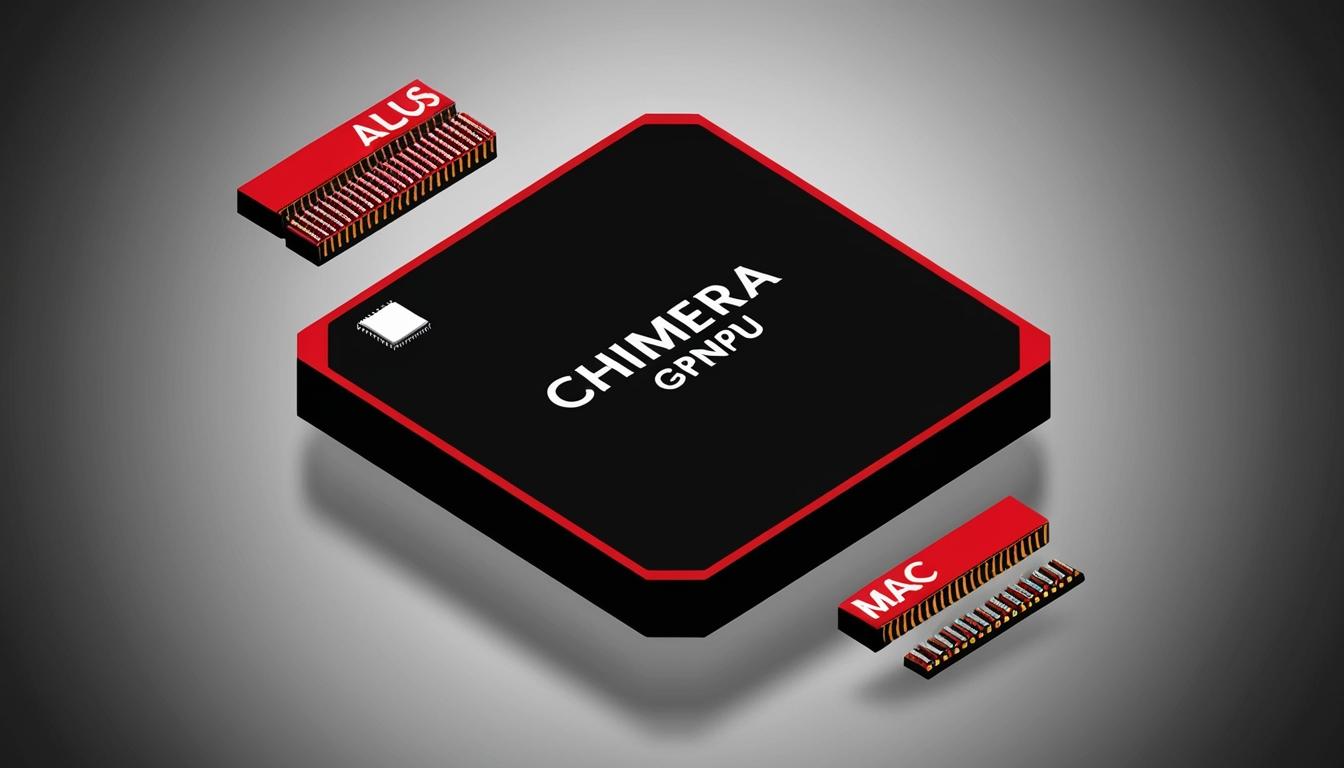
# The future of AI chip architecture: Rethinking the MAC model



In recent years, the integration of artificial intelligence (AI) and machine learning (ML) into business practices has accelerated, with companies increasingly turning to advanced chip design for on-device inference capabilities. For almost five years, chip designers have gravitated towards a common architecture comprising a “multiply-accumulate” (MAC) accelerator, supplemented by a legacy programmable core. However, not all stakeholders agree this model is sufficient for the complexities of modern networks.

Quadric, a technology company, has been vocal in its criticism of this traditional approach, positing that the MAC-centric architectures fail to meet the demands of contemporary networks that include a wide range of operators beyond simple MAC functions. The company argues that while the partitioned architecture may have sufficed for earlier convolutional neural networks (CNNs) such as Resnet-50, the more complex structures seen today require a broader range of functional units. "Hey, you’re not looking at the new wave of networks. MACs are not enough! It’s the fully functional ALUs that matter!" indicated a representative from Quadric at various forums.

Quadric’s perspective gained traction last November at the Automotive Compute Conference held in Munich, where Qualcomm presented findings that echoed some of Quadric’s assertions. Qualcomm’s keynote highlighted an analysis of over 1,200 AI and ML networks, revealing a trend that diverges from the traditional MAC-centric models. The Y-axis of Qualcomm's presentation diagram illustrated a breakdown of performance demands across different networks, with only 50% of them being MAC-dominated, while many featured minimal to no traditional MAC layers.

The Qualcomm analysis does, however, raise critical considerations regarding performance in multi-core systems. While Qualcomm advocates for a tri-core heterogeneous solution—pairing MAC engines with digital signal processors (DSPs)—they did not account for the potential latencies in data transfer between cores as the computing needs fluctuate. Moreover, the disparity in processing power between the matrix accelerators and programmable DSPs is striking; typical AI devices boast 40 trillion operations per second (TOPs) from their MAC accelerators, compared to the much slower DSPs that might manage 16 parallel operations at once. According to Quadric, the resulting bottleneck during operations that alternate between MAC and DSP tasks could hinder performance.

An alternative presented by Quadric is the Chimera GPNPU processor, designed to streamline the computational process by integrating a full-function 32-bit ALU with clusters of MACs. This design boasts the ability to house up to 1,024 ALUs within a single core, optimising throughput and enabling both MAC and ALU tasks to execute efficiently. This proposition suggests that businesses using Quadric’s architecture can expect a seamless experience in processing AI/ML tasks across diverse and complex network types.

For those interested in evaluating machine learning processing solutions, Quadric has made their DevStudio tool publicly accessible. This platform offers extensive performance data across various AI benchmark models, detailing source models, compilation results, and exhaustive simulation data to facilitate comparison across different scenarios and memory configurations.

As businesses navigate the evolving landscape of AI automation, understanding these emerging technologies and their implications on operational efficiency will be crucial. The shift in focus from traditional MAC-centric processing to more integrated systems could significantly impact how organisations implement AI-driven solutions in the future, thus representing a potential turning point in the industry’s trajectory.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

* <https://www.mckinsey.com/industries/semiconductors/our-insights/scaling-ai-in-the-sector-that-enables-it-lessons-for-semiconductor-device-makers> - Corroborates the use of AI/ML in semiconductor manufacturing, including cost reductions and efficiency improvements in chip design and manufacturing.
* <https://www.acldigital.com/blogs/how-ai-transforming-semiconductor-industry-2024-and-beyond> - Supports the role of AI in chip design, verification, and testing, highlighting advancements and challenges in AI-driven semiconductor design.
* <https://www.acldigital.com/blogs/how-ai-transforming-semiconductor-industry-2024-and-beyond> - Details AI-powered advancements in chip design, including the use of neural networks and genetic algorithms, and the challenges such as data quality and tool integration.
* <https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2023/ai-in-chip-design.html> - Discusses the use of AI in chip design for improving fault detection, managing supply chains, and addressing complex design challenges.
* <https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2023/ai-in-chip-design.html> - Highlights the potential of AI in developing new services and co-designing vertical-specific chips, aligning with the trend of moving beyond traditional MAC-centric models.
* <https://www.mckinsey.com/industries/semiconductors/our-insights/scaling-ai-in-the-sector-that-enables-it-lessons-for-semiconductor-device-makers> - Provides insights into AI/ML use cases in research and chip design, including automation of physical-layout design and verification processes.
* <https://www.acldigital.com/blogs/how-ai-transforming-semiconductor-industry-2024-and-beyond> - Explains the integration of AI in semiconductor manufacturing, including predictive maintenance, yield optimization, and defect detection.
* <https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2023/ai-in-chip-design.html> - Supports the idea that AI can address supply chain challenges and improve fault detection in semiconductor manufacturing.
* <https://www.acldigital.com/blogs/how-ai-transforming-semiconductor-industry-2024-and-beyond> - Discusses the challenges in AI-driven chip design, such as data quality, complex algorithms, and interpretability issues.
* <https://www.mckinsey.com/industries/semiconductors/our-insights/scaling-ai-in-the-sector-that-enables-it-lessons-for-semiconductor-device-makers> - Highlights the potential of AI/ML in reducing R&D costs and improving efficiency in the semiconductor industry.
* <https://www.acldigital.com/blogs/how-ai-transforming-semiconductor-industry-2024-and-beyond> - Describes the future trends in AI and semiconductor chip design, emphasizing innovation, efficiency, and adaptability.
* <https://semiengineering.com/macs-are-not-enough-why-offload-fails/> - Please view link - unable to able to access data