# Demystifying GPU architecture for AI processing

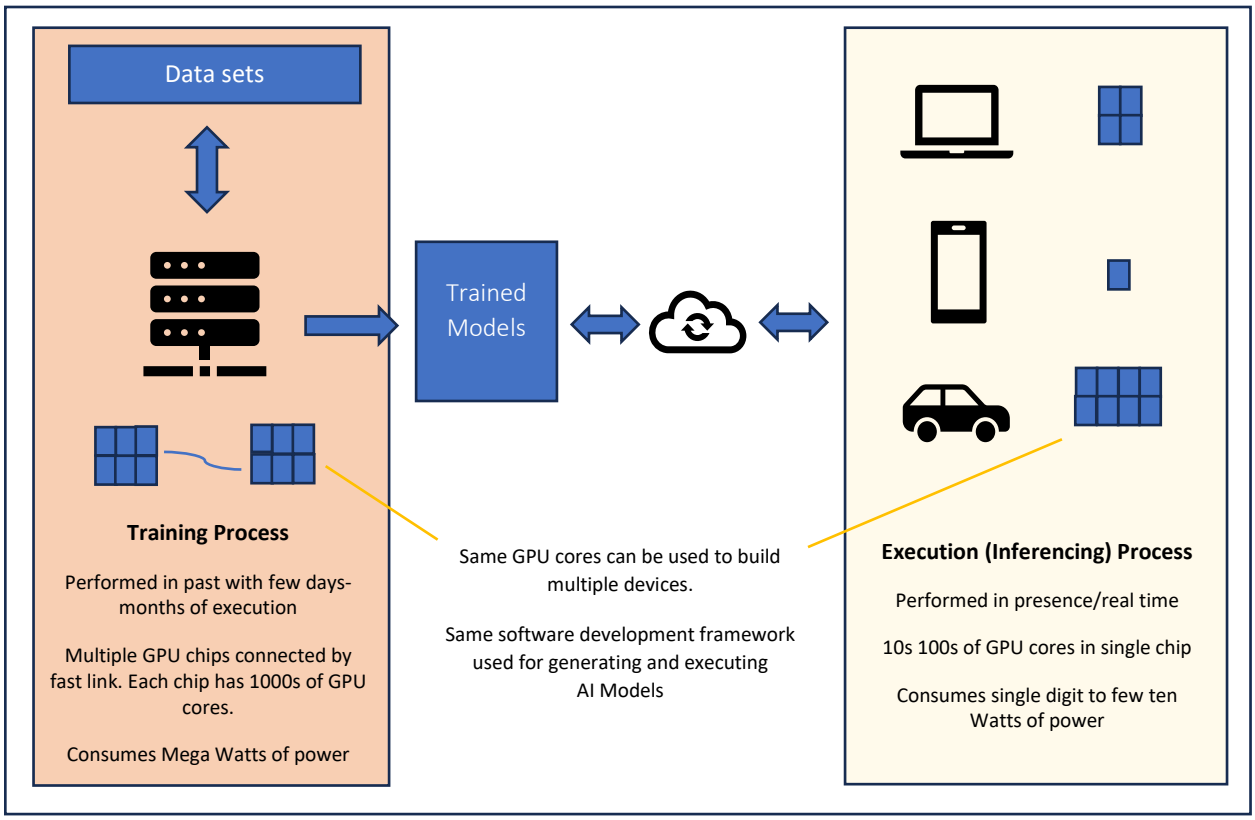
**Blog 5 : GPUs scalability across various platforms and conclusions**

In previous blogs, we discussed how the GPU's parallel processing architecture, programming model, and developer ecosystem provide an ideal foundation for AI processing. In this blog, we'll explore additional system-level aspects and benefits of GPUs. We'll conclude the series with a summary.

**GPU offers Modularity and Scalability across various range of devices and platforms.**

AI use cases, like our brains, alternate between learning and performing phases. Massive AI accelerators in data centres, which can consume megawatts of power, are used for training AI models—a process that can take days to months. Once trained, these models are deployed for inferencing—performing specific tasks based on user interactions, such as using ChatGPT for editing or content generation on PCs. The required processing power varies by application and processor size. For example, cell phones use less than 5 watts, while AI PCs consume 10-30 watts.

Building a brand-new architecture for every device is very costly. A single, scalable processor architecture that can adapt to various needs is highly desirable for cost efficiency and programming model compatibility. The GPU architecture meets these requirements effectively. Today, GPUs are used as accelerators in everything from cell phones to data servers, naturally aligning with the scalability needs of AI processing.




**GPU scales across multiple platforms**

**Conclusion and forward-looking possibilities:**

In conclusion, AI's evolution in data models, algorithms, and use cases has created new computational demands that challenge traditional CISC and RISC architectures. As scalar processors, CISC and RISC architectures do not scale well for AI processing needs. GPUs, with their parallel processing capabilities (vector processing, SIMD architecture), scalability, compatibility with existing programming models, and widespread ecosystem adoption, stand out as a promising solution. Their architecture allows for rapid adaptation to new AI requirements, significantly contributing to their current popularity.

However, will GPUs be the sole solution for AI processing?

Currently, the trend is toward hybrid processors that use RISC/CISC processors for user interfaces and offload AI workloads to GPUs. GPUs excel in scaling for both training and inferencing, offering high adaptability across various devices and programming models. NPUs (which we

haven't discussed in these blogs) are also emerging as effective options for low-power AI inferencing. There is potential to build custom processors tailored to optimally handle proprietary algorithms. The industry will continue to see a balance between custom and generic processors, each with its own advantages and drawbacks. Both will coexist and complement each other in the market. As AI continues to develop, it will be interesting to see how future emerging needs will challenge hardware developers. We are excited and ready for this challenge!!!!