Advances in Processor Architecture Driving HPC/AI Convergence for Next-Generation Exascale Systems

November 2022

Tachyum™

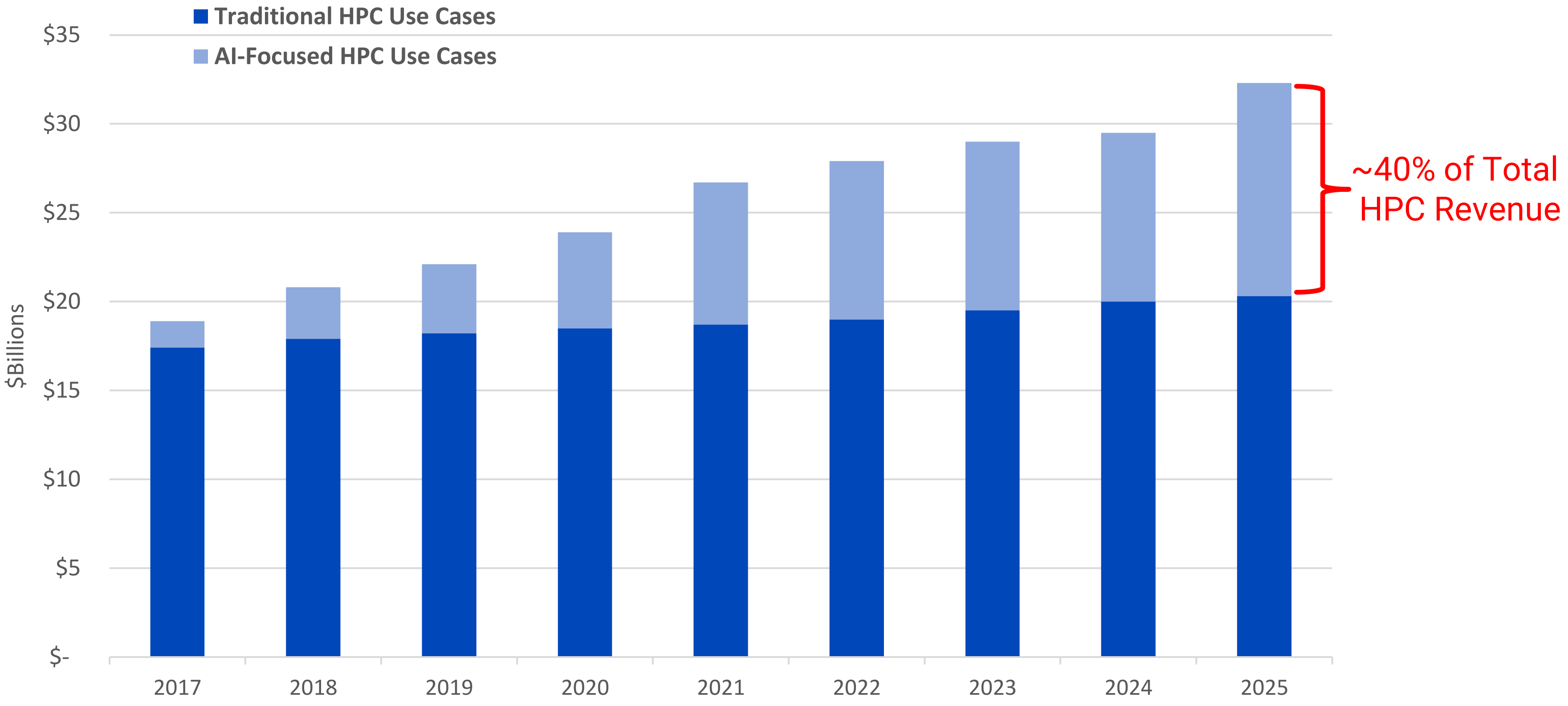
Robert Reiner

Director of Product Marketing

Worldwide HPC Revenue

2017 - 2025

AI-Focused HPC Use Cases to Account for ~40% of Total by 2025



Source: Tractica

Trends Driving HPC/AI Convergence

Key AI Applications are Growing in the HPC Space

- Simulation steering with trained AI models
- Data preparation and cleansing
- Training Neural Networks to do Simulations

Emerging Government Requirements

- Recent Dept. of Energy Request for Information for 2025 and beyond specifies both HPC and AI performance projections in a converged environment

Commercial IT Convergence

- IT departments moving away from disparate architectures for HPC and AI to reduce TCO
- Keeps common data localized

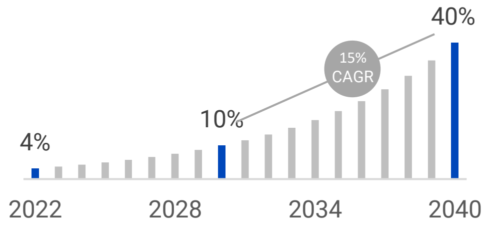
HPC vs. AI

Workload Characteristic	HPC	AI/ML
High Performance Parallel Processing	Very Important	
FP Precision	High Precision	Low Precision
Vector vs. Matrix Processing	HPC typically uses vectors	Deep learning typically uses matrixes
Sparsity and Quantization	Not Used	Very Important to Optimize Performance and Memory Footprint
Memory Bandwidth	Very Important	
Memory Latency	Important to the extent it affects effective bandwidth	
Scalable Processor and Memory	Very Important	
Cost and Power Efficient	Very Important	

Serious Issues Facing Data Centers

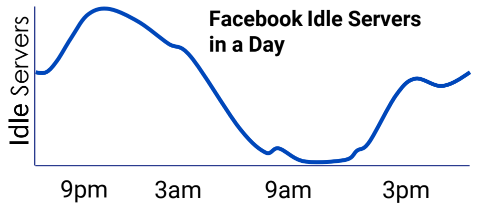
Data Center Power Consumption

- Currently data centers consume ~4% of the planet's power
- At ~15% annual growth this becomes a serious problem
- Power consumption could limit data center expansion



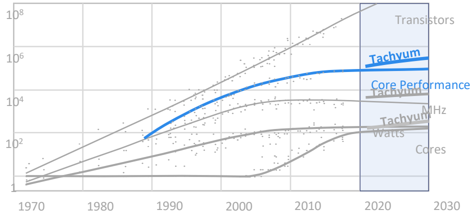
Low Server Utilization

- Average server utilization is frequently less than ~30%
- Facebook's study: <50% server utilization per 24-hours
- Low server utilization costs billions of dollars per year


Facebook Idle Servers in a Day

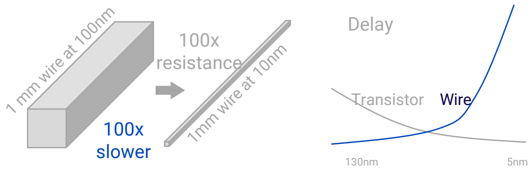
Performance Plateau and Moore's Law

- Performance increase of processors has slowed down
- Moore's law no longer holds with process shrinks



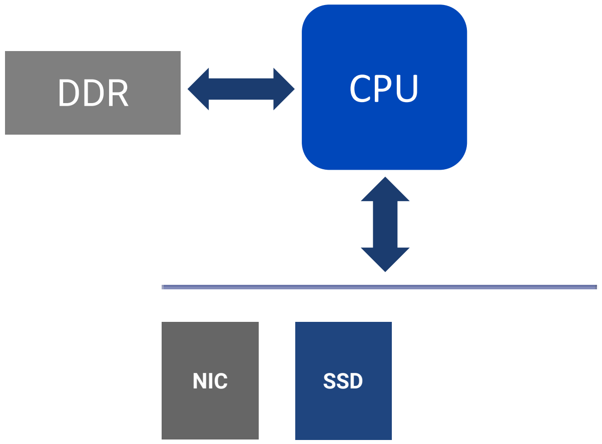
Wires Are Slower as Process Shrinks

- With process shrink transistors are faster but wires are slower
- 10x smaller process would results in 100x slower wire
- Using copper and low-K materials reduced slow down to ~20x
- Wire delays are now limiting performance of functional blocks



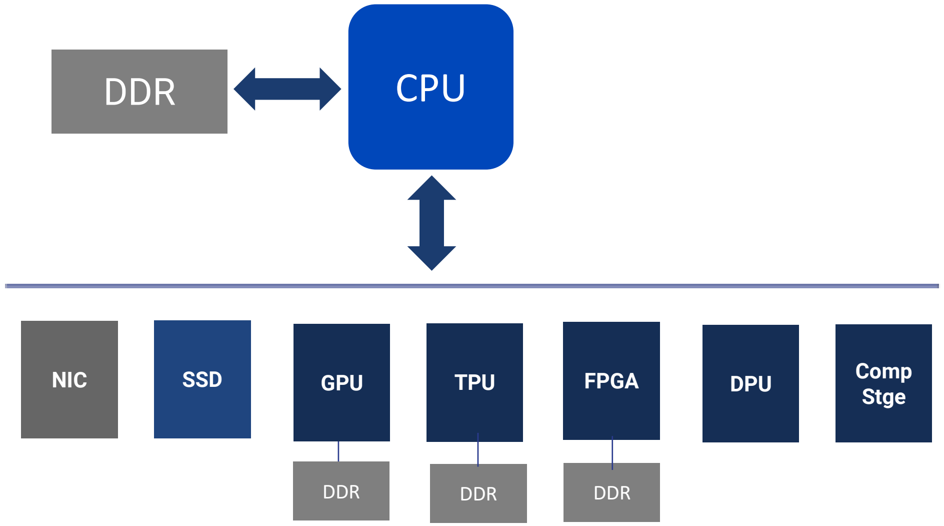
Homogeneous vs. Heterogeneous Systems

Homogeneous



Pros	Cons
• General Purpose, Flexible • Easy Deployment/ Maintenance	• Not Designed for HPC or AI • Low Parallel Performance for Modern Workloads

Heterogeneous

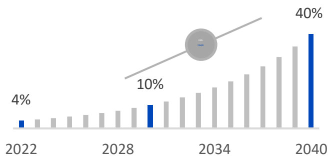


Pros	Cons
• Accelerates specific workloads, including HPC and AI • Scalable	• Needs special programming • Expensive, power-hungry • Under-utilized – contrary to software-defined data center

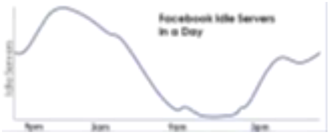
Tachyum Prodigy – The World’s First Universal Processor

Problems

Data Center Pain Points

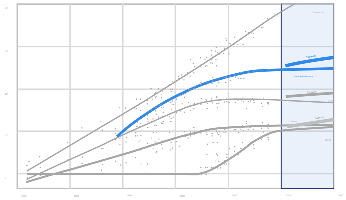


Data Center Power Consumption

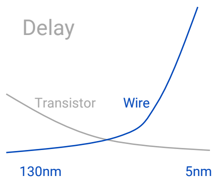


Low Server Utilization

Industry Transformation

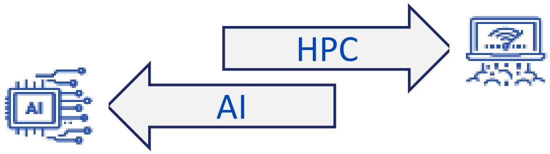


Performance Plateau

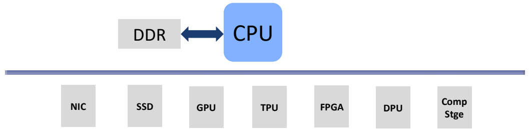


Slow Wires

HPC/AI Divergence



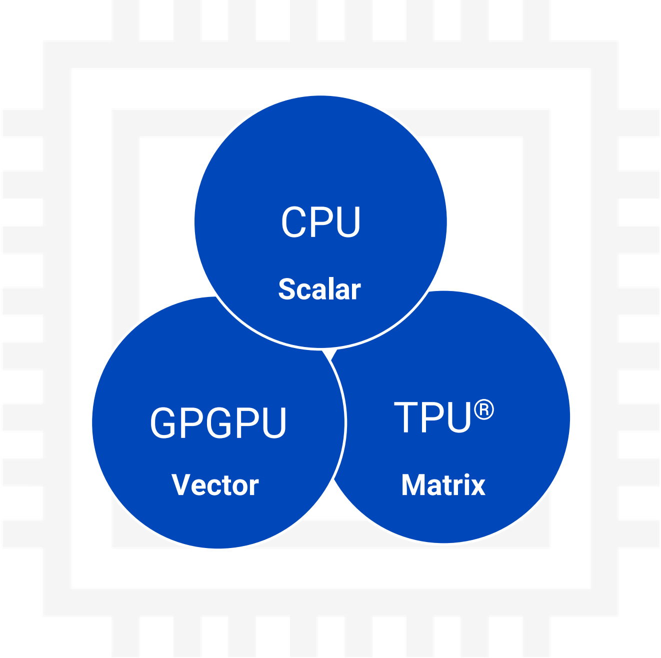
Accelerator Sprawl



Solution

Tachyum Prodigy Cloud / AI / HPC Supercomputer Chip

Unifies the Functionality of CPU, GPU, and TPU®

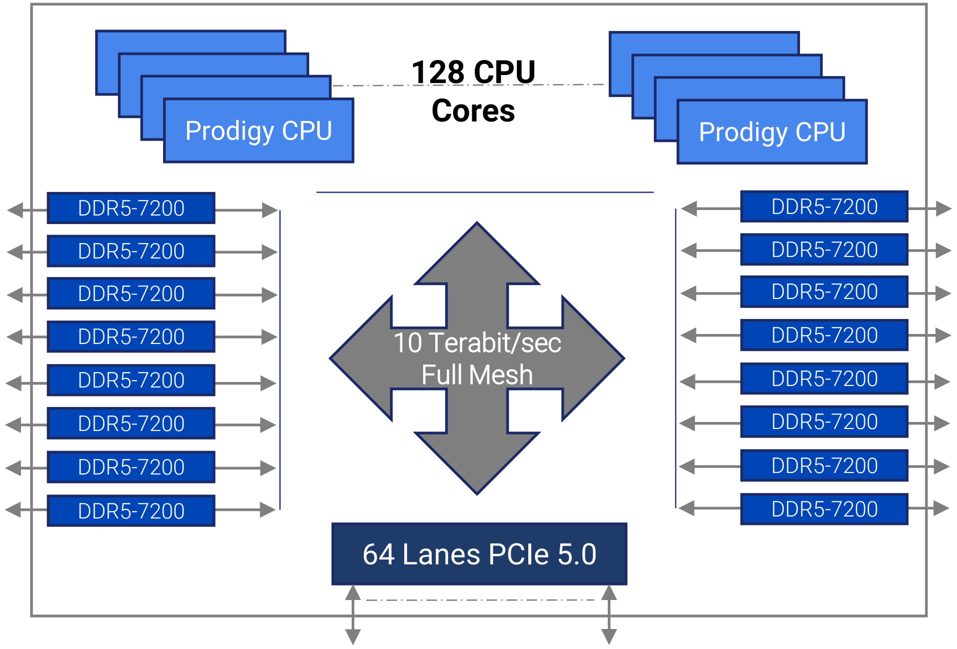


- Over 3x performance of Xeon
- Up to 10x performance at same power
- Faster than NVIDIA H100 in HPC and AI

Prodigy Feature Summary

High Performance CPU – HPC and AI for Free

High-Performance Processor	• 128 Custom-designed 64-bit cores running at 5.7+ GHz
	• Hardware Coherency Supports 2 and 4-socket Systems
High-Throughput Memory and I/O	• 16 DDR5-7200+ Memory Controllers
	• 1TB / 2TB* of Memory Bandwidth (2-4x of x86)
	• 64 Lanes of PCIe 5.0
Advanced Process	• 5nm Process Technology
Emulation for Other ISAs	• Runs Native and x86, Arm, and RISC-V Binaries
HPC and AI Features	• 2 x 1024-bit Vector Units per Core
	• 4096-bit Matrix Processors per Core
	• FP64, FP32, TF32, BF16, Int8, FP8, TAI Data Types
	• Sparse Matrix Multipliers Optimizes Efficiency
	• Quantization Support Using Low Precision Data Types
	• Scatter/Gather for efficient storing and loading matrices



Samples 3Q, 2023

* Bandwidth Amplification Technology

Prodigy Core Architecture

High Throughput Pipeline

- Fetch and decode up to 8 instructions per clock
- 8 wide x 6 deep instruction queue

Advanced Functional Units

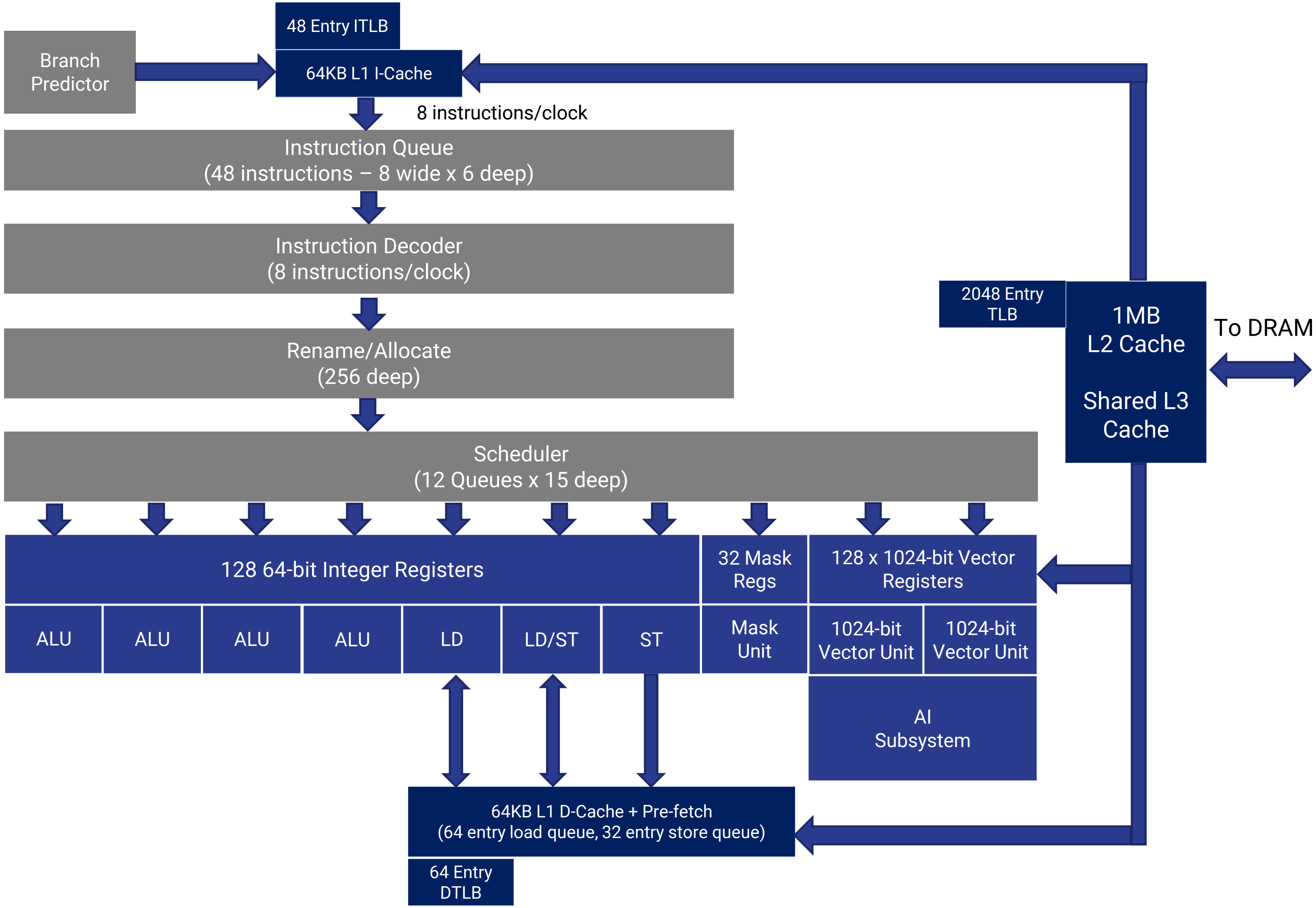
- 4 ALUs
- 1 LD, 1 LD/ST, 1 ST
- 2 x 1024-bit Vector Units

High Performance Cache

- 64KB I-Cache
- 64KB D-Cache
- 1MB L2 Cache
- Shared L3 Cache up to 128MB
  - L2 from Idle cores available as L3

RAS Features

- I-Cache, D-Cache: SECDED
- L2 Cache: DECTED


48 Entry ITLB
Branch Predictor
64KB L1 I-Cache
8 instructions/clock
Instruction Queue (48 instructions – 8 wide x 6 deep)
Instruction Decoder (8 instructions/clock)
Rename/Allocate (256 deep)
Scheduler (12 Queues x 15 deep)
2048 Entry TLB
1MB L2 Cache
Shared L3 Cache
To DRAM
128 64-bit Integer Registers
32 Mask Regs
128 x 1024-bit Vector Registers
ALU
ALU
ALU
ALU
LD
LD/ST
ST
Mask Unit
1024-bit Vector Unit
1024-bit Vector Unit
AI Subsystem
64KB L1 D-Cache + Pre-fetch (64 entry load queue, 32 entry store queue)
64 Entry DTLB

* Technology to be disclosed at a future date

Matrix / Vector Processing Built from the Ground Up - ***Not Bolted On***

Prodigy Treats Vectors and Matrices As 1st Class Citizens

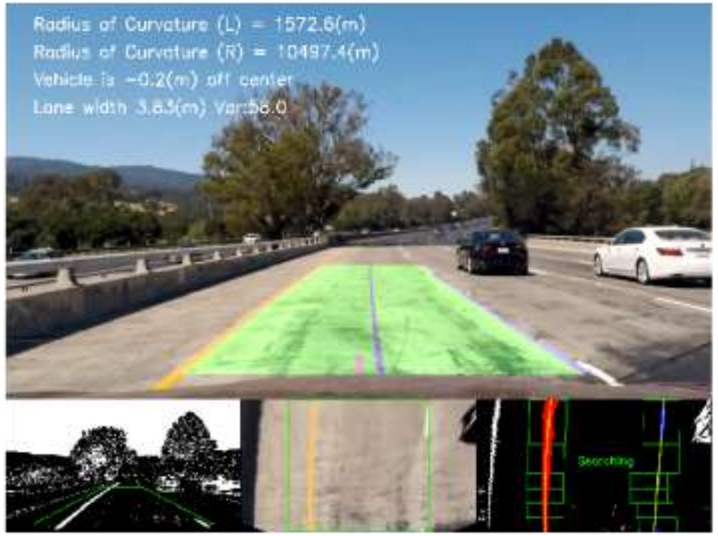
		CPUs		GPUs		
Feature	**Tachyum Prodigy**	**intel 8380**	**AMD 7763**	**NVIDIA H100**	**AMD MI250**	**Comments**
Support for FP8	✓			✓		High performance for training and inference
Support for TAI	✓					Increases performance and reduces memory utilization
2 x 1024-bit Vector Units	✓			N/A	N/A	• Prodigy 2x wider than Intel 2x512 vector units • Prodigy 4x wider than AMD 2 x 256 vector units
No Penalty for Misaligned Vector Loads/Stores	✓			N/A	N/A	Intel AVX-512 misaligned LOAD/STORE at half speed
AI Sparsity Support	✓			✓		
Super-Sparsity Support	✓					
Native Matrix Support	✓	*		✓	✓	* Intel matrix support is off the main execution path

Tachyum Prodigy Software Ecosystem

Applications	- Broad range of applications compiled to run natively on Prodigy	RabbitMQ, git, APACHE, gRPC, SQLite, slurm, Vim, mongoDB, MariaDB, RocksDB, POSTFIX, DOVECOT
Programming Languages	- Prodigy supports a broad spectrum of programming languages encompassing a wide array of applications and workloads	GO, php, python, F, Java, ERLANG, Lua, R, Perl, C
Frameworks & Libraries	- Support for major AI frameworks and scientific libraries for cutting-edge matrix and vector performance	TensorFlow, PyTorch, TensorFlow Lite, Google, QUANTUMESPRESSO, SELinux, LAMMPS, YOLO
System Software	- GCC, Linux and FreeBSD are ported to Prodigy along with the GNU libraries	KVM, FreeBSD, GCC, SELinux, SAMBA, docker, OpenMP, OpenBMC, XJTAG
Software Roadmap	- Tachyum's roadmap adds key applications for big data, containers, and virtualization	LLVM, ceph, U-Boot, Xen Project, Valgrind, APACHE Spark, OpenJDK, coreboot, LinuxBoot
Emulation	- SW Emulation with QEMU and C-model - Prodigy Hardware FPGA Emulation - Prodigy Runs x86, Arm, & RISC-V binaries	QEMU x86 Arm RISC-V

Scaling Deep Learning

- Prodigy addresses continuing trends in AI models, explosion in complexity as demanded by more complex NLP models and more accurate conversational AI.
- NLP transformer models (BERT, GPT-3, Megatron …) requires **billions of parameters**
- Computer vision models (ResNet-50, Fast R-CNN, SSD) requires **real-time processing of 4k video**
- Training these massive models in FP32 precision can take **days or even weeks**


Radius of Curvature (L) = 1572.6(m)
Radius of Curvature (R) = 10497.4(m)
Vehicle is -0.2(m) off center
Lane width 3.83(m) Var:56.0
Searching

Tachyum’s Solution:

- providing **native low precision datatypes (bf16, int8, fp8 …)**
- matrix multipliers utilizing low precision data types deliver an **order-of-magnitude higher performance**
- sparse matrix multipliers pushing the performance
- **16 DDR5 interfaces** to maximize memory bandwidth and capacity

Quantization and Mixed Precision Training

Quantization

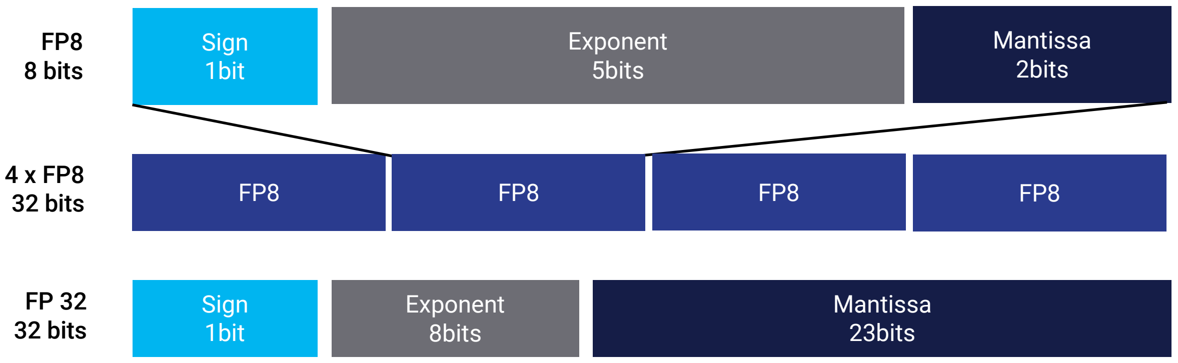
- Reduces memory footprint and inference time of Neural Networks
- Reduces numerical precision of both the weights and the operations in the network

FP8 Compared to FP32

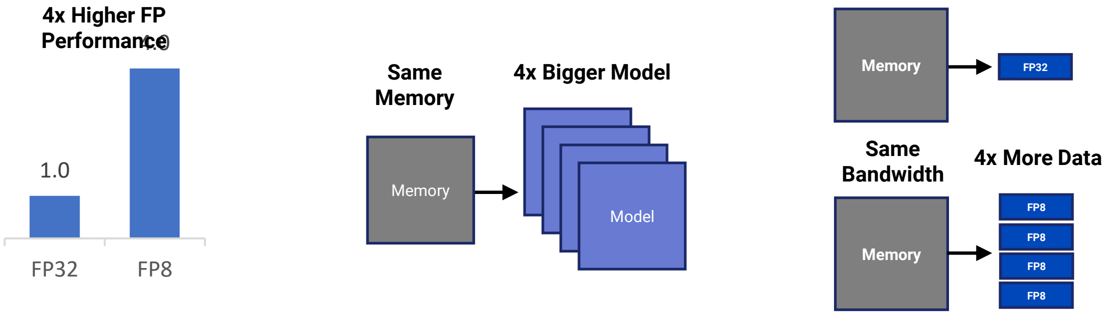
- 4x higher performance
- 4x memory reduction
- 4x higher memory bandwidth efficiency

Prodigy Mixed Precision Training using FP8

- FP8 used for all arrays
  - Weights, activations, errors, and gradients
- GEMM operations accumulate to BF16
- Master copy of weights stored in BF16



FP8 vs. FP32



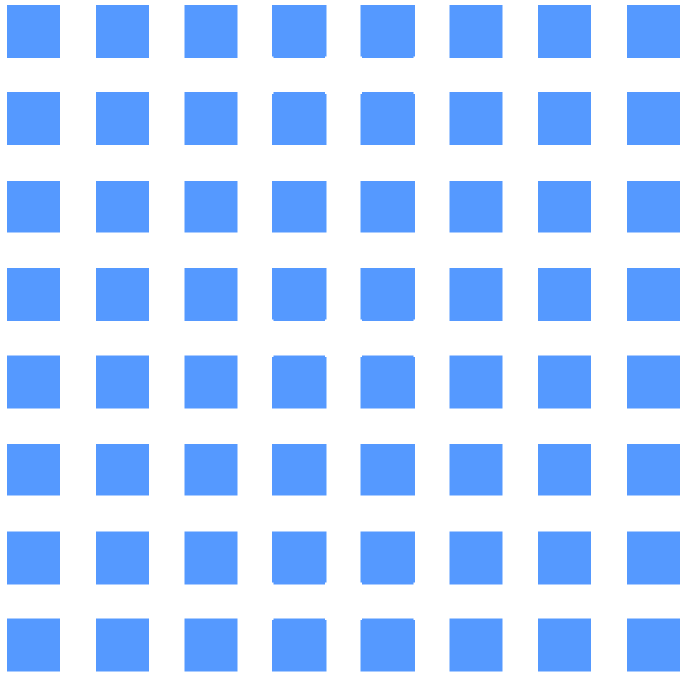
Sparsity and Super-Sparsity

Sparsity

- Pruning or compression of neural networks is another important approach for scaling deep learning
- Prodigy supports block structured sparsity, which Reduces memory and computation requirements
- Prodigy incorporates special instructions to efficiently store, load, and multiply sparse matrices

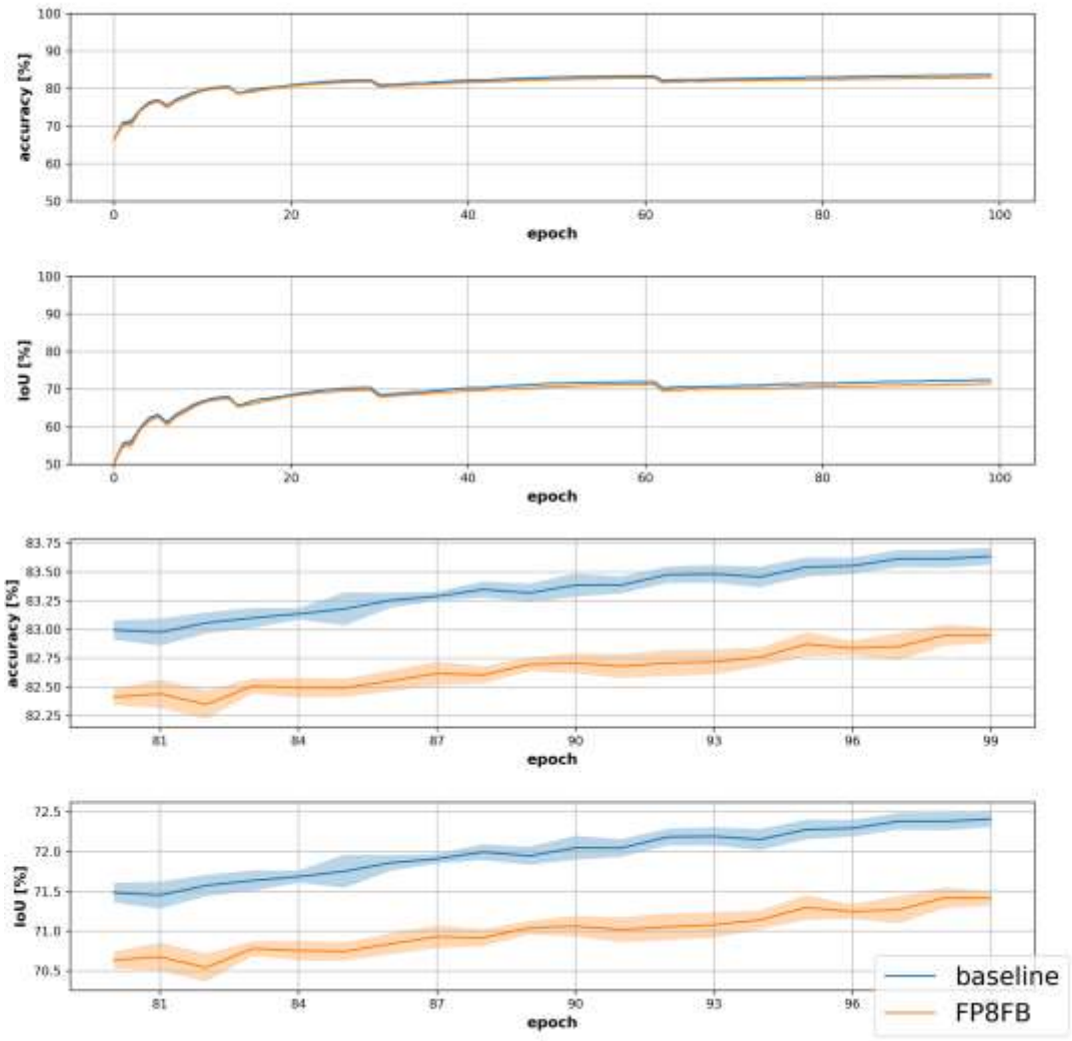
Prodigy Sparse Matrix Multipliers

- Sparsity
  - 4:2 compression ratio
  - Currently supported by others in the industry
- Super-Sparsity
  - 8:3 compression ratio
  - Introduced by Tachyum
  - Maximizes compute and memory efficiency

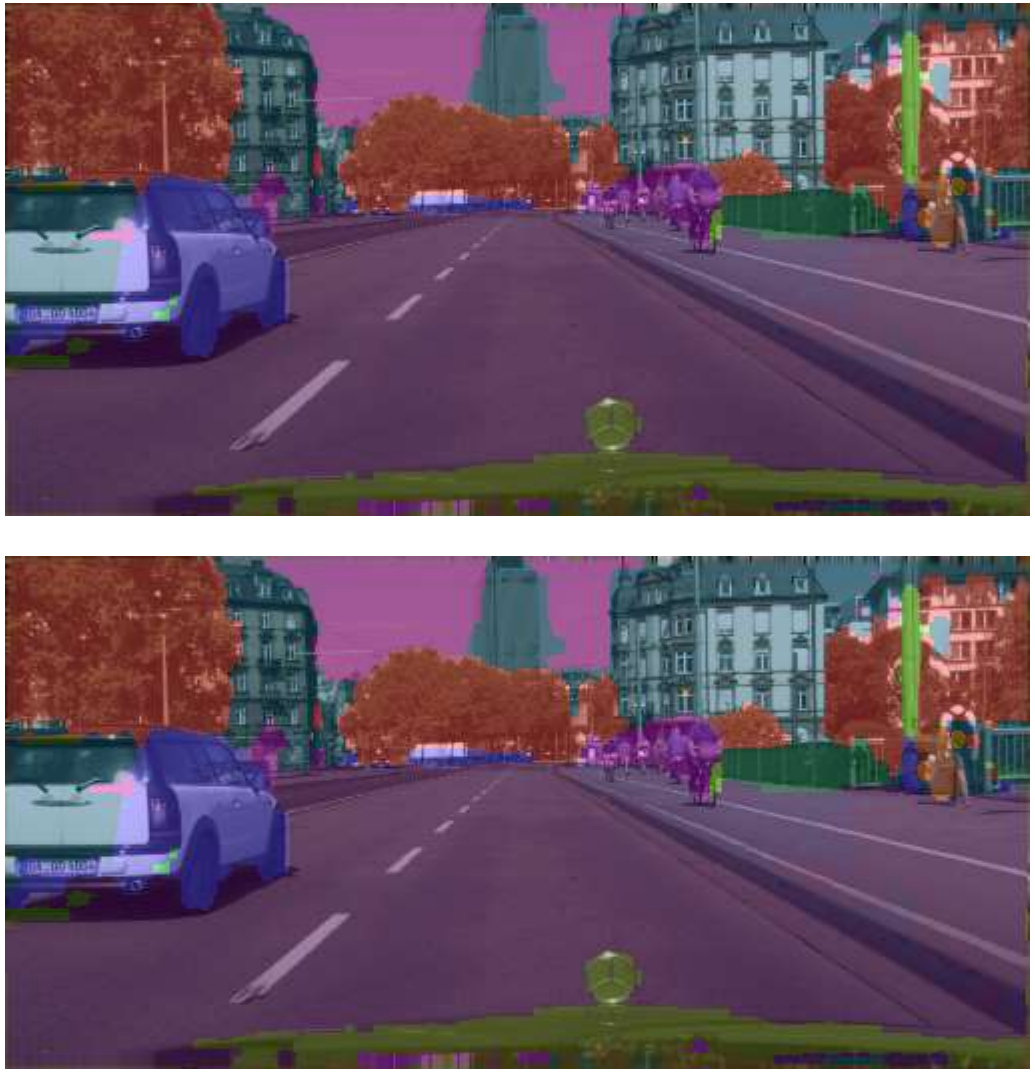


FP8 Instance Segmentation – ConvMixer

IoU** FP32 72% vs **FP8 71.5%

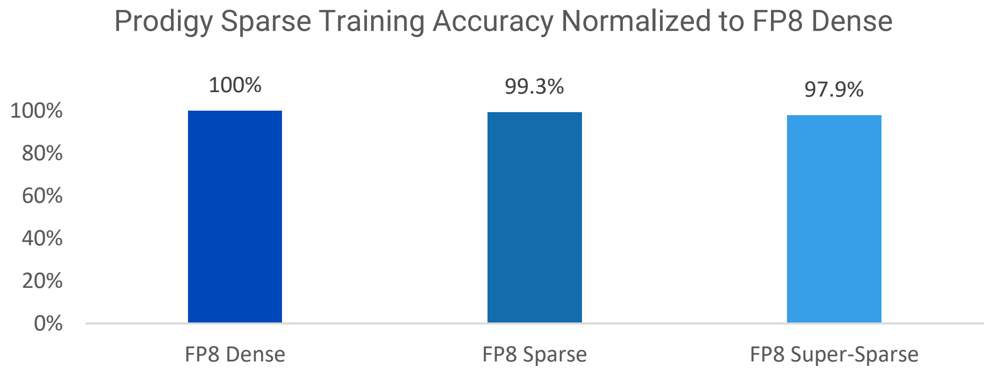


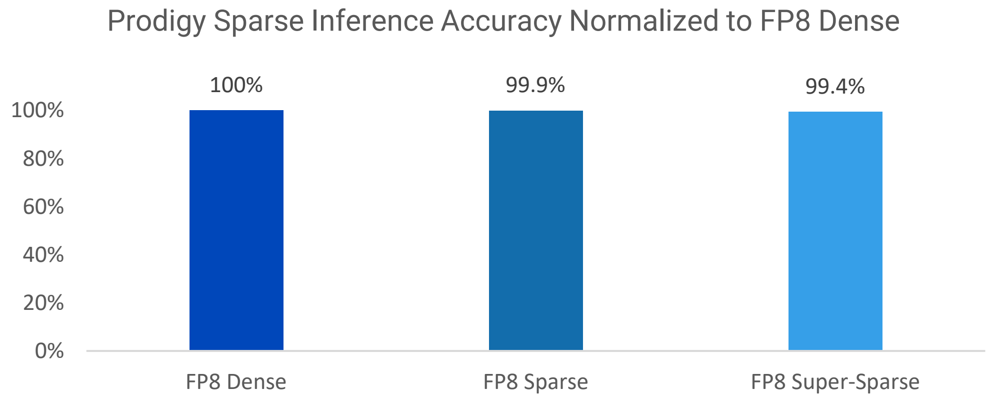
Target vs Predicted

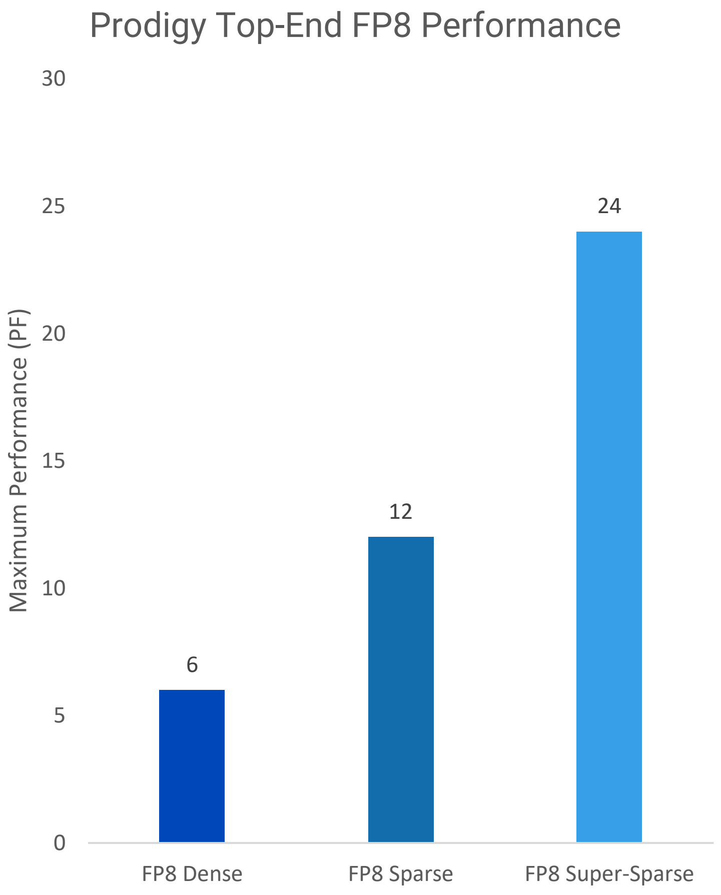


Scaling Deep Learning – Sparsity and Super-Sparsity

FP8 Quantized Resnet20 Model on CIFAR 10







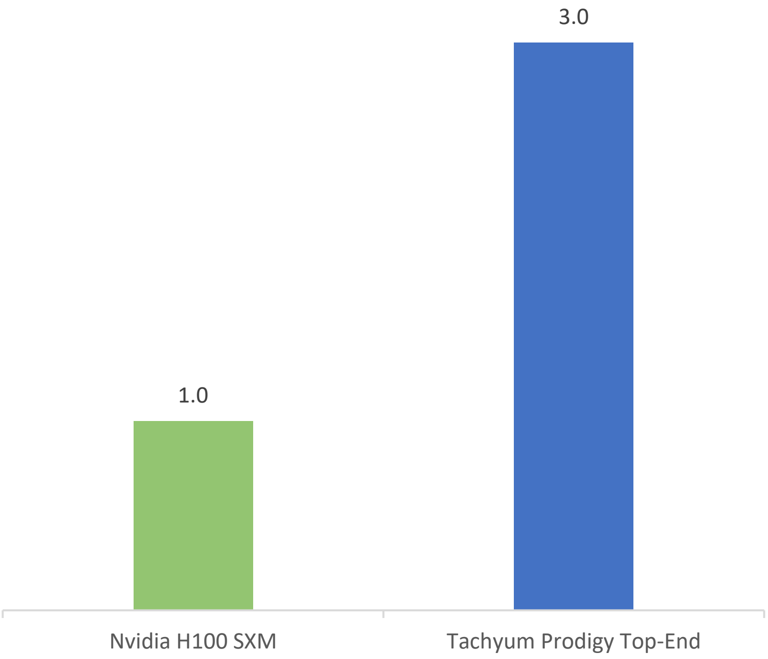
Super-Sparsity Performance **4x Greater** than Dense with Relatively Small Degradation in Accuracy

Prodigy vs. Nvidia H100

HPC and AI Performance

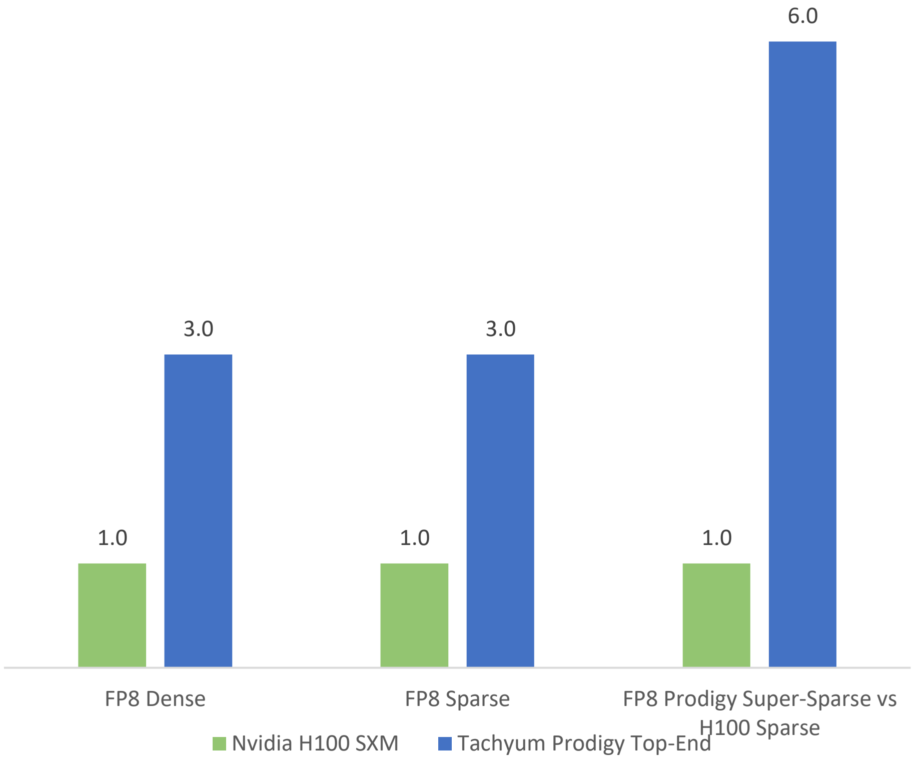
Tachyum Prodigy Performance
Normalized to H100

FP64: Prodigy Top-End vs. Nvidia SXM



Prodigy **3x Higher HPC** Performance

FP8: Prodigy Top-End vs. Nvidia H100 SXM



Prodigy **3x - 6x Higher AI** Performance

Prodigy Evaluation Platform

High Scalability with Multiple Configurations

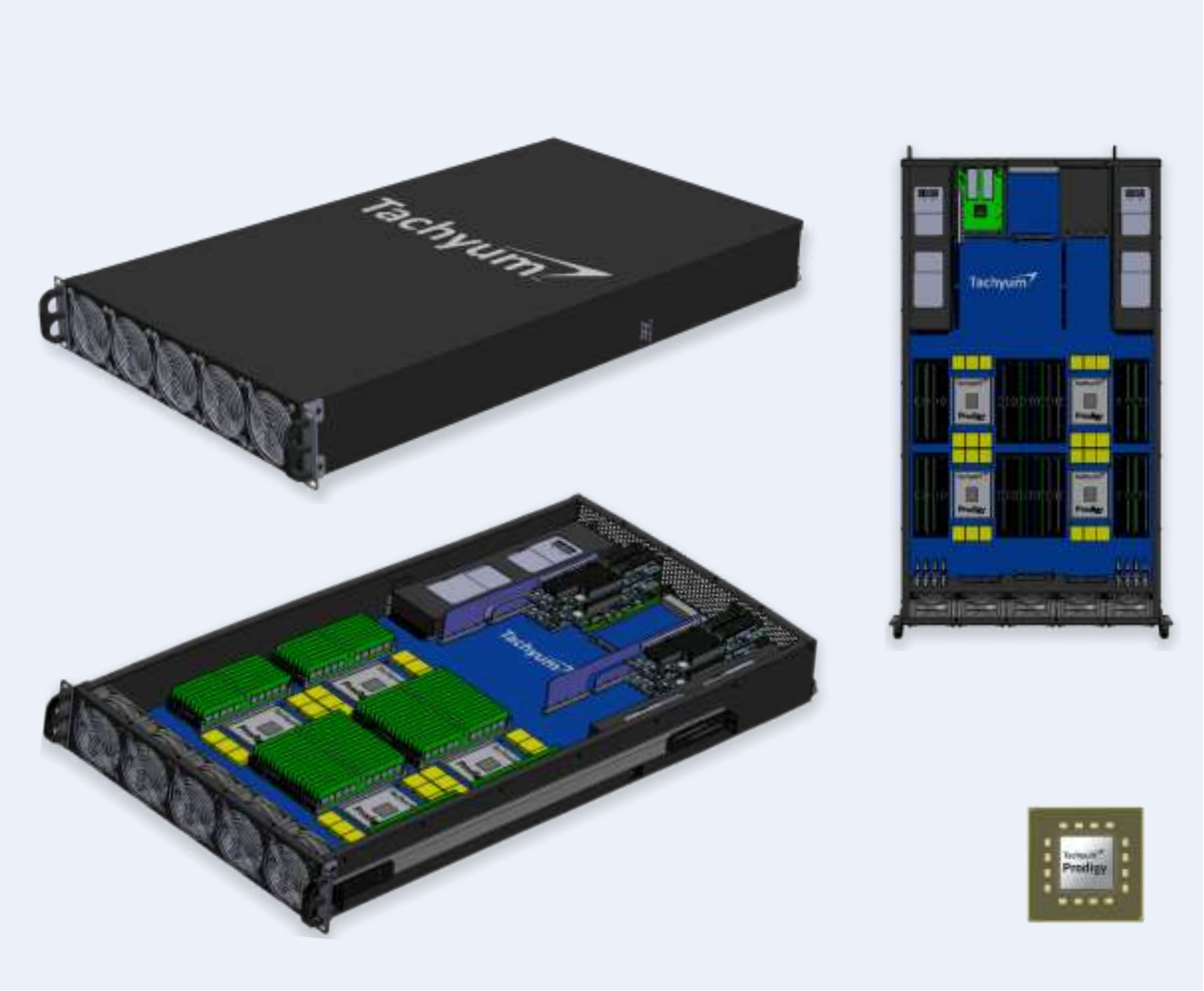
- 128, 64, and 32-core devices running up to 5+ GHz
- 4-socket and 2-socket hardware coherent multiprocessor configurations in addition to single socket
- PCIe 5.0 slots support standard and OCP form factors

Leading-Edge Memory Subsystem Provides Large Footprint for AI Processing

- Up to 64 DDR5 DIMM Modules
- Up to 64 TB memory capacity with 1TB DIMMs by 2024
- Increases to 128 TB with availability of 2TB DIMMs
- FP8 with super-sparsity in 128 TB is equivalent to 512 TB legacy model

Simple Out-of-the-Box Evaluation

- Powerful SDK includes Tachyum Linux, gcc compiler, and wide array of software libraries
- Runs native and x86, Arm, and RISC-V binaries
- Large software ecosystem of applications that have been compiled to run natively on Prodigy



Single Prodigy Platform can Process NLP Models in Memory – **Big AI**

Summary

Prodigy Feature	HPC	AI/ML
High Performance Parallel Processing	✓	✓
Range of Floating-Point Precision	✓	✓
High Performance Vector and Matrix Operations	✓	✓
Support for Quantization and Mixed-Precision Training		✓
Sparsity and Super-Sparsity Support		✓
Hardware Acceleration for Sparse Operations		✓
Scalable, including large memory footprint	✓	✓
High Memory Bandwidth	✓	✓
Simple Programming Model	✓	✓
Software Composable for 24/7 server on time	✓	✓
Easy Deployment and Maintenance	✓	✓
Cost and Power Efficient	✓	✓
AI Futures: **Tachyum AI** Continues to Scale AI Performance and Efficiency – STAY TUNED		✓

Thank You

Visit us at Booth #330

www.tachyum.com

Tachyum™