MOOR INSIGHTS & STRATEGY

Tenstorrent's Holistic Stack Of AI Innovation

INTRODUCTION

The explosive growth of AI processing in data center and edge environments has induced AI startups and established firms alike to develop silicon to handle the massive processing demands of neural networks. Inference processing, in particular, is an emerging opportunity, wherein a trained deep neural network is processed to predict characteristics of new data samples. This processing is typically performed on CPUs. However, that situation will have to change to handle the exponential growth in model size and new applications that depend on multiple neural networks to solve complex problems. We believe that the market for inference processing will exceed that of data center AI training in 3-4 years, surpassing $5B in annual chip sales by 2025.

The AI industry group OpenAI estimates that AI model sizes are doubling every 3.5 months, outpacing the improvement rate seen in silicon by several orders of magnitude. Canadian startup Tenstorrent, co-founded by AMD veteran Ljubisa Bajic, believes this challenge represents a unique opportunity to deliver performance that increases a full order of magnitude. To achieve this pace, Tenstorrent is taking a holistic approach to AI processing. This research brief will examine Tenstorrent's strategy, technology, differentiation and challenges. Given the breadth and depth of Tenstorrent's novel approach to hardware, we have deferred the critical discussion of software and firmware to a future brief.

TENSTORRENT'S HOLISTIC STRATEGY

Tenstorrent CEO Bajic realizes that incremental advancements in faster AI chips will only result in linear performance improvements. Such a meager pace will fail to meet the demands of exponentially larger AI models, which will increase by four orders of magnitude over the next five years. Consequently, Tenstorrent will approach AI acceleration hardware holistically on three levels: faster chips, improved scaling of parallel algorithms and computation load reduction through dynamic execution. The Tenstorrent approach appears to be unique in the industry, with novel designs to increase performance, improve memory utilization and reduce problem size to achieve impressive early results.

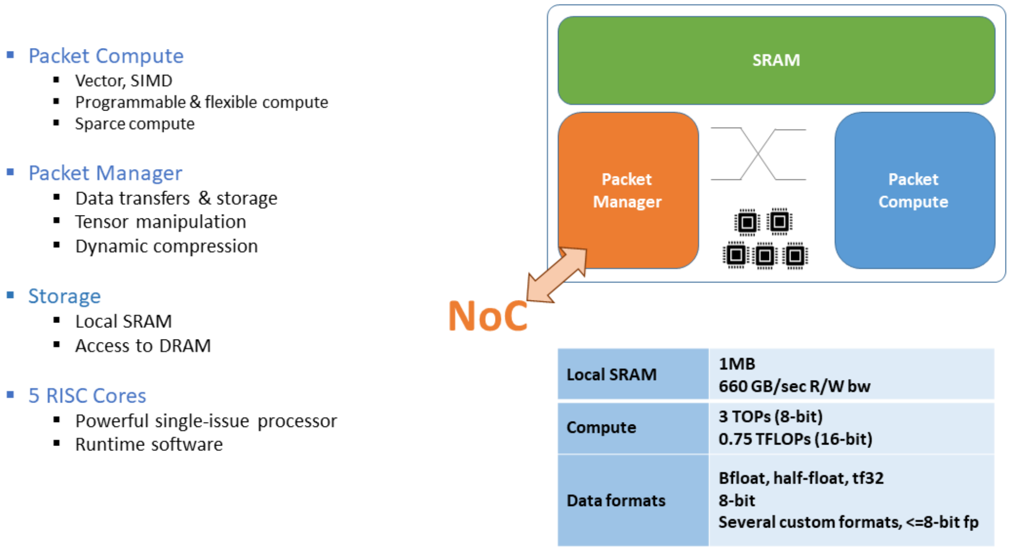
Copyright ©2020 Moor Insights & Strategy

START WITH AN EFFICIENT CHIP

The first production Tenstorrent chip, Grayskull, will target the market for inference processing in the data center and heavy-duty edge environments such as smart factories, retail locations and edge cloud computing. As we will explore, the Grayskull chip employs new concepts we haven't seen in AI silicon, such as packetization, dynamic memory allocation and dedicated hardware for tensor manipulations. The 65-watt, 368-TOPS (8-bit float) Grayskull chip contains 120 processing elements (PEs), each of which includes a packet manager, a compute engine, 1MB of SRAM and access to the chip's NOC, PCIe Gen 4 and 8 LPDDR4 controllers. Each PE also contains 5 RISC-V CPUs for runtime control support.

The fully programmable compute engine is a custom vector SIMD core that supports sparse operations and a variety of precisions and formats, including 16-bit floating-point operations at 92 TFLOPS.

FIGURE 1: TENSTORRENT GRAYSKULL PROCESSING ELEMENT (SINGLE CORE)



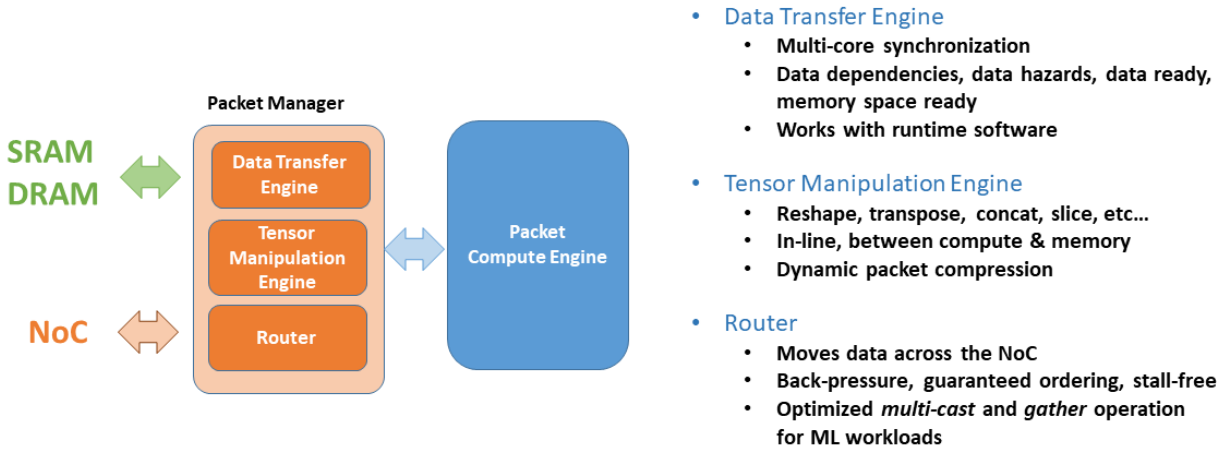
- Packet Compute
  - Vector, SIMD
  - Programmable & flexible compute
  - Sparce compute
- Packet Manager
  - Data transfers & storage
  - Tensor manipulation
  - Dynamic compression
- Storage
  - Local SRAM
  - Access to DRAM
- 5 RISC Cores
  - Powerful single-issue processor
  - Runtime software

Local SRAM	1MB 660 GB/sec R/W bw
Compute	3 TOPs (8-bit) 0.75 TFLOPs (16-bit)
Data formats	Bfloat, half-float, tf32 8-bit Several custom formats, <=8-bit fp

Source: Tenstorrent

A first in the industry, the Tenstorrent design is based on communication between PEs with packets, containing an instruction queue and mini-tensor elements decomposed from the model's full-size tensors. This approach should improve fine-grained parallelism and speed performance. The Packet Manager interacts with other PEs via the NOC and provides dynamic compression, tensor manipulations such as matrix transposition and storage. The data resides on the PE's SRAM, with overflow storage available on the LPDDR4 DRAM. When data preparation is complete, the Packet Manager then transfers control to the Packet Computer Engine for processing.

FIGURE 2: PACKET MANAGER



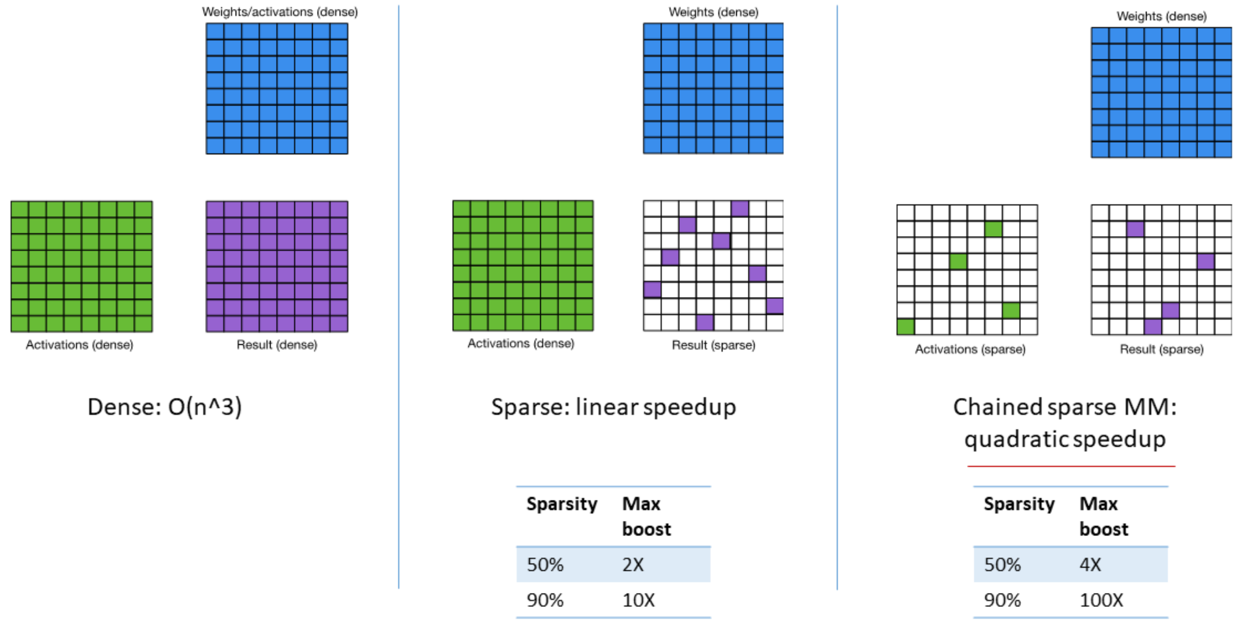
- Data Transfer Engine
  - Multi-core synchronization
  - Data dependencies, data hazards, data ready, memory space ready
  - Works with runtime software
- Tensor Manipulation Engine
  - Reshape, transpose, concat, slice, etc...
  - In-line, between compute & memory
  - Dynamic packet compression
- Router
  - Moves data across the NoC
  - Back-pressure, guaranteed ordering, stall-free
  - Optimized *multi-cast* and *gather* operation for ML workloads

Source: Tenstorrent

While an apples-to-apples comparison with real-world workloads is not yet available, Grayskull, at least on the surface, compares favorably to alternatives such as next year's NVIDIA Ampere-based Drive AGX Level 2+ SOC, which delivers some 200 TOPS at 45 watts. However, we must note that the near-decade lead NVIDIA enjoys in the software and development ecosystem cannot be matched by any startup, so we consider Grayskull a good starting point for that R&D, not a plug-and-play alternative to established vendor products.

The Grayskull device also supports dynamic sparsity in matrix multiplication; most chips only support a fixed level of sparsity, usually 50% for a 2X speedup. With Tenstorrent, a 90% sparse matrix multiplication will see a 10X performance boost. Tenstorrent also supports sparsity for activations, which can result in up to a 100X performance improvement in 90% sparse matrices.

FIGURE 3: O(N) MATRIX MULTIPLICATION



Sparsity	Max boost
50%	2X
90%	10X

Sparsity	Max boost
50%	4X
90%	100X

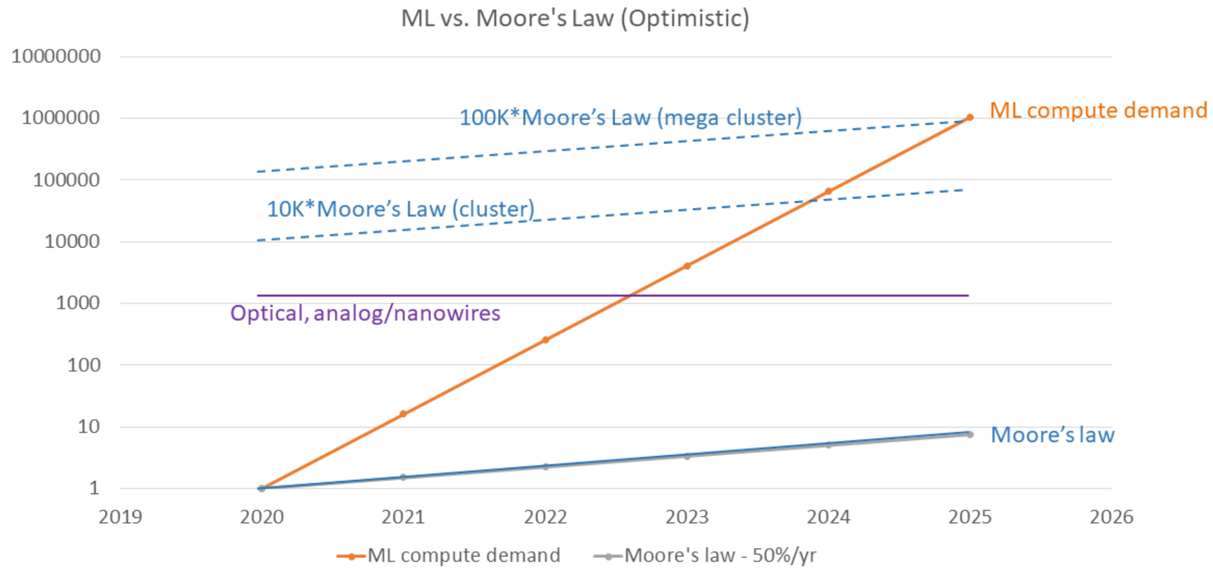
Source: Tenstorrent

SCALE-OUT: NECESSARY BUT INSUFFICIENT

AI chips are already astonishingly fast. However, even hundreds or thousands of trillion operations per second is inadequate for training massive AI models that consist of billions of parameters. Hence, AI processing is achieved through aggressive scaling, requiring perhaps a thousand accelerators to complete a training run in less than an hour.

While a fast chip is an essential building block, efficient scaling is needed to solve larger AI problems, especially in training. If one uses the growth of AI models as a proxy for the required compute performance to handle these larger models, a reasonable assumption, the situation ahead becomes quite apparent. Larger models will require an increase in performance of six orders of magnitude over the next 4-5 years, a level not achievable by any imagined chip advances. Consequently, near-linear scaling across enormous clusters is becoming the benchmark by which AI hardware must compete, all else being equal.

FIGURE 4: ML VS MOORE'S LAW (OPTIMISTIC)

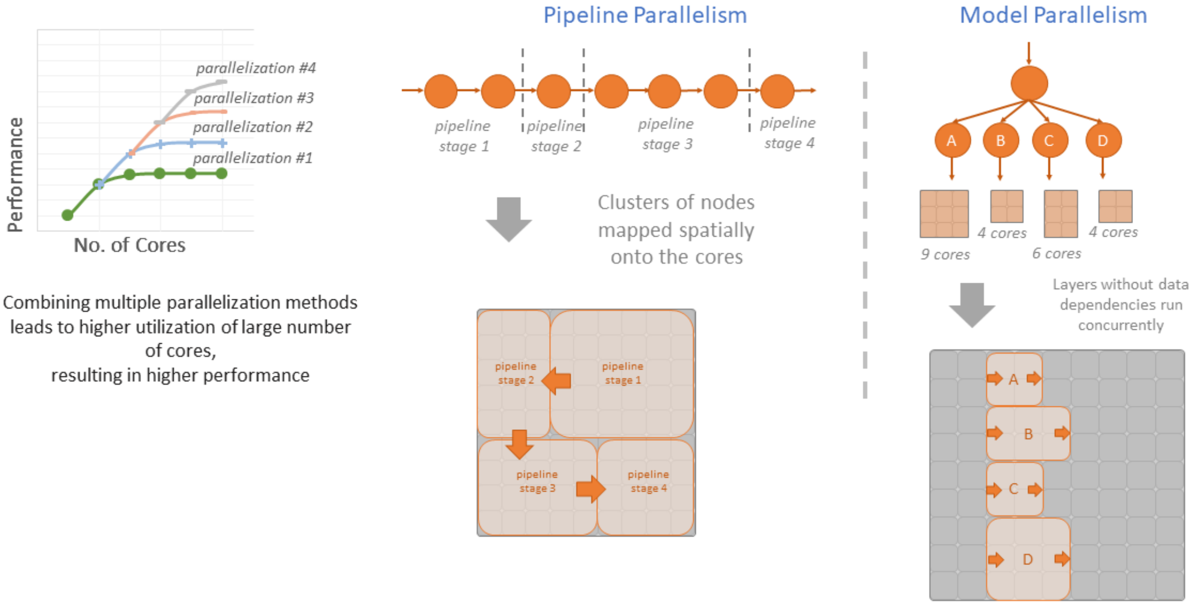


Source: Tenstorrent

As we shall see in the section covering the Tenstorrent roadmap, the company's training chip will support massive clusters, with 16 ports of 100Gb Ethernet providing the required bandwidth. The intent is to enable enormous levels of scaling typically only realized in supercomputer research facilities. To enable easy-to-use scaling, the Tenstorrent compiler will provide write-once, deploy-at-any-scale capability. If this works as designed, data scientists will not need to explicitly manage deployment across a cluster.

Additionally, we note that traditional shared-memory architectures cannot support the required level of parallelism. Thus, many architectures turn to private memory hierarchies, be it HBM2 memory on a GPU or faster but far smaller SRAM on an AI ASIC. These approaches all require explicit data movement from larger memory stores. The Tenstorrent Packet Manager uses run-time dynamic memory management to reduce data movement and lower latencies across the SRAM, LPDDR on the Grayskull board and host memory. The Tenstorrent training chip will support GDDR6 for increased private memory bandwidth and capacity at low latencies.

FIGURE 5: FLEXIBLE SCHEDULING & PARALLELIZATION



Combining multiple parallelization methods leads to higher utilization of large number of cores,
resulting in higher performance

Source: Tenstorrent

REDUCE THE PROBLEM SIZE WITH DYNAMIC EXECUTION

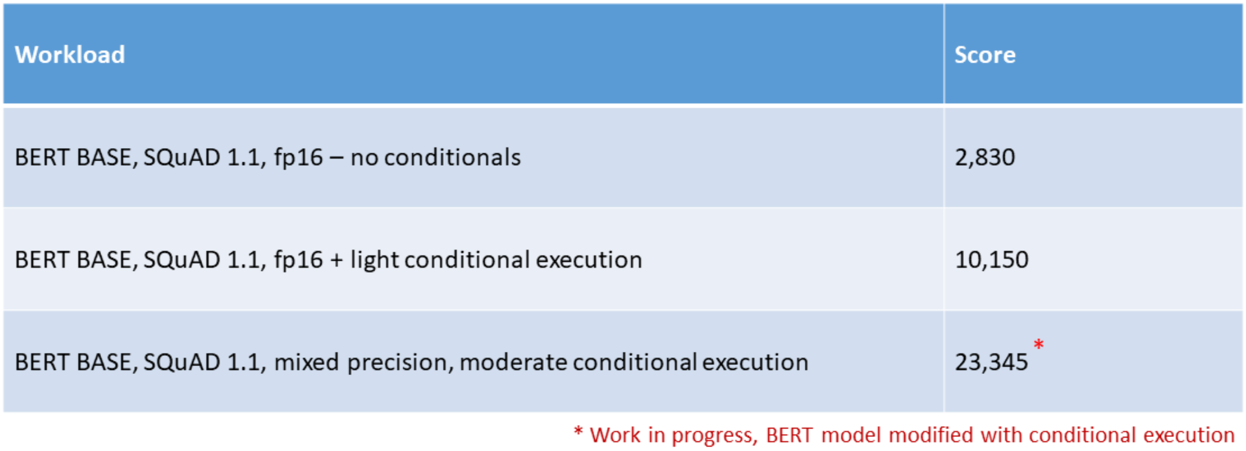
Even aggressive scaling will fail to deliver the performance that AI applications will require over the coming years. Since fast chips and scaling up will not adequately address the challenge, Tenstorrent focuses on "dynamic execution," which reduces the computation required to train and execute neural networks. The term applies to several optimizations the company exploits in software and hardware: control flow, compression, sparsity, dynamic precision and conditional execution.

Three exciting aspects of dynamic execution are dynamic precision, run-time compression and conditional execution. Dynamic precision optimizes the processing of network layers with the least amount of precision that still produces the desired accuracy of predictions, fine-tuning among various integer and floating-point formats in individual layers of the net. Run-time compression can reduce the data set size and minimize data transfer time and power consumption. Finally, conditional execution enables the early termination of an inference run when the output of the process adequately converges to a stable state.

The benchmarks below demonstrate the benefits of conditional execution in natural language processing (NLP). Early tests show that inferencing on Google's Bidirectional

Encoder Representation from Transformers (BERT BASE) produced an eightfold increase in performance on Grayskull when both dynamic precision and conditional execution are employed. It is important to note that one must modify a model to be self-aware of execution shortcut opportunities; this is not just a compile-and-go feature and exploitation will require a deep understanding of the inner workings of the neural network. But improving performance by an order of magnitude justifies some hard work.

TABLE 1: 65W GRAYSKULL BERT INFERENCE PERFORMANCE



Workload	Score
BERT BASE, SQuAD 1.1, fp16 – no conditionals	2,830
BERT BASE, SQuAD 1.1, fp16 + light conditional execution	10,150
BERT BASE, SQuAD 1.1, mixed precision, moderate conditional execution	23,345*

* Work in progress, BERT model modified with conditional execution

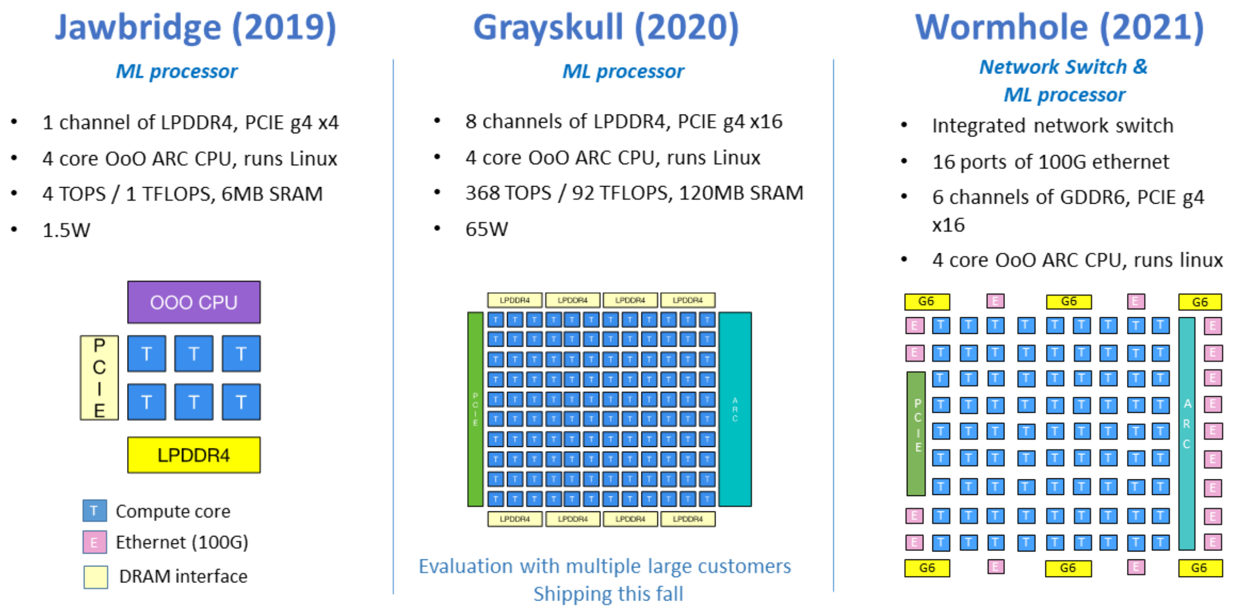
Source: Tenstorrent

TENSTORRENT PRODUCT ROADMAP

Tenstorrent has disclosed that the Grayskull chip for inference processing will be available for evaluation later this year to large potential customers. Details about the delivery platform(s) will be available soon. While Grayskull enables efficient inference processing, including conditional execution, the company's second production chip, Wormhole, will support the massive scale needed to train large models, with 16 ports of 100Gb Ethernet and fast local memory. Armed with this bandwidth, the company is eyeing the network switch market as well as AI training. Interestingly, the company decided to forego the higher bandwidth and much higher costs of HBM2 memory used in GPUs and selected GDDR6 instead, the fast memory used in workstation GPUs.

On the software front, Tenstorrent has described three critical areas of development: compilers/frameworks, firmware for the packet manager and run-time support to execute on the RISC CPUs. Software is the key to turning good hardware into great solutions in the data center, so this piece of the puzzle will be vital to its success. The company's software roadmap is still a work in progress; we will review its plans in a future publication.

FIGURE 6: TENSTORRENT SILICON ROADMAP



Jawbridge (2019)

ML processor

- 1 channel of LPDDR4, PCIE g4 x4
- 4 core OoO ARC CPU, runs Linux
- 4 TOPS / 1 TFLOPS, 6MB SRAM
- 1.5W

T Compute core
E Ethernet (100G)
□ DRAM interface

Grayskull (2020)

ML processor

- 8 channels of LPDDR4, PCIE g4 x16
- 4 core OoO ARC CPU, runs Linux
- 368 TOPS / 92 TFLOPS, 120MB SRAM
- 65W

Evaluation with multiple large customers
Shipping this fall

Wormhole (2021)

Network Switch & ML processor

- Integrated network switch
- 16 ports of 100G ethernet
- 6 channels of GDDR6, PCIE g4 x16
- 4 core OoO ARC CPU, runs linux

Source: Tenstorrent

CONCLUSIONS: THE HOLISTIC APPROACH HOLDS TREMENDOUS PROMISE AND CHALLENGES

Like many AI hardware startups, Tenstorrent has an excellent story to tell, full of innovation and unique strategies. Only time and hard work will tell if the company can capitalize on its technology by gaining initial traction in flagship accounts. In Tenstorrent's case, we see several areas of potential differentiation and associated challenges:

1. At a high level, we find the company's holistic viewpoint refreshing and realistic. Bajic recognizes that GPUs are excellent tools for AI. He envisions a top-to-bottom portfolio of potential long-term sources of differentiation, knowing that competitors can and will adopt similar ideas.
2. The Packet Manager is a new concept that enables dynamic memory management, tensor manipulations and compression. This implementation appears to be transparent to the developer and has the potential for further exploitation in future designs.

3. The concept of dynamic execution is powerful; reducing the size of the problem is the only way forward as models outgrow silicon capabilities by several orders of magnitude. The challenge, of course, is that many competitors are exploring these same ideas, including NVIDIA, Qualcomm and other startups. Tenstorrent will need to continue to innovate in this area to stay ahead.
4. Tenstorrent supports fine-grained parallelism enabled by mini-tensors. The abilities of the compiler and run-time support will need validation.
5. "Write-once, deploy-at-any-scale" is essential for massive scale model parallelism. We note that this feature has become a common theme for other startups, such as Graphcore.

To turn these potential differentiators into a successful company, Tenstorrent must tackle two rather tricky challenges. First, it must successfully engage critical customers and research institutions to gather feedback and hopefully gain traction in testing and production deployments at scale. Early customer engagement is a considerable challenge, especially for any startup focused on engineering and development. The second challenge is easier for bright engineers: continue to find new approaches to problem-size reduction. We have not heard many companies discuss this avenue publicly, aside from Qualcomm, but this approach is inevitable given the gap between AI's needs and the pace of silicon advancements. This area is ripe for collaboration if Tenstorrent can motivate research institutions to explore these new approaches with its hardware and software.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR

Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

This document, including any supporting materials, is owned by Moor Insights & Strategy. This publication may not be reproduced, distributed, or shared in any form without Moor Insights & Strategy's prior written permission.

DISCLOSURES

This paper was commissioned by Tenstorrent, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.

©2020 Moor Insights & Strategy. Company and product names are used for informational purposes only and may be trademarks of their respective owners.