The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

Tenstorrent Scales AI Performance

New Multicore Architecture Leads in Data-Center Power Efficiency

By Linley Gwennap (April 13, 2020)

Tenstorrent wants to scale its AI architecture from milliwatts to kilowatts, but it's beginning with a powerful data-center chip. Hoping to find an efficient middle ground between large monolithic architectures and arrays of tiny cores, the startup is sampling a chip with 120 cores, each capable of executing three trillion operations per second (TOPS). At a peak rate of 368 TOPS, the chip runs on just 65W. When inferencing popular neural networks, Tenstorrent expects its flexible architecture to outperform competing devices that burn as much as 300W. At the recent Linley Spring Processor Conference, it disclosed initial results for the ResNet-50 and Bert models to bear out this claim.

To increase power efficiency, the startup designed its Tensix architecture to take advantage of sparsity, meaning the cores don't waste power on operations that produce no meaningful results. Although the architecture isn't neuromorphic, it employs conditional execution to achieve similar savings in software. To further improve efficiency, it features "dense math units" that can perform thousands of operations based on a single instruction. Finally, Tenstorrent optimized the core-to-core interconnect to minimize packet overhead.

The basic architecture unit is the Tensix core, which is built around a large compute engine that produces most of its 3 TOPS from a single dense math unit. The core contains 1MB of SRAM to hold weight values and feed data to the compute engine, as Figure 1 shows. Five simple CPUs handle scalar processing and manage the conditional execution. Each tile connects to adjacent tiles via four bidirectional connections that form a double 2D torus, reducing the number of hops required to move data. A packet engine processes data from the torus, decompressing it on the fly.

Founded in late 2016, Tenstorrent has raised a total of $34 million, using it to build a team of more than 50 engineers in Toronto, Canada, as well as Austin, Texas. CEO Ljubisa Bajic is also the lead architect. Early last year, the team produced a five-core test chip to validate the architecture. In December, it received silicon for its first product, code-named Grayskull, which is now sampling. We expect the chip to reach production in 2H20.

Conditioning the Core

Whereas accelerators such as the Alibaba HanGuang 800, Google TPU, and Intel Habana Goya feature a few massive compute arrays, Tenstorrent divided the compute capacity across 120 cores. This approach reduces the size of each compute unit, making it easier to efficiently fill. It also opens the door to conditional computing. Massive arrays are good for large computations, but once started, they must continue to completion. By spreading the task across many cores, each core can test the data and stop computing if the results

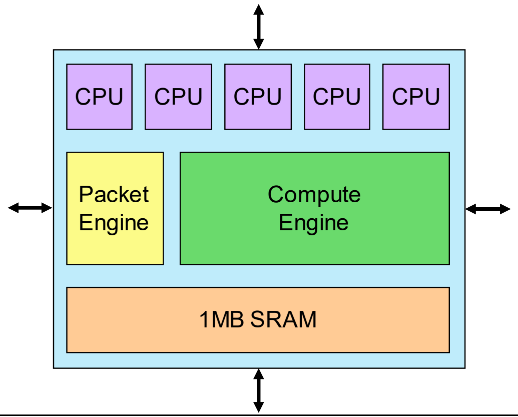

Figure 1. Tensix core. Each core features five scalar RISC CPUs that can all access the packet engine, which links to the fabric interconnect, and the compute engine, which can generate 3 TOPS.

are no longer useful, even as others continue to work. Conditional computing can save power and even boost throughput in some situations.

The CPUs in each Tensix core execute a single instruction per cycle using a proprietary RISC instruction set. They're programmable in C++, allowing developers to easily implement loops and conditional execution. These programs can also include API calls that execute vector and tensor operations using the compute unit. The five CPUs must arbitrate for access to the single compute unit. Since they're much smaller than the compute unit, the company provided several of them to keep the compute unit as busy as possible.

The compute unit is the heart of the Tensix core. It includes two components: a SIMD vector engine and a matrix/tensor engine. The SIMD unit performs basic and advanced math; it can accelerate both AI and non-AI computations (for example, signal processing). It operates on 64 values at a time, expanding any of the supported data types to a nonstandard FP format that can accumulate many FP16 products without losing precision. The tensor engine can accelerate both convolution operations and general matrix multiplication (GEMM). It produces one set of results per cycle that can be stored in memory or redirected back into the unit for further computation. Tenstorrent withheld further details, but to achieve the 3-TOPS rating, the tensor engine likely contains about a thousand 8-bit MAC units.

The compute engine operates on a variety of data types, but it's optimized for 8-bit integers (INT8). It also handles half-precision floating point (FP16) and Bfloat16 for customers that prefer greater dynamic range; these operations proceed at one-fourth the throughput of INT8 operations, however. A proprietary FP8 format permits the same throughput as INT8.

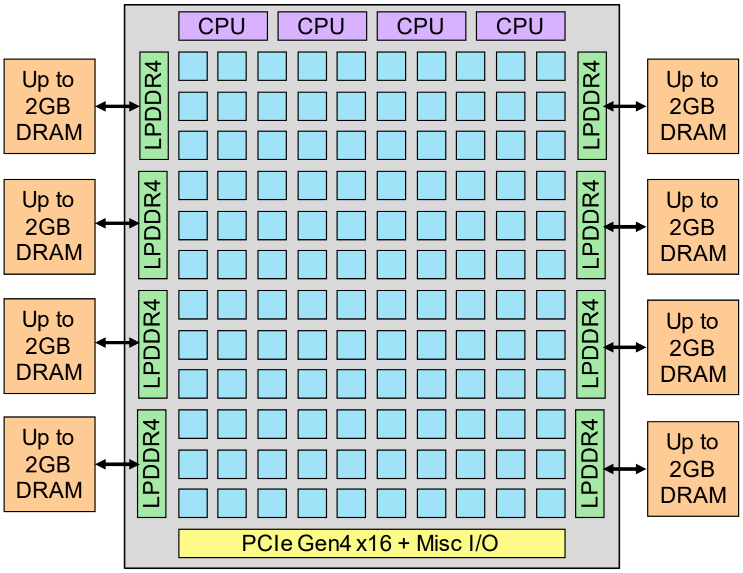

Figure 2. Grayskull accelerator. The Tenstorrent chip implements a 12x10 array of Tensix cores with a peak performance of 368 INT8 TOPS. Eight DRAM chips feed the cores, and a PCI Express interface connects to the host processor. Four superscalar CPUs manage the overall data flow.

To save memory space, the design implements a block FP format in which groups of 16 values share the same 8-bit exponent. Tensix defines block FP formats with 8-, 4-, or 2-bit mantissas, trading off throughput for precision. Once the core loads values from memory, it expands them to FP16 before any computation. In its first Nervana chip, Intel implemented a block FP mode called FlexPoint, but that approach provided only one shared exponent for all active data. The Tenstorrent design also differs in that hardware computes the shared exponent before storing the values in memory, so the compression is invisible to software.

Making a Big Mesh

Like most other AI-accelerator vendors, Tenstorrent rejects the GPU's complex memory model of large register files and small, multilevel caches. Having a small register file and 1MB of SRAM per core, Tensix creates a flat memory hierarchy that keeps data close to the compute units. At 120MB, the Grayskull chip offers about five times as much on-chip memory as Nvidia's V100 GPU, even when including the GPU's huge register storage. Although the compute unit can operate only on the local memory, each core can easily access data in other cores using the network-on-a-chip (NoC) interconnect.

The packet engine implements hardware data compression. It compresses data before transferring it across the NoC. Depending on the number of zeroes in the data, this compression typically shrinks the data by 50–75%, but the percentage can be even greater on sparse data. When the packets arrive at their destination, the compressed data is stored in the local SRAM. Once the compute engine is ready for the data, the packet engine decompresses it. After the processing completes, the packet engine compresses the data again. This approach not only increases the NoC's effective bandwidth, but it also crams more data into the core's memory.

The NoC is designed for moving huge tensors, so it uses large packets to minimize the overhead of packet headers. The packet protocol simplifies the software overhead of moving data. The NoC also implements sophisticated multicast modes to, for example, load a set of fixed weights into several or all cores at once. Software running on the RISC cores can manually transfer data, or it can configure the packet engine to autonomously move blocks of data.

In addition to the 120 Tensix cores, the NoC connects to eight DRAM controllers, as Figure 2 shows. Each supports a single LPDDR4-4266 chip for a maximum memory size of 16GB. Employing a 32-bit interface with packet-based ECC, each chip delivers 16.5GB/s of user data for a total peak bandwidth of 132GB/s. Commodity DRAM reduces cost relative to a higher-bandwidth choice such as GDDR or HBM, and LPDDR consumes less power than standard DDR memory.

Grayskull integrates four Synopsys ARC CPUs that supervise the overall operation. They can also handle special layers and functions that are poorly suited to the Tensix cores. The accelerator connects to the host processor through a 16-lane PCI Express Gen4 interface that can operate at 32GB/s. It's manufactured in Global Foundries' 12nm technology. Even including the DRAM and other components, the design fits within the 75W limit of a bus-powered PCIe card.

Tenstorrent is developing a software stack for its new architecture. The current software accepts neural networks from only the Pytorch framework directly, although it can work with other frameworks that export in the ONNX format. It handles quantization and then compiles the model to the Tensix instruction set and distributes the code and data across the 120 cores, scheduling all data movement as well as optimizing NoC and DRAM bandwidth. The CPUs in each core execute a small run-time engine instead of a standard RTOS; this code controls access to the compute and packet engines and handles core-to-core communication. Customers can also program in C++ and use the compiler to generate executable code for the processor.

Master of the Universe

Tenstorrent's chip is superficially similar to Nvidia's recent GPUs, such as the Titan RTX implementation of the Turing architecture. Each Turing core has a TOPS rating similar to that of the Tensix core. The Titan RTX features 72 cores, about two-thirds as many as Grayskull, so it falls a bit short in total TOPS. Both core types are easily programmable and flexible enough to handle a wide range of models. But the Nvidia product has far less on-chip memory, so it requires a much higher-bandwidth connection to DRAM, driving up cost and power. By optimizing for sparsity and neural networks while eliminating GPU overhead, the Tensix architecture is also far more power efficient, delivering more TOPS at less than a third of the total power, as Table 1 shows.

Groq's TSP is the most powerful merchant AI chip, generating at least 800 TOPS. It employs a radically different design with a single enormous core (see *MPR 1/6/20,* "Groq Rocks Neural Networks"). The TSP packs 220MB of on-chip SRAM—nearly twice as much as Grayskull—but it lacks external memory and is therefore best suited to models that fit in its internal memory. It requires far more power and exhibits poor utilization of its peak TOPS rating.

Price and Availability

Tenstorrent is sampling its first chip, code-named Grayskull; we expect production in 2H20. The startup withheld product details and pricing. For more information, access *tenstorrent.com*. Video of Tenstorrent's presentation at the Linley Spring Processor Conference will be at *www.linleygroup.com/SPC20*.

Tenstorrent has run initial benchmarks on its chip and expects to improve on them as it nears production. Even with preproduction software, Grayskull achieves 22,431 images per second (IPS) when running ResNet-50 inference with a batch size of 20. This score exceeds that of the Titan RTX by nearly 30% and beats the TSP as well. Grayskull delivers about 50% of this score even when running at batch=1, whereas the GPU loses about 80% of its peak performance. Having a single core, Groq is the champion for batch=1.

Tenstorrent measured the much larger Bert-base model at 2,830 sentences per second (SPS) with a sequence length of 128; Nvidia's V100 achieves 2,891 SPS on the same test. The startup also tested an "early exit" algorithm that simply stops after module five (about halfway through) if the confidence level is already high enough. This approach increases throughput by 1.76x to 5,632 SPS while reducing the model's accuracy by only 0.5%, which is acceptable for MLPerf testing and many end applications. The company achieved another 2x gain by adapting to Bert's variable sequence lengths. These gains demonstrate the value of Tensix's conditional execution, but they require some changes to the model.

Grayskull's ResNet-50 inference throughput exceeds that of all other merchant accelerators, as Figure 3 shows, but trails that of the HanGuang 800 ASIC, which Alibaba deploys in house (see *MPR 3/2/20,* "Alibaba Uses Convolution Architecture"). Most of the chips in this chart, however, require 200–300W—far more than Grayskull. The green

	Tenstorrent Grayskull	Groq TSP	Nvidia Titan RTX
Core Count	120 cores	1 core	72 cores
Max Clock Speed	1.3GHz*	1.0GHz*	1.7GHz
Peak FP16 Performance	92Tflop/s	205Tflop/s	130Tflop/s
Peak INT8 Performance	368 TOPS	820 TOPS	261 TOPS
Chip Memory	120MB	220MB	18MB reg files
DRAM Channels	8x LPDDR4	None	2x GDDR6
DRAM Bandwidth	132GB/s	None	672GB/s
Host Interface	PCIe Gen4 x16	PCIe Gen4 x16	PCIe Gen3 x16
Coherent Interconnect	None	Undisclosed	NVLink
ResNet-50 Inference†	22,431 IPS	20,400 IPS	17,400 IPS‡
ResNet-50 Utilization‡	23% of TOPS	11% of TOPS	24% of TOPS
Board Power (TDP)	75W	300W	280W
ResNet-50 Efficiency	393 IPS/W	68 IPS/W	62 IPS/W
IC Process	GF 12nm	14nm	TSMC 12nm
Die Area	620mm^2	725mm^2	754mm^2
Production	2H20	Mid-2020‡	4Q18

Table 1. Deep-learning accelerators for inference. IPS=images per second. Tenstorrent delivers by far the best power efficiency (IPS/W) despite Groq's leading TOPS rating. *Final-product speed could be higher; †best batch size. (Source: vendors, except ‡The Linley Group estimate)

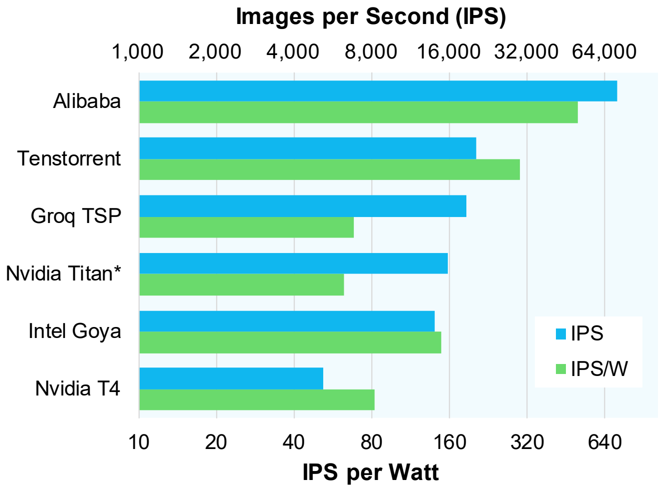

Figure 3. Accelerator inference performance. Tenstorrent offers better ResNet-50 throughput and far better performance per watt than leading merchant accelerators, but it falls behind Alibaba's HanGuang 800 ASIC. (Data source: vendors, except *The Linley Group estimate)

bars compare the power efficiency (performance per watt) of each product.

For power efficiency, Alibaba still comes out ahead using a low-voltage mode in HanGuang, but Tenstorrent's product is about twice as efficient as Intel's Goya chip, the nearest merchant competitor, and four times better than the Titan RTX and Groq's TSP. Figure 3 also shows Nvidia's power-efficient T4, which has about the same TDP as Grayskull. By running its Turing cores at a lower speed than the Titan RTX, the T4 improves its power efficiency slightly, but Tenstorrent more than triples the efficiency of Nvidia's best product.

Tops in Architecture Efficiency

Tenstorrent is already working on its next-generation product, code-named Wormhole. This chip will extend the Tensix architecture for neural-network training, likely by adding FP32 support. Wormhole will also include coherent interfaces, which Grayskull lacks, to connect multiple accelerators in order to train large models. The startup hopes to tape out its training chip later this year and ship production units in 2021.

Tenstorrent has already made strong progress in developing and validating the Tensix architecture, which combines common elements, such as a mesh of cores and a large on-chip memory, with unusual capabilities. The dense math units help the chip deliver better MAC utilization than traditional architectures. The larger core count improves utilization relative to monolithic designs. And the emphasis on scalar code and conditional execution provides intriguing opportunities for performance enhancement, as company's testing of early-exit algorithms indicates.

These capabilities, along with other optimizations, help Tenstorrent achieve leadership performance and power efficiency compared with all other announced inference accelerators. The power efficiency is particularly impressive, as the company can deliver more performance in a 75W PCIe card than other vendors achieve in a 300W product. Although Alibaba's in-house ASIC can provide better efficiency than Grayskull, it does so only when running at low voltage; Tenstorrent could probably outdo Alibaba using the same trick. Other accelerator vendors will be hard pressed to catch up on this metric.

Like other AI startups, Tenstorrent's biggest challenge is software. It currently supports only a single framework directly, and even then for just a few models. The company has raised a respectable amount of money, but it'll need considerably more to build a competitive software stack while continuing to develop its hardware roadmap. The successful tapeout and excellent benchmark performance of the initial product should attract interest from both investors and customers, giving Tenstorrent the ability to scale for the future. ♦