

Microprocessor Architectures for Web Networks: A Review of SmartNICs, DPUs, and Heterogeneous Computing Platforms

Haider W. Olewi¹, Saad T. Y. Alfalahi^{2,*}, Zahraa J. Mahdi³

^{1,2,3} Department of Computer Engineering Techniques, Madenat Alelem University College, Baghdad, Iraq

*Corresponding author: saad.t.yasin@mauc.edu.iq

Received: ----

Accepted: ----

Published: ----

Abstract

The rapid growth of cloud computing, hyperscale data centers, artificial intelligence services, and programmable web infrastructures has significantly transformed the design requirements of modern microprocessors. Traditional server-centric architectures based solely on general-purpose CPUs are increasingly challenged by the rising demands of networking, virtualization, storage management, and large-scale distributed computing workloads. This paper presents a systematic review of microprocessor architectures for modern web networks, with emphasis on heterogeneous computing platforms integrating infrastructure-class CPUs, SmartNICs, Data Processing Units (DPUs), and programmable network processors. The study reviews recent developments in cloud-oriented processor architectures such as Arm Neoverse, NVIDIA Grace, and Google Axion, alongside emerging SmartNIC and DPU technologies designed for infrastructure offloading and programmable data-plane acceleration. The reviewed architectures are comparatively analyzed in terms of throughput, latency, energy efficiency, scalability, programmability, and infrastructure offloading capability. The study further discusses the role of programmable networking technologies in enabling adaptive packet processing, virtualization, security enforcement, and AI-assisted web services. The review demonstrates that heterogeneous and network-centric computing architectures provide substantial improvements in scalability, resource utilization, and cloud-service efficiency compared with conventional host-centric systems. Finally, current challenges and future research directions are highlighted, including interoperability, orchestration complexity, security management, and the integration of AI-native acceleration technologies within next-generation web-network infrastructures.

Keywords:

Smart Network Interface Cards, Data Processing Units, Heterogeneous Computing, Web Network Architectures, Infrastructure-Class Microprocessors.

I. INTRODUCTION

THE rapid expansion of cloud computing, hyperscale data centers, web services, and content delivery platforms has significantly transformed the architecture of modern computing systems. Traditional monolithic server

infrastructures are increasingly being replaced by highly distributed and scalable web network environments that support billions of connected users and devices worldwide. Applications such as artificial intelligence services, large-scale data analytics, video streaming, virtualization, and edge computing require extremely high throughput, low latency, and efficient resource management. Consequently, the role of

microprocessors has evolved beyond conventional general-purpose computation toward specialized and heterogeneous processing architectures optimized for modern network-centric workloads [1].

Historically, central processing units (CPUs) were designed primarily for sequential and general-purpose tasks. However, the explosive growth of network traffic and cloud-native applications has introduced substantial processing overhead associated with packet handling, virtualization, encryption, storage management, and network orchestration. As a result, relying solely on conventional CPUs has become increasingly inefficient in large-scale web infrastructures. Recent studies have shown that network-related workloads consume a significant portion of server processing resources, reducing computational efficiency for application-level services [2].

To address these limitations, modern web network architectures increasingly employ heterogeneous computing models that combine infrastructure-class CPUs with specialized acceleration hardware. Technologies such as Smart Network Interface Cards (SmartNICs), Data Processing Units (DPUs), programmable network processors, and hardware accelerators have emerged as key components in next-generation data center and cloud infrastructures. These architectures enable computational offloading of networking, security, storage, and virtualization tasks from host CPUs, thereby improving overall system efficiency and scalability.

SmartNICs and DPUs represent a major shift in network-aware microprocessor design. Unlike traditional network interface controllers, SmartNICs integrate programmable processors, hardware accelerators, and memory resources directly within the network interface layer. This enables advanced packet processing, traffic management, security filtering, and virtualization support at the network edge. Similarly, DPUs combine multicore processors, high-speed networking interfaces, and dedicated acceleration engines into highly integrated platforms capable of handling infrastructure services independently of host CPUs. NVIDIA BLUEFIELD, Intel IPU architectures, and AMD PENSANDO platforms are prominent examples of this emerging class of network-centric processors. In parallel, infrastructure-class CPUs have also evolved to support cloud-scale workloads more efficiently. Recent architectures such as Arm Neoverse-based systems, NVIDIA Grace processors, and Google Axion emphasize energy efficiency, high memory bandwidth, scalable multicore processing, and optimized cloud-native execution environments. These processors are specifically designed to support modern distributed computing frameworks and high-density server deployments while reducing operational power consumption.

Several foundational studies have contributed significantly to the understanding of modern computer architecture and heterogeneous systems. Hennessy and Patterson (2019) emphasized the growing importance of domain-specific architectures and specialized accelerators in addressing performance and energy challenges in large-scale computing

systems [2]. Pfister (2015) further discussed the evolution of networked computing infrastructures and scalable multiprocessing systems required for modern cloud environments [3]. More recent studies on SmartNICs and DPUs have demonstrated the effectiveness of offloading network and infrastructure services to dedicated processing units, reducing CPU bottlenecks and improving data center efficiency.

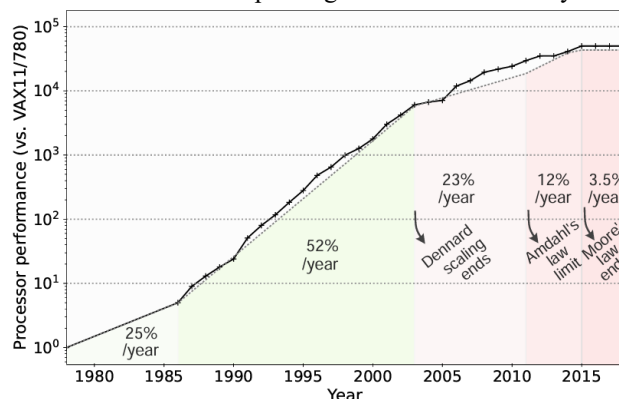


Fig. 1. Growth in processor performance over 40 years [4].

Motivated by these developments, this paper presents a systematic review of microprocessor architectures for modern web networks, focusing on heterogeneous computing models, infrastructure-class CPUs, SmartNICs, DPUs, and programmable network processors. The review highlights architectural trends, hardware offloading mechanisms, performance optimization strategies, and emerging challenges associated with next-generation web infrastructure design. The paper also examines how specialized microprocessors are reshaping cloud computing, edge networking, and hyperscale data center architectures in response to the increasing demands of modern internet-scale applications.

II. LITERATURE REVIEW

The evolution of modern web networks has significantly influenced the design of microprocessor architectures for cloud computing, hyperscale data centers, and programmable networking infrastructures. Traditional server-centric computing models based solely on general-purpose CPUs are increasingly unable to efficiently support modern workloads characterized by massive network traffic, virtualization, artificial intelligence (AI), and large-scale distributed services. Consequently, recent research has focused on heterogeneous computing architectures that combine CPUs with specialized network-centric processors such as SmartNICs, Data Processing Units (DPUs), and programmable accelerators.

Recent literature identifies SmartNICs and DPUs as key technologies for improving infrastructure efficiency by offloading networking, storage, virtualization, and security functions from host CPUs. Kfoury et al. presented a comprehensive survey of SmartNIC architectures, highlighting their role in accelerating packet processing, traffic management, and cloud infrastructure services through programmable hardware integration. The study emphasized that

modern SmartNICs increasingly incorporate embedded multicore processors, FPGA-based acceleration, and programmable data-plane support to reduce CPU bottlenecks and improve scalability [4].

Similarly, recent surveys on heterogeneous computing using SmartNICs and DPUs demonstrated that these devices enable distributed processing models where network services are executed closer to the data path rather than inside centralized CPUs. Tibbetts et al. discussed how DPUs extend SmartNIC functionality by integrating dedicated acceleration engines and independent processing cores capable of supporting AI inference, communication offloading, and cloud-native orchestration services [5].

Programmable network processors and programmable data planes have also become an important research direction in modern web infrastructure. Recent studies emphasized the importance of programmable data-plane technologies in enabling flexible packet processing, adaptive network control, and software-defined networking services. These architectures provide significant advantages for traffic engineering, telemetry, network security, and cloud service optimization [6].

Another major trend in recent literature is the emergence of infrastructure-class CPUs specifically optimized for cloud-scale workloads. Arm Neoverse-based architectures have received substantial attention due to their energy efficiency, scalability, and suitability for hyperscale data centers. Recent industrial and academic studies indicate that processors such as NVIDIA Grace and Google Axion are designed to provide high memory bandwidth, multicore scalability, and improved performance-per-watt for AI and cloud-native workloads [7]. Arm-based server architectures are increasingly adopted by hyperscale providers because of their reduced power consumption and efficient parallel processing capabilities. Communication offloading has also emerged as a critical research topic in DPU-enabled systems. Wahlgren et al. demonstrated that SmartNIC DPUs can significantly reduce CPU communication overhead by executing networking services directly at the network interface layer. Their quantitative evaluation showed improved application performance and reduced CPU resource consumption in large-scale computing environments [8].

Recent research has additionally explored SmartNIC-assisted memory disaggregation and distributed computing acceleration. Wahlgren et al. proposed a SmartNIC-enabled architecture for disaggregated memory systems using NVIDIA BLUEFIELD DPUs. Their experimental implementation demonstrated considerable performance improvements and reduced network traffic for graph-processing applications, highlighting the growing role of programmable network-centric processors in future cloud infrastructures [9].

Security-oriented SmartNIC applications have also become increasingly important in recent web network architectures. Elizalde et al. reviewed SmartNIC-based security applications including intrusion detection systems (IDS), distributed denial-of-service (DDoS) mitigation, and privacy-preserving

mechanisms. Their study highlighted the potential of SmartNICs to strengthen network defense capabilities while simultaneously reducing computational overhead on host servers [10].

Overall, recent literature confirms that heterogeneous computing architectures based on SmartNICs, DPUs, programmable network processors, and infrastructure-class CPUs are becoming fundamental components of next-generation web networks. These technologies collectively improve scalability, workload acceleration, network efficiency, and cloud infrastructure optimization while supporting the increasing computational demands of AI-driven and internet-scale applications.

III. METHODOLOGY

This study adopts a systematic review methodology to investigate recent microprocessor architectures developed for modern web-network infrastructures. Unlike experimental hardware studies, the methodology focuses on literature collection, architectural classification, comparative analysis, and trend evaluation of heterogeneous computing platforms used in cloud and hyperscale environments.

Relevant publications were collected from major scientific databases including IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, arXiv, and industrial technical reports. The review primarily considers recent works published between 2020 and 2025 related to SmartNICs, Data Processing Units (DPUs), programmable network processors, infrastructure-class CPUs, and heterogeneous computing systems for web networks.

The selected architectures were classified into four major categories:

1. Infrastructure-class CPUs,
2. SmartNIC architectures,
3. DPUs/IPUs,
4. Programmable network processors.

This classification facilitates a structured comparison between traditional host-centric processing and emerging network-centric offloading approaches. Additionally, the comparative evaluation considered several technical parameters, including processing throughput, latency, energy efficiency, scalability, programmability, infrastructure offloading capability, AI acceleration support, and security functionality. In addition, recent architectural trends such as cloud-native acceleration, programmable data planes, and heterogeneous computing integration were critically analyzed to identify future directions for web-network microprocessor design. Figure 2 illustrates the general classification of modern web-network microprocessor architectures, while Table 1 summarizes the main evaluation parameters used throughout comparative analysis.

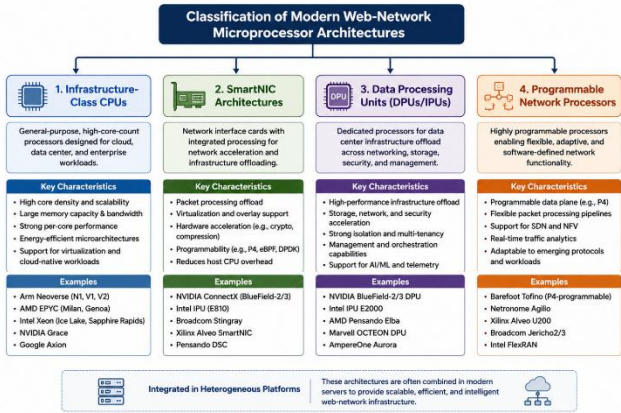


Fig. 2. General classification of modern web-network microprocessor architectures.

TABLE I
EVALUATION PARAMETERS USED THROUGHOUT
COMPARATIVE ANALYSIS

Evaluation Parameter	Purpose in Comparative Analysis
Processing Throughput	Measures the ability of the architecture to handle high-volume network and data workloads efficiently.
Latency	Evaluates the response delay and real-time performance of processing and communication operations.
Energy Efficiency	Assesses power-performance optimization for cloud and hyperscale data-center environments.
Scalability	Determines the capability of the architecture to support large-scale distributed web infrastructures.
Programmability	Measures flexibility in implementing custom packet processing, virtualization, and network functions.
Infrastructure Offloading Capability	Evaluates the ability to reduce CPU workload by offloading networking, storage, and security tasks.
AI Acceleration Support	Assesses compatibility with AI-driven workloads and machine-learning acceleration mechanisms.
Security Functionality	Evaluate support for encryption, intrusion detection, isolation, and secure infrastructure management.
Memory Bandwidth	Measures the capability of the processor to support high-speed data movement and intensive applications.
Cloud-Native Support	Assesses suitability for virtualization, containerization, orchestration, and cloud-native deployment environments.

IV. ARCHITECTURAL ANALYSIS AND DISCUSSION

The rapid evolution of web-scale computing has significantly transformed the architectural design of modern microprocessors. Traditional server architectures based solely on general-purpose CPUs are increasingly insufficient for handling the intensive networking, virtualization, storage, and AI workloads generated by

hyperscale cloud infrastructures. Consequently, modern web-network systems are progressively adopting heterogeneous computing architectures that integrate infrastructure-class CPUs with SmartNICs, DPUs/IPUs, programmable accelerators, and specialized networking processors.

A. Infrastructure-Class CPU Architectures

Infrastructure-class CPUs remain the computational foundation of modern cloud and web-network infrastructures. Recent processors such as AMD EPYC, Intel Xeon, Arm Neoverse platforms, NVIDIA Grace, and Google Axion are specifically designed to support large-scale distributed computing environments. These processors emphasize:

- high core density,
- large memory bandwidth,
- energy efficiency,
- and scalable multiprocessing capability.

Figure 3 illustrates a Grace-style monolithic mesh architecture and its role in supporting modern web and cloud workloads. Arm-based architectures have recently gained significant attention in hyperscale data centers because of their improved performance-per-watt characteristics.

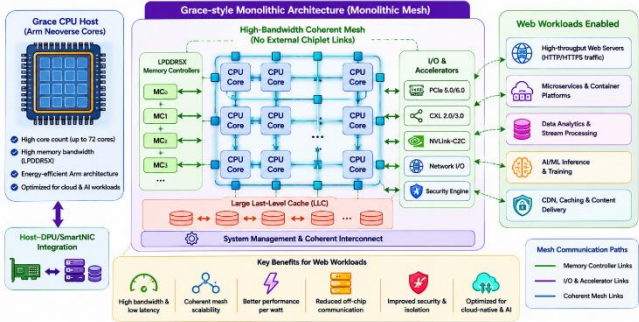


Fig. 3. Grace-style Monolithic Mesh and Its Role in Web Workloads.

NVIDIA Grace and Google Axion processors demonstrate how cloud-oriented CPU architectures are increasingly optimized for AI-driven workloads, virtualization, and high-density server deployment. Compared with traditional x86 architectures, Arm-based infrastructures often provide lower power consumption and improved scalability for cloud-native applications.

Despite these advantages, infrastructure CPUs still experience considerable overhead when executing networking and infrastructure-management tasks directly on host processors. This limitation has accelerated the development of network-centric offloading technologies. The major classes of microprocessors used in modern web-network infrastructures are summarized in Table 2.

TABLE II
CLASSES OF MICROPROCESSORS FOR WEB NETWORKS

Class	Typical Location	Primary Role in Web Networks	Example Architectures/Products
Infrastructure re CPU	Server nodes,	Application logic, web	Arm Neoverse N1/N2, NVIDIA

	control planes	servers, microservices, orchestration	Grace, Google Axion, x86 server CPUs
Network Processor (NP)	Routers, switches, middleboxes	Line-rate packet forwarding, classification, QoS, tunneling	NPU100-class designs, carrier-grade NPs
SmartNIC	Server NIC with on-board compute	Offload vSwitch, transport, encryption, telemetry, storage	AMD Pensando, Intel IPU, NVIDIA BlueField, FPGA-based SmartNICs
Data Processing Unit (DPU)	SmartNIC-style SoC, disaggregated	Heterogeneous compute for data, network, and security services	NVIDIA BlueField-2/3, various DPU vendors
On-Chip Interconnect / NoC	Inside CPUs, DPUs, accelerators	High-bandwidth, low-latency data movement across cores/engine	NVIDIA SCF, Arm CMN-600

B. SmartNIC Architectures

SmartNICs represent one of the most significant developments in modern web-network processor design. Unlike conventional network interface cards, SmartNICs integrate programmable processing units and acceleration hardware directly within the network interface layer. Their primary purpose is to offload networking and infrastructure services from host CPUs.

Modern SmartNIC platforms commonly support packet processing, virtualization acceleration, storage management, encryption, telemetry, and software-defined networking functions. Figure 4 illustrates the interaction between the host system, SmartNIC processing pipeline, and network infrastructure in modern web-network architectures.

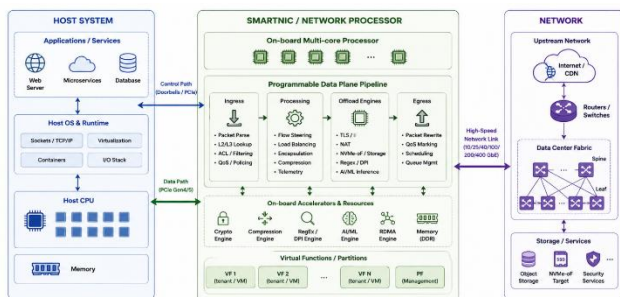


Fig. 4. Host-SmartNIC-network processor pipeline for Web traffic.

Technologies such as NVIDIA BlueField, Intel IPU, Pensando DSC, and FPGA-based SmartNICs provide programmable packet-processing pipelines capable of reducing

CPU utilization while improving network throughput and latency performance.

A major strength of SmartNIC architectures lies in their programmability. Support for technologies such as: P4, eBPF, DPDK, and programmable data planes enables highly flexible network-service deployment. However, SmartNICs may introduce architectural complexity, software integration challenges, and increased development requirements compared with traditional fixed-function NICs.

C. Data Processing Units (DPUs) and Infrastructure Offloading

DPUs extend the SmartNIC concept by integrating multicore processors, memory subsystems, networking engines, and dedicated accelerators into highly autonomous infrastructure-processing platforms. DPUs are specifically designed to execute infrastructure services independently from host CPUs.

Their major functions include network acceleration, storage virtualization, security enforcement, telemetry, orchestration, and AI-assisted infrastructure management. Figure 5 illustrates the simplified internal architecture of a modern SmartNIC/DPU platform used for infrastructure offloading in web-network environments.

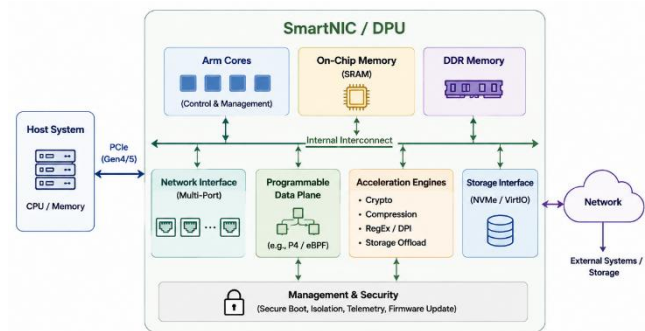


Fig. 5. SmartNIC/DPU Internal Block Diagram.

Unlike conventional SmartNICs, DPUs provide stronger computational isolation and can execute infrastructure workloads with minimal host-CPU intervention. This significantly reduces communication overhead and improves scalability in hyperscale cloud environments.

Recent studies demonstrate that DPU-based offloading substantially improves data-center efficiency, infrastructure scalability, workload isolation, and application-level performance.

Nevertheless, DPUs introduced additional hardware cost, architectural complexity, and software orchestration requirements. Efficient integration between CPUs, DPUs, accelerators, and virtualization frameworks remains a major research challenge.

D. Programmable Network Processors

Programmable network processors and programmable data planes are increasingly important in software-defined networking (SDN), cloud-native infrastructures, and adaptive web-network services. Unlike fixed-function networking

hardware, programmable processors allow dynamic implementation of custom packet-processing pipelines and network functions. Modern programmable data-plane technologies support adaptive traffic engineering, network telemetry, packet filtering, dynamic routing, intrusion detection, and virtualization services. Architectures based on P4-programmable pipelines, FPGA acceleration, and software-defined infrastructure frameworks provide considerable flexibility for future web-network applications.

However, fully programmable architectures often require sophisticated development tools, compiler support, runtime management systems, and advanced orchestration frameworks. Balancing programmability, security, and performance remains a key challenge in programmable network infrastructure research.

E. Comparative Discussion

The reviewed architectures collectively demonstrate the transition from host-centric computing toward heterogeneous and network-centric computing models. Infrastructure CPUs continue to provide general-purpose computational capability, while SmartNICs and DPUs increasingly assume responsibility for networking, security, storage, and infrastructure-management tasks.

This heterogeneous approach offers several advantages:

- reduced CPU overhead,
- improved scalability,
- lower latency,
- enhanced infrastructure isolation,
- and better energy efficiency.

In addition, programmable networking technologies provide flexibility for evolving cloud-native applications and AI-driven infrastructures.

Despite these benefits, several challenges remain unresolved, including:

- hardware/software integration complexity,
- interoperability between heterogeneous platforms,
- orchestration overhead,
- security management,
- and standardization limitations.

Future web-network infrastructures are expected to increasingly combine:

- infrastructure-class CPUs,
- SmartNICs,
- DPUs,
- AI accelerators,
- and programmable networking frameworks

within unified heterogeneous computing ecosystems capable of supporting next generation of hyperscale cloud services and intelligent distributed applications. A qualitative comparison between representative web-network processing platforms is presented in Table 3.

TABLE III
REPRESENTATIVE PLATFORMS FOR WEB/NETWORK
WORKLOADS

Feature / Platform	Arm Neoverse N1 (CPU)	NVIDIA Grace (CPU)	Google Axion (CPU)	Smart NIC / DPU (generic)	Classic NP (router)
Target workloads	Cloud-native apps, web services, edge compute	Mixed cloud + AI adjacency, high BW data center workloads	General web, AI-adjacent cloud workloads	Network, security, storage, virtualization offload	Packet forwarding, QoS, tunneling, routing
Core type and count	64-bit Arm cores; scalable multi-core designs	72 Arm Neoverse cores in a monolithic NUMA-free mesh	Arm Neoverse v2-based, up to 96 vCPUs per instance	Multiple ARM/RISC cores + domain accelerators	Tens of simple RISC cores or pipelines
Memory subsystem	Large, shared caches, DDR-based memory; power-efficient per core	LPDDR5X with very high aggregate bandwidth, unified L3 cache	DDR5-5600 MT/s, large L2/L3 cache, UMA per node	On-board DRAM, SRAM/TCAM for tables, NIC buffers	SRAM/DRAM hierarchies tuned for per-packet operations
Network interface	Attached via NIC (can be SmartNIC/DPU)	Integrated with high-speed NVLink-C2C, NICs, accelerators	High-bandwidth networking (e.g., 50–100 Gbps per instance)	Integrated NIC ports (10–400 Gbps), programmable data planes	High-speed front-panel ports in routers/switches
Offload / accelerators	None (or limited) for networking; relies on NICs	Potential integration with GPUs/TPUs; designed to feed accelerators efficiently	Works with TPUs and Titanium controllers for networking/security offload	Crypto, compression, parsing, match-action, virtualization offload, storage offload	Fixed-function accelerators for classification, QoS, and encryption

Program mability	General - purpose OS; C/C++; containe rs, VMs	Same as Neove rse, with tuning for NUM A-free domai n	General cloud stack; Kubernet es, microserv ices, etc.	Mix of general - purpos e cores (Linux) + P4 pipelin es + vendor SDKs	Typicall y, proprieta ry microco de/SDKs ; some P4- capable designs
---------------------	---	---	--	---	---

- [8] Wahlgren, J., Hu, A., Pearce, R., Gokhale, M., & Peng, I. (2026). Communication offloading on SmartNIC DPUs: A quantitative approach. arXiv. <https://arxiv.org/abs/2605.04842>
- [9] Wahlgren, J., Schieffer, G., Gokhale, M., Pearce, R., & Peng, I. (2024). Disaggregated memory with SmartNIC offloading: A case study on graph processing. arXiv. <https://arxiv.org/abs/2410.02599>
- [10] Elizalde, S., AlSabeh, A., Mazloun, A., Choueiri, S., Kfoury, E., Gomez, J., & Crichigno, J. (2025). A survey on security applications with SmartNICs: Taxonomy, implementations, challenges, and future trends. Journal of Network and Computer Applications, 242, 104257. <https://doi.org/10.1016/j.jnca.2025.104257>

V. CONCLUSION

This review examined the evolution of modern microprocessor architectures designed for web-network infrastructures, with emphasis on infrastructure-class CPUs, SmartNICs, DPUs/IPUs, and programmable network processors. The analysis demonstrated that heterogeneous computing platforms have become essential for supporting the increasing demands of cloud computing, AI-driven services, hyperscale data centers, and programmable networking environments. The reviewed studies showed that SmartNICs and DPUs significantly reduce host-CPU overhead by offloading networking, storage, security, and infrastructure-management tasks, while programmable network processors improve flexibility and adaptability for software-defined web services. In parallel, modern infrastructure-class CPUs such as Arm Neoverse, NVIDIA Grace, and Google Axion provide scalable and energy-efficient processing capabilities optimized for cloud-native workloads. Despite the advantages of heterogeneous architectures, several challenges remain, including software integration complexity, orchestration overhead, interoperability, and security management. Future web-network infrastructures are expected to increasingly integrate CPUs, SmartNICs, DPUs, AI accelerators, and programmable data-plane technologies within unified heterogeneous computing ecosystems capable of supporting next-generation intelligent and scalable digital services.

REFERENCES

- [1] Hennessy, J. L., & Patterson, D. A. (2019). A new golden age for computer architecture: Domain-specific hardware/software co-design, enhanced security, open instruction sets, and agile chip development. ACM/IEEE International Symposium on Computer Architecture (ISCA). <https://doi.org/10.1145/3307650.3322271>
- [2] Patterson, D. A., & Hennessy, J. L. (2021). Computer organization and design RISC-V edition: The hardware/software interface (2nd ed.). Morgan Kaufmann.
- [3] Pfister, G. F. (2015). In search of clusters (3rd ed.). Prentice Hall.
- [4] Kfoury, E., Choueiri, S., Mazloun, A., AlSabeh, A., Gomez, J., & Crichigno, J. (2024). A comprehensive survey on SmartNICs: Architectures, development models, applications, and research directions. IEEE Access, 12, 17297–17335. <https://doi.org/10.1109/ACCESS.2024.3437203>
- [5] Tibbetts, N., Ibtisum, S., & Puri, S. (2025). A survey on heterogeneous computing using SmartNICs and emerging data processing units. Future Generation Computer Systems. <https://arxiv.org/abs/2504.03653>
- [6] Michel, O., Keller, E., & Sherry, J. (2022). Programmable data planes: A survey of applications and techniques. IEEE Communications Surveys & Tutorials, 24(3), 1985–2012.
- [7] Arm Ltd. (2024). Arm Neoverse platforms for cloud and hyperscale computing. Arm Research Reports.