Heterogenous Computing Architectures in SoCs for 5G Communication

Tejaswini T[1], Tarun S N[2], Manjunath Umesh Kudalagi[3], Dr. Govinda Raju M[4]
Department of Electronics and Communication Engineering RV College of Engineering Bangalore, India

***Abstract*—The advent of 5G communication systems has enormously raised the need for low-power, high-performance computing solutions that can manage massive data processing in real-time. Conventional homogeneous computing structures fail to cope with these needs because they lack adequate parallel processing capacity and high-power consumption. System-on- Chip (SoC) heterogeneous computing architectures provide a viable solution with multiple specialized processing blocks such as Central Processing Units (CPUs), Graphics Processing Units (GPUs), Digital Signal Processors (DSPs), Field-Programmable Gate Arrays (FPGAs), and Application-Specific Integrated Cir- cuits (ASICs). Heterogeneous architectures allow for optimized task allocation, boosting computational performance, power man-agement, and scalability for 5G functions like signal processing, beamforming, channel estimation, and error correction. Never- theless, hurdles like software complexity, memory management, interconnect overhead, and thermal issues have to be dealt with to enable the complete promise of heterogeneous SoCs in 5G. This paper discusses design methodologies, optimization strategies, and future directions of heterogeneous computing architecture, emphasizing the pivotal role played by them in next-generation wireless communication.**

***Index Terms*—5G Communication, Heterogeneous Computing, SoCs, CPUs, GPUs, DSPs, FPGAs, ASICs, Parallelism, Energy Efficiency, Beamforming, Channel Estimation, Error Correction.**

I. INTRODUCTION

The introduction of 5G wireless communication systems has opened the doors to a new world of low-latency, high- speed wireless connectivity, transforming many applications and industries. The real-time processing of massive data poses daunting challenges to the conventional computing models. Heterogeneous computing architectures in SoC design have stepped up to mitigate these challenges with the capability of hosting several optimized processing units to cater to a wide range of workloads.

Heterogeneous computing platforms utilize diverse process- ing units, such as Central Processing Units (CPUs), Graphics Processing Units (GPUs), Digital Signal Processors (DSPs), Field-Programmable Gate Arrays (FPGAs), and Application- Specific Integrated Circuits (ASICs). These elements are indi-vidually important in terms of maximizing the computational efficiency, power usage, and performance of SoCs that are designed for 5G communication. In contrast to homogeneous architectures that depend only on general-purpose CPUs, heterogeneous SoCs allocate workloads between special pro- cessors such that parallelism and efficiency are achieved in the processing of 5G-specific functions like signal processing, beamforming, channel estimation, and error correction.

5G networks impose unprecedented requirements, such as ultra-reliable low-latency communication (URLLC), upgraded mobile broadband (eMBB), and gigantic machine-type com- munication (mMTC). The requirements call for highly effi- cient, scalable, and flexible SoCs capable of driving compli- cated data-rich operations while still ensuring low energy con- sumption. Heterogeneous architecture satisfies these demands by allocating tasks to the appropriate processing units for better system performance and energy conservation.

In addition, embedding AI and ML algorithms in 5G networks further reinforces the need for heterogeneous com- puting. Optimizations of network management, traffic fore- casting, and resource allocation via AI gain leverage from the ability of GPUs to perform parallel processing and from FPGAs' and ASICs' hardware-assisted specialization. This AI- heterogeneous synergy allows

5G systems to run optimally, with capabilities to adapt to changing network traffic and user load dynamically Although it has its benefits, heterogeneous computing in SoCs for 5G communication has various chal- lenges. They involve software development complexity, parti- tioning of tasks, memory management, interconnect overhead, and thermal issues. Overcoming these challenges involves the development of programming frameworks, compiler technolo-gies, and hardware-software co-design approaches.

Overall, heterogeneous computing architecture in SoCs is critical for the successful adoption and performance tuning of 5G communication systems. With integrated specialized processing cores and the power of parallelism, these archi- tectures provide high efficiency, scalability, and flexibility and are therefore imperative for future wireless communication. Through ongoing research and technological progress, hetero- geneous computing is set to lead the way toward the future of 5G and beyond.

II. LITERATURE SURVEY

[1] This paper surveys the challenges and key technologies shaping 5G wireless communication networks, addressing the limitations of 4G, such as spectrum scarcity, high energy consumption, and increasing demand for higher data rates and mobility. The authors propose a novel 5G cellular architecture that separates indoor and outdoor environments, leveraging technologies like Massive MIMO for enhanced spectral ef- ficiency, Cognitive Radio Networks for dynamic spectrum utilization, Visible Light Communication (VLC) for high- speed short-range transmission, and Mobile Femtocells for improved connectivity in high-mobility scenarios. While these technologies offer significant advantages, the paper highlights challenges such as complex signal processing in massive MIMO, realistic channel modeling for high-speed users, and effective interference management in cognitive radio networks. The study underscores the need for an optimized balance be- tween performance, efficiency, and implementation complexity to achieve the full potential of 5G. [2] This paper surveys the key design considerations for 5G network architecture, fo- cusing on emerging communication needs such as high-speed connectivity, low latency, and massive device support. It high- lights critical challenges like spectrum scarcity, increased data traffic, and the need for enhanced reliability and robustness. The study explores enabling technologies, including ultra-dense small cell deployment, massive MIMO, network func- tion virtualization (NFV), and software-defined networking (SDN) to improve network efficiency and flexibility. A two- layer architecture is proposed, integrating radio networks with cloud-based functionalities to enhance scalability and cost- effectiveness. The paper concludes with a proof-of-concept evaluation, demonstrating the feasibility of these technologies in meeting future 5G demands [3] This paper examines the fundamental limits of wireless network densification, a key driver of increased data rates in modern communication systems. It discusses how densification—deploying more base stations and access points—has historically improved spectral efficiency and capacity, but may soon reach a plateau due to interference and physical constraints. The study highlights key challenges such as signal-to-interference-plus-noise ratio (SINR) degradation and path loss subduction, which could reduce the effectiveness of ultra-dense networks. It explores solutions like millimeter-wave technology, dynamic spectrum access, and interference management to extend the benefits of densification into future wireless networks [4] This paper pro-vides an overview of Cloud Radio Access Network (C-RAN), a centralized and virtualized mobile network architecture de- signed to enhance network efficiency and reduce operational costs. C-RAN consolidates Baseband Units (BBUs) into a cen- tralized pool, improving resource utilization, energy efficiency, and network performance. The study highlights key advan- tages, such as reduced power consumption, simplified network maintenance, and enhanced spectral efficiency through co- operative processing techniques like Coordinated Multi-Point (CoMP). However, challenges like high fronthaul bandwidth requirements and synchronization complexities must be ad- dressed for successful implementation. The paper serves as a comprehensive guide for understanding C-RAN and its role in future mobile networks [5] This paper presents the Xhaul architecture, an integrated fronthaul and backhaul solution designed to meet the stringent requirements

of 5G networks. With the expected surge in mobile data traffic and network densification, traditional transport networks will struggle to handle the increasing bandwidth demands. The study explores how high-capacity switches, heterogeneous transmission links (fiber, mmWave, and wireless optics), and software-defined networking (SDN) can create a flexible, reconfigurable, and cost-effective transport network. It highlights key challenges such as latency, jitter, and synchronization while emphasizing the role of network function virtualization (NFV) in improving efficiency. The paper also discusses the standardization efforts required to ensure multi-vendor interoperability and smooth deployment of the Xhaul architecture

III. EXISTING ARCHITECTURE

A. Cell BEA

The current trend in enhancing computing performance focuses more on parallelism rather than increasing clock fre- quency. This review examines parallelism within a single node, where instruction-level parallelism has been extensively uti- lized . Consequently, improved performance must be achieved through multi-chip, multi-core, or multi-context parallelism. Flynn's taxonomy categorizes hardware parallelism into four types:

1. Single Instruction Single Data (SISD)
2. Single Instruction Multiple Data (SIMD)
3. Multiple Instruction Single Data (MISD)
4. Multiple Instruction Multiple Data (MIMD)

Additionally, MIMD is further divided into Single Program
Multiple Data (SPMD) and Multiple Program Multiple Data (MPMD). These terms are utilized to describe the different architectures.

The single-chip Cell Broadband Engine Architecture (CBEA), as depicted in Figure 1, comprises a conventional CPU core and eight SIMD accelerator cores. This architecture is highly flexible, allowing each core to execute separate programs in an MPMD manner and communicate through a rapid on-chip bus. Its primary design goal is to maximize performance while minimizing power consumption. Figure 2 illustrates a GPU featuring 30 highly multi-threaded SIMD accelerator cores alongside a standard multi-core CPU. The
GPU excels in bandwidth and computational performance, optimized for executing SPMD programs with minimal syn- chronization. It is tailored for high-performance graphics, pri- oritizing data throughput. Lastly, Figure 3 represents an FPGA, which includes an array of logic blocks and a standard multi- core CPU. FPGAs can also integrate conventional CPU cores on the chip, making them inherently heterogeneous. They function as reconfigurable, user-defined application-specific integrated circuits (ASICs), offering deterministic performance and designed for high throughput, such as in telecommunica- tions.

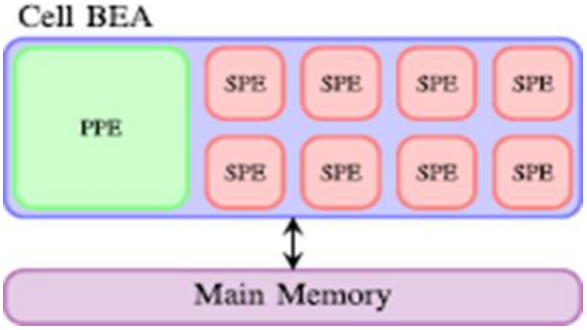

Fig. 1. CBEA

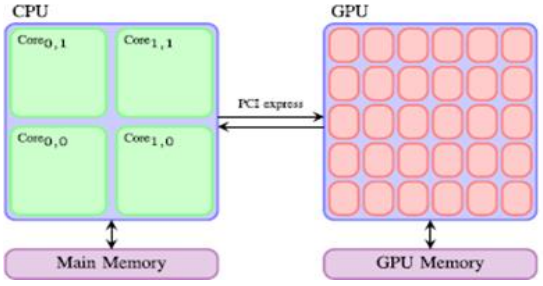

Fig. 2. CPU in computation with GPU

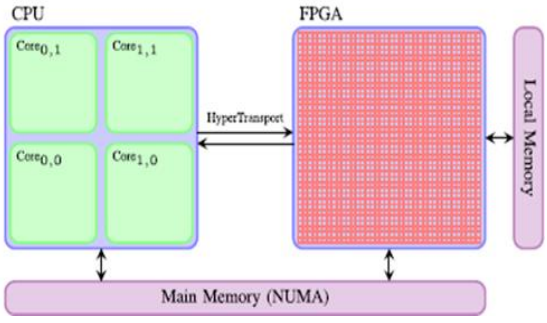

Fig. 3. CPU in computation with FPGA

The subsequent sections provide a more detailed overview of the hardware of these three heterogeneous architectures. Fi- nally, we conclude with a discussion comparing key attributes

such as the necessary level of parallelism, communication capabilities, performance, and cost.

B. Workflow in Heterogenous Computing Systems

In heterogeneous computing systems, various types of pro- cessors (e.g., CPUs, GPUs, FPGAs, and AI accelerators) are combined to execute specific tasks according to their capabilities the Fig.4 represents the workflow in detail. The workflow in these systems generally consists of various stages, each of which might need different processing resources. The following is a general workflow for heterogeneous computing systems, especially for 5G communication systems:

1. Task Decomposition - The workload is broken down into independent tasks that can be processed separately. These tasks are usually classified into computation-intensive, data-parallel tasks (e.g., signal processing), and control tasks (e.g., network management).
2. Task Assignment - After the decomposition of tasks, as- signing them to proper processing units follows. This is carried out depending on the requirement of the task, i.e., compute power, latency, and memory bandwidth. Task assignment is usually controlled by a scheduler or resource manager.
3. Task Execution - The tasks assigned are run on the selected processing unit. The tasks are executed in parallel on various processors, utilizing their specialized capabilities.
4. Inter-Processor Communication - Various processing units in heterogeneous systems must communicate with one another to share data. Efficient interconnects and memory architectures are important for reducing communication over- head.
5. Task Synchronization and Coordination - Coordination and synchronization of tasks are required to ensure that the execution flow remains consistent. Depending on the work- load, different tasks may need to be synchronized to avoid issues like race conditions or data corruption.
6. Result Collection and Output - The final output is generated by gathering the results from various processors once all the tasks are done. The output is utilized for additional actions or for updating system parameters.

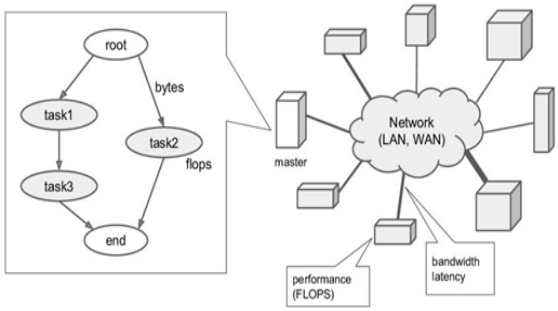

Fig. 4. Workflow in Heterogenous Computing Systems

C. Programming Models for Heterogenous Computing

Heterogeneous computing is used to process different kinds of processors or cores in a single system to enhance perfor-
mance and efficiency. Following are some popular program- ming models for heterogeneous computing:

CUDA (Compute Unified Device Architecture): CUDA is a parallel computing platform and programming model devel- oped by NVIDIA. It enables developers to program NVIDIA GPUs for general-purpose processing.

OpenCL (Open Computing Language): An open, cross- platform standard for parallel programming heterogeneous systems, OpenCL provides support for a broad variety of processors such as CPUs, GPUs, and other accelerators.

HIP (Heterogeneous-Compute Interface for Portability): AMD has created HIP as a C++ runtime API and kernel language which enables developers to write portable AMD and NVIDIA GPU code.

These models of programming make it possible for develop- ers to tap into the potential of heterogeneous computing sys- tems by offering abstractions and tools to control parallelism and optimize performance.

1) Graphics processing unit architecture: Traditionally, GPUs were tailored for rendering 2D images of 3D objects in computer games, processing each pixel's color simultaneously using a set of processors. However, recent advancements have broadened their applications beyond rendering, with GPUs now offering theoretical peak performances close to three teraflops, making them appealing for high-performance computing tasks. Despite their power, GPUs are typically connected via the PCI express bus, with second-generation PCI express×16 allowing data transfers of up to 8 GB/s between CPU

and GPU memory, though benchmarks often achieve around5.2 GB/s.

Functioning as symmetric multi-core processors, GPUs are exclusively controlled by the CPU, forming a heterogeneous system. They operate asynchronously from the CPU, enabling concurrent execution and memory transfer. Among the ma- jor GPU vendors, AMD and NVIDIA dominate the high-performance gaming market, while Intel has plans to release its high-performance gaming GPU.

Graphics Processing Units (GPUs) are central to current computing, driving innovation in gaming, artificial intelli- gence, and data processing. Market leaders in semiconductor companies—Qualcomm, Intel, and AMD—have built unique GPU architectures to support different applications. What follows is an overview of the GPU architecture from each company along with its dominant applications.

Qualcomm's Adreno GPUs are core elements of its Snap- dragon System-on-Chip (SoC) platforms, which are primarily employed in mobile devices like smartphones and tablets. The Adreno architecture is designed to provide high-performance graphics with low power consumption, important for mobile computing.

Mobile Gaming and Graphics Rendering: Adreno GPUs support immersive and smooth gaming on mobile devices with high frame rates and high-quality graphics. Augmented Reality (AR) and Virtual Reality (VR): The platform facil- itates AR and VR applications through real-time rendering computational capability and sensor support. Multimedia Pro-

cessing: Adreno GPUs are responsible for processing image processing, video playback, and encoding/decoding, delivering high-quality media experience on mobile. Architecture: The Adreno GPUs have been engineered with attention towards parallel processing abilities, making use of a unified shader architecture to efficiently manage various graphical tasks. The architecture delivers the best performance and power efficiency within the limitations of mobile devices.

Intel Xe architecture is a scalable GPU design cover- ing several microarchitectures that are designed for varied performance requirements such as integrated graphics, high- performance gaming, and data center compute.

Integrated Graphics: Xe-LP (Low Power) is used as inte- grated graphics in Intel processors, providing improved visual performance for general computing tasks. High-Performance Gaming: Xe-HPG (High Performance Graphics) is aimed at gaming and enthusiast segments, providing support for hardware-accelerated ray tracing and DirectX 12 Ultimate. Data Center and High-Performance Computing: Xe-HP and Xe-HPC (High Performance Compute) are intended for data centers, with a focus on parallel processing workloads and scientific calculations. Architecture: The Xe architecture fea- tures a modular architecture with parts like vector and ma-trix engines, unified cache hierarchies, and scalability across varying levels of performance. For example, Xe-HPG includes Xe-cores with multiple execution units for efficient parallel processing.

AMD's Radeon DNA (RDNA) architecture drives its Radeon graphics cards for gaming consoles, desktops, and data centers. RDNA is designed to offer high performance and efficiency while supporting complex graphical features.

RDNA architecture supports high frame rates and high- detail images, enabling technologies such as real-time ray tracing and variable rate shading. Professional Graphics: Employed in workstations for applications like 3D render- ing, video editing, and CAD. Compute Workloads: Parallel processing capability in RDNA recommends it for use in compute-demanding workloads such as data analysis and scientific simulations. Architecture: RDNA incorporates a re- designed graphics pipeline, better cache hierarchy, and greater execution units with higher performance and efficiency as their cumulative effect. Scalability in the architecture benefits the company as it can adapt to all classes of markets starting from gaming consoles up to ultra-high-end GPU.

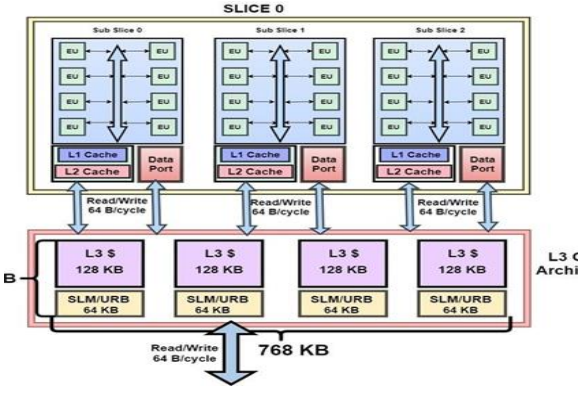

Fig. 5. Intel integrated GPU architecture

2) Digital Signal Processing: Digital Signal Processing (DSP) is a key element in the technological innovation of major semiconductor players like Qualcomm, Intel, and AMD. All companies incorporate DSP technology into their solutions to improve performance, power management, and overall sys- tem capability. Following is a summary of specific DSP imple- mentations and architectures for each company. Qualcomm's Hexagon, Intel's Movidius Myriad X and AMD's RDNA GPU Architecture. These architectures share similarities featuring highly multi- threaded cores. Qualcomm's Hexagon DSP is integral to its Snapdragon System-on-Chips (SoCs), which are widely used in smartphones, tablets, and IoT devices. The Hexagon architecture is designed to efficiently handle computationally intensive tasks, supporting advancements in wireless communication, audio processing, computer vision, and artificial intelligence (AI).

5G Communication: The Hexagon DSP is dedicated to processing baseband signals in 5G networks, performing tasks such as modulation, demodulation, error correction, and beam- forming to ensure high-performance, low-latency connectivity. Audio and Voice Processing: It enhances voice and audio qual- ity through noise cancellation, sound equalization, and voice recognition capabilities.AI and Machine Learning: With built- in AI processing capabilities, the Hexagon DSP facilitates real-time on-device AI operations like computer vision and facial recognition. Architecture: The Hexagon DSP features a Very Long Instruction Word (VLIW) architecture optimized for parallel signal processing. It includes dedicated vector and scalar processing cores for real-time tasks such as audio, video, and communication signal processing. This parallelism and low-latency operation enable efficient execution of multiple tasks while maintaining low power consumption, which is critical for mobile and edge devices.

Intel integrates DSP functionality across various product lines, notably in its Movidius Vision Processing Units (VPUs). The Movidius Myriad X VPU is optimized for computer vision

applications built using deep neural networks, offering low power consumption, low cost, and suitability for small form factors.

AI and Machine Vision: The Myriad X VPU is designed for high-throughput, low-power, real-time vision processing, making it highly suitable for edge devices and AI applications. Architecture: The Myriad X VPU is a 16 nm system-on-a- chip that integrates vision accelerators, imaging accelerators, and the Movidius Neural Compute Engine. It features 16 SHAVE vector processors paired with a CPU, providing a programmable and flexible platform for DSP operations in data centers and 5G networks.

AMD incorporates DSP-like features in its Radeon GPUs, offering robust computing resources for applications ranging from gaming and video processing to AI and telecommuni- cations. Gaming and Graphics Processing: AMD's Radeon GPUs are widely used for real-time rendering of graphics and encoding/decoding of video in gaming and multimedia. The GPUs perform tasks associated with signal processing such as scaling video, image enhancement, and motion compensation, which are typically DSP tasks.

AMD Radeon GPUs incorporate stream processors opti- mized for parallel data processing, which is best suited for applications like image processing, video encoding, and deep learning inference, which have traditionally been addressed by DSPs. Both the Vega and RDNA architectures used in Radeon GPUs incorporate integrated DSP blocks for decoding, encoding, and enhancing videos, positioning them as efficient for handling real-time communication and media processing. These DSP implementations exemplify how Qualcomm, Intel, and AMD leverage specialized architectures to address the growing demands for efficient and high-performance signal processing in modern applications.

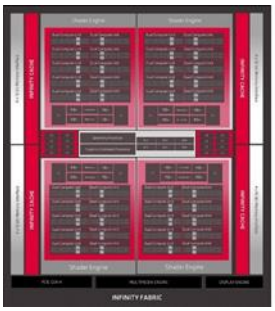

Fig. 6. Digital Signal Processing

D. Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs)

Heterogeneous computing platforms in System-on-

Chips (SoCs) are now a necessity for 5G communication, incorpo- rating Digital Signal Processors (DSPs), Graphics Processing Units (GPUs), Field-Programmable Gate Arrays (FPGAs), and Application-Specific Integrated Circuits (ASICs). These platforms facilitate efficient processing of sophisticated 5G workloads, such as baseband signal processing, AI-based network optimizations, and multimedia streaming. Through the integration of various processing units, SoCs can effec- tively balance performance, power consumption, and latency, providing smooth 5G connectivity across mobile devices, edge computing, and cloud infrastructure. Top semiconduc- tor firms—Qualcomm, Intel, and AMD—have implemented heterogeneous architectures strategically to achieve optimal efficiency in 5G systems. Qualcomm's Snapdragon SoCs use Hexagon DSPs for real-time baseband processing, Adreno GPUs for multimedia acceleration, and dedicated AI accel- erators for network optimization and power efficiency. Intel implements Xeon processors, Movidius VPUs, and Altera FPGAs in telecom infrastructure to support the deployment of 5G networks and AI-driven network automation. AMD, with its Xilinx FPGAs and adaptive computing platforms, provides programmable hardware acceleration for 5G network functions and edge computing workloads.

The future of 5G communications will be based more and more on heterogeneous architectures, with AI-powered net- work optimization, ultra-low-latency processing, and power- efficient computing becoming essential. The integration of DSPs, GPUs, FPGAs, and ASICs into SoCs will keep ad- vancing with developments in machine learning, edge com- puting, and real-time data analytics. As 5G networks become widespread, heterogeneous architectures will be at the center of providing high-speed, ultra-reliable, and power-efficient solutions for future-generation communication systems.

TABLE I-COMPARISON BETWEEN DIFFERENT COMPANIES

Features	Mediatek(Di mensity)	Qualcomm(Sna pdragon)	NXP(Layerscape)
Market Focus	Smartphones, Tablets, IOT	Smartphones, XR, Automotive	Automotive, Industrial, Networking
CPU Type	ARM Cortex (Stock)	Kryo (Customized ARM)	ARM Cortex (Stock)
GPU	Mail/ Immortalis	Adreno (Superior for gaming)	Vivante (Basic)
5G Modem Not a focus (Net- working SoCs)	Integrated 5G (Sub-6	mmWave)	Leading-edge 5G (X75/ X80)
AI Accelera tion	APU	Hexagon AI	Limited AI
Use Case Industria l, Auto- motive, Network - ing	Consumer devices	High-end consumer	enterprise
Power Efficienc y	High (TSMC 4nm)	High (TSMC 4nm)	Moderate (28nm/ 16nm)
Long- term Sup- port	3-5 years	3-5 years	10-15 years

IV. DISCUSSION

The fast development of 5G communication systems has brought forth an urgent need for efficient and high- performance computing solutions. The conventional homoge- neous architecture based only on general-purpose CPUs cannot satisfy the huge data processing demands of 5G applications like signal processing, beamforming, and channel estimation. Heterogeneous computing architectures-based SoCs provide an efficient solution using a variety of specialized processing blocks such as CPUs, GPUs, DSPs, FPGAs, and ASICs. The architectures offload tasks based on the capability of each processing block, resulting in enormous improvement in parallelism, efficiency, and powermanagement.

In heterogeneous SoCs, GPUs are superior in parallel pro- cessing, and thus best suited for AI-based applications and machine learning operations. DSPs

are best suited for real-time signal processing, necessary to maintain the dependability and low latency of 5G networks. FPGAs, though providing a re- configurable computing platform, are best suited for hardware acceleration of dynamic network optimizations. ASICs are best suited for fixed-function processing, with maximum efficiency and minimum power consumption achievable for dedicated 5G processes like error correction and baseband processing. With such specialized processors, heterogeneous computing is capable of providing greater computational throughput and power efficiency optimization for wireless networks in the future.

In the future, AI and machine learning integration into 5G network optimization will further drive the adoption of heterogeneous computing. Hardware will be required for AI accelerator-specialized, AI-based resource management, traffic prediction, and dynamic network optimization. Hybrid archi- tecture use will be facilitated by this hardware. With the 5G transitioning into 6G, heterogeneous SoCs will still be the base for ultra-low-latency, high-speed, and energy-efficient communications.

V. CONCLUSION

(SoC) heterogeneous computer platforms are central to satisfying 5G communications systems' demands for high performance and low power computation. By leveraging spe- cialized processing engines such as CPUs, GPUs, DSPs, FP- GAs, and ASICs, heterogeneous SoCs enable parallel process- ing, workload optimization, and power-conscious execution. Heterogeneous SoCs allocate work based on the ability of the individual processor compared to homogeneous systems, making it possible to process data more quickly, with reduced latency, and better scalability to mission-critical 5G applica- tions such as beamforming, signal processing, and AI-driven networkmanagement.

While these benefits exist, thermal dissipation, interconnect overhead, memory management, and software complexity pose serious challenges to the effective execution of heterogeneous architectures. Mitigating such challenges requires high-level programming frameworks, compiler optimizations, and im- proved task scheduling to be capable of utilizing the complete potential of heterogeneous architectures. AI and machine learning-based network optimizations even further highlight the necessity of in-house accelerators in SoCs to handle real- time decision-making and resource allocation.

As the 5G networks expand and evolve to 6G, hetero- geneous computing will remain at the forefront of wireless communication for the next generation. Edge computing, AI- tailored optimizations, and ultra-low-power processing will remain driving heterogeneous SoC design innovation, with high-speed, dependable, and power-efficient communications for the evolving digital world.

REFERENCES

[1] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architec- ture and key technologies for 5G wireless communication networks," Communications Magazine, IEEE, vol. 52, no. 2, pp. 122–130, 2014.

[2] P. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjeb- bour,"Design considerations for a 5G network architecture," Commu- nicationsMagazine, IEEE, vol. 52, no. 11, pp. 65 75, 2014.

[3] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, "Are We Approaching the Fundamental Limits of Wireless Network Densifica- tion?" ArXiv e-prints, Dec. 2015.

[4] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, A. Neal et al., "Mobile-Edge Computing Introductory Technical White Paper," White Paper, Mobile-edge Computing (MEC) industry initiative, 2014.

[5] Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks a technology overview," Communications Surveys & Tutorials, IEEE, vol. 17, no. 1, pp. 405–426, 2015.

[6] De la Oliva, X. Costa Perez, A. Azcorra, A. Di Giglio, F. Cav- aliere, D. Tiegelbekkers, J. Lessmann, T. Haustein, A. Mourad, and P. Iovanna, "Xhaul: toward an integrated fronthaul/backhaul architecture in 5g networks," Wireless Communications, IEEE, vol. 22, no. 5, pp. 32–40,