Other - Colloquium on Artificial Intelligence Research and Optimization

The Groq Tensor Streaming Processor (TSP) and the Value of Deterministic Instruction Execution

Utham Kamath, Groq

Director of Machine Learning Systems

Virtual- REGISTRATION REQUIRED (SEE ABSTRACT) Zoom
May 05, 2021 - 01:00 pm

Abstract:

The explosion of machine learning and its many applications has motivated a variety of new domain-specific architectures to accelerate these deep learning workloads. The Groq Tensor Streaming Processor (TSP) is a functionally-sliced microarchitecture with memory units interleaved with vector and matrix functional units. This architecture takes advantage of dataflow locality of deep learning operations. The TSP is built based on two key observations: (1) machine learning workloads exhibit abundant data parallelism, which can be readily mapped to tensors in hardware, and (2) a deterministic processor with a stream programming model enables precise reasoning and control of hardware components to achieve good performance and power efficiency. The TSP is designed to exploit parallelism inherent in machine-learning workloads including instruction-level parallelism, memory concurrency, data and model parallelism. It guarantees determinism by eliminating all reactive elements in the hardware, for example, arbiters and caches. The instruction ordering is entirely software controlled and the underlying hardware cannot reorder these events and they must complete in a fixed amount of time. This has several consequences for system design: zero variance latency, low latency, high throughput at batch size 1 and reduced total cost of ownership (TCO) for data centers with diverse service level agreements (SLAs). Early ResNet50 image classification results demonstrate 20.4K processed images per second with a batch size of one. This is a 4X improvement compared to other modern GPUs and accelerators. The first ASIC implementation of the TSP architecture yields a computational density of more than 1 TOp/s per square mm of silicon. The TSP is a 25x29mm 14nm chip operating at a nominal clock frequency of 900MHz. In this talk we discuss the TSP and the design implications of its architecture. The talk will cover our work published at ISCA 2020: https://groq.com/isca-2020-conference/

REGISTRATION IS FREE AND REQUIRED TO RECIEVE ZOOM ID:
https://docs.google.com/forms/d/e/1FAIpQLSdLcIBvX8tiTkxm9HbvBdyPx70fhpKpzsAQmCf0AXgn44cqcQ/viewform

Speaker's Bio:

Utham Kamath is Director of Machine Learning Systems at Groq where he works on the implementation and optimization of ML models for Groq's hardware and performance analysis of ML workloads. He has twenty years of industry experience including technical and management roles at Qualcomm, Atheros Communications and Hewlett Packard. He has a Bachelor's degree in Engineering from Bangalore University and an MS and PhD from the University of Southern California.

Center for Computation & Technology
2003 Digital Media Center • Telephone: +1 225/578-5890 • Fax: +1 225/578-8957