system. As a first step, individual components must be engineered to resemble the properties of the individual components in the brain.

Synapse/Memristor. The synapses are the most advanced elements that have thus far been simulated and constructed. These have two important properties: switching and plasticity. The implementation of a synapse is *frequently* accomplished in a two-terminal device such as a memristor. This type of devices exhibits a pinched (at V=0), hysteretic I-V characteristic.

Soma/Neuristor. These types of devices provide two important functions: integration and threshold spiking. Unlike synapses, they have not been investigated much. A *possible* implementation of such a device consists of a capacitor in parallel with a memristor. The capacitance ("integration") and spiking function can be engineered into a single two-terminal memristor.

Axon/Long wire. The role of the axon has commonly (perhaps wrongly) been assumed simply to provide a circuit connection and a time delay line. Consequently, little research has been done on this element despite the fact that much of the dissipation may occur in the transmission of information. Recent research indicates that the axon has an additional signal-conditioning role. Therefore, much more research is needed to understand its role and how to construct a device that resembles its function.

Dendrite/Short wire. The role of dendrites is commonly believed simply to provide signals from multiple neurons into a single neuron. This in fact emphasizes the three-dimensional nature of the connectivity of the brain. While pseudo-3D systems have been implemented in multilayer (~8) CMOS-based architecture, a truly 3D implementation needs further research and development. In addition, recent work in neuroscience has determined that dendrites can also play a role in pattern detection and subthreshold filtering. Some dendrites have been shown to detect over 100 patterns.

Fan-in/Fan-out. Some neurons have connections to many thousands of other neurons. One axon may perhaps connect to ten thousand or more dendrites. Current electronics is limited to fan-in/fan-out of a few tens of terminals. New approaches to high-radix connections may be needed; currently, crossbars are used in most neuromorphic systems but they have scaling limits.

Many of the needed functions can be (and have been) implemented in complex CMOS circuits. However, these not only occupy much real estate, but also are energy inefficient. The latter perhaps is a crucial fundamental limitation as discussed above. *Thus, for the next step in the evolution of brain-like computation, it is crucial to build these types of devices from a single material that is sufficiently flexible to be integrated at large-scale and have minimal energy consumption.*

PROPOSED IMPLEMENTATION

Architecture

Ultimately, an architecture that can scale neuromorphic systems to “brain scale” and beyond is needed. A brain scale system integrates approximately 10^{11} neurons and 10^{15} synapses into a single system. The high-level neuromorphic architecture illustrated in Figure 1 consists of several large-scale synapse arrays connected to soma arrays such that flexible layering of neurons (including recurrent networks) is possible and that off-chip communication uses the address event representation (AER) approach to enable digital communication to link spiking analog circuits. Currently, most neuromorphic designs implement synapses and somata as discrete sub-circuits connected via wires implementing dendrites and axons. In the future, new materials and new devices are expected to enable integrated constructs as the basis for neuronal connections in large-scale systems. For this, progress is needed in each of the discrete components with the primary focus on identification of materials and devices that would dramatically improve the implementation of synapses and somata.

One might imagine a generic architectural framework that separates the implementation of the synapses from the soma in order to enable alternative materials and devices for synapses to be tested with common learning/spiking circuits (see Figure 6). A reasonable progression for novel materials test devices would be the following: (1) single synapse-dendrite-axon-soma feasibility test devices, (2) chips with dozens of neurons and hundreds of synapses, followed by (3) demonstration chips with hundreds of neurons and tens of thousands of synapses.

Once hundreds of neurons and tens of thousands of synapses have been demonstrated in a novel system, it may be straightforward to scale these building blocks to the scale of systems competitive with the largest CMOS implementations.

State-of-the-art neural networks that support object and speech recognition can have tens of millions of synapses and networks with thousands of inputs and thousands of outputs. Simple street-scene recognition needed for autonomous vehicles require hundreds of thousands of synapses and tens of thousands of neurons. The largest networks that have been published—using over a billion synapses and a million neurons—have been used for face detection and object recognition in large video databases.

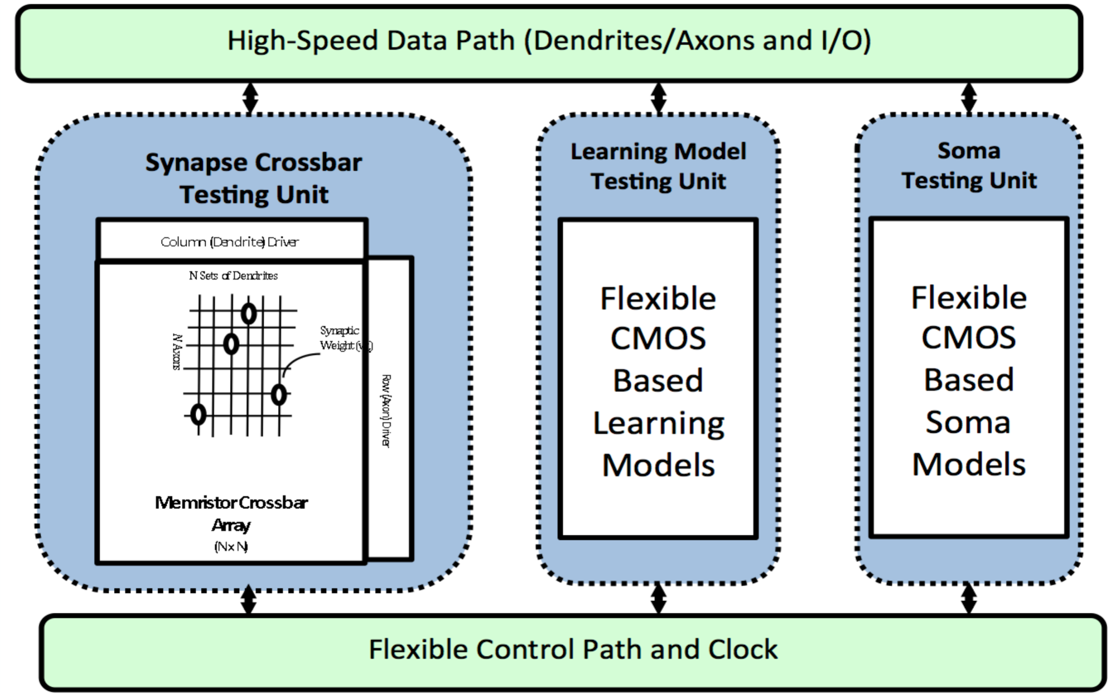

High-Speed Data Path (Dendrites/Axons and I/O)
Synapse Crossbar Testing Unit
Column (Dendrite) Driver
N Sets of Dendrites
N Axons
Synaptic Weight (wij)
Row (Axon) Driver
Memristor Crossbar Array
(N x N)
Learning Model Testing Unit
Flexible CMOS Based Learning Models
Soma Testing Unit
Flexible CMOS Based Soma Models
Flexible Control Path and Clock

Figure 6. Block diagram of a hybrid neuromorphic processor for synapse materials testing. The idea is that novel materials could be tested in a “harness” that uses existing CMOS implementations of learning and soma. A framework such as this could be used to accelerate testing of materials at some modest scale.

Properties

Development of neuromorphic computers, materials and/or devices are needed that exhibit some (or many) of the following properties:

1. ***Multistate behavior***, in which a physical property may have different values for the same control parameters, depending on past history.
2. ***Sensitivity*** to external stimuli such as current, voltage, light, H field, temperature or pressure to provide desirable functionalities.

3. ***Threshold behavior***, in which the material may drastically change its properties after repetitive application of the same stimulus.
4. ***Fault tolerance,*** so that the responses are repeatable in a statistical sense, but not necessarily with extremely high reliability.
5. ***Nonvolatility***, such that the properties are maintained for a long time without the need for refreshing processes or energy dissipation to hold state.
6. ***Temperature window,*** in which the properties can be controlled and maintained.
7. ***Insensitivity to noise,*** through phenomena such as stochastic resonances caused by inherent nonlinearities in the material. Counter intuitively; the addition of noise to a periodic signal may enhance its intensity.
8. ***Low energy***, in which switching from one state to another is obtained and maintained with low dissipation without the need for energy-costly refreshing.
9. ***Compatibility*** with other materials already in use in these types of systems.

In order to make this program a reality, several material-specific needs must be met. In general, the types of material systems that have been investigated in this context include strongly correlated oxides, phase change materials, metal inclusions in insulators, spin torque devices, ionic liquid-solid interfaces, and magnetic nanostructures. In addition, many of the complex materials are close to some type of electronic and/or structural instability. The use of these materials for neuromorphic applications requires extensive knowledge of their behavior under highly nonlinear, nonequilibrium conditions in heterogeneous structures at the appropriate nanometer scale, presenting an ambitious materials/condensed matter challenge. Because of the broad range of materials and systems, this cannot be cast as a single, universal aim.

Specifically, researchers must quantitatively address issues related to synthesis, characterization (e.g., static, dynamic, and in operando), measurements at short time scales, interactions with different types of electromagnetic radiation, and nanoscale inhomogeneities and defects. The problem is complex enough that it requires the attention of multiple investigators with different expertise and techniques. Much of this work can be done in small-scale laboratories such as at universities; however, certain resources are available and accessible only at major facilities such as national laboratories.

Devices

The main basis for the current digital computers is a three-terminal device that has gain: the transistor. In this device, the drain current is controlled by the application of a gate voltage, as shown in Figure 7. For a fixed gate voltage, the I-V characteristic is reversible. The control is provided by the changes in the output current as a function of gate voltage. In this case, for fixed parameters the output is always the same. Typically, these devices are built from ultra-pure semiconductors (e.g., Si, GaAs) where extreme control has to be exercised to minimize the defect density.

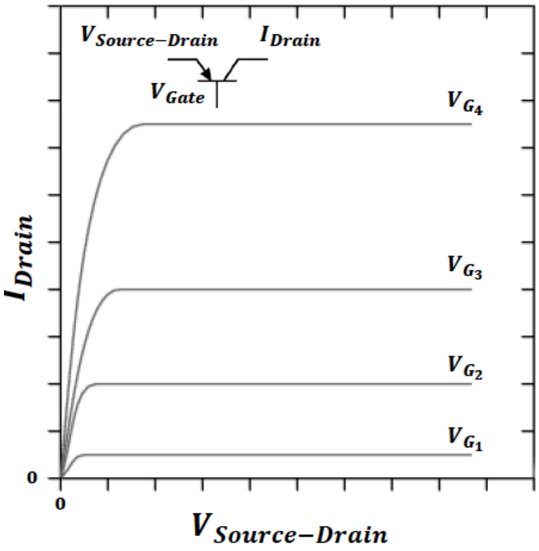


Figure 7. I-V characteristic of a transistor, the basis of von Neumann architecture. This I-V is controlled by the voltage applied to the gate.

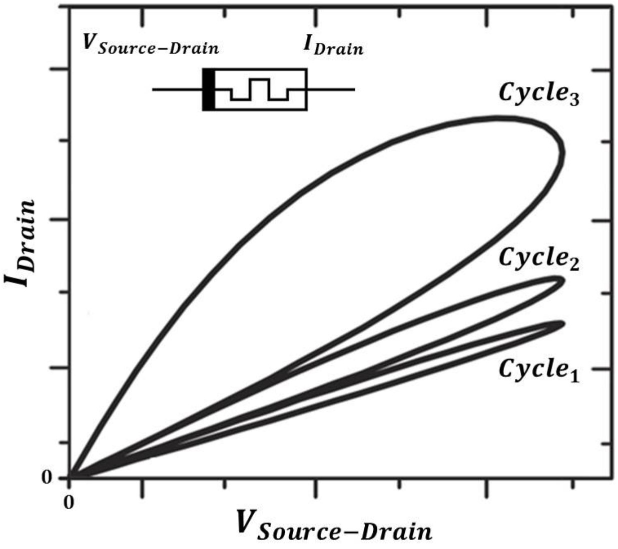


Figure 8. I-V characteristic of a memristor, the basis of a possible neuromorphic architecture. This I-V is controlled by the history of the device.

One possible implementation of a neuromorphic synapse is a hysteretic, two-terminal device: the memristor (see Figure 8). Typically, these types of devices exhibit a pinched (at zero voltage) I-V characteristic that is hysteretic. The type of hysteresis depends on the particular device implementation and material. An experimentally, easily accessible memristor material can be built from a strongly correlated oxide such as TiO_2. The behavior of a memristor has been well established, although the detailed mechanism and how to modify it are still being investigated. This type of I-V characteristics lends itself to plasticity, namely, changing properties depending on the past history.

It is expected that data transfer will be considerably reduced in the type of new neuromorphic architectures, with the consequent energetic savings. However, this cannot be completely eliminated as exemplified by the preponderance of connections through dendrites and axons in the brain. As a consequence, the interesting issue arises whether it would be possible to replace some of the electrically based data transfer to more energy efficient optical communications using reconfigurable optical elements and antennas. This may even be the basis of the development of a largely optically based data processor.

Materials

The development of neuromorphic devices also requires the use of strongly correlated, possibly complex, heterogeneous materials and heterostructures that are locally active so that they can be used to create an action potential. In *many* systems that are used as a basis for unconventional neuromorphic computing, the ultimate aim is to control the dynamical resistance (i.e., I-V characteristic) of a material or device. The control parameters can be categorized generally into two major classes: electronically driven and defect driven.

Within each class, many materials systems and devices have been investigated in different contexts and depth. A (probably) non-exhaustive list is given in Tables 5 and 6 in the Appendix.

Electronically driven materials rely on changes of a physical property due to variation in some control parameter such as temperature, current, voltage, magnetic field, or electromagnetic irradiation. Generally, they become useful if they exhibit negative differential resistance or capacitance and/or "threshold switching." While for some systems the basic physics is well understood, for others, even the underlying phenomena are controversial. The types of materials that are actively being investigated and the fundamental issues that are being addressed are related to strong electronic correlations, phase transitions, charge, and spin propagation. In addition, efforts are underway to identify applications in the areas of spintronics, ferroelectric storage, and sensors.

Electronically driven transitions include spin torque, ferroelectric, phase change, and metal-insulator transition based devices. The detailed physics of these systems differs significantly depending on the phenomena and the material system. In general, however, a scientific bottleneck exists because the detailed atomic-scale dynamics of these systems is still largely unknown.

Defect driven materials take advantage of some of the fundamental properties, which are strongly affected by the presence of inhomogeneities. These include the formation of metallic filaments, inhomogeneous stress, and uniform and nonuniform oxygen diffusion. Strong debates arise concerning how well these phenomena can be controlled. Nevertheless, a number of experiments have proven that the inhomogeneities can be controlled by control parameters such as voltage, temperature, or electric fields. Typical systems of this type are oxygen ionic diffusion in oxides, formation of metallic filaments, and ionic front motion across metal-insulator-metal trilayers. Moreover, the endurance of these types of devices has been shown to exceed one trillion cycles—not as much as DRAM or SRAM—although research is expected to significantly increase this.

Optically controlled materials and devices had a much more limited use in this context. This probably is because the eventual system that will emerge will likely be mostly electronic. On the other hand, the development of tunable optical elements may add an additional functionality to materials, which may be useful partially in this context.

Fault tolerance is one of the principal requirements of the materials to be used. Defects and their effect play a crucial role in the behavior of these materials especially when incorporated into devices. This has not been the case with the materials that have been used until now in silicon-based technology, which strives for ever higher perfection Thus, the new, fault-tolerant materials may actually improve device performance.

OPEN ISSUES

The most important general issue that needs extensive research and is not clearly defined is how to integrate individual devices into a working (although limited) system ("accelerator") that will serve as a proof of concept. Moreover, this system should be potentially scalable, although the exact way to do this may not be known at present time. Below we list some of the open issues that arise when considering materials/devices and systems almost independently.

Materials/Devices

Many open issues and questions remain regarding properties of materials and devices used and proposed for neuromorphic implementations. Because it is practically impossible to produce a comprehensive literature review in this brief report, we list here a few striking examples.

Resistive switching. Some memristor devices use as a basis the metal-insulator transition of simple transition-metal oxides. The switching mechanism in these is based on a first-order electronic transition that is generally coincident with a symmetry-changing structural phase transition. Resistive devices based on these materials, so-called locally active memristors, exhibit unusual hysteretic I-V characteristics of different types. They can show oscillations in the presence of a DC bias, can inject previously stored energy into a circuit to provide power amplification or voltage amplification for electrical pulses or spikes, and can exhibit chaos under controlled conditions. The physics of these materials is still being investigated, although the properties can be controlled well. For these types of strongly correlated materials, researchers continue to debate intensely regarding the role that Mott, Peierls, Hubbard, or Slater mechanisms play in the transition, the relevant time scales, the correlation of structural and resistive transitions, the effect of proximity, and the way these effects may be controlled (e.g., by epitaxial clamping). These issues are related to the specific properties of the materials in questions and are being investigated in many other contexts. Nevertheless, their properties can be controlled sufficiently well that a number of neuromorphic devices have been implemented.

Filament formation. Many of the neuromorphic devices utilized as synaptic memory rely on the formation of filaments or conductive channels in the material between two metallic electrodes. These may occur because of an intrinsic phase transition present, such as amorphous-crystalline or metal-insulator transitions, or because of redistribution of defects/ions that modulates local electrical properties. Understanding the mechanisms in the formation and destruction of filaments and the effect of preexisting defects is crucial for understanding the reproducibility of these devices. Particularly important is the role of pre-existing defects and the way these modify the formation of filaments.

Spin torque switching. In spin torque devices, the resistive transition is controlled by the magnitude of a current through (in principle) a four-layer device. Spin torque devices already have been implemented for unrelated spintronics applications and are being incorporated into commercial nonvolatile magnetic random access memory. The magnitude of the spin torque effect and the role the Oersted field are subjects of important research, especially when the size of the devices is reduced to the nanoscale or when devices are closely packed, as needed here. Understanding defects and their effect on the stability of ferromagnetic materials is important in order to improve the endurance of materials.

***Ionic diffusion**.* Defect-induced devices and materials rely on the controlled formation of filaments or the diffusion of ionic fronts between two metallic electrodes to control the conductance and thus provide large resistance switching. Examples are the formation of metallic shorts across metal-insulator-metal trilayers, the change in the resistance of an insulator due to an ionic front diffusion between two metallic electrodes, and the diffusion of oxygen induced by a metallic tip from an ionic liquid into an oxide. For these types of defect-based devices and phenomena, basic research is needed on several major issues: the importance of preexisting defects on the diffusion of light elements, thermophoresis, the formation of filaments, the reproducibility from cycle to cycle, and electromigration and the consequent effects on endurance.

***Nano and mesoscale**.* A number of open issues straddle both the materials and devices contemplated here. In particular, incorporating these materials into computational systems will require reducing them to nanometer scale in functional structures. Therefore, understanding the materials and especially their interface behavior at the nano and mesoscale is crucial. While at large-scale these phenomena may be well understood, when reduced to the nanoscale, additional effects become important. For instance, Oersted fields and dipolar interactions may become more important than at the micron scale in spin torque devices; filament formation may become highly inhomogeneous and uncontrolled in defect-based devices; and the importance and magnitude of enhanced fields in connection with roughness become especially important in nanoscale devices based on ionic conduction. A key issue in scaling down to the nanoscale is the fact that while the device heat capacity decreases, its thermal resistance increases, which can lead to huge Joule heating-induced temperature increases and enormous temperature gradients when electrically biased. This issue requires detailed studies of materials properties, ideally *in operando*. In many cases, complex interactions will appear at different nano and meso scales which can only be solved by the experimental capabilities available at DOE facilities. This will also necessarily include some capabilities for fabricating test structures using what is traditionally called back-end-of-line (BEOL) processing techniques (i.e., not using full CMOS capabilities). The size scales and the reproducibility required for meaningful analysis will mean that some lithographic capabilities at the 50 nm and smaller scales will be required, which can be obtained by electron beam lithography or refurbished and therefore relatively inexpensive UV steppers.

System

As we consider the building of large-scale systems from neuron like building blocks, there are a large number of challenges that must be overcome. One challenge arises from the need for dense packaging of neurons in order to achieve comparable volumes to brains. This implies dense 3D packing with a range of problems associated with assembly, power delivery, heat removal and topology control. Another set of challenges arises from the abstract nature of neuro-inspired computation itself. How close to nature must we build to gain the benefits that evolution has devised? Can we develop computational abstractions that have many of the advantages of biology but are easier to construct with non-biological materials and non-biological assembly processes? How will such systems be designed? How will they be programmed and how will they interact with the vast computational infrastructure that is based on conventional technologies?

A number of critical issues remain as we consider the artificial physical implementation of a system that partially resembles a brain-like architecture:

1. What are the minimal physical elements needed for a working artificial structure: dendrite, soma, axon, and synapse?
2. What are the minimal characteristics of each one of these elements needed in order to have a first proven system?
3. What are the essential conceptual ideas needed to implement a minimal system: spike-dependent plasticity, learning, reconfigurability, criticality, short- and long-term memory, fault tolerance, co-location of memory and processing, distributed processing, large fan-in/fan-out, dimensionality? Can we organize these in order of importance?
4. What are the advantages and disadvantages of a chemical vs. a solid-state implementation?
5. What features must neuromorphic architecture have to support critical testing of new materials and building block implementations?
6. What intermediate applications would best be used to prove the concept?

These and certainly additional questions should be part of a coherent approach to investigating the development of neuromorphic computing systems. The field could also use a comprehensive review of what has been achieved already in the exploration of novel materials, as there are a number of excellent groups that are pursuing new materials and new device architectures. Many of these activities could benefit from a framework that can be evaluated on simple applications.

At the same time, there is a considerable gap in our understanding of what it will take to implement state-of-the-art applications on neuromorphic hardware in general. To date, most hardware implementations have been rather specialized to specific problems and current practice largely uses conventional hardware for the execution of deep learning applications and large-scale parallel clusters with accelerators for the development and training of deep neural networks. Moving neuromorphic hardware out of the research phase into applications and end use would be helpful. This would require advances which support training of the device itself and to show performance above that of artificial neural networks already implemented in conventional hardware. These improvements are necessary both regarding power efficiency and ultimate performance.

INTERMEDIATE STEPS

This section identifies the major milestones needed toward the development of a neuromorphic computer. We should highlight that every step must be based on earlier steps and connected to eventual implementation of next steps. This can be considerably advanced through the construction of appropriate compact theoretical models and numerical simulations that are calibrated through experimentation. It is also important to point out that this field is in its earlier stages of development and therefore sufficient flexibility should be maintained at every stage. This should not be viewed as a well-defined development task but as a research project. Therefore, it is important that at every stage several competing projects are implemented to allow for the best solution to emerge. The key ingredients in these intermediate steps could be:

General Aim. As a general goal, it would be desirable to develop well-defined intermediate application such as needed in the fields of vision, speech, and object recognition to prove the reality of a program as described here.

Simulations. There are opportunities to leverage large-scale computing in the development of simulators for neuromorphic designs and to develop a deep understanding of materials and device. These simulations could be used to refine architectural concepts, improve performance parameters for materials and devices, and to generate test data and signals to help support accelerated testing as new materials, devices and prototypes are developed.

Devices. Development and engineering of novel devices perhaps based on some type of memristive or optically bistable property is needed. This should include incorporation into well-defined systems and be based on well-understood materials science.

Material Science. Synthesis, characterization and study of new functional, tunable materials with enhanced properties are needed to integrate into novel neuromorphic devices.

We envision the following stages in the development of such a project:

1. Identify conceptual design of neuromorphic architectures
2. Identify devices needed to implement neuromorphic computing
3. Define properties needed for prototype constituent devices
4. Define materials properties needed
5. Identify major materials classes that satisfy needed properties

6. Develop a deep understanding of the quantum materials used in these applications
7. Build and test devices (e.g., synapse, soma, axon, dendrite)
8. Define and implement small systems, and to the extent possible, integrate/demonstrate with appropriate research and development results in programming languages, development and programming environments, compilers, libraries, runtime systems, networking, data repositories, von Neumann-neuromorphic computing interfaces, etc.
9. Identify possible "accelerator" needs for intermediate steps in neuromorphic computing (e.g., vision, sensing, data mining, event detection)
10. Integrate small systems for intermediate accelerators
11. Integrate promising devices into end-to-end system experimental chips (order 10 neurons, 100 synapses)
12. Scale promising end-to-end experiments to demonstration scale chips (order 100 neurons and 10,000 synapses)
13. Scale successful demonstration chips to system scale implementations (order millions of neurons and billion synapses)
14. Scale successful demonstration chips to system scale implementations (order millions of neurons and billion synapses)

We have outlined in this report many of the open issues and opportunities for architecture and materials science research needed to realize the vision of neuromorphic computing. The key idea is that by adopting ideas from biology and by leveraging novel materials, we can build systems that can learn from examples, process large-scale data adjust their behavior to new inputs and do all of these with the power efficiency of the brain. Taking steps in this direction will continue the development of data processing in support of science and society.

CONCLUSIONS

The main conclusions of the roundtable were:

1. Creating the architectural design for neuromorphic computing requires an integrative, interdisciplinary approach between computer scientists, engineers, physicists, and materials scientists
2. Creating a new computational system will require developing new system architectures to accommodate all needed functionalities
3. One or more reference architectures should be used to enable comparisons of alternative devices and materials
4. The basis for the devices to be used in these new computational systems require the development of novel nano and meso structured materials; this will be

accomplished by unlocking the properties of quantum materials based on new materials physics

5. The most promising materials require fundamental understanding of strongly correlated materials, understanding formation and migration of ions, defects and clusters, developing novel spin based devices, and/or discovering new quantum functional materials
6. To fully realize open opportunities requires designing systems and materials that exhibit self- and external-healing, three-dimensional reconstruction, distributed power delivery, fault tolerance, co-location of memory and processors, multistate—i.e., systems in which the present response depends on past history and multiple interacting state variables that define the present state
7. The development of a new brain-like computational system will not evolve in a single step; it is important to implement well-defined intermediate steps that give useful scientific and technological information

Acknowledgements

We thank all participants of the workshop for their valuable input under much time constraints. We thank Juan Trastoy for critical reading and input for Table 5. We thank Emily Dietrich for editorial support.

APPENDICES

Acronyms

3D: Three-dimensional

AER: Address Event Representation

ALU: Arithmetic/Logic Unit

BEOL: Back-end-of-line (BEOL)

CMOS: Complementary Metal Oxide Semiconductor

CNN: Convolutional Neural Network

CPU: Central Processing Unit

ColdRAM: Memory organized for infrequent access patterns

DNN: Deep Neural Network

DMA: Logic for supporting Direct Memory Access

HotRAM: Memory organized for frequent access patterns

InstRAM: Instruction Memory

LIF: Leaky Integrate and Fire (neuron)

LTPS: Long-Term Plasticity Synapses

MLU: Machine Learning Unit, functional units optimized for ML operations

MU: Memory Unit

OutputRAM: Memory for holding output of operations

STDP: Spike Time-Dependent Plasticity

STPS: Short-Term Plasticity Synapses

VLSI: Very Large-Scale Integration

Glossary

Axon: Transmitting signals to other neurons, axons provide long-distance output connection between a presynaptic to a postsynaptic neuron

Charge trapping: Mobile electrons maybe trapped sometimes in defects invariably present in materials

Crystalline-amorphous transition: Some materials change their physical structure from an ordered crystalline to a completely disordered (amorphous) phase; the electronic properties of these two phases maybe radically different

Deep learning: Deep learning is the branch of machine learning that builds tools based on deep (multilayer) neural networks

Dendrite: Providing multiple, distributed inputs into the neurons, the role of dendrites is commonly believed simply to provide signals from multiple neurons into a single neuron

Depress: To increase resistance

Domain wall motion: In some cases, a magnetic material reverts its magnetization by the motion of a separation ("wall") in between two well-defined magnetization areas

Filament formation: The resistance of two electrodes separated by an insulator may change drastically if a conducting filament forms; this can form because of intrinsic reasons or due to the motion of atoms in the insulator

Hebbian learning: Change occurs in the synapse resistance after repeated spiking by the presynaptic neuron before the postsynaptic neuron; this is also sometimes referred to as spike time-dependent plasticity (STDP)

Ionic motion: Certain solids and liquids ions may move under different driving forces such as high voltages, currents and temperature

Learning: Conditioning; process by which the various elements of the system change and readjust depending on the type of stimuli they receive

Machine learning: The branch of computer science that deals with building systems that can learn from and make predictions on data

Magnetic tunnel junction: A device based on quantum mechanical tunneling that changes it resistance depending on the relative magnetization of the magnetic electrodes; the basis for most spintronics applications

Parallel computing: Differing from serial von Neumann, parallel computing is distinguished by the kind of interconnection between processors and between processors and memory; most parallel computers today are large-ensembles of von Neumann processors and memory

Plasticity: Synaptic resistance; key property that allows the complex neuromorphic circuits to be modified ("learn") as they are exposed to different signals

Potentiation: Increased in conductance

Soma: Neuron body, where primary input integration and firing occurs

Synapse: Space between axon and dendrites that allows for signals to be transmitted from the presynaptic to the postsynaptic neuron.

Vacancy motion: In certain solids and liquids, the absence of an ion ("vacancy") may move under different driving forces such as high voltages, currents and temperature

Tables

Table 2 References (Table 2 above)

1. Dean, Mark E., Catherine D. Schuman, and J. Douglas Birdwell. "Dynamic adaptive neural network array." *Unconventional Computation and Natural Computation*. Springer International Publishing, 2014. 129-141.
2. Merolla, Paul A., et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface." *Science* 345.6197 (2014): 668-673.
3. Benjamin, Ben Varkey, et al. "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations." *Proceedings of the IEEE* 102.5 (2014): 699-716.
4. Brüderle, Daniel, et al. "A comprehensive workflow for general-purpose neural modeling with highly configurable neuromorphic hardware systems." *Biological cybernetics* 104.4-5 (2011): 263-296.
5. Furber, Steve B., et al. "Overview of the spinnaker system architecture." *Computers, IEEE Transactions on* 62.12 (2013): 2454-2467.
6. Shen JunCheng, et. al. "Darwin: a Neuromorphic Hardware Co-Processor based on Spiking Neural Networks". *SCIENCE CHINA Information Sciences*, DOI: 10.1360/112010-977

Neural networks digital hardware implementations

Name	Architecture	Learn	Precision	Neurons	Synapses	Speed
Slice architectures						
Micro devices MD-1220[a]	Feedforward, ML	No	1 x 16 bits	8	8	1.9 MCPS
NeuraLogix NLX-420[a]	Feedforward, ML	No	1-16 bits	16	Off-chip	300 CPS
Philips Lneuro-1	Feedforward, ML	No	1-16 bits	16 PE	64	26 MCPS
Philips Lneuros-2.3	N.A.	No	16-32 bits	12 PE	N.A.	720 MCPS
SIMD						
Inova N64000[a]	GP, SIMD, Int	Program	1-16 bits	64 PE	256k	870 MCPS 220 MCUPS
Hecht-Nielson HNC 100-NAP[b]	GP, SIMD, FP	Program	32 bits	4 PE	512k Off-chip	250 MCPS 64 MCUPS
Hitachi WSI	Wafer, SIMD	Hopfield	9 x 8 bits	576	32k	138 MCPS
Hitachi WSI	Wafer, SIMD	BP	9 x 8 bits	144	N.A.	300 MCUPS
Neuricam NC3001 TOTEM	Feedforward, ML, SIMD	No	32 bits	1-32	32k	1 GCPS
Neuricam NC3003 TOTEM	Feedforward, ML, SIMD	No	32 bits	1-32	64k	750 MCPS
RC Module NM6403	Feedforward, ML	Program	1-64 x 1-64 bits	1-64	1-64	1200 MCPS
Systolic array						
Siemans MA-16	Matrix ops	No	16 bits	16 PE	16 x 16	400 MCPS
Radial basis functions						
Nestors/Intel NI1000[c]	RBF	RCE, PNN, program	5 bits	1 PE	245 x 1024	40 kpat/s
IBM ZISC036	RBF	ROI	8 bits	36	64 x 36	250 kpat/s
Silicon recognition ZISC78	RBF	KNN, L1. LSUP	N.A.	78	N.A.	1 Mpat/s
Other chips						
SAND/1	Feedforward, ML, RBF	No	40 bits	4 PE	Off-chip	200 MCPS
MCE MT 19003	Kohonen Feedforward, ML	No	13 bits	8	Off-chip	32 MCPS

Table 3. Historical hardware implementations of neural networks.

	Description
Clock Free	Fully asynchronous
Scale Free	Activity can vary from local to system level scales depending upon context
Symbol Free	No single neuron or synapse represents any single item/concept
Grid Free	Small world network geometry allows feature integration from heterogeneous and non local brain areas
Dendritic Neuron	Nonlinear signal processing via dendrites in each neuron
Synaptic Plasticity	Most synapses exhibit plasticity at various time scales (secs to hrs)
Synaptic Path Length	Approx. constant number of hops between different brain areas
Dense Connectivity	Each neuron connects to between 1000-10000 other neurons
Modular Cortex	Six layered modular architecture that repeats across architecture
Broadcasting	Brain areas that broadcast signals (neuromodulatory) to all other parts

Table 4. Neuromorphic system level architecture features.

CATEGORY	SYSTEM	MECHANISM	WRITE SPEED	ON / OFF	DATA RETENTION (@ 1 Hz)	TEMP (ºC)	POWER (W)	ISSUES	REFERENCE
Oxides	$HfO_X/TiO_X/HfO_X/TiO_X$	Vom-front	10s ns	1000	> 15 min	20-85	10^{-4}	Abrupt SET process (only depression is possible)	http://onlinelibrary.wiley.com/doi/10.1002/adma.201203680/abstract
	VO_2	Ff	1s	100		70	10^{-2}	Temperature, memory duration, 50 V pulses, only potentiation	http://scitation.aip.org/content/aip/journal/apl/95/4/10.1063/1.3187531
	Nb_2O_5/Pt	Vom	100s ns	10	> 500 years	RT	10^{-4}	Simple planar micron scale structure	http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1425686
	WO_X	Vom-front	100s μs	1.4	> 3 hours	RT	10^{-5}	Low ON/OFF ratio	http://dx.doi.org/10.1109/TED.2014.2319814
	Nb-doped-a-STO	Vom-filament	10s μs	1000	> 1 day	27-125	10^{-4}	Electroforming needed	http://onlinelibrary.wiley.com/doi/10.1002/adfm.201501019/abstract
	$Pt/TiO_2/Pt$	Vom	10 ns	10	> 35,000 years	RT	$<10^{-4}$	30 nm wire width crossbar structure	http://onlinelibrary.wiley.com/doi/10.1002/adfm.201202170/abstract
Phase Change	GeSbTe	CAt	1 ns	100	> 3 hours	RT	10^{-3}		https://www.sciencemag.org/content/336/6088/1566.full
Optical	C-Nanotubes	Photo/Electrical gating	10s	200	> 2 days	RT	10^{-6}	Very slow and difficult to implement	http://onlinelibrary.wiley.com/doi/10.1002/adma.200902170/full
	GaLaSO	Photodark	Ms	1.1		RT	10^{-1}	Proof of concept	http://onlinelibrary.wiley.com/doi/10.1002/adom.201400472/abstract
Metal Inclusions	Ag on a-Si	Ff	10s ns	1000	11 days	RT	10^{-8}	Electroforming, short endurance	http://pubs.acs.org/doi/abs/10.1021/nl073225h
	Cosputtered a-Si and Ag	Iom-front	100s μs	8	5 years	RT	10^{-6}	Low ON/OFF ratio	http://pubs.acs.org/doi/abs/10.1021/nl904092h
Organic	Au/Pentacene/Si NWs/Si	Ct		120		RT	10^{-5}	Speed is not clear	http://scitation.aip.org/content/aip/journal/apl/104/5/10.1063/1.4863830
Ferro electric	BTO/LSMO Tunneling	Ps	10s ns	300	> 15 min	RT	10^{-6}		http://www.nature.com/nmat/journal/v11/n10/full/nmat3415.html
Magnetic	MgO-based MTJ	DWm		1.1		RT	10^{-4}	Needs external magnetic field, Low ON/OFF ratio	http://www.nature.com/nphys/journal/v7/n8/full/nphys1968.html
Liquid-Solid	Ionic liquid/$SmNiO_3$	Iom	10s ms	11		35-160		Requires gating circuit, slow	https://doi.org/10.1038/ncomms3676

Table 5. List of materials systems for neuromorphic applications. Characteristics obtained from the literature of the different material systems put forward for neuromorphic applications. The following abbreviations are used: Vom=Vacancy motion; Ff=Filament formation; Iom=ionic motion; DWm=domain wall motion; Cat=crystalline amorphous transition; Ct=Charge trapping; Ps=Polarization switching; RT=room temperature; SET=change from high to low resistance state, and MTJ=magnetic tunnel junction.

Summary of common inorganic storage media and corresponding switching characteristics

Storage medium	Switching mode	ON/OFF ration	Operation speed	Endurance (cycles)
Binary oxides				
MgO_x	Unipolar, bipolar	$>10^{5}$	-	$>4 \times 10^{2}$
AlO_x	Unipolar, bipolar	$>10^{6}$	<10 ns; <10 ns	$>10^{4}$
SiO_x	Unipolar, bipolar	$>10^{7}$	<100 ps; <100 ps	$>10^{8}$
TIO_x	Unipolar, bipolar	$>10^{5}$	<5 ns; <5 ns	$>2 \times 10^{6}$
CrOx	Bipolar	$>10^{2}$	<4 µs; <5 µs	$>6 \times 10^{4}$
MnO_x	Unipolar, bipolar	$>10^{4}$	<100 ns; <200 ns	$>10^{5}$
FeO_x	Bipolar	$>10^{2}$	<10 ns; <10 ns	$>6 \times 10^{4}$
CoO_x	Unipolar, bipolar	$>5 \times 10^{3}$	<20 ns; <20 ns	$>10^{3}$
NiO_x	Unipolar, bipolar	$>10^{6}$	<10 ns; <20 ns	$>10^{6}$
CuO_x	Unipolar, bipolar	$>10^{5}$	<50 ns; <50 ns	$>1.2 \times 10^{4}$
ZnO_x	Unipolar, bipolar	$>10^{7}$	<5 ns: <5 ns	$>10^{6}$
GaO_x	Bipolar	$>10^{2}$	<400 ns; <600 ns	$>10^{4}$
GeO_x	Unipolar, bipolar	$>10^{9}$	<20 ns; <20 ns	$>10^{6}$
ZrO_x	Unipolar, bipolar	$>10^{6}$	<10 ns; <10 ns	$>10^{4}$
NbO_x	Unipolar, bipolar	$>10^{8}$	<100 ns; <100 ns	$>10^{7}$
MoO_x	Unipolar, bipolar	>10	<1 µs; <1 µs	$>10^{6}$
HfO_x	Unipolar, bipolar	$>10^{5}$	<300 ps; <300 ps	$>10^{10}$
TaO_x	Unipolar, bipolar	$>10^{9}$	<105 ps ; <120 ps	$>10^{12}$
WO_x	Unipolar, bipolar	$>10^{4}$	<300 ns; <50 ns	$>10^{8}$
CeO_x	Unipolar, bipolar	$>10^{5}$	<1 µs : <200 ns	$>10^{4}$
GdO_x	Unipolar, bipolar	$>5 \times 10^{5}$	<1 ns; <1 ns	$>10^{7}$
YbO_x	Unipolar, bipolar	$>10^{5}$	-	$>10^{5}$
LuO_x	Unipolar, bipolar	$>10^{4}$	<10 ns; <30 ns	$>8 \times 10^{2}$
Ternary and more complex oxides				
$LaAlO_3$	Bipolar	$>10^{4}$	-	$>10^{2}$
$SrTiO_3$	Bipolar	$>10^{5}$	<5 ns: <5 ns	$>10^{6}$
$BaTiO_3$	Unipolar, bipolar	$>10^{4}$	<10 ns; <70 ns	$>10^{5}$
LC(or S)MO	Bipolar	$>10^{3}$	<25 ns; <25 ns	$>10^{3}$
PCMO	Bipolar	$>10^{3}$	<8 ns; <8 ns	$>10^{10}$
$BiFeO_3$	Unipolar, bipolar	$>10^{5}$	<50 ns; <100 µs	$>10^{3}$
Chalcogenides				
Cu_2S	Bipolar	$>10^{6}$	<100 µs; <100 µs	$>10^{5}$
GeS_x	Bipolar	$>10^{5}$	<50 ns; <50 ns	$>7.5 \times 10^{6}$
Ag_2S	Bipolar	$>10^{6}$	<200 ns; <200 ns	-
Ge_xSe_y	Bipolar	$>10^{6}$	<100 ns; <100 ns	$>3.2 \times 10^{10}$
Nitrides				
AlN	Unipolar, bipolar	$>10^{3}$	<10 ns; <10 ns	$>10^{8}$
SiN	Unipolar, bipolar	$>10^{7}$	<100 ns; <100 ns	$>10^{9}$
Others				
a-C	Unipolar, bipolar	$>3 \times 10^{2}$	<50 ns; <10 ns	$>10^{3}$
a-Si	Bipolar	$>10^{7}$	<5 ns; <10 ns	$>10^{8}$
AgI	Bipolar	$>10^{6}$	<50 ns; <150 ns	$>4 \times 10^{5}$

Table 6. Comprehensive list of relevant properties for interesting materials. The operation speed is written as ‘set (write) speed; reset (erase) speed’. The symbol ‘-‘ means that no data concerning that characteristic is found. (after *F. Pan, S. Gao, C. Chen, C. Song, F. Zeng, Materials Science and Engineering R 83 (2014) 1–59.*)

WORKSHOP LOGISTICS

Roundtable Participants

Co-chairs:

Ivan K. Schuller	University of California, San Diego
Rick Stevens	Argonne National Laboratory and University of Chicago

Participants and Observers:

Nathan Baker	Pacific Northwest National Laboratory
Jim Brase	Lawrence Livermore National Laboratory
Hans Christen	Oak Ridge National Laboratory
Mike Davies	Intel Corporation
Massimiliano Di Ventra	University of California, San Diego
Supratik Guha	Argonne National Laboratory
Helen Li	University of Pittsburgh
Wei Lu	University of Michigan
Robert Lucas	University of Southern California
Matt Marinella	Sandia National Laboratories
Jeff Neaton	Lawrence Berkeley National Laboratory
Stuart Parkin	IBM Research – Almaden
Thomas Potok	Oak Ridge National Laboratory
John Sarrao	Los Alamos National Laboratory
Katie Schuman	Oak Ridge National Laboratory
Narayan Srinivasa	HRL Laboratories LLC
Stan Williams	Hewlett-Packard
Philip Wong	Stanford University

Roundtable Summary

Roundtable on Neuromorphic Computing:
From Materials to Systems Architecture

Co-chairs:

Ivan K. Schuller	University of California, San Diego
Rick Stevens	Argonne National Laboratory and University of Chicago

DOE Contacts:

Robinson Pino	Advanced Scientific Computing Research
Michael Pechan	Basic Energy Sciences

Purpose

The Neuromorphic Computing: From Materials to Systems Architecture Study Group convened national laboratory, university, and industry experts to explore the status of the field and present future research opportunities involving research challenges from materials to computing, including materials science showstoppers and scientific opportunities. The output goal of the roundtable is a symbiotic report between systems, devices and materials that would inform future ASCR/BES research directions.

Logistics

Gaithersburg, MD, Montgomery Ballroom
Thursday, October 29, 2015 (5:00pm – 8:00pm)
Friday, October 30, 2015 (8:00am – 5:00pm)

Participants

Participation and observation, by invitation only, was approximately 20 external scientists (DOE laboratories, university and industry). Two co-chairs helped select participants and helped lead the discussion. Several—approximately 10—Federal Program Managers from DOE attended as observers. The total meeting size was approximately 30.

Agenda

The agenda comprised two days and included a keynote address, overview talks, discussion sessions, breakout sessions, and a closing summary.

Roundtable Report

A draft will be delivered by December 18, 2015.

Roundtable Agenda

Roundtable on Neuromorphic Computing:
From Materials to Systems Architecture

Co-chairs:

Ivan K. Schuller	University of California, San Diego
Rick Stevens	Argonne National Laboratory and University of Chicago

DOE Contacts:

Robinson Pino	Advanced Scientific Computing Research
Michael Pechan	Basic Energy Sciences

Agenda:

Thursday, October 29, 2015

5:00pm	Registration
6:00pm	Working Dinner with Keynote Speaker Stanley Williams
8:00pm	Adjourn

Friday, October 30, 2015

8:00am	Continental Breakfast and Registration
8:30am	Morning Session: Overview talks
10:30am	Break
10:45am	Guided Discussion Session
12:00pm	Working Lunch
1:00pm	Breakout Sessions
3:00pm	Reports-outs
4:00pm	Closing Summary
5:00pm	Adjourn

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

This page intentionally left blank