Neuro-inspired computing chips

Article *in* Nature Electronics · July 2020

DOI: 10.1038/s41928-020-0435-7

9 authors, including:

Wenqiang Zhang
Tsinghua University
20 PUBLICATIONS **484** CITATIONS
SEE PROFILE

Bin Gao
Peking University
155 PUBLICATIONS **2,717** CITATIONS
SEE PROFILE

Jianshi Tang
IBM
99 PUBLICATIONS **4,013** CITATIONS
SEE PROFILE

Peng Yao
Tsinghua University
30 PUBLICATIONS **518** CITATIONS
SEE PROFILE

nature electronics

REVIEW ARTICLE

https://doi.org/10.1038/s41928-020-0435-7

Neuro-inspired computing chips

Wenqiang Zhang[1,6], Bin Gao[1,2,6], Jianshi Tang[1,2], Peng Yao[1], Shimeng Yu[3], Meng-Fan Chang[4], Hoi-Jun Yoo[5], He Qian[1,2] and Huaqiang Wu[1,2] ✉

The rapid development of artificial intelligence (AI) demands the rapid development of domain-specific hardware specifically designed for AI applications. Neuro-inspired computing chips integrate a range of features inspired by neurobiological systems and could provide an energy-efficient approach to AI computing workloads. Here, we review the development of neuro-inspired computing chips, including artificial neural network chips and spiking neural network chips. We propose four key metrics for benchmarking neuro-inspired computing chips — computing density, energy efficiency, computing accuracy, and on-chip learning capability — and discuss co-design principles, from the device to the algorithm level, for neuro-inspired computing chips based on non-volatile memory. We also provide a future electronic design automation tool chain and propose a roadmap for the development of large-scale neuro-inspired computing chips.

Artificial intelligence (AI) can already exceed the capabilities of humans in certain specific classes of problems, including image/speech classification and gaming[1,2]. However, the practical applications of AI, which require the processing of large amounts of data and thus place significant demands on computing speed and power efficiency, are limited by computing hardware[3]. Conventional computing hardware is based on a von Neumann architecture in which processing and memory units are physically separated (Fig. 1a,b). This architecture is well suited to scientific computing, with high-precision data representation and accurate Boolean calculations, but it is inefficient in handling AI tasks. As a result, the application of AI technology in power-constrained environments, such as edge computing and the Internet of Things (IoT), requires the development of new chip architectures and device technologies.

Neuro-inspired computing chips emulate the structure and working principles of the biological brain, and offer a promising approach to the development of intelligent computing[4,5]. Compared to conventional systems, these neuro-inspired computing chips are expected to offer advantages in terms of energy efficiency and computing power when processing AI workloads. Many different types of neuro-inspired computing chips have been developed over the last few years, and various neuro-inspired features have been implemented in these chips, from the device level to the circuit and architecture level. Neuro-inspired computing chips are though still only at an early stage of development and thus it is important to explore the challenges and opportunities for the field.

In this Review Article, we examine the origins of neuro-inspired computing chips and discuss recent progress in the field. We identify four key metrics for evaluating the performance of the chips: computing density, energy efficiency, computing accuracy, and learning capability. We then examine the challenges and co-design principles involved in developing large-scale chips based on non-volatile memory (NVM). We also discuss a future electronic design automation (EDA) tool chain and propose a technological roadmap for the development of large-scale neuro-inspired computing chips.

History of neuro-inspired computing chips

Over the past century, our understanding of the physical structure and working mechanisms of the brain has greatly developed. The improved insight of neuroscience gives rise to the idea of developing electronic hardware to emulate the neural system of the brain. The human brain, which consists of more than 10^{11} neurons and 10^{15} synapses, is much more efficient than conventional von Neumann computers for recognition and decision-making tasks, and it consumes only 20 watts (ref. [6]). Information is stored in the human brain in the form of synaptic weights. Synaptic plasticity (for example, the ability to increase or decrease these weights by means of changes in conductance) is the underlying mechanism responsible for knowledge-based learning. A neuron integrates all incoming weighted spikes to determine whether a spike should be fired. Based on the development of neuroscience and the emergence of very-large-scale integration (VLSI), neuro-inspired computing chips have been proposed to emulate the widely consensual neuron–synapse structure and/or some of the working mechanisms of the brain, such as spike encoding, parallel in-memory computing (Fig. 1).

Research on neuro-inspired computing chips started as early as the 1980s, but the idea of neuro-inspired computing hardware can be traced even farther back to the introduction of the Perceptron[7] in 1957, which was a dedicated analogue computer using motorized potentiometers as the weights. In 1961, ADALINE[8] was constructed by assembling discrete variable resistors into an adaptive hardware perceptron with training capabilities. However, without the capability of high-density device integration, the size of these original pieces of hardware was huge, while the number of contained synaptic devices was quite limited.

With the development of silicon-based integration technologies for VLSI and the emergence of strong demands for computing capability, interest greatly arose in neuro-inspired computing chips during the 1980s, including works on analogue spiking integrated neural systems and analogue non-spiking mixed-signal neural network hardware. The spiking approach was based on the analogy between the physical behaviours of integrated circuits and the biological spiking behaviours in the brain. A representative example

[1]Institute of Microelectronics, Beijing Innovation Center for Future Chips (ICFC), Tsinghua University, Beijing, China. [2]Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China. [3]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [4]Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan. [5]Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea. [6]These authors contributed equally: Wenqiang Zhang, Bin Gao. ✉e-mail: wuhq@tsinghua.edu.cn

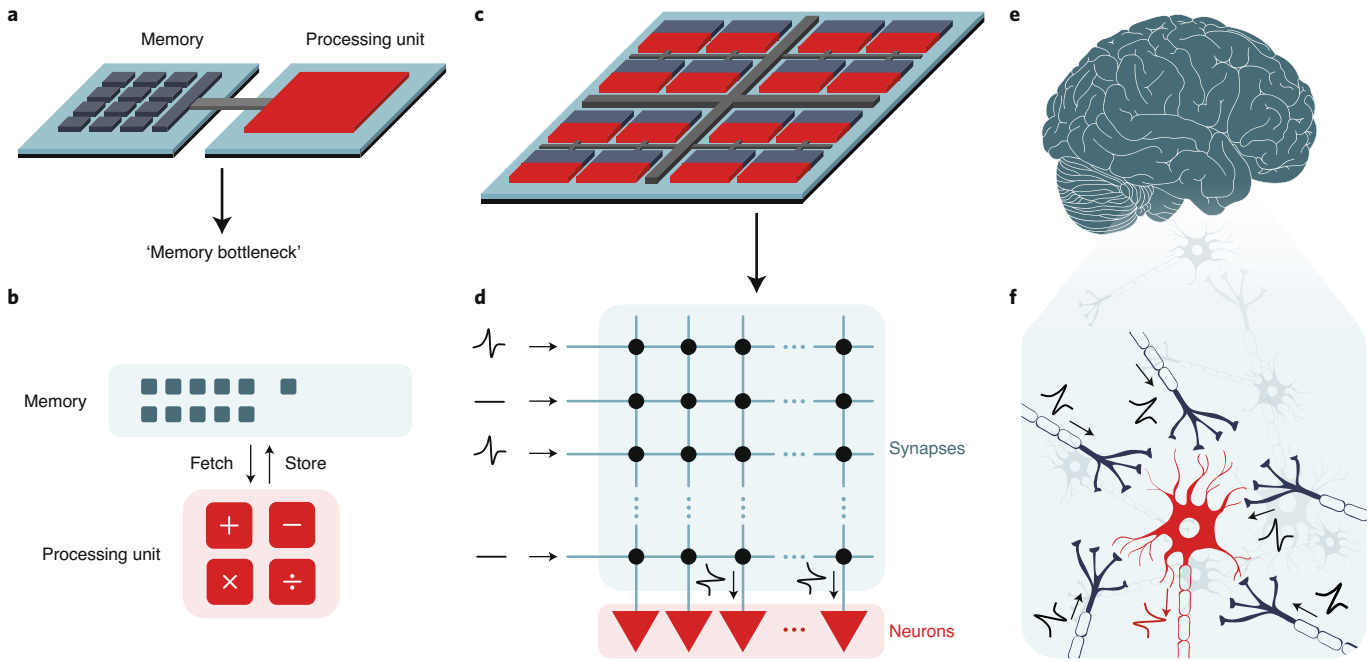

Fig. 1 | Computing architectures and paradigms. a, Schematic of the von Neumann architecture and the corresponding 'memory bottleneck' between the memory and processing units. **b**, Program–storage computing paradigm. The input data are fetched from memory and processed in the processing unit, and the results are then stored back in memory. **c**, Schematic of the neuro-inspired architecture. **d**, Neuro-inspired computing paradigm. The fine-grained synapse–neuron core enables arbitrary connections between input and output electronic neurons through electronic synapses. **e**, Schematic of the biological brain architecture. **f**, Schematic of biological neurons and synapses. A neuron integrates incoming spiking signals from other neurons to which it is connected through synapses until it reaches a certain threshold and 'fires'.

is Mead's work[9], which was based on metal–oxide–semiconductor field-effect transistor (MOSFET) circuits emulating neuronal behaviours. In comparison, non-spiking approach utilized physical laws to accelerate matrix multiplication in neural network operations. For instance, a group at Bell Laboratories fabricated a 12×12 unprogrammable synaptic resistor array to perform vector–matrix multiplication in a neural network[10]. However, these pioneering chips were constrained to small-scale demonstrations due to practical limitations of the available semiconductor technology at that time, such as a lack of compact devices emulating biological behaviors.

When HP Labs first linked the resistive switching behaviour of a solid-state device to the conceptual memristor in 2008[11], a fundamental revolution started by utilizing the basic device capabilities to develop neuro-inspired computing chips. In 2010, an analogue memristive device was fabricated to demonstrate synaptic functions, as the first attempt to explore NVM-based neuro-inspired computing[12]. Since then, increasing research efforts have been devoted to the devices, architectures, chips and algorithms for NVM based neuro-inspired computing.

Progress in neuro-inspired computing chips

The breakthroughs in algorithms and powerful hardware brought the AI revolution. As the demand for intelligent IoT and smart edge devices continues to grow and the neural network algorithms become increasingly complex, neuro-inspired computing chips are attracting more and more attention for their potential superior performance. Before reviewing the recent progress in this field, it is necessary to clarify the difference between present neuro-inspired and non-neuro-inspired computing chips.

Recently developed application-specific integrated circuit (ASIC) accelerators[13,14] towards parallel multiply-and-accumulate (MAC) computations do not belong to the category of neuro-inspired computing chips. They utilize structures involving multiple processing elements (PEs), which partially exploit parallel computation capabilities to speed up the execution of AI workloads. These ASIC accelerators exhibit better inference speed and energy efficiency than conventional graphics processing units (GPUs) when processing artificial neural network (ANN) algorithms, especially deep neural network (DNN) algorithms. However, they still perform MAC computations by means of digital circuits in each PE. The need for intensive data transfer between the MAC units and data buffers still limits their energy efficiency. Hence, such ASIC accelerators are still based on the von Neumann architecture, and hence they cannot be classified as neuro-inspired computing chips.

Here, the neuro-inspired computing chips are defined as the chips that emulate the structure and part of the working mechanisms of the biological brain. The current neuro-inspired computing chips can be loosely classified as ANN (DNN) chips and spiking neural network (SNN) chips, because most of them work with either ANN or SNN algorithms. In an ANN chip, the neuron states are encoded as digital bits, clock cycles or voltage levels, while in an SNN chip, information is encoded into spike timing, as in a bio-plausible neural network[15]. Although there is distinct difference in these two types of neuro-inspired computing chips, both have some features similar to those of the brain, such as a neuron-synapse structure, in-memory computation and learning capabilities. A number of medium-scale functional neuro-inspired computing chips for ANN and SNN applications have emerged. Among these chips, neuro-inspired ANN hardware realizations aim to improve energy efficiency by computing weighted-sum tasks in memory, while SNN hardware realizations are engineered to seek ultralow power consumption and run at relatively low frequencies to emulate realistic biological behaviours.

ANN chips. Neuro-inspired computing chips such as Conv-RAM[16] and 1-Mb NVM macros[17] are inspired by the in-memory computing

feature in brain to accelerate ANN computations. Both digital and analogue memory devices have been exploited for neuro-inspired ANN chips. MIT's Conv-RAM chip consists of a 256×64 1-bit SRAM array. This chip performs 1-bit weight multiplication and average through charge sharing of bit-lines. It achieves a >16× improvement in energy efficiency compared with prior ASIC accelerator implementations[18]. Since SRAM is volatile, NVM-based chips were developed. NTHU's macro chip integrates one million units of 1-bit resistive random-access memory (RRAM). By dividing the weights into positive and negative memory arrays and subtracting the output results in digital domain, it has the ability to implement a binary-input-ternary-weight network and can perform one MAC operation within less than 16 ns. In a more recent work, a 2-b-input 3-b-weight network has been presented by subtracting positive and negative weights in the same array but different columns[19]. Analogue computation can significantly improve the computing performance and area efficiency by combining several binary memory devices into one analogue memory device. Panasonic designed a chip with 2Mb analogue RRAM to realize a perceptron neural network[20]. The on-chip memory devices showed a well-controlled analogue cell current with a linear dynamic range and low variation.

In addition to these macro chips, recent analogue NVM arrays without peripheral circuits have demonstrated the ability to execute various ANN algorithms for both learning and inference with high energy efficiency. For example, UCSB's perceptron was the first to utilize a 12×12 crossbar to classify 3×3 binary images[21]. THU's 128×8 array accomplished face recognition tasks with an improvement of three orders of magnitude in energy efficiency compared to a processor with off-chip memory[22]. IBM's mixed hardware-software multilayer neural network on 204,900 analogue memory devices showed the equivalent accuracy of a pure software implementation with an improvement of more than two orders of magnitude in energy efficiency compared with a GPU[23]. UMass's 128×64 RRAM crossbar system could solve real-world problems involving the regression of the number of airline passengers and the classification of individual humans by their gaits with long short-term memory (LSTM), one variant of recurrent neural network[24].

Although these macro and array chips employed various types of synaptic memory, such as SRAM[25,26], DRAM[27], flash memory[28] and analogue NVM[29], all of them eliminate the energy-intensive and time-consuming data transfer by means of the neuro-inspired in-memory computing paradigm, which provides an alternative approach to the program–storage paradigm of the von Neumann architecture. In addition, in the biological brain, the information stored in the synapses is represented by analogue states, each of which is quite loose and noisy. Most of these chips employ 1-b or few-b devices to store synaptic weights. A more radical concept is to perform analogue computing by storing information in analogue devices and processing purely in the analogue domain[30].

SNN chips. Other neuro-inspired computing chips such as BrainScaleS[31], TrueNorth[32] and Loihi[33] implement hardware SNNs to emulate the spiking-based information encoding and processing, which is one of the important neuronal behaviors. All of these chips adopt synapse–neuron architecture. Heidelberg's BrainScaleS system is based on uncut silicon wafers, with 384 high input count analogue neural network (HICANN) dies per wafer. One HICANN die consists of 512 analogue neurons, similar to the leaky integrate-and-fire (LIF) model, and 100,000 4-b SRAM-based synapses. The entire system can run 10,000 times faster than the real biological frequency (~kHz) but consumes 500 W per wafer.

Besides the circuit-level design, some chips also involved architecture level innovation. For example, IBM's TrueNorth chip integrates one million digital neurons and more than two hundred million transposable 1-b SRAM synapses, which are grouped into 4,096 distributed synapse–neuron cores. The central design of a TrueNorth synapse–neuron core is a 256×256 crossbar. In the crossbar, the incoming neural spike events are selectively connected to outgoing digital neurons to implement a kind of integrate-and-fire algorithm with 23 configurable parameters. TrueNorth has demonstrated reasonable classification accuracy on vision and speech datasets while maintaining its remarkable energy efficiency at a near-real-biological frequency (on the order of kilohertz)[34]. Intel's Loihi chip includes up to 130,000 neurons and 130 million 1-b to 9-b reconfigurable SRAM synapses. Loihi supports several on-chip learning rules and can solve optimization problems with an energy-delay product that is superior to that of a typical CPU-based solver by over three orders of magnitude. Recently, analogue NVM system was also developed to demonstrate SNN algorithm and unsupervised learning, exhibiting the potential for future in-memory computing SNN chips[35].

These SNN hardware realizations use a massively parallel architecture with short-distance connections for local communications and long-distance connections for global communications (Fig. 1c), which is similar to the architecture of the brain (Fig. 1e). Input signals, weighted by either electronic synapses (Fig. 1d) or biological synapses (Fig. 1f), are sent to neurons. These neurons can be implemented in the form of complementary metal–oxide–semiconductor (CMOS) circuits[32] or threshold switching devices[36]. At present, the available SNN chips can only mimic the structure of the brain at a rudimentary level and perform basic AI tasks.

Benchmarking metrics

A diverse range of approaches to develop neuro-inspired computing chips have been proposed, however, each of these hardware realizations has its own pros and cons. Based on the hardware realizations discussed above, several key metrics that represent important attributes of neuro-inspired computing chips are proposed, including the computing density, energy efficiency, computing accuracy, and on-chip learning capability. Beside these key metrics, throughput and bandwidth are also critical metrics for those real-time applications which need communication with environments.

Computing density. State-of-the-art DNNs contain millions of weights or more, which must be mapped to on-chip synaptic memories. As the weight capacity of neural networks rapidly grows[37], the density of the synapses within the limited on-chip synaptic area need improve accordingly. Thus, the computing density, which reflects the area efficiency of a chip, is a crucial metric for large-scale neuro-inspired designs. Notably, the corresponding peripheral circuits for synaptic memory should also be included when calculating the computing density. Here, we evaluate the trends of the computing density of neuro-inspired computing chips based on the reported numbers of synapses and areas of synapse–neuron core, as shown in Fig. 2a.

In recent years, the growth in the computing density of chips based on CMOS technology has slowed down due to the large area of SRAM cells and the increasing difficulty of CMOS scaling. By contrast, the NVM approach shows greater potential in terms of computing density even for a chip fabricated using a less advanced technology node. With the superior integration density, NVM-based neuro-inspired computing chips could also reach the ultimate performance density in learning phase (Fig. 2b).

Energy efficiency. Energy efficiency is a critical metric in hardware realizations to bridge the gap of power consumption between neuro-inspired systems and the biological brain. During the learning phase, the majority of the energy consumption is due to the programming operations of the synaptic memory. During the inference phase, currently, either the peripheral circuits (typically analogue-to-digital converters) and/or the read operations of the synaptic memory dominate the system's energy consumption.

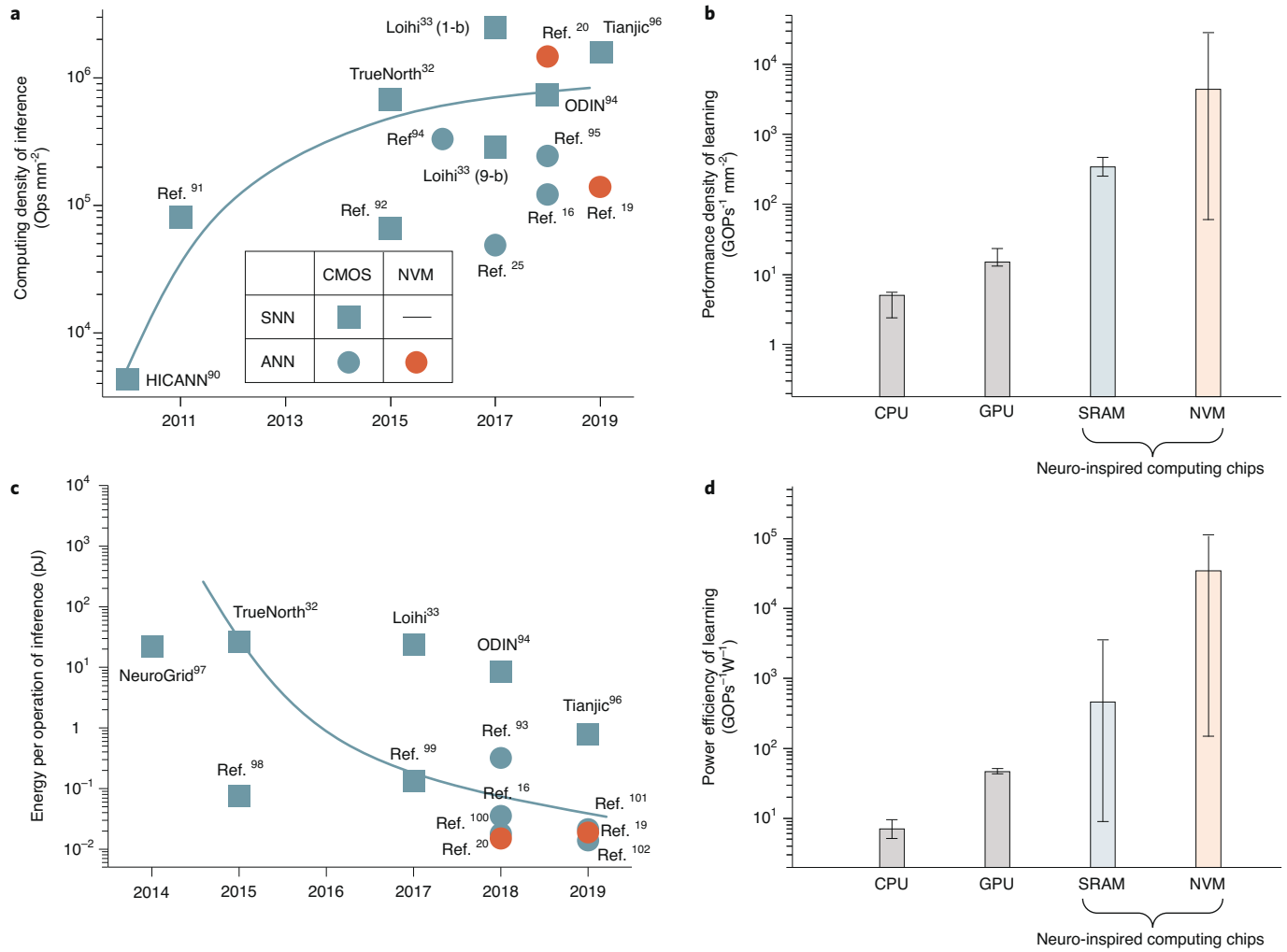

Fig. 2 | Benchmarks. a, Benchmarking computing density. The computing densities of representative neuro-inspired ANN and SNN chips based on CMOS and NVM technologies are evaluated. The computing density is defined as the number of on-chip synaptic elements per synapse–neuron core divided by the area of synapse–neuron core. The line is an approximation to indicate the trend of computing density. Data are from refs. [16,19,20,25,32,33,90–96]. **b**, The typical performance densities for learning phase of various chips. The bar indicates the average value and the error bars indicate the highest and lowest values. **c**, Benchmarking synaptic operation energy. The amount of synaptic energy consumed per operation in inference phase for an SNN or ANN chip is the energy dissipation per spike event or MAC operation, respectively. The red square shows the synaptic operation energy in learning phase. The line is an approximation to indicate the trend of energy per operation. Data are from refs. [16,19,20,32,33,93,94,96–102]. **d**, The typical power efficiency values for learning phase of various chips. The bar indicates the average value and the error bars indicate the highest and lowest values. The data of performance densities and power efficiency of CPUs and GPUs in learning phase is generated from the commercial products in recent years[103,104]. The benchmarking data of neuro-inspired computing chips in learning phase is generated from literature-reported results[23,24,33,95].

However, with the optimization of the architecture and an increasing computing density, the read energy of the synaptic memory will become dominant. Therefore, the energy per operation is another key factor for neuro-inspired computing chips to estimate the energy efficiency. Figure 2c shows the energy efficiency of recent neuro-inspired computing chips in inference phases. Figure 2d illustrates that the neuro-inspired computing chips demonstrates superior energy efficiency compared to a CPU- or GPU-based implementation in learning phases.

Computing accuracy. High computing accuracy is necessary throughout the lifetime of a chip. The computing accuracy of neuro-inspired computing chips will be influenced by device non-ideal factors and circuit noises, such as thermal noises and reliability issues. As a result, compared to defect-free simulation, the hardware implementation would show a lower computing accuracy[38]. For AI applications, the computing accuracy can be suitably represented by the final output accuracy when processing a specific task. In particular, when benchmarking a chip, one should evaluate the accuracy by running widely used models on standard datasets, such as the ImageNet[39] for object classification or Microsoft COCO[40] for object detection, and so on. In addition, any neuro-inspired DNN hardware realization is supposed to reach similar accuracy as the software results. Since the SNNs currently lack appropriate bio-plausible benchmark models and datasets, the computing accuracy of SNN chip is generally evaluated by running a DNN-based transformed spiking neural model and using the traditional datasets.

Learning capability. In most prior works, the learning phase has been implemented in the cloud using public datasets, and the learned parameters have then been downloaded to edge devices to perform inference tasks. However, for future edge workloads, the capability of local on-chip learning is necessary for personalization

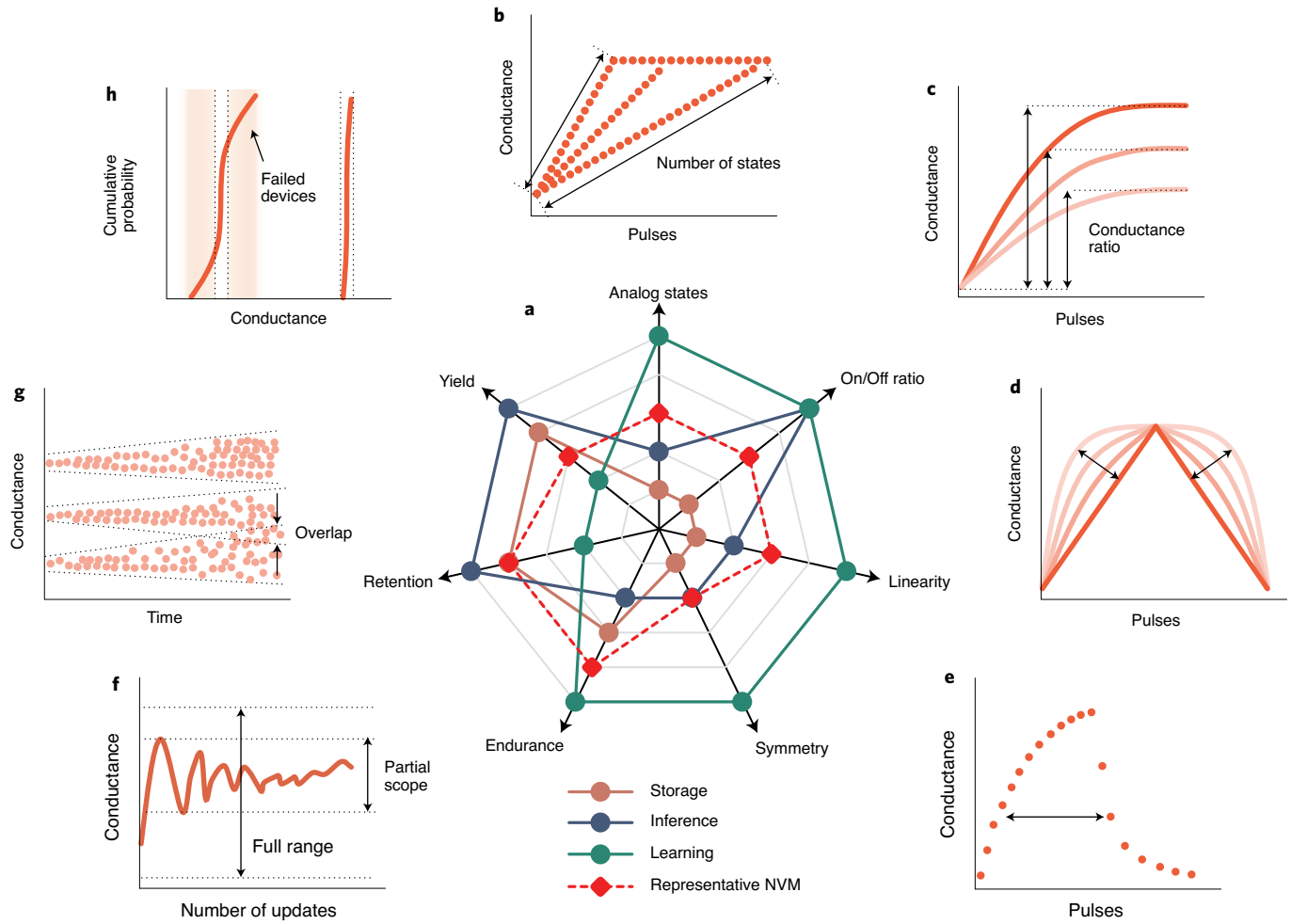

Fig. 3 | Application-dependent device metric requirements. a, Ranking of the qualitative device requirements for three prospective applications. A larger value on a given axis indicates a higher requirement in terms of the corresponding metric. The red line represents experimental NVM data that have been previously reported in representative works. **b**–**h**, Schematic illustrations of device requirements for computing: analogue states (**b**), on/off ratio (**c**), linearity (**d**), symmetry (**e**), endurance (**f**), retention (**g**) and yield (**h**). The dashed and solid curves in **b**–**e** indicate the conductance tuning of an analogue NVM device. The conductance updates of a NVM device in the training process are usually in the partial scope rather than in the full range of the conductance window (**f**). After NVM devices are tuned to different conductance levels, the conductance of the devices can fluctuate over time and two levels may overlap (**g**). The NVM devices that cannot be tuned to the target conductance level are considered failed devices (**h**). The device requirements are qualitatively estimated in Supplementary Table 1.

and privacy protection in domain-specific scenarios. In addition, learning is a sufficient driver of rapid adaptation to environmental changes and regarded as the foundation of intelligence[41]. Therefore, the learning capability is another critical metric for a neuro-inspired computing chip.

Co-design principles

As shown in Fig. 2, NVM-based chips exhibit obvious advantages compared to CMOS-based chips. NVM devices also have the potential to be scaled down to sub-2-nm size[42] and integrated into ultrahigh-density three-dimensional (3D) array[43], as well as the capability to emulate bio-plausible behaviors[44]. Therefore, NVM is a very promising candidate for future mainstream technology of large-scale neuro-inspired computing chips. Several small-scale NVM-based chips, from 1,000 cells to 1 million cells, have already been developed. However, the implementation of a neuro-inspired computing chip integrated with multiple large-scale analogue NVM arrays still faces many challenges. Extensive research efforts, from the device level to the chip level and to the algorithm level, need to be carried out. Future innovations of cross-layer co-design are highly desired, since the high complexity of neuro-inspired computing chips makes it difficult to reach the target performance through a specific level optimization only. In the following, we will elaborate the challenges and design considerations at each level, as well as cross-layer co-design methodology and future requirement of an EDA tool chain.

NVM devices. The performance requirements of NVM devices for neuro-inspired computing chips are largely dependent on specific systems and applications. Figure 3a shows the general requirements for different chip applications, including data-storage-only, inference-only, and on-chip learning applications, in comparison with the current representative NVM performance. The number of analogue states (Fig. 3b) determines the weight tuning precision. It has been reported that a precision of at least 8 equivalent bits is required for training a relatively large neural network[45], such as ResNet[46]. More than 256 analogue states have been realized by optimization of device material stack and write circuit design[20,47].

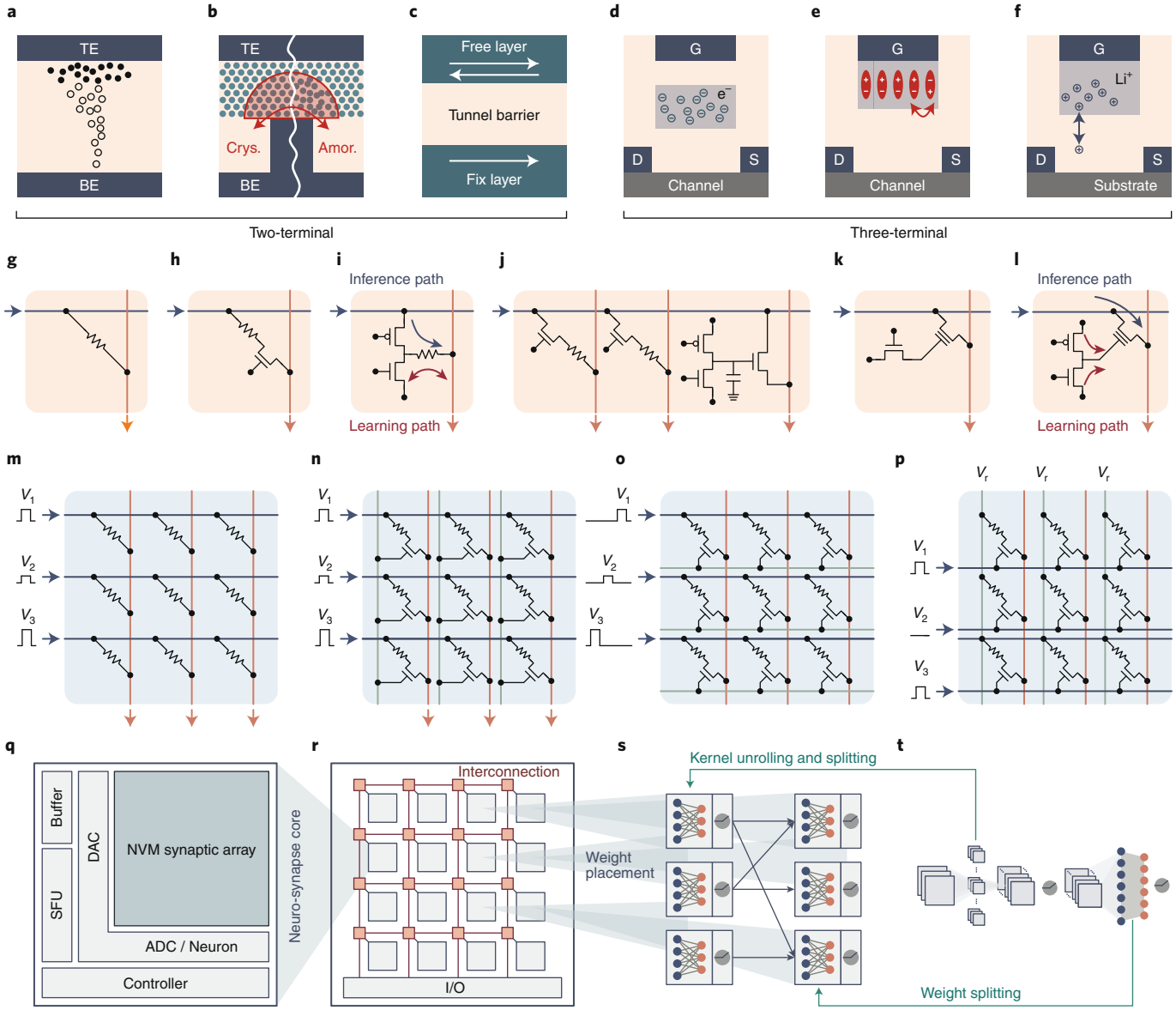

Fig. 4 | Analogue computing with NVM. a–c, Two-terminal NVM devices for synaptic weights: RRAM (**a**), PCM (**b**) and MRAM (**c**). The two-terminal synaptic devices exhibit both of the weight tuning and synaptic inference operation in the TE-BE path. **d–f**, Three-terminal NVM devices for synaptic weights: flash memory (**d**), FeFET (**e**) and ECRAM (**f**). The three-terminal synaptic devices exhibit weight tuning and read operation based on the gate-channel and drain-source paths, respectively. The FeFET device (**e**) utilizes the partial polarization switching within the ferroelectric gate oxide to change conductance. The conductance tuning of an ECRAM device (**f**) is based on the motion of Li ions between the solid-state electrolyte and tungsten oxide. **g–l**, NVM-based computing unit cell configurations: 1R (**g**), 1T1R (**h**), 2T1R (**i**), 2T2R+3T1C (**j**), 1T+1TriR (**k**) and 2T+1TriR (**l**). The blue arrows and red arrows indicate inputs and outputs, respectively. **m–p**, Array structures for computing: passive crossbar (**m**), parallel pseudo-crossbar (**n**), row-by-row pseudo-crossbar (**o**) and memory-like array (**p**). In the computing phase, the pulses with V_1, V_2 and V_3 amplitudes are applied in parallel or in sequence to row terminals of the array. **q**, NVM-based computing core. **r**, Generic architecture of NVM-based neuro-inspired computing chips. **s**, Mapped memory-centric intermediate representation. **t**, A typical neural network application. TE, top electrode; BE, bottom electrode; G, gate terminal; D, drain terminal; S, source terminal; SFU, special function units.

The on/off ratio (Fig. 3c), which is defined as the dynamic range in the analogue switching regime, determines the capability of mapping the weights in the algorithms to the device conductance. In contrast to binary switching, most NVM devices only have on/off ratios of less than 10 in the analogue switching regime. The linearity (Fig. 3d) in conductance tuning refers to the linearity of the curve relating the device conductance to the number of programming pulses. Generally, a programming scheme using identical pulses is preferred; otherwise, the cost of the pulse determination process would incur extra circuit overhead. The trajectory of the weight increase process usually differs from that of the weight decrease process, also resulting in asymmetry (Fig. 3e).

The overall requirements for SNN are similar to those for ANN, but the specific requirements for SNN might be different in terms of some device metrics, such as linearity or symmetry. In addition, additional requirements should be introduced for SNN, such as timing-related plasticity.

Different types of NVM devices have been proposed and fabricated for neuro-inspired computing applications. Based on the relation between the weight tuning path and the weight read path,

these synaptic weight devices can be categorized into two-terminal devices, such as RRAM[12], phase change memory (PCM)[23], magnetic random-access memory (MRAM)[48], and three-terminal devices, such as flash memory[28], ferroelectric field-effect transistor (FeFET)[49] and electrochemical random-access memory (ECRAM)[50]. RRAM (Fig. 4a) exhibits superior conductance tuning performance but faces challenges in terms of device yield and uniformity. In the literature, RRAM devices are sometimes also referred to as "memristor"[11]. The PCM (Fig. 4b) and MRAM (Fig. 4c) technologies are relatively more mature in term of large-scale manufacturing. Further engineering efforts will still be required to achieve controllable analogue conductance tuning behaviour. Three-terminal NVM devices provide a better control of conductance tuning capability through the modulation from additional gate terminal. In comparison, flash memory (Fig. 4d) is the most mature NVM technology. The main challenge regarding flash memory technology for computing application lies in its slow programming speed, which significantly increases the learning energy. FeFET (Fig. 4e) has gained attention due to its faster speed, lower programming voltage, and nearly symmetric channel conductance tuning ability using carefully designed programming schemes[49]. Very recently, an increasing number of emerging electrochemical synaptic devices have been developed (Fig. 4f), showing excellent analogue conductance tuning performance[50].

Device reliability and yield. Reliability issues are also important for NVM-based neuro-inspired computing chips. Endurance (Fig. 3f) is essential for on-chip learning, which requires a large number of weight tuning operations for each cell[51]. Some preliminary results have shown promise for sufficient endurance in the analogue switching regime, at least for small-scale datasets. Retention (Fig. 3g) is crucial for inference applications. Since an inference application requires a stable conductance value on each cell[52], any random change in conductance may cause the computing accuracy to degrade. In contrast, conventional data storage applications require only the maintenance of a certain resistance window and are not as sensitive to conductance drift as inference applications are.

Although relatively high yields have been achieved for binary NVMs, a practical yield for analogue NVMs is not currently available. The device yield for analogue NVM can be defined as the average percentage of the fabricated devices that function with analogue switching behavior, which is limited by the occurrence of stuck-state and abrupt conductance change phenomena. Particularly, it is usually observed that some NVM cells cannot be programmed to a specific intermediate state (Fig. 3h), even though these cells can be effectively switched between bi-stable states. The unprogrammable states vary from cell to cell due to the intrinsic stochastic mechanism of NVM devices. The inference applications usually require a high device yield, while most on-chip training applications could relax the requirement on device yield[53]. Due to that reinforcement learning is largely dependent on the intrinsic programming stochasticity of the device, the device requirement for reinforcement learning will be strict[54].

Computing unit. Various unit structures have been proposed for neuro-inspired computing chips with the abovementioned NVM devices. For two-terminal devices, the 1R cell in a crossbar array[55] (Fig. 4g), which may suffer from write disturbance, shows the best area efficiency of $4F^2$ (where F is the feature size associated with a given technology node), whereas the others have area efficiencies of $\geq 6F^2$. The 1T1R cell[56] (Fig. 4h), which uses a transistor as the selective device, is the most common unit cell used in neuro-inspired computing chips due to the balance between the integration density and scalability. In the 2T1R cell (Fig. 4i), the inference and learning paths are separated, avoiding additional complexity and power consumption associated with timing schemes[57]. Recently, to increase the dynamic range and enable linear and symmetric conductance tuning, a pair of NVM devices providing long-term storage and a capacitor with near-linear update capabilities and varying significance were combined to form the 2T2R+3T1C unit cell[23] (Fig. 4j). For three-terminal devices, the 1T+1TriR[58] (triple terminal resistive device) unit cell (Fig. 4k) enables the weight tuning of the device by means of applying positive or negative pulses on the transistor. The 2T+1TriR unit cell[59] (Fig. 4l) can also achieve linear and symmetric conductance tuning by modulating the gate voltage through PFET path or NFET path with identical pulses.

To overcome the limited precision of NVM devices, a general approach is using multiple devices with proportional significance to represent a multi-bit number[60,61]. The output result is generated by merging the results of adjacent columns of different significances with digital shift and add circuits. However, in this approach, the readout circuits and digital shift and add circuits need to be customized, and the extended bit width of weights lacks reconfigurability. To configure the weight bit width, the weights can be represented by multiple arrays with varied significance[62]. The results of multiple arrays are merged with a bit-width-reconfigured joint module. In this scenario, the merging operation need take the shifted device variation into consideration. Differential NVM devices in the same row[60] or column[63] can be used to implement one signed bit of weight. The latter one could reduce the voltage drop along the output line but increase the design complexity, especially for the input offset of the ADC circuits.

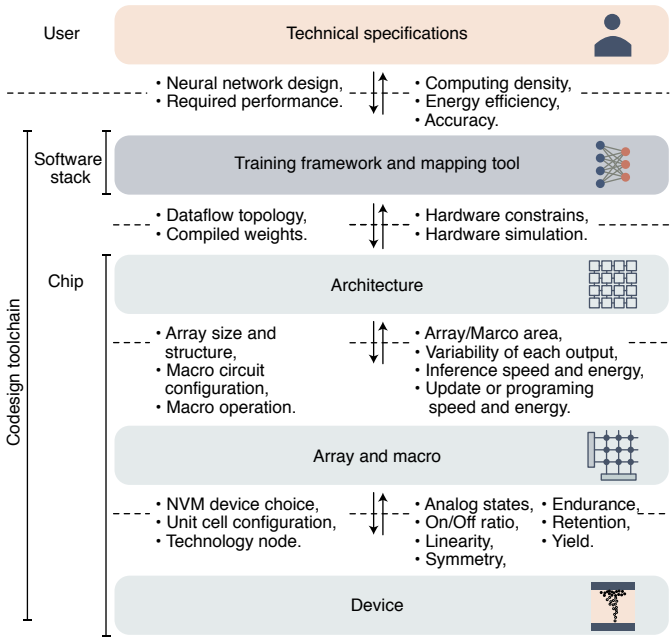

Fig. 5 | EDA tool chain. Hierarchical tools with the corresponding interfaces and the co-design flow for the NVM-based neuro-inspired chips. The user first feeds the specifications of the intelligent workloads and chip to the EDA tool. Within the EDA tool chain, the software stack is used to train the neural networks with/without the constraints of device and array, and then compile the optimized weights into memory-centric intermediate representations to chip optimization and evaluation tools. The evaluation tool of chip should cover the three-level designs and configurations including architecture, array/macro and device. These device models should integrate most of the device non-idealities and be compact and accurate enough for industry requirements. Finally, when the EDA tool finishes the optimization and evaluation of chip after the co-design exploration, the simulated on-chip computing accuracy and chip-specified performance will be presented to the user.

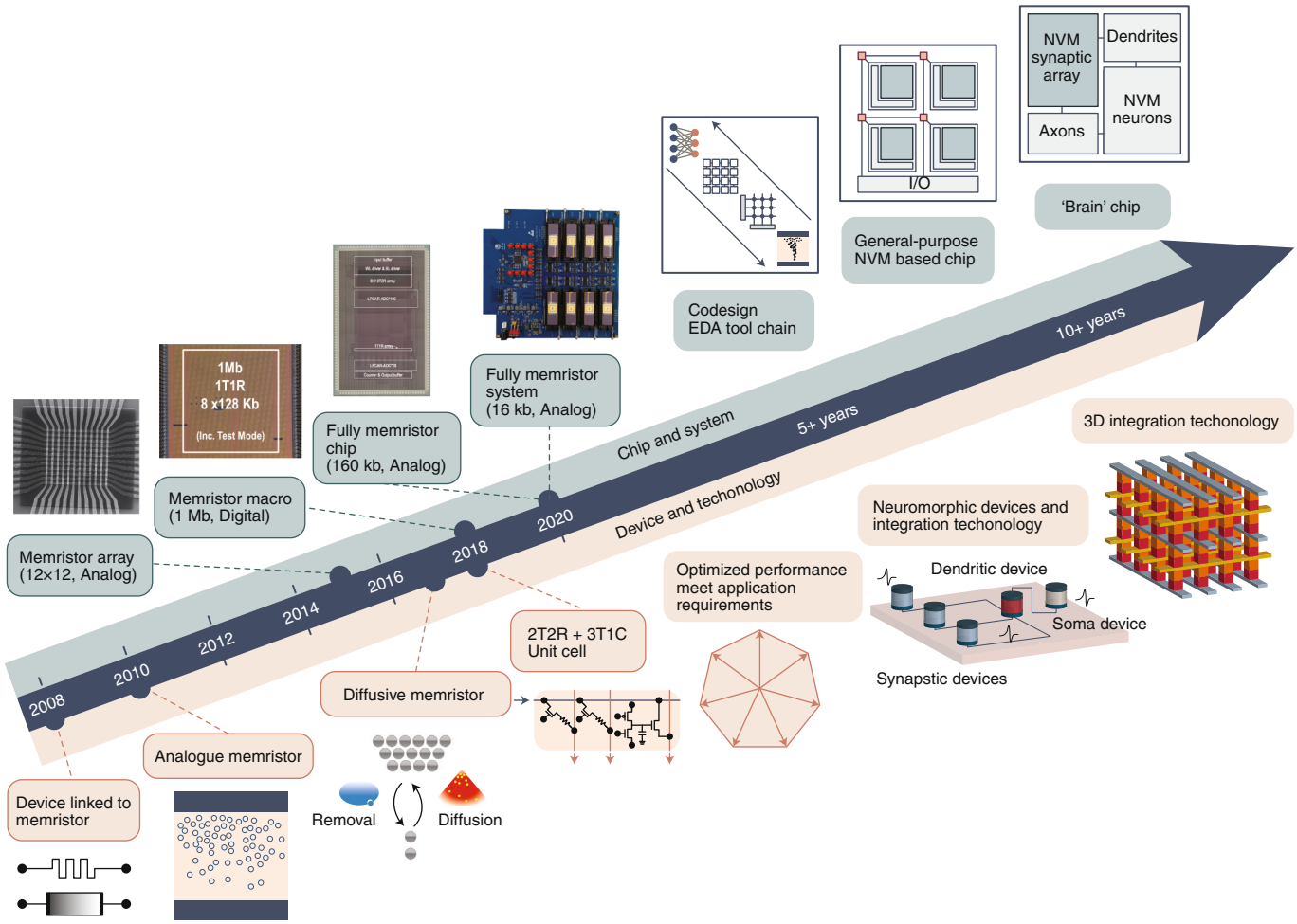

Fig. 6 | Roadmap. The past milestones[11,12,17,21,23,63,71,105] and future prospects are illustrated. For the future devices and technology, the NVM device should be optimized to meet the application requirements. More neuromorphic devices are exploited to demonstrate a wide range of bio-plausible behaviours, and the integration technology should be developed to demonstrate a monolithic system. The 3D integration of NVM devices is a more advanced approach to design future 3D large-scale neuro-inspired computing chips, which have the potential to exceed the capabilities of the human brain. For the chip and system side, the co-design EDA toolchain from device to algorithm is a foundation to design efficient neuro-inspired computing chips. The general-purpose neuro-inspired computing chip will have a complete architectural design and integrates sufficient on-chip synaptic memory devices to support state-of-the-art applications. In addition, the general-purpose chip should realize an improvement of several orders of magnitude in energy efficiency while achieving the same computing accuracy as the conventional chips. The future 'brain' chip is expected to adopt the more intelligent working mechanisms of the brain to realize high-level intelligence. Images in the figure adapted or reproduced with permission from: 'Device linked to memristor', ref. [11], 'Memristor array', ref. [21], '2T2R+3T1C Unit cell', ref. [23], 'Fully memristor system', ref. [71], 'Diffusive memristor', ref. [105], Springer Nature Ltd; 'Memristor macro', ref. [17], 'Fully memristor chip', ref. [63], IEEE; and 'Analogue memristor', ref. [12], ACS.

Array and macro. Analogue computing for vector–matrix multiplication (VMM) operations exploits the fact that a matrix can be mapped onto an NVM crossbar array. The weight matrix is stored in the form of the NVM conductance, and the inputs are mapped to input voltage amplitudes or pulse numbers. In a passive crossbar array[21] (consisting of 1R cells), as shown in Fig. 4m, or a parallel pseudo-crossbar array[22] (consisting of 1T1R cells), as shown in Fig. 4n, when input voltages are applied to the row electrodes of the array, the output currents of the column electrodes that are generated via Ohm's law and Kirchhoff's law represent the arithmetic results of the VMM operation. To overcome the problems of sneak paths and variability, a row-by-row pseudo-crossbar array[64] (Fig. 4o) sequentially adds partial sums of multiplications in the digital domain with the help of peripheral circuits. Considering the compatibility issue between the storage and computing mechanisms, a memory-like array structure[17] (Fig. 4p), in which the input signals determine the on and off states of the selective devices, is more attractive from the foundry manufacturing perspective. Row-by-row readout is implemented in most NVM-based neuro-inspired chip designs that have been developed to date because the computing accuracy that can be achieved with this method is much better than that of the fully parallel readout method. However, row-by-row readout will slow down the computing speed when the number of rows is increased.

At the macro level of design, the parasitic effects of the NVM array and the overhead of the peripheral circuits will limit the computing performance. The parasitic resistance inevitably causes a voltage drop along the interconnect wires, which decreases the analogue readout currents. The existence of sneak paths in the crossbar causes the current reduction to be superlinearly dependent on the array size and interconnect resistance[65]. This is one reason why such arrays are difficult to scale up to the megabit level. Although applying heuristic compensation by adding scaling factors to the output terminal can mitigate this effect for some specific neural networks[66],

a universal compensation method for various types of neural networks is still lacking. In addition, a large number of conversion circuit modules are utilized to transform signals between the analogue and digital domains. The latency and energy consumption of these peripheral circuits play a critical role in the overall system performance on inference tasks. Low-precision computing approaches such as binary inputs and outputs can simplify the design of conversion circuits and decrease the area and power overhead[67]. Fully analogue computing design is another way to eliminate conversion circuits[68]. However, computation within the analogue domain across multiple macros or multiple neural network layers still requires further investigation from the circuit/architecture perspective.

Architecture. A generic architecture comprises many distributed synapse–neuron cores (Fig. 4q), which are connected with a reconfigurable on-chip interconnection fabric (Fig. 4r). Generally, the synapse–neuron core is comprised of NVM array, DACs, ADCs, drivers, buffer, special function units (SFU), and controller. To increase the system throughput, a data pipeline and fine-grain parallelism should be introduced into the computing architecture[69,70]. The inter-frame parallelism could also be achieved by replicating the convolutional kernels where each kernel handle one frame[71]. However, the execution times of different layers are usually unbalanced and the pipeline would be blocked in the computing-intensive layers. With inter-layer parallelism and weight replication, a continuous dataflow can be achieved by balancing the execution time of different layers[70]. To enable a flexible and efficient pipeline, a high-bandwidth interconnect fabric between synapse–neuron cores, such as shared memory bus[70], network-on-chip[69,72] and switch matrix[73], needs to be employed to reduce the communication bound. In addition, sparse in-memory computing architecture could reduce the inefficient computation by exploiting the sparsity of neural network[74].

Since the learning phase of a neural network involves weight updates and intricate data dependencies, most existing neuro-inspired architectures support only the inference phase. Compared to the implementation of DNNs, the progress towards implementing SNNs on NVM-based neuro-inspired architectures has thus far been limited.

Learning and mapping algorithms. To access the full potential of analogue NVM computing, co-design of hardware and algorithms, including learning and mapping algorithms, is important. Generally, the learning algorithms for NVM-based neuro-inspired chips can be roughly divided into three categories: off-chip learning[75], on-chip learning from scratch[38], and hybrid learning[71]. The off-chip learning trains the weights using standard learning algorithm in an external computer with or without considering non-idealities of devices and circuits, such as conductance variation, voltage drop. This approach has the smallest hardware overhead but shows poor tolerance of non-idealities. In addition, the programming circuits for off-chip learning must be carefully designed to overcome the problems of conductance drift and retention. On-chip learning approaches could enhance the environmental adaptability of the system and the error tolerance of device and circuit. If on-chip learning from scratch, the hardware would consume too much energy and time to train the neural network after weight initialization. To strike a balance between the advantages and disadvantages of the above learning algorithms, new learning strategies involving fine tuning a chip that has been preprogramed with off-chip trained weights are proposed. For example, a hybrid training algorithm, which only fine tunes a small portion of the neural network, can increase the immunity to device non-idealities and significantly reduce the overhead of learning energy and time.

The on-chip learning phase requires high-precision arithmetic and frequent weight updates according to the learning strategy. However, the endurance of NVM devices is usually limited and the learning precision is influenced by the fluctuations and nonlinearity during device switching. Hence, for neuro-inspired computing chips, the corresponding learning algorithms should be modified to adapt to the non-idealities of these chips as well as their simplified peripheral circuits for learning. Several efficient learning algorithms, such as those with localized unsupervised learning rules, for example, STDP[76], or simplified gradient-based algorithms, for example, Manhattan rule algorithms[77] or sign-based backpropagation[78], are more suitable than others for on-chip learning. In addition, reinforcement learning or self-learning could be more useful for neuro-inspired computing chips. To adapt learning algorithms to practical neuro-inspired chips, both the accuracy of the algorithms and the costs of learning and programming circuits need to be considered.

The mapping algorithm includes three stages: compiling, weight placement and programming. In the compiling stage, the memory-centric intermediate representation (Fig. 4s) is synthesized from a neural network task (Fig. 4t) by satisfying both on-chip constraints and optimization target. For example, the execution latency of a specified application could be optimized by increasing the crossbar resource allocation of the computing-intensive layers by spatially replicating these layers under the constraints of restricted on-chip memory and routing resources. In the weight placement stage, the intermediate representation is scheduled onto the physical neuron-synapse cores and transformed into physical device computing units. In the programming stage, the conductance of device is written into the target conductance window within the preset error margin by programming algorithms, such as close-loop programming algorithms[79]. To resolve the conductance drift or relaxation of NVM devices, the programming precision of neuro-inspired computing chips can be improved by periodically refreshing the devices.

EDA tool chain. The major obstacle of cross-layer co-design is the lack of an EDA tool chain. The EDA tool chain can facilitate the design space exploration of a large-scale neuro-inspired computing chip, while considering the compromises among the non-idealities of the NVM devices, the peripheral circuits in the synapse–neuron cores, and the co-design of the architectures and algorithms.

Although such a universal EDA tool chain is lacking, several simulator platforms[72,80] for the algorithm and chip co-design have been proposed in recent years that address this need to some extent. In addition, architecture-level simulator platforms support the system-level design of neuro-inspired computing chips; however, they offer limited considerations regarding device non-idealities. Circuit-level macro-model simulator platforms can be used to estimate area, latency, and energy of synapse–neuron cores for on-chip learning or inference with realistic device properties[81]. Based on the above-mentioned challenges of device, array/macro, architecture and algorithm, the hardware-software co-design flow should be supported in the future EDA tool chain in a bottom-up and top-down fashion for large-scale neuro-inspired computing chips (Fig. 5). Besides, this tool chain strongly demands an accurate NVM device model that can meet industry requirements like the widely used Berkeley short-channel IGFET model (BSIM) family for MOSFET.

Roadmap

As CMOS technology approaches its physical limits, NVM-based neuro-inspired computing chips offer a promising route forward and are the focus of increasing research effort in terms of device engineering and chip prototyping. Based on the typical reported devices (Fig. 6), future device research should focus on implementing analogue NVMs with improved performance, and exploring more bio-plausible characteristics. On the chip and system side, and benefitting from device development efforts, on-chip capacity

of synapses and neurons has increased considerably in recent years (Fig. 6); and further full chip implementations of DNNs and more brain-like algorithms are expected to use more complete architectural designs.

Ultimately, and through integration with high-density 3D NVM, future neuro-inspired computing chips should be more efficient in terms of area and should reach a much larger integration scale. One key challenge in 3D NVM devices is the realization of high-performance selector devices to eliminate sneak current paths. With 3D NVM, the architecture and datapath of monolithic 3D large-scale neuro-inspired computing chip should be carefully designed to accommodate the high-bandwidth and fine-grained interconnection across multiple functional layers[82]. Furthermore, device reliability remains a key issue and future applications of NVM-based neuro-inspired computing chips should focus on situations that can tolerate the non-ideal characteristics and limited reliability of the devices.

It is also worth noting other approaches, including quantum and unconventional computing, that could also have an impact in the future development of AI-focused hardware. Quantum computing approaches, such as the quantum Boltzmann machine[83], could potentially exponentially speed up machine learning techniques compared with the best classical computers. However, to make quantum computers a practical reality much more work is required, particularly in terms of robust large-scale system design[84]. Alternatively, unconventional computing approaches could help accelerate many specific computing tasks, such as one-step equation solvers using negative-feedback-connected memristor arrays[85] and combinatorial optimization problem solvers using intrinsic noise[86]. Furthermore, there are several neuro-inspired techniques that could be employed in future chips, including neuro-fuzzy networks[87], probabilistic networks[88] and hierarchical temporal memory[89].

Neuro-inspired computing chips already offer capabilities in certain tasks that extend beyond the capabilities of conventional chips, and various bio-inspired technologies have been adopted in recent designs in order to find new ways to build efficient and robust systems. Although neuro-inspired computing chips do not yet meet the ideal requirements of our proposed benchmarking metrics, various encouraging initiatives are already underway in the field — at the device, architecture and algorithm level — and the chances of reaching them look promising.

Received: 31 March 2019; Accepted: 2 June 2020;
Published online: 21 July 2020

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
3. Hoi-Jun, Y. Intelligence on Silicon: From Deep-Neural-Network Accelerators to Brain Mimicking AI-SoCs. In *2019 IEEE International Solid - State Circuits Conference - (ISSCC)* 20–26 (IEEE, 2019).
4. Roy, K., Jaiswal, A. & Panda, P. Towards spike-based machine intelligence with neuromorphic computing. *Nature* **575**, 607–617 (2019).
5. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* https://doi.org/10.1038/s41565-020-0655-z (2020).
6. Kandel, E. R. et al. *Principles of Neural Science* vol. 4 (McGraw-hill New York, 2000).
7. Rosenblatt, F. *The Perceptron, a Perceiving and Recognizing Automaton Project Para* (Cornell Aeronautical Laboratory, 1957).
8. Widrow, B., Pierce, W. H. & Angell, J. B. Birth, Life, and Death in Microelectronic Systems. *IRE Trans. Mil. Electron.* **MIL–5**, 191–201 (1961).
9. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78**, 1629–1636 (1990).
10. Jackel, L. D. Artificial neural networks for computing. *J. Vac. Sci. Technol. B* **4**, 61 (1986).
11. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453**, 80–83 (2008).
12. Jo, S. H. et al. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10**, 1297–1301 (2010).
13. Chen, Y.-H., Krishna, T., Emer, J. & Sze, V. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. In *2016 IEEE Int. Solid-State Circuits Conference (ISSCC)* 262–263 (IEEE, 2016).
14. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proc. 44th Annual Int. Symposium on Computer Architecture* https://doi.org/10.1145/3140659.3080246 (ACM, 2017).
15. Yu, S. Neuro-inspired computing with emerging nonvolatile memorys. *Proc. IEEE* **106**, 260–285 (2018).
16. Biswas, A. & Chandrakasan, A. P. Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications. In *2018 IEEE Int. Solid - State Circuits Conference - (ISSCC)* 488–490 (IEEE, 2018).
17. Chen, W.-H. et al. A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors. In *2018 IEEE Int. Solid - State Circuits Conference - (ISSCC)* 494–496 (IEEE, 2018).
18. Sim, J. et al. A 1.42TOPS/W deep convolutional neural network recognition processor for intelligent IoE systems. In *2016 IEEE Int. Solid-State Circuits Conference (ISSCC)* 264–265 (IEEE, 2016).
19. Xue, C.-X. et al. A 1Mb multibit reram computing-in-memory macro with 14.6ns parallel MAC computing time for CNNBased AI edge processors. In 2019 *IEEE Int. Solid - State Circuits Conference - (ISSCC)* 388–389 (IEEE, 2019).
20. Mochida, R. et al. A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture. In *2018 IEEE Symposium on VLSI Technology* 175–176 (IEEE, 2018).
21. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
22. Yao, P. et al. Face classification using electronic synapses. *Nat. Commun.* **8**, 15199 (2017).
23. Ambrogio, S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60–67 (2018).
24. Li, C. et al. Long short-term memory networks in memristor crossbar arrays. *Nat. Mach. Intell.* **1**, 49–57 (2019).
25. Zhang, J., Wang, Z. & Verma, N. In-memory computation of a machine-learning classifier in a standard 6T SRAM array. *IEEE J. Solid-State Circ.* **52**, 915–924 (2017).
26. Srinivasa, S. et al. Monolithic 3D$^+$ -IC based reconfigurable compute-in-memory SRAM macro. In *2019 Symposium on VLSI Technology* T32–T33 (IEEE, 2019).
27. Li, S. et al. DRISA: a DRAM-based reconfigurable in-situ accelerator. In *Proc. 50th Annual IEEE/ACM Int. Symposium on Microarchitecture - MICRO-50 '17* 288–301 (ACM, 2017).
28. Guo, X. et al. Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology. In *2017 IEEE Int. Electron Devices Meeting (IEDM)* 6.5.1-6.5.4 (IEEE, 2017).
29. Cai, F. et al. A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations. *Nat. Electron.* **2**, 290–299 (2019).
30. Wu, H., Yao, P., Gao, B. & Qian, H. Multiplication on the edge. *Nat. Electron.* **1**, 8–9 (2018).
31. Schmitt, S. et al. Neuromorphic hardware in the loop: training a deep spiking network on the BrainScaleS wafer-scale system. In *2017 Int. Joint Conference on Neural Networks (IJCNN)* 2227–2234 (IEEE, 2017).
32. Vaquer-Sunyer, R. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
33. Davies, M. et al. Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99 (2018).
34. Esser, S. K. et al. Convolutional networks for fast, energy-efficient neuromorphic computing. *Proc. Natl Acad. Sci. USA* **113**, 11441–11446 (2016).
35. Wang, W. et al. Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses. *Sci. Adv.* **4**, eaat4752 (2018).
36. Gao, L., Chen, P.-Y. & Yu, S. NbOx based oscillation neuron for neuromorphic computing. *Appl. Phys. Lett.* **111**, 103503 (2017).
37. Xu, X. et al. Scaling for edge inference of deep neural networks. *Nat. Electron.* **1**, 216–222 (2018).
38. Li, C. et al. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. *Nat. Commun.* **9**, 2385 (2018).
39. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comp. Vis.* **115**, 211–252 (2015).
40. Lin, T.-Y. et al. Microsoft COCO: common objects in context. In *Computer Vision – ECCV 2014* (eds. Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) 740–755 (Springer, 2014).

41. Jeongwoo, P., Juyun, L. & Dongsuk, J. A 65nm 236.5nJ/classification neuromorphic processor with 7.5% energy overhead on-chip learning using direct spike-only feedback. In *2019 IEEE Int. Solid - State Circuits Conference - (ISSCC)* 140–141 (IEEE, 2019).
42. Pi, S. et al. Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension. *Nat. Nanotechnol.* (2018).
43. Lin, P. et al. Three-dimensional memristor circuits as complex neural networks. *Nat. Electron.* **3**, 225–232 (2020).
44. Prezioso, M. et al. Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits. *Nat. Commun.* **9**, (2018).
45. Jacob, B. et al. quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2704–2713 (IEEE, 2018).
46. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
47. Wang, Z. et al. Engineering incremental resistive switching in TaO_x based memristors for brain-inspired computing. *Nanoscale* **8**, 14015–14022 (2016).
48. Schneider, M. L. et al. Ultralow power artificial synapses using nanotextured magnetic Josephson junctions. *Sci. Adv.* **4**, e1701329 (2018).
49. Jerry, M. et al. Ferroelectric FET analog synapse for acceleration of deep neural network training. In *2017 IEEE Int. Electron Devices Meeting (IEDM)* 6.2.1-6.2.4 (IEEE, 2017).
50. Tang, J. et al. ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing. in *2018 IEEE Int. Electron Devices Meeting (IEDM)* 13.1.1-13.1.4 (IEEE, 2018).
51. Zhao, M. et al. Characterizing Endurance Degradation of Incremental Switching in Analog RRAM for Neuromorphic Systems. in *2018 IEEE Int. Electron Devices Meeting (IEDM)* 20.2.1–20.2.4 (IEEE, 2018).
52. Zhao, M. et al. Investigation of statistical retention of filamentary analog RRAM for neuromophic computing. In *2017 IEEE Int. Electron Devices Meeting (IEDM)* 39.4.1–39.4.4 (IEEE, 2017).
53. Wu, H. et al. Device and circuit optimization of RRAM for neuromorphic computing. In *2017 IEEE Int. Electron Devices Meeting (IEDM)* 11.5.1–11.5.4 (IEEE, 2017).
54. Wang, Z. et al. Reinforcement learning with analogue memristor arrays. *Nat. Electron.* **2**, 115–124 (2019).
55. Sheridan, P. M. et al. Sparse coding with memristor networks. *Nat. Nanotechnol.* **12**, 784–789 (2017).
56. Li, C. et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).
57. Kim, S. et al. NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning. In *2015 IEEE Int. Electron Devices Meeting (IEDM)* 17.1.1–17.1.4 (IEEE, 2015).
58. Jerry, M. et al. A ferroelectric field effect transistor based synaptic weight cell. *J. Phys. D.* **51**, 434001 (2018).
59. Sun, X., Wang, P., Ni, K., Datta, S. & Yu, S. Exploiting hybrid precision for training and inference: a 2T-1FeFET based analog synaptic weight cell. In *2018 IEEE Int. Electron Devices Meeting (IEDM)* 4 (2018).
60. Chi, P. et al. PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. In *2016 ACM/IEEE 43rd Annual Int. Symposium on Computer Architecture (ISCA)* 27–39 (IEEE, 2016).
61. Boybat, I. et al. Neuromorphic computing with multi-memristive synapses. *Nat. Commun.* **9**, (2018).
62. Zhu, Z. et al. A configurable multi-precision CNN computing framework based on single bit RRAM. In *Proc. 56th Annual Design Automation Conference 2019 on - DAC '19* 0738-100X (ACM, 2019).
63. Liu, Q. et al. A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing. In *2020 IEEE Int. Solid - State Circuits Conference - (ISSCC)* 500–502 (IEEE, 2020).
64. Sun, X. et al. XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks. In *2018 Design, Automation & Test in Europe Conference & Exhibi*tion (DATE) 1423–1428 (IEEE, 2018).
65. Cassuto, Y., Kvatinsky, S. & Yaakobi, E. Sneak-path constraints in memristor crossbar arrays. In *2013 IEEE Int. Symposium on Information Theory* 156–160 (IEEE, 2013).
66. Jeong, Y., Zidan, M. A. & Lu, W. D. Parasitic effect analysis in memristor-array-based neuromorphic systems. *IEEE Trans. Nanotechnol.* **17**, 184–193 (2018).
67. Yu, S. et al. Binary neural network with 16 Mb RRAM macro chip for classification and online training. In *2016 IEEE Int. Electron Devices Meeting (IEDM)* 16.2.1-16.2.4 (IEEE, 2016).
68. Wang, Z. et al. Fully memristive neural networks for pattern classification with unsupervised learning. *Nat. Electron.* **1**, 137–145 (2018).
69. Shafiee, A. et al. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. in *2016 ACM/IEEE 43rd Annual Int. Symposium on Computer Architecture (ISCA)* 14–26 (IEEE, 2016)
70. Song, L., Qian, X., Li, H. & Chen, Y. PipeLayer: A pipelined ReRAM-based accelerator for deep learning. In *2017 IEEE Int. Symposium on High Performance Computer Architecture (HPCA)* 541–552 (IEEE, 2017).
71. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
72. Ankit, A. et al. PUMA: a programmable ultra-efficient memristor-based accelerator for machine learning inference. In *Proc. Twenty-Fourth Int. Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS* '19 715–731 (ACM, 2019).
73. Ji, Y. et al. FPSA: a full system stack solution for reconfigurable ReRAM-based NN accelerator architecture. In *Proc. Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS '19* 733–747 (ACM, 2019).
74. Yang, T.-H. et al. Sparse ReRAM engine: joint exploration of activation and weight sparsity in compressed neural networks. In *Proc. 46th International Symposium on Computer Architecture - ISCA '19* 236–249 (ACM, 2019).
75. Hu, M. et al. Memristor-based analog computation and neural network classification with a dot product engine. *Adv. Mater.* **30**, 1705914 (2018).
76. Song, S., Miller, K. D. & Abbott, L. F. Competitive Hebbian learning through spike-timing-dependent synapticplasticity. *Nat. Neurosci.* **3**, 919–926 (2000).
77. Zamanidoost, E., Bayat, F. M., Strukov, D. & Kataeva, I. Manhattan rule training for memristive crossbar circuit pattern classifiers. In *2015 IEEE 9th Int. Symposium on Intelligent Signal Processing (WISP) Proc.* https://doi.org/10.1109/WISP.2015.7139171 (IEEE, 2015).
78. Zhang, Q. et al. Sign backpropagation: An on-chip learning algorithm for analog RRAM neuromorphic computing systems. *Neural Netw.* **108**, 217–223 (2018).
79. Hu, M. et al. Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication. In *Proc. 53rd Annual Design Automation Conference on - DAC '16* 1–6 (ACM, 2016).
80. Zhang, W. et al. Design guidelines of RRAM based neural-processing-unit: a joint device-circuit-algorithm analysis. In *Proc. 56th Annual Design Automation Conference 2019 on - DAC '19* 1–6 (ACM, 2019).
81. Chen, P.-Y., Peng, X. & Yu, S. NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Trans. Computer-Aided Design Int. Circ. Syst.* 1–1 (2018).
82. Sabry Aly, M. M. et al. The N3XT approach to energy-efficient abundant-data computing. *Proc. IEEE* **107**, 19–48 (2019).
83. Amin, M. H., Andriyash, E., Rolfe, J., Kulchytskyy, B. & Melko, R. Quantum Boltzmann machine. *Phys. Rev. X* **8**, 021050 (2018).
84. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
85. Sun, Z. et al. Solving matrix equations in one step with cross-point resistive arrays. *Proc. Natl Acad. Sci. USA* **116**, 4123–4128 (2019).
86. Mahmoodi, M. R., Prezioso, M. & Strukov, D. B. Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization. *Nat. Commun.* **10**, 5113 (2019).
87. Merrikh-Bayat, F. & Shouraki, S. B. Memristive neuro-fuzzy system. *IEEE Trans. Cybern.* **43**, 269–285 (2013).
88. Serb, A. et al. Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses. *Nat. Commun.* **7**, 12611 (2016).
89. Krestinskaya, O., Dolzhikova, I. & James, A. P. Hierarchical temporal memory using memristor networks: a survey. *IEEE Trans. Emerg. Top. Comput. Intell.* **2**, 380–395 (2018).
90. Schemmel, J. et al. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Proc. 2010 IEEE International Symposium on Circuits and Systems* 1947–1950 (IEEE, 2010).
91. Seo, J. et al. A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons. In *2011 IEEE Custom Integrated Circuits Conference (CICC)* https://doi.org/10.1109/CICC.2011.6055293 (IEEE, 2011).
92. Kim, J. K., Knag, P., Chen, T. & Zhang, Z. A 640M pixel/s 3.65mW sparse event-driven neuromorphic object recognition processor with on-chip learning. In *2015 Symposium on VLSI Circuits (VLSI Circuits)* C50–C51 (IEEE, 2015).
93. Kang, M., Gonugondla, S. K., Patil, A. & Shanbhag, N. R. A Multi-Functional In-Memory Inference Processor Using a Standard 6T SRAM Array. *IEEE J. Solid-State Circuits* **53**, 642–655 (2018).
94. Frenkel, C., Lefebvre, M., Legat, J. & Bol, D. A 0.086-mm^2 12.7-pJ/SOP 64k-Synapse 256-Neuron Online-Learning Digital Spiking Neuromorphic Processor in 28-nm CMOS. *IEEE Trans. Biomed. Circ. Syst.* **13**, 145–158 (2019).
95. Gonugondla, S. K., Kang, M. & Shanbhag, N. A 42pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training. In *2018 IEEE Int. Solid - State Circuits Conference - (ISSCC)* 490–492 (IEEE, 2018).
96. Pei, J. et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106–111 (2019).

97. Benjamin, B. V. et al. Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations. *Proc. IEEE* **102**, 699–716 (2014).
98. Qiao, N. et al. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. *Front. Neurosci.* **9**, (2015).
99. Moradi, S., Qiao, N., Stefanini, F. & Indiveri, G. A scalable multi-core architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs). *IEEE Trans. Biomed. l Circ. Syst.* **12**, 106–122 (2018).
100. Khwa, W.-S. et al. A 65nm 4Kb algorithm-dependent computing-in -memory SRAM unit-macro with 2.3ns and 55.8TOPS/W fully parallel product-sum operation for binary DNN edge processors. In *2018 IEEE Int. Solid - State Circuits Conference - (ISSCC)* 496–498 (IEEE, 2018).
101. Si, X. et al. A Twin-8T SRAM computation-in-memory macro for multiple-Bit CNN-based machine learning. In *2019 IEEE Int. Solid - State Circuits Conference - (ISSCC)* 396–397 (IEEE, 2019).
102. Yang, J. et al. Sandwich-RAM: an energy-efficient in-memory BWN architecture with pulse-width modulation. In *2019 IEEE International Solid - State Circuits Conference - (ISSCC)* 394–395 (IEEE, 2019).
103. List of Intel microprocessors. *Wikipedia* (2020); https://en.wikipedia.org/wiki/List_of_Intel_microprocessors
104. List of Nvidia graphics processing units. *Wikipedia* (2020); https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units
105. Wang, Z. et al. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat. Mater.* **16**, 101–108 (2017).

Acknowledgements

This work is supported in part by the National Key R&D Program of China (2019YFB2205104), National Natural Science Foundation of China (61851404), Beijing Municipal Science and Technology Project (Z191100007519008), Huawei Project (YBN2019075015), and Beijing Tsinghua and Hsinchu Tsinghua Joint Project.

Author contributions

W.Z. and B.G. contributed to data collection, analysis and writing. All the authors contributed to discussion and revised the manuscript at all stages. H.W. contributed to work planning and development.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41928-020-0435-7.

Correspondence should be addressed to H.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.