The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

# GROQ ROCKS NEURAL NETWORKS

## *Startup Creates New Architecture: One Core, 1,000 TOPS*

*By Linley Gwennap (January 6, 2020)*

Groq has taken an entirely new architectural approach to accelerating neural networks. Instead of creating a small programmable core and replicating it dozens or hundreds of times, the startup designed a single enormous processor that has hundreds of function units. This approach greatly reduces instruction-decoding overhead, enabling the initial Tensor Streaming Processor (TSP) chip to pack 220MB of SRAM while computing more than 400,000 integer multiply-accumulate (MAC) operations per cycle. The result is performance of up to 1,000 trillion operations per second (TOPS), four times faster than Nvidia's best GPU. Initial ResNet-50 results show a similar advantage. The chip can also handle floating-point data, allowing it to perform both inference and training. The startup is now sampling the TSP, and we expect production shipments to begin around midyear.

Even general-purpose processors have long since abandoned large monolithic CPUs in favor of multicore designs. One challenge with creating a physically large CPU is that clock skew makes it difficult to synchronize operations. Groq instead allows instructions to ripple across the processor, executing at different times in different units. As Figure 1 shows, they flow first into a set of function units called Superlane 0, which executes these instructions. In the next cycle, they execute in Superlane 1, while Superlane 0 executes a second group of instructions. This technique simplifies the design and routing, eliminates the need for synchronization, and is easily scalable; the TSP chip features 20 superlanes.

Within each superlane, data flows horizontally. In fact, the TSP continually pushes it across the chip on every clock cycle. Again, this simplifies the routing and allows a natural flow of data during neural-network calculations. As the figure shows, memory is embedded with the function units, providing a high-bandwidth data source and eliminating the need for external memory. The result is somewhat like a systolic array of heterogeneous function units, but the data only moves horizontally while the instructions move vertically.

Because the architecture lacks caches, branch prediction, and similar mechanisms, it's fully deterministic, meaning programmers can calculate throughput even before running their applications on the chip. But this approach places a heavy burden on the compiler, which must comprehend the twin pulses of instruction flow and data flow while optimizing function-unit utilization. The compiler must schedule all data movement, manage the memory and function units, and even manually fetch instructions. Groq's compiler already enables utilization similar to that of Nvidia GPUs, and the company expects further improvements as its software matures. The software tools accept models developed

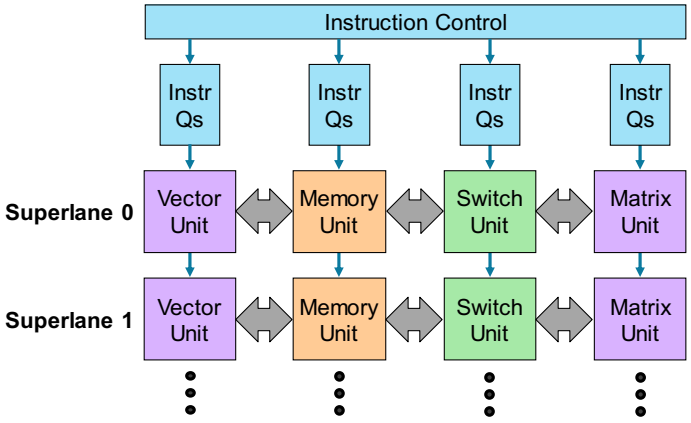


**Figure 1. TSP conceptual diagram.** In this new architecture, instructions flow downward through identical function units, pipelining the operations. Data flows across the processor, allowing the program to perform different operations.

in TensorFlow, and the company is developing drivers for other popular frameworks.

## Difficult to Grok

The entire TSP executes a single instruction stream, so we consider it one processor core. But it's effectively a 144-wide VLIW architecture, issuing one gross instructions per cycle to control the superlane. As Figure 2 shows, the TSP superlane is actually two sets of mirrored function units divided into what Groq calls the east hemisphere and the west hemisphere. Each function unit contains multiple subunits that accept instructions, as the figure shows. For example, the vector unit contains 16 ALUs that are individually controlled.

From a programmer's perspective, data is organized into streams, which physically comprise one byte per lane (320 bytes). The architecture supports 32 eastward streams and 32 westward streams. Each stream automatically progresses in its designated direction on every cycle, moving 32 bytes per lane. An instruction typically operates on data from different streams. For example, ADD S1, S2, S3 adds each value in stream 1 to the corresponding value in stream 2 and stores the results in stream 3. Thus, instead of a fixed set of 32 registers, each function unit operates on a moving set of 32 values.

A superlane comprises 16 lanes. Each instruction is performed on all 16 lanes at once, and then in the next superlane in the subsequent cycle, and so forth. Thus, over 20 cycles, each instruction executes on all 320 lanes across the 20 superlanes, so it effectively becomes a 320-byte SIMD operation having a 20-cycle pipeline. Because the architecture lacks register files, the compiler must ensure the streaming data is available to the function unit at the designated time to execute the designated instruction.

The lane structure is optimized for INT8 data, but larger operands (INT16, INT32, FP16, or FP32) can be formed by combining streams. This approach enables the compiler to operate on 320-element vectors for all data types. To simplify the hardware, the wider data types must be assigned to adjacent streams (e.g., S0, S1, S2, S3) along naturally aligned boundaries. For high reliability, the superlane applies a 9-bit error-correction code (ECC) across all 16 lanes, correcting nearly all errors; the chip logs these errors and reports them to host software.

## Look Ma, No Register Files!

The central vector unit contains 16 ALUs per lane. Each ALU can perform a 32-bit calculation using aligned groups of four stream bytes as operands. In addition to the usual arithmetic and logical operations, these ALUs can convert between integer and floating-point formats. They also perform the common normalization functions ReLU and tanh as well as exponentiation and reciprocal square roots, allowing programmers to build their own normalization functions. Although the chip's 5,120 vector ALUs can produce a total of 20 TOPS, we exclude them from the total TOPS value because they aren't MAC operations.

The matrix units handle the heavy computation. Each one contains 320 MAC units per lane that can be grouped into 20 supercells. Each MAC unit has two 8-bit weight registers and two 32-bit accumulators. On each cycle, it multiplies the stored weight values by a pair of activation values from the streaming data. Each 16x16 supercell can compute an integer partial sum in one cycle and a complete 320-element fused dot-product in 20 cycles. The MAC unit can instead perform a single FP16 MAC, but these operations require two cycles, reducing throughput by 75% relative to INT8 operations. Each hemisphere has 320x320 MAC units producing 409,600 INT8 operations or 102,400 FP16 operations per cycle. Using all 32 streams in each direction, the TSP can load all 409,600 weight registers in less than 40 cycles.

The switch units can reshape tensor data to better suit the compute units. For example, it can rotate or transpose a stream of data across the lanes. The unit can duplicate bytes to fill a vector or zero any of the vector elements to pad values. The switch units also perform an important function as the only units that can communicate between superlanes. Every unit has two sets of lane shifters that can move data up or down (north/south) to adjacent superlanes.

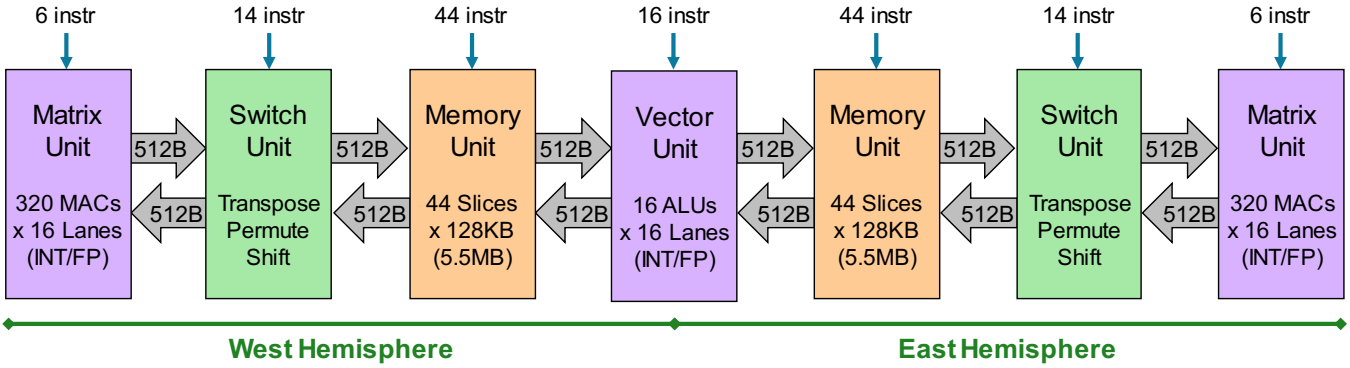


**Figure 2. TSP superlane block diagram.** Every superlane is bilaterally symmetric with an east side and a west side. It contains 16 lanes, each of which is 8 bits wide. Data flows from east to west and from west to east.

Each memory unit contains 5.5MB of SRAM divided into 44 slices (banks) of 128KB apiece. The memory unit can perform two 16-byte reads and two 16-byte writes per cycle, as long as they access different banks, allowing it to both source and sink data in two directions (east and west) across all lanes in a superlane. Combining the 20 superlanes produces 110MB of SRAM per hemisphere. Like the data moving through the superlanes, the SRAM is also ECC protected, creating a highly robust design.

## A 320-Lane Superhighway

The memory units also store VLIW instructions, which are 2,304 (144x16) bytes wide. The program manually fetches instructions when the memory units are otherwise idle; instruction fetches require less than 10% of the total SRAM bandwidth. Instructions are decoded and loaded into queues, allowing the program to prefetch. To reduce code size, the REPEAT N instruction repeats the previous instruction *N* times. Since NOP is the most common instruction, the program can specify it to last for *N* cycles.

The initial TSP chip contains 26.8 billion transistors on a massive 725mm$^2$ die built in 14nm ASIC technology. As Figure 3 shows, the die area splits about evenly between memory and compute units (not counting I/O). Instruction control requires only 3% of the die area. The chip contains an extra superlane that's normally unused, but it can map into the system to replace any superlane that has a manufacturing defect; this redundant feature ensures high yield for such a large die while adding only 4% to the die area.

The TSP's large on-chip SRAM avoids the need for external memory, so it lacks DRAM controllers and interfaces. The chip includes a x16 PCI Express Gen4 interface to connect to the host processor. (Current Xeon processors are limited to PCIe Gen3, but we expect Intel to introduce a Gen4-capable Xeon later this year.) Groq provides software that runs on an x86 host processor to download the neural-network program (instructions) and data to the TSP; the accelerator can then autonomously execute the model and return results to the host.

## Groq Rolls Over Competitors

Although the TSP provides an impressive 205 teraflop/s for FP16 data (with FP32 accumulators) at 1.0GHz, it provides four times more INT8 operations than FP16 operations. Thus, we expect the chip will serve mainly for inference. Some users prefer INT16 or FP16 for inferencing certain models, so the TSP can handle these data types as well. The 220MB of on-chip memory can hold models as big as Bert-base (110 million parameters); larger models must be divided across multiple chips.

After just a few months of testing initial silicon, Groq achieved 20,400 images per second (IPS) for ResNet-50 inference. As Figure 4 shows, that performance is more than double Nvidia's best reported score for its flagship V100 GPU and also exceeds results for Habana's Goya, the fastest

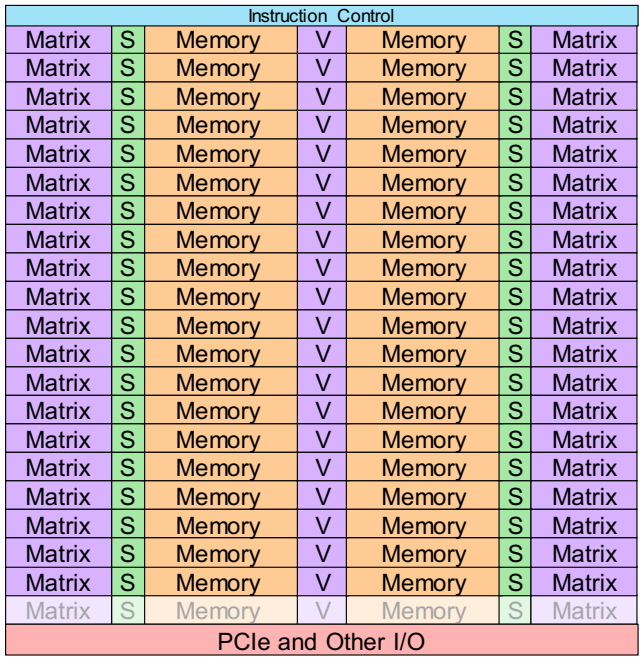


**Figure 3. Groq TSP die layout.** Built in 14nm CMOS, the 725mm$^2$ chip has 20 superlanes plus an extra superlane for redundancy. Instruction control requires only 3% of the die area. Diagram not to scale.

merchant accelerator available on the basis of this metric. The TSP still trails Alibaba's HanGuang 800, but because the Chinese cloud vendor uses that ASIC only for its own data centers, it doesn't compete directly with Groq. The start-up achieved this ResNet-50 score while running the TSP at 900MHz, so it hopes to do better over time with improvements in both clock speed and software. Its touted 1,000 TOPS requires a 1.25GHz clock speed.

One advantage of the TSP architecture is that it doesn't require large batches for optimal performance. The single-core chip achieves peak throughput while processing one image at a time. By contrast, Nvidia requires a 128-image

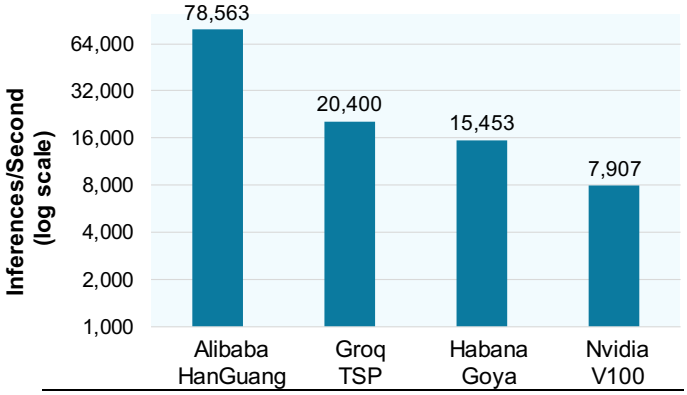


**Figure 4. ResNet-50 inference performance.** Groq's initial ResNet-50 score exceeds that of all merchant products shipping today, but it falls short of Alibaba's in-house ASIC. All data is for maximum batch size, but the TSP operates at the same IPS even for batch=1. (Data source: vendors)

### Price and Availability

Groq is now sampling the TSP to select customers; we expect production availability around mid-2020. The company withheld pricing. For more information, access www.groq.com.

batch to reach 7,907 IPS; for batch=1, the V100 delivers just 1,156 IPS, giving the TSP a much bigger advantage of 18x. Small batches are important for real-time applications to minimize latency: for ResNet-50, the TSP's latency is just 0.05ms, versus 0.87ms for the V100 at batch=1 and 16ms at batch=128. Even Goya requires 0.24ms for batch=1.

## Microarchitecture Extremes

Both Alibaba and Habana have released limited architecture information, so we compared Groq's architecture to the V100 and to Graphcore's C2 accelerator, which contains two of that company's custom deep-learning chips. All three accelerators dissipate about 300W, as Table 1 shows, and all three feature huge die built in the same process node.

The microarchitectures are different, however. The core counts range from one for the TSP to more than a thousand for the Graphcore chip (2,432 total for the card). Core count affects the software's complexity: Graphcore's compiler must divide a single neural network into thousands of tasks that can be allocated among the many tiny cores (see *MPR 9/17/18,* "Graphcore Makes Big AI Splash"). The Groq compiler can instead run a single task on the TSP, although it must still allocate work across the many parallel function units. Nvidia takes a middle road, instantiating 80 cores of moderate complexity in its Volta architecture (see *MPR 6/12/17,* "Nvidia's Volta Upgrades HPC, Training").

Unlike the other two accelerators, however, the V100 has relatively little on-chip memory (256KB of registers per core and 6MB of shared cache), so it requires an expensive high-speed memory subsystem. By eliminating most of the hardware-scheduling logic and relying on short data connections instead of registers, Groq's chip provides far more memory and compute performance than Nvidia's within similar die area and power.

The Graphcore chip is optimized for floating-point performance, and the two-chip card achieves slightly better flop/s than the TSP at the same power, although it requires twice as much silicon to do so. Graphcore mainly targets neural-network training, which exploits its FP capability, but the startup has produced lackluster benchmark results using its initial software stack and hasn't even published ResNet-50 numbers (see *MPR 12/9/19,* "Graphcore Up and Running"). For integer-based inference tasks, Groq offers far more operations per second than the C2 card.

| | Groq TSP | Nvidia V100 | Graphcore C2 |
|---|---|---|---|
| **Core Count** | 1 core | 80 cores | 2x 1,216 cores |
| **Max Clock Freq** | 1.0GHz* | 1.5GHz | 1.6GHz |
| **Peak Tflop/s (FP16)** | 205Tflop/s | 125Tflop/s | 250Tflop/s |
| **Peak TOPS (INT8)** | 820 TOPS | 250 TOPS | Not applicable |
| **Chip Memory** | 220MB | 20MB† | 2x 300MB |
| **Board Memory** | None | 32GB HBM2 | None |
| **Host Interface** | PCIe Gen4 x16 | PCIe Gen3 x16 | PCIe Gen4 x16 |
| **Board Power (TDP)** | 300W | 300W | 300W |
| **ResNet-50 Inference** | 20,400 IPS | 7,907 IPS | Undisclosed |
| **ResNet-50 Latency** | 0.04ms | 0.87ms | Undisclosed |
| **IC Process** | 14nm | TSMC 12nm | TSMC 16nm |
| **Die Area** | 725mm$^2$ | 815mm$^2$ | 2x 806mm$^2$ |
| **Production** | Mid-2020‡ | 4Q17 | 4Q19† |

**Table 1. Deep-learning-accelerator comparison.** Even though all three designs consume similar die area and power, the TSP stands out in both peak performance and ResNet-50 throughput. *Final product speed could be higher; †register-file capacity. (Source: vendors, except ‡The Linley Group estimate)

## Not a Google Clone

Jonathan Ross, a Google TPU architect, and other members of the TPU team founded Groq in 2016. The startup (located in Mountain View, of course) has raised $67 million to fund its initial chip, and the staff has grown to about 60 people. Many watchers expected Groq to quickly crank out a TPU clone, but it instead took the time to develop a unique architecture. The design bears some resemblance to the TPU—both have a single massive core comprising systolic function units—but it achieves greater performance per transistor and per watt. Compared with the TPU, the TSP's heterogeneous function units provide more flexibility, although its 320x320 MAC array is even bigger than the TPU's (see *MPR 5/8/17,* "Google TPU Boosts Machine Learning").

Nvidia, the dominant vendor of deep-learning accelerators, is Groq's primary competition. The GPU architecture is widely derided because it isn't optimized for neural networks, but the V100 delivers strong AI performance, particularly when using its tensor cores. On the basis of ResNet-50 scores, however, the TSP more than doubles the V100's best performance, and it's an order of magnitude faster for latency-sensitive workloads. Indeed, on this model, Groq's accelerator is the fastest available on the merchant market. By the time the TSP reaches production, however, Nvidia is likely to have debuted its next-generation Ampere architecture, which should come closer to the TSP's throughput.

Having delivered its first chip, Groq's new challenge is to demonstrate a wider range of models on its architecture. Although the TSP hardware is deterministic, throughput depends on the compiler's ability to optimally schedule operations. The architecture simplifies this task in some ways, but software must contend with a 144-wide statically scheduled VLIW machine comprising 320-byte SIMD units. Fully utilizing its huge MAC arrays to compute various-size tensors is challenging. And like all new accelerator vendors, Groq must broaden its

software support beyond the handful of TensorFlow operations that ResNet-50 requires, enabling customers to run production-level models on the TSP. Building out the software will take time and another funding round.

Groq rejects the manycore approach, which makes the hardware easier to design but the software more difficult. Achieving the opposite extreme—a single gigantic core—required substantial innovation. The startup has delivered on its vision with a chip that achieves excellent integer and floating-point performance and outperforms all merchant competitors on at least one neural network. The single-core design is particularly well suited to real-time cloud services that require low-latency inferences. That's something everyone should be able to grok. ♦