Neuromorphic computing is a critical future technology for the computing industry, but it has yet to achieve its promise and has struggled to establish a cohesive research community. A large part of the challenge is that full realization of the potential of brain inspiration requires advances in both device hardware, computing architectures, and algorithms. This simultaneous development across technology scales is unprecedented in the computing field. This article presents a strategy, framed by market and policy pressures, for moving past these current technological and cultural hurdles to realize its full impact across technology. Achieving the full potential of brain-derived algorithms as well as post-complementary metal–oxide-semiconductor (CMOS) scaling neuromorphic hardware requires appropriately balancing the near-term opportunities of deep learning applications with the long-term potential of less understood opportunities in neural computing.

1. Introduction

The core value proposition for brain-inspired computing can be stated as follows.

Leveraging knowledge of how the brain processes information will impact a wide range of science and technology applications.

The bulk of this article will focus on the first half of this statement, providing a strategic approach to overcome the challenge of getting insights of the brain into computing. There is no reason to expect that this will be easy, and we must resist the temptation to place too much emphasis on limited near-term solutions. But, advances in fields ranging from neuroscience to materials science provide encouragement that true brain-inspired computing is achievable with the right approach.

However, it is important to consider the second half of the statement as well. Many fields stand to benefit immensely if neural computing can achieve its potential. More brain-like computing can potentially revolutionize brain–machine interfaces, facilitating both neurological interventions and measurements. Many efforts in pursuing autonomy, from self-driving cars to advanced robotics, stand to benefit from more energy-efficient and effective brain-derived approaches. Also, this same energy efficiency from neuromorphic hardware has the potential to decrease the extreme power costs associated with large high-performance computing systems required for scientific computing applications for climate modeling to drug discovery. Arguably, the most immediate application driver for neural inspiration is artificial intelligence (AI), and this white paper will, thus, emphasize the hidden growing needs out of AI as the primary driver for this technology.

1.1. A Market Perspective on AI

The recent resurgence of AI has been dramatic and unprecedented. Core to the rise of AI has been the success of machine learning (ML), and particularly deep artificial neural networks (ANNs), which have shown considerable promise in tasks ranging from image classification to game playing to natural language processing.

ANNs had been a focus of research for many years, with only limited success. In hindsight, it is evident that in previous decades, ANN performance saturated below the thresholds required for real-world impact due to the considerable computational and data costs associated with ANNs. However, two parallel trends changed the economics of ANNs. First, Moore’s Law led to larger ANNs; second, the rise of the Internet and the associated proliferation of sensors and social networks provided a sufficient large volume of data necessary for training. Combined, these two trends shifted the economics of ANNs to a point where they are not only economically viable, but one of the dominant forms of ML used today (Figure 1).

While ANNs have been tremendously successful, there remain many domains where we require higher levels of performance from AI systems than available today. While there is hope that continued scaling of ANNs and similar algorithms will yield further improvements, it is prudent to consider whether these economic drivers will continue.

The risk to future computing scaling is more immediate, as Moore’s Law has been slowing by several metrics over the last decade (Figure 2, left). At the same time, AI algorithms,
including large ANNs, are increasing in computational cost at an exponential rate.\cite{6} These large computing costs are already being cited as an environmental concern,\cite{7} and there are increasing worries that the required resources represent a financial burden that is limiting the ability for the broader research community to compete with those with access to large compute.

Less apparent, a similar worry can be seen in the realm of data. While many domains are “drowning in data,” the irony is that in many cases, these data are ill-suited for use in current AI techniques. Most ANNs today require large volumes of expensive-labeled annotated data and generally demand that the training data be comprehensively reflective of data it will see in deployment. Furthermore, increasing data are neither uniformly available nor can they be expected to the representative of the real-world, or worse, captures undesirable elements of human behaviors (e.g., discrimination codified into models through biased training data). The growing value of data has led to debates as to whether data are the “new oil” and need to be regulated as a public resource.\cite{8,9} As shown on the right-hand side of Figure 2, it is reasonable to expect that the availability to high-quality data can increasingly become a significant barrier to enter in AI, with those with access to large proprietary data sources (e.g., social networks) having an intrinsic advantage in AI training compared with those who do not.

Coupled with these technological reasons for concern, both the power and data challenges have a political dimension as well. It is reasonable to expect that concern about climate will continue to grow, with increased attention to the carbon footprint of training AI algorithms risking putting a brake on the field. Likewise, skepticism about social media and personally identifiable information (PII) persistence in AI algorithms has led to movements to regulate data sharing and increased consumer privacy that may ultimately lead to government restrictions on data availability.\cite{10} For instance, the countries of Germany and Austria do not allow Google Maps Street View. Such changes may have considerable implications for future data-centric algorithm development.

2. A Role for Brain Inspiration for the Next Generation of AI

For these and other reasons, it is important for the broader computing community to begin to look toward where the next advances in AI will come from. There are, of course, many potential futures for AI; aside from a continuation of ANN dominance, we may see a renewal of attention on expert systems or statistical methods,\cite{11} or perhaps the rise of techniques such as causal graphs\cite{12} or Bayesian methods.\cite{13} This article is premised though on the perspective that the natural successor to build on the ANN’s successes while moving beyond its limitations is to pursue deeper brain inspiration within AI.\cite{14} The brain not only lends itself to a physical instantiation that can exist at very low power, but it operates with incredible data efficiency and is by definition capable of many of the human capabilities desired in AI applications.
Brain inspiration in AI is often cited as a leading driver for the recent rise in ANNs. However, the neuroscience that has had the most influence on AI is from over a half-century ago; the considerable advances in neuroscience over recent decades have only had a fringe influence on AI methods to date. This lack of direct influence of neuroscience in modern AI represents an opportunity. For both AI algorithms and future hardware, there are compelling reasons to strengthen the pipeline from neuroscience into AI.[14]

This article proposes a specific technological roadmap for achieving this end goal. The nature of neural computing is that advances are required for both algorithms and hardware, and the value of these advances is contingent on advances in the other. Arguably, this chicken-and-egg problem has limited the real-world utility of neural computing to date; however, if approached appropriately, the current rise and requirements of deep ANNs may be the catalyst required to advance the field.

The roadmap proposed here has the following five stages, which will be explored in more detail in subsequent sections: 1) design flexible low-power AI implementations using spiking neuromorphic hardware; 2) extend reach of neural hardware by designing efficient spiking neural algorithms for 1); 3) design hybrid analog-spiking architectures for biological-scale algorithms from 2); 4) achieve more powerful AI capabilities by implementing brain-derived neural algorithms on 3); and 5) look to exotic hardware approaches to achieve full neuromorphic capabilities.

This notional roadmap is shown in Figure 3. Today, spiking neuromorphic hardware is becoming a reality (e.g., TrueNorth, SpiNNaker, Loihi).[15–17] Importantly, while steps 2–4) are described sequentially, opportunities exist to start making initial progress on each of these so as to enable the most effective co-design between algorithms and hardware going forward. Finally, this roadmap emphasizes that the final step of developing long-term hardware solutions must be focused on addressing the challenges posed by long-term algorithm development as opposed to addressing today’s problems.

2.1. Current State of Neuromorphic Hardware Community

Brain-inspired hardware has been investigated to some extent since the dawn of computing, however, the concept of neuromorphic systems, where the hardware directly emulates some aspects of biological systems, started in the 1980s with circuits designed to use analog complementary metal–oxide-semiconductor (CMOS) circuits to emulate neuron dynamics.[18] Today, complementing the continued research into analog systems, which are primarily academic-based, there are a number of significant efforts focused on scalable spiking neuromorphic hardware, which primarily achieve advantages from the event-driven communication and parallel operation of neurons. Adding to the confusion, at times, the neuromorphic umbrella is often scoped to include accelerator architectures for neural networks, though this report will focus exclusively on the former more directly neuromorphic (analog and spiking) architectures.

At the time of writing, there is an interesting divergence of these approaches in terms of both technological maturity and perceived value. As spiking architectures can be designed using conventional CMOS devices and logic, they have been able to rapidly reach impressive scales at near-commercial reliability. The space of potential analog hardware is considerably larger, and while most researchers agree that, in principle, analog would confer some fundamental advantages unavailable to exclusively digital spiking systems, considerable research remains necessary to achieve a scalable, production-suitable analog system.

2.2. Spiking Neuromorphic Algorithms

We can define spiking neuromorphic hardware as any platform that seeks performance benefits primarily through event-driven communication and is implemented primarily using digital CMOS circuits. Currently, there are three systems that meet this criteria and have shown to be capable of existing at large scale (multi-chip platforms can reach over 10^6 neurons): the SpiNNaker

![Figure 3. Roadmap for true brain-derived computation. Co-design between algorithms and hardware is required to advance neuromorphic field, but as an emerging technology, it is critical to resist the temptation to pursue novel hardware or algorithms to solve problems already successfully addressed by existing technology and instead target future applications.](image-url)
platform from the University of Manchester,[17] the IBM TrueNorth chip,[14] and the Intel Loihi chip[15] (but see the University of Heidelberg BrainScales for a hybrid analog–digital approach[16]). These spiking neuromorphic platforms, and similar ones being developed, are attractive both due to their configurability (similar to field-programmable gate arrays [FPGAs]) and ability to operate with far lower power than conventional architectures that use equivalent fabrication methods.

While there has long been some research in computing with spiking neurons, the rapid success of achieving large-scale spiking neuromorphic systems has put the hardware technology ahead the required algorithm development. This has left the overall technology in a somewhat awkward place, and there are several notable challenges facing the spiking neuromorphic community (see Box 1). Of note is the relevance of spiking architectures for use in existing ML applications. While these architectures were generally not developed for ANNs, in the absence of other impactful applications, there is a growing urgency to show that these approaches are not only compatible but efficient at large-scale ANN tasks. Research into high-performing spiking neural networks (SNNs) has shown promise but has slowly developed.[20–22] Converting an ANN to an SNN is non-trivial, as the continuous-valued activation functions of artificial neurons need to be converted to a discrete “1” or “0” spike (Figure 4). To overcome this reduced precision, SNNs typically have several options. It can encode the same information in the time regime (i.e., a rate code), which increases latency and spike counts (thus offsetting some of the energy advantage of spiking), or it can be trained to more efficiently distribute its information across the population of neurons, which at minimum requires a different approach to training and often requires additional neurons. These mitigations all introduce complications for existing spiking neuromorphic platforms, which typically have slower clock speeds for individual neurons (thus more sensitive to any added latency cost) and have spike-routing paradigms optimized for sparse event communication. For this reason, maximizing the value of SNNs on neuromorphic likely requires faster neurons, higher throughput spike routing, or an algorithm advance to more efficiently train SNNs to be suitable for existing hardware constraints. Without such advances, there remains a strong possibility that specialized application-specific integrated circuits (ASICs) for ANNs will outperform less-specialized spiking architectures.

**For configurable spiking architectures to be commercially viable, it is, therefore, important they have utility beyond ANNs.** There are two clear application paths available to spiking hardware. The most obvious is to pursue more the brain-like algorithms for which these architectures were originally intended, such as neural-inspired navigation[23] or neural-based olfaction circuits on Loihi.[24] It is, furthermore, entirely reasonable that cognitive tasks currently emphasized by ANN research (e.g., image processing, text generation) can be reimplemented as more brain-like SNNs, prioritizing the routing of information as spikes and local learning as opposed to approaching neural computing with linear algebra and global optimization techniques. However, there are considerable market pressures of AI to stay the course with algorithmic approaches compatible with the dominant hardware available,[25] and as stated earlier, this prospect of brain-derived computing offers tremendous potential impact, but as explained in the following, the timeline for this algorithm development is complicated, making it a longer-term solution the market pressures facing this technology now.

In the near term, the algorithm diversity required to spur spiking hardware technology development likely has to be provided by existing applications. For this reason, here, we advocate for a third, perhaps, unorthodox, application of spiking neuromorphic hardware: numerical computing. The last two years have seen a rise in proposed algorithms for using spiking architectures for

![Figure 4](https://www.advancedsciencenews.com/doi/10.1002/adv.2000191)

**Figure 4.** Differences between continuous valued ANNs and SNNs. Top left: ANNs typically use continuous activation functions, such as sigmoids (blue) or rectified linear units (ReLUs; red dashed), to transform a neuron’s input to output. Top right: SNNs use discrete thresholds, whereby the neuron’s internal voltage (sum of inputs and potentially a decayed previous state) is compared with a threshold to determine whether the output is a “1” or a “0.” Bottom left: different inputs streaming through network will change ANN neuron activities continuously. Bottom right: in a rate-coded spiking neuron, different inputs will change firing rate of neuron, and the rate approximation is shown as a dashed line.
scientific computing tasks. In particular, there are three classes of conventional algorithms that appear well suited to implement on spiking architectures: graph algorithms (e.g., dynamic programming), Monte Carlo algorithms (e.g., solving stochastic differential equations), and some linear algebra tasks (e.g., Strassen matrix multiplication).

Combined, the three domains have to the potential to form a compelling argument for spiking neuromorphic approaches. This diversity of applications justifies a configurable architecture, as opposed to specialized ASICs. The utility within both ML and numerical computing justifies exploration for use in embedded systems (e.g., mobile devices) as well as a component in future heterogeneous platforms (e.g., high-performance computing systems). Finally, the prospect of enabling new classes of brain-derived algorithms makes it appealing as a research platform.

2.3. Scaling Neuromorphic Systems with Hybrid Analog-Spiking Architectures

The spiking applications mentioned earlier leverage two computational advantages of neuromorphic hardware: event-driven communication and extreme parallelism. Combined, these two provide considerable theoretical advantages; however, these benefits will likely be most impactful at larger scales. Similarly, while the field points to mammalian neural systems such as cortex as a source of algorithm inspiration, it is important to note that the smallest mammalian brains (e.g., mice) are still roughly 100× larger and significantly more configurable than the biggest single neuromorphic chips (≈1 million neurons).

For this reason, scaling neuromorphic platforms should continue to be seen as a priority on the hardware-design side. While there has been a lot of research on technologies that can help scaling, such as analog devices for synapses and higher density neurons, it is crucial that this scaling be targeted toward a broad class of algorithms (or similarly risk being outclassed by conventional ASICs). To date, much of the research in scaling analog neuromorphic systems has focused on ANNs, which is natural given their market demand; however, it is unclear whether a prioritization on more efficient vector-matrix multiplication (VMM; the dominant kernel in many ANN applications) is sufficiently impactful. For this reason, as with spiking hardware, there is growing interest in non-deep learning applications of resistive memory-based hardware, with attention primarily focused on numerical applications that are heavily linear algebra centric.

In general, the challenges stated earlier for spiking neuromorphic hardware (Box 1) can be similarly stated for analog approaches; however, the neuromorphic field has been challenged by the fact that the analog and spiking communities have addressed these challenges with some degree of opposition to one another. This opposition by the respective communities is unfortunate, as the spiking field likely requires analog memory for scaling, and the analog community likely requires the diversity of applications amenable to spiking approaches. For this reason, it is more viable for both communities to explore the potential of a truly hybrid spiking/analog neuromorphic architecture.

A natural first step hybrid architecture is to combine analog VMM crossbar accelerators with very simple spiking representations (i.e., binary “1” or “0” activations) to minimize costs associated with communication and analog-to-digital conversion. Long term, this combination is probably too limited to scale to biological sizes; the threshold gate model is likely not sufficiently rich to overcome the precision limitations that emerge in large analog crossbars. However, short-term evaluations of these architectures can provide insights into how analog noise and spiking precision interact with one another in novel hardware architectures (as they do in the brain), among other questions (Box 2).

---

**Box 1. Challenges for Spiking Neuromorphic Hardware**

- Can spiking hardware be used for mainstream AI techniques? Is it efficient?
- Does spiking have anything to offer beyond neural models?
- Is there a theoretical value to using temporal-coding with spikes?
- Can spiking and learning be used together practically?
- Can spiking architectures be improved?
- How can people program spiking architectures?

**Box 2. Challenges for Hybrid Analog-Spiking Neuromorphic Hardware**

- How do challenges of analog and spiking interact with one another?
  - *Example:* How does low-precision of spiking interact with noise of analog memory in ANNs?
- Are short-term hybrid architectures realistic? What does a true hybrid architecture look like?
- What is the right theoretical perspective to combine analog with event-driven communication?
- What materials/devices make sense? Can they be used for the spiking components as well?
- At what scales does this approach begin to open new possibilities?
- Do we have the requisite hardware pieces for learning? What else is needed?
2.4. Driving Toward Brain-Derived Algorithms

While neuroscience has been noted as an inspiration for computing since its earliest days, there has been a renewal of optimism about the prospects of more extensive inspiration from the brain. Much has been written about the goal of achieving brain-inspired AI, so this article focuses on the potential impact of developing successful neuromorphic hardware on facilitating the long-term transfer of neuroscience knowledge into computing.

Despite the significant interest in linking neuroscience and computing, there has long remained a technological and cultural gap between the two fields (Figure 5). There are a number of explanations for this gap, but in recent years, the successes of AI and the increased sophistication of neuroscience tools have amplified pressure to bridge this gap. While the brain could influence computing algorithms in many distinct ways, the recent success of sensory cortex-inspired neural networks for AI applications has been the dominant interface where neuroscience and AI is being explored.

The neuromorphic hardware trajectory described earlier is premised on the conclusion that enabling ANNs is not a sufficiently powerful market force to develop this new technology. However, if neuromorphic computing can successfully have utility beyond ANNs, it can similarly enable a broader influence of neuroscience on computing. The presence of a hardware platform that is naturally suited for brain-like algorithms will enable the development of a new class of brain-derived algorithms. These novel algorithms, in turn, can help drive an increased market pressure for neuromorphic hardware.

The use of the term “derived” as opposed to the more conventional “inspired” here is intentional. While intent varies across researchers, brain-inspired algorithms typically are existing and understood algorithms that have been extended or adjusted to incorporate some aspects of the brain for improved performance or efficiency. Such brain inspiration can be powerful and, for many applications, is sufficient to achieve the desired performance. An example is sparse coding—originally, a theoretical abstraction of how the visual cortex encodes information and has turned into an ML algorithm on its own.

In contrast, brain-derived algorithms speak to those potential approaches that target problems solved by the brain in a manner that is essentially distinct from existing algorithmic approaches. For example, the Neural Turing Machine algorithm was proposed to capture hippocampus-inspired memory in AI frameworks, yet relies on a very conventional computing implementation. In contrast, an algorithm based on the actual implementation within the mammalian hippocampus that encodes experiences in an episodic-like way would provide a novel capability that currently eludes conventional AI approaches and would likely look different than any current ANN architecture. For instance, an effectively implemented brain-derived memory could potentially combine the desirable associative capabilities of content-addressable memories with the capacity of random-access memories while scaling efficiently on neuromorphic hardware.

Figure 5. Eliminating the technical gap between neuroscience and computer science. Top: historically, there has been notional but limited interaction between neuroscience and computing communities. Middle: the recent successes in neural network approaches to AI have allowed some interchange of ideas between computing fields and neuroscience. Bottom: collaboratively pursuing truly brain-derived computing can provide a more natural means of interaction.
There historically are few fully brain-derived approaches to solving computing tasks, but, of course, the extent of the influence of neuroscience is a spectrum and often debatable. The intent of this label is not a litmus test, but rather simply to emphasize that one goal of the neuromorphic computing community should be to facilitate the development of algorithms that may push the boundaries of brain derivation. At the simplest level, the value of a neuromorphic architecture would be to provide an alternative path to ANNs to facilitate the design of new algorithms. However, if approached well, a robust neuromorphic community can become the translation layer between the existing neuroscience and computing fields, in large part, because the neuromorphic computing community is currently facing many of the fundamental problems (programming paradigms, benchmarks, etc.) that will be required going forward.

Achieving successful brain-derived algorithms that impact computing broadly will require researchers from both fields to help bridge the gap with neuroscience. While the neuroscience and algorithms communities must work together to ease the translation of biology into computationally suitable descriptions, the neuromorphic computing community must contribute to the process if their technology is to help this translation.

Box 3 highlights several challenges facing the neuromorphic field if it is to be useful in developing brain-derived algorithms. Most of these challenges are straightforward to describe, without the software and theoretical tools to translate biology into hardware effectively, the process will stall. However, in practice, these challenges will require long-term commitments and investments. Addressing current gaps such as identifying useful benchmarks must be approached with care, as a poorly chosen benchmark can be more detrimental than no benchmark at all. Furthermore, the requirement of cross-disciplinary buy-in makes provides a fundamental cultural hurdle that must be overcome for this approach to be successful.

Box 3. Challenges for Developing Brain-Derived Algorithms

- Can neuromorphic computing identify an advantage for implementing more brain-like ANNs?
- What are the right programming paradigms to translate neuroscience into neuromorphic hardware?
- Can we design mechanisms to translate biological complexity into neuromorphic kernels?
- How should neuromorphic computing extend to scales above and below the spiking neural model scale?
- What are appropriate hardware-cognizant benchmarks for emerging brain-derived tasks?
- How should brain-derived algorithms interface with conventional AI algorithms?
- Can neuroscience provide unified insights for computing from both systems and cellular scales?

Box 4. Long-Term Hardware and Materials Challenges Neuromorphic Hardware

- Can event-driven communication plus low-power devices enable true 3D technology?
- Is there a way to get scalable “point-to-point” synaptic connectivity?
- Can optical communication and computing enable neuromorphic computing to achieve super-biological speeds and efficiency?
- Can hardware enable structural plasticity and global modulation?
- Can analog device research avoid the temptation of focusing exclusively on linear algebra?
- How do we evaluate emerging materials and device technologies before the architectures and algorithms are well understood?
- Can bio-compatible materials suitable for brain-machine interface be developed at scale?
with the existing technologies already on the market, such as linear algebra accelerators or graphics processing unit for deep learning. A key argument to this article is that most arguments for neuromorphic computing have had too many “if’s”: if a proposed hardware approach scales, if the algorithms work, and if an important application space emerges, then neuromorphic hardware will be invaluable. The conservative hardware roadmap described earlier argues that there exists a justification today for neuromorphic computing without that level of uncertainty. However, it is worth considering how following this roadmap may help guide more technologically exotic hardware approaches (Box 4).

From the brain-derived algorithms perspective the neuromorphic system technologies described earlier have clear appeal over conventional technologies, but they still lack a clear relationship to how the brain processes information. As an example, most current hardware technologies are effectively 2D, effectively imposing a computational penalty on algorithms that require more complex embedding. Current neuromorphic solutions likewise impose discrete quantized values to weights, timing, and thresholds, which is typical (and in fact a necessity) for any computer architecture, but remains a somewhat unnatural concept when considering how the brain processes information.

Similarly, from a novel hardware perspective, the aforementioned technologies are interesting, but they arguably fail to truly represent a post-Moore’s Law technology, defined as one that may scale in a fundamentally different way than CMOS transistors. Much of the appeal of neuromorphic computing to the materials and devices community is that neuromorphic approaches could provide a path forward to allow these technologies to truly disrupt computing, as opposed to simply judging whether a new device could be an effective drop-in replacement for silicon in transistor technologies. Implicit in such considerations is that the materials properties that make silicon transistors so advantageous for computing may not be as relevant.

Analog computing has long embraced neuromorphic approaches as an opportunity for impact computing. Much of this attention has focused on resistive random access memory (ReRAM) crossbars, and other non-volatile memories, which are a natural fit for artificial synapses in ANNs.[51] Despite the relatively long timescales for these emerging devices to become commercially viable, there has been a temptation to optimize devices for the current trends in AI described earlier. While some examination of ReRAM devices considered their potential for in situ learning (such as STDP) and neural dynamics,[52–54] the emphasis on deep learning inference acceleration has prioritized finding devices that show high reliability and linearity in responses.[55–57]

Regardless of how neural algorithms develop, there are identifiable trends in neural networks and known features of neurobiological circuits that can be used to motivate long-term hardware design. Moving away from the ANN motivation and the goal of achieving higher density synaptic weights, there has been interest in emerging devices that capture other aspects thought to be important for neural computation. For instance, while the functional importance of stochasticity in the brain is an area of ongoing exploration, several proposed neuromorphic devices may have intriguing stochastic switching properties that may provide controllable randomness in neural dynamics.[58,59] Neuristors built from Mott memristors have been shown to be highly effective at implementing the ion channel dynamics of neurons.[58]

Arguably, the biggest need emerging from the current understanding of neuroscience contrasted with today’s neuromorphic hardware is the requirement of a more robust communication paradigm. The brain’s circuits do not naturally reduce to sequences of simple linear algebra transformations; rather, the brain’s circuits are highly recurrent and interconnected—demonstrating highly complex graphical structures that do not embed well into two dimensions (for example, see the previous studies [60,61]). Furthermore, while the brain’s microcircuits are heavily biased toward local connections, the non-local connectivity is incredibly important for achieving higher order functions.[62]

For this reason, proposed technologies that have the potential to shift how information is communicated in neuromorphic systems are important to develop. Optical neuromorphic computing has long been considered one such opportunity.[63] While still immature, several recent optical approaches have shown considerable potential in impacting neural computing.[64,65] While optical approaches are an example of a technology whose baseline costs may limit competitiveness at small scales; it may be a fundamentally critical component of large-scale neuromorphic systems that leverage long-distance communication. Other pathways to plausible point-to-point 3D communication, such as carbon nanotubes,[66] organic synapses,[67] and silver nanowires,[68] are similarly positioned to change how we approach neuromorphic architectures long term.

Beyond communication, there remain considerable opportunities on the hardware side to emulate the robust plasticity of biological systems, which likely will be important for applications such as the brain-derived memory proposed earlier. While most device research has focused on synaptic plasticity, the brain experiences learning and plasticity at many spatial and temporal scales.[14] For instance, the human brain continues to add neurons throughout life in a few regions, and continual neurogenesis is pervasive in many non-mammalian brains.[69] Likewise, the addition and removal of synapses are an ongoing process throughout the brain.[70] While the computational importance of these specific processes is still a subject of investigation, it is notable that the deep learning community has adopted similar processes, such as Drop Out, as regularizers in training ANNs.[71] Less dramatically, the brain also leverages global modulation through transmitters, such as serotonin and dopamine, whereby a fairly coarse signal is provided to many synapses and neurons in a local region.[72] Neither structural plasticity nor modulation has yet received significant attention from hardware design; however, both will likely be of growing importance as neuroscience, and the AI communities look beyond the synapse for learning.

4. Conclusion

The above-mentioned analysis highlights a trajectory of neuromorphic computing to address the looming needs of the AI community for more capable, efficient hardware by enabling it to reach both the scalability and algorithm capabilities that are often assumed but have yet to be realized. This will require a multi-disciplinary effort, with contributions from many
technical fields, and it will similarly require that many communities expand beyond their comfort zones to contribute to this technology.

Beyond these cultural challenges, this analysis accounts for three competing pressures. First, the development of new hardware paradigms requires a commercially viable market for their success and ongoing development. Second, neuromorphic hardware must exist on what has historically been an unstable equilibrium point between specialized hardware and general-purpose hardware. Finally, designers of neuromorphic algorithms face a challenge of whether to build off of the successes of ANNs or to seek entirely novel capabilities.

For each of these cases, there is a tension between seeking near-term wins and building toward the long-term vision outlined earlier. For any technology in isolation, it is tempting to pursue a near-term strategy, but neuromorphic computing is not well positioned for near-term wins: ASICs will likely win any accelerator market, and ANNs have been heavily optimized for the tasks they perform well without much neural inspiration. In contrast, as the likely market demand for energy-efficient and scalable hardware as well as new capabilities is close but not yet upon us, there is an opportunity to consider the longer term opportunities illustrated here.

Importantly, the market analysis here was cast specifically toward the needs of the AI field, as this has been the most heavily explored. However, neuromorphic computing has a potentially much wider reach than what is currently pursued within the AI community today, and these applications similarly will benefit from the advances considered here.

Acknowledgements

The authors acknowledge financial support from the DOE Advanced Simulation and Computing program and Sandia National Laboratories’ Laboratory Directed Research and Development Program. Sandia National Laboratories is a multiprogram laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. This article describes objective technical results and analysis. Any subjective views or opinions that might be expressed do not necessarily represent the views of the U.S. Department of Energy or the U.S. Government.

Conflict of Interest

The author declares no conflict of interest.

Keywords

artificial intelligence, brain, neural algorithms, neuromorphic computing

Received: August 14, 2020
Revised: September 23, 2020
Published online: November 16, 2020
James Bradley Aimeone is a Principal Member of the Technical Staff in the Center for Computing Research at Sandia National Laboratories. He received his Bachelor of Science and Masters in chemical engineering from Rice University, and his Ph.D. degree in computational neuroscience from the University of California, San Diego. He currently investigates the development of brain-derived AI, neural algorithms for numerical computing, and neuromorphic computing architectures.