STAP REVIEW

Systems and circuits for AI chips and their trends

Japanese Journal of Applied Physics **59**, 050502 (2020)
https://doi.org/10.35848/1347-4065/ab839f

STAP REVIEW
Physics and devices toward artificial intelligence

Systems and circuits for AI chips and their trends

Hiroshi Momose*, Tatsuya Kaneko*, and Tetsuya Asai*

Graduate School/Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060-0814, Japan

*E-mail: hiroshi.momose@ist.hokudai.ac.jp; kaneko.tatsuya.cy@ist.hokudai.ac.jp; asai@ist.hokudai.ac.jp

Received December 31, 2019; revised March 6, 2020; accepted March 25, 2020; published online April 21, 2020

The trends in AI chip development since 2013 can be divided into two categories of applications: server and edge. As for the server, recently, ASIC chips designed for learning functionalities have become mainstream, with competitively computational performance. However, the limitations of Moore's Law have started to impose themselves on this emerging type of AI chips, creating the need for new technological innovation. Meanwhile, regarding edge AI chips, research on data compression technology is advancing to lower power consumption while maintaining high performance. Further improvements have been made since 2017 in recognition accuracy in binary/ternary; moreover, there has been research on in-memory processing to configure 1 bit by combining memory and arithmetic element, where non-volatile memory can achieve higher performance and lower power consumption. With these backdrops, this paper summarizes the progress made to date in the field of AI chip technology while also identifying the future direction of next-generation technologies. © 2020 The Japan Society of Applied Physics

1. Introduction

Because of the numerous advancements made in the field of deep learning technology in 2010, various novel artificial intelligence (AI) chips have been developed and various products based on these chips have been released since 2013. In this study, we explain the technical content through several trends. We define an AI chip as a "chip specializing in realizing the operations of the brain" and targets deep learning chips with advanced abstraction as well as some neuromorphic chips.[1)]

As shown in Fig. 1, the evolution and emergence of AI chips can be broadly divided into two periods: one from 2013 to 2015, and the second from 2015 onward. The first period is the basic research (Basic) phase, which can be seen as the period during which the implementation method of the basic net models was explored, and the second period can be considered as the practical application research phase. This can be further divided into two phases, one in which high efficiency was pursued, and the other phase that is more diversified (Versatile) than the second phase.[2)] This phase is the period in which products with low power consumption and high performance were developed, particularly targeting edge applications such as mobile/IoT applications. This paper will discuss various specific chip types.

The Basic Phase has four activity flows. The first is the pursuit of the circuit configuration of the convolutional neural network (CNN). This is from the LeCun-led New York University and Purdue University, as well as from the TeraDeep group, which took over from Neuflow[3)] in 2011 to nn-X[4)] and later extended to ShiDianNao[5)] and Eyeriss.[6)] The second activity flow constitutes a wide range of activities related to the Chinese Academy of Science's (CAS) DianNao series, which took place over a relatively brief period of 18 months starting 2014. Extensive studies were conducted on four different chips: basic configuration (DianNao),[7)] learning (DaDianNao),[8)] vision (ShiDianNao),[5)] and multi-purpose (PuDianNao).[9)] The third activity flow is Google's Tensor Processing Unit (TPU),[10)] whose basic configuration and basic design were considered to have been realized between 2013 and 2014. This basic chip design is dedicated to server inferences. The three activity flows mentioned above are explained further in Chapter 2 (AI chip for servers) and Chapter 3 (edge AI chips). The fourth activity flow is IBM's TrueNorth,[1)] which gained attention in 2014. This is described in Chapter 4 as a type of neuromorphic chip that aims to acquire intelligence by imitating human brain cells.

During the second half of the 2nd phase (high efficiency), research on quantization and compression technology aiming at achieving smaller size and lower power consumption, with a focus on application to edge-type mobile devices worldwide, was actively pursued from mid-2015. Table I shows the specifications of each chip. As shown in Table I, many studies have been published, for examples, Eyeriss[6)] of the Massachusetts Institute of Technology (MIT), Energy Efficient Engine (EIE)[11)] of Stanford University, ENVISION 11 of KU Leuven University in 2017, and the deep neural processing unit (DNPU) of KAIST.[12)] The circuit technology for quantization and compression of these chips is discussed in Chapter 3 (edge AI chips), and chip characteristics are discussed in Chapter 5.

Furthermore, AI chips have diversified since the latter half of 2017. The authors classified this stage of technological development as Phase 3 (Versatile). For server applications, ASIC chips (for learning) include Google's TPU-v2/v3[13)] and preferred network's (PFN) MN-Core. Regarding edge chips, several IPs for incorporation into smartphones have now debuted, such as Kirin970 by Huawei and Apple's A12 chip. The pursuit for high efficiency continues, and KAIST's UNPU[14)] and the QUEST,[15)] developed at Hokkaido University, which adopted a circuit architecture optimized for lower bits (4 bits or fewer), were announced at academic conferences. Many other presentations from around 2016 have also focused on a form of in-memory processing as a lower-bit technology that is closer to the mechanisms of the synapses and neurons of brain cells, such as 6T-SRAM[16)] developed at Princeton University, 8T1C- SRAM,[17)] and BRein[18)] developed at Hokkaido University (BRein is a near-memory configuration). Although bits are reduced to binary (1 bit)/ternary (2 bit), deep learning algorithms to increase recognition accuracy have also been actively proposed (e.g. IBM's PACT-SAWB-fpsc technology[19)]). Similar to brain cells, these chips require analog sensing (readout) technology. Ultimately, it is more efficient and desirable to replace SRAM with NMV. Since 2018, non-volatile memory (NVM) technologies have been adopted on AI chips. Many announcements that incorporate NVM, such as Panasonic's

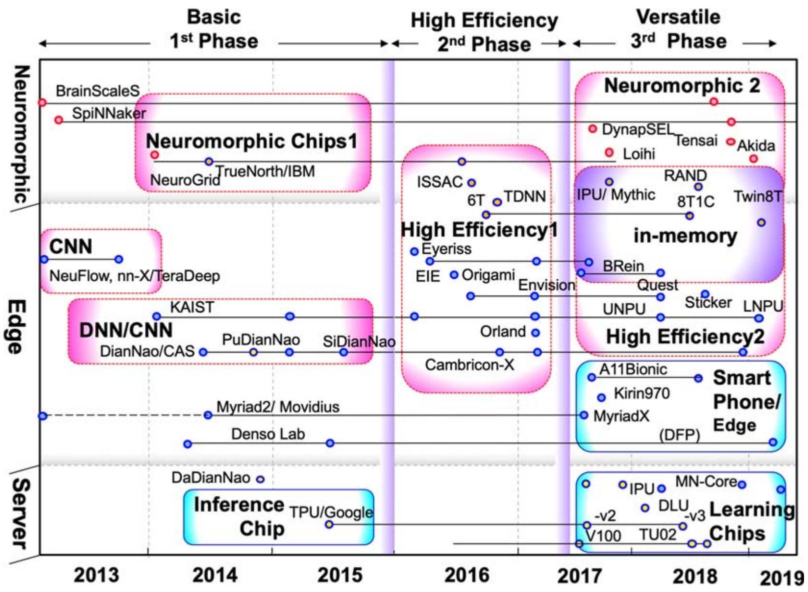


Fig. 1. (Color online) AI chip trend.

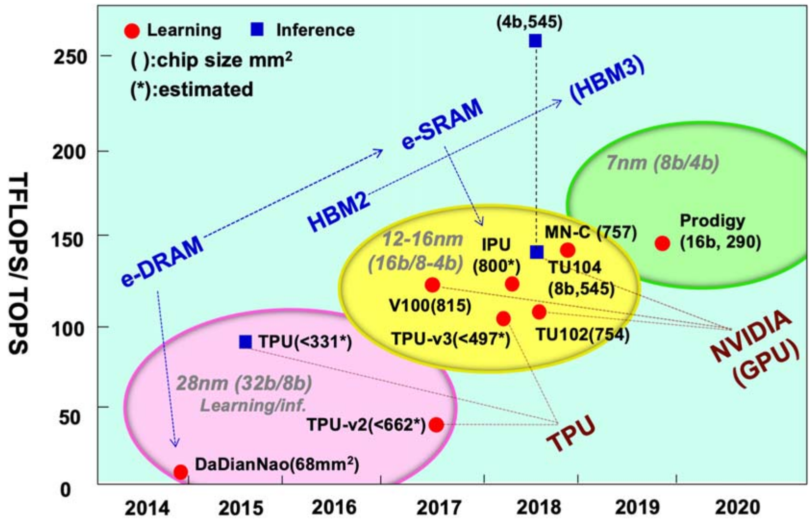
Table I. (Color online) Typical AI chip specifications.

Chip name			True-North		Da-DianNao#	TPU	Eyeriss	EIE	QUEST	RAND	8T1C SRAM	MN-Core *3	LNPU
Basic		Organaization	IBM		China CAS	Google	MIT/NVID	Stanford-U	Hokkaido-U	Panasonic	Princeton-U	PFN *3	KAIST
		Published Date	2014-8		2014-12	(2017-04)	2016-2	2016-5	2018-2	2018-6	2018-6	2018-12	2019-2
		Paper/Conference	Science		MICRO	ISCA	ISSCC	ISCA	ISSCC	VLSI-T	VLSI-C	Web	ISSCC
		Application	Edge		Server	Server	Mobile	Smart Phone	HE/MR-Edge	Edge/IoT	IoT	Server	Edge
		Technology(nm)	28		28	28	65	45	40	40	65	12	65
Net		Net Layer	High*1	Ave.	CONV/FC	CONV/FC	CONV	FC	CONV/FC	FC	CONV/FC	CONV/FC	CONV/FC
		Learning/Inference	Inference		Learning	Inference						Learning	
		MACs/chip	(268M)		4096	65,536	168	64	12,288	2M*9	(2,359,296)*10	(131,072)	768
Cir		Special circuit1	L-IF Type		e-DRAM	Systolic	Rec.*6	Compression	Bit-serial	A-ReRAM	A-SRAM	MAB	FGMP
		Special circuit2	Cross bar		BackProp	Batch*5	ZeroSkip	ZeroDetect	Log repre.	FNA	8T1C	Die to Die	ZeroDetect
Compression		bit precision (Fixed Point)	1b+2b(LUT)		16	8/(16)INT	16	5.4+4(ind) *7	1-log, 4-log	1	1(XNOR)	16	8 Float*8
	Quan-tization	Weight	LUT		---	---	---	LUT	Log	Aanlog(>2)	1	---	---
		Activation	1b(spike)		---	---	---	---	Log	1	1	---	FGMP(learn)
	Sparsity	Weight (Pruning)	connection *2		---	---	---	○	---	---	---	---	---
		Activation (Zero Skip)	Spike		---	---	○	○(Detect)	---	---	---	---	○(Balance)
		Lossless compression	---		---	---	○	(○)	---	---	---	---	---
Structure		Mmory	eSRAM		eDRAM	ext, DRAM	ext. DRAM	eSRAM	3DSRAM/TCI	ReRAM	eSRAM	ext. DRAM	ext.DRAM
		On chip	Near		Near				Near(3D stack)	InMemory	InMemory		
		Total onchip memory (MB)	54		37	(24(IM))	0.182	10.4	96MB(3D*4)	(24(IM))	0.259	ND	0.372
		onchip mem(Weight)	32		32		0/(0.74RF)	8.2	6.1	---	0.259	ND	---
		buffer (Activation)			4	24	0.108	0.128	0.9	---	ND	ND	---
		Memory BW (GB/sec)Weight	---		5,000	30	NA	---	29 (TCI)	---	---	PCIe-G3	---
		Frequency (MHz)	180Hz	20Hz	606	700	250	800	330	>10	100	(500)*3	50-200
		Chip area (mm^2)	430		67.7	300	12.3	40.8	122	2.71	17.6	757.0	16
Results		Power(mW) w/memory	140	72	10,000	75,000	278	590	2,700	9.9	14.34	(150,000)	376 (43.1*8)
		Peak Throughput(GOPS/s)	112	6.6	5,600	91,750	84	3,000	1960 /4b-log	660	9,438	131,000	1,090 (Inf)*8
		Power Efficiency(GOPS/W)	800	92	560	1,223	302	5,085	877 /4b-log	66,667	658,159	(873)*3	25,300/1,360*8
			*1,2			*5	*6	*7	*4	*9	*10	*3	*8

*1:High speed,L-IF=Leaky Integrated Fire, *2: Connection information, *3:Preferred Networks, Inc.,MAB=Matrix Arithmatic Block.500MHz/150W are estimates due to undisclosure.
*4:TCI(ThruChip interface), onchip memory 7MB, Log representation.
*5:Batch=Batch Processing, Systolic PE array, *6: Reco=Reconfigurable, RS(Row Stationary)/NOC, *7: ind=index for compressed sparse column format,
*8:FGMP=Fine-grained Mixed Precision (8bit Floating Point). At 0.78V/50MHz⇒43.1mW, 1.09TFOPS, 25.3/1.36 TFOPS/W=Inf/Learn.
*9: Resistive Analog Neuro Device, 4M Cell, Analog ReRAM with Cross bar in memory computing(analog sensing), FNA=Flexible network architecture. Several 10MHz
*10: A-SRAM(MixSignal), In memo Processing.

ReRAM's resistive analog neuro device (RAND) chip[20] and IBM's Projection PCM[21] technology, have received much attention. Moreover, Intel's Loihi[22] project, which aims to explore the practical application of spike-timing-dependent plasticity, a learning principle specific to spiking neural networks (SNNs) is gaining research attention for its ability to create a new trend in the future.

Chapter 2 describes AI chips for servers, Chapter 3 describes edge AI chips, and Chapter 4 describes neuromorphic chips. Chapter 4 discusses the recent configurations



CAS(DaDianNao), Graphcore(IPU), MN-C chip(PFN), Tachyum(Prodigy)

Fig. 2. (Color online) Throughput Trend for server AI chips.

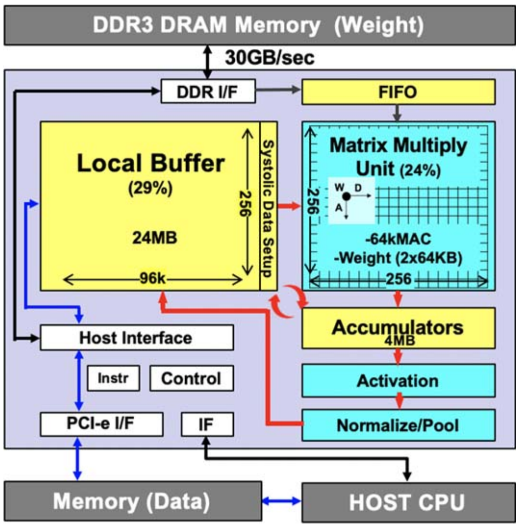


Fig. 3. (Color online) Block diagram of the TPU circuit.

made with respect to in-memory processing as a function designed to mimic brain operations.

2. Trends in AI chips for servers

In this chapter, we will first discuss the typical chip throughput [tera flops per second (TFLOPS) or tera operation per second (TOPS)] for each generation of AI chips used for full-scale inference and learning in servers. Next, the basic circuit system configuration of the AI chip will be described using Google's TPU[10] as a motif. An overview of the two major factors used to determine performance, computation, and the data transfer memory bandwidth will be outlined. After describing the DaDianNao DRAM-embedded chip,[8] announced in 2014, the current state of typical GPUs and ASICs will be explained.

2.1. AI computing

Before embarking on the main discussion, a list and specifications summarization of different AI chip technologies are provided in Table I. The different chips are arranged from left to right in the chronological order of their development year. However, for TPU, we defined the development year as 2015, as this was the year in which the technology was first used in data centers. Chip with an asterisk (#) in the chip name in Table I are evaluated by only CAD simulations, and no actual chips of this type have been manufactured. Only characteristic circuits are described in the circuit column. Next, the table describes the compression technology, numerical values of the LSI configuration, and performance specifications.

Figure 2 describes server chip throughput since 2014. The subscript is the code name or chip name. ■ indicates chips for inference-making, and • indicates chips for learning. Although learning and inference-making cannot be easily diverted to each other, large differences are usually indistinguishable in the basic configuration. The suffix number is the chip size and the unit used is mm^2. All chips are ASICs except for the NVIDIA GPUs. Typical examples include DRAM-embedded DaDianNao by CAS, TPU for inference by Google, and also TPU-v2, v3[12] for learning. The MN-Core was also recently developed by PFN in Japan. Furthermore, IPU chip incorporating large SRAM developed by Graphcore and Prodigy chip by Tachyum can also be included. Although their performance is improving, the design rule is already 7 nm in view, and the cutting-edge design is always used. The biggest task is the placement of a large amount of AI-processing elements (PEs) into a single chip; thus, the chip size of approximately 800 mm^2 is used, and competition with respect to performance centers on capability to achieve the maximum performance near the exposure area. This is the same for GPUs after V100 (Volta100) in 2017. It is easier to understand that GPUs after V100 are considered to be composed of AI parts and traditional GPU parts.

2.2. Basic circuit diagram

Google's TPU[10] is an accelerator that specializes in 8 bit integer and inference processing for servers. Around 2013–2014, Google anticipated exponential rise in demand in the future and accordingly started designing for the early introduction to data centers in 2015. As such, as the developers have stated, the mounting technology is currently orthodox and by-the-book, except the presence of a local buffer for batches. Using a design rule of 28 nm, the chip size is estimated to be 200–300 mm^2, with an operating frequency of 700 MHz and power consumption of 40 W (thermal design power is 75 W).

2.2.1. Circuit operation. Figure 3 shows a circuit block diagram of the TPU chip. Data (e.g. image data) is transferred from the host CPU/main memory from the lower side of the chip and stored in a local buffer. Meanwhile, the weights are loaded from the DDR3 DRAM (8 GB) on the upper side and are transferred into the chip and two-dimensionally deployed (position fixed: two weights are stored on each element) in a matrix multiply unit. The data and weight transfer rates are 10 GB s^{-1} and 30 GB s^{-1}, respectively. In the unit, PEs that handle 8 bit multiply-accumulate (MAC) operations are arranged in an array ($256 \times 256 = 64$k). A total of 256 bits of data are input from the left side of the unit, and the multiplications with the weights at the first column of the array are executed. This is a vector–vector multiplication of one of the so-called matrix (weight)-vector (input data) multiplication operations, which is usually suitable to process a fully connected layer used for multilayer perceptron (MLP). The data movements in the next step (one clock) are shown in the unit with two arrows. The data (D) are transferred in the x direction clock by clock in a horizontally systolic manner, and the multiplied results (A) are transferred in the $-y$ direction by one grid and are summed (accumulated) in the $-y$ direction through 256 clocks. As a result, one summed value is output from the lower side of the unit after 256 clocks. Similar processing is performed independently for each column, and 256 summed values are output. It should be noted that if 256 data bits of the one vector are simultaneously input from the left side of the unit, the summation of 256 multiplied results (A) in the one column cannot be easily performed. Therefore, a slightly complicated control operation is performed; when entering the unit, each element of the input vector is input with a one-clock delay each other in the $-y$ direction by the systolic data setup circuit. Such a data flow is called systolic data flow in the vertical direction. The 256 outputs of the unit are temporarily stored in an accumulator. When the number of input data is greater than 256, the results are summed in the accumulator.

Then, after performing activation, normalization, and pooling, calculations for the layer are completed; the data are transferred to the local buffer, and they become the input for the subsequent layer. In other words, the data go through one round to the next layer. The workload of the activation function and pooling processing is as small as a few percent of the total workload.[23] The calculation for the convolution layer can be performed by one-dimensionally arranging weights of the filters and input feature map pixels in the $-y$ direction following the above-mentioned procedure.

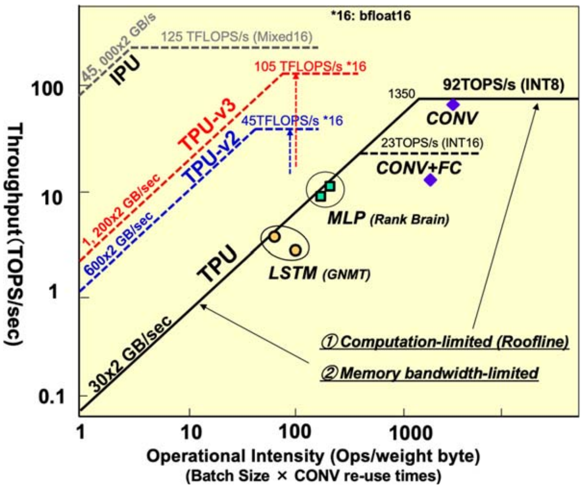


Fig. 4. (Color online) TPU Roofline model.

2.2.2. Roofline model. The performance limits for a normal processor can be described using a Roofline model, as depicted in Fig. 4. The applied workload (application: Alpha-Go, CNN, MLP, LSTM, etc.) is subject to the following five rate-determining conditions depending on the situation, and finally, the operation point is determined.

① Computation limitation (Roof): number of MACs = 65k, frequency = 700 MHz

② Weight transfer rate limitation: memory bandwidth (line) = 30 GB s^{-1}

③ Input data transfer rate limitation: bandwidth = 10 GB s^{-1}

④ Input data buffer size limitation: 29% of surface area

⑤ Response time limitation: bandwidth, batch size

Note that the typical limiting factors ① and ② are abbreviated in Fig. 4. The computation rate for ① represents the peak value of 92 TOPS s^{-1} (Tera Operations Per Second), and the measured value is approximately 90%.[9] The weight transfer rate of ② is determined by a bandwidth of 30 GB s^{-1}, and therefore, only 60 GOPS s^{-1} (Giga Operations Per Second) can be performed. This rate is very small; it is one thousandth of the computation rate. Therefore, the index of operational intensity on the horizontal axis of the figure is of significance as it represents the number of batches, or in the case of a convolutional (CONV) layer, it represents a value obtained by multiplying the number of batches by the number of reuses of the weights and its value is usually in several hundreds, which is equal to the number of nodes of the output layer (feature maps) in the CONV layer. The weight transfer rate limit is denoted as diagonal lines, as shown in Fig. 4. The intersection with the roof is 1350 batches. To increase the number of batches, a sufficient local buffer must be designed in advance. For example, when the size of batch is 100, the weights located in the unit are reused 100 times for 100 kinds of input data which have to be pre-stored in the local buffer. Then, the throughput is improved up to around 5 TOPS s^{-1} from less than 0.1 TOPS s^{-1} as shown in Fig. 4.

The method of selecting the number of batches is explained with respect to each application (Fig. 4). First, regarding the CONV layer (Alpha-Go), when the number of squares on the Go board is 19×19, and the zero-padding state and a number of filter striding are considered, the number of weight reuses (horizontal axis) is $19 \times 19 = 361$ for each layer. In this state, the throughput is limited by ②, and as such, eight batch processes are performed. Consequently, operational intensity in Fig. 4 becomes 2888 (361×8), which is laid on the computation limited roof, ①. Since CONV + FC (based on Inception V2) for image recognition is composed of convolutional layers and with four fully connected layers, it is presumed that this network is affected by the combined effects of ②, ③, and ④. Since the MLP and the LSTM are mainly composed of fully connected layers, the throughput is limited by memory bandwidth limit, ②, and the number of batches is limited to approximately 100.

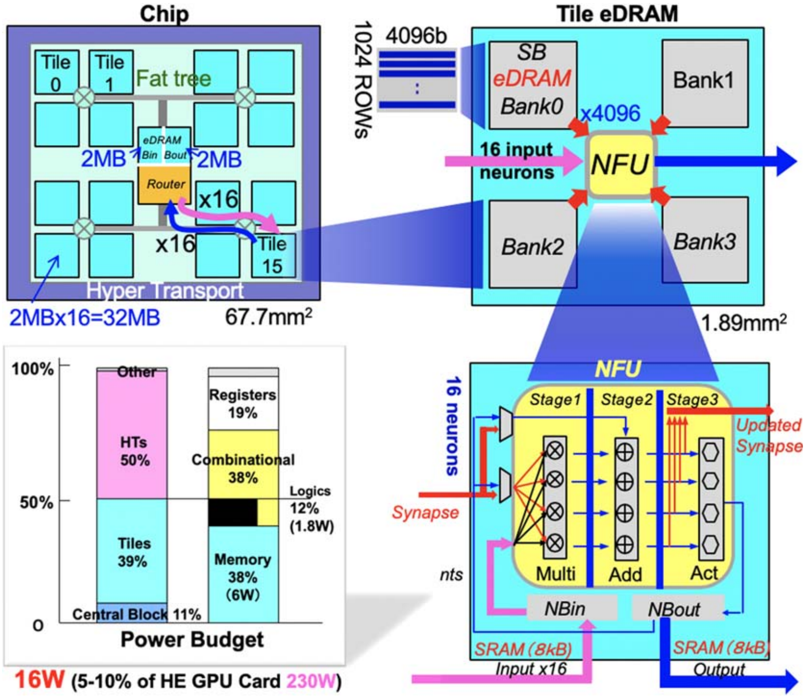


Fig. 5. (Color online) Block diagram of the DaDianNao.

This can be attributed to the buffer size limitation of ④, but as the buffer is reported to be in the range 20%–40%,[10] the remaining constraint is the response time limitation, ⑤. With LSTM, the application developer requested 7 ms, and as such, the number of batches was limited to approximately 100.

As described in the paper,[10] while numerous specific issues (learning, countermeasures to memory bandwidth limitation during operation of fully connected layers, sparsification, quantization, and architecture optimized or reconfigured for the convolutional layer, etc.) were left without implementing on the TPU, the issues were addressed after the later part of 2014. Subsequent sections of this paper will outline developments over these issues.

2.3. Memory-embedded AI chips (DaDianNao/IPU)

In particular, DaDianNao,[8] which was announced at the end of 2014 and attracted attention as a machine learning supercomputer, was developed with the aim of eliminating weight transfer limitations especially in the fully connected layers and for developing a chip suitable for learning tasks. This is the second of the four chips of CAS's DianNao series and is a DRAM-embedded chip, as shown in Fig. 5. A total of sixteen tiles are placed, and each tile contains four memory banks, at the center of which is an arithmetic unit (neural function unit). The operation unit has a built-in 16 × 16 matrix operation unit. The weight memory capacity is 32 MB for the entire chip, and the total input/output buffer capacity is 4 MB. A 16 bit fixed-point is used. The chip size is 68 mm^2 at 28 nm, which is approximately 1/4th of that of the TPU. In this configuration, the peak operation performance is 5.6 TOPS s^{-1} (16 bits), the weight transfer rate (memory bandwidth) is 5 TB s^{-1}, and the data transfer rate is 25.6 GB s^{-1}. The chip itself consumes 10 W and 6 W for external data transfer. If the size is approximately the same as the TPU, the weight transfer is 20 TB s^{-1}, which is approximately 700 times the TPU 30 GB s^{-1}. However, on the contrary, the peak throughput is approximately 45 TOPS s^{-1} (8 bits), which is approximately 1/2 of that for the TPU. The Roofline is reached in two batches. If the size is approximately the same as the TPU, 32 MB × 4 = 128 MB of memory can be included on-chip, and therefore, if used in a Google data center, the maximum weight will be approximately 100 MB, thereby allowing for on-chip implementation with a majority of the models.[10] If the number of MACs is further multiplied, LSTM can achieve 100 TOPS s^{-1}, 20–30 times the performance of the TPU.

Another feature of DaDianNao is that it employs a reconfigurable circuit architecture and can perform backpropagation for the learning operation. However, enough information regarding this backpropagation is not available. The chip used was designed only using CAD and was not planned to be manufactured. Thereafter, CAS established a venture capital firm and was conducting activities based on the Cambricon[24] (ISCA2016), launched in 2016.

Another chip that has recently attracted attention as a memory embedded AI chip is the IPU chip, which was launched by Graphcore in 2018. As shown in Fig. 6, 304 MB SRAMs are distributed on a huge 800 mm^2 class chip. The chip is configured such that the weight memories are dispersed and closely located with the MAC logics in the PE array in near-memory fashion (not in-memory fashion which will be described in Chapter 4). The chip achieves a throughput of 122 TFLOPS. It should be noted, however, that the weight memory achieves an effective bandwidth of 45 TB s^{-1}. Consequently, as shown in the upper left portion of Fig. 4, there is almost no bandwidth limitation, and a throughput in learning of 100 TFLOPS can be obtained even with a small number of batches. Graphcore has also verified that the smaller the number of batches, the higher the accuracy of learning.

Meanwhile, in Fig. 4, rooflines for two learning chips of TPU-v2 and v3 by Google are depicted. Although the

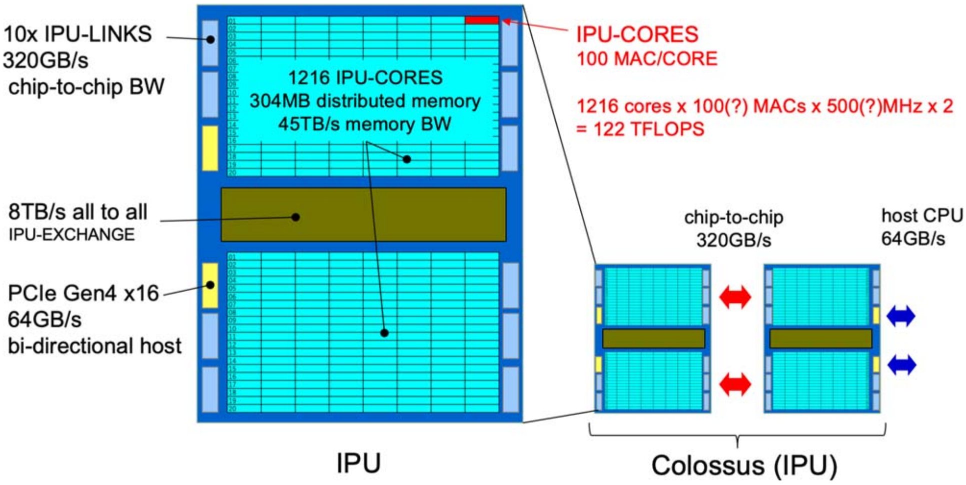


Fig. 6. (Color online) Block diagram of IPU-CORES.

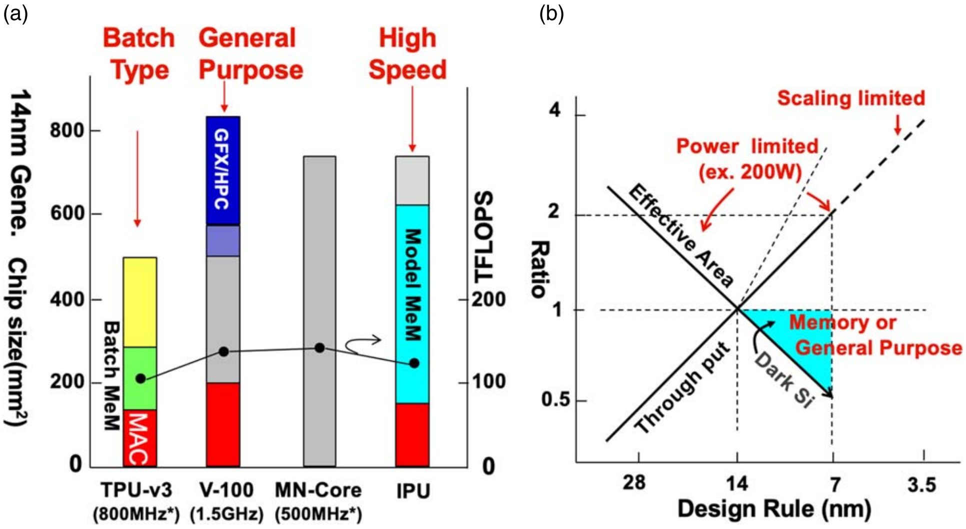


Fig. 7. (Color online) Representative server chips; (a) Architecture comparison, (b) Scaling and constraints.

bandwidths of these chips are improved up to 600 GB s^{-1} and 1.2 TB s^{-1} using high-bandwidth technologies of DDR5 interface and HBM2, respectively, higher batch sizes of several tens up to 100 are required and several tens of percentages of chip area have to be prepared for the buffers for the batch operation. Overall, Fig. 4 shows the advantage of the Memory-embedded AI chips on the performance.

2.4. Current status and future trends

In this section, we examine typical server chips for learning, TPU-v3, V-100 (GPU), MN-Core, and IPU chips, in terms of the throughput and the circuit configuration. As shown in Fig. 7, the throughput is distributed at approximately 100–150 TFLOPS, and there is not a big difference among the chips. However, the circuit configuration differs greatly; the memory for batch processing is woven in TPU-v3, and the block for graphics or HPC is woven in V-100. Regarding the IPU, the memory for the parameters (Model) takes up a large area. These chips can be divided into batch-type, general-purpose-type, and high speed-type.

With this backdrop of the appearance of such a variety of chip types, there is actually a rate-limiting factor on throughput that becomes more remarkable due to miniaturization, that is, the capacitance between wires is not scaled. Consequently, power constraints become even more severe, and the number of the AI PEs cannot increase as the miniaturization. This is shown in Fig. 7(b). Dark Si region highlighted with blue region is generated with proceeding miniaturization down to 7 nm. To effectively use the Dark Si area, it is replaced as a batch memory (TPU), general-purpose or graphic-processing parts (V-100), and model parameter (weight) SRAM (IPU). Although only two to three years have passed since its appearance in commercialization, the learning chips with simple CMOS technology have already reached its limitations on throughput. In the near future, the learning chip is expected to exploit quantization[25] down to 8 bit and even 4 bit and sparsification technologies, along with full-scale memory embedded technology.

3. Trends in edge AI chips

This section starts by briefly explaining the range of applications of edge AI; it then presents an overview of the first representative edge AI chip type, Eyeriss.[6,26] Next, a specific technology of data compression as the most important aspect of edge AI chips is explained in terms of its detailed techniques, implementation methods on the LSIs, and their trend.

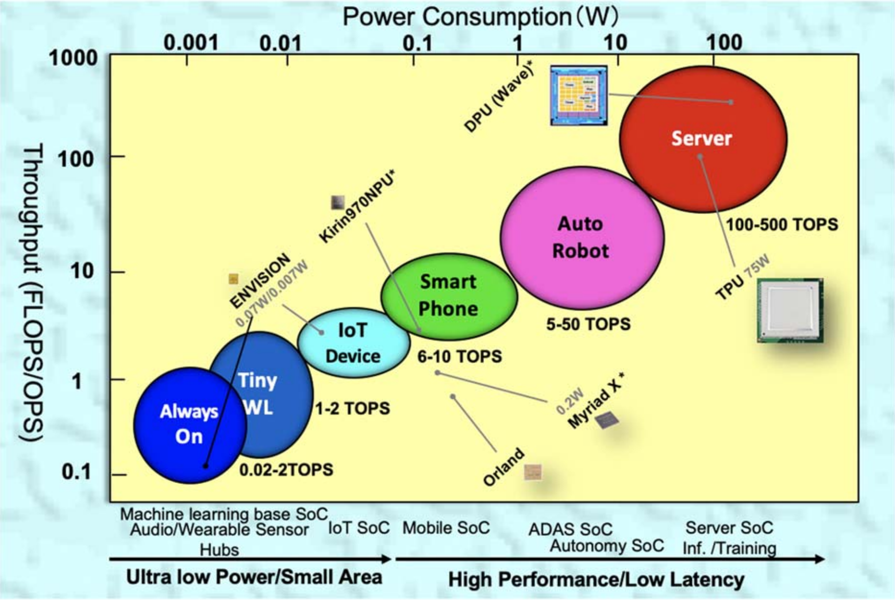


Fig. 8. (Color online) Application domains for AI chips and their throughputs and power consumptions.

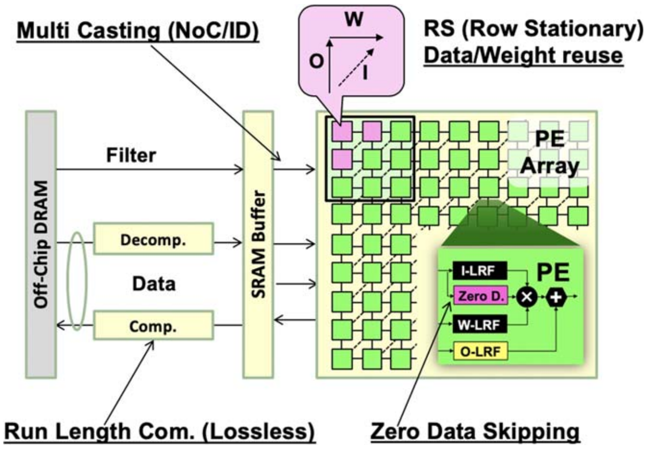


Fig. 9. (Color online) Eyeriss circuit diagram and technologies involved.

3.1. Edge AI chips

As shown in Fig. 8, the field of applications of edge chips is broad, extending from high performance automotive and robotic applications to smartphones, IoT, wearables, and always-on products; thus, required performances of throughput and power consumption are also widely spread. In automobiles, the throughput of 10 TOPS and more is required, and the power consumption is approximately 10 W. Conversely, for smartphones, it is essential that the power is 1 W or lower. Improving the throughput under this power restriction is critical. Therefore, implementation of the data compression technology on the edge AI that can increase the throughput and reduce the power consumption is the key. It is noteworthy that with always-on products, the ultimate compression rate is assumed. The focus of this section is on the technology, particularly for smartphones and always-on products.

3.2. Basic circuit diagram (Eyeriss)

Eyeriss chip was announced by MIT/NVIDIA in February 2016 at the International Solid-State Circuits Conference (ISSCC).[6] It targets convolutional layer-oriented models such as Network in Network/GoogleNet at that time. In this chip, DRAM is external. As shown in Fig. 9, it incorporates four circuit technologies.

The first technique is a dataflow control method that incorporates a proprietary technique while keeping in mind the reuse of both the data and filter weights in convolutional computing. In the PE array shown in Fig. 9 above, the weights (W) are transferred in the x direction, input activations (I) are transferred diagonally, and the accumulated outputs (O) is transferred in the y direction. Each PE is responsible for calculating the one-dimensional convolutional computation (1D Conv) with the weights (one row of filter) and input activations (one row of input feature map); the weights, input activations, and intermediate outputs of the 1D Conv are stored in local register files (or SRAM) of W-LRF, I-LRF, and O-LRF, respectively, and reused with changing one of the weights or the input activations. Consequently, repeated data flow to and from the PE array can be prevented, and thus power wastage due to data transfer is eliminated. This technique is named the row stationary dataflow (RS) method. Typical dataflow methods, other than the RS method, include the weight stationary (WS) method, which is suitable for batch processing used in TPU,[10] and the output stationary (OS) method, which is relatively suitable for sparse compression (sparsification) used in ShiDianNao.[5]

The second technique is a method of detecting the zero value of the input activations and skipping the multiplication operation. In the CONV layer, ReLU is usually used as an activation function, and thus, more than 50% of the activations are zero data, and wasteful power consumption is reduced by 45%. This technology is currently used in most AI chips.

The third is a network on chip (NoC) function placed in the SRAM buffer block. Recognition ID numbers are given to each PE, and using the numbers, the NoC transfer effectively data and weights to the PEs. For example, data and/or weights multicasting can be performed efficiently. This technology controls the RS dataflow as explained above. Furthermore, the size of the logical two-dimensional PE array can be efficiently reconfigured with respect to changes in

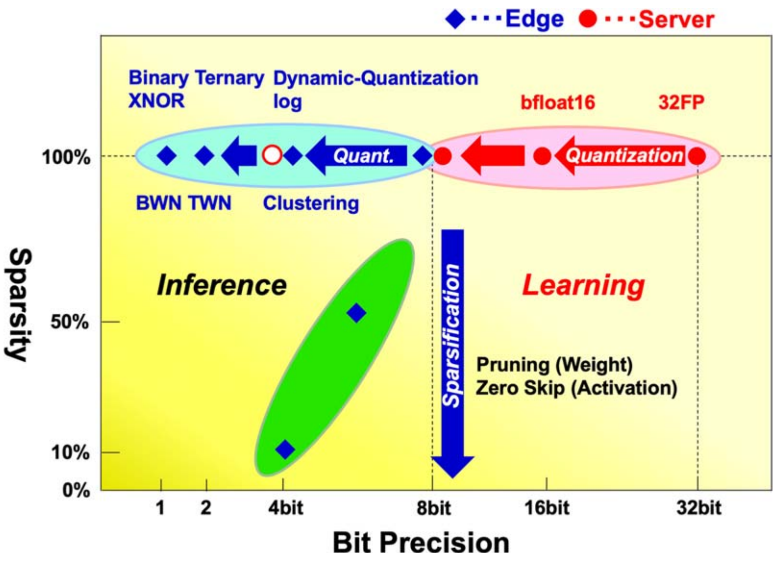
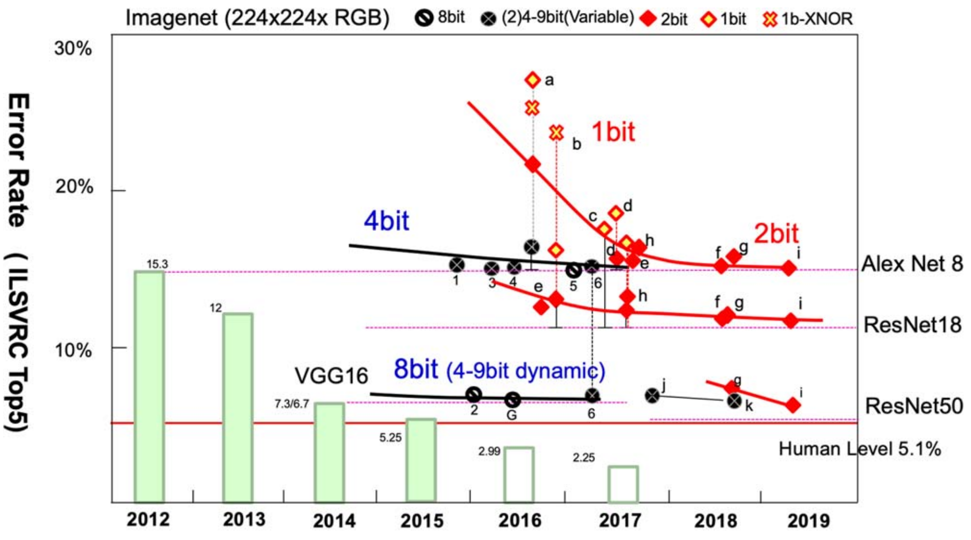


Fig. 10. (Color online) Trend for data compression.



1:Deep Compression, 2:Angeleye, 3.Squeeze Net, 4:Moon/VLSI'16(Alex), 5:Orlando(Alex), 6:Envision(Alex/VGG16), G: Google
a:Bengio Group(AlexNet), b:TWN(ResNet18), c:YodaNN(ResNet18), d:STD大(Alex/ResNet18), e:DoReFa, f:QIP, g:LQ-NETs,
h:HWGQ, i:PACT-SAWB-fpsc, j:Apprentice(W2-A8), k:UNIQ(W4-A8)

Fig. 11. (Color online) Reduction of bit precision.

model size (node size, number of channels, number of filters, etc.).

The fourth point is that lossless compression/decompression processing (run length compression) shown in Fig. 9 is performed in the input/output data to the external DRAM, and the compression efficiency is approximately 1/2.

3.3. Quantization and sparsification

3.3.1. Trends in data compression. The practical application research phase began around late 2015, quickly following the basic research phase. Research on low power consumption and high performance, especially for mobile applications, became the most active area of research. The subject of the study was PE. As shown in Fig. 10, corresponding dedicated compression circuits are incorporated to achieve low-bit quantization for data and weights, and sparsification for data and weights, respectively.

The trend of research activities regarding data compression can be divided into two areas: one is an algorithm-oriented movement represented by binary connect/binarized net,[27] led by Professor Bengio of the University of Montreal. Promising results were obtained with binary (1 bit) and ternary (2 bit) between 2015 and 2018 as shown in Fig. 11, and since the ternary-based PACT-SAWB-fpsc net[19] was announced by the IBM in 2019, performance comparable to 32 bit floating point has become available.

Another trend is the circuit architecture-oriented movement creating novel circuit architectures to adapt high-precision models of Image Net class used for server systems on the edge AI of mobile or wearable devices, under the condition that the increase in target error rate is maintained below 1%.

Regarding the first trend depicted in Fig. 11, the details are not discussed here, but the favorable results with Ternary/Binary give a good momentum to the circuit architecture research. It can be considered as a boost to the recent progress of In-memory computing architecture especially suited for low-bit networks. In-memory computing will be discussed in the Chapter 4.

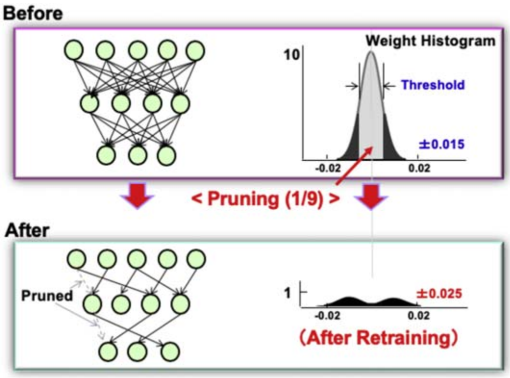


Fig. 12. (Color online) Pruning in weights.

This paper focuses on the second trend, in which power consumption is required to be several hundreds of 100 mW for the mobile or less than 10 mW to achieve always-on operation. As shown in Fig. 10 (please refer also to the compression techniques listed in Table I), quantization and sparsification are further divided into weights and activations, and they can be roughly categorized into four types.

① Quantization (weights): TWN (Ternary Weight Network) and BWN (Binary Weight Network), in which only weights are quantized and whose weights are 2 bits and 1 bit, are representative examples. Although there is a feature that the accuracy is less deteriorated, the calculation cost is not greatly reduced. One unique technique is weight-clustering (grouping) during/after learning and weights are referenced in a look-up table (LUT) during inference.[12]

② Quantization (activations/weights): Various techniques have been published. ENVISION[28] performs bit-precision reductions off-line by alternating activations/weights layer by layer, from the first layer of the CONV net. In DNPU,[12] only activations are quantized, but numbers of bits are optimized dynamically and layer by layer, by increasing or decreasing the decimal digits of the fixed-point format while detecting the overflow of the MSB of accumulated results (intermediate values) on-line. The problem is how to increase speed while reducing the number of bits. In ENVISION,[28] one multiplier is operated in sub-word units, thus realizing a 2× (4bit)/4× (2bit) speedup compared with that for 8bit.

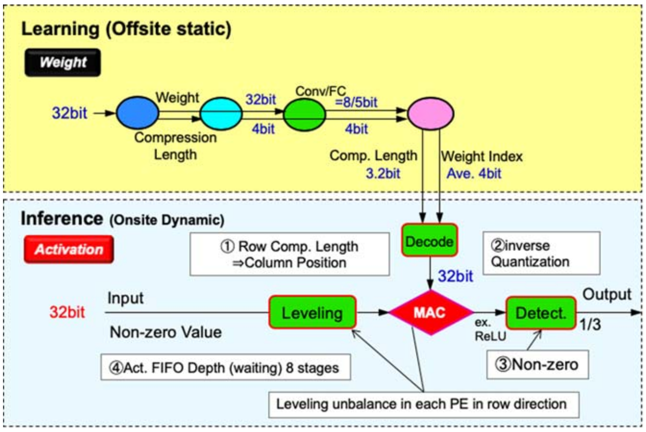


Fig. 13. (Color online) EIE processing flow for compression and decompression.

③ Sparsification (weights: pruning): A well-known technique is the pruning technique adopted by EIE,[11] shown in Fig. 12. In this technique, connection between the neuron of the previous layer and the neuron of the current layer is disconnected (pruned: the weight is set to zero). It can be done assuming that the weight with value closed to zero and less than a pre-defined threshold has a negligible effect on the dot products between weights and activations, as shown in the top of Fig. 12. Weight histogram after retraining is shown in the bottom of Fig. 12. The weight compression effect is 9 times in the FC layer (usually larger than in the CONV layer). The problem is how to memorize a large amount of the positions of the pruned connections (zero weights). The important point is how to compress the zero positions before inference, and decompress them in real time during inference. The EIE uses a compressed sparse column method after learning in the top of Fig. 13. The top of the figure depicts the flow of the weight compression. In the inference, a dedicated decompression circuit (decoder) is installed in the PE, as shown in the bottom/right diagram of Fig. 13. Particularly effective applications are speech recognition and neural machine translation on smartphones that use MLP and LSTM networks, which are mainly composed of FC layers.

④ Sparsification (activations: zero skipping): When the input activations to each PE are zero, the built-in zero detection circuit issues a flag and stops the operation of that PE, thereby reducing power consumption by 50% or more. This technique is used in most chips, but from a PE array utilization point of view, more than 50% of PEs are stopped down, so the PE array is not being used effectively. Thus, a circuit (non-zero detection circuit: at the bottom side of Fig. 13) was introduced in the EIE for detecting non-zero activations, leveling the operation status among PEs, and improving the effective utilization rate of the PE array.

As described in ①–④ above, a special feature of this field is the use of specialized circuit design techniques to achieve goals such as optimal quantization using new algorithms, solving problems during or after sparsification, or further improvement.

3.3.2. Implementation status of data compression on edge AI chips.

Significant research and development has been conducted to resolve the issues of how to incorporate data compression technology into AI chips since Eyeriss in 2016, as shown in Fig. 14. As indicated in the inset of Fig. 4, each AI chip is represented by a black or a red colored bar with the name or code, the height and width of which denote the bit precision range of the chip and the contribution of the chip on the sparsification, respectively; the red color indicates the contribution of the chip on the development of the dataflow method.

Research activities related to 2 bit (Ternary) and 1 bit (Binary) technologies have been conducted since 2017. The goal is to develop a more broadly applicable and more flexible type of reconfigurable AI chip. KAIST's UNPU[14] and Hokkaido University's QUEST,[15] both announced in

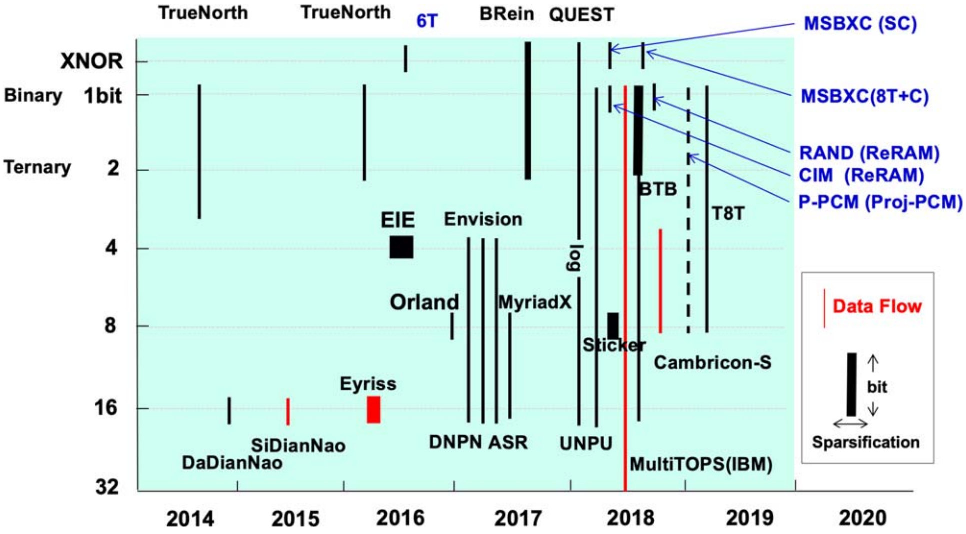


Fig. 14. (Color online) Trend for AI chips and bit precision.

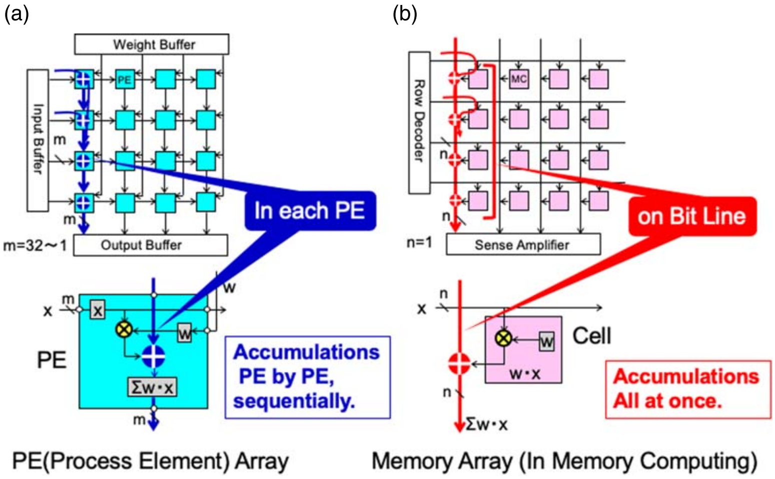


Fig. 15. (Color online) PE array versus memory array (In-memory computing).

2018, achieved this goal by introducing bit-serial technology that performs operations on a bit-by-bit scheme. Using this scheme, a wide range of bit precision can be easily reconstructed. As for the sparsification, a typical example is Eyeriss that applies sparse compression for input activity, and EIE that applies pruning for weights. The red line indicates a chip for which dataflow has been studied. The OS method was used for ShiDianNao,[5] and the row stationary (RS) method was considered for application to a part of Eyeriss and MultiTOPS.[29]

As shown in Fig. 14, much research has been conducted on the implementation of 2 bit (Ternary), 1 bit (Binary), and XNOR networks on LSIs, i.e. BRein,[18] 8T + C,[17] RAND,[20] and T8T,[30] to handle the rapid progression in the algorithms aiming to reduce bit precision as introduced in Fig. 11. This movement has led to a new trend of in-memory computing from around 2017, which will be described in Chapter 4.

4. Neuromorphic chips

This section will describe the trends in AI chip development related to "Neuromorphic Engineering", a research and academic field that aims to faithfully reconstruct neuroscience-based models of brain cells on integrated circuits. This section will focus on AI chips that incorporate in-memory computing or/and network based on spiking neurons. In this paper, in-memory computing is described as an imitation of the operation of brain cells.

4.1. In-memory computing using SRAM or NVM

Figure 15 shows the comparison between the ordinal WS-type PE array (a) and memory array for in-memory computing (b). The memory cells include single MOSFET and comprise MAC operation with weight W stored in memory. These two configurations of the PE array and Memory array will be compared in the following paragraphs.

(1) Array configuration: In the PE array, the output of the PE is sent vertically to the PE directly below. In contrast, in the memory array, all the output of the cells are connected to the one bit-line.[31]

(2) Multiplication result: The output of each PE is in digital. Meanwhile, in memory array, the output is in analog.

(3) Accumulation: In the PE, after sending the output of the PE into the successive one every cycle in a digital manner, the final PE outputs the accumulated value. In the in-memory processing, in contrast, the values of the products of all cells can be gathered on the bit-line and are read as the accumulated value with the sense amplifier simultaneously.

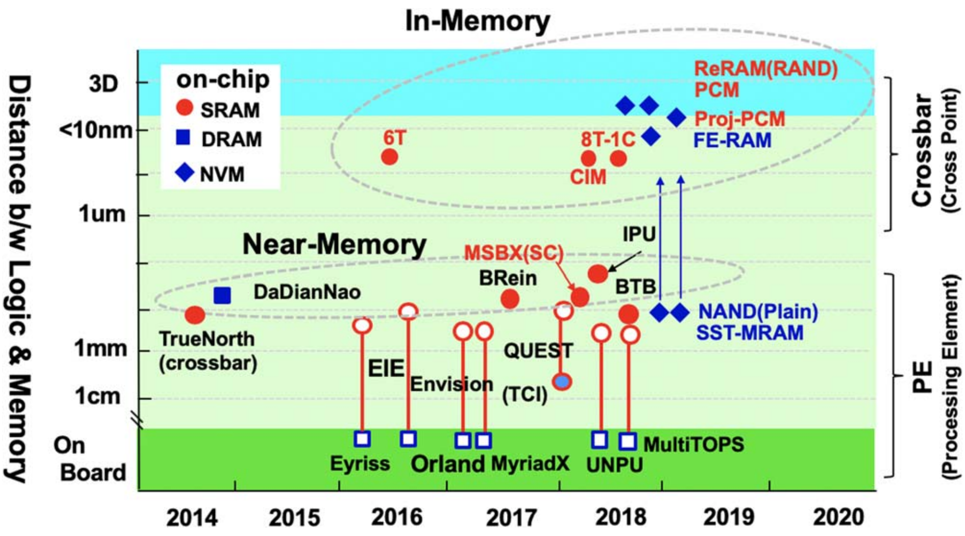


Fig. 16. (Color online) Distance between logic elements and memory.

In in-memory computing, power consumption is greatly reduced and computing speed much improved.

The other advantages in the in-memory computing are negligible power wastage due to transfer weights to each PE and extremely high weight memory bandwidth due to negligible distance between logic and weight memory. Figure 16 shows the trend in the distance between the logic (MAC) and weight memory. The distance to the external memory on the board (external memory) is several centimeters. The distance to the internal buffer or internal SRAM varies from 1 mm to several mm (usually referred to as near-memory). In in-memory computing, these can be further reduced down to several nm classes. Both 6T SRAM[16,32] and 8T-1C SRAM[17] were announced in 2016 and 2018, respectively, by Princeton University. The latter evolved into a 2T (selection transistor) and capacitor C to prevent erroneous writing during reading, which occurs in the 6T SRAM case. Because the number of bits per cell is 1, as long as the SRAM cell is used, to increase bit precision, it becomes necessary to bundle and process numerous cells simultaneously.[30]

As mentioned above, research, development, and commercialization of in-memory processing by NVM have been performed in many cases. A typical example is the development of ReRAM, which is already mass-produced as a pure memory chip, for in-memory processing applications, such as in RAND.[20] This is an analog-type NVM AI chip, whose memory represents several bits per cell. Each column-sensing circuit on the chip reads a digital result of 1 or 0 by sensing and comparing the difference in states between pairs of two complementary bit-lines. It has demonstrated an extremely low power efficiency of 66 TOPS/W (explained in Chapter 5). Similarly, the development of phase change memory (PCM)[21] is also advancing. This development can be characterized by the pursuit of multiple bits per memory, and to date, highly favorable results have been realized.[21] In addition, the commercialization of products using in-memory processing is also progressing rapidly. Further, the shipping of Mythic chips was announced in 2018.

In offering a brief overview of the deployment of NVM in other types of devices, FeRAM,[33] NAND,[34] and SST-MRAM[35] have each garnered attention. For NAND, the ability to read at a high speed is essential. As MRAM can be four or more times denser than SRAM, expectations for a high-end inference chip are increasing. Again, it has been shown that highly accurate inference is possible with 2 bits (Ternary),[19] and this will become an increasingly important aspect in the future.

4.2. Spiking neuron AI chips

The feature of the SNN is that each element (cell) operates (spike emission) only when a large amount of information through an axon is input. As such, operations are relatively rare and low power consumption is achieved. The challenge is that it is possible to implement the principle of synaptic plasticity (spike-timing-dependent plasticity), which mimics the learning mechanism of the human brain relatively easily, and that new learning task use cases can be realized. The following paragraphs will illustrate these two well-known examples by using two AI chips.

4.2.1. Inference AI chips. This section focuses on the TrueNorth[1] technology published in Science in August 2014 as a result of the synapse project funded by the U.S. Department of Defense's Defense Advanced Research Projects Agency (DARPA). This chip is a neuromorphic chip that mimics the mechanism of brain cells. However, the following paragraphs will attempt to explain the calculation method used by TrueNorth and performance with respect to deep learning rather than to a scientific emulator. This chip is used only for inferencing.

The configuration of this chip is quite large, at 17 × 25 mm^2 (28 nm), as shown on the left side of Fig. 17. There are 4096 cores laid out in a 2D array. Communication is performed by spikes among cores, chips, and boards. This chip is a crossbar-like virtual synapse incorporated within a massively parallel real network using packet communications comprising pulses generated by connected neurons. The majority of other chips constitute a virtual neural network adopting a time-share method in which the majority of other chips use the circuit (PE) at high speed (several hundred MHz to 1 GHz). In the core depicted on the right side of Fig. 17, 256-axon inputs are selectively connected at synapses and transmitted to neurons located at the bottom of the array.

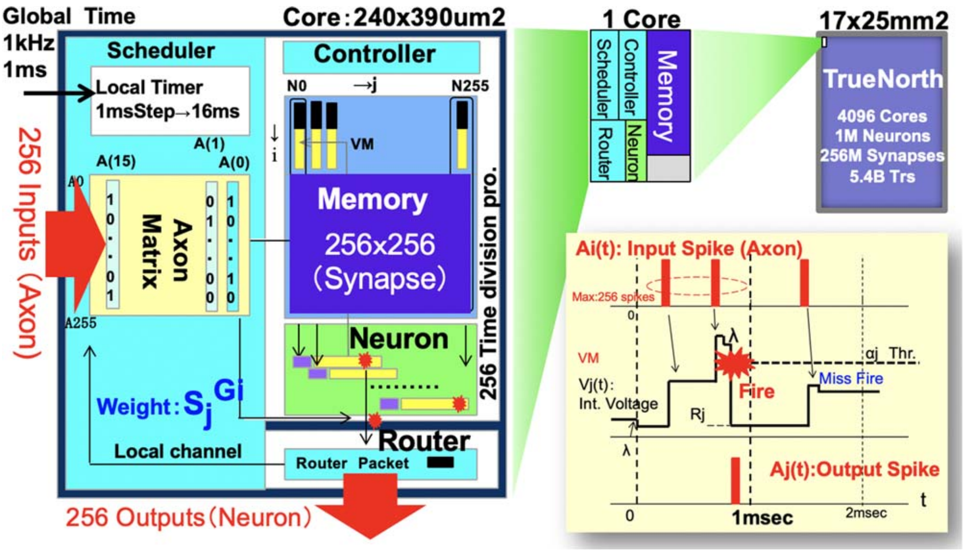


Fig. 17. (Color online) TrueNorth design architecture.

Information (pulses) is transmitted in the 256 independent neurons according to the neuron model. The architecture can be configured similar to that of a 256 × 256 bit SRAM. The grid points 0 and 1 of the crossbar serve as synapse connection information. The crossbar portion performs a matrix operation (matrix multiplication). However, as the neuron circuit is large, a virtual neuron method that performs time-division processing (256 divisions) by using one neuron circuit repeatedly is adopted.

When calculating the computational performance, the number of MACs is $256 \times 256 \times 64 \times 64 = 268\text{M}$ for the entire chip, and massive parallel processing of approximately 4,000 times is possible for 65k TPUs. The operating frequency is 1 kHz to simulate the operation of cranial nerves. Table I shows the average number of spikes (180 Hz/20 Hz). The computational performance is 536 GOPS/s, which is considerably lower than that of the TPU (92 TOPS/s); nevertheless, it is possible to achieve the same performance by increasing the operating frequency by 1 kHz. However, in practice, there is thought to be a limitation as a considerably large circuit utilizes time-division multiplexing. The rate of power consumption is 70 mW, though the actual leakage current is approximately half of this, resulting in a very high performance of 7.7 TOPS/W. At that time, the effective axis input is 1/2 and the pulse-generation probability is usually 20 Hz/1 kHz.[1] The probability is controlled by the threshold of the model.

The key aspects that enable low power consumption are quantization and sparsification. Considering the similarities with quantization discussed in the previous section, TrueNorth uses 1 bit for data, 1 bit for connections, and 2 bits or more clustered weight. In sparsification, the threshold value of the neuron model is similar to the pruning threshold value. As the spike output does not transmit a zero value, this can be interpreted as the function of zero skip having been performed in the previous layer. The above 7.7 TOPS/W is only the best reference value; this is because whether or not the performance is the same between the dense and sparse operations is not shown. Using the performance in the sparse state as is and adding the condition (1 SOPS = 2 OPS) results in a very low value of 92 GOPS/W as shown in Table I; this is close to that proposed by IBM. Conversely, the compression ratio can be considered to be approximately 100-fold. As the

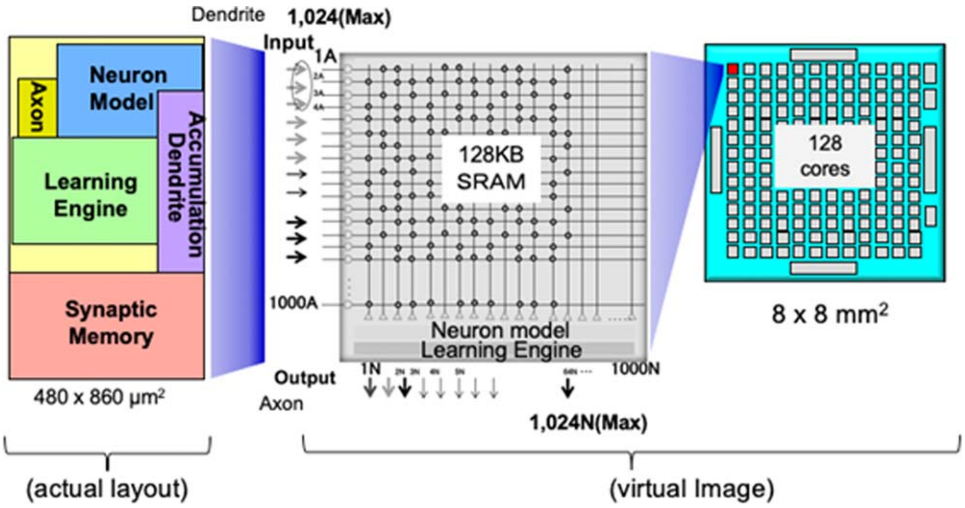


Fig. 18. (Color online) Loihi circuit diagram.

compression ratio of EIE is 30–50×, it can be regarded as the same, and the degree of similarity is quite high. Some TOPS/W can be expected.

To summarize, neuromorphic chips achieve higher performance (lower power consumption) by emulating the brain more faithfully, whereas virtual neural networks with advanced abstraction initially perform only basic MAC operations. Nevertheless, in the last 1–2 years, quantization and sparsification have been introduced to realize higher performance.

In addition to the unavoidable time-division multiplexing of neuron circuits, there is also the second issue of spiking output when attempting to use deep learning techniques in mounting a neuromorphic chip. Regarding the first issue, it is not possible to add more than the size of the core to increase the scale of the network; therefore, it is necessary to devise a method that, for example, limits the model itself. Regarding the second issue, the differential processing utilized by the backpropagation method is not compatible with spikes; nevertheless, this can be overcome by approximating the pulse to a triangular wave.[36] In this case, favorable results were obtained during demonstrations of CNN (MNIST/CIFAR100) and MLP/BLSTM, which express weights with ternaries (−1, 0, +1) using axons of adjacent pairs as one input.

4.2.2. Learning AI chip with spike-timing-dependent plasticity. This subsection will serve to introduce AI chips designed for learning tasks. These include Intel's Loihi[22] as shown in Fig. 18 and CEA-Leti's DynapSEL chips. These are equipped with self-learning functions based on synaptic plasticity. The Loihi was officially released during the first half of 2018, and Intel has been actively working to strengthen its collaboration with academia while promoting challenges to bring out new learning effects. The configuration of this chip utilizes a neuron model and a weight (synaptic) memory similar to those of TrueNorth, but differs in that it incorporates a large-scale learning engine shown in Fig. 18. Although the number of bits per weight is variable, it typically ranges from 3 to 5 bits. The chip has a Near Memory configuration with a pseudo-crossbar configuration, a memory capacity of 128 kB SRAM, and contains 128 cores.

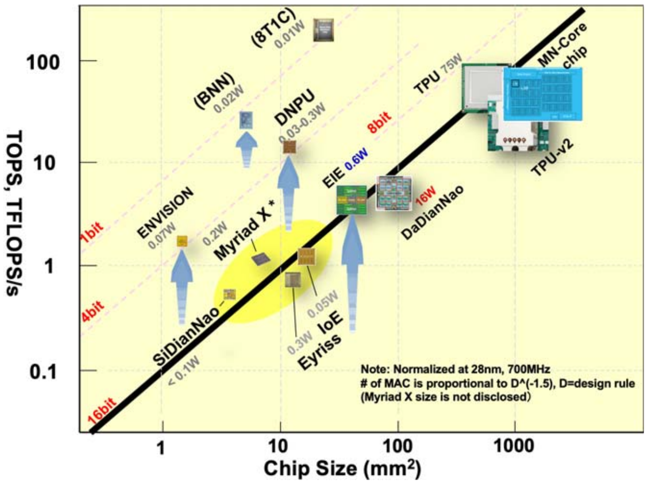


Fig. 19. (Color online) Relationship between throughput and chip size.

5. Discussion

This section will first discuss throughput and power consumption, which are important performance factors for AI chips. It will finally characterize the relationship between applications and weights that represent the scale of the model, and discuss the importance of memory embedding.

Throughput and chip size: Fig. 19 shows the relationship between the peak throughput (TOPS or FLOPS/s) and the chip area. The area size was used by estimating the area of the circuit related to the operation of the neural network. In addition, only throughputs were regularized to 28 nm and 700 MHz. Therefore, it is clear that the throughput is uniquely determined in proportion to the area and is inversely proportional to the number of bit precision. Most typically, 4 bits constitutes 1 TOPS mm^{-2} (28 nm, 700 MHz). If the weight memory is embedded, the calculation portion will occupy approximately 10% of the whole chip (with DaDianNao, calculation logic = 6%, buffer = 5%); consequently, if the figure is shifted one digit (10 to 1) to the left, the result will become almost consistent with the rule. With EIE[11]/DNPU[12]/ENVISION,[28] the speed can be increased almost proportionally by selecting a low-bit on the circuit. By contrast, with the TPU,[9] the operation speed does not change even after changing to 4 bits. EIE can improve the performance up to 30-fold by incorporating the functions of pruning and non-zero detection function.

Power efficiency and throughput: Fig. 20 displays the relationship between the power efficiency and throughput. Power efficiency can be improved by reducing the number of bits, and increases to approximately 10 TOPS/W at 4 bits. Orland and DNPU use the existing dynamic-voltage-frequency scaling circuit technology that can change the voltage and frequency for each circuit, and ENVISION realizes a 40-fold increase in efficiency while maintaining performance at 76 GOPS/s. This was achieved by introducing the dynamic-voltage-accuracy-frequency scaling circuit technology, which is capable of accurate scaling in addition to providing a fully depleted SOI substrate. As mentioned above, numerous types of chips can also be mounted on mobile devices, and a 10 mW class chip (ENVISION: mean 7 mW), capable of always-on operation, is also possible by dynamic operation of the circuit. Each chip type represented by a 1 bit Analog in the figure incorporating the in-memory processing technology described previously in Sect. 4 has recently realized performances of 100 TOPS/W or more, including the RAND.[20]

Estimation of memory embedding: Fig. 21 shows the relationship between the number of input dimensions and the number of weights representing the scale of the model. There is a fixed relationship between the network model and the number of input dimensions. The size of the model of the CNN, which comprises primarily the Convolutional Layer frequently used in image recognition applications, increases in proportion to the number of input dimensions. Meanwhile, in MLP or RNN/LSTMs, which are mainly composed of fully connected layers suitable for simple identification or recognition of temporal sequences [recurrent neural network (RNN)], the number increases in proportion to the square of the number of input dimensions.

Particularly when used for natural language processing or automatic machine translation (Neural Machine Translation),

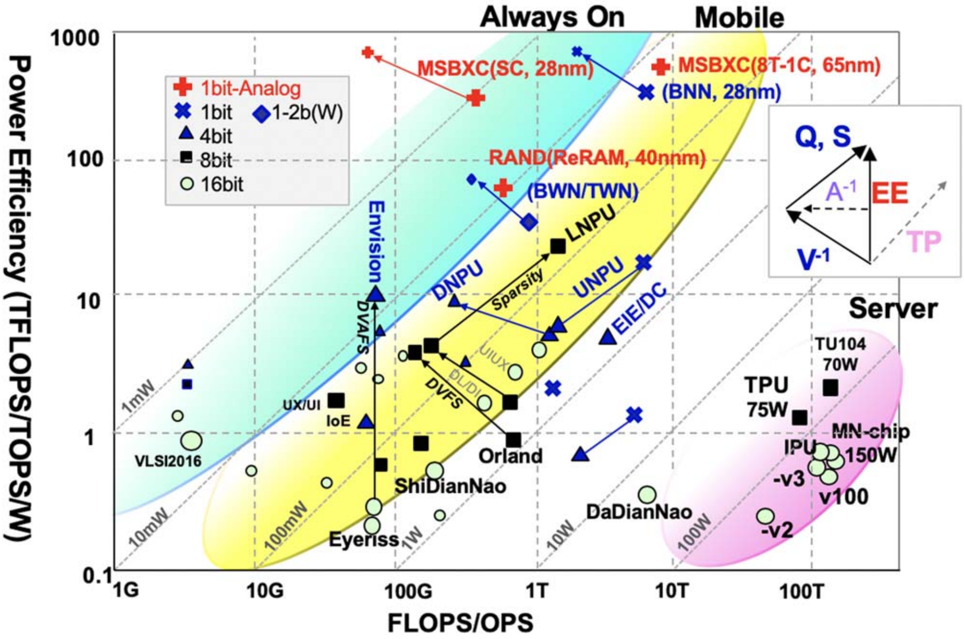


Fig. 20. (Color online) Power efficiency and throughput.

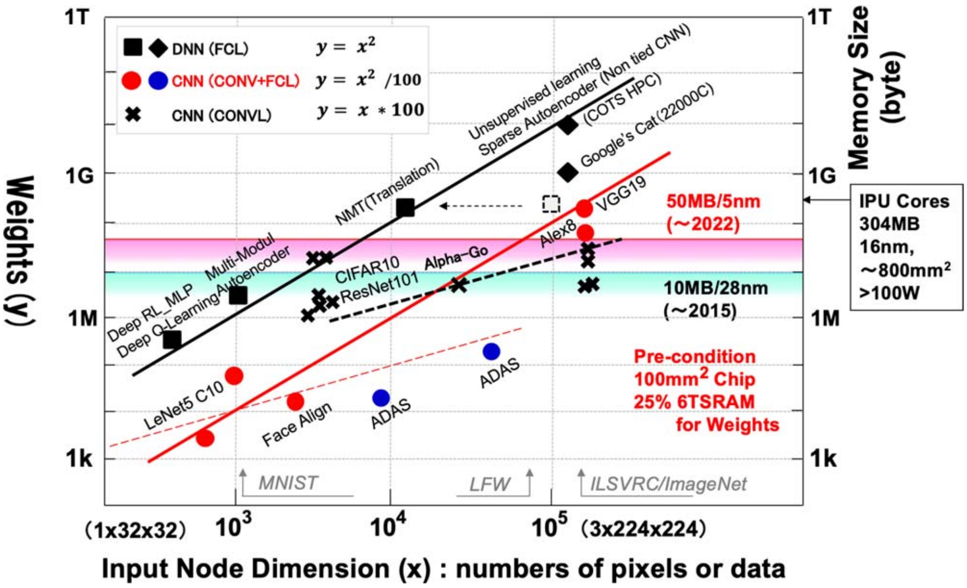


Fig. 21. (Color online) Relationship between numbers of input nodes and weights.

and as the number of dimensions increases to approximately 10^5 or higher, because language is used as input, the scale reaches approximately 1G. At present, IPU Cores, which have the most memory embedded at the practical level, have a size of 16 nm, 800 mm^2, and an ultimate capacity of approximately 304 MB (equivalent to a 1G weight when reduced to 2 bits). However, this type is only used for server or high-end edge AI chips of approximately 3 cm. Nevertheless, its power is comparatively high, whereas the requirement is low power consumption. Various applications will also become possible with Edge AI, with the development of in-memory processing and miniaturization using NVM with a capacity of several 1 G bits or more.

6. Conclusions

AI chip technology entered a new third developmental phase in mid-2017, following a basic research phase that commenced in 2013 and a recent commercialization phase. Server systems, such as Google's TPU-v2/v3[13] for Cloud, PFN's MN-Core, and Graphcore's IPU Cores, are all meant for practical use or are in a stage just prior to commercialization, and all are designed for learning tasks. However, the limitations of Moore's law, particularly the limitations on power consumption, are already visible, and these limitations remain imminent even with further optimization (e.g. low-bit learning[25]). Additional developmental activities are necessary, and NVM is now regarded as a candidate. For example, multi-bit Projection-PCM[21] is one possibility; however, it appears that many pending issues remain that must be resolved until practical application will be possible. With the rapid increase in request on learning, there are high expectations for the emergence of a savior technology.

Edge AI is believed to be close to practical use with optimization at the CMOS PE level covered to some extent.

The development of ASICs that are fairly circuit-oriented is in progress, and optimization for bit serial processing and even LUT-based applications is expected to occur in the future. However, the trend of bit reduction is in the limelight, and edge products in the 2 bit class will attract more attention in the future. There has been progress in in-memory processing configuration and miniaturization of the NVM toward practical use. Regardless, success will depend on the model structure and the determination of the optimal number of bits, and as such, it is necessary to identify future technological trends and work on solving problems while promoting collaboration with research and development activities in the field of AI algorithms. We believe that a new vision for AI chip technology beyond the current CMOS SRAM base will be born from ongoing research.

Acknowledgments

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

1) P. Merolla et al., Science **345**, 668 (2014).
2) H. Momose and T. Asai, J. Jpn. Soc. Artificial Intell. **33**, 23 (2018) [in Japanese].
3) C. Farabet, E. Culurciello, and Y. LeCun, Conference, 2011, p. 109.
4) V. Gokhale, J. Jin, A. Dundar, B. Martini, and E. Culurciello, Computer Vision and Pattern Recognition Workshops (CVPRW 2014), 2014, p. 23.
5) Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, The 42nd Int. Symp. on Computer Architecture (ISCA), 2015.
6) Y. Chen, T. Krishna, J. Emer, and V. Sze, Proc. 2016 IEEE Int. Solid-State Circuits Conf. (ISSCC), Session 14.5, 2016, p. 262.
7) T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, ASPLOS '14 Proc. 19th Int. Conf. on Architectural Support for Programming Languages and Operating Systems, 2014, p. 269.
8) Y. Chen et al., Proc. 47th IEEE/ACM Int. Symp. on Microarchitecture (MICRO'14), 2014, p. 609.
9) D. Fu Liu, T. Chen, S. Liu, J. Zhou, S. Zhou, O. Temam, X. Feng, X. Zhou, and Y. Chen, ASPLOS '15 Proc. 20th Int. Conf. on Architectural Support for Programming Languages and Operating Systems, 2015, p. 369.
10) N. Jouppi et al., Proc. 44th Annual Int. Symp. on Computer Architecture (ISCA), 2017, p. 1.
11) S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. Horowitz, and W. Dally, Proc. 2016 ACM/IEEE 43rd Annual Int. Symp. on Computer Architecture (ISCA), 2016, p. 243.
12) D. Shin, J. Lee, J. Lee, and H. Yoo, Proc. 2017 IEEE Int. Solid-State Circuits Conf. (ISSCC), Session 14.2, 2017, p. 240.
13) J. Dean, NIPS 2017 Workshop, Deep Learning at Supercomputer Scale, Panel Discussion, 2017.
14) J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H. J. Yoo, Proc. 2018 IEEE Int. Solid-State Circuits Conf. (ISSCC), Session 13.3, 2018, p. 218.
15) K. Ueyoshi, K. Ando, K. Hirose, S. Takamaeda-Yamazaki, J. Kadomoto, T. Miyata, M. Hamada, T. Kuroda, and M. Motomura, Proc. 2018 IEEE Int. Solid-State Circuits Conf. (ISSCC), Session 13.2, 2018, p. 216.
16) J. Zhang, Z. Wang, and N. Verma, Proc. 2016 Symp. on VLSI Circuits Digest of Technical Papers, 2016, p. C252.
17) H. Valavi, P. Ramadge, E. Nestler, and N. Verma, Proc. 2018 Symp. on VLSI Circuits Digest of Technical Papers, 2018, p. C141, C13-5.
18) K. Ando et al., Proc. 2017 Symp. on VLSI Circuits Digest of Technical Papers, 2017, p. C24, C2-1.
19) J. Choi, S. Venkataramani, V. Srinivasan, K. Gopalakrishnan, Z. Wang, and P. Chuang, The Conf. on Systems and Machine Learning (SysML) 2019, 2019 [https://mlsys.org/Conferences/2019/doc/2019/168.pdf].
20) R. Mochida, K. Kouno, Y. Hayata, M. Nakayama, T. Ono, H. Suwa, R. Yasuhara, K. Katayama, T. Mikawa, and Y. Gohou, Proc. 2018 Symp. on VLSI Technology Digest of Technical Papers, 2018, p. T175, T16-4.
21) I. Giannopoulos, A. Sebastian, M. Le Gallo, V. P. Jonnalagadda, M. Sousa, M. N. Boon, and E. Eleftheriou, Proc. Int. Electron Device Meeting 2018, 2018, p. 629, 27.7.
22) M. Davies et al., IEEE Micro **38**, 82 (2018).
23) S. Venkataramani, A. Ranjan, and A. Raghunathan, Proc. 44th Annual Int. Symp. on Computer Architecture (ISCA), 2017, p. 13.
24) S. Liu, Z. Du, J. Tao, D. Han, T. Luo, Y. Xie, Y. Chen, and T. Chen, Proc. 43rd Int. Symp. on Computer Architecture, 2016, p. 393.
25) N. Wang, J. Choi, D. Brand, C. Chen, and K. Gopalakrishnan, 32nd Conf. on Neural Information Processing Systems (NeurIPS 2018), 2018.
26) Y. Chen, J. Emer, and V. Sze, 2016 ACM/IEEE 43rd Annual Int. Symp. on Computer Architecture (ISCA), 2016, p. 367.
27) M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, (2016), arXiv:1602.02830.
28) B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, Proc. 2017 IEEE Int. Solid-State Circuits Conf. (ISSCC), Session 14.5, 2017, p. 246.
29) B. Fleischer et al., Proc. 2018 Symp. on VLSI Circuits Digest of Technical Papers, 2018, p. C35, C4-2.
30) X. Si et al., Proc. 2019 IEEE Int. Solid-State Circuits Conf. (ISSCC), Session 24.5, 2019, p. 396.
31) T. Morie, J. Jpn. Soc. Artificial Intell. **33**, 39 (2018) [in Japanese].
32) J. Zhang, Z. Wang, and N. Verma, IEEE J. Solid-State Circuits **52**, 915 (2017).
33) X. Sun, P. Wang, K. Ni, S. Datta, and S. Yu, Proc. Int. Electron Device Meeting 2018, 2018, p. 55, 3.1.
34) Y. Lin, F. Lee, M. Lee, W. Chen, H. Lung, K. Wang, and C. Lu, Proc. of Int. Electron Device Meeting 2018, 2018, p. 39, 2.4.
35) N. Xu, Y. Lu, W. Qi, Z. Jiang, X. Peng, F. Chen, J. Wang, W. Choi, S. Yu, and D. Kim, Proc. of Int. Electron Device Meeting 2018, 2018, p. 348, 15.3.
36) S. Esser et al., Proc. Natl Acad. Sci. **113**, 11441 (2016).

Hiroshi Momose was born in Nagano, Japan, in 1954. He received B.S, M.S and Ph.D. degrees in electrical and electronics engineering from Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979 and 2000, respectively. In 1979, he joined the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan, where he has been engaged in the research and development of the advanced CMOS/ BiCMOS device and LSI technology. In 2009, he joined Semiconductor Technology Academic Research Center (STARC), Yokohama, Japan, where he has surveyed and analyzed the status and trend of the emerging technologies of IoT and AI. Since 2016, he has been a research fellow of Graduate School of Information Science and Technology, Hokkaido University, Hokkaido, Japan. His current research interest involves the highly efficient edge AI devices and the analysis of the trend of them.

Tatsuya Kaneko received B.E. and M.E. degrees from Hokkaido University, Sapporo, Japan, in 2018 and 2020, respectively, where he is currently pursuing a Ph.D. degree. His current research interests include hardware-aware training algorithm and its hardware architecture for an edge device.

Tetsuya Asai received B.S. and M.S. degrees in electronic engineering from Tokai University, Hiratsuka, Japan, in 1993 and 1996, respectively, and a Ph.D. degree from the Toyohashi University of Technology, Toyohashi, Japan, in 1999. He is currently a Professor with the Graduate School/Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan. His current research interests include developing intelligent integrated circuits and their computational applications, emerging research architectures, deep learning accelerators, and device-aware neuromorphic very large-scale integrations.