Samir Rahman

Title: Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture

URL: https://doi.org/10.1145/3352460.3358302

The paper *"Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture"* explores how deep learning inference can be efficiently scaled using a multi-chip-module (MCM)-based system composed of smaller chiplets rather than relying on large, monolithic chips. The specific behavior the researchers study is the execution of deep neural network (DNN) inference on such MCM architectures, focusing on how to optimize performance, energy efficiency, and scalability when compute and memory resources are distributed across many chiplets. The paper frames its central research question as: *"This work investigates and quantifies the costs and benefits of using MCMs with fine-grained chiplets for deep learning inference, an application area with large compute and on-chip storage requirements."* The challenges addressed include managing non-uniform latency and bandwidth between on-chip and inter-chiplet communication, reducing tail latency (where the slowest chiplet limits performance), and mapping DNN workloads efficiently across the system. The main problem tackled is the performance bottleneck that arises from these non-uniformities in communication and computation when scaling inference across many chiplets. The authors claim that Simba, their 36-chiplet prototype, successfully achieves up to 128 TOPS (throughput) and 6.1 TOPS/W (energy-efficiency), offering low-latency inference suitable for data center workloads. To support these claims, they provide evidence from silicon-based evaluations using workloads like ResNet-50 and DriveNet. Optimizations such as non-uniform work partitioning, communication-aware data placement, and cross-layer pipelining yield up to 16% performance gains and 2.3× throughput improvement. Statistical analysis is embedded in detailed experimental results, showing normalized latency, energy usage, and throughput across layers, chiplet configurations, and communication parameters. Overall, the paper demonstrates that with thoughtful architectural and mapping strategies, MCM-based systems can scale deep learning inference effectively while overcoming the inherent limitations of inter-chiplet communication.