Chip and Package-Scale Interconnects for General-Purpose, Domain-Specific, and Quantum Computing Systems—Overview, Challenges, and Opportunities

Abhijit Das, *Member, IEEE*, Maurizio Palesi, *Senior Member, IEEE*, John Kim, *Senior Member, IEEE*, and Partha Pratim Pande, *Fellow, IEEE*

***Abstract*— The anticipated end of Moore's law, coupled with the breakdown of Dennard scaling, compelled everyone to conceive forthcoming computing systems once transistors reach their limits. Three leading approaches to circumvent this situation are the chiplet paradigm, domain customisation and quantum computing. However, architectural and technological innovations have shifted the fundamental bottleneck from computation to communication. Hence, on-chip and on-package communication play a pivotal role in determining the performance, energy efficiency and scalability of general-purpose, domain-specific and quantum computing systems. This article reviews the recent advances in chip and package-scale interconnects due to the change in architecture, application and technology. The primary objective of this article is to present the current status, key challenges, and impact-worthy opportunities in this research area from the perspective of hardware architectures. The secondary objective of this article is to serve as a tutorial providing an overview of academic and industrial explorations in chip and package-scale communication infrastructure design for general-purpose, domain-specific and quantum computing systems.**

***Index Terms*— Network-on-chip (NoC), network-in-package (NiP), interconnects, silicon interposer, cryo antenna.**

Manuscript received 10 June 2024; revised 22 July 2024; accepted 8 August 2024. Date of publication 19 August 2024; date of current version 13 September 2024. The work of Abhijit Das was supported in part by the European Union's Horizon Europe Program through European Research Council (ERC) under Grant 101042080 (WINC project) and European Innovation Council (EIC) PATHFINDER Scheme under Grant 101099697 (QUADRATURE project). The work of Maurizio Palesi was supported in part by the Spoke 1 "FutureHPC and BigData" of Italian Research Center on High-Performance Computing, Big Data and Quantum Computing (ICSC). This article was recommended by Guest Editor W.-H. Peng. *(Corresponding author: Abhijit Das.)*

Abhijit Das is with the Department of Computer Architecture, Universitat Politécnica de Catalunya, 08034 Barcelona, Spain (e-mail: abhijit.das@upc.edu).

Maurizio Palesi is with the Department of Electrical, Electronics and Computer Engineering, University of Catania, 95124 Catania, Italy (e-mail: maurizio.palesi@unict.it).

John Kim is with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea (e-mail: jjk12@kaist.edu).

Partha Pratim Pande is with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164 USA (e-mail: pande@wsu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JETCAS.2024.3445829.

Digital Object Identifier 10.1109/JETCAS.2024.3445829

I. Introduction

THE advent of many-core systems in the early 2000s marked a revolutionary shift in computer architecture. This transformation was driven by the breakdown of Dennard scaling [1], the limitations of Instruction-Level Parallelism (ILP) [2], and the continued scaling of Moore's law [3]. Ideally, more cores should mean more computing power. However, this isn't automatic, as efficient data sharing among cores requires a low-latency and high-throughput communication infrastructure. Consequently, communication, rather than computation, has emerged as the principal bottleneck within contemporary and future computer architectures [4], [5], [6], [7].

Global wires, shared buses, and monolithic crossbars initially served as the chip-scale communication infrastructure in many-core systems. However, as core counts increased, their limitations in bandwidth and scalability quickly became apparent. Enters the Network-on-Chip (NoC) [32], [33], a groundbreaking packet-based communication infrastructure that interconnects cores through routers. Subsequently, different technologies, like 3D NoC [34], [35], [36], Optical NoC (ONoC) [37], [38], [39], [40], Wireless NoC (WiNoC) [16], [41], [42], etc. are proposed. With its high transfer bandwidth, scalability and reliability, NoC swiftly became the de facto infrastructure for chip-scale communication.

The anticipated end of Moore's law has necessitated a reimagining of future computer architectures as transistor scaling reaches its physical limits. Three prominent approaches have emerged to address this challenge: the chiplet paradigm, domain customisation, and quantum computing. Nevertheless, these innovations significantly impact the existing communication infrastructure. For instance, the chiplet paradigm has led to multi-chiplet systems, integrating multiple chiplets within a single package. This development demands a package-scale communication infrastructure akin to the NoC, leading to the rise of Network-in-Package (NiP) [44]. Similarly, increasing qubit counts to scale quantum computers will require cryogenic and quantum-coherent communication infrastructures at both the chip and package scales [45]. While the path forward is challenging, it is also rife with opportunities for groundbreaking designs in communication infrastructures.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

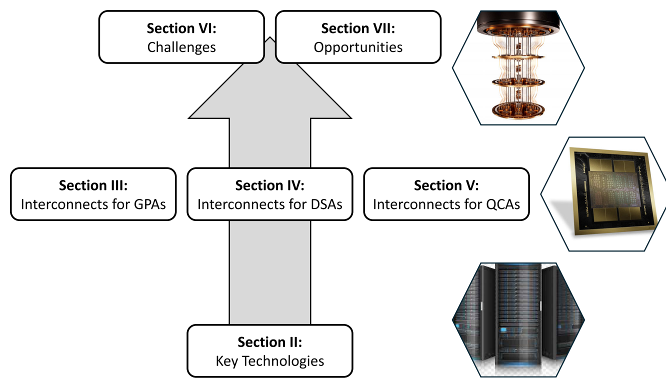


Fig. 1. Organization of the article[1], highlighting the three figures that depict the general-purpose, domain-specific, and quantum computing systems discussed.

This article aims to provide a comprehensive overview of the designs of chip and package-scale communication infrastructure due to changes in architecture, applications, and technology. As shown in Table I, previous surveys have either considered only chip-scale communication or a specific class of applications. The major contributions of this article are:

1) It presents the recent advances in chip and package-scale communication infrastructures for general-purpose, domain-specific and quantum computing systems.
2) It outlines the prevailing challenges in designing chip and package-scale interconnects with respect to performance, scalability and implementation overhead.
3) It proposes potential opportunities for future research and exploration towards designing chip and package-scale communication-aware computer architectures.

The remainder of the article is organised in a bottom-up approach as depicted in Figure 1. In Section II, key interconnection technologies are highlighted to help the reader understand the communication infrastructures discussed in the subsequent sections. Popular and emerging chip and package-scale interconnects for General-Purpose Architectures (GPAs) are presented in Section III. Similarly, Section IV describes the interconnects for major Domain-Specific Architectures (DSAs). Next, in Section V, the emerging interconnects for Quantum Computing Architectures (QCAs) are presented. Then, the prevailing challenges and the potential opportunities in designing chip and package-scale interconnects with respect to performance, scalability and implementation overhead are discussed in Sections VI and VII, respectively. Finally, the discussion is concluded in Section VIII. Table II lists frequently used abbreviations and acronyms throughout the article. There are numerous groundbreaking interconnect designs for GPAs, DSAs, and QCAs. This article aims to provide a comprehensive overview of both academic and industrial interconnects, highlighting a selection of particularly intriguing examples.

II. Key Technologies for Chip and Package-Scale Communication Infrastructures

This section highlights key interconnection technologies widely explored in academia and industry. It is not an exhaustive list nor a detailed description. Instead, it offers a brief overview to help understand the interconnect infrastructures covered in the subsequent Sections III, IV, and V.

A. Metallic Wires

Wired interconnects have long been the backbone of chip-scale communication. Copper-based metallic wires are used to transmit signals by charging and discharging them. This technology, also known as Resistive-Capacitance (RC) interconnects, has been the standard due to its simplicity and cost-effectiveness. However, as transistors scale down, the challenges with RC interconnects are exposed. The continuous decrease in wire thickness and spacing increases resistance and capacitance [46], leading to higher delays [47], [48] and power consumption [49], [50], [51]. These drawbacks are especially severe in global wires, as their lengths stay the same or even increase with scaling, while local wires face fewer issues due to their decreasing lengths. Enhancements like fat wires, mixed wire geometry, and repeaters temporarily masked the drawbacks of metallic wired interconnects [22]. However, the increasing demand for a higher bandwidth and scalable interconnect technology has revealed their inherent limitations.

In summary, although metallic wired interconnects remain prevalent in embedded and mobile systems, server-scale and domain-specific computer architectures are increasingly adopting alternative technologies to address the needs of future many-core systems. Among the forefront technologies driving this shift are photonics, wireless, and 3D integration solutions.

B. Nanophotonics

Optical or photonic interconnects are widely used in supercomputers and data centers due to their high bandwidth density and speed-of-light propagation over long distances. Despite advances in photonic interconnects promising CMOS-compatible nanophotonic solutions for chip and package-scale communication [52], [53], [54], [55], solid evidence was lacking until Hummingbird [56]. This breakthrough, the world's first many-core system powered by an ONoC, features a low-latency optical all-to-all broadcast network spanning 64 cores. Hummingbird exemplifies ONoC's potential to revolutionise High-Performance Computing (HPC), Artificial Intelligence (AI) and Machine Learning (ML) applications, offering a powerful and efficient solution for modern computing systems.

The main components of ONoC include light sources, waveguides, modulators, and detectors. Light sources, typically off-chip lasers, couple into on-chip waveguides made of materials with a high refractive index core and low refractive index cladding to guide the light. Modulators convert electrical signals into optical ones, while detectors perform the reverse operation. Micro-ring resonators are commonly used for modulation and detection, enabling Dense Wavelength Division Multiplexing (DWDM) to increase overall throughput by allowing multiple signals to share the same optical waveguide.

Unlike metallic wired interconnects, which struggle with bandwidth and power efficiency limitations, photonic interconnects excel with high bandwidth per channel, ultralow

[1]The article organisation diagram is inspired from [43].

TABLE I
LITERATURE SURVEYS ON CHIP AND PACKAGE-SCALE COMMUNICATION INFRASTRUCTURES

Year	Reference	Architecture		Application		Technology	
		Chip-scale	Package-scale	General-Purpose	Domain-Specific	Classical	Quantum
2005	[8]	✓	✗	✓	✗	✓	✗
2005	[9]	✓	✗	✓	✗	✓	✗
2006	[10]	✓	✗	✓	✗	✓	✗
2007	[11]	✓	✗	✓	✗	✓	✗
2008	[12]	✓	✗	✓	✗	✓	✗
2008	[13]	✓	✗	✓	✗	✓	✗
2008	[14]	✓	✗	✓	✗	✓	✗
2009	[15]	✓	✗	✓	✗	✓	✗
2009	[16]	✓	✗	✓	✗	✓	✗
2010	[17]	✓	✗	✓	✗	✓	✗
2012	[18]	✓	✓	✓	✗	✓	✗
2012	[19]	✓	✗	✓	✗	✓	✗
2013	[20]	✓	✗	✓	✗	✓	✗
2015	[21]	✓	✗	✓	✗	✓	✗
2016	[22]	✓	✗	✓	✗	✓	✗
2017	[23]	✓	✗	✓	✗	✓	✗
2017	[24]	✓	✗	✓	✗	✓	✗
2020	[6]	✓	✓	✗	✓	✓	✗
2021	[25]	✓	✓	✗	✓	✓	✗
2022	[26]	✓	✓	✓	✓	✓	✗
2022	[27]	✓	✓	✓	✗	✓	✗
2023	[28]	✓	✓	✓	✓	✗	✓
2023	[29]	✓	✓	✓	✗	✗	✓
2024	[30]	✓	✗	✓	✗	✓	✗
2024	[31]	✓	✓	✓	✗	✗	✓
2024	This work	✓	✓	✓	✓	✓	✓

TABLE II
FREQUENTLY USED ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ASIC	Application Specific Integrated Circuit
BEOL	Back-End-Of-Line
CMOS	Complementary Metal Oxide Semiconductor
CXL	Compute Express Link
DDR	Double Data Rate
DNN	Deep Neural Network
DSA	Domain-Specific Architecture
DWDM	Dense Wavelength Division Multiplexing
EIC	Electrical Integrated Circuit
FEOL	Front-End-Of-Line
FPGA	Field Programmable Gate Array
GPA	General-Purpose Architecture
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
HPC	High-Performance Computing
IMC	In-Memory Computing
MAC	Multiply-Accumulate
MCM	Multi-Chip Module
ML	Machine Learning
NiP	Network-in-Package
NoC	Network-on-Chip
ONoC	Optical NoC
PE	Processing Element
QCA	Quantum Computing Architecture
QML	Quantum Machine Learning
RF	Radio Frequency
SoC	System-on-Chip
TSV	Through-Silicon Via
WiNoC	Wireless NoC

latency, and low power consumption. However, they also face scalability challenges due to the non-reconfigurable nature of point-to-point waveguides. Despite this, photonic interconnects have the potential to transform chip and package-scale communication through meticulous design and integration. Currently, both standalone photonic and hybrid photonic-electric interconnects are increasingly explored for NiP in multi-chiplet systems [57], [58], highlighting their promising future.

C. Wireless

Wireless technology offers a compelling alternative to traditional wired interconnects by using Radio-Frequency (RF) signals for data transmission. It is increasingly seen as cost-effective compared to photonic interconnects due to RF circuitry's compatibility with CMOS technology, resulting in lower area and power consumption. Despite advances in designing CMOS-compatible nano-scale wireless antennas and transceivers [59], [60], the technology hasn't matured enough for industrial adoption. Consequently, there's no consensus on its viability for chip and package-scale communication, and unlike Hummingbird, no many-core system powered by a WiNoC exists yet. Nevertheless, this technology shows potential for even greater performance and efficiency if transmission frequencies can be increased to the THz/optical range [19].

Wireless interconnects use free space as the communication medium, eliminating the need for physical wires or channels. A WiNoC essentially converts electrical signals into electromagnetic waves via integrated transceivers. These waves propagate through free space using the antennas, enabling one-hop communication to surrounding cores at nearly the speed of light, significantly reducing the transmission latency. The natural broadcast capability of WiNoC is ideal for systems with high multicast communication demands and enhances scalability. The design of WiNoC utilising THz frequencies is becoming feasible and presents numerous advantages [26].

Wireless interconnects address the multi-hop communication bottlenecks of traditional interconnects, providing low latency, but their bandwidth is limited by antenna operational frequencies. Thus, they are proposed to supplement rather than replace wired interconnects, leading to hybrid wired-wireless interconnects. Additionally, wireless interconnects face challenges related to interference and signal integrity, impacting their reliability. Despite these challenges, advancements in materials and designs, such as graphene-based solutions, are improving the technology [61], [62]. Ongoing research highlights both the benefits and design challenges that must be addressed for industrial adoption [42], [63], [64], [65].

D. 2D, 2.5D and 3D Integration

The ever-increasing demand for higher computing power and memory bandwidth, coupled with the slowdown of Moore's law and the escalating cost of silicon, has driven the adoption of advanced packaging and die-stacking technologies in modern computing systems. Key innovations include chiplets, Through-Silicon Vias (TSVs), micro-bumps, silicon interposers, and silicon bridges. These advancements enable more powerful and efficient integration of semiconductor devices, addressing the limitations of traditional 2D scaling.

1) 2D Integration: 2D Multi-Chip Module (MCM) technology addresses the challenge of declining semiconductor manufacturing yields by partitioning a large die into smaller, more manageable chips. This technology enhances yield and reduces costs due to the non-linear relationship between die size and yield [66]. For instance, AMD's first-generation EPYC™ CPU uses MCM to divide a 32-core CPU into four 8-core dies, achieving approximately 40% cost reduction compared to a monolithic design [67], [68]. More recently, NVIDIA's latest Blackwell GPUs use MCM to provide up to 1.8 TB/s of bidirectional bandwidth between GPUs, significantly enhancing performance for HPC and AI applications [69]. These advancements underscore the growing adoption of 2D MCM technology within contemporary architectures in the industry.

2D chiplet technology addresses the increasing costs of newer silicon nodes by partitioning a System-on-Chip (SoC) into multiple smaller dies, each optimised for specific tasks. Unlike MCM, this approach allows for better performance optimisation and cost efficiency using the most appropriate process technology for each chiplet. For instance, AMD's second-generation EPYC™ CPU employs eight smaller chiplets, each implementing eight CPU cores in a 7nm technology node, surrounding a larger input and output (I/O) die manufactured in a more cost-effective 12nm node [67]. The industry is also standardising chiplet interfaces with Universal Chiplet Interconnect Express (UCIe) [70], allowing seamless integration of chiplets from different manufacturers, which is crucial for maintaining performance gains and enabling diverse applications [71]. It is important to note that the terms "MCM" and "chiplet" are now often used interchangeably in contemporary architectures.

2) 2.5D Integration: In MCM and chiplet-based designs, multiple dies are interconnected within a single package to build a large computing system. Typical die-to-die communication links have bandwidths of 10s of GB/s [68]. However, integrating a memory die within the package necessitates bandwidths reaching 100s of GB/s. This requirement calls for technology capable of providing high-speed, high-density interconnections. Interposer technology has garnered significant attention as a leading solution for heterogeneous component integration. Various interposer substrate materials exist, including silicon, organic, and glass [72]. Silicon interposers are widely used due to their superior performance, electrical properties, and compatibility with TSVs. As shown in Figure 2a, a silicon interposer is a passive die with high-density TSVs and metal routing layers that interconnect multiple active dies. It acts as an interface, providing high-bandwidth communication while minimising latency and signal loss. The silicon interposer uses conventional Back-End-Of-Line (BEOL) processes to construct its metal routing layers, offering interconnect densities comparable to any other die and supporting higher bandwidth in a relatively small area [66]. For instance, Samsung's I-Cube4 [73] integrates multiple logic and memory dies on a silicon interposer, reducing latency and enhancing data throughput. This integration is termed "*2.5D*" because, while the dies are 3D stacked on top of the interposer, they remain organised in a 2D layout relative to each other.

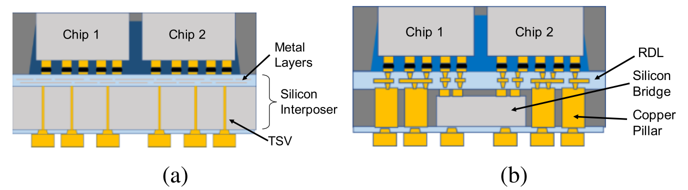
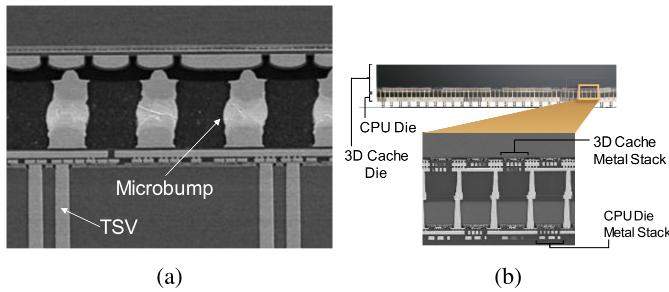


Fig. 2. Cross-section schematic of two chips interconnected by (a) a silicon interposer, and (b) a silicon bridge [66].

One of the constraints with silicon interposers is their size; they must be large enough to accommodate all the active dies intended for 2.5D stacking. In large computing systems, this size requirement can exceed the interposer reticle limit (typically between 800 - 900mm^2), necessitating costly reticle stitching to construct large interposers [74]. A silicon bridge offers an alternative 2.5D integration technology, providing silicon-level wire density with much smaller silicon pieces. As shown in Figure 2b, a silicon bridge is a passive die, like an interposer but significantly smaller, only needing to cover the die-to-die connection interfaces. Outside the bridge's region, conventional copper pillar technology can provide I/O, power, and ground signals directly, eliminating the need for additional manufacturing steps for TSVs. For instance, Intel's EMIB [75] and AMD's EFB [76] technologies provide high-density interconnects in a smaller form factor, enabling efficient data transfer without large interposer cost and complexity.

3) 3D Integration: 3D stacking technology enhances integration density and die-to-die bandwidth by placing multiple active dies directly on top of each other. Microbump bonding is one such technology that uses very small solder connections for vertical die stacking. Figure 3a shows a cross-sectional micrograph of two dies connected vertically with microbumps,



(a) (b)

Fig. 3. Cross-section micrograph of two chips interconnected by (a) microbumps, and (b) hybrid bonding [66].

where the bottom die supports TSVs for external connections. This process can be repeated to create stacks with multiple dies. For instance, industry's largest capacity High Bandwidth Memory (HBM) to date by Samsung [77] used microbump bonding technology to stack 12 memory dies, raising both performance and capacity by more then 50%. However, microbump bonding presents challenges, including higher thermal resistance due to underfill and the additional height from the microbumps and metal connection pads [66].

A recent 3D stacking technology employs a two-phase hybrid bonding process, eliminating microbumps by fusing dies directly [78]. In the first phase, covalent bonds form between the oxides of the two dies' surfaces. The second phase involves a high-temperature copper-copper bonding process, fusing the metal pads on each die directly together. Figure 3b shows a cross-sectional micrograph of a cache die hybrid bonded on top of a CPU die. This approach supports higher interconnect densities (below 10 microns) and superior thermal performance. For instance, AMD's use of TSMC's 3D SoIC technology (hybrid bonding) [79] for its 3D V-Cache stacking achieves 15 times higher interconnect density and 3 times better energy efficiency compared to microbump bonding [80]. Similarly, Graphcore's Bow IPU is the world's first 3D Wafer-on-Wafer (WoW) processor using TSMC's 3D SoIC technology [81]. These advancements position hybrid bonding as crucial for future semiconductor devices, offering significant density, performance, and energy efficiency improvements.

Table III compares the key technologies for communication infrastructures across various metrics (extended from [26]), highlighting their respective capabilities and limitations.

III. Interconnects in General-Purpose Architectures

Interconnects enable seamless data exchange within GPAs between CPUs, memory and peripherals, and are pivotal in determining system performance, efficiency, and scalability. Innovations in NoC and NiP designs, including exploring optical, wireless and packaging technologies, have significantly enhanced interconnect capabilities. Standards like Compute Express Link (CXL) [82] are emerging to streamline the data flow between diverse computing resources further. Besides, the field of secure interconnect design is constantly evolving to address the ever-growing threat of hardware attacks, which can be corroborated by the launch of Intel Trust Domain Extensions (TDX) [83] and AMD Infinity Guard [84] for enhancing data security and privacy. These advancements ensure that modern GPAs meet the growing demands of HPC, AI and other more complex and powerful domain-customised applications [85].

This section presents popular and emerging interconnect infrastructures for processors, cache-coherence, and security.

A. Processor Interconnects

Modern general-purpose processors are designed to meet diverse computing needs with advanced features and architectures. These processors, from leading manufacturers like Intel, AMD, and ARM, utilise multi-core designs, often incorporating four to dozens of cores to enable parallel processing and enhance performance. Adopting chiplet architectures further enhances scalability and customisation, allowing for modular designs that balance cost and performance in the GPAs.

With modern workstations and high-end servers already integrating hundreds of cores [86], [87], [88] and academic efforts showing the feasibility of thousand-core GPAs [89], the role of chip and package-scale interconnects becomes critical. Arguably, the most popular NoC is the mesh-based interconnects due to their simple layout, high path diversity and scalability. A mesh network connects cores in a grid pattern, where each core is connected to its four immediate neighbours (north, south, east, and west). For example, Ampere One [87] uses a mesh interconnect to facilitate communication among its 192 cores. Similarly, Intel® Xeon® scalable processors [86] adopted mesh interconnects to mitigate the increased latencies and bandwidth constraints associated with its previous ring-based interconnect. Another notable mention is ARM's mesh-based interconnect technology, Neoverse CMN-700 [90].

Arteris FlexNoC [91] is a widely utilised third-party interconnect IP by SoC manufacturers who opt not to develop proprietary interconnects. FlexNoC allows configurable topologies and supports various protocol standards, facilitating the seamless integration of heterogeneous components.

The industry is increasingly adopting chiplet-based GPAs due to their scalability and cost-efficiency. Each chiplet can be considered a multi-core processor with its own NoC, and multiple chiplets are interconnected through a Printed Circuit Board (PCB) or an interposer forming an NiP [67], [92], [93]. While it reduces cost, the NiP communication is much slower than NoC [94]. This has spurred significant efforts in industry and academia to develop faster package-scale interconnect technologies. For example, AMD EPYC™ processors utilise a chiplet-based design interconnected via AMD Infinity Fabric™ [95], offering up to 64 GB/s bidirectional bandwidth per link. Similarly, Intel's Ultra Path Interconnect (Intel® UPI) [96] enables scalable systems with multiple processors sharing a single address space, providing two or three high-speed, low-latency links for Intel Xeon processors. NVIDIA employs a mesh-based NiP in Simba, a prototype homogeneous multi-chiplet accelerator system [94]. Moreover, academic research is exploring advanced NiP infrastructures, including wireless [26], [97] and optical

TABLE III
COMPARISON OF KEY TECHNOLOGIES FOR COMMUNICATION INFRASTRUCTURES. THE BEST VALUES FOR EACH METRIC ARE HIGHLIGHTED

Metric	Metallic Wires	Nanophotonics	Wireless	2.5D Integration	3D Integration
Medium Frequency	Wires Baseband	Waveguides Optical	Package Terahertz	Interposers/bridges Baseband	Bumps/metal fusing Baseband
Latency	10–100 ns	10–100 ns	1–10 ns	10–100 ns	1–10 ns
Link Density	Poor	Very good	Not applicable	Very good	Excellent
Bisection Bandwidth	0.1–1 Tb/s	1–100 Tb/s	0.1–1 Tb/s	1–10 Tb/s	1–100 Tb/s
Energy Efficiency	1–100 pJ/bit	0.1–10 pJ/bit	1–10 pJ/bit	1–10 pJ/bit	0.1–10 pJ/bit
Scaling Mechanism	More wires	More waveguides	Frequency-space channels	Larger interposer/more bridges	More stacking
Area Overhead	Low	High	Medium	Medium	Low
Broadcast Capability	Poor	Expensive	Native	Poor	Poor
Implementation Cost	Moderate	High	Medium	Medium	High
CMOS Compatibility	High	Low/BEOL	High/FEOL	High/BEOL	High/BEOL
Design Complexity	Low	High	Low	Medium	High

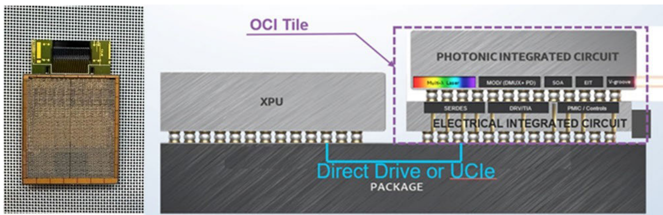


Fig. 4. Top and cross-sectional view of Intel's OCI [100].

technologies [98], [99]. The nascent chiplet paradigm has led to the establishment of multiple standards, such as UCIe [70] and CXL [82], to streamline interconnect infrastructure designs across vendors.

One of the recent breakthroughs in package-scale interconnects is the Intel® Optical Compute Interconnect (OCI) [100]. As shown in Figure 4, Intel® OCI is a fully integrated die stack consisting of a single Intel® Silicon Photonics Integrated Circuit (PIC) with on-chip DWDM lasers and an advanced node CMOS Electrical Integrated Circuit (EIC). No external laser source or optical amplification is required. The first-generation chiplet supports 4 Tbps bidirectionally, with a roadmap to tens of Terabits per second per device.

B. Cache-Coherent Interconnects

Cache coherence in many-core systems ensures that multiple caches maintain a consistent view of memory. As each core in a many-core system has its own private cache, inconsistencies can arise when different caches store different copies of the same memory location. Data inconsistencies can lead to unpredictable behaviour and software errors. Cache coherence protocols address this issue, ensuring that any read of a memory location returns the most recent write.

Cache coherence protocols necessitate the exchange of numerous messages, including multicast, among cores to maintain data consistency. These messages traverse the chip or package-scale interconnection infrastructures. As the number of cores increases, the volume of messages passing through these interconnects also rises, thereby placing additional strain on their capacity. Consequently, there has been extensive research into designing cache-coherent interconnects. For example, the industry has established standards like AMBA Coherent Hub Interface (AMBA CHI) [101] for designing efficient cache-coherent interconnects. As a result, all the latest interconnect technologies, as discussed in Section III-A, are designed to support cache coherence. However, those chiplet-based interconnects restrict cache sharing within individual chiplets to minimise chiplet-to-chiplet communication and simplify coherence management in modern GPAs.

Like FlexNoC [91], Ncore Cache Coherent Interconnect [102] is a popular third-party IP by Arteris. Ncore supports heterogeneous coherency, configurable snoop filters, I/O cache and system memory cache for designing scalable systems.

Academia has long been investigating cache-coherent interconnects, as evidenced by numerous studies and proposals [103], [104], [105], [106], [107], [108], [109], [110]. For instance, the use of optical and RF transmissions via shared nanophotonic waveguides or transmission lines can facilitate broadcast capabilities, which can be utilised to implement traditional snoopy coherence [38] or custom directory coherence protocols [110]. Additionally, the inherent broadcast capability and one-hop latency of wireless technology present opportunities for designing innovative coherence protocols [103], [104], [105].

A notable example is the recent work on WiNoC, which proposed a new directory-based coherence protocol called WiDir [104]. As shown in Figure 5, this approach integrates a WiNoC within a conventional flip-chip package, featuring one vertical monopole antenna and transceiver per core. When the number of sharers for a cacheline exceeds a predefined threshold, the cacheline transitions to a Wireless (W) state, enabling sharing between cores via the WiNoC. Conversely, when the number of sharers drops below the threshold, the cacheline reverts to using the original wired NoC infrastructure.

C. Secure Interconnects

The design of GPAs, particularly SoCs, increasingly relies on third-party IPs to reduce manufacturing and validation costs and meet stringent time-to-market requirements. Despite

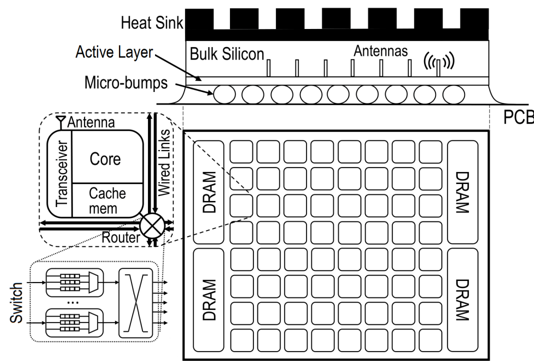


Fig. 5. Top and cross-sectional view of WiDir [104].

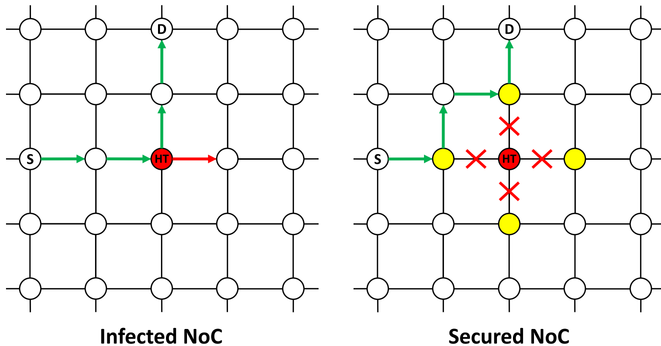


Fig. 6. Attack and its countermeasure in SECTAR [113].

these benefits, incorporating third-party IPs introduces security risks, such as the potential inclusion of malicious implants like hardware Trojans, hidden backdoors, and undocumented vulnerabilities [30]. The Downfall attack [111] is the latest example of a hardware vulnerability identified in contemporary GPAs. Therefore, hardware security must be a fundamental design requirement alongside performance and efficiency [112].

Interconnects enable communication among different components in a GPA and have access to all the shared information of the system. Hence, due to their positional advantage, interconnects are a prime target for security attacks. Over the recent years, academia has proposed a lot of interconnect attacks and possible countermeasures. Due to the limited bandwidth and multi-hop communication of electrical NoC, Denial-of-Service (DoS) is a typical attack [113], [114], [115], [116], [117]. DoS is a scenario where an NoC packet gets indefinitely delayed in the path and never reaches its destination. For example, a recent work, SECTAR [113], proposed trusted NoC communication in the presence of potentially untrusted IPs capable of initiating a DoS attack. As shown in Figure 6, an intermittent Hardware Trojan (HT) implanted on an NoC router tampers with the routing algorithm and enables the misrouting of packets, which results in a DoS. SECTAR proposed a dynamic shielding technique to isolate the HT and reroute the affected packets. In general, using different topologies with redundant paths to critical resources can minimise the effect of DoS attacks.

In many-core systems, multiple applications contend for shared interconnect resources such as links and buffers. The performance characteristics - latency, throughput, and power consumption, of one application can be observed by another co-located application, creating a potential side channel for adversaries to gather and exploit sensitive information. Consequently, side-channel attacks are prevalent in interconnects [118], [119], [120], [121], [122]. Effective countermeasures to these threats include implementing side-channel-aware encryption and authentication mechanisms, employing secure communication protocols, and utilising physical security measures.

The shared and unguided WiNoC mediums present unique security challenges, making them susceptible to eavesdropping, spoofing, and jamming attacks. In contrast, ONoC waveguides offer superior physical security, providing robust protection against interference and signal leakage.

IV. Interconnects in Domain-Specific Architectures

Interconnects are crucial to the performance and efficiency of modern DSAs, especially in AI and ML. These specialised networks enable low-latency, high-throughput communication between Processing Elements (PEs), essential for parallel processing tasks like matrix multiplications and convolutions, as seen in Google's TPUs [123] and NVIDIA's Tensor Cores [124]. The shift to chiplet-based designs is a notable trend, leveraging 2D, 2.5D, and 3D integration technologies to interconnect heterogeneous components with high bandwidth and low latency. Furthermore, the adoption of HBM in GPUs and other chips illustrates the importance of advanced packaging technologies. HBM uses microbumps to stack memory dies, providing significantly higher bandwidth and energy efficiency than traditional DDR memory. These innovations in interconnect technologies are critical as DSAs evolve to meet the increasing demands of AI, ML, and other domain-specific applications. They ensure that PEs can operate at peak performance without bottlenecks, supporting the rapid advancement of computationally intensive applications.

This section presents the most popular interconnect infrastructures for major DSAs; ASIC, FPGA, GPU and IMC.

A. ASIC Interconnects

Application Specific Integrated Circuits (ASICs), as the name implies, are custom-designed for particular tasks, offering specialised optimisations that result in higher speeds and lower power consumption compared to general-purpose hardware. Consequently, ASICs are the preferred choice for creating powerful and efficient DSAs. For example, Google's TPUs are ASICs specifically developed to accelerate ML tasks.

Due to the consistent data flow in Deep Neural Network (DNN) operations, array-based interconnects are very common in modern DNN accelerators [125], [126], [127], [128], [129]. This array-level operation allows efficient reuse of computing data, such as weights and partial sums, significantly enhancing performance. Array-based interconnects utilise a structured grid of PEs where each PE can directly communicate with its

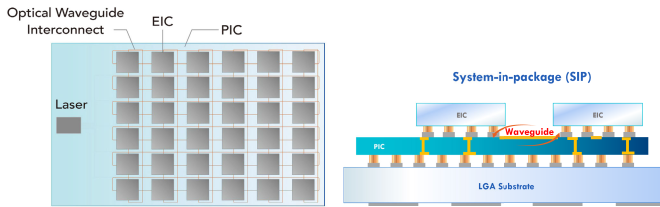


Fig. 7. Top and cross-sectional view of Hummingbird [56].

neighbours, reducing the latency and energy cost associated with accessing higher levels of memory. For example, the Matrix Multiply Unit (MXU) in Google's TPU [125] is arranged as a systolic array of PEs to pass intermediate results to its neighbours in a pipelined fashion, akin to the beats of a heart, which is where the term "systolic" has originated.

Another popular communication infrastructure is the mesh-based interconnects [94], [130], [131], [132]. Among other benefits like simplicity and scalability, mesh networks supports complex communication patterns in DNNs, such as one-to-many and many-to-one, while maintaining the bisection bandwidth. For example, Cerebras uses a 2D mesh fabric for chip-scale communication within its Wafer Scale Engine (WSE) [130]. This fabric connects all the cores on the wafer, enabling efficient data exchange. Similarly, Eyeriss-v2 [131] uses a hierarchical mesh-based NoC to interconnect its PEs.

With emerging DNN workloads showing diverse dataflow preferences [133], reconfigurable interconnects are gaining popularity [134], [135], [136], [137]. These interconnects can dynamically adapt their configuration to match the specific needs of different DNN workloads. For example, MAERI [135] employ building blocks that can be reconfigured via tiny switches, supporting diverse dataflow. The activations and weights are supplied to the PEs via a fully configurable distribution tree, while a configurable adder tree gathers the multiplier outputs.

One of the most recent breakthroughs in ASIC interconnects came with Hummingbird [56]. As shown in Figure 7, it uses 3D stacking technology to integrate a photonic and electronic die into one single package. Each EIC is an AI chip, and all 64 of them are interconnected through optical waveguides. The ONoC is a low latency all-to-all broadcast network enabling higher utilisation of chip computing power due to more efficient communication.

B. FPGA Interconnects

Field Programmable Gate Arrays (FPGA)-based DSAs are gaining popularity due to their balance of performance, energy efficiency and flexibility. They allow for rapid prototyping and customisation, making them ideal for varied and evolving DNN workloads. For example, Xilinx's DPUs [139] are programmable hardware accelerators designed specifically for video and image processing algorithms. They leverage the inherent parallelism of FPGAs to achieve high-performance processing for various video and image workloads.

FPGA-based DSAs typically consist of an FPGA board, a host processor and a memory device (DRAM/HBM), where the FPGA board executes specific workloads. This board requires frequent communication with the host processor for instruction and the memory device to access weights, partial sums, etc. Consequently, Advanced eXtensible Interface (AXI) [140] is one of the most popular interconnects in FPGA-based accelerators [138], [139], [141], [142], [143]. It is a crossbar-based interconnect that allows high bandwidth and low latency connections between the different components of the accelerators. For example, Xilinx's DPU uses a slave AXI interface to access configuration and status registers and a configurable AXI master interface with 64 or 128 bits for data.

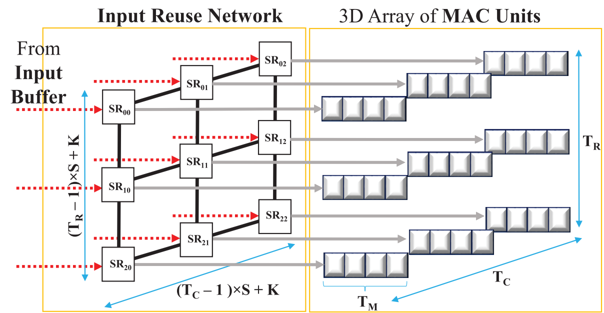


Fig. 8. Input reuse network and compute tile of ICAN [138].

When it comes to the FPGA board, the PEs are usually arranged in a 1D or 2D systolic array, as using short and local interconnects allows the design to realise high frequency [144], [145], [146], [147]. For example, the computing array inside Xilinx's DPU also appears as an array-based interconnection between the PEs. One of the recent works, ICAN [138] slightly deviated from it and instead proposed a 2D-torus interconnect. As shown in Figure 8, ICAN's compute tile is a 3D array of $T_M T_R T_C$ MAC units with no connection among them. The input reuse network is a set of registers connected in a 2D-torus interconnect, where one register is connected to T_M MAC units of the compute tile. The input reuse network can quickly be loaded from the input buffer to provide data to the MAC units. This design mitigates the complexity of internal wiring and input reuse patterns.

Due to the limited resources in a single FPGA and the increasing size of DNN models, Multi-FPGA accelerators are on the rise [139], [148]. Peripheral Component Interconnect Express (PCIe) [149] is preferred to interconnect multiple FPGAs due to its high-speed point-to-point communication. For example, Microsoft's Project Brainwave [148] uses PCIe Gen3 $\times 16$ to interconnect its FPGAs. Besides, a recent work built an accelerator prototype with 7 FPGAs connected using fast and high-bandwidth serial links in a ring network [150].

C. GPU Interconnects

Graphics Processing Units (GPUs) are specialised hardware designed to handle parallel processing tasks efficiently. They contain thousands of small, efficient cores designed for simultaneous multi-threading, making them exceptionally good at handling large-scale parallel tasks, such as matrix multiplications in AI. Hence, GPUs are used to design some of the most powerful DSAs in the industry. For example, NVIDIA's Blackwell and Hopper are GPU-powered AI superchips [124].

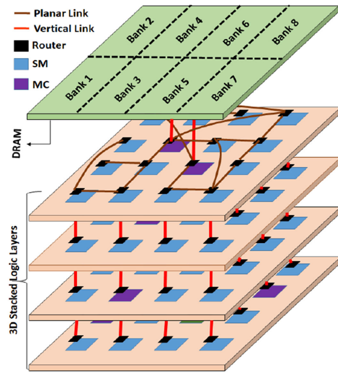


Fig. 9. Block diagram of 3D SWNoC accelerator [151].

GPUs are the powerhouse driving DNNs, excelling in both inference and, most critically, training. Modern GPUs are built for scalability, with multi-GPU setups and distributed training frameworks enabling handling very large models and datasets, significantly speeding up the training process. However, training on multi-GPU clusters faces communication challenges due to the rapid increase in GPU computing power, which creates a gap between computation and communication capacity. Additionally, the rise of larger, more complex DNNs with billions of parameters further strains the network's ability to distribute data efficiently. To overcome this bottleneck, NVLink from NVIDIA [152] is the leading technology in the industry [153], [154], [155], [156]. It is a 1.8 TB/s bidirectional, direct GPU-to-GPU interconnect that scales multi-GPU I/O within a server. For example, NVIDIA Dual GH200 Grace Hopper Superchip [153] uses NVLink to connect two GH200 Superchips. Within a superchip, NVLink-C2C interconnects the NVIDIA Grace CPU and the Hopper GPU. However, since NVLink is a proprietary interconnect, PCIe continues to be the dominant choice for communication between multiple GPUs and the host system (CPU, motherboard) [157], [158], [159], [160]. Nevertheless, PCIe gets slower when used for direct GPU-to-GPU interconnection [161].

Within the GPU chip, the Streaming Multiprocessors (SMs) are usually interconnected using shared buses or mesh networks. One of the recent works proposed a 3D NoC-based GPU accelerator for graph computation [151]. As shown in Figure 9, it uses 3D integration to design a Small-World NoC (SWNoC) that interconnects SMs and Memory Controllers (MCs). The GPU chip is based on the NVIDIA Volta architecture, considering 56 SMs and 8 MCs to give rise to a system with 64 cores. The proposed design reduces average hop count which reduces latency, making SMs available for more computation and thereby improving the throughput.

D. IMC Interconnects

In-Memory Computing (IMC) or Near-Memory Computing (NMC) technology directly integrates processing capabilities within the memory. This eliminates the data movement bottleneck, leading to significant speedups and lower power consumption. While the technology is still maturing, it holds immense potential for applications where efficient AI processing and low power consumption are crucial, like real-time AI tasks and battery-powered edge devices. For example, Mythic AMP™ [163] is an analog IMC-based accelerator for edge.

Bitwise parallel computing and crossbar arrays are the leading architectures in IMC [164]. In bitwise parallel computing, memory rows with operands activate simultaneously, allowing interaction at the sense-amplifier. The sense-amplifier logic is designed according to the required operation to produce the output. The operands and the outputs are placed along the same column (bitline). Activating multiple bitlines in parallel creates a powerful bitwise parallel computing engine [165], [166], [167], [168]. Bitwise parallel computing based IMC cores are usually more area efficient than crossbar arrays.

Crossbar arrays are preferred in resistance-based IMC cores that are analog in nature. Analog input signals from the Digital-to-Analog Converters (DACs) are transmitted through the wordlines and are multiplied by the resistance values of the memory cells. These products accumulate along the bitlines, passing the analog results through Analog-to-Digital Converters (ADCs) to convert them back to the digital domain. Maxtrix-vector multiplication, which is at the center of any ML algorithm, can be easily performed in these IMC cores making them suitable for DNN accelerators [169]. For example, Mythic AMP™ M1076 [163], the industry's first analog IMC chip, is powered by crossbar array-based Mythic ACE cores. It can perform low-power, high-performance matrix-vector multiplication without needing any external DRAM.

Supporting large-scale DNNs necessitates scaling an IMC-based accelerator. This can be achieved in two main ways: (a) increasing the size of an IMC core, or (b) deploying multiple moderately-sized IMC cores. Increasing the size of an IMC core presents significant challenges related to array impedance and programming complexity [170], [171]. As a result, combining multiple IMC cores using interconnects is becoming more popular. One common interconnect method is the mesh network [163], [172], [173], [174]. For instance, available information about the Mythic AMP™ M1076 suggests that a 2D mesh NoC is used to interconnect its 76 Mythic AMP tiles. Besides, there have been recent explorations with wireless NiP-based interconnection among the IMC cores [162], [175], [176]. WHYPE [162], as shown in Figure 10, uses wireless NiP communication technology to interconnect a large number of physically distributed IMC cores capable of executing Hyperdimensional Computing (HDC) algorithms. The controller cores have wireless Transmitters (Tx), while the IMC cores have wireless Receivers (Rx). The package is built on a 2.5D silicon interposer and enclosed in a metallic lid. WHYPE achieves a joint broadcast distribution and computation with performance and efficiency unattainable with wired interconnects, enabling massive system-level parallelisation.

V. Interconnects in Quantum Computing Architectures

Interconnects are vital to QCAs, enabling the seamless transmission of quantum information between qubits, processors, and classical control systems. Essential for maintaining

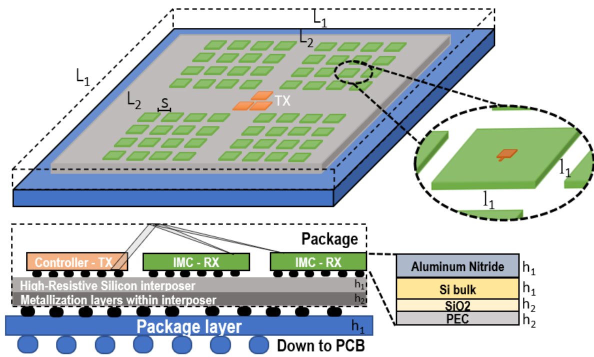


Fig. 10. Top and cross-sectional view of WHYPE [162].

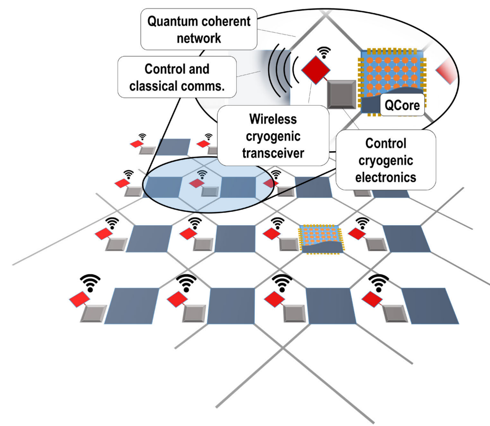


Fig. 11. Vision of a wireless-enabled cryo interconnect [45].

coherence and fidelity, interconnects directly impact the performance and scalability of QCAs. Recent advancements underscore their critical role, highlighting innovations across various quantum systems. For instance, companies like IonQ have achieved significant milestones in ion trap and photonic systems by generating ion-photon entanglement, paving the way for robust quantum networks capable of entangling qubits over longer distances [177]. These photonic interconnects are crucial for linking multiple quantum processing units, enhancing computational power and enabling complex quantum algorithms. Besides, Pasqal and Welinq are developing tailored interconnects for neutral-atom quantum computing, which promises to bring practical quantum advantages to real-world problems [178]. Continued advancements in this field are fundamental to unlocking the full potential of quantum technology, as projected in IBM's Quantum Roadmap [179], [180].

This section presents the promising chip and package-scale interconnect infrastructures for major emerging QCAs.

A. Cryo-Compatible Interconnects

The basic unit of information in a quantum computer, called a qubit, is the quantum equivalent of a bit in a classical computer but way more versatile. The consensus is that addressing any real-world problem using QCAs will require upscaling them to thousands or even millions of qubits [181]. Working towards this goal, Atom Computing recently created the world's first 1000+ qubit QCA, currently populated with 1180 nuclear spin qubits [182]. Similarly, IBM has set its sights on building a 100,000 qubit QCA by 2033 [183].

Scaling up QCAs faces a critical challenge in managing communication between the quantum processor, which hosts the qubits, and the classical computer, which controls and reads them. Most qubit technologies rely on RF signals for state control and readout [184], [185], [186], [187], [188]. Controlling a single qubit typically requires two input lines, while state readout needs multiple I/O lines [189]. Consequently, the extensive wiring necessary for current QCAs quickly becomes a bottleneck as the number of qubits increases [190]. To address this, researchers have proposed cryogenic RF switch [191] and crossbar [192] designs. Beyond wiring, bandwidth and power consumption also pose significant issues. Recent advancements have explored alternative technologies like optical interconnects [193] and wireless readout schemes [194] to enhance bandwidth and reduce power usage. For instance, Atom Computing's Phoenix wirelessly controls its nuclear spin qubits using lasers [195].

Besides the wiring bottleneck, scaling a QCA by densely packing a monolithic chip with a large number of qubits will give rise to issues like cross-talk, complexity of control, and limited yield. Inspired by the classical computers, an alternative to large monolithic chip is a modular multi-core QCA [196], [197], [198]. This approach requires interconnecting multiple moderately sized quantum cores via classical and quantum-coherent links [199]. Microwave link technology is one of the most popular choices for interconnecting quantum cores while designing scale-out QCAs [196], [200], [201]. Here, a microwave pulse excites the source qubit to a higher energy state, allowing it to emit a microwave photon into a resonator. This photon travels through a transmission line to another resonator on a different chip, enabling the entanglement of qubits on separate chips by ensuring the photon emitted by one qubit is absorbed by another. Photonic interconnects are another popular alternative, where optical photons can be transmitted through waveguides or optical fibres over long distances with low loss and interference [202], [203], [204], [205].

One of the latest works envisioned a compact, high-bandwidth and highly reconfigurable wireless NiP for multi-core QCAs [45]. As shown in Figure 11, its main idea is to integrate on-chip cryogenic antennas with RF cryogenic transceivers, creating a wireless network within the quantum computing package. This network is system-wide and broadcast, enabling swift reconfiguration of the underlying architecture for enhanced performance and flexibility.

B. Quantum ML Interconnects

Combining the power of quantum computing and ML can lead to two significant advancements: (a) ML-assisted quantum systems, and (b) quantum-assisted ML models. ML-assisted quantum systems involve applying ML models to the data generated by quantum systems to reconstruct an unknown quantum state [206], optimise quantum error correction [207], etc. Conversely, quantum-assisted ML models, commonly

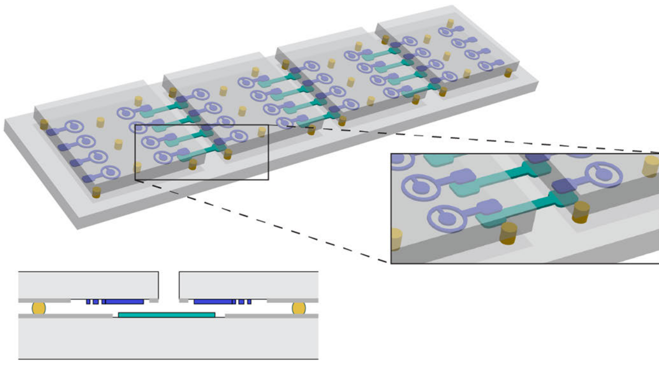


Fig. 12. Cross-sectional view of a modular QCA [199].

known as Quantum ML (QML), leverage quantum systems to speed up the execution of classical ML models [208], [209].

Quantum annealers are widely used for optimisation problems to find the global minimum of an objective function from a pre-defined space [28]. Similarly, universal quantum computers, when integrated with classical systems, excel at implementing clustering algorithms [210], [211]. Nevertheless, the restricted qubit-to-qubit interactions in QCAs impose substantial overhead on both quantum annealers and universal quantum computers, hindering their performance [212]. A wireless technology-based quantum-coherent interconnect infrastructure could facilitate an all-to-all qubit interaction.

QML offers increased speed but is currently limited to smaller datasets [213]. A workaround is to decompose classical ML tasks into subroutines and delegate them to quantum computers [212]. For example, TensorFlow Quantum (TFQ) [214] is a QML library for rapidly prototyping hybrid quantum-classical ML models. However, increasing the qubit count is crucial to handling larger ML datasets effectively on QCAs, highlighting the importance of interconnects.

While the existing literature is rich with QML algorithms and their potential applications, it lacks guidance on designing scalable QCAs to accommodate the increasing size of ML datasets, such as Large Language Models (LLMs). The authors suggest that QML-capable systems could draw inspiration from classical ML accelerators. Chiplet-based QCAs offer the flexibility to customise individual chiplets for specific operations or dataflows, enabling easy scalability. A few works have suggested using chiplets to create modular QCAs [197], [199]. As shown in Figure 12, superconducting qubits (depicted as blue circular structures) on Quantum Integrated Circuit (QuIC) dies are flip-chip bonded on an interposer [199]. This technology has achieved inter-chip coupling rates and entanglement quality comparable to intra-chip qubit-to-qubit coupling. Besides, the parallel progress in cryo-compatible interconnects will enable large-scale deployment of QCAs that could be exploited by ML algorithms in the near future.

VI. Challenges

This section presents some key challenges that should be addressed in designing chip and package-scale interconnects.

- **Rigid NiP:** The move towards specialised chiplets has dramatically increased the complexity of package-scale communication needs, with each chiplet requiring unique interconnection solutions [94]. Meanwhile, the current NiP infrastructures remain rigid and overprovisioned, failing to address these critical interconnection demands.
- **Timing and Synchronisation:** In heterogeneous computing systems, various chiplets often operate on different clocks, creating a significant challenge in maintaining timing accuracy and synchronisation across these distinct clock domains. This issue becomes even more pronounced in high-speed interconnects like photonics.
- **Resource Constrained Vulnerability:** Electrical, optical, and wireless interconnects have distinct vulnerabilities, including DoS, side channels, snooping, and jamming [30]. Implementing security measures to counter these threats within the stringent performance, power, and area constraints of NoCs remains a formidable challenge.
- **High Communication Latency:** Mesh-based interconnects are prevalent in both GPAs and DSAs, yet they often lead to high communication latency, especially for distant cores or PEs [215]. This latency significantly impacts the performance of DNN accelerators, particularly during multicast traffic, hindering overall system efficiency.
- **Energy and Power at the Edge:** Edge devices running specific AI and ML workloads, such as autonomous cars and AR/VR hardware, face significant energy and power constraints. Such devices demand lightweight interconnect infrastructures that seamlessly balance high performance with stringent energy and power requirements.
- **IMC Bandwidth Demand:** IMC-based accelerators have the potential to overcome memory bottlenecks and deliver exceptional performance. However, their full potential is often hampered by limited interconnect capacity [162]. High-bandwidth interconnect infrastructures are crucial to fully leveraging IMC technology in modern DSAs.
- **Scaling Quantum Systems:** Scaling up to millions of qubits involves balancing connectivity and minimising crosstalk and interference between qubits [198]. Efficient routing and scheduling algorithms are needed to handle the growing quantum states moving across multiple chips, along with precise synchronisation mechanisms.
- **Qubit Coherence Time:** Quantum operations must be executed within the coherence time of qubits, rendering QCAs highly sensitive to latency. All operations, including inter-qubit communication, must be completed before decoherence sets in [31]. This constraint is particularly critical in Noisy Intermediate-Scale Quantum (NISQ) devices, which have inherently limited coherence times.
- **Cryogenic Power Budget:** The power consumption of interconnects must be extremely low to align with the limitations of cryogenic environments [216]. The energy required for data transmission should not exceed the heat dissipation capacity of the cryocoolers, necessitating highly efficient interconnect infrastructure designs.

VII. Opportunities

This section presents some impact-worthy opportunities that could be explored in chip and package-scale interconnects.

- **Domain-Specific Interconnects:** With the emergence of DSAs that are designed for a particular application domain, such as HPC, AI and ML, it remains to be seen if the interconnect infrastructure should be general-purpose or domain-specific [217].
- **Wireless-enabled NiP:** Utilising package-scale wireless networking technology can provide efficient and versatile interconnect fabrics, potentially bypassing the physical limitations of wired networks. Wireless interconnects offer dynamic bandwidth sharing, native broadcast support, and reduced latency for long-distance data transfers.
- **Advanced Network Interfaces:** The development of sophisticated network interfaces that integrate protocol adaptivity for synchronisation, error management, and Quality-of-Service (QoS) directly into hardware can enhance the performance and reliability of interconnects.
- **Lightweight Encryption:** Employing lightweight encryption and authentication protocols designed for resource-constrained architectures can effectively counter many security threats in interconnects without imposing significant performance, power, or area overhead.
- **Hybrid Security:** Implementing hybrid interconnects that combine electrical, optical, and wireless technologies can provide robust security by leveraging the strengths of each medium while mitigating individual weaknesses.
- **Reconfigurable Interconnects:** Reconfigurable interconnects, like those used in MAERI [135], can support different dataflows and improve resource utilisation significantly. Photonics technology could be exploited for higher bandwidth, while wireless technology could be exploited for lower latency to cater to application needs.
- **Approximate Computing:** Approximate computing can boost interconnect performance and energy efficiency by tolerating minor inaccuracies. This approach leverages the inherent error resilience characteristic of large DNN models to improve performance and energy metrics.
- **High Bandwidth Interconnects:** Emerging technologies such as photonic interconnects offer higher bandwidth and energy efficiency. This technology can address the limitations of conventional electrical interconnects, making them suitable for DSAs like IMC-based accelerators.
- **Innovative Quantum Interconnects:** Exploring technologies like optical and wireless interconnects can enhance bandwidth and reduce power consumption. They can mitigate thermal disturbances in cryogenic environments and support the scalability of quantum systems.
- **Quantum Hardware-Software Co-Design:** Integrated hardware-software design approaches can optimise quantum system performance. Techniques involving mapping and resource allocation are essential, considering algorithm structure, qubit connectivity, and gate fidelities.
- **Quantum Simulation Tool:** Developing comprehensive simulation tools that integrate various chip and package-scale communication protocols can bridge the gap in current quantum computing research and exploration.

VIII. Conclusion

The evolving computing landscape has shifted from computation-centric to communication-centric challenges, highlighting the critical role of chip and package-scale interconnects in shaping performance, energy efficiency, and scalability. Exploring diverse paradigms like chiplet designs, DSAs, and quantum computing necessitates advanced, flexible interconnect solutions. However, significant challenges persist, including the rigidity of current NiP designs, synchronisation complexity in heterogeneous systems, and security vulnerabilities of electrical, optical, and wireless interconnects. Latency issues in mesh-based interconnects, energy and power constraints of edge devices, and bandwidth demands of IMC-based accelerators further complicate the landscape. QCAs face additional hurdles with coherence times, cryogenic power budgets, and the need for scalable, low-latency interconnects for quantum operations and QML. Addressing these multifaceted challenges is crucial for developing high-bandwidth, low-latency chip and package-scale communication infrastructures that integrate seamlessly with emerging technologies, driving advancements across general-purpose, domain-specific, and quantum computing systems.

References

[1] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, “Design of ion-implanted MOSFET’s with very small physical dimensions,” *IEEE J. Solid-State Circuits*, vol. JSSC-9, no. 5, pp. 256–268, Oct. 1974.

[2] D. W. Wall, “Limits of instruction-level parallelism,” in *Proc. 4th Int. Conf. Architectural Support Program. Lang. Operating Syst. (ASPLOS-IV)*, 1991, pp. 176–188.

[3] G. E. Moore, “Cramming more components onto integrated circuits,” *Electronics*, vol. 38, no. 8, pp. 1–4, 1965.

[4] R. Farjadrad. (2023). *Interconnect is the Root of Generative AI*. Accessed: May 20, 2024. [Online]. Available: https://embeddedcomputing.com/technology/ai-machine-learning/interconnect-is-the-root-of-generative-ai

[5] M. S. Smith, “Single-chip processors have reached their limits-chiplets seem to be the future, but interconnects remain a battleground,” *IEEE Spectr.*, vol. 59, no. 7, p. 11, 2022.

[6] S. M. Nabavinejad, M. Baharloo, K.-C. Chen, M. Palesi, T. Kogel, and M. Ebrahimi, “An overview of efficient interconnection networks for deep neural network accelerators,” *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 10, no. 3, pp. 268–282, Sep. 2020.

[7] S. Salahuddin, K. Ni, and S. Datta, “The era of hyper-scaling in electronics,” *Nature Electron.*, vol. 1, no. 8, pp. 442–450, Aug. 2018.

[8] R. Marculescu, J. Hu, and U. Y. Ogras, “Key research problems in NoC design: A holistic perspective,” in *Proc. 3rd IEEE/ACM/IFIP Int. Conf. Hardw./Softw. Codesign Syst. Synth.*, Sep. 2005, pp. 69–74.

[9] P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, “Performance evaluation and design trade-offs for network-on-chip interconnect architectures,” *IEEE Trans. Comput.*, vol. 54, no. 8, pp. 1025–1040, Aug. 2005.

[10] T. Bjerregaard and S. Mahadevan, “A survey of research and practices of network-on-chip,” *ACM Comput. Surv.*, vol. 38, no. 1, p. 1, Jun. 2006.

[11] J. D. Owens, W. J. Dally, R. Ho, D. N. Jayasimha, S. W. Keckler, and L.-S. Peh, “Research challenges for on-chip interconnection networks,” *IEEE Micro*, vol. 27, no. 5, pp. 96–108, Sep. 2007.

[12] E. Salminen, A. Kulmala, and T. D. Hamalainen, “Survey of network-on-chip proposals,” *White Paper, OCP-IP*, vol. 1, p. 13, Mar. 2008.

[13] D. Atienza, F. Angiolini, S. Murali, A. Pullini, L. Benini, and G. De Micheli, “Network-on-chip design and synthesis outlook,” *Integration*, vol. 41, no. 3, pp. 340–359, May 2008.

[14] R. Marculescu, U. Y. Ogras, L.-S. Peh, N. E. Jerger, and Y. Hoskote, "Outstanding research problems in NoC design: System, microarchitecture, and circuit perspectives," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 28, no. 1, pp. 3–21, Jan. 2009.

[15] R. Marculescu and P. Bogdan, "The chip is the network: Toward a science of network-on-chip design," *Found. Trends Electron. Des. Autom.*, vol. 2, no. 4, pp. 371–461, 2009.

[16] L. P. Carloni, P. Pande, and Y. Xie, "Networks-on-chip in emerging interconnect paradigms: Advantages and challenges," in *Proc. 3rd ACM/IEEE Int. Symp. Netw.-Chip*, May 2009, pp. 93–102.

[17] G. De Micheli, C. Seiculescu, S. Murali, L. Benini, F. Angiolini, and A. Pullini, "Networks on chips: From research to products," in *Proc. 47th Design Autom. Conf.*, Jun. 2010, pp. 300–305.

[18] J. Kim, K. Choi, and G. Loh, "Exploiting new interconnect technologies in on-chip communication," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 2, pp. 124–136, Jun. 2012.

[19] S. Deb, A. Ganguly, P. P. Pande, B. Belzer, and D. Heo, "Wireless NoC as interconnection backbone for multicore chips: Promises and challenges," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 2, pp. 228–239, Jun. 2012.

[20] M. Radetzki, C. Feng, X. Zhao, and A. Jantsch, "Methods for fault tolerance in networks-on-chip," *ACM Comput. Surv.*, vol. 46, no. 1, pp. 1–38, Oct. 2013.

[21] D. Bertozzi, G. Dimitrakopoulos, J. Flich, and S. Sonntag, "The fast evolving landscape of on-chip communication: Selected future challenges and research avenues," *Des. Autom. Embedded Syst.*, vol. 19, nos. 1–2, pp. 59–76, Mar. 2015.

[22] A. Karkar, T. Mak, K. Tong, and A. Yakovlev, "A survey of emerging interconnects for on-chip efficient multicast and broadcast in many-cores," *IEEE Circuits Syst. Mag.*, vol. 16, no. 1, pp. 58–72, 1st Quart., 2016.

[23] A. Ben Achballah, S. Ben Othman, and S. Ben Saoud, "Problems and challenges of emerging technology networks-on-chip: A review," *Microprocess. Microsyst.*, vol. 53, pp. 1–20, Aug. 2017.

[24] S. Werner, J. Navaridas, and M. Luján, "A survey on optical network-on-chip architectures," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–37, Nov. 2018.

[25] F. P. Sunny, E. Taheri, M. Nikdast, and S. Pasricha, "A survey on silicon photonics for deep learning," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 17, no. 4, pp. 1–57, Oct. 2021.

[26] S. Abadal et al., "Graphene-based wireless agile interconnects for massive heterogeneous multi-chip processors," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 162–169, Aug. 2023.

[27] H. Mekawey, M. Elsayed, Y. Ismail, and M. A. Swillam, "Optical interconnects finally seeing the light in silicon photonics: Past the hype," *Nanomaterials*, vol. 12, no. 3, p. 485, Jan. 2022.

[28] Z. Yang, M. Zolanvari, and R. Jain, "A survey of important issues in quantum computing and communications," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1059–1094, 2nd Quart., 2023.

[29] P. Escofet, S. B. Rached, S. Rodrigo, C. G. Almudever, E. Alarcón, and S. Abadal, "Interconnect fabrics for multi-core quantum processors: A context analysis," in *Proc. 16th Int. Workshop Netw. Chip Architectures*, Oct. 2023, pp. 34–39.

[30] H. Weerasena and P. Mishra, "Security of electrical, optical, and wireless on-chip interconnects: A survey," *ACM Trans. Design Autom. Electron. Syst.*, vol. 29, no. 2, pp. 1–41, Mar. 2024.

[31] D. Barral et al., "Review of distributed quantum computing. From single QPU to high performance quantum computing," 2024, *arXiv:2404.01265*.

[32] W. J. Dally and B. Towles, "Route packets, not wires: On-chip interconnection networks," in *Proc. 38th Design Autom. Conf.*, 2001, pp. 684–689.

[33] L. Benini and G. De Micheli, "Networks on chips: A new SoC paradigm," *Computer*, vol. 35, no. 1, pp. 70–78, 2002.

[34] C. Addo-Quaye, "Thermal-aware mapping and placement for 3-D NoC designs," in *Proc. IEEE Int. SOC Conf.*, Sep. 2005, pp. 25–28.

[35] V. F. Pavlidis and E. G. Friedman, "3-D topologies for networks-on-chip," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 15, no. 10, pp. 1081–1090, Oct. 2007.

[36] B. S. Feero and P. P. Pande, "Networks-on-chip in a three-dimensional environment: A performance evaluation," *IEEE Trans. Comput.*, vol. 58, no. 1, pp. 32–45, Jan. 2009.

[37] D. Huang, T. Sze, A. Landin, R. Lytel, and H. L. Davidson, "Optical interconnects: Out of the box forever?" *IEEE J. Sel. Topics Quantum Electron.*, vol. 9, no. 2, pp. 614–623, Mar. 2003.

[38] N. Kirman et al., "Leveraging optical technology in future bus-based chip multiprocessors," in *Proc. 39th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Dec. 2006, pp. 492–503.

[39] D. A. B. Miller, "Device requirements for optical interconnects to silicon chips," *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, Jul. 2009.

[40] K. Ohashi et al., "On-chip optical interconnect," *Proc. IEEE*, vol. 97, no. 7, pp. 1186–1198, Jul. 2009.

[41] O. K. Kihong et al., "The feasibility of on-chip interconnection using antennas," in *Proc. Int. Conf. Comput. Aided Design (ICCAD)*, 2005, pp. 979–984.

[42] A. Ganguly, K. Chang, S. Deb, P. P. Pande, B. Belzer, and C. Teuscher, "Scalable hybrid wireless network-on-chip architectures for multicore systems," *IEEE Trans. Comput.*, vol. 60, no. 10, pp. 1485–1502, Oct. 2011.

[43] S. Abadal, C. Han, V. Petrov, L. Galluccio, I. F. Akyildiz, and J. M. Jornet, "Electromagnetic nanonetworks beyond 6G: From wearable and implantable networks to on-chip and quantum communication," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 8, pp. 2122–2142, Aug. 2024.

[44] K. Lee et al., "Networks-on-chip and networks-in-package for high-performance SoC platforms," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Nov. 2005, pp. 485–488.

[45] E. Alarcón et al., "Scalable multi-chip quantum architectures enabled by cryogenic hybrid wireless/quantum-coherent network-in-package," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2023, pp. 1–5.

[46] (2011). *International Technology Roadmap for Semiconductors (ITRS)*. Accessed: May 20, 2024. [Online]. Available: https://www.semiconductors.org/resources/2011-international-technology-roadmap-for-semiconductors-itrs/

[47] W. J. Dally and J. W. Poulton, *Digital Systems Engineering*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[48] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.

[49] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in *Proc. Int. Workshop Syst. Level Interconnect Predict.*, Feb. 2004, pp. 7–13.

[50] M. B. Taylor et al., "Evaluation of the raw microprocessor: An exposed-wire-delay architecture for ILP and streams," *ACM SIGARCH Comput. Archit. News*, vol. 32, no. 2, p. 2, 2004.

[51] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar, "A 5-GHz mesh interconnect for a teraflops processor," *IEEE Micro*, vol. 27, no. 5, pp. 51–61, Sep. 2007.

[52] A. H. Atabaki et al., "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature*, vol. 556, no. 7701, pp. 349–354, Apr. 2018.

[53] N. Dupuis et al., "30 Gbps optical link utilizing heterogeneously integrated III-V/Si photonics and CMOS circuits," in *Proc. OFC*, Mar. 2014, pp. 1–3.

[54] J. E. Cunningham et al., "Scaling hybrid-integration of silicon photonics in freescale 130 nm to TSMC 40 nm-CMOS VLSI drivers for low power communications," in *Proc. Opt. Interconnects Conf.*, May 2012, pp. 1–7.

[55] X. Zheng et al., "Ultra-low power arrayed CMOS silicon photonic transceivers for an 80 Gbps WDM optical link," in *Proc. Opt. Fiber Commun. Conf. Expo. Nat. Fiber Optic Engineers Conf.*, Mar. 2011, pp. 1–3.

[56] M. Steinman, "Hummingbird low-latency computing engine," in *Proc. IEEE Hot Chips 35 Symp. (HCS)*, Aug. 2023, pp. 1–20.

[57] M. Wade et al., "TeraPHY: A chiplet technology for low-power, high-bandwidth in-package optical I/O," *IEEE Micro*, vol. 40, no. 2, pp. 63–71, Mar. 2020.

[58] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy, "Silicon-photonic network architectures for scalable, power-efficient multi-chip systems," *ACM SIGARCH Comput. Archit. News*, vol. 38, no. 3, pp. 117–128, Jun. 2010.

[59] K. K. Tokgoz et al., "A 120Gb/s 16QAM CMOS millimeter-wave wireless transceiver," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 168–170.

[60] A. Townley et al., "A fully integrated, dual channel, flip chip packaged 113 GHz transceiver in 28 nm CMOS supporting an 80 Gb/s wireless link," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Mar. 2020, pp. 1–4.

[61] S. Abadal, E. Alarcón, A. Cabellos-Aparicio, M. C. Lemme, and M. Nemirovsky, "Graphene-enabled wireless communication for massive multicore architectures," *IEEE Commun. Mag.*, vol. 51, no. 11, pp. 137–143, Nov. 2013.

[62] S.-J. Han, A. V. Garcia, S. Oida, K. A. Jenkins, and W. Haensch, “Graphene radio frequency receiver integrated circuit,” *Nature Commun.*, vol. 5, no. 1, pp. 1–6, Jan. 2014.

[63] S. Abadal, C. Han, and J. M. Jornet, “Wave propagation and channel modeling in chip-scale wireless communications: A survey from millimeter-wave to terahertz and optics,” *IEEE Access*, vol. 8, pp. 278–293, 2020.

[64] M. S. Shamim, N. Mansoor, R. S. Narde, V. Kothandapani, A. Ganguly, and J. Venkataraman, “A wireless interconnection framework for seamless inter and intra-chip communication in multichip systems,” *IEEE Trans. Comput.*, vol. 66, no. 3, pp. 389–402, Mar. 2017.

[65] D. W. Matolak, A. Kodi, S. Kaya, D. Ditomaso, S. Laha, and W. Rayess, “Wireless networks-on-chips: Architecture, wireless channel, and devices,” *IEEE Wireless Commun.*, vol. 19, no. 5, pp. 58–65, Oct. 2012.

[66] G. H. Loh and R. Swaminathan, “The next era for chiplet innovation,” in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Apr. 2023, pp. 1–6.

[67] S. Naffziger et al., “Pioneering chiplet technology and design for the AMD EPYC and Ryzen processor families : Industrial product,” in *Proc. ACM/IEEE 48th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2021, pp. 57–70.

[68] N. Beck, S. White, M. Paraschou, and S. Naffziger, “‘Zeppelin’: An SoC for multichip architectures,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 40–42.

[69] (2024). *NVIDIA Goes MCM and Officially Announces Blackwell At GTC 2024: Details of the Monster Chip*. Accessed: May 20, 2024. [Online]. Available: https://tinyurl.com/nvdiagtc2024

[70] D. Das Sharma, G. Pasdast, Z. Qian, and K. Aygun, “Universal chiplet interconnect express (UCIe): An open industry standard for innovations with chiplets at package level,” *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 12, no. 9, pp. 1423–1431, Sep. 2022.

[71] (2024). *Chiplets: 10 Breakthrough Technologies 2024*. Accessed: May 20, 2024. [Online]. Available: https://tinyurl.com/chiplets2024

[72] A. Usman et al., “Interposer technologies for high-performance applications,” *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 7, no. 6, pp. 819–828, Jun. 2017.

[73] (2021). *Samsung Electronics Announces Availability of Its Next Generation 2.5D Integration Solution ‘I-Cube4’ for High-Performance Applications*. Accessed: May 20, 2024. [Online]. Available: https://tinyurl.com/samsungicube2021

[74] W. Flack et al., “Large area interposer lithography,” in *Proc. IEEE 64th Electron. Compon. Technol. Conf. (ECTC)*, May 2014, pp. 26–32.

[75] R. Mahajan et al., “Embedded multi-die interconnect bridge (EMIB)—A high density, high bandwidth packaging interconnect,” in *Proc. IEEE 66th Electron. Compon. Technol. Conf. (ECTC)*, May 2016, pp. 557–565.

[76] R. Swaminathan, M. J. Schulte, B. Wilkerson, G. H. Loh, A. Smith, and N. James, “AMD InstinctTM MI250X accelerator enabled by elevated fanout bridge advanced packaging architecture,” in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2023, pp. 1–2.

[77] (2024). *Samsung Develops Industry-First 36GB HBM3E 12H DRAM*. Accessed: Jul. 20, 2024. [Online]. Available: https://news.samsung.com/global/samsung-develops-industry-first-36gb-hbm3e-12h-dram

[78] (2024). *Hybrid Bonding Plays Starring Role in 3D Chips*. Accessed: Jul. 20, 2024. [Online]. Available: https://spectrum.ieee.org/hybrid-bonding

[79] (2021). *TSMC-SoIC*. Accessed: May 20, 2024. [Online]. Available: https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/SoIC.htm

[80] (2022). *AMD 3D V-CacheT Technology*. Accessed: May 20, 2024. [Online]. Available: https://www.amd.com/en/products/processors/technologies/3d-v-cache.html

[81] (2022). *BOW IPU Processor*. Accessed: May 20, 2024. [Online]. Available: https://www.graphcore.ai/bow-processors

[82] D. D. Sharma. (2019). *Compute Express Link*. Accessed: May 20, 2024. [Online]. Available: https://docs.wixstatic.com/ugd/0c1418_d9878707bbb7427786b70c3c91d5fbd1.pdf

[83] (2023). *Intel Trust Domain Extensions*. Accessed: May 20, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/developer/tools/trust-domain-extensions/overview.html

[84] (2021). *AMD Infinity Guard*. Accessed: May 20, 2024. [Online]. Available: https://www.amd.com/en/products/processors/server/epyc/infinity-guard.html

[85] J. Kim and N. S. Kim, “Special issue on emerging system interconnects,” *IEEE Micro*, vol. 43, no. 2, pp. 6–8, Mar. 2023.

[86] (2024). *Intel® Xeon® 6780E Processor*. Accessed: May 20, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/240362/intel-xeon-6780e-processor-108m-cache-2-20-ghz/specifications.html

[87] (2023). *Ampereonet 64-Bit Multi-Core Processors*. Accessed: May 20, 2024. [Online]. Available: https://amperecomputing.com/briefs/ampereone-family-product-brief

[88] (2023). *AMD EPYCT 9754S*. Accessed: May 20, 2024. [Online]. Available: https://www.amd.com/en/products/processors/server/epyc/4th-generation-9004-and-8004-series/amd-epyc-9754s.html

[89] A. Olofsson, “Epiphany-V: A 1024 processor 64-bit RISC system-on-chip,” 2016, *arXiv:1610.01832*.

[90] (2021). *ARM Neoverse CMN-700*. Accessed: May 20, 2024. [Online]. Available: https://developer.arm.com/Processors/Neoverse

[91] (2009). *ARTERIS FlexNoC Interconnect IP*. Accessed: May 20, 2024. [Online]. Available: https://www.arteris.com/products/non-coherent-NoC-ip/flexnoc/

[92] J. Xia, C. Cheng, X. Zhou, Y. Hu, and P. Chun, “Kunpeng 920: The first 7-nm chiplet-based 64-core ARM SoC for cloud services,” *IEEE Micro*, vol. 41, no. 5, pp. 67–75, Sep. 2021.

[93] A. Kannan, N. E. Jerger, and G. H. Loh, “Enabling interposer-based disintegration of multi-core processors,” in *Proc. 48th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2015, pp. 546–558.

[94] Y. S. Shao et al., “Simba: Scaling deep-learning inference with multi-chip-module-based architecture,” in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 14–27.

[95] (2019). *AMD Infinity Architecture*. Accessed: May 20, 2024. [Online]. Available: https://www.amd.com/en/technologies/infinity-architecture

[96] (2017). *Intel® Ultra Path Interconnect (Intel® UPI)*. Accessed: May 20, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/silicon-innovations/6-pillars/interconnect.html

[97] R. Medina et al., “System-level exploration of in-package wireless communication for multi-chiplet platforms,” in *Proc. 28th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2023, pp. 1–6.

[98] E. Taheri, S. Pasricha, and M. Nikdast, “ReSiPI: A reconfigurable silicon-photonic 2.5D chiplet network with PCMs for energy-efficient interposer communication,” in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Oct. 2022, pp. 1–9.

[99] Z. Wang et al., “CAMON: Low-cost silicon photonic chiplet for many-core processors,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 9, pp. 1820–1833, Jun. 2020.

[100] (2024). *Intel Optical Compute Interconnect (OCI)*. Accessed: May 20, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/products/details/network-io/silicon-photonics.html

[101] (2017). *AMBA Coherent Hub Interface (CHI)*. Accessed: May 20, 2024. [Online]. Available: https://developer.arm.com/documentation/102407/0100/Introducing-the-AMBA-Coherent-Hub-Interface

[102] (2016). *ARTERIS Ncore Cache Coherent Interconnect IP*. Accessed: May 20, 2024. [Online]. Available: https://www.arteris.com/products/coherent-NoC-ip/ncore/

[103] S. H. Gade, M. Sinha, M. Kumar, and S. Deb, “Scalable hybrid cache coherence using emerging links for chiplet architectures,” in *Proc. 35th Int. Conf. VLSI Design 21st Int. Conf. Embedded Syst. (VLSID)*, Feb. 2022, pp. 92–97.

[104] A. Franques, A. Kokolis, S. Abadal, V. Fernando, S. Misailovic, and J. Torrellas, “WiDir: A wireless-enabled directory cache coherence protocol,” in *Proc. IEEE Int. Symp. High-Performance Comput. Archit. (HPCA)*, Feb. 2021, pp. 304–317.

[105] S. H. Gade and S. Deb, “A novel hybrid cache coherence with global snooping for many-core architectures,” *ACM Trans. Design Autom. Electron. Syst.*, vol. 27, no. 1, pp. 1–31, Jan. 2022.

[106] B. K. Daya, L.-S. Peh, and A. P. Chandrakasan, “Low-power on-chip network providing guaranteed services for snoopy coherent and artificial neural network systems,” in *Proc. 54th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2017, pp. 1–6.

[107] W.-C. Kwon and L.-S. Pehy, “A universal ordered NoC design platform for shared-memory MPSoC,” in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD)*, Nov. 2015, pp. 697–704.

[108] B. K. Daya et al., “SCORPIO: A 36-core research chip demonstrating snoopy coherence on a scalable mesh NoC with in-network ordering,” *ACM SIGARCH Comput. Archit. News*, vol. 42, no. 3, pp. 25–36, 2014.

[109] D. Vantrease, M. H. Lipasti, and N. Binkert, "Atomic coherence: Leveraging nanophotonics to build race-free cache coherence protocols," in *Proc. IEEE 17th Int. Symp. High Perform. Comput. Archit.*, Feb. 2011, pp. 132–143.

[110] G. Kurian et al., "ATAC: A 1000-core cache-coherent processor with on-chip optical network," in *Proc. 19th Int. Conf. Parallel Architectures Compilation Techn. (PACT)*, Sep. 2010, pp. 477–488.

[111] D. Moghimi, "Downfall: Exploiting speculative data gathering," in *Proc. 32nd USENIX Secur. Symp. (USENIX)*, Aug. 2023, pp. 7179–7193.

[112] R. B. Lee, *Security Basics for Computer Architects*. Berlin, Germany: Springer, 2022.

[113] R. Manju, A. Das, J. Jose, and P. Mishra, "SECTAR: Secure NoC using trojan aware routing," in *Proc. 14th IEEE/ACM Int. Symp. Netw. Chip (NOCS)*, Sep. 2020, pp. 1–8.

[114] C. Sudusinghe, S. Charles, and P. Mishra, "Denial-of-service attack detection using machine learning in network-on-chip architectures," in *Proc. 15th IEEE/ACM Int. Symp. Netw.-Chip (NOCS)*, Oct. 2021, pp. 35–40.

[115] M. K. Jyv, A. K. Swain, S. Kumar, S. R. Sahoo, and K. Mahapatra, "Run time mitigation of performance degradation hardware trojan attacks in network on chip," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Jul. 2018, pp. 738–743.

[116] T. Boraten and A. Kodi, "Mitigation of hardware trojan based denial-of-service attack for secure NoCs," *J. Parallel Distrib. Comput.*, vol. 111, pp. 24–38, Jan. 2018.

[117] L. Fiorin, G. Palermo, and C. Silvano, "A security monitoring service for NoCs," in *Proc. 6th IEEE/ACM/IFIP Int. Conf. Hardw./Softw. Codesign Syst. Synth.*, Oct. 2008, pp. 197–202.

[118] P. Guo, W. Hou, L. Guo, Z. Cao, and Z. Ning, "Potential threats and possible countermeasures for photonic network-on-chip," *IEEE Commun. Mag.*, vol. 58, no. 9, pp. 48–53, Sep. 2020.

[119] T. H. Boraten and A. K. Kodi, "Securing NoCs against timing attacks with non-interference based adaptive routing," in *Proc. IEEE/ACM Int. Symp. Netw. Chip (NOCS)*, Apr. 2018, pp. 1–8.

[120] C. Reinbrecht, A. Susin, L. Bossuet, and J. Sepúlveda, "Gossip NoC—Avoiding timing side-channel attacks through traffic management," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Jul. 2016, pp. 601–606.

[121] M. J. Sepúlveda, J.-P. Diguet, M. Strum, and G. Gogniat, "NoC-based protection for SoC time-driven attacks," *IEEE Embedded Syst. Lett.*, vol. 7, no. 1, pp. 7–10, Mar. 2015.

[122] Y. Wang and G. E. Suh, "Efficient timing channel protection for on-chip networks," in *Proc. IEEE/ACM 6th Int. Symp. Netw. Chip*, May 2012, pp. 142–151.

[123] (2023). *Google Cloud TPUv5*. Accessed: May 20, 2024. [Online]. Available: https://cloud.google.com/tpu/

[124] (2022). *NVIDIA Tensor Cores*. Accessed: May 20, 2024. [Online]. Available: https://www.nvidia.com/en-us/data-center/tensor-cores/

[125] N. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in *Proc. 44th Annu. Int. Symp. Comput. Architect.*, 2017, pp. 1–12.

[126] A. Parashar et al., "SCNN: An accelerator for compressed-sparse convolutional neural networks," *ACM SIGARCH Comput. Archit. news*, vol. 45, no. 2, pp. 27–40, 2017.

[127] S. Han et al., "EIE: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 243–254, 2016.

[128] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[129] Y. Chen et al., "DaDianNao: A machine-learning supercomputer," in *Proc. 47th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2014, pp. 609–622.

[130] (2024). *Cerebras Wafer Scale Engine-3*. Accessed: May 20, 2024. [Online]. Available: https://www.cerebras.net/product-chip/

[131] Y.-H. Chen, T.-J. Yang, J. Emer, and V. Sze, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 2, pp. 292–308, Jun. 2019.

[132] W. Lu, G. Yan, J. Li, S. Gong, Y. Han, and X. Li, "FlexFlow: A flexible dataflow accelerator architecture for convolutional neural networks," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2017, pp. 553–564.

[133] A. Das, E. Russo, and M. Palesi, "Multi-objective hardware-mapping co-optimisation for multi-DNN workloads on chiplet-based accelerators," *IEEE Trans. Comput.*, vol. 73, no. 8, pp. 1883–1898, Aug. 2024.

[134] E. Qin et al., "SIGMA: A sparse and irregular GEMM accelerator with flexible interconnects for DNN training," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2020, pp. 58–70.

[135] H. Kwon, A. Samajdar, and T. Krishna, "MAERI: Enabling flexible dataflow mapping over DNN accelerators via reconfigurable interconnects," *ACM SIGPLAN Notices*, vol. 53, no. 2, pp. 461–475, Nov. 2018.

[136] H. Kwon, A. Samajdar, and T. Krishna, "Rethinking NoCs for spatial neural network accelerators," in *Proc. 11th IEEE/ACM Int. Symp. Netw.-Chip (NOCS)*, Oct. 2017, pp. 1–8.

[137] S. Carrillo et al., "Advancing interconnect density for spiking neural network hardware implementations using traffic-aware adaptive network-on-chip routers," *Neural Netw.*, vol. 33, pp. 42–57, Sep. 2012.

[138] A. Rahman, J. Lee, and K. Choi, "Efficient FPGA acceleration of convolutional neural networks using logical-3D compute array," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2016, pp. 1393–1398.

[139] (2020). *Xilinx DPU*. Accessed: May 20, 2024. [Online]. Available: https://www.xilinx.com/products/intellectual-property/dpu.html

[140] (2017). *AXI Interconnect V2.1*. Accessed: May 20, 2024. [Online]. Available: https://docs.amd.com/r/en-U.S./pg059-axi-interconnect/AXI-Interconnect-v2.1-LogiCORE-IP-Product-Guide

[141] J. Wang and S. Gu, "FPGA implementation of object detection accelerator based on vitis-AI," in *Proc. 11th Int. Conf. Inf. Sci. Technol. (ICIST)*, May 2021, pp. 571–577.

[142] A. Aimar et al., "NullHop: A flexible convolutional neural network accelerator based on sparse representations of feature maps," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 644–656, Mar. 2019.

[143] H. Sharma et al., "From high-level deep neural models to FPGAs," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2016, pp. 1–12.

[144] J. Tong, A. Itagi, P. Chatarasi, and T. Krishna, "FEATHER: A reconfigurable accelerator with data reordering support for low-cost on-chip dataflow switching," 2024, *arXiv:2405.13170*.

[145] X. Wei et al., "Automated systolic array architecture synthesis for high throughput CNN inference on FPGAs," in *Proc. 54th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2017, pp. 1–6.

[146] D. J. M. Moss et al., "High performance binary neural networks on the Xeon+FPGA platform," in *Proc. 27th Int. Conf. Field Program. Log. Appl. (FPL)*, Sep. 2017, pp. 1–4.

[147] Y. Ma, Y. Cao, S. Vrudhula, and J.-S. Seo, "Optimizing loop operation and dataflow in FPGA acceleration of deep convolutional neural networks," in *Proc. 2017 ACM/SIGDA Int. Symp. Field-Programmable Gate Arrays*, 2017, pp. 45–54.

[148] J. Fowers et al., "A configurable cloud-scale DNN processor for real-time AI," in *Proc. Int. Symp. Comput. Architecture (ISCA)*, Jun. 2018, pp. 1–14.

[149] (2022). *Peripheral Component Interconnect Express (PCIe) V6*. Accessed: May 20, 2024. [Online]. Available: https://pcisig.com/

[150] C. Zhang, D. Wu, J. Sun, G. Sun, G. Luo, and J. Cong, "Energy-efficient CNN implementation on a deeply pipelined FPGA cluster," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2016, pp. 326–331.

[151] D. Choudhury, R. Barik, A. S. Rajam, A. Kalyanaraman, and P. P. Pande, "Software/hardware co-design of 3D NoC-based GPU architectures for accelerated graph computations," *ACM Trans. Design Autom. Electron. Syst.*, vol. 27, no. 6, pp. 1–22, Nov. 2022.

[152] (2014). *NVLink and NVLink Switch*. Accessed: May 20, 2024. [Online]. Available: https://www.nvidia.com/en-us/data-center/nvlink/

[153] (2024). *NVIDIA GH200 Grace Hopper Superchip*. Accessed: May 20, 2024. [Online]. Available: https://www.nvidia.com/en-us/data-center/grace-hopper-superchip/

[154] A. Li et al., "Evaluating modern GPU interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 1, pp. 94–110, Jan. 2020.

[155] D. Narayanan et al., "PipeDream: Generalized pipeline parallelism for DNN training," in *Proc. 27th ACM Symp. Operating Syst. Princ.*, Oct. 2019, pp. 1–15.

[156] D. Foley and J. Danskin, "Ultra-performance Pascal GPU and NVLink interconnect," *IEEE Micro*, vol. 37, no. 2, pp. 7–17, Mar. 2017.

[157] Y. You, A. Buluç, and J. Demmel, "Scaling deep learning on GPU and knights landing clusters," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Nov. 2017, pp. 1–12.

[158] S. A. Mojumder et al., "Profiling DNN workloads on a volta-based DGX-1 system," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Sep. 2018, pp. 122–133.

[159] J. Guo et al., "AccUDNN: A GPU memory efficient accelerator for training ultra-deep neural networks," in *Proc. IEEE 37th Int. Conf. Comput. Design (ICCD)*, Nov. 2019, pp. 65–72.

[160] M. Rhu, N. Gimelshein, J. Clemons, A. Zulfiqar, and S. W. Keckler, "vDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2016, pp. 1–13.

[161] Q. Xu, H. Jeon, and M. Annavaram, "Graph processing on GPUs: Where are the bottlenecks?" in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2014, pp. 140–149.

[162] R. Guirado, A. Rahimi, G. Karunaratne, E. Alarcón, A. Sebastian, and S. Abadal, "WHYPE: A scale-out architecture with wireless over-the-air majority for scalable in-memory hyperdimensional computing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 13, no. 1, pp. 137–149, Mar. 2023.

[163] (2021). *Mythic AMPT*. Accessed: May 20, 2024. [Online]. Available: https://mythic.ai/product/

[164] S. Bavikadi, P. R. Sutradhar, K. N. Khasawneh, A. Ganguly, and S. M. P. Dinakarrao, "A review of in-memory computing architectures for machine learning applications," in *Proc. Great Lakes Symp. VLSI*, Sep. 2020, pp. 89–94.

[165] N. Hajinazar et al., "SIMDRAM: A framework for bit-serial SIMD processing using DRAM," in *Proc. 26th ACM Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2021, pp. 329–345.

[166] C. Eckert et al., "Neural cache: Bit-serial in-cache acceleration of deep neural networks," in *Proc. ACM/IEEE 45Th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2018, pp. 383–396.

[167] S. Li et al., "SCOPE: A stochastic computing engine for DRAM-based in-situ accelerator," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2018, pp. 696–709.

[168] S. Li, D. Niu, K. T. Malladi, H. Zheng, B. Brennan, and Y. Xie, "DRISA: A DRAM-based reconfigurable in-situ accelerator," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2017, pp. 288–301.

[169] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 529–544, Jul. 2020.

[170] Z. Yan, D.-C. Juan, X. S. Hu, and Y. Shi, "Uncertainty modeling of emerging device based computing-in-memory neural accelerators with application to neural architecture search," in *Proc. 26th Asia–South Pacific Design Autom. Conf. (ASP-DAC)*, 2021, pp. 859–864.

[171] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *IEDM Tech. Dig.*, Dec. 2015, p. 17.

[172] A. Ankit et al., "PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Proc. 24th Int. Conf. Architectural Support Program. Lang. Operating Syst.*, Apr. 2019, pp. 715–731.

[173] A. Shafiee et al., "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Comput. Archit. News*, vol. 44, no. 3, pp. 14–26, 2016.

[174] X. Liu et al., "RENO: A high-efficient reconfigurable neuromorphic computing accelerator design," in *Proc. 52nd ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2015, pp. 1–6.

[175] R. Guirado, A. Rahimi, G. Karunaratne, E. Alarcón, A. Sebastian, and S. Abadal, "Wireless on-chip communications for scalable in-memory hyperdimensional computing," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[176] N. Bruschi et al., "Scale up your in-memory accelerator: Leveraging wireless-on-chip communication for AIMC-based CNN inference," in *Proc. IEEE 4th Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Jun. 2022, pp. 170–173.

[177] (2024). *IonQ Achieves Critical First Step Towards Developing Future Quantum Networks*. Accessed: May 20, 2024. [Online]. Available: https://tinyurl.com/ionq2024

[178] (2024). *Pasqal and Welinq Partner to Develop Tailored Quantum Interconnects for Neutral-Atom Quantum Computing*. Accessed: May 20, 2024. [Online]. Available: https://www.pasqal.com/news/pasqal-and-welinq-partnership/

[179] (2020). *IBM Quantum Roadmap*. Accessed: May 20, 2024. [Online]. Available: https://www.ibm.com/roadmaps/quantum/

[180] J. Gambetta. (2020). *IBM's Roadmap for Scaling Quantum Technology*. Accessed: May 20, 2024. [Online]. Available: https://www.ibm.com/quantum/blog/ibm-quantum-roadmap

[181] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, Aug. 2018.

[182] (2023). *Quantum Startup Atom Computing First to Exceed 1,000 Qubits*. Accessed: May 20, 2024. [Online]. Available: https://atom-computing.com/quantum-startup-atom-computing-first-to-exceed-1000-qubits/

[183] (2023). *Charting the Course to 100,000 Qubits*. Accessed: May 20, 2024. [Online]. Available: https://www.ibm.com/quantum/blog/100k-qubit-supercomputer

[184] R. Srinivas et al., "High-fidelity laser-free universal control of trapped ion qubits," *Nature*, vol. 597, no. 7875, pp. 209–213, Sep. 2021.

[185] Y.-Y. Liu et al., "Radio-frequency reflectometry in silicon-based quantum dots," *Phys. Rev. Appl.*, vol. 16, no. 1, Jul. 2021, Art. no. 014057.

[186] F. J. Schupp et al., "Sensitive radiofrequency readout of quantum dots using an ultra-low-noise SQUID amplifier," *J. Appl. Phys.*, vol. 127, no. 24, Jun. 2020, Art. no. 244503.

[187] P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, "A quantum engineer's guide to superconducting qubits," *Appl. Phys. Rev.*, vol. 6, no. 2, Jun. 2019, Art. no. 021318.

[188] J. J. Pla et al., "High-fidelity readout and control of a nuclear spin qubit in silicon," *Nature*, vol. 496, no. 7445, pp. 334–338, Apr. 2013.

[189] M. Malinowski, D. T. C. Allcock, and C. J. Ballance, "How to wire a 1000 -Qubit trapped-ion quantum computer," *PRX Quantum*, vol. 4, no. 4, Oct. 2023, Art. no. 040313.

[190] X. Xue et al., "CMOS-based cryogenic control of silicon quantum circuits," *Nature*, vol. 593, no. 7858, pp. 205–210, May 2021.

[191] A. Potoč nik et al., "Millikelvin temperature cryo-CMOS multiplexer for scalable quantum device characterisation," *Quantum Sci. Technol.*, vol. 7, no. 1, Jan. 2022, Art. no. 015004.

[192] R. Li et al., "A crossbar network for silicon quantum dot qubits," *Sci. Adv.*, vol. 4, no. 7, Jul. 2018, Art. no. eaar3960.

[193] F. Lecocq, F. Quinlan, K. Cicak, J. Aumentado, S. A. Diddams, and J. D. Teufel, "Control and readout of a superconducting qubit using a photonic link," *Nature*, vol. 591, no. 7851, pp. 575–579, Mar. 2021.

[194] J. Wang et al., "THz cryo-CMOS backscatter transceiver: A contactless 4 Kelvin-300 Kelvin data interface," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2023, pp. 504–506.

[195] (2021). *Atom Computing: Our Technology Advantage*. Accessed: May 20, 2024. [Online]. Available: https://atom-computing.com/quantum-computing-technology/

[196] H. Jnane, B. Undseth, Z. Cai, S. C. Benjamin, and B. Koczor, "Multicore quantum computing," *Phys. Rev. Appl.*, vol. 18, no. 4, Oct. 2022, Art. no. 044064.

[197] K. N. Smith, G. S. Ravi, J. M. Baker, and F. T. Chong, "Scaling superconducting quantum computers with chiplet architectures," in *Proc. 55th IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2022, pp. 1092–1109.

[198] S. Rodrigo, S. Abadal, E. Alarcón, M. Bandic, H. V. Someren, and C. G. Almudéver, "On double full-stack communication-enabled architectures for multicore quantum computers," *IEEE Micro*, vol. 41, no. 5, pp. 48–56, Sep. 2021.

[199] A. Gold et al., "Entanglement across separate silicon dies in a modular superconducting qubit device," *npj Quantum Inf.*, vol. 7, no. 1, p. 142, Sep. 2021.

[200] P. Magnard et al., "Microwave quantum link between superconducting circuits housed in spatially separated cryogenic systems," *Phys. Rev. Lett.*, vol. 125, no. 26, Dec. 2020, Art. no. 260502.

[201] Y. Zhong et al., "Deterministic multi-qubit entanglement in a quantum network," *Nature*, vol. 590, no. 7847, pp. 571–575, Feb. 2021.

[202] B. Marinelli et al., "Dynamically reconfigurable photon exchange in a superconducting quantum processor," 2023, *arXiv:2303.03507*.

[203] Q.-Y. Zhang, P. Xu, and S.-N. Zhu, "Quantum photonic network on chip," *Chin. Phys. B*, vol. 27, no. 5, May 2018, Art. no. 054207.

[204] C. Monroe et al., "Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects," *Phys. Rev. A, Gen. Phys.*, vol. 89, no. 2, Feb. 2014, Art. no. 022317.

[205] S. Ritter et al., "An elementary quantum network of single atoms in optical cavities," *Nature*, vol. 484, no. 7393, pp. 195–200, Apr. 2012.

[206] S. Yu et al., "Reconstruction of a photonic qubit state with reinforcement learning," *Adv. Quantum Technol.*, vol. 2, nos. 7–8, Aug. 2019, Art. no. 1800074.

[207] H. P. Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, "Optimizing quantum error correction codes with reinforcement learning," *Quantum*, vol. 3, p. 215, Dec. 2019.

[208] M. Schuld, I. Sinayskiy, and F. Petruccione, "An introduction to quantum machine learning," *Contemp. Phys.*, vol. 56, no. 2, pp. 172–185, 2015.

[209] P. Wittek, *Quantum Machine Learning: What Quantum Computing Means to Data Mining*. Cambridge, MA, USA: Academic Press, 2014.

[210] J. S. Otterbach et al., "Unsupervised machine learning on a hybrid quantum computer," 2017, *arXiv:1712.05771*.

[211] E. Aïmeur, G. Brassard, and S. Gambs, "Quantum clustering algorithms," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 1–8.

[212] A. Perdomo-Ortiz, M. Benedetti, J. Realpe-Gómez, and R. Biswas, "Opportunities and challenges for quantum-assisted machine learning in near-term quantum computers," *Quantum Sci. Technol.*, vol. 3, no. 3, Jun. 2018, Art. no. 030502.

[213] Z. Abohashima, M. Elhosen, E. H. Houssein, and W. M. Mohamed, "Classification with quantum machine learning: A survey," 2020, *arXiv:2006.12270*.

[214] (2020). *TensorFlow Quantum is a Library for Hybrid Quantum-classical Machine Learning*. Accessed: May 20, 2024. [Online]. Available: https://www.tensorflow.org/quantum

[215] G. H. Loh, N. E. Jerger, A. Kannan, and Y. Eckert, "Interconnect-memory challenges for multi-chip, silicon interposer systems," in *Proc. Int. Symp. Memory Syst.*, Oct. 2015, pp. 3–10.

[216] A. Ganguly et al., "Interconnects for DNA, quantum, in-memory, and optical computing: Insights from a panel discussion," *IEEE Micro*, vol. 42, no. 3, pp. 40–49, May 2022.

[217] D. Abts and J. Kim, "Enabling artificial intelligence supercomputers with domain-specific networks," *IEEE Micro*, vol. 44, no. 2, pp. 41–49, Mar. 2024.

Abhijit Das (Member, IEEE) received the Ph.D. degree in computer science and engineering from Indian Institute of Technology (IIT) Guwahati, India, in 2021. He is currently the Director of Research and Group Leader with the NaNoNetworking Center in Catalunya (N3Cat), Universitat Politécnica de Catalunya, Spain. His current research interests include chip and package-scale networks, memory systems, DNN accelerators, and quantum systems. He won the Best Thesis Award.

Maurizio Palesi (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in communication and computer engineering from the University of Catania, Italy, in 1999 and 2003, respectively. He is currently an Associate Professor with the University of Catania and a Visiting Associate Professor with Indian Institute of Technology Guwahati, India. His current research interests include domain-specific architectures (DSAs). He was a guest editor for 20 special issues in top-tier journals. He was the general chair and the TPC co-chair for several international conferences and workshops. He is an associate editor of 12 international journals.

John Kim (Senior Member, IEEE) received the B.S. and M.Eng. degrees in electrical engineering from Cornell University in 1997 and 1998, respectively, and the Ph.D. degree in electrical engineering from Stanford University in 2008. He is currently a Full Professor with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology. His research interests include computer architecture, interconnection networks, security, and mobile systems.

Partha Pratim Pande (Fellow, IEEE) is currently a Professor and the Holder of the Boeing Centennial Chair in computer engineering with the School of Electrical Engineering and Computer Science, Washington State University (WSU), Pullman, USA. He is also the Interim Dean of the Voiland College of Engineering and Architecture, WSU. His current research interests include novel interconnect architectures for manycore chips, on-chip wireless communication networks, heterogeneous architectures, and ML for EDA. He serves as the Editor-in-Chief of *IEEE Design and Test*. He is on the editorial boards of IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, *ACM Journal of Emerging Technologies in Computing Systems*, and IEEE EMBEDDED SYSTEMS LETTERS.