micromachines

Review

Architecture of Computing System based on Chiplet

Guangbao Shan [1], Yanwen Zheng [1], Chaoyang Xing [2], Dongdong Chen [1,*], Guoliang Li [1,*] and Yintang Yang [1]

[1] School of Microelectronics, Xidian University, Xi'an 710071, China; gbshan@xidian.edu.cn (G.S.); ywzheng@stu.xidian.edu.cn (Y.Z.); ytyang@xidian.edu.cn (Y.Y.)
[2] Beijing Institute of Aerospace Control Devices, Beijing 100039, China; 20111110257@stu.xidian.edu.cn
* Correspondence: ddchen@xidian.edu.cn (D.C.); guoliangli@stu.xidian.edu.cn (G.L.)

Abstract: Computing systems are widely used in medical diagnosis, climate prediction, autonomous vehicles, etc. As the key part of electronics, the performance of computing systems is crucial in the intellectualization of the equipment. The conflict between performance, efficiency, and cost can be solved by choosing an appropriate computing system architecture. In order to provide useful advice and instructions for the designers to fabricate high-performance computing systems, this paper reviews the Chiplet-based computing system architectures, including computing architecture and memory architecture. Firstly, the computing architecture used for high-performance computing, mobile, and PC is presented and summarized. Secondly, the memory architecture based on mainstream memory and emerging non-volatile memory used for data storing and processing are introduced, and the key parameters of memory are compared and discussed. Finally, this paper is concluded, and the future perspectives of computing system architecture based on Chiplet are presented.

Keywords: computing system; computing architecture; memory architecture; Chiplet

Citation: Shan, G.; Zheng, Y.; Xing, C.; Chen, D.; Li, G.; Yang, Y. Architecture of Computing System based on Chiplet. *Micromachines* **2022**, *13*, 205. https://doi.org/10.3390/mi13020205

Academic Editors: Ran Peng and Shuailong Zhang

Received: 26 December 2021
Accepted: 24 January 2022
Published: 28 January 2022

1. Introduction

Electronic equipment is becoming more intellectualized with the development of 5G, artificial intelligence (AI), and big data. It has been widely used in medical diagnosis, automotive, electronic product design, Industry 4.0 Internet of things, etc. In the medical field, computer-aided diagnosis can improve efficiency and accuracy by preprocessing and classifying pathological images [1]. Explosive data from vehicles sensors and high-precision avigraph are also processed by computing systems for safety [2]. In addition, computing systems have been used to analyze data from the Internet of Things (IoT) to improve efficiency in smart factories [3]. Precision equipment can be designed by using a computing system; therefore, high-performance computing systems are crucial in electronic equipment [4].

Traditionally, the performance of computing systems can be improved by increasing transistors and frequency of integrated circuits (IC) [5]. In order to meet the requirements of the higher computing power, energy efficiency, and the lower cost of diversified applications, architectural innovation and technology scaling have been proposed to achieve these goals. The computing systems have been developed from single-core to multi-core, including homogenous multi-core and heterogeneous multi-core. In the data-centered applications, the traditional approach is facing some problems: (1) explosive costs [6]; (2) rapid increase in leakage power; (3) scalability degrades; (4) system design complexity increases, which can affect the improvement of the computing system. Due to the advantages of Chiplet, it has been used in the architecture of computing systems [7,8]. Chiplet is a small-scale hard IP with high yield and reusability [9–11]. The computing system architecture design based on Chiplet glues together the advantages of technology scaling, three dimensions (3D) integration technology, and a new device to construct a high-performance computing system, which has some merits (1) reducing the design cost via a smaller area

and higher yield [12]; (2) avoiding the dark silicon effect [13]; (3) shortening the design cycle by Chiplet reuse; (4) improving the system scalability by flexible Chiplet combinations [14].

The appropriate computing system architecture can effectively utilize the advantages of Chiplet technology in specific applications. This paper aims to summarize the characteristics and performance of Chiplet-based computing and memory architectures to provide instructions for the design of a high-performance computing system. This paper mainly introduces the computing system architectures based on Chiplet, as shown in Figure 1, which mainly includes computing architectures and memory architectures. In computing architecture, 2.5D and 3D computing architectures based on Chiplet are presented and compared. In-memory architecture, near-processor memory architecture, and processing-in-memory architecture based on mainstream and emerging memory are presented and analyzed. Finally, the future perspectives of the computing system architectures based on Chiplet are discussed.

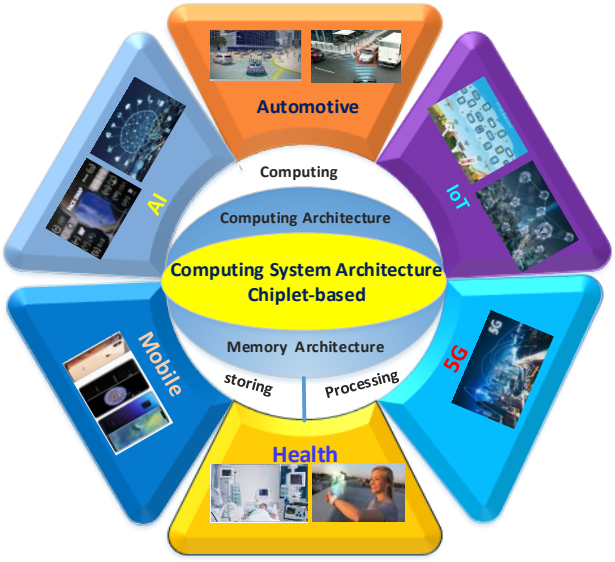

Figure 1. Computing system architectures and application.

2. Computing Architecture Based on Chiplet

As shown in Figure 2a, heterogeneous multi-core architecture has a higher efficiency than single-core and homogenous multi-core architectures [15]. In order to further improve multi-core architectures performance, more transistors were integrated into a limited area of a die; however, the leakage power of the transistor increases as the technology scaling, which severely reduces the energy efficiency of the multi-core architectures. Moreover, in order to ensure the thermal reliability of the computing system, some hardware resources in a die cannot be utilized; that is, the dark silicon effect is more obvious. The architecture of the computing system prepares the computing unit and the memory on one substrate with the same advanced technology, which is a monolithic System on Chip (SoC) and can improve its performance; however, the integration of computing, memory, control, and other IPs into the chip significantly increase the complexity of design and verification. Further, the analog and digital circuits are fabricated using different processes, so multi-manufacturing equipment has to be used in the same process, which dramatically increases the costs. In order to further improve the SoC performance, many chips are designed as dedicated chips, so the chip scalability deteriorates. For example, the performance of Apple mobile SoC processors was significantly improved through technology scaling and architecture updates while the costs obviously raised, as shown in Figure 2b [16].

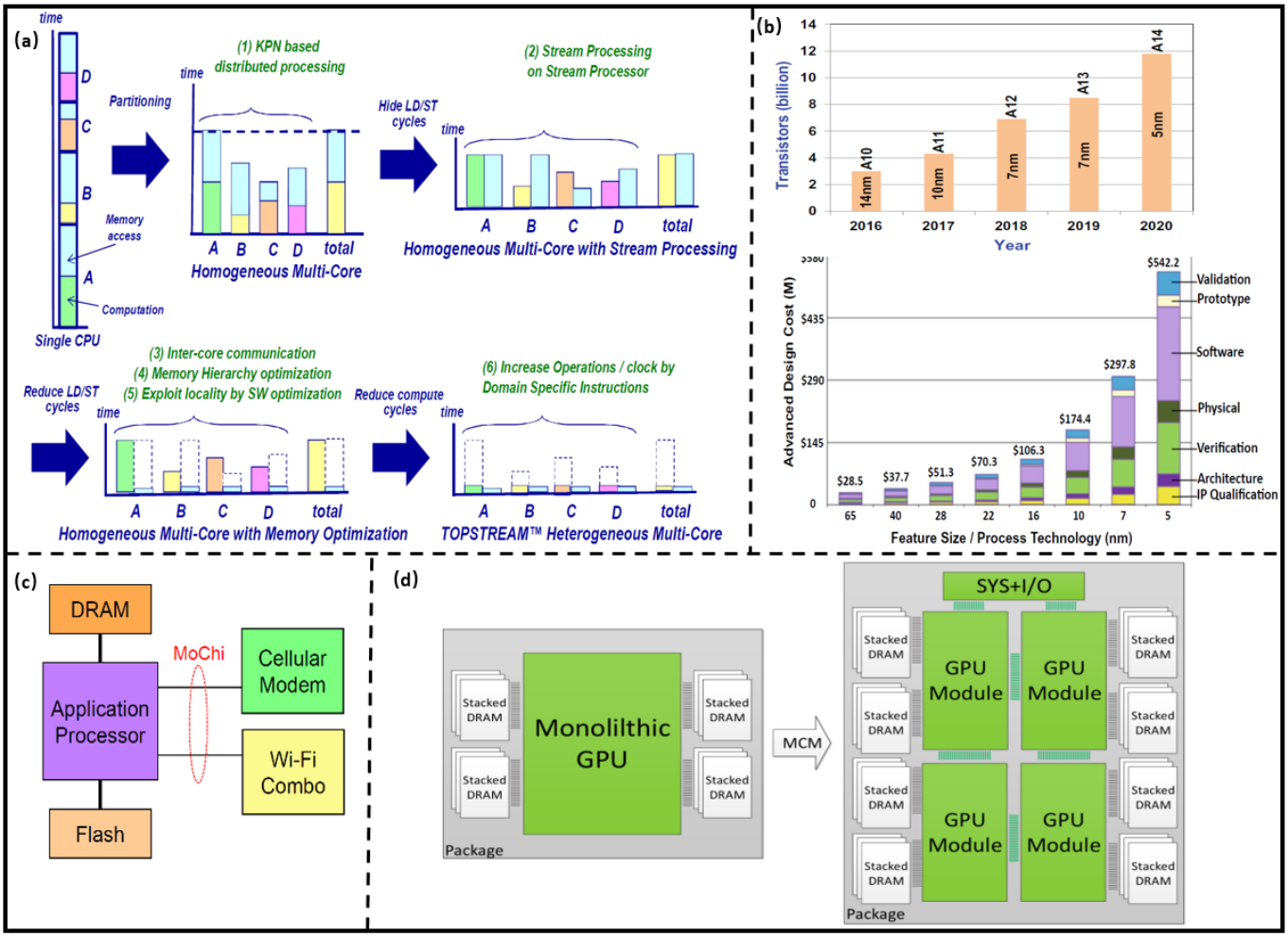

Figure 2. (**a**) Performance comparison of homogeneous and heterogeneous multi-core architectures. (Reprinted from [15], Copyright 2012, with permission from IEEE); (**b**) iPhone processor development and cost analysis. (Reprinted from [16], Copyright 2021, with permission from Springer); (**c**) Mochi Processor architecture. (Reprinted from [17], Copyright 2019, with permission from IEEE); (**d**) GPU design technology based on Chiplet. (Reprinted from [16], Copyright 2021, with permission from Linley Group, Inc.)

The computing system architecture designed by modularization and generalization Chiplet can achieve higher performance, lower complexity and cost, and it can reduce the parasitic effect by using 2.5D or 3D integration technology. The computing system architectures based on Chiplet are a key research aspect of computing architecture.

2.1. Computing Architecture Integrated with 2.5D Technology

Nurvitadhi et al. [17] compared the performance of GPU (NVIDIA Volta, 10 nm) and Chiplet-based Field Programmable Gate Array (FPGA) integrated with 2.5D integrated technology (Intel Stratix 10, 14 nm) in a given computing task (FP32, INT8). The results show that the computing powers of GPU and Chiplet-based FPGA are 6% and 57% of their peak, respectively. The delay and energy efficiency of FPGA are 1/16 and 34× of GPU, respectively. It shows that the computing system architecture based on Chiplet has higher performance and hardware utilization, as well as lower cost. Sehat [18] proposed the mobile architecture called MoChi, which is integrated by computing Chiplet (such as CPU) with advanced technology and other Chiplets with mature technology. The system resource sharing and communication can be achieved by the MoChi interface, as shown in Figure 2c. The architecture has lower design complexity compared with traditional

monolithic SoC architectures. The core of the computing architecture is the integration of Chiplet from different vendors through the interface, and the advanced 2.5D integration technology can be used to reduce the number of pins and packaging costs.

Arunkumar et al. [19] decomposed a single-chip multi-core GPU into multiple GPU Chiplets to design a high-performance computing architecture, as shown in Figure 2d, which can improve computing speed by 22.8% with an energy efficiency of 0.5 pj/bit. The utilization ratio of hardware resources is increased for the GPU and DRAM Chiplet, so the dark silicon effect is alleviated. Further, the yield of the wafer is improved for the larger GPU is decomposed into multiple GPU Chiplet with a smaller area.

Based on the requirements of the National Aeronautics and Space Administration (NASA) in reconfigurable computing architecture, Mounce et al. [20] proposed a high-performance spatial heterogeneous computing architecture for space applications, as shown in Figure 3b. High-speed communication between Chiplets is implemented by standard communication protocol and bus. In addition, they proposed that the Chiplet-based approach can build more powerful heterogeneous systems with radio frequency (RF) Chiplet and FPGA, and further achieve a smaller size and lower cost. This indicates that the computing system architecture based on Chiplet can take advantage of different hardware resources and achieve higher system scalability. The system performance can be further improved through advanced packaging. Vijayaraghavan et al. [21] designed a Chiplet-based computing system for climate prediction, as shown in Figure 3a. It integrates high-throughput and energy-efficient GPU Chiplet, high-performance multi-core CPU Chiplet, and large capacity 3D memory. The system can achieve a bandwidth of 3 TB/s and power consumption of 160 W at 1 GHz. Lin et al. [22] designed a Chiplet-based high-performance computing architecture, which integrates four 7 nm ARM Cortex-A72 cores in two computing Chiplets. The Chiplet communication can be achieved through the parallel channels formed by Low-voltage-InPackage-INterCONnect technology. The bandwidth rate and density are 320 GB/s and 1.6 Tb/s/mm^2 under 4 GHz, respectively. The lower roughness and smaller line spacing for the Chiplet connection can be achieved by InFO_SoW technology. The bandwidth density and power distribution network (PDN) impedance are 2× and 30% more than flip-chip multi-chip-module (MCM) interconnection, and interconnection power consumption is reduced by 15% [23], as shown in Figure 3c. In Agilex series FPGAs, the core Chiplet and other Chiplets were interconnected using Embedded Multidie Interconnect Bridge (EMIB). Compared to Stratix10, the delay is reduced by 2.5×, and the bandwidth density and energy efficiency are improved 5.68× and 2.84×, respectively [24]. The interconnect technology has no limitation on the Chiplet area compared to industrial standard 2.5D multi-chip interconnection, which permits flexible placement. The technology can improve the signal and power integrities by isolating signal and power paths, and reducing the cost due to without addition through silicon via (TSV) [25], as shown in Figure 3d. The power consumption of data transfer takes up a large proportion of the total computing system energy. One promising way to improve energy efficiency and bandwidth is to optimize the Chiplet interconnection. The commonality between InFO_SoW and EMIB lies in the preparation of high-density TSV and re-distribution layer (RDL), within the interposer. The Chiplets, interposer, substrate, and printed circuit board (PCB) were integrated by 2.5D technology, so the bandwidth, energy efficiency, signal, and power integrities were improved effectivity.

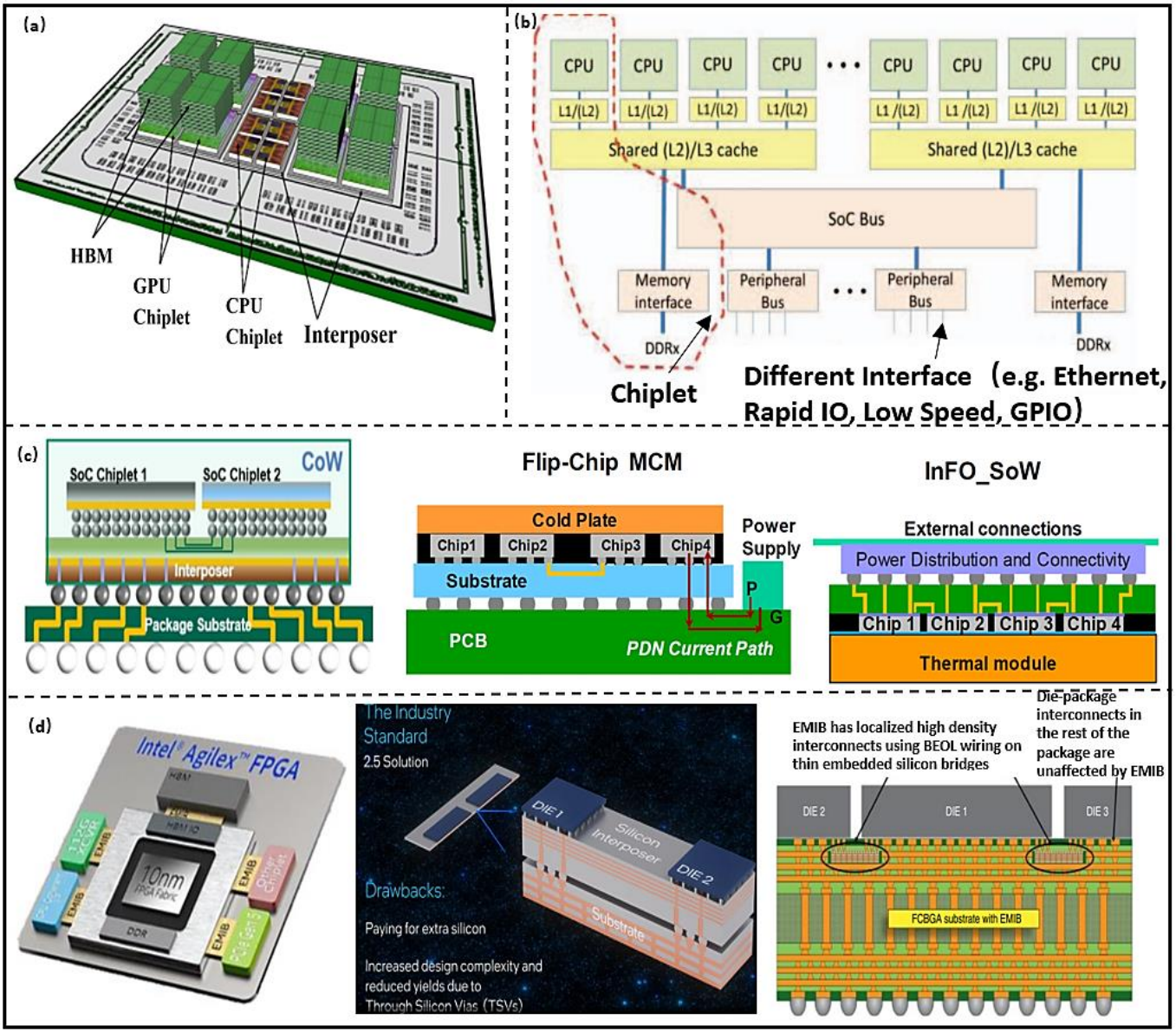

Figure 3. (**a**) Computing system architecture model based on Chiplet. (Reprinted from [19], Copyright 2017, with permission from IEEE); (**b**) Chiplet planning technology. (Reprinted from [20], Copyright 2016, with permission from IEEE) [20]; (**c**) TSMC high performance computing architecture based on Chiplet. (Reprinted from [21], Copyright 2017, with permission from IEEE); (**d**) Chiplet-based integration architecture. (Reprinted from [22], Copyright 2020, with permission from IEEE).

Zaruba et al. [26] used four computing Chiplets and high bandwidth memory (HBM) Chiplet (8 GB L1 Cache and 27 MB shared L2 memory) to construct computing architecture for high-precision floating-point computing. The computing architecture can be switched between high-performance and high-efficiency modes by reconfiguration. The peak efficiency is larger than 4 TDPflop/s, and power consumption is 25% lower than NVIDIA Volta (7 nm). The efficiency of the architecture is two times and three times that of Intel i9-9900K (14 nm) and ARM N1 (7 nm), respectively. The results show that computing architectures based on Chiplet are more easily integrated with large memory and have a high configurability. Due to the higher modularity of Chiplet, the computing system architecture can be configured in various modes according to the applications. The computing architecture has higher reconfigurability and scalability compared with the traditional SoC-based computing system. It requires co-design of software and hardware,

and there is a certain design complexity. Fortunately, there are already solutions for these problems; therefore, the Chiplet-based reconfigurable computing system design technology has obvious technical advantages.

2.2. Computing Architecture Integrated with 3D Technology

Since technology scaling cannot improve the performance of digital Chiplet (CPU compute die) and analog Chiplet (IO Chiplet and memory Chiplet) in the same proportion without increasing the cost. The design method of computing architecture based on Chiplet achieves the optimization of performance and cost by selecting the combination of Chiplet with the best technology. Further, it is necessary to reduce the size of electronics driven by small form factors and the lightweight of wearable (motion watch, bodily function devices, etc.), portable electronics (mobile, laptop, etc.); therefore, more and more computing systems are designed with 3D architectures. The computing system performance can be improved by co-design of 3D architectures and advanced packing technology.

This approach is widely used by AMD in high-performance computing (HPC) system design, enabling rapid development of two products through a different number of Chiplets combinations, such as Rome and Matisse [9], as shown in Figure 4a. The most obvious advantages are that the design of the computing system is simplified and the time to market of product is reduced. The other merits of the architecture include the fact that the digital Chiplet is backward compatible with complex interfaces and the memory Chiplet; that is, the optimal combination of computing and memory Chiplets can be selected according to the computing ability requirements, which has higher scalability and reconfigurability compared with the traditional multi-core architecture and SoC computing system architectures. In order to improve energy efficiency, Kadomoto et al. [27] proposed a method to realize Chiplet communication using the mutual coupling effect of on-chip inductor coils, and fabricated a communication network using 0.18 μm process. The maximum bandwidth can reach 1.6 Gb/s, and the time variation is 3%. The total power consumption is 14.5 mW. The computing architecture has potential in medical microrobots. Although the inter-chip communication based on mutual inductance simplifies the routing design; however, electromagnetic coupling in a small volume leads to signal timing deterioration; therefore, this method requires a sufficient shielding design, which can increase the design difficulty. Burd et al. [28] proposed the infinity fabric (IF) technology to connect Chiplets for higher scalability and configurability in a computing system. It combines scalable data fabric (SDF) and scalable control fabric (SCF) as a critical enabler and utilizes 3D package routing layers to support more complex connections. The in-package bandwidth can achieve 256 GB/s with 534 IFs, and its energy efficiency is 1.2 pj/bit (2 pj/bit for EMIB). CEA-LETI [29] developed a 96-core processor by stacking 28 nm computing Chiplet on the 65 nm interposer with a power management module. The Chiplet interconnected with μbump (20 μm pitch), TSV (depth to width ratio of 10:1 and 40 μm pitch) and RDL (10 μm width and pitch of 20 μm). The Chiplets communication can be achieved by extendable Network on Chip (NoC), and the bandwidth is above 3 Tbit/s/mm^2, delay below 0.6 ns/mm [30], as shown in Figure 4b. The Lakefield mobile processor also adopted multiple Chiplets design technology, which consists of the computing and memory Chiplets prepared with optimal technology (10 nm and 22 FFL). All Chiplets were bonded face to face with micro-bumps in 50 μm pitch (Foveros technology) [31]. The parasitic capacitance and resistance are below 250 fF and 70 mΩ, respectively. The data transfer rate bandwidth is up to 500 Mb/s with an energy efficiency of 0.2 pj/b. Foveros technology has good compatibility with EMIB and can be used for high-density interconnection of the same system for more flexible interconnection [32]. IF, NoC, and Foveros are all based on 3D electrical interconnection, and the preparation technology is relatively mature. The performance of the computing system is highly predictable. The computing system can obtain a high bandwidth and energy efficiency at a certain working frequency (The typical value is 1.15 GHz, as shown in Table 1); however, with the increase in operating frequency, the parasitic resistor, capacitor, and inductor of TSV and RDL can degrade the signal integrity. In addition, Joule heat

produced by TSV and RDL can reduce the system reliability; therefore, more optimized interconnect technologies are needed.

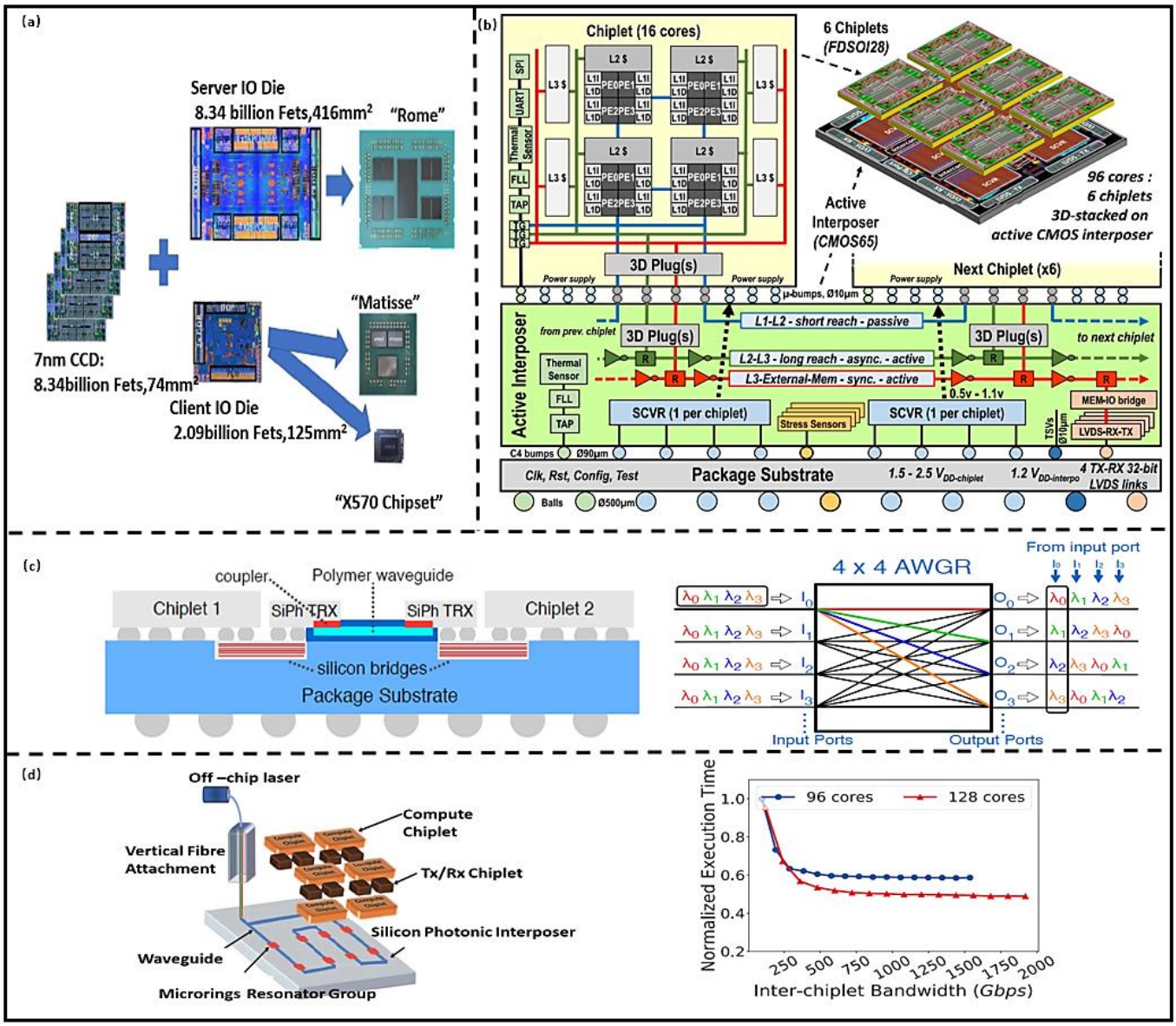

Figure 4. (**a**) AMD processors design technology based on Chiplet. (Reprinted from [9], Copyright 2020, with permission from IEEE); (**b**) INTACT computing architecture based on Chiplet. (Reprinted from [29], Copyright 2019, with permission from IEEE); (**c**) Hybrid optical–electrical interconnection. (Reprinted from [31], Copyright 2020, with permission from IEEE); (**d**) POPSTAR interconnection architecture. (Reprinted from [19], Copyright 2019, with permission from IEEE).

Fotouhi et al. [33] proposed a 3D integration architecture that uses the hybrid Chiplet interconnect technology, as shown in Figure 4c. Silicon bridge is used for a short distance electrical interconnect transceivers (TRXs) Chiplet, and an arrayed waveguide grating router (AWGR) is used for long interconnection in wavelength division multiplexing (WDM). The computing performance is improved by 23%, while the power is reduced by 30%. Narayan et al. [34] designed an optical communication structure for data-parallel transmission between Chiplets by wavelength selection, which can save 38% energy with 1% performance degeneration, and peak bandwidth of 1750 Gb/s, as shown in Figure 4d. AWGR in [34] and interconnection technology in [35] are based on silicon photonic technology, which can realize the selective routing of optical signals by adjusting wavelengths.

The higher data bandwidth, smaller signal delay, less heat, and higher energy efficiency can be achieved compared with the electrical interconnection; however, silicon photonic communication requires a high-power laser source, which is difficult to be integrated on the chip. In addition, the performance of optical devices is greatly affected by the fluctuation of the process, so the reliability is lower than the electrical interconnection. Due to the difficulty of fabrication and integration of silicon photonic devices, optical interconnection technology cannot be widely used; however, the advantages of the technology will drive the development of the integration technology, and it will be more widely used in future computing systems.

Table 1. Comparison of computing architectures based on Chiplet.

	Intel [24]	**TSMC** [22]	**AMD** [9]	**CEA-Leti** [30]	**Intel** [25]	**Bologna** [26]
Product Name	Agilex	-	Ryzen	INTACT	Lakefield	Manticore
Launched Time	201904	201908	201908	202002	202006	202012
Chiplet Technology (nm)	10	7	7 + 12	FDSOI 28	10 + 22 FFL	GF 22 FDX
Chiplet Number	scalable	2	>2	6	1	4
Number of cores/Chiplet	Cortex-A53	4 Cortex-A72	64 (Server) 16 (Cilient)	16	1 Core+ 4 Atom	1024 RISC-V
Area (mm^2)	-	4.4 × 6.2	-	4 × 5.6	-	9
Bandwidth (Max)	32 Gb/s	320 GB/s	~55 GB/s	527 GB/s	~34 GB/s	1 TB/s
Bandwidth density		1.6 $Tb/s/mm^2$	-	3 $Tbit/s/mm^2$	-	-
Frequency (GHz)	1.5	4	~1	1.15	~1	1
Integrated type	2.5D	2.5D	3D	3D	3D	2.5D
Interposer type	Passive	Passive	N/A	Active	Active	Yes
Interconnect pitch (μm)	55	40	-	20	50	20
Delay	~60 ps	-	<9 ns	0.6 ns/mm	-	-
Integration technology	EMIB	CoWoS		F2F	Foveros	-
Yield	High	High	High	High	High	High
Scalability	High		High	High		-
Configurability	Good	Yes	Yes	Yes	alternative	High efficiency/performance
Reusability	High	High	High	High	High	High
Testability			Good	Good	Good	
Power efficiency	-	0.56 pJ/b	2 pJ/b	0.59 pj/b	0.2 pJ/b	50 Gdopflop/sW
Application	Data Center, Networking, Edge Computing	HPC	Server and Desktop Products	Cloud Computing Accelerators	Mobile, PC	Data Center, Networking, Edge Computing.

2.3. Summary

Single-core and homogeneous multi-core architectures handle task parallelization and computing acceleration under lightweight workloads. Heterogeneous computing architectures can improve energy efficiency by integrating the merits of different computing cores, such as CPU–GPU/CPU–NPU; however, multi-core architectures cannot improve computing performance and energy efficiency as further increasing intensive workloads and scaling of technology and the dark silicon effect are made worse as cores increase in number. It can achieve a single optimization for performance, energy efficiency, or scalability. In the Chiplet-based computing system, the Chiplet is prepared with the optimized technology and further integrated with 2.5/3D advanced packing technology, which has high bandwidth and energy efficiency and low data delay. As shown in Table 1, in [22], the computing architecture was constructed with the four Chiplets using 2.5D Chip on Wafer on a substrate (CoWoS) technology, and the bandwidth can be improved to 1.6 $Tb/s/mm^2$ in high-performance computing. In [24], the delay of Agilex can be reduced to 60 ps by using 2.5D integration technology, and the architecture has high configurability and reusability. In [25], the energy efficiency of Lakefield can be improved to 0.2 pJ/b, and the architecture can be configured for PC and mobile processors. In [28], the Chiplets were prepared with the most mature technology among all computing systems; however, the delay can be reduced to 0.6 ns/mm and the bandwidth can be improved to 527 GB/s through 3D integration. In [26], the interconnect pitch between μbumps can be reduced to 20 μm through 2.5D integration, and the maximum bandwidth reaches 1 TB/s. Due to the mature preparation technology of electrical interconnection and higher energy efficiency of

silicon photon interconnection, these two technologies have obvious application advantages in Chiplet-based computing system architecture.

The Chiplet-based 2.5D and 3D integrated architectures have obvious advantages; however, the diversified applications have different focuses. In terms of data bandwidth, the 3D integrated architecture is better, which requires better thermal design. This architecture is more suitable for high-performance computing, for example, data center, networking, server, etc. In terms of cost, the 2.5D integrated architecture does not require a multi-layer Interposer with high-density TSVs; thus, the process is less difficult. The architecture is more suitable for applications such as mobile, laptop, wearable electronics, etc. In terms of Chiplet materials, due to the same thermal expansion coefficient, multiple homogeneous Chiplets adopt the 3D integrated architecture, which is beneficial to improve mechanical reliability; heterogeneous Chiplets are more suitable for the 2.5D integrated architecture (such as EMIB integration technology), which has the higher performance of system heat dissipation, while its area will be increased.

3. Memory Architecture Based on Chiplet

The explosive data eagerly demands memory with larger capacity, bandwidth, and energy efficiency [35–37]; however, the mainstream memory has the relatively matured preparation technology, while the poor integration density and energy efficiency, the emerging memory is just the opposite. Thus, the problems can be solved by optimizing the current memory architecture and introducing emerging non-volatile memory. This section introduces mainstream memory architecture and emerging non-volatile memory architecture, as shown in Figure 5.

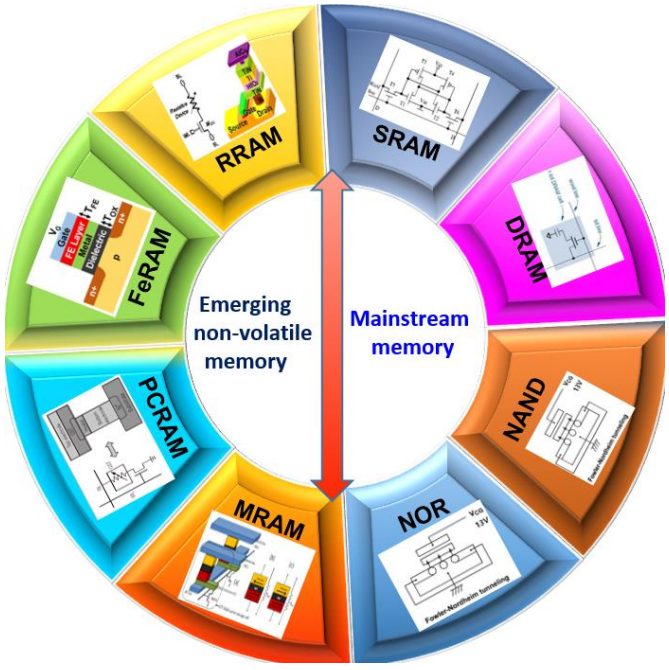

Figure 5. Memory architecture based on different memory. Note: SRAM (Static Random-Access Memory); DRAM (Dynamic RAM); MRAM (Magnetoresistive Random Access Memory); PCRAM (Phase Change RAM); FeRAM (Ferroelectric RAM); RRAM (Resistive RAM).

3.1. Memory Architecture for Storing Data

Due to the big cell area, 2D architectures for mainstream storage have a large form factor, which cannot meet the minimization of electronics. Moreover, the cost of mainstream memory is increasing as the technology shrinks while its endurance is decreasing [38,39]. Koh et al. [40] proposed Flash retention, which is decreased as the technology scales, as

shown in Figure 6a. Cai et al. [41] found the error proportions in NAND rising with the increase in write/read cycles, as shown in Figure 6b.

Therefore, the 2D architecture of mainstream storage cannot meet the needs of high-performance storage, and the mainstream storage architecture needs to be optimized. Loi et al. [42] proved 3D memory architecture has a smaller delay than 2D architecture with a bus model, and the performance is significantly improved in intensive applications, as the frequency increases. Jun et al. [43] designed the 3D HBM architecture for data storage in parallel computing, as shown in Figure 6c. The 2N-Prefetch data mechanism was also used for improving bandwidth (up to 256 GB/s) in data acquisition. Lee et al. [44] further optimized the 3D memory data channel, which can increase the bandwidth of HBM 1 and HBM 2 by 4.6× and 9.1× compared with DDR5, and its power is reduced by 42%. They also proposed the self-repair structure of TSV for increasing testability and reliability, as shown in Figure 6d. Thus, the mainstream memory with 3D architecture can effectively improve bandwidth; however, another difficulty in 3D architecture is the test structure of the memory. Kirihata et al. [45] designed 3D DRAM using TSV for high-density interconnection and developed the electromechanical system (MEMS) probe card for rapid detections, as shown in Figure 6e. A wider and faster bus for data movement was proposed by Micron to simplify the memory control mechanism, which can avoid the complex scheduler and deep queue [46], as shown in Figure 6f. The energy efficiency and bandwidth are 10.82 pj/bit and 128 GB/s, respectively. Shulaker et al. [47] designed a computing system for integrated storage, calculation, and perception of the Chiplet. The 3D integration architecture was adopted to reduce the transmission distance between the data in the Chiplet and improve the signal and power integrities. The whole system was developed by CMOS technology with low preparation difficulty, as shown in Figure 6g. Sandhu et al. [48] proposed a hierarchical memory system, as shown in Figure 6h; it combines the merits of non-volatile memory (NVM) and mainstream memory to improve bandwidth and decrease power.

3.2. Memory Architecture for Processing Data

The energy for data transfer between memory and computing is about 4× that of computing in Von Neumann computing architecture, which reduces the energy efficiency significantly [49,50]. Processing-in-memory (PIM) can complete the data computing and storage in memory with high power efficiency in computation-intensive applications [51–53]. In addition, 3D memory architecture shortens the data transmission path by vertically stacking multiple Chiplets compared with 2D storage architecture and effectively reduces the energy consumption and improves the thermal reliability [47].

3.2.1. PIM Architectures Based on Mainstream Memory

Agrawal et al. [54] designed an 8 TB SRAM Chiplet, which uses parasitic capacitance for accumulating voltages and dot product calculation. The energy-delay product (EDP) is 38% lower than that of Von Neumann computing systems within the acceptable accuracy degradation range (1–5%), as shown in Figure 7a. Sinangil et al. [55] developed the SRAM PIM architecture, which can simultaneously perform multiply and sum computation with the average energy efficiency of 3511 TOPS/W, as shown in Figure 7b. They prepared the SRAM Chiplet with an area of 0.0032 mm^2 using 7 nm technology. Ali et al. [56] designed and prepared a 65 nm SRAM PIM Chiplet, which dynamically uses sparsity of workload to configure the output precision of peripheral circuits to keep data accuracy, as shown in Figure 7c. The energy efficiency is above 120 TOPS/W at 1.1 V, 100 MHz. Srinivasa et al. [57] designed SRAM PIM Chiplet with 3D architecture, as shown in Figure 7d. The read and write stabilities are improved 6.6% and 17.6%. The read and write delay times are reduced 17.5% and 6.6%, and EDP is decreased by 1.6× compared with baseline. The design and preparation technology of SRAM Chiplet is relatively mature. As shown in Table 2, it has the fastest read and write speed and the lowest read and write power consumption; however, the cell is large since it requires four or six transis-

tors to store 1 bit of data. Moreover, the volatility of SRAM requires a continuous power supply, and the transistor generates high static power consumption, which hinders its widespread application.

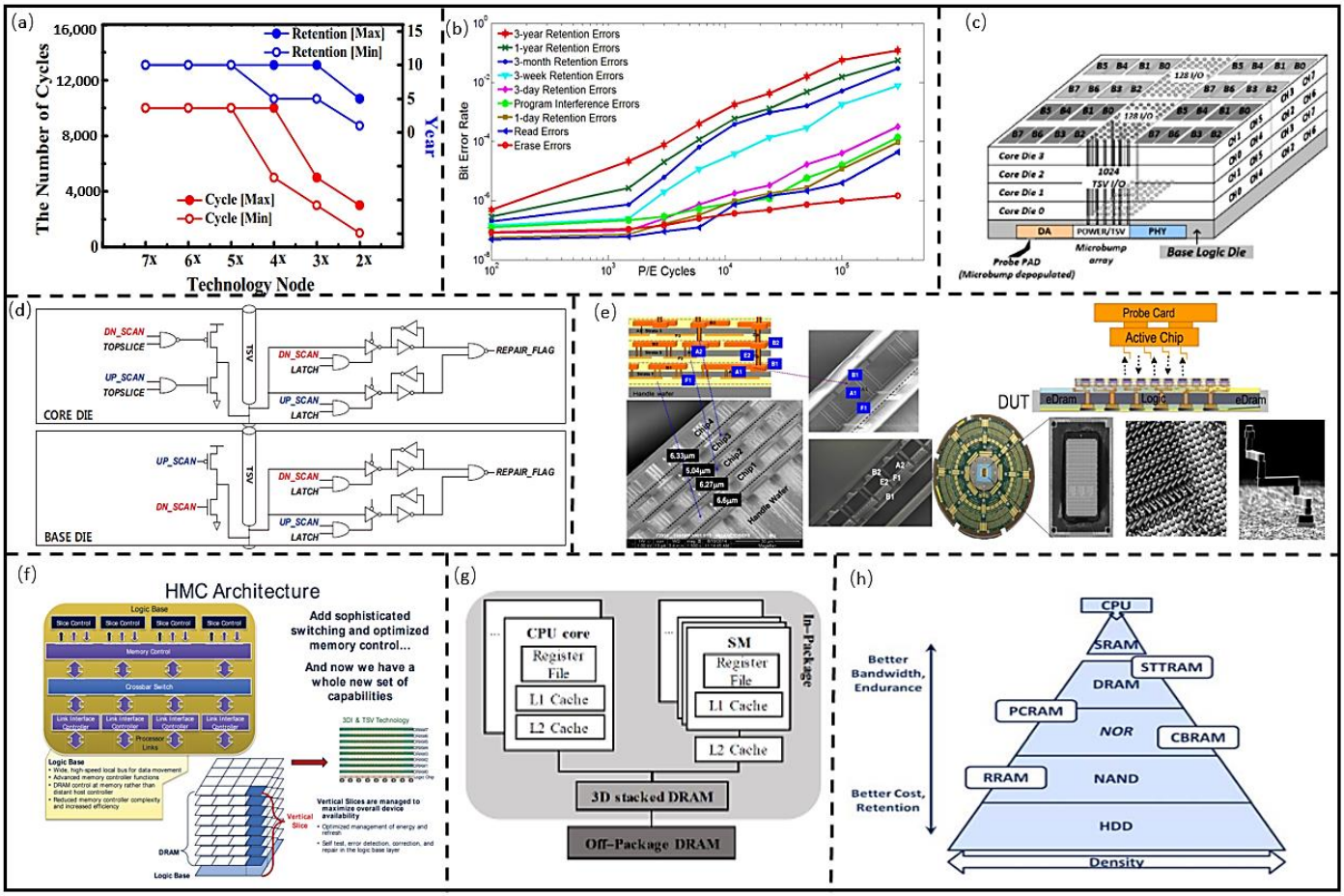

Figure 6. (**a**) Flash performance change with technology. (Reprinted from [40], Copyright 2019, with permission from IEEE); (**b**) NAND performance vary with process technology. (Reprinted from [41], Copyright 2012, with permission from IEEE); (**c**) HBM memory architecture. (Reprinted from [43], Copyright 2017, with permission from IEEE); (**d**) HBM interconnect architecture optimization [44]; (**e**) 3D DRAM Architecture and test architecture (Reprinted from [45], Copyright 2016, with permission from IEEE); (**f**) micron hybrid memory architecture. (Reprinted from [46], Copyright 2011, with permission from IEEE); (**g**) 3D computing systems integrates memory and sensor. (Reprinted from [47], Copyright 2017, with permission from Nature); (**h**) computing systems with hybrid memories (Reprinted from [48], Copyright 2013, with permission from IEEE).

Yu et al. [58] designed the embedded DRAM PIM Chiplet for vector-matrix operation in a neural network, as shown in Figure 8a. In the proposed architecture, the memory node capacitance is increased to improve retention time, which can improve the system energy efficiency up to 552.5 TOPS/W. Werner et al. [59] used vertical optical interconnects (VOIs) to connect the DRAM Chiplet, which eliminates heavily coupling between TSVs, as shown in Figure 8b. Ali et al. [60] designed a DRAM Chiplet to perform data operation in odd rows simultaneously. It improves the parallelism of operation and data throughput, and its performance improves 11.5× compared with baseline. Salkhordeh et al. [61] proposed an analysis model based on the Markov decision method to evaluate the hit ratio and the average lifetime of hybrid memory (DRAM-NVM), as shown in Figure 8c. Compared to the latest simulator, the error is decreased by 2.93%, and speed is improved by 10×. It is a promising way to use the parasitic capacitor of DRAM Chiplet to improve retention time as

well as signal quality. Since DRAM needs to be constantly refreshed, an efficient evaluation method is needed to predict the reliability of DRAM Chiplet, and Markov evaluation technology has merits in evaluation efficiency.

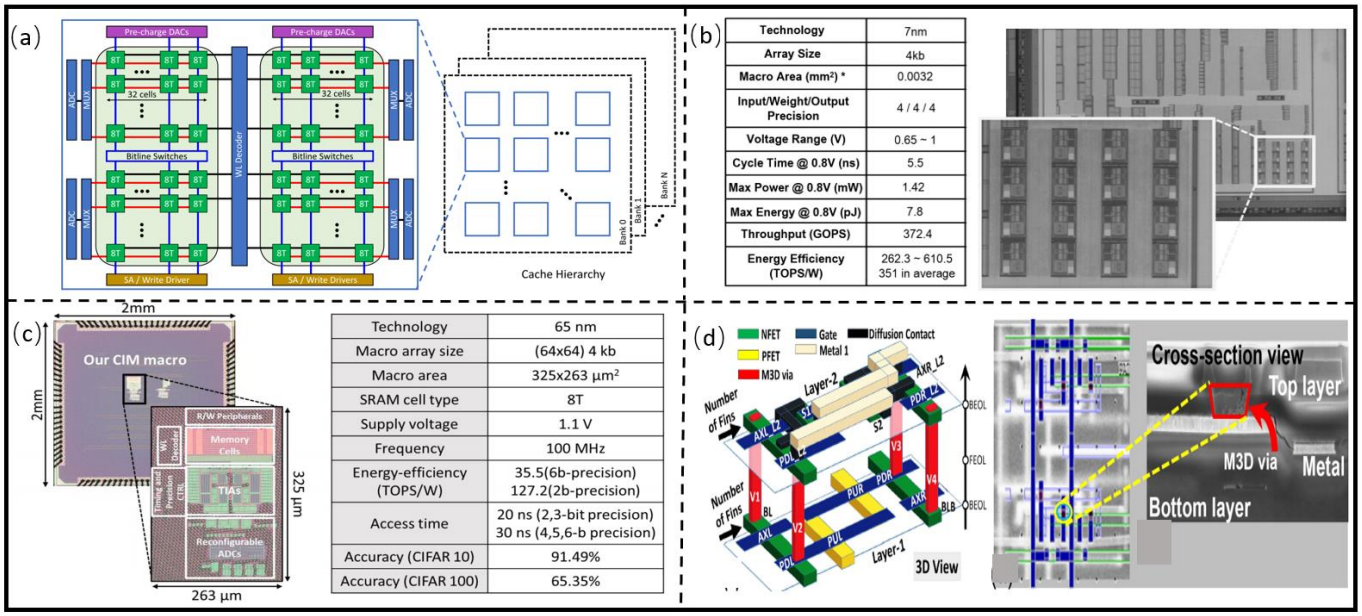

Figure 7. (**a**) An 8T SRAM PIM Chiplet. (Reprinted from [55], Copyright 2020, with permission from IEEE) (**b**) 3D SRAM PIM Chiplet. (Reprinted from [56], Copyright 2021, with permission from IEEE); (**c**) reconfigurable SRAM PIM Chiplet. (Reprinted from [57], Copyright 2021, with permission from IEEE); (**d**) 3D SRAM PIM Chiplet architecture and SEM photo. (Reprinted from [58], Copyright 2019, with permission from IEEE).

3.2.2. PIM Architectures Based on Emerging Nonvolatile Memory

The RRAM architecture designed by Liang et al. [62] could be reconstructed into logical and memory modes. They further developed the adaptive layout and routing algorithms to improve efficient utilization. The power consumption and delays of the proposed architecture are reduced by 1.9× and 2.8×, respectively, and its performance is improved by 5.6× compared with FPGA. Li et al. [63] designed the 3D PIM architecture based on RRAM Chiplet, as shown in Figure 8d, which uses four Chiplets for stacked, and ferroelectric field effect transistor (FeFET) is used as selectors. The voltage, EDP, and area are reduced by 74%, 55%, and 4× compared to 2D memory, respectively. RRAM has good compatibility with CMOS technology and is suitable for high-density integration; however, as a logic Chiplet, the conductive filaments in its structure are affected by the randomness of metal atoms, which generates random noise in the logic mode. The RRAM Chiplet is more suitable as memory. Due to the unique characteristics of hysteretic, FinFET can be designed either as a switch or memory. Yin et al. [64] designed a PIM Chiplet based on FeRAM, whose area and power consumption are 58% and 64% of the SRAM, respectively. Soliman et al. [65] prepared an FeRAM Chiplet with 28 nm CMOS technology; the energy efficiency and latency are 13714 TOPS/W and 0.5 ns, respectively, when 2-bit data operations are performed, as shown in Figure 8e. FeRAM has low read/write time and power consumption and has the best compatibility with CMOS technology; however, the FeRAM Chiplet has a high cost because its electrode materials are noble metals (Pt, Ir). Angizi et al. [66] designed the Chiplet-based MRAM to solve the multi-period logic problem in PIM architecture. Its energy efficiency and speed are 1.7× and 11.2× than those of ASIC, respectively. Shreya et al. [67] designed the Spin-Orbit Torque MRAM PIM Chiplet based on the voltage control technique. The power and data transfer energy consumption are reduced 53.98% and 2.7%, respectively, compared with traditional structures. The read and

write time and current of MRAM are small, and it is expected to be used as an L2 cache, that is, to supplement the existing cache. Dong et al. [68] proposed a 3D PCRAM Chiplet used for checkpointing in parallel computing, which incurs less than 6% overhead in an exascale computing system by making near-instantaneous checkpoints, as shown in Figure 8g. The PCRAM Chiplet requires a large write current to melt the phase change material. As shown in Table 2, the data retention is affected by the amorphous resistance drift of the phase change material, and the power consumption and speed are inferior to RRAM.

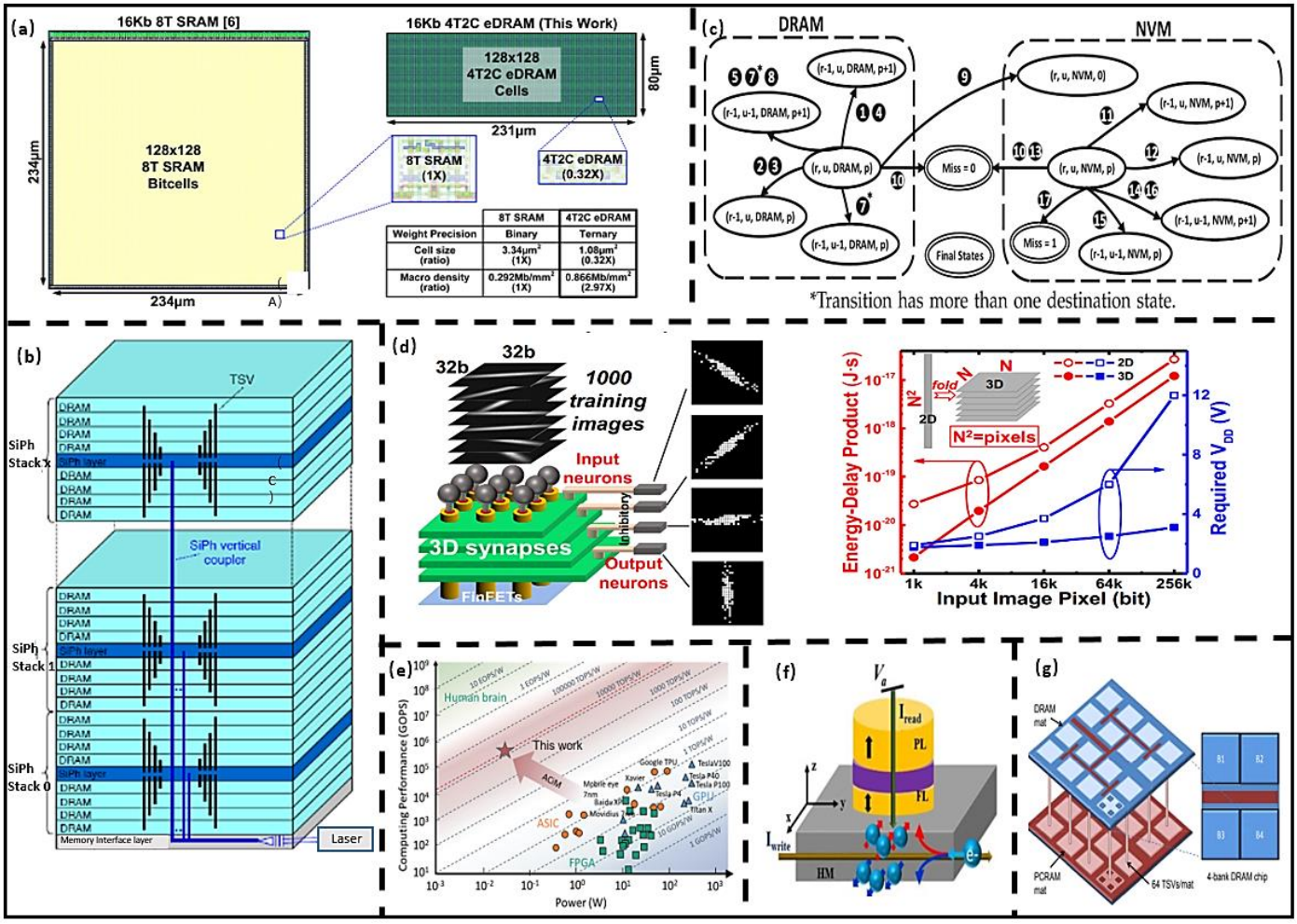

Figure 8. (**a**) Embedded DRAM Chiplet. (Reprinted from [58], Copyright 2021, with permission from IEEE); (**b**) 3D DRAM based on optical interconnect. (Reprinted from [59], Copyright 2019, with permission from IEEE); (**c**) Chiplet evaluation technology. (Reprinted from [61], Copyright 2019, with permission from IEEE); (**d**) 3D RRAM Chiplet, (Reprinted from [62], Copyright 2020, with permission from IEEE); (**e**) FeRAM Chiplet vs. traditional computing system. (Reprinted from [65], Copyright 2020, with permission from IEEE); (**f**) MRAM Chiplet (Reprinted from [68], Copy-right 2021, with permission from ELSEVIER); (**g**) 3D PCRAM Chiplet, (Reprinted from [68], Copy-right 2009, with permission from IEEE).

3.3. Summary

Due to its mature design and manufacturing technology, the mainstream memory Chiplet has been widely used in IoT, PC, mobile, etc. With the technology scaling, the current memory architecture design is dealing with issues regarding a compromise of bandwidth, capacity, power consumption, and cost. As shown in Table 2, due to the shortest read/write time (1 ns), the SRAM Chiplet is used as a cache; however, the cell area is above 160 F^2, which hinders the miniaturization of memory systems. Because of

the mature design and fabrication techniques, the SRAM Chiplet is still used as a small capacity, fast read/write storage (cache). The working voltage and cell area of MRAM and DRAM are similar (voltage: 1 V, 1.5 V, cell area: 10 F^2). The static current of MRAM is smaller than that of DRAM, and it can be used as the main memory. NAND and NOR Flash are preferred for large-capacity memory due to their long read/write and lower cost. The PCRAM and RRAM are expected to complement the existing large capacity storage with a smaller static current (~10^{-4} A), and their cell areas are similar to that of NOR and NADN (10 F^2, 4 F^2). The low read/write energy of FeRAM makes it more promising in low-power applications; however, the current immature technology seriously affects the volume manufacture of NVM. Thus, the Chiplet-based 3D integration technology is an effective method to design high-performance memory. The 3D PIM architecture based on mainstream memory and emerging memory can effectively reduce the distance of data movement, and complete data storage and calculation at the same time, which has obvious application advantages in data-centric computing systems.

Table 2. Memory Chiplet comparison.

	SRAM Chiplet [69,70]	**DRAM Chiplet [71,72]**	**NOR Chiplet [69]**	**NAND Chiplet [73]**	**MRAM Chiplet [74]**	**PCRAM Chiplet [75,76]**	**RRAM Chiplet [77,78]**	**FeRAM Chiplet [67]**
Technology [79]	7 nm	14 nm	28 nm	32 nm	28 nm	28 nm	28 nm	-
Cell area	160–280 F^2	10 F^2	10 F^2	4 F^2	10–20 F^2	5–20 F^2	4–10 F^2	15–20 F^2
Voltage (V)	<1	~1	~10	~15	<1.5	<2	1–3	~1
Current (A)	~10^{-5}	~10^{-5}	~10^{-7}	~10^{-7}	~10^{-5}	~10^{-4}	~10^{-4}	~10^{-6}
Read time (ns)	~1	~10	~10	~10	~10	~10	~10	<10
Write time	~1 ns	~10 ns	10 µs–1 ms	~1 ms	~10 ns	~50 ns	~10 ns	~10 ns
Write energy	~fj	~10 fj	~100 pj	~10 pj	0.1 pj	~10 pj	~0.1 pj	~0.1 pj
Endurance	~10^{16}	~10^{16}	~10^{5}	~10^{5}	~10^{15}	~10^{9}	10^{6}–10^{12}	~10^{10}
Retention	N/A	~64 ms	>10 y	>10 y	>10 y	>10 y	>10 y	>10 y
Static power	High	High	Medium	Medium	Low	Low	Low	Low
Dynamic power	Low	Low	Medium	Medium	Medium	Medium	Medium	Medium
Anti-radiation	Low	Low	Very low	Very low	High	High	High	High
No volatility	NO	NO	Yes	Yes	Yes	Yes	Yes	Yes

Note: F is feature sizes.

4. Conclusions and Perspectives

In this paper, the Chiplet-based computing system architectures with 2.5D and 3D integration technology are introduced, and their characteristics and performance indexes are summarized. The mainstream and emerging NVM memory architectures are also introduced, and their structures and key parameters are summarized. The advantages and disadvantages of the three computing architectures, including single-core, multi-core, and Chiplet-based, are summarized and compared, and their applications are shown. The single-core computing system architecture has a short design cycle and low cost and is mainly applied to light-load computing. Multi-core computing system architecture can meet the requirements of multi-task parallelization and high-precision computing; however, the performance improvement gradually slows down as technology scaling, and the dark silicon effect is obvious. The Chiplet-based computing system architecture has merits of high scalability, energy efficiency, and low cost. With the driven by diversified applications, these computing system architectures will develop in parallel, and the Chiplet-based architectures design method will gradually become the mainstream method of computing system architecture in HPC Mobile, etc. The future development and perspectives for the computing system based on Chiplet are summarized as follows:

(1) Advanced integration technology. The Chiplet-based 2.5D and 3D integration technologies will be widely used in high-performance computing systems. The AI-based optimization layout technology for Chiplet can not only improve the integration density but also enhance the thermal routing capability of computing systems.
(2) Standardized interconnection protocols. The standardized interconnection protocols can achieve the normalization and modularization of Chiplet in computing systems, which can decrease the research and development cycle and cost for Chiplet-based computing systems.

(3) Scalable and reconfigurable architecture design technology. The scalable and reconfigurable technology can effectively improve the utilization efficiency of Chiplet, and then improve the utilization range of computing systems, which can also decrease the research and development cycle and cost.

Author Contributions: Conceptualization, G.S. and Y.Z.; resources, D.C. and G.S.; data curation, Y.Z. and C.X.; writing—original draft preparation, G.S. and Y.Z.; writing—review and editing, C.X. and D.C.; visualization, Y.Z. and G.L.; supervision, G.S. and Y.Y.; project administration, G.S. and D.C.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (no 62021004), the State Key Program of National Natural Science of China (no 62134005), the National High Technology Research and Development Program of China (no 2019YFB2204402).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mosquera-Lopez, C.; Agaian, S.; Velez-Hoyos, A.; Thompson, I. Computer-Aided Prostate Cancer Diagnosis From Digitized His topathology: A Review on Texture-Based Systems. *IEEE Rev. Biomed. Eng.* **2015**, *8*, 98–113. [CrossRef] [PubMed]
2. Traub, M.; Maier, A.; Barbehön, K.L. Future Automotive Architecture and the Impact of IT Trends. *IEEE Softw.* **2017**, *34*, 27–32. [CrossRef]
3. Okeme, P.A.; Skakun, A.D.; Muzalevskii, R.A. *Transformation of Factory to Smart Factory*; IEEE ElConRus: Moscow, Russia, 2021; pp. 1499–1503.
4. Design and Visualization. Available online: https://www.nvidia.cn/design-visualization/solutions/engineering-simulation/ (accessed on 26 November 2021).
5. The Tick-Tock Model Through the Years. Available online: https://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html (accessed on 26 November 2021).
6. Nothing Stacks up to EPYC. Available online: https://www.amd.com/zh-hans (accessed on 26 November 2021).
7. Vangal, S.; Paul, S.; Hsu, S.; Agarwal, A.; Kumar, S.; Krishnamurthy, R.; Krishnamurthy, H.; Tschanz, J.; De, V.; Kim, C.H. Wide-Range Many-Core SoC Design in Scaled CMOS: Challenges and Opportunities. *IEEE Trans. VLSI Syst.* **2021**, *29*, 843–856. [CrossRef]
8. IEEE Electronics Packaging Society. Available online: https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2019-edition.html (accessed on 26 November 2021).
9. Naffziger, S.; Lepak, K.; Paraschou, M.; Subramony, M. 2.2 AMD Chiplet Architecture for High-Performance Server and Desktop Products. In Proceedings of the IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 16–20 February 2020; pp. 44–45.
10. Moore, S.K. Chiplets are the future of processors: Three advances boost performance, cut costs, and save powe. *IEEE Spectr.* **2020**, *55*, 11–12. [CrossRef]
11. Stow, D.; Xie, Y.; Siddiqua, T.; Loh, G.H. Cost-effective design of scalable high-performance systems using active and passive interposers. In Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Jose, CA, USA, 13–16 November 2017; pp. 728–735.
12. Schulte, M.J.; Ignatowski, M.; Loh, G.H. Achieving Exascale Capabilities through Heterogeneous Computing. *IEEE Micro* **2015**, *35*, 26–36. [CrossRef]
13. Esmaeilzadeh, H.; Blem, E.; Amant, R.S.; Sankaralingam, K.; Burger, D. Dark Silicon and the End of Multicore Scaling. In Proceedings of the 38th International Symposium on Computer Architecture (ISCA), San Jose, CA, USA, 4–8 June 2011; IEEE: Washington, DC, USA; pp. 365–376.
14. Pal, S.; Petrisko, D.; Kumar, R.; Gupta, P. Design Space Exploration for Chiplet-Assembly-Based Processors. *IEEE Trans. VLSI Syst.* **2020**, *8*, 1062–1073. [CrossRef]
15. Matsumoto, Y.; Morimoto, T.; Hagimoto, M.; Uchida, H.; Hikichi, N.; Imura, F.; Nakagawa, H.; Aoyagi, M. Cool System scalable 3-D stacked heterogeneous Multi-Core / Multi-Chip ar chitecture for ultra low-power digital TV applications. In Proceedings of the IEEE COOL Chips XV, Yokohama, Japan, 18–20 August 2012; pp. 1–3.
16. Lau, J.H. *Semiconductor Advanced Packaging*; Springer: Berlin, Germany, 2021; pp. 414–415.
17. Nurvitadhi, E.; Kwon, D.; Jafari, A.; Boutros, A.; Sim, J.; Tomson, P.; Sumbul, H.; Chen, C.; Knag, P.; Kumar, R.; et al. Why Compete When You Can Work Together: FPGA-ASIC Integration for Persistent RNNs. In Proceedings of the IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), San Diego, CA, USA, 28 April–1 May 2019; pp. 199–207.
18. Microprocessor Report. Available online: https://www.linleygroup.com/mpr/archive.php?j=MPR&year=2015 (accessed on 26 November 2021).

19. Arunkumar, A.; Bolotin, E.; Cho, B.; Milic, U.; Ebrahimi, E.; Villa, O.; Jaleel, A.; Jean, C.J.; Nellans, D. MCM-GPU: Multi-chip-module GPUs for continued performance scalability. In Proceedings of the ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), Toronto, ON, Canada, 24–28 June 2017; pp. 320–332.
20. Mounce, G.; Lyke, J.; Horan, S.; Powell, W.; Doyle, R.; Some, R. Chiplet based approach for heterogeneous processing and packaging architectures. In Proceedings of the IEEE Aerospace Conference, Big Sky, MT, USA, 5–12 April 2016; pp. 1–12.
21. Vijayaraghavan, T.; Eckert, Y.; Loh, G.H.; Schulte, M.J.; Ignatowski, M.; Beckmann, B.M.; Brantley, W.C.; Greathouse, J.L.; Huang, W.; Karunanithi, A.; et al. Design and Analysis of an APU for Exascale Computing. In Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA), Austin, TX, USA, 4–8 February 2017; pp. 85–96.
22. Lin, M.-S.; Huang, T.-C.; Tsai, C.-C.; Tam, K.-H.; Hsieh, K.C.-H.; Chen, C.-F.; Huang, W.-H.; Hu, C.-W.; Chen, Y.-C.; Goel, S.K.; et al. A 7-nm 4-GHz Arm[1]-Core-Based CoWoS[1] Chiplet Design for High-Performance Computing. *IEEE. J. Solid-State Circuits* **2020**, *55*, 956–966. [CrossRef]
23. Chun, S.R.; Kuo, T.H.; Tsai, H.Y.; Liu, C.-S.; Wang, C.-T.; Hsieh, J.-S.; Lin, T.-S.; Ku, T.; Yu, D. InFO_SoW (System-on-Wafer) for High Performance Computing. In Proceedings of the 2020 IEEE 70th Electronic Components and Technology Conference (ECTC), Orlando, FL, USA, 3–30 June 2020; pp. 1–6.
24. Ganusov, K.; Iyer, M.A.; Cheng, N.; Meisler, A. Agilex™ Generation of Intel® FPGAs. In Proceedings of the 2020 IEEE Hot Chips 32 Symposium (HCS), Palo Alto, CA, USA, 16–18 August 2020; pp. 1–26.
25. Keser, B.; Kroehnert, S. Embedded Multi-die Interconnect Bridge. In *Advances in Embedded and Fan-Out Wafer Level Packaging Technologies*; Keser, B., Kroehnert, S., Eds.; Wiley-IEEE Press: Chandler, AZ, USA, 2019; Volume 23, pp. 487–499.
26. Zaruba, F.; Schuiki, F.; Benini, L. A 4096-core RISC-V Chiplet Architecture for Ultra-efficient Floating-point Computing. In Proceedings of the IEEE Hot Chips 32 Symposium (HCS), Palo Alto, CA, USA, 16–18 August 2020; pp. 1–24.
27. Kadomoto, J.; Irie, H.; Sakai, S. A RISC-V Processor with an Inter-Chiplet Wireless Communication Interface for Shape-Changeable Computers. In Proceedings of the IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS), Kokubunji, Japan, 15–17 April 2020; pp. 1–3.
28. Burd, T.; Beck, N.; White, S.; Paraschou, M.; Naffziger, S. Zeppelin: An SoC for multichip architectures. *IEEE J. Solid-State Circuits* **2019**, *54*, 40–42. [CrossRef]
29. Coudrain, P.; Charbonnier, J.; Garnier, A.; Vivet, P.; Vélard, R.; Vinci, A.; Ponthenier, F.; Farcy, A.; Segaud, R.; Chausse, P.; et al. Active Interposer Technology for Chiplet-Based Advanced 3D System Architectures. In Proceedings of the 2019 IEEE 69th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, USA, 28–31 May 2019; pp. 569–578.
30. Vivet, P.; Guthmuller, E.; Thonnart, Y.; Pillonnet, G.; Fuguet, C.; Miro-Panades, I.; Moritz, G.; Durupt, J.; Bernard, C.; Varreau, D.; et al. IntAct: A 96-Core Processor With Six Chiplets 3D-Stacked on an Active Interposer With Distributed Interconnects and Integrated Power Management. *IEEE. J. Solid-State Circuits* **2021**, *56*, 79–97. [CrossRef]
31. Gomes, W.; Khushu, S.; Ingerly, D.B.; Stover, P.N.; Chowdhury, N.I.; O'Mahony, F.; Balankutty, A.; Dolev, N.; Dixon, M.G.; Jiang, L.; et al. 8.1 Lakefield and Mobility Compute: A 3D Stacked 10 nm and 22FFL Hybrid Processor System in 12 × 12 mm^2, 1 mm Package-On-Package. In Proceedings of the IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 16–20 February 2020.
32. Ingerly, D.B.; Enamul, K.; Gomes, W.; Jones, D.; Kolluru, K.C.; Kandas, A.; Kim, G.-S.; Ma, H.; Pantuso, D.; Petersburg, C.; et al. Foveros: 3D Integration and the use of Face-to-Face Chip Stacking for Logic Devices. In Proceedings of the IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 7–11 December 2019; pp. 19.6.1–19.6.4.
33. Fotouhi, P.; Werner, S.; Lowe-Power, J.; Yoo, S.J.B. Enabling scalable chiplet-based uniform memory architectures with silicon photonics. In Proceedings of the International Symposium on Memory Systems (MEMSYS '19), New York, NY, USA, 30 September–3 October 2019.
34. Narayan, A.; Thonnart, Y.; Vivet, P.; Joshi, A.; Coskun, A.K. System-level Evaluation of Chip-Scale Silicon Photonic Networks for Emerging Data-Intensive Applications. 2020 Design. In Proceedings of the Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 9–13 March 2020.
35. Ausavarungnirun, R.; Chang, K.K.; Subramanian, L.; Loh, G.H.; Mutlu, O. Staged memory scheduling: Achieving high performance and scalability in heterogeneous systems. In Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA), Portland, OR, USA, 9–13 June 2012; pp. 416–427.
36. Mutlu, O.; Moscibroda, T. Stall-Time Fair Memory Access Scheduling for Chip Multiprocessors. In Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007), Chicago, IL, USA, 1–5 December 2007; pp. 146–160.
37. Subramanian, L.; Seshadri, V.; Kim, Y.; Jaiyen, B.; Mutlu, O. MISE: Providing performance predictability and improving fairness in shared main memory systems. In Proceedings of the IEEE 19th International Symposium on High Performance Computer Architecture (HPCA), Shenzhen, China, 23–27 February 2013.
38. Hong, S. Memory technology trend and future challenges. In Proceedings of the International Electron Devices Meeting, San Francisco, CA, USA, 6–8 December 2010.
39. Kim, K. Future memory technology: Challenges and opportunities. In Proceedings of the International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA), Taipei, Taiwan, 3–5 December 2007; pp. 5–9.
40. Koh, Y. NAND Flash Scaling Beyond 20 nm. In Proceedings of the IEEE International Memory Workshop, Monterey, CA, USA, 10–14 May 2009; pp. 1–3.

41. Cai, Y.; Haratsch, E.F.; Mutlu, O.; Mai, K. Error Patterns in MLC NAND Flash Memory: Measurement, Characterization, and Analysis. In Proceedings of the 2012 Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, Germany, 12–16 March 2012.
42. Loi, G.L.; Agrawal, B.; Srivastava, N.; Lin, S.; Sherwood, T.; Banerjee, K. A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy. In Proceedings of the 2006 43rd ACM/IEEE Design Automation Conference, San Francisco, CA, USA, 24–28 July 2006; pp. 991–996.
43. Jun, H.; Cho, J.; Lee, K.; Son, H.-Y.; Kim, K.; Jin, H.; Kim, K. HBM (High Bandwidth Memory) DRAM Technology and Architecture. In Proceedings of the 2017 IEEE International Memory Workshop (IMW), Monterey, CA, USA, 14–17 May 2017.
44. Lee, J.C.; Kim, J.; Kim, K.W.; Ku, Y.J.; Kim, D.S.; Jeong, C.; Yun, Y.S.; Kim, H.; Cho, H.S.; Oh, S.; et al. High bandwidth memory(HBM) with TSV technique. In Proceedings of the 2016 International SoC Design Conference (ISOCC), Jeju, Korea, 23–26 November 2016; pp. 181–182.
45. Kirihata, T.; Golz, J.; Wordeman, M.; Batra, P.; Maier, G.W.; Robson, N.; Graves-Abe, T.L.; Berger, D.; Iyer, S.S. Three-Dimensional Dynamic Random Access Memories Using Through-Silicon-Vias. *IEEE.J. Emerg. Sel. Top. Circuits Syst.* **2016**, *6*, 373–384. [CrossRef]
46. Pawlowski, J.T. Hybrid memory cube (HMC). In Proceedings of the IEEE Hot Chips 23 Symposium (HCS), Stanford, CA, USA, 17–19 August 2011; pp. 1–24.
47. Shulaker, M.; Hills, G.; Park, R.; Howe, R.T.; Saraswat, K.; Wong, H.-S.P.; Mitra, S. Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature* **2017**, *547*, 74–78. [CrossRef] [PubMed]
48. Sandhu, G.S. Emerging memories technology landscape. In Proceedings of the 2013 13th Non-Volatile Memory Technology Symposium (NVMTS), Minneapolis, MN, USA, 12–14 August 2013; pp. 1–5.
49. Pedram, A.; Richardson, S.; Horowitz, M.; Galal, S.; Kvatinsky, S. Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era. *IEEE Des. Test* **2017**, *34*, 39–50. [CrossRef]
50. Han, S.; Liu, X.Y.; Mao, H.Z.; Pu, J.; Pedram, A.; Horowitz, M.A.; Dally, W.J. EIE: Efficient Inference Engine on Compressed Deep Neural Network. In Proceedings of the 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, Korea, 18–22 June 2016; pp. 243–254.
51. Deering, S.; Estrin, D.L.; Farinacci, D.; Jacobson, V.; Liu, C.; Wei, L. The PIM architecture for wide-area multicast routing. *IEEE/ACM Trans. Netw.* **1996**, *4*, 153–162. [CrossRef]
52. Yantır, H.E.; Eltawil, A.M.; Salama, K.N. Efficient Acceleration of Stencil Applications through In-Memory Computing. *Micromachines* **2020**, *11*, 622. [CrossRef] [PubMed]
53. Santoro, G.; Turvani, G.; Graziano, M. New Logic-In-Memory Paradigms: An Architectural and Technological Perspective. *Micromachines* **2019**, *10*, 368. [CrossRef]
54. Agrawal, A.; Kosta, A.; Kodge, S.; Kim, D.E.; Roy, K. CASH-RAM: Enabling In-Memory Computations for Edge Inference Using Charge Accumulation and Sharing in Standard 8T-SRAM Arrays. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2020**, *10*, 295–305. [CrossRef]
55. Sinangil, M.E.; Erbagci, B.; Naous, R.; Akarvardar, K.; Sun, D.; Khwa, W.-S.; Liao, H.-J.; Wang, Y.; Chang, J. A 7-nm Compute-in-Memory SRAM Macro Supporting Multi-Bit Input, Weight and Output and Achieving 351 TOPS/W and 372.4 GOPS. *IEEE J. Solid-State Circuits* **2021**, *56*, 188–198. [CrossRef]
56. Ali, M.; Chakraborty, I.; Saxena, U.; Agrawal, A.; Ankit, A.; Roy, K. A 35.5–127.2 TOPS/W Dynamic Sparsity-Aware Reconfigurable-Precision Compute-in-Memory SRAM Macro for Machine Learning. *IEEE Solid-State Circuits Lett.* **2021**, *4*, 129–132. [CrossRef]
57. Srinivasa, S.R.; Ramanathan, A.K.; Li, X.; Chen, W.-H.; Gupta, S.K.; Chang, M.-F.; Ghosh, S.; Sampson, J.; Narayanan, V. ROBIN: Monolithic-3D SRAM for Enhanced Robustness with In-Memory Computation Support. *IEEE Trans. Circuits Syst. I* **2019**, *66*, 2533–2545. [CrossRef]
58. Yu, C.; Yoo, T.; Kim, H.; Kim, T.T.H.; Chuan, K.C.T.; Kim, B. A Logic-Compatible eDRAM Compute-In-Memory with Embedded ADCs for Processing Neural Networks. *IEEE Trans. Circuits Syst. I* **2021**, *68*, 667–679. [CrossRef]
59. Werner, S.; Sebastian, P.; Xian, F.X. 3D photonics as enabling technology for deep 3D DRAM stacking. In Proceedings of the International Symposium on Memory Systems, Washington, DC, USA, 30 September–3 October 2019.
60. Ali, M.F.; Jaiswal, A.; Roy, K. In-Memory Low-Cost Bit-Serial Addition Using Commodity DRAM Technology. *IEEE Trans. Circuits Syst. I* **2020**, *67*, 155–165. [CrossRef]
61. Salkhordeh, R.; Mutlu, O.; Asadi, H. An Analytical Model for Performance and Lifetime Estimation of Hybrid DRAM-NVM Main Memories. *IEEE Trans. Comput.* **2019**, *68*, 1114–1130. [CrossRef]
62. Liang, Y.; Yin, L.; Xu, N. A Field Programmable Process-In-Memory Architecture Based on RRAM Technology. In Proceedings of the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 25–27 December 2020; pp. 2323–2326.
63. Li, H.; Li, K.-S.; Lin, C.-H.; Hsu, J.-L.; Chiu, W.-C.; Chen, M.-C.; Wu, T.-T.; Sohn, J.; Eryilmaz, S.B.; Shieh, J.-M.; et al. Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing. In Proceedings of the IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 14–16 June 2016; pp. 1–2.
64. Yin, X.; Chen, X.; Niemier, M.; Hu, X.S. Ferroelectric FETs-Based Nonvolatile Logic-in-Memory Circuits. *IEEE Trans. VLSI Syst.* **2019**, *27*, 159–172. [CrossRef]

65. Soliman, T.; Muller, F.; Kirchner, T.; Hoffmann, T.; Ganem, H.; Karimov, E.; Ali, T.; Lederer, M.; Sudarshan, C.; Kampfe, T.; et al. Ultra-Low Power Flexible Precision FeFET Based Analog In-Memory Computing. In Proceedings of the IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 12–18 December 2020; pp. 29.2.1–29.2.4.
66. Angizi, S.; He, Z.; Awad, A.; Fan, D. MRIMA: An MRAM-Based In-Memory Accelerator. *IEEE Trans. CADICS* **2020**, *39*, 1123–1136. [CrossRef]
67. Shreya, S.; Jain, A.; Kaushik, B.K. Computing-in-memory using voltage-controlled spin-orbit torque based MRAM array. *Microelectronics* **2021**, *109*, 1–8. [CrossRef]
68. Dong, X.; Muralimanohar, N.; Jouppi, N.; Kaufmann, R.; Xie, Y. Leveraging 3D PCRAM technologies to reduce checkpoint overhead for future exascale systems. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, Portland, OR, USA, 14–20 November 2009; pp. 1–12.
69. Vetter, J.S.; Mittal, S. Opportunities for Nonvolatile Memory Systems in Extreme-Scale High-Performance Computing. *Comput. Sci. Eng.* **2015**, *17*, 73–82. [CrossRef]
70. Mittal, S.; Vetter, J.S. A survey of software techniques for using non-volatile memories for storage and main memory systems. *IEEE Trans. Parallel Distrib.* **2016**, *27*, 1537–1550. [CrossRef]
71. Xia, F.; Jiang, D.; Xiong, J.; Sun, N. A Survey of Phase Change Memory Systems. *J. Comput. Sci. Technol.* **2015**, *30*, 121–144. [CrossRef]
72. Boukhobza, J.; Rubini, S.; Chen, R.; Shao, Z. Emerging NVM: A Survey on Architectural Integration and Research Challenges. *ACM Trans. Des. Autom. Electron. Syst.* **2018**, *23*, 1–32. [CrossRef]
73. Shim, W.; Yu, S. System-Technology Codesign of 3-D NAND Flash-Based Compute-in-Memory Inference Engine. *IEEE J. Explor. Solid-State Comput. Devices Circuits* **2021**, *7*, 61–69. [CrossRef]
74. Koike, H.; Tanigawa, T.; Watanabe, T.; Nasuno, T.; Noguchi, Y.; Yasuhira, N.; Yoshiduka, T.; Ma, Y.; Honjo, H.; Nishioka, K.; et al. 40 nm 1T–1MTJ 128 Mb STT-MRAM with Novel Averaged Reference Voltage Generator Based on Detailed Analysis of Scaled-Down Memory Cell Array Design. *IEEE Trans. Magn.* **2021**, *57*, 1–9. [CrossRef]
75. Dong, Q.; Sinangil, M.E.; Erbagci, B.; Sun, D.; Khwa, W.-S.; Liao, H.-J.; Wang, Y.; Chang, J. 15.3 A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7 nm FinFET CMOS for Machine-Learning Applications. In Proceedings of the IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 16–20 February 2020; pp. 242–244.
76. Endoh, T.; Koike, H.; Ikeda, S.; Hanyu, T.; Ohno, H. An Overview of Nonvolatile Emerging Memories— Spintronics for Working Memories. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2016**, *6*, 109–119. [CrossRef]
77. Hsieh, M.C.; Liao, Y.C.; Chin, Y.W.; Lien, C.-H.; Chang, T.-S.; Chih, Y.-D.; Natarajan, S.; Tsai, N.-J.; King, Y.-C.; Lin, C.J. Ultra high density 3D via RRAM in pure 28nm CMOS process. In Proceedings of the IEEE International Electron Devices Meeting, Washington, DC, USA, 9–11 December 2013; pp. 10.3.1–10.3.4.
78. Akinaga, H.; Shima, H. Resistive Random Access Memory (ReRAM) Based on Metal Oxides. *Proc. IEEE* **2010**, *98*, 2237–2251. [CrossRef]
79. Marinella, M.J. Radiation Effects in Advanced and Emerging Nonvolatile Memories. *IEEE Trans. Nucl. Sci.* **2021**, *68*, 546–572. [CrossRef]