Literature Survey on GPU Chip Design

Introduction

Graphics Processing Units (GPUs) have evolved from specialized graphics accelerators to versatile parallel processors critical for a wide range of applications, including scientific computing, deep learning, and high-performance computing (HPC). The design of GPU chips encompasses numerous challenges such as managing massive parallelism, optimizing memory architectures, ensuring energy efficiency, and addressing thermal constraints. Additionally, emerging workloads like graph neural networks (GNNs), large language models (LLMs), and diffusion-based AI models impose new demands on GPU architecture, further complicating design considerations. This literature survey synthesizes recent advances in GPU chip design, highlighting innovations in architectural scalability, memory systems, interconnect technologies, energy efficiency, and thermal management. It also explores how domain-specific requirements and heterogeneous integration influence modern GPU architectures.

Architectural Innovations and Scalability

Modern GPU chip design increasingly focuses on architectural flexibility and scalability to meet diverse workload demands. Lee et al. (2005) proposed a scalable GPU architecture leveraging dynamically reconfigurable embedded processors, enabling on-the-fly adaptation of processing units to optimize resource utilization across varying workloads. This approach enhances both scalability and efficiency, providing a foundation for flexible GPU designs.

Building upon this, Fu et al. (2021) introduced the Composable On-Package GPU (COPA-GPU) architecture, which disaggregates GPU components into multi-chip modules specialized for different domains such as high-precision HPC and low-precision deep learning (DL). This modular design enables tailored augmentation of GPU capabilities, including increased off-die bandwidth and larger on-package cache, resulting in over 30% performance improvements in DL tasks and reducing the number of GPUs needed for scale-out training by half. Such domain specialization addresses the divergent requirements of modern GPU workloads within a unified chip design framework.

To support large-scale GPU systems, Li et al. (2024) proposed GROOT, a silicon photonic chiplet-based network that partitions GPU systems into groups connected via optical interconnects. This design overcomes transistor count limitations per die and improves chiplet communication efficiency, achieving a 48% performance boost and 24.5% energy savings. Similarly, Bashir et al. (2019) developed a power-efficient photonic network-on-chip (NoC) tailored for GPUs, employing clusterbased optical communication and traffic separation to reduce latency and energy consumption significantly. These photonic interconnect innovations represent promising directions for scalable and energy-efficient GPU chip design.

Zhang et al. (2023) studied multi-chip GPU data sharing, identifying the bandwidth gap between inter-chip and intra-chip communication as a critical bottleneck. Their characterization of shared datasets and sensitivity to workload properties informs GPU chip design choices such as cache placement and thread scheduling, which are vital for optimizing multi-chip GPU systems. Pitliya and Palecha (2020) further contributed a shared buffer crossbar architecture for CPU-GPU on-chip networks, addressing many-to-few traffic patterns in GPUs to reduce area by 28% and power by 32% compared to traditional mesh networks. Together, these works underscore the importance of scalable and efficient interconnect architectures in modern GPU design.

Memory Architecture and Energy Efficiency

Memory architecture remains a central focus in GPU chip design due to its impact on performance and power consumption. Dai et al. (2017) emphasized the need for a sound baseline in GPU memory architecture research to standardize evaluation and foster meaningful comparisons. Building on this, Zhang et al. (2023) introduced G10, a unified GPU memory and storage architecture integrating host memory, GPU memory, and flash storage into a single address space. G10 leverages predictable tensor behaviors in deep learning to schedule proactive data migrations, achieving up to 1.75× performance improvement without code modification. This approach simplifies memory management and addresses the GPU memory wall challenge.

Exploring emerging memory technologies, Shua (2013) evaluated Phase Change Memory (PCM) integration in GPU memory, proposing architectural optimizations to mitigate latency and endurance issues. Wang et al. (2013) investigated hybrid memory architectures combining different memory technologies through software-hardware co-design, demonstrating significant energy savings while maintaining performance. These studies highlight the potential of novel memory technologies and management strategies to enhance GPU energy efficiency.

Compiler and software techniques also play a crucial role in optimizing memory utilization. Voitsechov et al. (2018) proposed software-directed methods to improve register file utilization in GPUs, increasing thread occupancy by 23% and yielding a 12% performance gain on register-bound applications. Gupta et al. (2017) developed compiler techniques to reduce synchronization overhead in GPU redundant multithreading, enhancing execution efficiency. Additionally, Lee et al. (2015) introduced the GREEN Cache architecture, which exploits the disciplined memory model of OpenCL to improve cache efficiency and power consumption in GPUs.

Marangoz et al. (2021) addressed memory bandwidth management by designing a GPU memory architecture extension for bandwidth reservation, enabling concurrent execution of multiple applications with predictable performance and reducing interference. This approach improved overall throughput by 20%, illustrating the importance of flexible memory resource allocation in multi-tenant GPU environments.

Thermal and Power Management

Thermal constraints and power efficiency are critical challenges in GPU chip design, particularly as transistor densities increase and workloads intensify. Wang and Chen (2023) proposed low-power schemes for large GPU chips using Unified Power Format (UPF) to partition power domains and integrate Dynamic Voltage Frequency Scaling (DVFS), achieving an 8.19% power reduction. Sharma and Al-Asaad (2021) reviewed low-power GPU techniques, emphasizing architectural optimizations, power gating, and workload-aware scheduling to balance performance and energy consumption.

Thermal management strategies have been extensively studied to maintain GPU reliability and performance. Sheaffer et al. (2005) analyzed heat generation in GPU architectures and proposed cooling optimizations, while Maity et al. (2022) developed a model predictive control (MPC) based scheduling for CPU-GPU embedded platforms to minimize peak temperatures via task mapping and frequency tuning. Chung et al. (2025) focused on thermal management in heterogeneously integrated HBM-GPU modules, demonstrating that combined cooling techniques can reduce maximum chip temperature by over 22%, improving operational stability.

Lu et al. (2025) contributed a real-time thermal map characterization method for commercial GPUs under AI workloads, revealing hotspot distributions and proposing GPUThermalMap, a neural network-based thermal estimation tool with high accuracy and low latency. Prakash et al. (2016) highlighted cooperative CPU-GPU thermal management to enhance mobile gaming performance by preventing thermal throttling through coordinated control.

Voltage scaling techniques also enhance energy efficiency. Leng et al. (2014) showed that reducing voltage guardbands in the Kepler GPU architecture yields significant power savings without performance loss. Possignolo et al. (2018) introduced GPU stacking based on voltage stacking to compensate for process variation in near-threshold computing, achieving a 37% performance improvement and better power delivery.

Specialized GPU Architectures and Emerging Workloads

The rise of AI and specialized workloads has driven innovations in GPU chip design to meet unique computational demands. Xie et al. (2023) developed Accel-GCN, a GPU accelerator architecture optimized for Graph Convolutional Networks (GCNs). By addressing workload imbalance, memory access irregularities, and metadata overhead, Accel-GCN achieves significant speedups over existing GPU implementations through novel block-level partitioning and combined warp strategies.

Jing et al. (2024) presented AIG-CIM, a tri-gear heterogeneous compute-in-memory chiplet module designed for diffusion model acceleration in AI-generated content. This architecture attains remarkable latency and energy efficiency improvements compared to traditional GPUs, indicating the potential of compute-in-memory techniques for specialized AI workloads.

Wang et al. (2025) proposed LEAP, integrating processing-in-memory with network-on-chip architectures to accelerate large language model inference. LEAP dynamically balances dataflow and parallelism, achieving over 2.5× throughput and nearly 72× energy efficiency improvements compared to NVIDIA A100 GPUs.

Joardar et al. (2020) introduced AccuReD, a heterogeneous 3-D architecture combining ReRAM arrays with GPU cores for high-accuracy CNN training, mitigating ReRAM noise through GPU integration and thermal-aware mapping. This design yields up to 12× acceleration without accuracy loss.

Hübner et al. (2025) evaluated Apple Silicon M-Series SoCs, highlighting unified memory architecture and energy-efficient GPU designs that provide competitive HPC performance despite limited double-precision support. Luo et al. (2024) benchmarked NVIDIA Hopper GPUs, revealing novel tensor core features and instruction sets tailored for AI workloads, offering insights into future GPU design directions.

Conclusion

GPU chip design has undergone significant advancements driven by the need for scalability, energy efficiency, thermal management, and specialization for emerging workloads. Architectural innovations such as dynamically reconfigurable processors, composable multi-chip modules, and silicon photonic interconnects have enhanced scalability and performance. Memory architecture research emphasizes unified memory spaces, hybrid technologies, and compiler optimizations to overcome bandwidth and capacity limitations while improving energy efficiency. Thermal and power management strategies, including dynamic voltage scaling, multi-domain power management, and sophisticated cooling techniques, are critical to maintaining reliability and performance under increasing computational loads.

Furthermore, specialized GPU architectures tailored for AI workloads, such as GCNs, diffusion models, and LLMs, demonstrate the importance of heterogeneous integration and compute-in-memory approaches. Comprehensive benchmarking and simulation frameworks continue to inform design decisions, while emerging technologies like ReRAM and photonic networks promise further gains.

Overall, the surveyed literature reflects a vibrant and multifaceted research landscape in GPU chip design, balancing general-purpose capabilities with domain-specific enhancements to meet the evolving demands of modern computing.

References

Bashir, Janibul and Sethi, Khushal and Sarangi, S. (2019). Power Efficient Photonic Network-On-Chip For A Scalable Gpu. *ACM/IEEE International Symposium on Networks-on-Chips*.

Chung, Euichul and Manley, Madison and Zeng, Wanshu and Bakir, Muhannad (2025). Thermal Management Of Heterogeneously Integrated Hbm-Gpu Module With Step Height Difference. *Electronic Components and Technology Conference*.

Dai, Hongwen and Li, C. and Lin, Zhen (2017). The Demand For A Sound Baseline In Gpu Memory Architecture Research.

Fu, Yaosheng and Bolotin, Evgeny and Chatterjee, Niladrish and Nellans, D. and Keckler, S. (2021). Gpu Domain Specialization Via Composable On-Package Architecture. *ACM Transactions on Architecture and Code Optimization (TACO)*.

Gupta, Manish and Lowell, Daniel and Kalamatianos, J. and Raasch, Steven and Sridharan, Vilas and Tullsen, D. and Gupta, Rajesh (2017). Compiler Techniques To Reduce The Synchronization Overhead Of Gpu Redundant Multithreading. *Design Automation Conference*.

Hübner, Paul and Hu, Andong and Peng, Ivy and Markidis, Stefano (2025). Apple Vs. Oranges: Evaluating The Apple Silicon M-Series Socs For Hpc Performance And Efficiency. *IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*.

Jing, Yiqi and Wu, Meng and Zhou, Jiaqi and Sun, Yiyang and Ma, Yufei and Huang, Ru and Jia, Tianyu and Ye, Le (2024). Aig-Cim: A Scalable Chiplet Module With Tri-Gear Heterogeneous Compute-In-Memory For Diffusion Acceleration. *Design Automation Conference*.

Joardar, B. and Doppa, J. and Pande, P. and Li, H. and Chakrabarty, K. (2020). Accured: High Accuracy Training Of Cnns On Reram/Gpu Heterogeneous 3-D Architecture. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

Lee, Won-Jong and Woo, Sangoak and Kwon, Kwon-Taek and Son, Sungjin and Min, Kyoungha and Jang, Gyeong-Ja and Lee, Choong-Hun and Jung, Seokyoon and Park, Chan-Min and Lee, Shihwa (2005). A Scalable Gpu Architecture Based On Dynamically Reconfigurable Embedded Processor.

Lee, Jaekyu and Woo, Dong and Kim, Hyesoon and Azimi, M. (2015). Green Cache: Exploiting The Disciplined Memory Model Of Opencl On Gpus. *IEEE transactions on computers*.

Leng, Jingwen and Zu, Yazhou and Reddi, V. (2014). Energy Efficiency Benefits Of Reducing The Voltage Guardband On The Kepler Gpu Architecture.

Li, Chengeng and Jiang, Fan and Chen, Shixi and Li, Xianbin and Liu, Jiaqi and Zhang, Wei and Xu, Jiang (2024). Towards Scalable Gpu System With Silicon Photonic Chiplet. *Design, Automation and Test in Europe*.

Lu, Jincong and Sachdeva, Sachin and Lin, Yuxuan and Tan, S. (2025). Real-Time Thermal Map Characterization And Analysis For Commercial Gpus With Ai Workloads. *IEEE International Symposium on Quality Electronic Design*.

Luo, Weile and Fan, Ruibo and Li, Zeyu and Du, Dayou and Wang, Qiang and Chu, Xiaowen (2024). Benchmarking And Dissecting The Nvidia Hopper Gpu Architecture. *IEEE International Parallel and Distributed Processing Symposium*.

Maity, Srijeeta and Roy, Rudrajyoti and Majumder, A. and Dey, Soumyajit and Hota, A. (2022). Future Aware Dynamic Thermal Management In Cpu-Gpu Embedded Platforms. *IEEE Real-Time Systems Symposium*.

Marangoz, Emir and Kang, K. and Shin, Seunghee (2021). Designing Gpu Architecture For Memory Bandwidth Reservation. *IEEE International Symposium on Performance Analysis of Systems and Software*.

Pitliya, Deepika and Palecha, Namita (2020). Design Of Shared Buffer Architecture For Cpu-Gpu On Chip Network. *INTERNATIONAL JOURNAL OF ELECTRICAL ENGINEERING & TECHNOLOGY*.

Possignolo, Rafael and Ebrahimi, E. and Ardestani, E. and Sankaranarayanan, Alamelu and Briz, J. and Renau, Jose (2018). Gpu Ntc Process Variation Compensation With Voltage Stacking.

Prakash, A. and Amrouch, Hussam and Shafique, M. and Mitra, T. and Henkel, J. (2016). Improving Mobile Gaming Performance Through Cooperative Cpu-Gpu Thermal Management. *Design Automation Conference*.

Sharma, P. and Al-Asaad, H. (2021). Brief Review Of Low-Power Gpu Techniques.

Sheaffer, J. and Skadron, K. and Luebke, D. (2005). Studying Thermal Management For Graphics-Processor Architectures.

Shua, Mu (2013). Evaluating And Optimizing Of Pcm Based Gpu Memory Architecture.

Voitsechov, Dani and Zulfiqar, A. and Stephenson, M. and Gebhart, Mark and Keckler, S. (2018). Software-Directed Techniques For Improved Gpu Register File Utilization. *ACM Transactions on Architecture and Code Optimization (TACO)*.

Wang, Wenjie and Chen, Qianli (2023). Research On Low-Power Schemes Based On Large Gpu Chip.

Wang, Bin and Wu, Bo and Li, Dong and Shen, Xipeng and Yu, Weikuan and Jiao, Yizheng and Vetter, J. (2013). Exploring Hybrid Memory For Gpu Energy Efficiency Through Software-Hardware Co-Design.

Wang, Bin and Wu, Bo and Li, Dong and Shen, Xipeng and Yu, Weikuan and Jiao, Yizheng and Vetter, Jeffrey (2013). Exploring Hybrid Memory For Gpu Energy Efficiency Through Software-Hardware Co-Design. *International Conference on Parallel Architectures and Compilation Techniques*.

Wang, Yimin and Chong, Yue and Fong, Xuanyao (2025). Leap: Llm Inference On Scalable Pim-Noc Architecture With Balanced Dataflow And Fine-Grained Parallelism.

Xie, Xi and Peng, Hongwu and Hasan, Amit and Huang, Shaoyi and Zhao, Jiahui and Fang, Haowen and Zhang, Wei and Geng, Tong and Khan, O. and Ding, Caiwen (2023). Accel-Gcn: High-Performance Gpu Accelerator Design For Graph Convolution Networks.

Zhang, Shiqing and Naderan-Tahan, Mahmood and Jahre, Magnus and Eeckhout, L. (2023). Characterizing Multi-Chip Gpu Data Sharing. *ACM Transactions on Architecture and Code Optimization (TACO)*.

Zhang, Haoyang and Zhou, Y. and Xue, Yu and Liu, Yiqi and Huang, Jian (2023). G10: Enabling An Efficient Unified Gpu Memory And Storage Architecture With Smart Tensor Migrations. *Micro*.

Zhang, Huaipeng and Ho, Nhut-Minh and Polat, Dogukan and Chen, Peng and Wahib, M. and Nguyen, Truong and Meng, Jintao and Goh, R. and Matsuoka, S. and Luo, Tao and Wong, W. (2023). Simeuro: A Hybrid Cpu-Gpu Parallel Simulator For Neuromorphic Computing Chips. *IEEE Transactions on Parallel and Distributed Systems*.