The Challenge of Batch Size 1: Groq Adds Responsiveness to Inference Performance

AUTHOR
Groq, Inc.

 400 Castro St., Suite 600, Mountain View, CA 94041 www.groq.com

Running a batch size of one, which refers to computations on a single image or sample during inference processing, is a valuable option for many machine learning workflows – particularly those that require real-time responsiveness. Near instantaneous inference is essential for safety-critical applications and creating new customer experiences from visual object detection, speech recognition or other forms of data input. For instance, the ability to auto-fill suggestions as a retail customer speaks or types product names in an e-commerce site is one example of where batch size 1 processing is required for optimal responsiveness. In an autonomous vehicle, the navigation system may need to react within milliseconds to avoid hitting an obstacle, so having a processing architecture where minimum response times are ensured each and every time is essential for safety. However, small batch sizes and batch size 1 introduce a number of performance and responsiveness complexities to machine learning applications, particularly with conventional inference platforms based on GPUs.

GPU architecture and batch size limitations

GPUs were historically designed for massively parallel processing, and built on multi-data, or multi-task, fixed-structure processing engines. These platforms were originally created for real-time 3-D graphics rendering for gaming applications. GPUs render graphics based on a set of parallelized processes that take place in hundreds or thousands of cores concurrently. The parallelism is dependent on a pipeline architecture with each processing stage adding new data to ongoing parallel calculations. Due to this pipeline architecture, GPUs can perform billions of graphics-related geometry calculations per second.

When used as a processing unit in machine learning applications, GPUs experience significant declines in performance when fed data in small batch sizes, because gaps in the data flow introduce stalls in GPU execution. This latency can introduce a dramatic impact on real-time inference performance at batch size 1, with associated impacts on TCO (total cost of ownership) for machine learning platform investments. If an application relies on batch size 1 processing for real-time responsiveness – for instance, for applications where fast reaction time is paramount – GPU-based platforms may not deliver inference performance fast enough. Responsiveness here is defined as the ability to service a workload in minimum execution time (latency).

Groq offers an alternative platform for machine learning that is not dependent on the performance limitations of GPU parallelism and long latency, offering machine learning engineers a superior inference platform with no performance penalty for small batch sizes – or workloads of any size. Groq's new tensor streaming processor (TSP) architecture is designed specifically for the performance requirements of machine learning and other computationally-intensive applications, demanding both low latency and high throughput.

A machine learning platform with peak performance, even at batch size 1

Significantly, the Groq architecture is hyper-focused on low latency, single-thread performance at batch size 1. This architecture, implemented in our Tensor Streaming Processor (TSP), is a single-threaded, single-core architecture that delivers maximum performance at any batch size. Groq's processor is 17.6 times faster than GPU-based platforms at batch size 1 and nearly 2.5 times faster at large batch sizes.

The Groq compiler schedules program execution on the TSP and provides a new paradigm for achieving both flexibility and massive parallelism without the limitations and communication overheads of conventional GPU architectures. The Groq compiler orchestrates streaming operands to the operators executing on the chip. Because it understands the hardware and the speed of each instruction, the compiler can tell the hardware efficiency exactly what to do, and when. This allows deterministic, repeatable and predictable performance. The flow of instructions to the hardware is completely orchestrated by the compiler, making processing efficient and predictable. This predictability of performance is extremely valuable, especially for safety-critical applications where real-time responsiveness is paramount.

Groq delivers the industry's highest level of compute processor responsiveness without compromising performance

Inference platform agility requires responsiveness *and* performance

The difficulty of achieving inference agility – the combination of responsiveness and performance – becomes more evident when we compare performance metrics from published benchmarks for a number of leading machine learning platforms.

Figure 1 on the following page illustrates metrics for performance of image recognition for ResNet-50 v1.5 benchmarks for leading inference platforms while processing small batch size workloads. The graph shows the amount of time required for the inference platform to process small batch size image data, with the X-axis indicating images per second per chip while the Y-axis denotes latency in milliseconds. These GPU-based inferencing platforms all trend down and to the right as batch sizes increase, indicating greater latency and reduced responsiveness. The leading GPU-based inference solution, NVIDIA V100, at batch size 1 can process just over 1000 images per second, with latency of 1 millisecond. At batch size 1 and running ResNet-50 v2, the Groq TSP100 chip is capable of processing 18,900 images per second, with latency less than 0.1 milliseconds. This level of performance is possible because with Groq, it is unnecessary to run larger batch sizes to achieve high throughput. At small batch sizes, the Groq chip demonstrates 3x to 20x performance superiority over the GPU-based platforms running ResNet-50 v1.5 benchmarks.

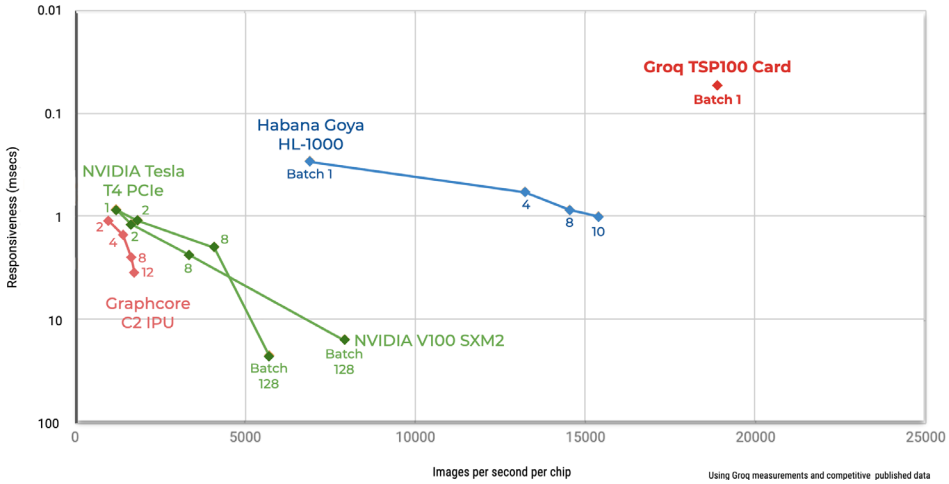

Figure 1.

Image recognition inference performance for ResNet-50 v2 benchmarking at small batch sizes

The performance disparity between different platforms is shown in Figure 2; they trade more throughput (inferences per second on the X-axis) at the cost of added latency (lower responsiveness, on the Y-axis). Figure 2 charts the impact that low-latency requirements, or responsiveness, have on performance of four leading inference platforms when running a benchmark ResNet-50 workload. The X-axis ranges from zero to 100 percent and indicates the percentage of the total workload that must be processed with sub-millisecond latency, in as close to real-time as possible. The Y-axis represents actual inference performance measured in inferences per second (IPS).

In the graph, the Groq TSP100 inference platform shows no reduction in performance as the proportion of responsiveness increases: whether running with zero

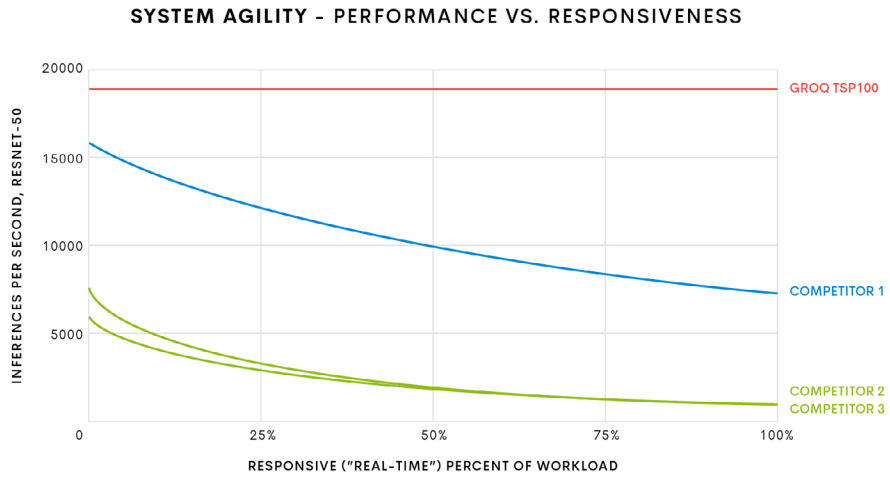

Figure 2.

200 server node with two accelerators per server, running ResNet-50. In the Groq configuration, all 400 cards can be used to yield both responsiveness and throughput, providing 100% cluster utilization

or 100 percent responsiveness, the Groq platform runs at peak performance. However, as the percentage of low latency, responsive workloads increase, the performance of the NVIDIA V100 and NVIDIA T4 GPU-based platform decreases

rapidly. Having even a small percentage of low latency, real-time workloads can strand precious resources in the cluster and reduce the utilization and overall performance of the cluster when using a GPU-based inference solution.

This dramatic decrease in performance when processing for low latency workloads impacts the size of machine learning clusters necessary to perform real-world inference processing. The Groq platform is capable of nearly 20,000 IPS no matter how much of the inference workload is responsive. However, the two NVIDIA platforms not only begin with a lower number of IPS at peak performance but also show diminishing IPS as soon as any real-time responsive workloads are added to the workflow. If running an inference platform for 100 percent, sub-1-millisecond latency, the Groq chip is almost 18 times faster than the NVIDIA GPUs.

Groq lowers TCO for machine learning investments

This performance and latency differential become especially meaningful when the above results are extrapolated into purchasing parameters for compute cluster design. To achieve the inference processing performance of a small number of Groq processors would require investment in a significantly larger number of NVIDIA GPUs.

For example, a data center engineer is deploying an 800-node compute cluster for image classification, with a portion of the node's workload distribution that requires highly responsive processing, while other workloads demand less. Figure 3 illustrates an 800-node cluster built on NVIDIA V100 GPUs and designed for ResNet-50 inference. As part of the node's performance requirements, 13 percent of the workload must be highly responsive, with latency of less than one millisecond; a further 44 percent of the workload distribution requires response times of less than 7 milliseconds; and 43 percent has no latency requirements and can be run at batch speed. Because of the low performance of GPUs at small batch sizes, to guarantee that 13 percent of the total workload runs with highly responsive inference would require dedicating over half of the entire cluster (401 nodes) to processing the low latency workloads.

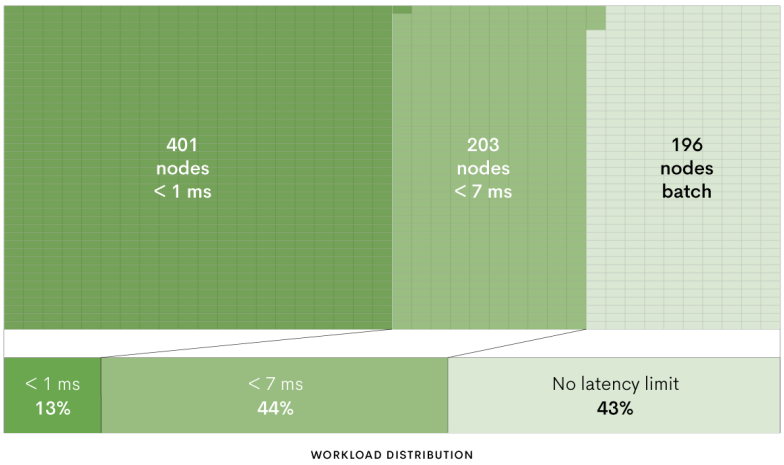

Figure 3.

Example of an 800-node compute cluster for image classification. For NVIDIA V100 ResNet-50 inference, if just 13 percent of the workload requires less than one millisecond latency, more than half of the nodes would be dedicated to serving the low latency workloads

The same latency requirements and workload distribution on Groq chips requires dramatically fewer nodes, due to the more efficient processing of small batch size workloads on Groq-based platforms; see Figure 4, below. The 13 percent of the workload that must be highly responsive, with latency of less than one millisecond, can be processed by only 25 Groq-based servers; the 44 percent of the workload distribution that requires response times of less than 7 milliseconds can be processed by 85 Groq nodes; and the 43 percent of workloads that can be run at batch speed can be handled by 83 Groq nodes. In total, deploying Groq-based nodes for the same ResNet-50 image classification workloads results in a cluster with 75 percent fewer servers while maintaining the same throughput, due to an overall 16x performance difference between NVIDIA V100 GPUs and Groq processors.

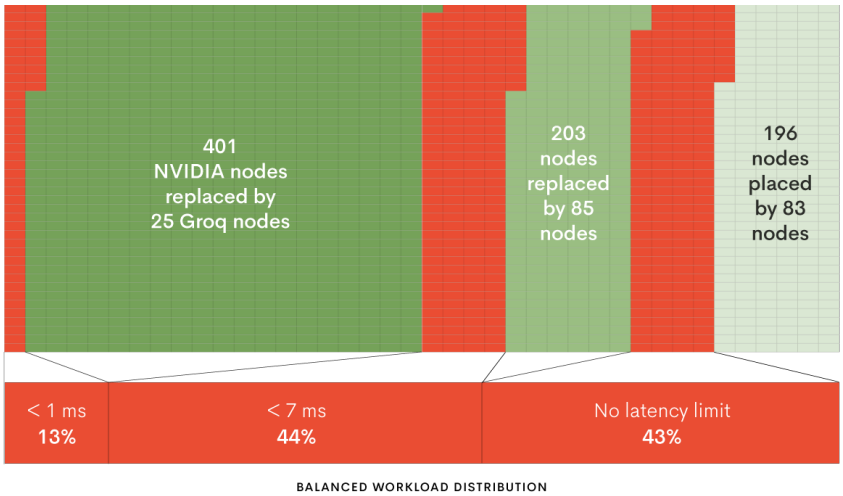

Figure 4.

193 Groq servers can replace 800 NVIDIA V100s while maintaining the same system throughput

In fact, with Groq deployed in the cluster, there is no need to make a distinction between performance and responsiveness, as Groq processes batch size one and large batch sizes with the same low latency and high efficiency. As a

result, just 25 Groq servers can replace 401 NVIDIA servers for workloads that require responsiveness to meet user expectations and satisfying demanding service level agreements (SLAs).

Conclusion

The ratio between performance and responsiveness for machine learning inference has dramatic real-world consequences for compute cluster design and investment. Other machine learning platforms require a tradeoff between latency and throughput. Performance quickly degrades in non-Groq platforms, even if a small percentage of the workload requires real-time, low latency response. Groq's revolutionary TSP architecture delivers industry leading performance and sub-millisecond latency with efficient, software-driven solutions for compute-intensive applications.

Learn more about Groq and get in contact by visiting www.groq.com.

