SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Review

A Comprehensive Review of Memory Architectures and Network Interfaces and Advancements in High-Bandwidth and Low-Latency Systems

Shaik Rajeena and Shaik Karimullah *

Department of Electronics and Communications, Annamacharya University, Rajampet 516115, India

* Correspondence: munnu483@gmail.com

Received: 22 October 2025; **Revised:** 5 December 2025; **Accepted:** 19 December 2025; **Published:** 15 April 2026

Abstract: This review discusses the significant advances in the architecture of memory systems, the enhancement of which has been among the most significant concerns in the process of creating systems with high bandwidth and low latency requirements. It has analyzed various fields that include virtually pipelined systems, millimeter-wave interface, and enhanced memory controllers, which would meet the requirements of high-performance computing (HPC). Among all the developments put in place, die-stacked DRAM (Dynamic Random Access Memory) systems offer a high bandwidth boost with the networks-on-chip scalable and meant to address the issues of congestion and incoherence in interconnecting cores. The literature survey also identifies some of the major developments in the mechanism of passing messages and memory management optimizations in distributed systems that have been found to be critical towards useful data transfer and processing in massive parallel computer systems. The advanced multi-port memory controllers are also observed to improve the bandwidth and efficiency in utilizing the resources. The review, however, points out some of the challenges that are still there, including the limitation of scalability of the centralized memory system, the latency issues of high-radix interconnects and the integration problems of heterogeneous computing systems. The evaluation highlights the need for new ways to tackle the limitations of current memory management techniques. Research will center on the possibilities of using machine learning to anticipate workloads and applying adaptive hybrid memory allocators to allocate across memory types dynamically. The goal of these techniques is to increase performance, bandwidth, latency, and energy efficiency in high-performance computing systems.

Keywords: Networks-on-Chip; Optimized Memory Management; Memory Controllers; Die-Stacked DRAM; Distributed Systems; Latency; Heterogeneous Computing Environments

1. Introduction

The increasing needs of data-intensive and compute-intensive applications are driving high-performance computing (HPC) systems, which demand memory architectures that are high-bandwidth, low-latency, and energy-efficient. Conventional memory hierarchies are effective for general-purpose computing but cannot meet the performance needs of modern workload classes such as machine learning, scientific simulations, and real-time analytics. Therefore, innovations across the entire memory subsystem are necessary. Hardware improvements include more advanced memory controllers and optimized interconnects that allow for faster access, improved contention, and throughput of data. Emerging memory technologies like die-stacked DRAM, non-volatile memory (NVM), and resistive RAM (ReRAM) represent potential solutions to bandwidth and latency performance issues in traditional

DRAM, plus they can sit even closer to the CPU or accelerator, improving the times it takes to access memory and decreasing idle cycles. Second, architectural methods for caching, prefetching, and memory scheduling will be significant in minimizing the effect of latency and contention on heterogeneous and multi-core computing platforms. Finally, hardware development has an opportunity to take advantage and pair with software optimizations, like NUMA-aware memory allocation and parallel memory access patterns, which could yield greater overall performance. Recent research into memory architecture combined highlights a strategic change for designing memory hierarchies designed for the high-performance computing (HPC) market today and will scale to the needs for tomorrow's applications. An emphasis on energy efficiency, low latency, and high throughput is situated in an integrated memory system maximizing workloads to be effective and efficient in an increasingly opaque and complex data scape. Research around virtual pipelining, multi-stream command scheduling, high bandwidth interconnects and millimeter wave communication channels has produced frameworks that appropriately manage energy consumption, design complexity, operational costs, and speeds of memory accesses. Together with hardware advancement and adaptation of memory subsystems, again shows an opportunity to provide efficient, sustained data delivery at high speeds, thereby effectively interacting with compute cores capable of high rates of data processing for applications in artificial intelligence, deep learning, and large-matrix analysis.

1.1. Background and Motivation

The methodological foundation for these advancements and innovations has been thoroughly evaluated to elaborate on the reasoning for each design choice. For instance, probabilistic state machines in pipelined memory controllers attempt to provide a balance of throughput and latency, delay buffers and memory banking exploit parallelism but at the expense of increasing the design complexity, multi-stream memory controllers and high-speed interconnects can improve the utilization of bandwidth, which comes with factoring the designs that are scalable as well as connections to the existing systems. In each of the design choices, designers are balancing performance and design constraints, with some of the most important design metrics comprising throughput, latency and energy efficiency.

1.2. Research Challenges in Conventional Memories

In summarizing the contributions, at a high level, improvements in memory systems don't just improve one subsystem, they also improve the performance of the overall HPC system. Memory system improvement creates more rapid computation, more efficient use of resources, and lower energy footprints, all of which contribute to development and improvement at the system level. Lastly, the innovations mentioned in this chapter also communicate future research opportunities including scalable architecture improvements, new memory technology and adaptive mechanisms of control. As computational workloads continue to be more demanding and complex, improved memory hierarchies will be paramount to unlocking new generations of computing systems that satisfy future HPC systems with respect to scientific, industrial and artificial intelligence domains.

As represented in **Figure 1**, the accelerator is designed to accommodate the entire data flow of a neural network layer, ranging from input, through computation, to output. The input is first fed into the accelerator. The accelerator then distributes the input to multiple processing elements (PEs) to perform parallelized operations (e.g., matrix multiplications and convolutions). By doing this, each PE only performs a part of the overall computations, which speeds the processing of large operations. After computation, the activation functions (e.g., ReLU, sigmoid, softmax) are applied directly in the accelerator to have low latency for the application and reduce the transfer of intermediate data. Data generated at the accelerator is either written to the memory or processed for the next layer in the flow of data memory.

This architecture is especially well-suited to classification, regression, and pattern recognition tasks, which are especially true in deep learning applications that require the simultaneous processing of large amounts of data and weights. By providing a dedicated memory subsystem to store weights and intermediate results, the accelerator limits expensive memory accesses to the external DRAM, which is often the performance bottleneck for general-purpose systems. The architecture provides high throughput and low-latency execution because wire latency is an important factor when processing workloads for real-time AI applications such as image recognition, natural language processing, and autonomous systems. Together, this neural network hardware accelerator provides a balanced architecture that leverages compute, memory management, and data flow into a hardware accelerator

application that is well-suited for AI workloads leveraging compute and energy efficiency.

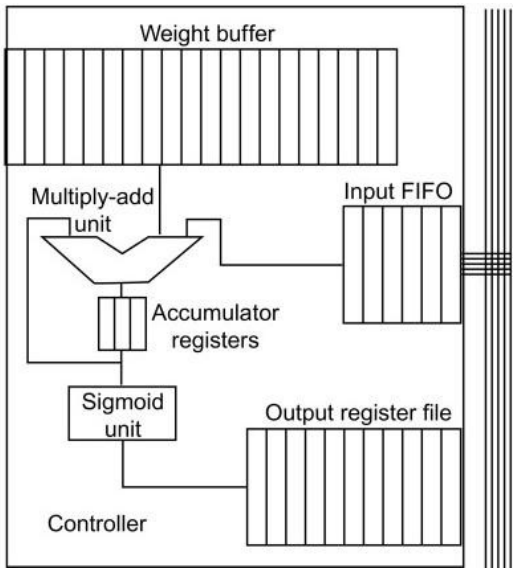

Figure 1. The architecture of a neural network hardware accelerator is designed to optimize the execution of machine learning computations by efficiently mapping neural network operations onto specialized hardware components.

2. Literature Review

Memory architectures are the building blocks of any high-performance computing system in which the performance of the system depends critically on system bandwidth, memory latency, and energy cost. To this end, Agrawal and Sherwood (2009) proposed a virtually pipelined memory architecture, which combines a probabilistic state machine with a delay storage buffer, to manage banked memory more effectively. The probabilistic state machine enables dynamic scheduling of memory requests, thereby reducing conflicts between requests and increasing utilization of available bandwidth [1]. The delay storage buffer allows memory operations to be held temporarily to enable pipelining of memory requests for a more uniform memory latency. These two mechanisms differ from classical packet buffer schemes because they provide higher bandwidth use and predictable access time, yielding a superior performance. While memory banking theoretically allows for the parallel execution of multiple memory operations, it comes along with complexities in the design of memory controllers and also limits scalability, which has impaired the use of such architectures at scale. The efficacy of the approach was established through simulation studies and mathematical modeling that demonstrated the viability of virtual pipelined memory usage in the development of next-generation memory systems.

2.1. Emerging Non-Volatile Technologies

To build upon memory performance, Wang et al. [2] demonstrated that significant end-to-end packet latency reductions were possible by redesigning network processors and operating on-chip memory to process packets. They stated that simply improving memory systems in a networked environment will not be sufficient for improving overall system performance unless the microarchitecture is fundamentally rethought at the networking processor level. Along a parallel thread of research, they studied mmWave interfaces as an option to improve memory speed to compute nodes. This research on mmWave memory speed had an opportunity to provide an extreme transfer of data speed for transferring memory, and posed challenges, with nonlinear phase distortions leading to inefficient propagation of data across the channel. Nonetheless, there was evidence that mmWave channels could accept advanced modulation schemes and propagate signals well enough to reach viable conclusions about mmWaves's value for reasonable future interconnects at high memory speeds. Expanding other components of memory use, Shao et al. [3] described a design for a memory controller that used multiple streams of commands to better utilize bandwidth with the intention of improving the efficiency of context switching between page accesses. Such a de-

sign allowed memory accesses to create overlap, which resulted in higher performance and fewer idle cycles during DRAM accesses, as it relates to the required timing for commands. The commands interleave (either in the same bank or adjacent banks), yielding very high effective bandwidth and lower latency, which is very desirable for data-centric workloads. However, scalability in this approach is not adequately explored, which could be a challenge in adopting to increase the scale required of next-generation computers [3]. Overall, this work illustrates a need to study and explore new means of controlling memory that include ideas such as probabilistic scheduling, pipelining, higher speed core interconnects, multi- stream command execution in parallel accessing memory commands elected for execution in the form of a parallel command.

2.2. Hybrid Memory Architectures

Although the architectural proposals enhance both the effective and latency measures to varying degrees, the following issues remain challenges for consideration: controller complexity, scalability, and finite physical storage channels. The problem for future work is to develop memory systems that provide performance measures with manageable energy utilization and also prove scalable for potential deployment in next-generation computing platforms, such as exascale platforms, AI accelerators, or data-intensive applications. Progress in these areas will be a significant aspect of the continuing evolution of modern memory architectures. Lin et al. [4] provide traffic classification, flit-based switching, and control mechanisms for pre-assigned paths with latency and bandwidth guarantees within NoCs. Although reasonably implemented, their work demonstrates shortcomings, including potential BE traffic starvation categories of isolation in traffic class queues. The work provides reliable solutions for many service requirements and contributes to a rich NoC design space [4]. Cheriton and Kutter [5] enhance the performance of communication in worst-case network messaging using memory hardware + software memory-based messaging systems. In the end, this work demonstrated the inefficiency of legacy operating system support when using architectural optimizations for large-scale systems that improved performance by a factor of three to five [5]. In 2023, Milton and Zarkesh-Ha used simulations in the SST framework to investigate network topologies under evaluation and demonstrated that the HyperX topology performed the best with high-scale bandwidth (>4 GBs) achieving the best topology maximal throughput and then even hard-real-time approaches were hampered due to the law of diminishing returns [6]. The research also carried on to officially define what configurations are best for high-performance computing systems too. Recently, there have been strides made in optimizing intra-data transfer in terms of efficiency, memory management/optimization and lastly broadcasting performance and mechanisms across High-Performance computing as well as wireless communication systems. One such contribution is Pandey et al. [7], who focused on optimizing wireless network throughput and performing accurate measurements. In their study, they introduced a throughput-optimal broadcast approach based on Mayfly optimization, a bio-inspired approach that can deal with the geometrical configurations of nodes in a network topology. The goal was to optimize throughput, while providing an accurate wireless link performance measurement, as it is critical for reliable communication in dynamic wireless networks. In their study even with the throughput optimized and improved accuracy in performance, there were still some implications, they found some limitations, specifically higher Bit Error Rates and increased retransmission, affecting overall system throughput performance. These challenges emphasized the demand for added approaches to balance reliability and throughput within wireless broadcasting systems, which led to more research in later publications [7]. While wireless networks continued to innovate, multi-core memory systems became a popular topic to study in distributed computing. Barthels et al. studied Remote Direct Memory Access (RDMA) as an option for moving data amongst distributed nodes. RDMA allows for memory-to-memory transfers to happen directly without using the CPU—improving latency and freeing up CPU resources for other processing. They showed that RDMA does improve the throughput and performance of distributed systems whilst recognizing some issues (registering ndr DA and the challenge of implementing in real-world systems [8]. To address some of the issues mentioned, Barthels et al. proposed new communication abstractions to facilitate better data transfers and to help improve processing resources and future development. Their paper highlights the interaction between high-speed memory access, low-latency communication within a system-wide optimized architecture in modern computing [8]. In 2017 further development of memory control architecture was presented, which investigated coordinated dual-clock, dual-port FIFOs combined with window-based arbitration techniques to optimize bandwidth utilization and memory flexibility. This design strategy enabled concurrent data access through multiple memory ports while coordinating interactions among clocks that were in distinct domains. With the adoption

of these devices, the design has realized an impressive operating frequency of 150 MHz and an average bandwidth utilization of 93.2%. These numbers indicate a very efficient use of memory. Even with these achievements, it was stated in the authored study that the design did require application-level changes to take full advantage of their enhanced memory control method, although, to be fair, many such hardware improvements do require more app-or system-level changes to software [9]. I think all of these studies really point to the larger movement of putting new memory management techniques together with the new compounding advantages of advanced optimization algorithms and hardware capabilities to improve both the overall energy utilization, as well as data efficiencies. From Mayfly-inspired broadcasting of data within wireless networks, to using RDMA to transfer data within a distributed memory system, and to implementing high-frequency, dual port FIFO memory controllers, these iterations of performance improvement provide a multi-faceted approach to improve and address performance bottlenecks. These concerns, however, demonstrate another common theme, specifically, trade-offs between performance improvements, reliability and system complexity. This affirms the need for and the shift towards Co-design efforts of the hardware, software and application-level requirements. Finally, in 2018, Benoit et al. [10] carried out a study that touched upon the focal issue of execution timing. In that study, they showed that optimizing workflows using tuned memory management makes a contribution in addressing that issue as well. In 2016, Fujiki et al. employed random topologies for optimizing latency-sensitive memory communication, thus reducing memory access cycles and improving energy efficiency. However, performance suffered from an overage of limited links at each node, necessitating further work [11]. Schulz et al. provided a mathematical model on latency distribution analysis for mobile networks, indicating potential routing strategy optimization possibilities for critical communications; however, the cost of testbed simulation was too complex and resource-heavy [12]. Cheema et al. supplied methods for worst-case analysis for real-time processing problems in the context of memory subsystems and incorporated an automated design flow. The work did result in delivering a guaranteed latency/bandwidth level, but scalability with respect to centralized policies proved lacking [13]. Setter et al. supplied an integrated memory on silicon that provided low latency and high throughput. The work was promising and attractive, but broad die-stacked scaling and cost were underestimated [14]. Santella and Vatalaro (2020) provided a discussion on an Edge Cloud Computing future with improvements to latency reduction while the proposed KPIs were within performance expectations for future networks. Though low-level protocol data rates increased, based on the emerging demands, it still proved insufficient for forthcoming services [15]. Li et al. 2016 examined topologies with a high-radix interconnect and provided an overview of various topological designs with regard to low latency and scalability. Despite the incremental efficiency in data movement, latency in current supercomputers became major issue to be addressed [16]. Wertz et al. [17] in 2016 proposed a high-data-rate modulator that could achieve a user input of 2 Gbps. Limitations at this time included sub-optimal amplifier non-linearity resulting in poor performance with respect to higher-order modulation schemes, showing that there is potential to improve available data capacities. Hagendorf and Pawletka discussed combining simulation, optimization, and model management to allow rapid dynamic reconfiguration of the model topology; while this expanded way of managing the methodology allowed some more optimized methodology, the methodology still focused on pre-determined methods of customizing specific applications [18]. Hylton et al. examined link management strategies (proactive and reactive) for space networks; while the proactive strategy increased data rates, it reflected how difficult it is to effectively integrate new protocols into legacy networks [19]. Agrawal and Sherwood [20] 2006 presented a virtually pipelined network memory architecture that would improve memory throughput for high- performance network systems. Traditional memory pipelines had a limitation in servicing several memory requests simultaneously. This restriction is particularly problematic for high packet rate routers and switches, in particular, when processing packets in batches. Instead of deepening the memory pipeline to overlap voice memory accesses, the virtually pipelined design, like instruction-level parallelism (ILP), facilitates memory accesses by pipelining memory accesses in a virtual sense. By pipelining per-flow queues, the agent-based approach achieves near maximum utilization of memory while virtually reducing the latency of each packet access. This method provided strong increases in overall network performance, particularly for systems that require high bandwidth, but it offered substantially higher packet processing times. Furthermore, the study establishes that the ordering and stall performance were improved substantially overall in high-speed routers, yielding an effective technique for managing scalable packet processing over high speed lines. The study can lay the foundation for work to exploit low latency memory access times. Nikologiannis and Katevenis [21] considered the problem of per-flow queueing at OC-192 line rates for a high-speed router by using out-of-order

execution techniques to manipulate DRAM in a way which enabled multiple flows to be processed successfully; importantly, without dropping packets or increasing latency. By dynamically reordering how memory was accessed, their technique enhanced memory bandwidth utilization while providing fairness in processing packets from many flows. The authors also showed the approach could support thousands of flows concurrently without excessive resources from the hardware, increasing the scalability of a network's interior infrastructure. The assignment to memory also minimized the negatives of flow contention, enhancing network performance overall. This work was a valuable contribution to memory design for routers so that per-flow queues could be managed efficiently in very high-speed networking contexts, as well as insight for future advances in DRAM-based packet buffering. Iyer, Kompella, and McKeown [22] considered packet buffer architectures for router linecards, which are memory constructs that manage bursts of traffic and have become important for fast networking. Their work had a primary focus on facilitating trade-offs between buffer size, speed of access, and cost without high loss or low throughput. They suggested designs that efficiently managed packets so they were retrieved from DRAM and stored with little or no contention. Their work also considered flow isolation and fairness such that the buffer could not be solely allocated to one traffic stream. The research underscored the necessity for architectural trade-offs in buffer design as well as mentioned factors relevant to routers with high-performance capabilities: reaching the memory with optimal bandwidth, latency, and scalability. This technical report became a source for realistic designs for high-speed packet buffering systems and impacted later work on memory architectures and routing optimization. Garcia et al. [23] explored high-performance memory systems for next-generation packet buffers in routers and switches. Recognizing that often the bottleneck in network throughput is the accessed memory, they proposed designs that utilized combinations of parallel memory banks, pipelining, and store intelligent scheduling to make memory more efficient. This system optimized latency, bandwidth, and buffer occupancy and allowed routers to handle very high traffic loads with low packet drop. Garcia et al. also took an abstract investigation of memory types and access types along with trade-offs between costs, complexity, and performance. Ultimately, their architecture provided systems that leveraged the previous techniques to run concurrent multi-flow packet processing with lower contention and latency. The research expanded the understanding of memory architecture in terms of optimization for networking equipment and assisted in later discoveries and developments of scalable and high-performance buffering for packet switching and also future hardware design in networks. There was a growing relevance for what routers could accomplish to establish their functionality in the greater needs of internet traffic and backbone networks.

Balusamy et al. [24] included detailed concepts of processing and management of data and cloud computing within big data. To facilitate data processing of enormous datasets, the chapter stressed efficient storing of data, processing frames, and scalable architecture. Further explained were the categorization of cloud computing services, such as Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS), supporting data-intensive applications. Data leadership, utilization of data management, partitioning of data, replication models, and consistency in building the parts for improving processing and reliability of processing were included.Interestingly, the authors discussed challenging issues in the best practices for real-time analytics and security against less costly access or hold on collective resource use in a non-centralized distributed environment. Kogge and colleagues, documented their examination of the technical challenges required to reach exascale computing aimed at achieving a system mindset capable of executing at least a quintillion (1,018) calculations per second. The study identifies some of the factors in the design of processors, memory bandwidth, interconnect networking, and other factors in their general cross-metrics form as identified [25]. Exascale systems require enormous parallelism featuring millions of cores, order of magnitude synchronization, complexity of fault tolerance, and programming complexity. The authors also emphasized the need for heterogeneous architectures of CPUs, GPUs, and accelerators to achieve the best performance and energy consumption. In addition, they also discussed innovations in the memory hierarchy that can be integrated into interconnect system designs to overcome performance bottlenecks that can limit scalability. The report offered a future direction that they hoped for researchers and engineers could accomplish as individualized hardware and software challenges in power-efficient computation, resilient design, and more sophisticated software frameworks, laying the top of future high-performance computing systems which, in turn, would satisfy scientific, industrial, and defense computational needs. Elnozahy et al. [26] explored resilience with an orientation toward a one large scale computing system in which probability of failure will increase with some kind of scaling to exascale. This assessment detailed faults in hardware and software, focusing on detection of faults, recovery schemes, and preventive fault resilient design. As examples, the authors analyzed checkpointing, replication, and

dynamic reconfiguration to support uninterrupted collective system operation. The report also clearly stated the energy efficiency element identifying trade-offs in resilience and energy or power consumption. For resilience components to be applied to system architecture, system developers may be able to maintain performance capabilities even failures without extreme overhead and degradation of performance from the failure design. This work was significant to designing reliable exascale computing platforms to provide assurance and reliability for scientific simulations or computational systems, operational cloud systems, or high-performance workloads that entail extended runtimes for which systematic failures and incorrect computations should be avoided as a practice or normal condition. This work also illustrates resilient system modelling, monitoring and predictive maintenance methodologies. Sarkar et al. [27], explored the challenges of software in extreme-scale computing which is a computer architecture designed around exascale with millions of cores. The delayed work reports several of the most anticipated bottlenecks associated with exascale related to identifying and exploiting parallelism, as well as issues associated with memory management and fault tolerance. Overall, the report identifies the need for scalable algorithms, effective scheduling, and adaptive runtime for more complex workloads. In addition, the report highlights the programming models' high levels of abstraction, yet still have control of raw performance from an exascale potential. This is captured in their work on anticipatory fault tolerance and resilience, as anyone building work at exascale is to assume that faults will happen. Moreover, energy-efficiency in software frameworks is also examined in computing that addresses the trade-offs among computation, communication, and memory access. This report discusses the many necessary challenges to establishing principles in developing software that can utilize exascale hardware, so that capabilities for scientific simulations, data analytics, and AI workloads can flourish. The report developed a framework for studies in parallel programming frameworks, middleware, and optimization strategies for extremely large scale computing systems. Jagasivamani [28] analyzed Resistive RAM (ReRAM) technology as a potential substitute to conventional DRAM for use in main memory systems. Because of its high density, low power consumption, and non-volatility, ReRAM holds promise for energy-efficient computing systems. The dissertation studied architectural tradeoffs that include the write endurance, latency, and reliability limitations of ReRAM, and included approaches for wear leveling, error correction, and hybrid memory architectures to mitigate these concerns. The research studied the performance implications of integrating ReRAM into, primarily DRAM-based, modern memory hierarchy for multi-core and parallel workloads. The experiments reported energy savings and performance improvements under certain access patterns, while also exposing the consistency of latency and the reliability of memory write characteristics challenges. This paper proposes a deep and broad framework for thinking about the integration of ReRAM, and will support the design of next generation main-memory systems for data-centric, low-power computing domains. Albutiu et al. [29] focused on efficient query processing in main-memory multi-core databases by massively parallel sort-merge join operations as a way to mitigate some of the inefficiencies caused by traditional join operations on analytical workloads caused by sharing of memory accesses and poor parallelization. In order to partition joining tasks across cores to maximize memory locality and minimize synchronization costs, multiple solutions were proposed. Their method offered high throughput and reliable scalability to accomplish real-time analytics over extremely large data sets. By using NUMA-aware memory patterns and by scheduling join tasks in a NUMA-aware fashion, their solution minimized memory latency and memory contention work. They demonstrated remarkable performance improvement compared with joining algorithms of conventional databases, their results confirmed the performance benefits of pooled in-memory processing. This paper contributes a valuable paper to the consideration of design of database systems with the consideration of performance reporting as it relates to operational performance, specifically, with respect to optimizing the actual effectiveness of fast join processing and allocation of memory resources to establish analytical workloads. Alverson et al. [30] describe the Cray XC series interconnect network, a high performance network designed for their supercomputer. The network allows for exceptional scalability, low-latency communication, and high-bandwidth communications to support exascale-class computations. Dragonfly topology is used to capitalize on local and global links to provide optimal efficiencies in communication and lower latency when moving data. The whitepaper included detailed sections on the hardware, algorithms, and error tolerance and redundancy that keep functionality and efficiency under heavy loads. The design also emphasizes energy and modularity which allow upgrades to be made before the system becomes totally useless or inefficient. The Cray XC is capable of removing network bottlenecks, allowing large-scale parallel applications like scientific simulations and climate models to have the potential to execute efficiently. This case study is relevant with respect to interconnect architectures being utilized to address the challenges of supercomputers in

the next generation, where performance, scaling, and reliability will be a requirement.

Balkesen et al. examined hash joins and performance in main memory database systems on state-of- the-art multi-core architectures. Hash joins are beneficial for analytical queries but may also present performance bottlenecks as a result of contention for memory bandwidth, cache contention, and overheads from coordination and communication [31]. The authors showed some optimizations for memory bandwidth through cache-aware layouts, NUMA-aware, and greater parallelism. One aspect of the proposed design was to reduce overall memory stalls by balancing load on processors, and this method saw significant speed- up performance compared to traditional methods. This research identified trade-offs between memory consumption and compute efficiency and as a result also supplied a number of recommendations for system architects. This study is useful for the implementation of high-performance in-memory database engines, notably those implemented in analytics-heavy and real-time processing environments.

Recent advances in emerging memory technologies have significantly enhanced the performance, scalability, and energy efficiency of modern computing systems [32]. Non-volatile memories such as Spin-Transfer Torque Random Access Memory (STT-RAM), Phase-Change Memory (PCM), Resistive Random Access Memory (ReRAM), and Ferroelectric RAM (FeRAM) have gained increasing attention due to their superior endurance, low latency, and compatibility with CMOS technology [33]. Recent studies have demonstrated improved write endurance in STT-RAM through adaptive write schemes, enhanced multilevel storage in PCM using optimized thermal programming, and reliable filament formation control in ReRAM through material engineering. Furthermore, hybrid memory architectures combining CMOS-compatible non-volatile memory with conventional DRAM have been shown to improve system-level performance and reduce power consumption in edge computing and AI-driven applications [34]. These recent developments highlight the growing maturity and practical feasibility of emerging memory technologies [35].

Innovations in memory and material technologies have the potential to help with challenges associated with computing after the end of the Moore Law. The use of processing-in-memory (PIM) architectures, such as the DUAL framework from Imani et al. [36], allows for improved efficiency by decreasing data movement, which can speed up clustering algorithms. Additionally, ferroelectric materials based on HfO_2 have been identified as good options for next-generation non-volatile memory devices due to their scalability and their ability to be combined with CMOS technology; see for an excellent overview of the state of the art [37]. At the level of the materials, Lee et al. [38] described how oxygen vacancies are critical to the ferroelectric and resistive switching properties of the materials, allowing for performance improvements via defect engineering. Kanda et al. [39] presented insights from the 2025 IEEE Symposium on VLSI Technology and Circuits, highlighting emerging trends in high-performance and energy-efficient integrated circuit design, including innovations in device scaling, circuit optimization, and system-level integration. Their work emphasizes the growing need for advanced design methodologies to address the challenges of modern semiconductor technologies. Similarly, Majumdar and Zeimpekis [40] explored the development of analog ferroelectric field-effect transistors (FeFETs) compatible with back-end and flexible substrates, enabling accurate online training in deep neural network accelerators. Their study demonstrates the potential of novel device architectures for enhancing the efficiency and adaptability of AI hardware systems. These contributions indicate a shift towards integrating advanced device technologies and intelligent design approaches to achieve improved performance and energy efficiency in next-generation VLSI systems.

This paper is an essential reference for cloud-native big data system design for researchers and engineers by providing modern data processing system architecture (conceptually and empirically), system management practices, and considerations or optimizations related to performance.

3. Methodological Analysis

This section describes the ways in which the work advanced the memory architecture:

1. Through Simulated Experimentation: Multiple seminars (Agrawal and Sherwood [1], with their work on a virtually pipelined system among them) used mathematical modeling and simulation environments to show improvements in bandwidth/latency.
2. Hardware Prototyping: Shao et al. [3] and Nguyen et al. [9] showed practical implementation by building hardware prototypes and enabling testing.

3. Improvement Algorithms: The application of higher level computational techniques, adopting Mayfly optimization to the broadcast methodology by Pandey et al. [7] and heuristics applied by Benoit et al [10] in the context of memory management.
4. Hybrid Approach: Lin et al.'s [4] NoC design method proved traffic classification in addition to path assignment both apply a hybrid methodology when desirable to settle complex problems.

Table 1 presents a summary of the advancements in memory systems and networks. The virtual-pipelined systems developed by Agrawal and Sherwood were able to achieve a high amount of bandwidth with a uniform latency across a greater distance while Wang et al. achieved a high throughput with efficient modulation despite distortion occurring in their metrics. Van Dyke improved on memory controllers by optimizing the controllers to address increasing bandwidth with minimized latency; and Lin et al. addressed the issue of latency guarantees in NoCs; however, they still experienced starvation of best-effort (BE) traffic. Similarly to expanded work on large-scale systems, optimized messaging as conducted by Cheriton and Kutter experienced a similar metric of improved efficiency by a factor of 3–5. Studies outside of the explored areas focused on specific areas in which Zarkesh-Ha showed off the efficiency of the HyperX topology when bandwidths were high (a limitation) and Barthels implemented RDMA for efficient data movement (throughput and latency but volume, the opposite of Bandwidth). Nguyen showed 93.2% utilization of bandwidths when lab testing flexible controllers within packets; Schulz was able to receive benefit in latency and manage constrained network resources during implementation on mobile networks. Ryoo developed die-stacked DRAMs and Santella researched edge cloud computing; both improved bandwidth and minimized latency but in doing so increased scalability and data rates. Wertz and Hylton produced higher data developments with efficient data rates from their high-rate modulators and proposed space networks. Although higher developments occurred the integration was the most glaring issue, that presented study limitations among others. Overall, these technical works represent additional improvements to memory optimizations and limited improvements to system network performance optimizations, respectively.

Table 1. Previous Contributions.

S.No	Study	Key Metrics	Performance
1	High-Bandwidth Virtual Pipelines	Bandwidth, Latency	High bandwidth; Uniform latency
2	Low Latency Throughput Interface	Channel Efficiency, Propagation	Efficient modulation; Challenges in distortion
3	High Bandwidth Memory Controller	Bandwidth, Latency	Simultaneous optimization achieved
4	NoC with Latency Guarantees	Latency, Bandwidth	Starvation issues in BE traffic
5	Optimized Memory Messaging	Performance Improvement	3–5× improvement in large-scale systems
6	Topology Impacts in Shared Systems	Bandwidth Utilization	HyperX topology optimal at >4 GB bandwidth
7	Broadcasting for Wireless Networks	Throughput, Link Precision	Improved throughput; Bit defect challenges
8	RDMA in Distributed Databases	Data Movement Efficiency	Reduced overhead; Enhanced efficiency
9	Flexible Multi-Port Memory Controller	Bandwidth Utilization	93.2% utilization; Adjustments required
10	Performance Model for Memory Systems	Execution Time, Fairness	Heuristics optimized workflows
11	Random Packet Memory Networks	Latency, Energy Efficiency	Reduced cycles; Node link constraints
12	Latency in Mobile Networks	Latency Distribution	Optimized routing; Resource-intensive testing
13	Scalable Memory Subsystem Design	Guaranteed Latency, Bandwidth	Centralized policies limit scalability
14	Die-Stacked DRAM Subsystems	Latency, Bandwidth	High bandwidth; Scalability concerns
15	Edge Cloud Computing Networks	Latency, Protocol Efficiency	Latency reduced; Insufficient data rates
16	High-Radix Interconnects	Latency, Scalability	Efficient data movement; Latency persists
17	High Data Rate Modulators	Data Rate, Modulation Efficiency	2 Gbps achieved; Amplifier non-linearity
18	Simulation-Based Optimization	Model Adaptability	Dynamic reconfiguration; Specifics lacking
19	Space Networks for High Data Rates	Link Management Efficiency	High rates; Protocol integration complexity

4. Conclusion

Recent progress in memory architecture design has resulted in changes to high-performance computing (HPC) systems that now more effectively address the growing demands of new applications. The analyzed studies have collectively identified multiple ways to improve memory efficiency, bandwidth, and latency. Pipelined memory architectures, as shown in work by Agrawal and Sherwood, show that it is possible to achieve high bandwidth and consistently uniform latency, which will support predictable and effective memory accesses for HPC workloads. These designs demonstrate that optimized memory pipeline design can produce substantial performance improvements to the memory system without introducing additional excessive complexity. Beyond the use of intelligent memory pipelines, intelligent memory controllers as proposed by Van Dyke et al. incorporate techniques to minimize bandwidth and latency together. Multistream memory command execution and intelligent page context switching

are the foundations of this research area, which can be optimized for bandwidth utilization in memory systems. Such a memory controller enables agile memory systems that are able to perform adaptive operations in real time to workloads to further support throughput and in particular, latency reduction. Also, Wang et al. clearly demonstrate that contemporary technologies (for example, very close millimeter-wave memory interfaces and increased channel modulation efficiency) can exhibit significant enhancements to memory performance, even with non-linear displacement and propagation challenges.

In addition, the evolution of dynamic memory subsystems, multi-ported flexible memory controllers and adaptive arbitration continues to further promote the transition to low-latency high-throughput architectures for memory subsystems. This is important for HPC applications since memory subsystems can encounter performance limitations that significantly impact computational efficiency. Together, they indicate that memory architecture is not a passive participant but rather a computational efficiency enabler. These advances articulate a pathway to a future HPC system that, when integrated with adaptive, high-performance, scalable memory architecture, will support the increasing complexity and data intensive workloads of the future while maintaining energy efficiency and system resilience. However, there still remain several questions to be answered, including factory-scaling options (e.g., die-stamped DRAM systems have scaling challenges for memory subsystem pathways), centralized policies for memory subsystems (both hardware-based memory subsystems and software-managed memory subsystems), cook-centricity, etc., even with habitual and complex architectures that could be encountered as a modern experience of contemporary architecture. Another integration point to the framework in this section, focuses on integrating additional coupling containers that are far beyond anything contemporary to scale since it is using a traditional hardware generated protocol that still has a union with the traditional hardware protocol even with the original and hardware scalable framework designed using variations from those presented in these protocols. Again optimization can be a powerful tool but as stated earlier it can also include negative effects of realities in real world that occur outside our lab for example, non-linear distortions in a scheme employed, i.e., starvation of best effort traffic in a network on chips sense.

These issues will be discussed in future investigations wherein hybrid approaches that combine hardware and software technologies will be investigated. Hybrid approaches may be on the level of predictive memory management approaches denoting artificial intelligence; machine learning integrated in models that dynamically optimize in real-time; modular architecture suited for flexible scaling. Filling these gaps will better prepare architectures of memory for the next generation HPC.

Author Contributions

Conceptualization, S.R. and S.K.; methodology, S.K.; software, S.K.; validation, S.R. and S.K.; formal analysis, S.K.; investigation, S.R.; resources, S.K.; data curation, S.R.; writing—original draft preparation, S.R.; writing—review and editing, S.K.; visualization, S.K.; supervision, S.K.; project administration, S.K.; funding acquisition, S.K. Both authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data used in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Agrawal, B.; Sherwood, T. High-bandwidth network memory system through virtual pipelines. *IEEE ACM Trans. Netw.* **2009**, *17*, 1029–1041. [CrossRef]
2. Wang, Q.; Guidotti, D.; Lin, F.; et al. Low latency high throughput memory-processor interface. In Proceedings of the 2012 IEEE 62nd Electronic Components and Technology Conference, San Diego, CA, USA, 29 May–1 June 2012. [CrossRef]
3. Shao, M.H.; Zhao, R.T.; Liu, H.; et al. Challenges and recent advances in HfO2-based ferroelectric films for non-volatile memory applications. *Chip* **2024**, *3*, 100101.
4. Lin, S.; Su, L.; Su, H.; et al. Design networks-on-chip with latency/bandwidth guarantees. *IET Comput. Digit. Tech.* **2009**, *3*. [CrossRef]
5. Cheriton, D.R.; Kutter, R.A. Optimized Memory-Based Messaging: Leveraging the Memory System for High-Performance Communication. *Comput. Syst.* **1996**, *9*, 179–215.
6. Milton, J.; Zarkesh-Ha, P. Impacts of Topology and Bandwidth on Distributed Shared Memory Systems. *Computers* **2023**, *12*, 86. [CrossRef]
7. Pandey, D.; Pandey, B.K.; Nassa, V.K.; et al. Optimized Throughput-Based Broadcasting for Next Generation Wireless Networks. In *Emerging Engineering Technologies and Industrial Applications*; IGI Global Scientific Publishing: Hershey, PA, USA, 2024. [CrossRef]
8. Barthels, C.; Alonso, G.; Hoefler, T. Designing Databases for Future High-Performance Networks. *IEEE Data Eng. Bull.* **2017**, *40*, 15–26.
9. Nguyen, X.-T.; Le, D.-H.; Bui, T.-T.; et al. A Flexible High-Bandwidth Low-Latency Multi-Port Memory Controller. *Vietnam J. Sci. Technol.* **2017**, *56*, 357–369. [CrossRef]
10. Benoit, A.; Perarnau, S.; Pottier, L.; et al. A Performance Model to Execute Workflows on High-Bandwidth-Memory Architectures. In Proceedings of the 47th International Conference on Parallel Processing, Eugene, OR, USA, 13–16 August 2018. [CrossRef]
11. Fujiki, D.; Matsutani, H.; Koibuchi, M.; et al. Randomizing Packet Memory Networks for Low-Latency Processor-Memory Communication. In Proceedings of the 2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), Heraklion, Greece, 17–19 February 2016. [CrossRef]
12. Schulz, P.; Ong, L.; Abdullah, B.; et al. End-to-End Latency Distribution in Future Mobile Communication Networks. *IEEE Access* **2020**. Available online: https://www.vodafone-chair.org/pbls/philipp-schulz/End-to-End_Latency_Distribution_in_Future_Mobile_Communication_Networks.pdf
13. Cheema, S.S.; Shanker, N.; Hsu, C.H.; et al. One nanometer HfO_2-based ferroelectric tunnel junctions on silicon. *Adv. Electron. Mater.* **2022**, *8*, 2100499.
14. Setter, N.; Damjanovic, D.; Eng, L.; et al. Ferroelectric thin films: Review of materials, properties, and applications. *J. Appl. Phys.* **2006**, *100*, 051606.
15. Santella, G.; Vatalaro, F. An approach to define Very High Capacity Networks with improved quality at an affordable cost. *arXiv preprint* **2020**, *arXiv:2011.03685*. [CrossRef]
16. Li, S.; Huang, P.-C.; Banks, D.; et al. Low Latency, High Bisection-Bandwidth Networks for Exascale Memory Systems. In Proceedings of the Second International Symposium on Memory Systems, Alexandria, VA, USA, 3–6 October 2016; pp. 62–73. [CrossRef]
17. Wertz, P.; Hespeler, B.; Kiessling, M.; et al. Next generation high data rate downlink subsystems based on a flexible APSK modulator applying SCCC encoding. In Proceedings of the 2016 International Workshop on Tracking, Telemetry and Command Systems for Space Applications (TTC), Noordwijk, The Netherlands, 13–16 September 2016.
18. Hagendorf, O.; Pawletta, T. An approach for simulation based structure optimisation of discrete event systems. In Proceedings of the 2008 Spring Simulation Multiconference, Ottawa, ON, Canada, 14–17 April 2008; pp. 431–438.
19. Hylton, A.; Raible, D.E.; Clark, G. On the Development and Application of High Data Rate Architecture (HiDRA) in Future Space Networks. In *AIAA 2017-5415*; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2017. [CrossRef]
20. Agrawal, B.; Sherwood, T. Virtually pipelined network memory. In Proceedings of the 39th Annual IEEE/ACM

International Symposium on Microarchitecture, Orlando, FL, USA, 9–13 December 2006; pp. 197–207.
21. Nikologiannis, A.; Katevenis, M. Efficient per-flow queueing in DRAM at OC-192 line rate using out-of-order execution techniques. In Proceedings of the IEEE International Conference on Communications, Helsinki, Finland, 11–14 June 2001; pp. 2048–2052.
22. Iyer, S.; Kompella, R.R.; McKeown, N. *Designing Packet Buffers for Router Linecards*; Stanford University: Stanford, CA, USA, 2002.
23. Garcia, J.; Corbal, J.; Cerda, L.; et al. Design and implementation of high-performance memory systems for future packet buffers. In Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture, San Diego, CA, USA, 7–11 December 2003; pp. 372–384.
24. Balusamy, B.; Abirami, R.N.; Kadry, S.; et al. Processing, Management Concepts, and Cloud Computing. In *Big Data: Concepts, Technology, and Architecture*; Wiley: Hoboken, NJ, USA, 2021; pp. 83–110.
25. Kogge, P.; Bergman, K.; Borkar, S.; et al. *Exascale Computing Study: Technology Challenges in Achieving Exascale Systems*; DARPA IPTO: Arlington, VA, USA, 2008.
26. Elnozahy, E.N.; Bianchini, R.; El-Ghazawi, T.; et al. *System Resilience at Extreme Scale*; DARPA IPTO: Arlington, VA, USA, 2008.
27. Sarkar, V.; Amarasinghe, S.; Campbell, D.; et al. *ExaScale Software Study: Software Challenges in Extreme Scale Systems*; DARPA IPTO: Arlington, VA, USA, 2009.
28. Jagasivamani, M. Resistive Ram Based Main-Memory Systems: Understanding the Opportunities, Limitations, and Tradeoffs. PhD Thesis, University of Maryland, College Park, MD, USA, 2020.
29. Albutiu, M.-C.; Kemper, A.; Neumann, T. Massively Parallel Sort-Merge Joins in Main Memory Multi-Core Database Systems. *Proc. VLDB Endow.* **2012**, *5*, 1064–1075.
30. Alverson, B.; Froese, E.; Kaplan, L.; et al. *Cray XC Series Network*; Cray Inc.: 2012.
31. Balkesen, C.; Teubner, J.; Alonso, G.; et al. Main-Memory Hash Joins on Modern Processor Architectures. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 1754–1766.
32. Zhou, X.; Chen, Y.; Kang, J. Recent progress in emerging non-volatile memories: Materials, devices, and applications. *IEEE Trans. Electron Devices* **2021**, *68*, 1023–1035.
33. Raoux, S.; Wełnic, W.; Lelmini, D. Phase change materials and their application to nonvolatile memories. *Chem Rev.* **2010**, *110*, 240–267.
34. Chen, A. ReRAM: History, status, and future. *IEEE Trans. Electron Devices* **2020**, *67*, 1420–1433.
35. Zhou, Z.; Li, L.; Feng, Y.; et al. Advancing the frontiers of HfO_2-based ferroelectric memories: innovative concepts from materials to applications. *Adv. Mater.* **2025**, *37*, e09525. [CrossRef]
36. Imani, M.; Pampana, S.; Gupta, S.; et al. DUAL: Acceleration of Clustering Algorithms Using Digital-Based Processing In-Memory. In Proceedings of the 2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), Athens, Greece, 17–21 October 2020; pp. 356–371. [CrossRef]
37. Liao, J.; Dai, S.; Peng, R.C.; et al. HfO2-based ferroelectric thin film and memory device applications in the post-Moore era: A review. *Fundam. Res.* **2023**, *3*, 332–345. [CrossRef]
38. Lee, J.; Yang, K.; Kwon, J.Y.; et al. Role of oxygen vacancies in ferroelectric or resistive switching hafnium oxide. *Nano Converg.* **2023**, *10*, 55. [CrossRef]
39. Kanda, M.; Tokuda, T.; Fan, Q. 2025 IEEE Symposium on VLSI Technology and Circuits [Conference Reports]. *IEEE Solid-State Circuits Mag.* **2025**, *17*, 112–126. [CrossRef]
40. Majumdar, S.; Zeimpekis, I. Back-End and Flexible Substrate Compatible Analog Ferroelectric Field-Effect Transistors for Accurate Online Training in Deep Neural Network Accelerators. *Adv. Intell. Syst.* **2023**, *5*, 2300391. [CrossRef]