1Wから1MWまで同一アーキテクチャでスケールする AIアクセラレータ

Tenstorrent Japan S.FAE Yasuhiro Ito

Oct 2023

tenstorrent

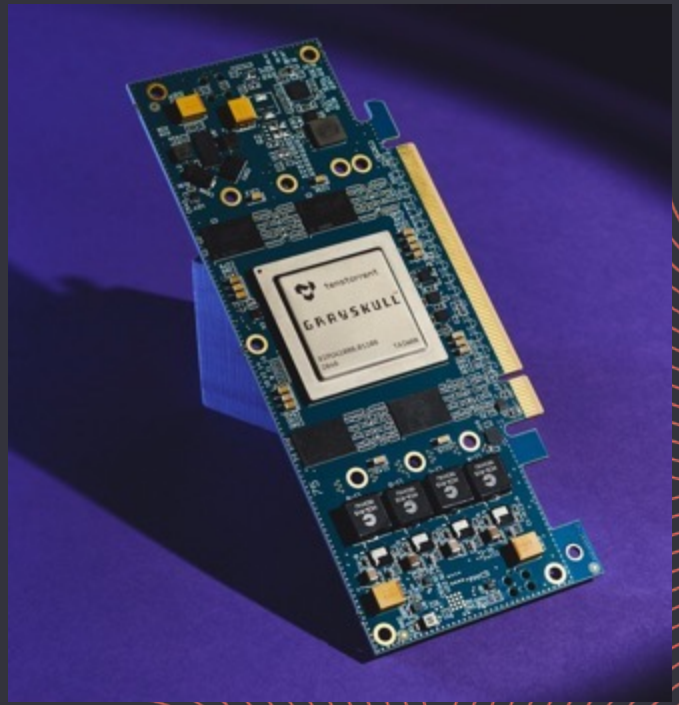
Topics

- About Tenstorrent & AI-chips + RISC-V CPU
- Scalable architecture and Software development kits
- Single Accelerator & SDK cover from Edge to Cloud .

Our company & products

About Tenstorrent

- Based in North America, with offices in Tokyo, Toronto, Santa Clara, Austin, Belgrade, and Bengaluru.
- Tenstorrent builds the most innovative AI products:
  - Inference and Training, CNNs, LLMs, and NLPs
  - Powerful software stacks for models & bare metal programming
- Tenstorrent created the highest performing RISC-V CPU technology in the world
- Led by industry veteran hardware engineer and CPU architect, CEO Jim Keller.



Tenstorrent CPU Team

Jim Keller

CEO, Digital Alpha processor, Apple A series, AMD Zen, Tesla Autonomous Driving system

Wei-Han Lien

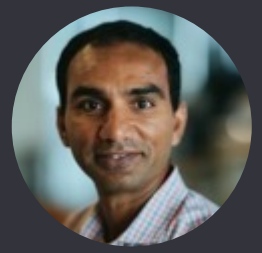
Chief CPU Architect: Apple, PA Semi, AMD

Jim Montanaro

PD Fellow: Apple, AMD

Dan Bailey

Senior Fellow: Tesla, AMD, DEC



Srikanth Arekapudi

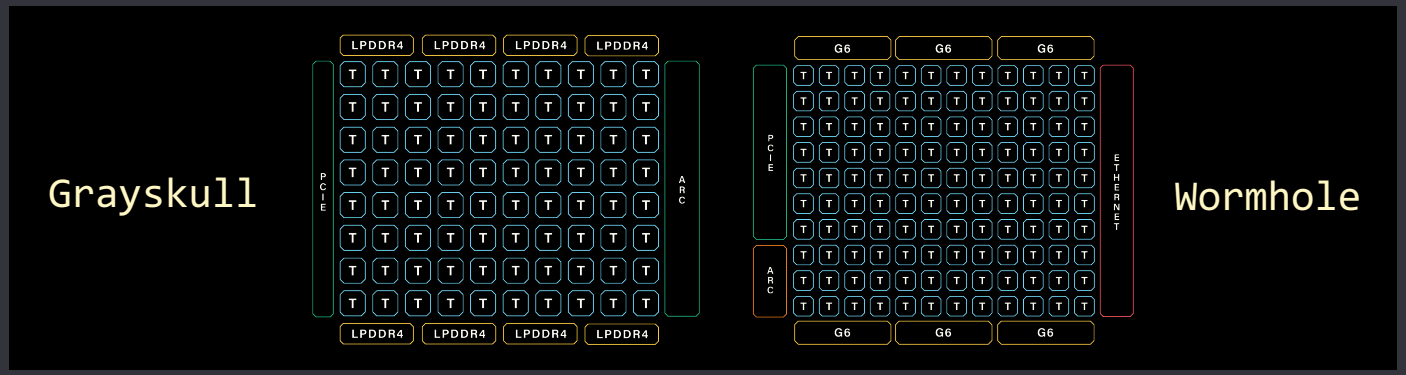
RTL/DV Fellow: Cerebras, AMD

Yasuo Ishii

Architecture Fellow: Arm, NEC

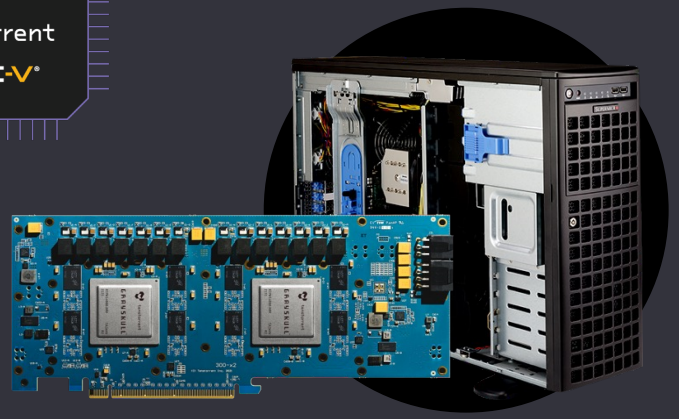
Current products
We deliver AI/ML to customers

Grayskull
For Infenrence



Wormhole
For Inference & Training

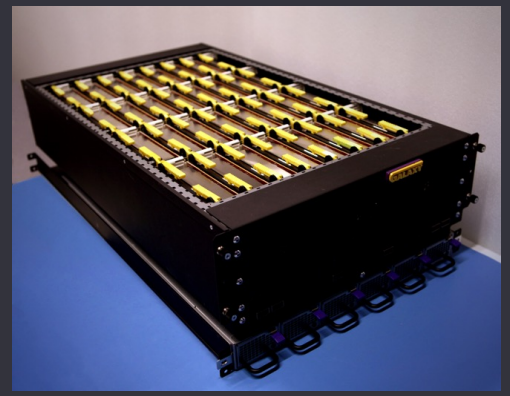
IP



Workstations

PCIe Cards

Servers

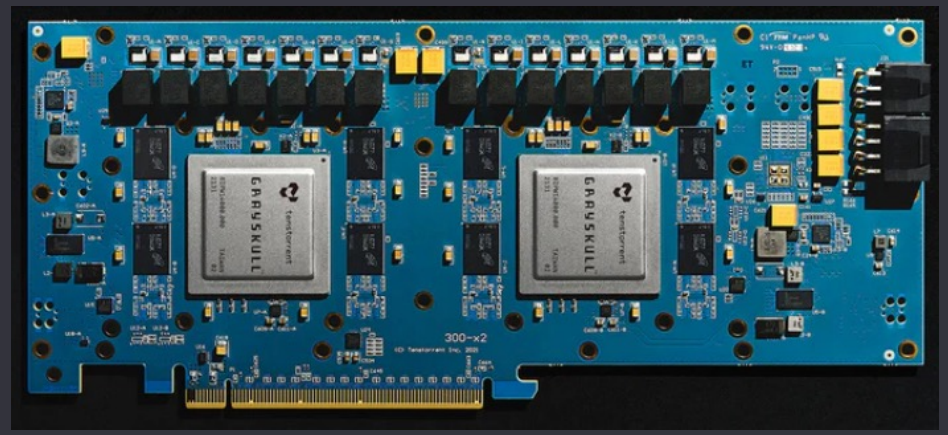


Galaxy

AI Cloud

tenstorrent

PCIe Cards



	n300	n150	e300	e150	e75
Technology	Wormhole, dual chip	Wormhole, single chip	Grayskull, dual chip	Grayskull, single chip	Grayskull, single chip
Form Factor	¾ length	¾ length	¾ length	¾ length	½ length
Performance	Silicon TOPS 540	Silicon TOPS 315	Silicon TOPS 442	Silicon TOPS 276	Silicon TOPS 220
DRAM	24 GB, 1152 GB/sec	12 GB, 576 GB/sec	16 GB, 200 GB/sec	8 GB, 100 GB/sec	8 GB, 100 GB/sec
Network	PCIe 4, 16x 100Gb	PCIe 4, 16x 100Gb	PCIe Gen 4	PCIe Gen 4	PCIe Gen 4
Power	300W max	150W max	300W max	150W max	75W max

Wormhole Products (2nd Gen device for AI at scale)

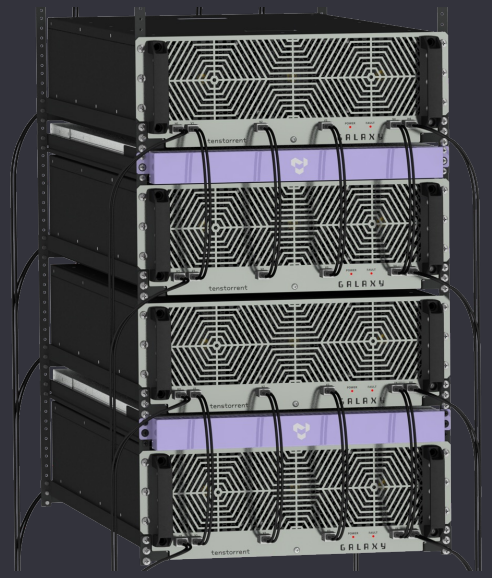
12nm AI Accelerator on PCIe Gen 4

N300s/d (Nebula, single or dual chip config available)

- Modular device with 1.6TB onboard ethernet
- Natively scalable to an arbitrary number of devices
- High performance at low cost

Galaxy

- High-density AI servers in 4U enclosures
- Comprised of 32 x n300s devices
- Includes backplane interconnect, active cooling
- 7PFLOP (BF8) at 7.5KW



tenstorrent

Tenstorrent wins on Perf/$ (and Perf/W)

Galaxy wins on OOB Large Language Model Performance and Price

nVidia DGX

8x A100
$120,000

(est. ~6.5 kW)

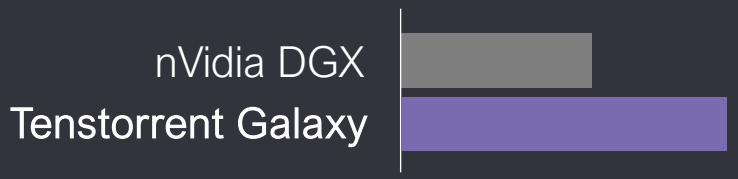
vs.

Tenstorrent Galaxy

32x Wormhole
$80,000

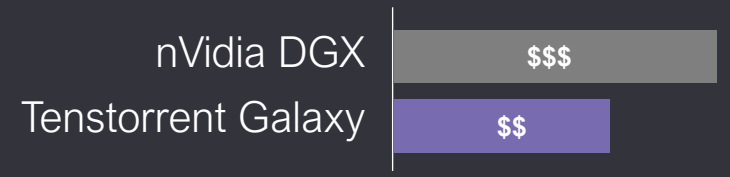
(est. ~6.5 kW)

Out of the Box Performance



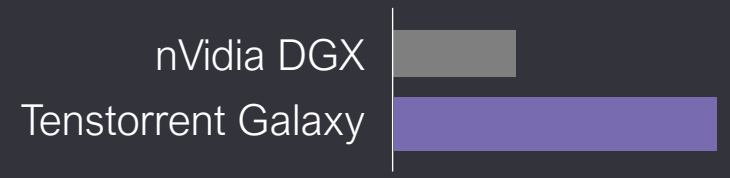
Delivers → **1.7x MORE Performance** on BERT-large model

Price



Delivers → **33% Lower Cost** on BERT-large model
* List price comparison

Price Performance



Delivers → **2.6x MORE Price Performance** on BERT-large model
* List price comparison

tenstorrent

Tensix: Scalable AI accelerator

AI Chip Roadmap

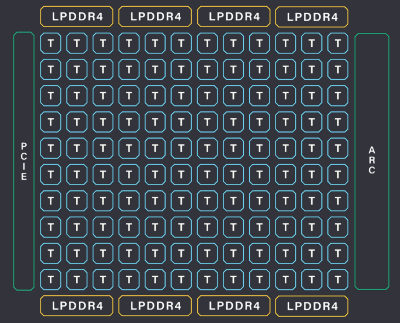
Heterogenous

Chiplet

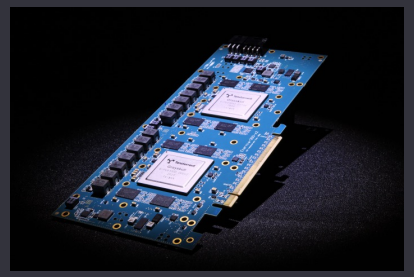
2021 → 2022 → 2023 → 2024

Grayskull

ML Processor



- 12nm, 276 TFLOP (FP8)



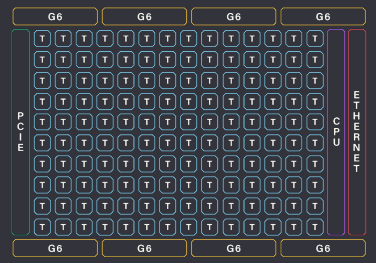
Wormhole

Networked ML Processor

- 12nm, 328 TFLOP (FP8)
- 200 GB/S Scale-out Ethernet

Black Hole

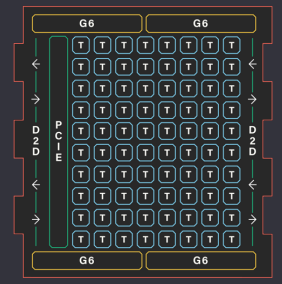
Standalone ML Computer



- 6nm, 600mm^2
- Small RISC-V cluster
- Heterogenous compute

Quasar

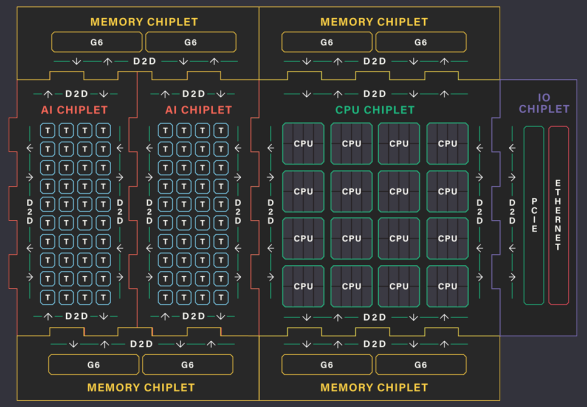
Low Power, Low Cost ML Chiplet



- ML Chiplet

Grendel

Highly Configurable and Performant ML Chiplet

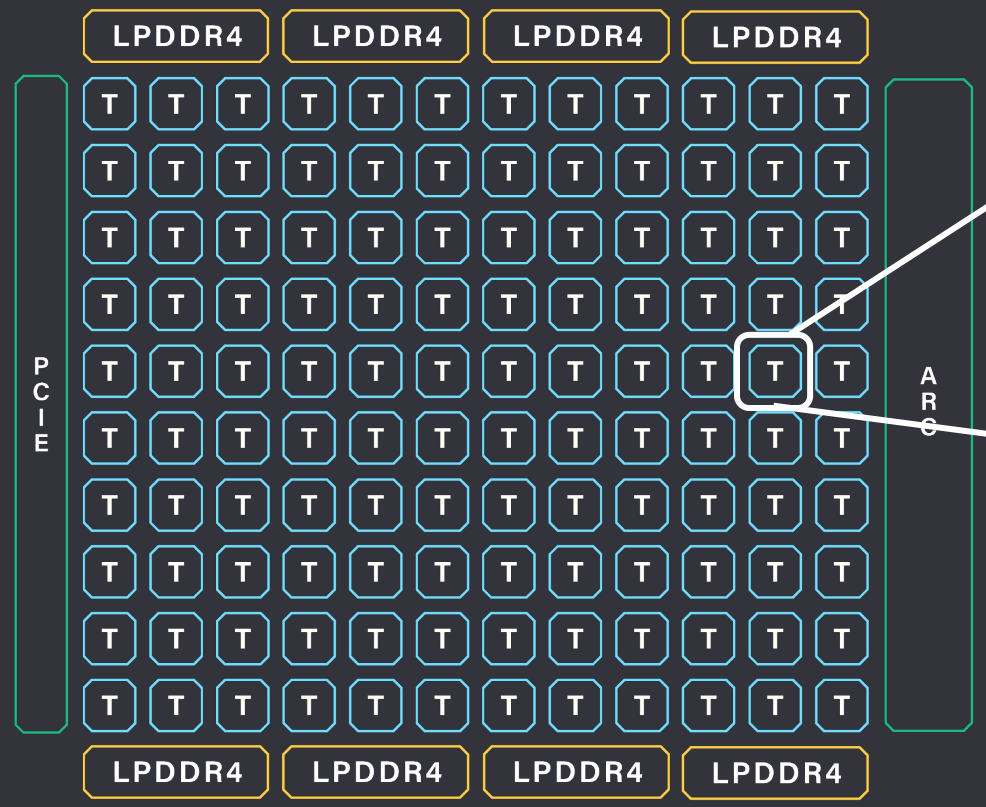


- CPU + ML chiplets

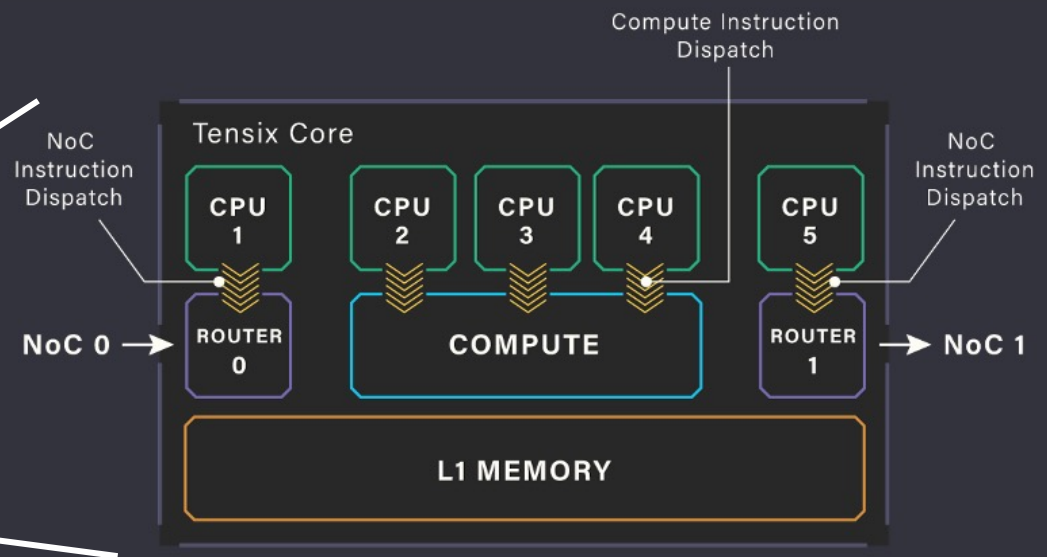
Multiple Generations of AI Accelerators
All Powered by the Scalable High Performance
Tensix AI Core

tenstorrent

Tenstorrent ML Accelerator Architecture



Grayskull: 120 Tensix cores

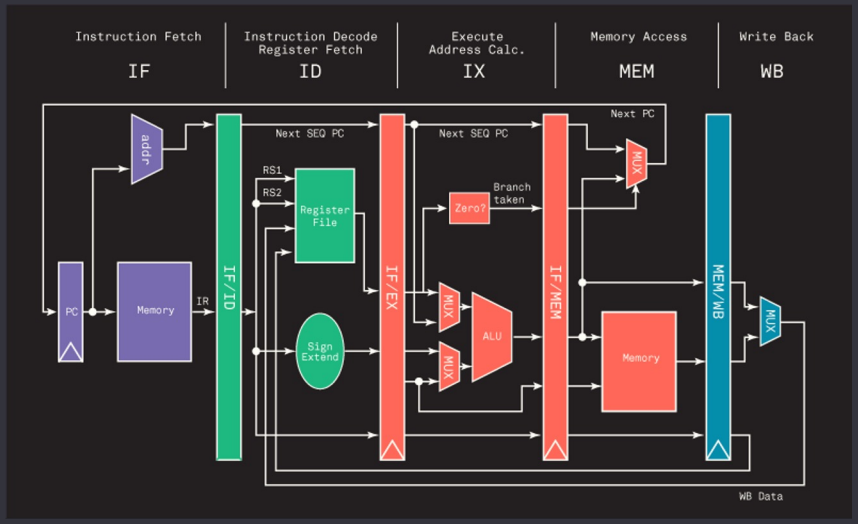


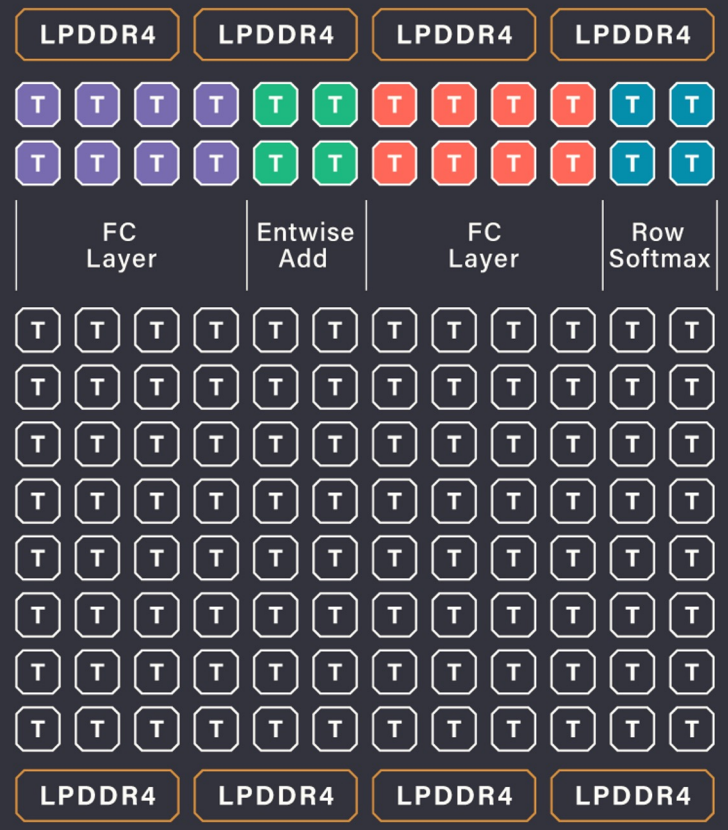
- RISC-V CPUs in a tile support flexible data manipulation
- Example
  - One for NoC receive data
  - One for NoC transmit data
  - Three for Compute engine

tenstorrent

Mapping Operation to Tensix

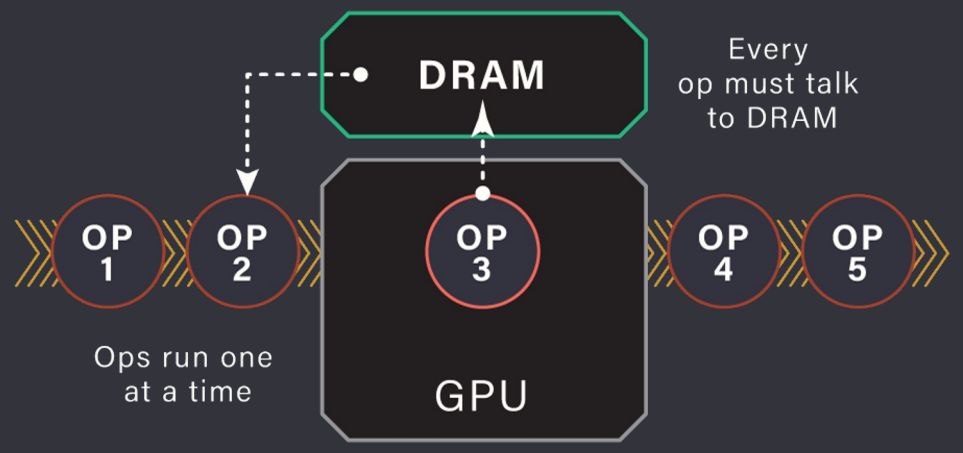
- Processing elements are assigned to each graph node
- Software balances the amount of assigned core per graph node
- Just like in CPU pipelines, the data moves through the pipe



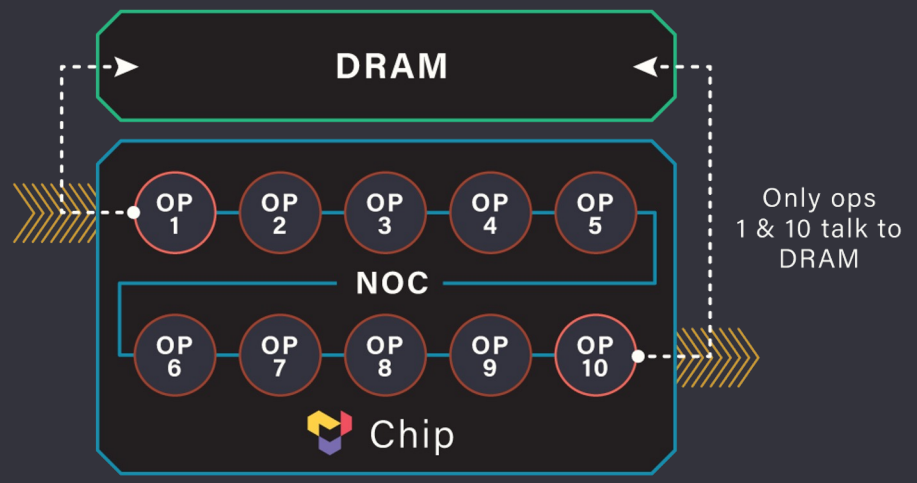


tenstorrent

GPU vs. Tenstorrent Memory Bandwidth



VS.



GPU Memory

A GPU needs to reload op parameters every time it switches to the next op

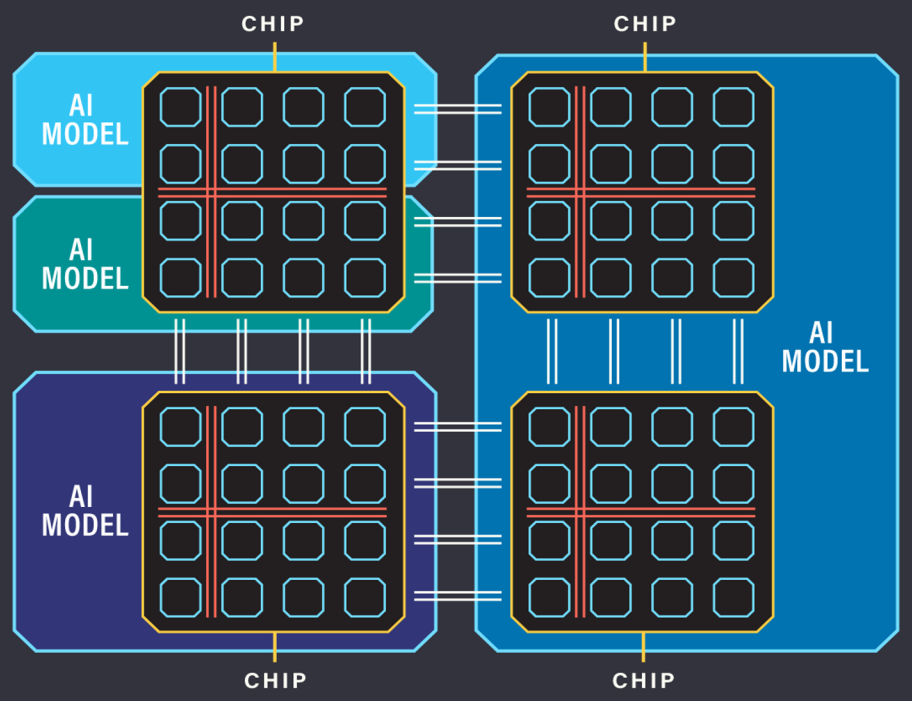
Tenstorrent Memory

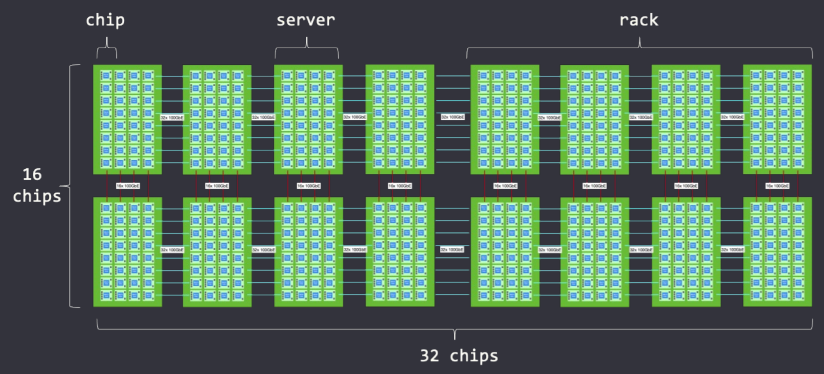
In a Tenstorrent chip, the outputs of the ops feed the next op over the NOC, and only the first and the last op on the chip communicate with DRAM.

tenstorrent

‘Scale-out’ Microarchitecture

- Large AI models can be mapped to multiple chips
- Data flows can go through chip to chip
- Scale-out multiple racks

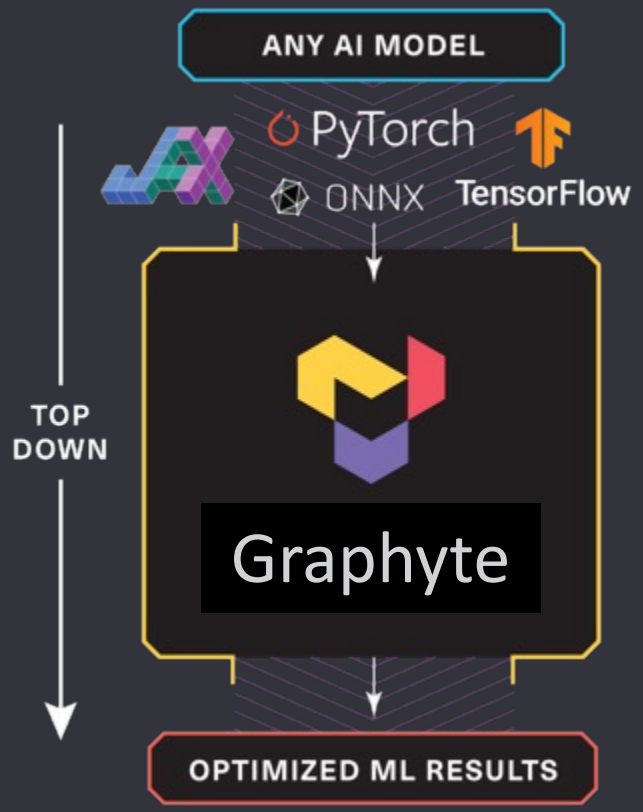




tenstorrent

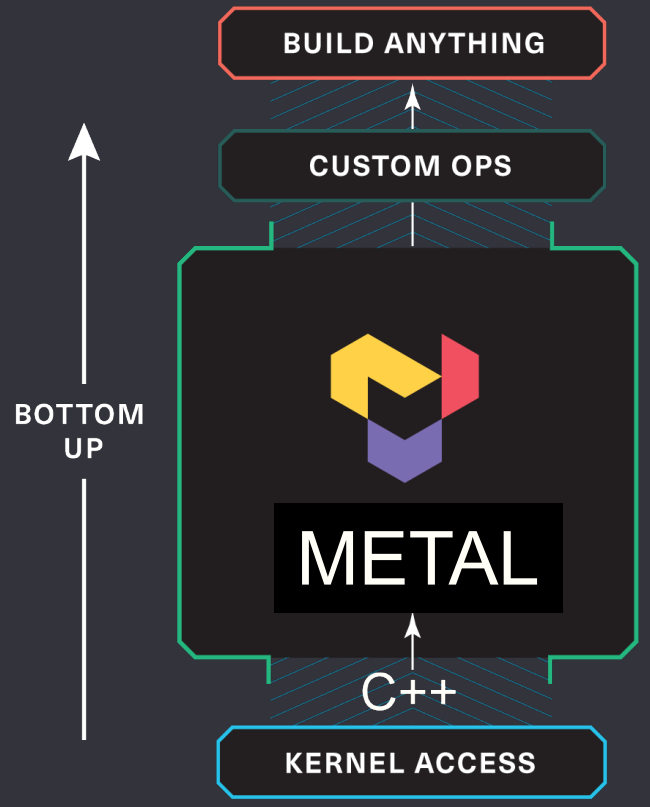
SDK

Tenstorrent Software – Two Distinct Approaches



Graphyte: Run any model right away

Great for production customers who want to get models up and running with ease. They want flexibility, but don't have time to program new operations or contact Nvidia every time something changes.

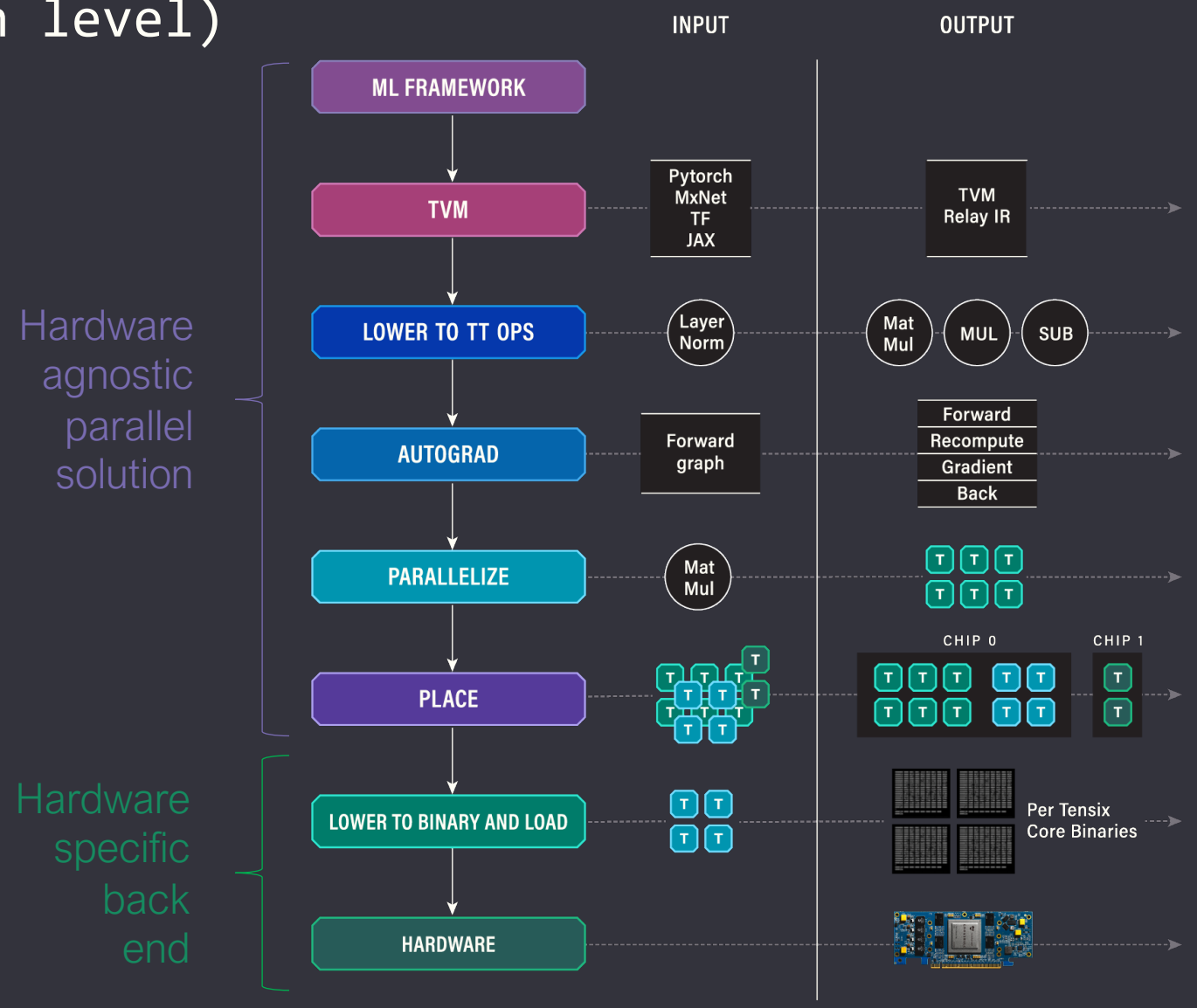


METAL: Open Access to Tenstorrent Hardware & Software

Great for development customers who want to customize their models, write new ones or even run non-machine learning code. No black boxes, encrypted APIs or hidden functions.

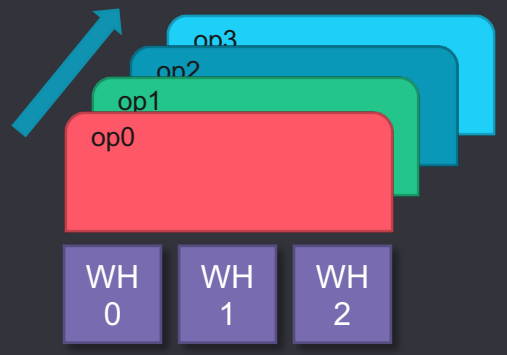
ML compiler stack (high level)

- Fully automated path from all popular ML framework to optimized implementation
- High quality results with no manual effort
- Same compiler targets one chip or many thousands of chips



tenstorrent

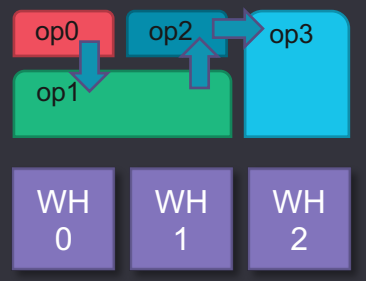
We support Temporal and Spatial Execution



Temporal: Ops run one after the other, each using all available resources

PROS:
- Every op runs as fast as it can

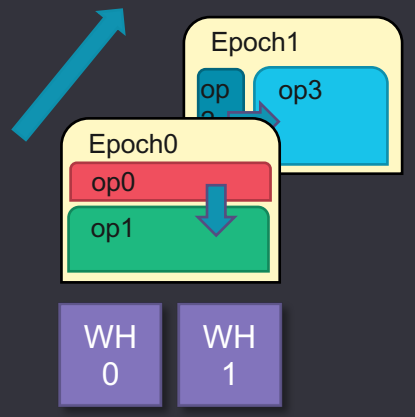
CONS:
- High DRAM overhead in r/w intermediate data
- Reconfigure penalty between ops
- Small ops do not use available hardware efficiently



Spatial: Each op is assigned fixed resources, according to their size. Model runs as a pipeline.

PROS:
- Local memory / cache reuse for model parameters
- Intermediate data travels on NOC
- Zero reconfiguration and intermediate data storage overhead
- Very high utilization of resources

CONS:
- Large models could require large number of devices to fit
- Pipeline overhead for fill/drain



Temporal/Spatial Mix: Model is divided into "epochs", where each epoch spatially fits within available hardware. Epochs run sequentially. Small ops are fused into larger ones.

PROS:
- Fits on available hardware
- High utilization within epochs due to spatial computation
- Reconfigure & intermediate storage overhead limited due to a smaller of number of transitions

tenstorrent

Running Huggingface Inference in Pytorch vs. Graphyte

```python
from transformers import BertForQuestionAnswering, BertTokenizer
import torch

# Load Bert tokenizer and model from HuggingFace
tokenizer = BertTokenizer.from_pretrained("bert-large-uncased-whole-word-masking-finetuned-squad")
model     = BertForQuestionAnswering.from_pretrained("bert-large-uncased-whole-word-masking-finetuned-squad")

# Load data sample
question, context = "Who was Jim Henson?", "Jim Henson was a nice puppet"

# Data preprocessing
input_tokens = tokenizer.encode(question, context, max_length=128, padding="max_length", return_tensors="pt")

# Run inference on CPU
with torch.inference_mode()
    output = model(input_tokens)

# Data postprocessing
answer_start_index = output.start_logits.argmax()
answer_end_index   = output.end_logits.argmax()

# answer = "nice puppet"
answer = tokenizer.decode(input_tokens[0, answer_start_index:answer_end_index+1], skip_special_tokens=True)
```

Running HF with PyTorch

```python
from transformers import BertForQuestionAnswering, BertTokenizer
import pybuda

# Load Bert tokenizer and model from HuggingFace
tokenizer = BertTokenizer.from_pretrained("bert-large-uncased-whole-word-masking-finetuned-squad")
model     = BertForQuestionAnswering.from_pretrained("bert-large-uncased-whole-word-masking-finetuned-squad")

# Load data sample
question, context = "Who was Jim Henson?", "Jim Henson was a nice puppet"

# Data preprocessing
input_tokens = tokenizer.encode(question, context, max_length=128, padding="max_length", return_tensors="pt")

# Run inference on Tenstorrent device
output_q = pybuda.run_inference(pybuda.PyTorchModule("bert_large_qa", model), inputs=[(input_tokens,)])
output   = output_q.get(timeout=0.5)

# Data postprocessing
answer_start_index = output[0].value().argmax().item()
answer_end_index   = output[1].value().argmax().item()

# answer = "nice puppet"
answer = tokenizer.decode(input_tokens[0, answer_start_index:answer_end_index+1], skip_special_tokens=True)
```

Running HF with Graphyte

tenstorrent

Over 40 standard models running

NLP
BERT
GPT-2
BART
T5
ALBERT
RoBERTa
DistilBERT
GPT Neo
GPT-J
OPT
XGLM
XLM
SqueezeBERT

Computer Vision	
ResNet	ResNeXt
YOLOv5	VideoPose
ViT	VGG
DenseNet	HRNet
U-Net	MNIST
MobileNetV3	DeepCoNN
MobileNetV2	DALLE VAE
MobileNetV1	ConvNeXt
EfficientNetV2	GhostNet
EfficientNet	FCN
VoVNet	OpenPose

Other
Wav2Vec
UniSpeech
ViLT
NBeats
DeepFM

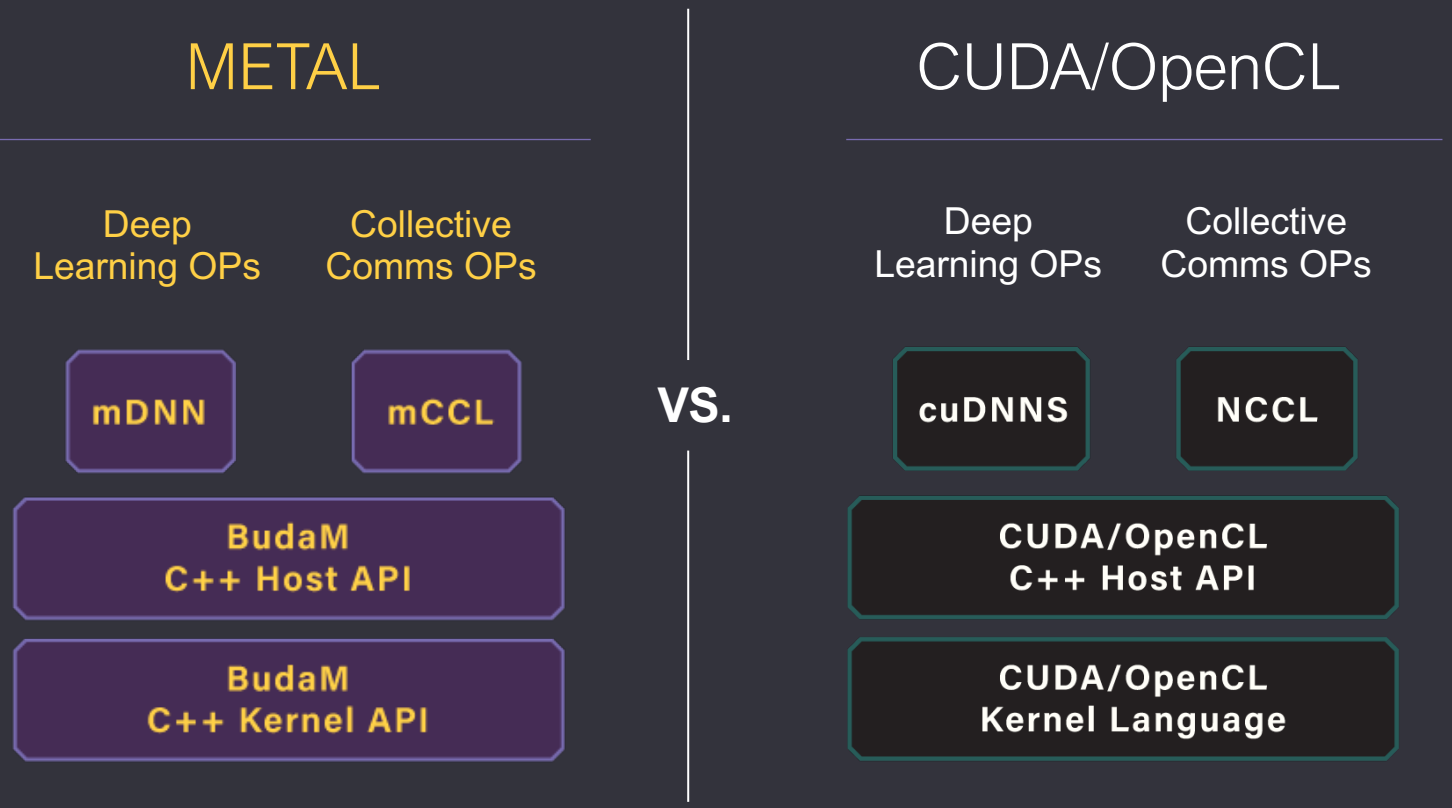
With another 50 in the pipeline…

tenstorrent

METAL Delivers More Control and Capabilities than CUDA

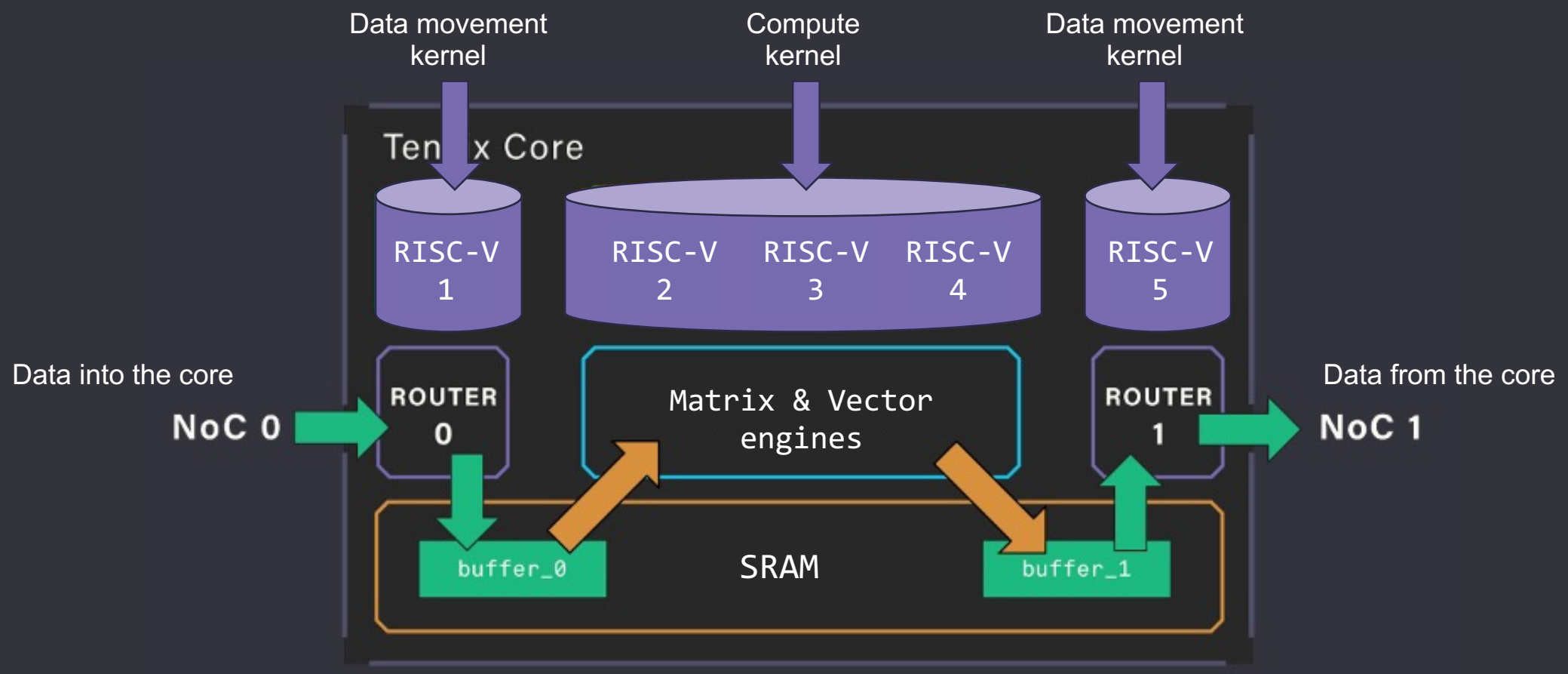
METAL Advantages over CUDA:

1. Kernels are pure C++ with APIs
2. De-coupled data movement & compute – optimize compute and data movement separately
3. Different Cores can run different kernels with pipelines connecting them
4. Direct Control of SRAM and DRAM



tenstorrent

Metal ”Read->Compute->Write” kernels running on a Tensix Core



tenstorrent

AI HW scalability : Datacenter ~ Edge

Auto/Electric industries

Tenstorrent、LGと提携し、将来のスマートテレビ向けにAIとRISC-V Chipletsを構築

テンストレント

2023年5月31日 09時00分

2

TenstorrentとLG Electronics Inc.（LG）は、新世代のRISC-V、AI、ビデオコーデックのチップレットを構築するために協力することを発表しました。これらは将来のLGプレミアムTVや車載製品に加え、Tenstorrentのデータセンター製品にも搭載される可能性があります。

革新的なコンシューマ電機及び家電製品製造のグローバルリーダーであるLGは、今回の協業を通じてTenstorrentからAIとRISC-V CPU技術の提供を受けることになります。これらの技術は、LGの将来の製品群 - プレミアムTV、高性能自動車用チップ、およびその他のスマート製品 - において、AIによる機能強化と高性能コンピューティングの導入に最適な技術となります。

業界を牽引するTenstorrentの革新的なAIとRISC-V CPUの技術は、LGの技術ポートフォリオを豊かにし、競争の激しい市場においてチップソリューションの差別化を可能にします。業界のベテランであり、伝説的なCPUアーキテクトであるCEOジム・ケラーが率いるTenstorrentは、LGが自社のシリコンロードマップをコントロールするた

ＡＩ半導体開発テンストレント、現代自などから1億ドル調達

By Reuters Staff

1 MIN READ

半導体業界のベテラン、ジム・ケラー氏が率いるカナダの人工知能（ＡＩ）半導体開発新興企業テンストレント（写真左）は２日、韓国の現代自動車グループやサムスン電子の投資ファンドなどから１億ドルを調達したと発表した。写真は８月２日にテンストレントが公表（２０２３年　Courtesy of Tenstorrent /Handout via REUTERS）

[２日 ロイター] - 半導体業界のベテラン、ジム・ケラー氏が率いるカナダの人工知能（ＡＩ）半導体開発新興企業テンストレントは２日、韓国の現代自動車グループやサムスン電子の投資ファンドなどから１億ドルを調達したと発表した。

同社は今回までに２億３４５０万ドルを調達しており、その際の評価額は１０億ドルだった

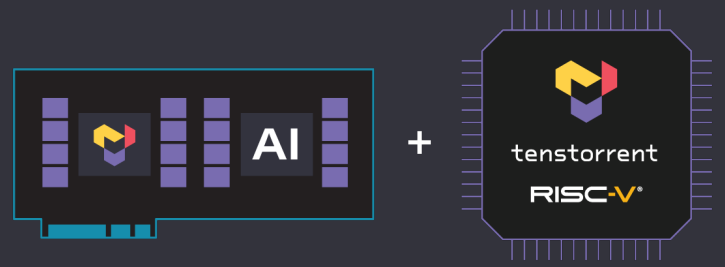
tenstorrent

Tenstorrent’s Offerings

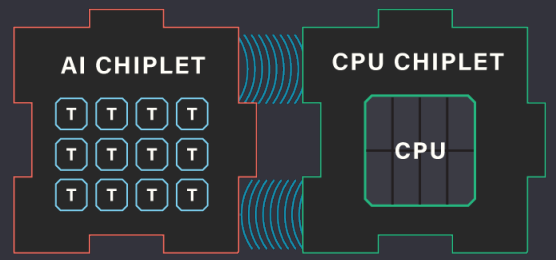
Product	IP	Chiplet	Chip	Card	Systems	Cloud
 RISC-V	✓	✓				
 AI Acc	✓	✓	✓	✓	✓	✓

tenstorrent

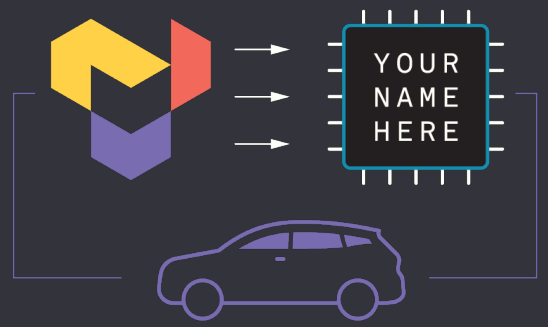
Tenstorrent Example Vertical: Automotive



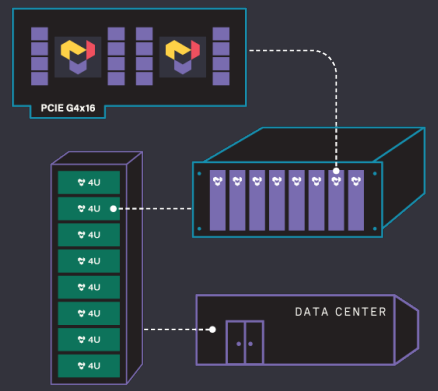
Tenstorrent AI and RISC-V IP deliver the compute power that ADAS and IVI require



Chiplet approach reduces cost while accelerating design and production schedules.



Automotive companies can own their own silicon working with Tenstorrent



Power Consumption is critical: Tenstorrent technology scales from MW to mW

tenstorrent

Summary

- Scaleable Tensix cluster + reference server appliance
- SDK handles complexity and parallelis
- Edge deployment, Training in datacenter with same Framework, Architecture

Thank You!

tenstorrent