The Convergence of HPC and AI in a Software-Composable Homogeneous Data Center

May 2022

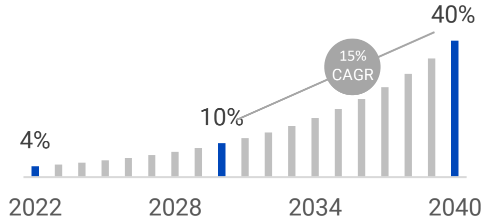
Robert Reiner

Director of Product Marketing

Serious Issues Facing Data Centers

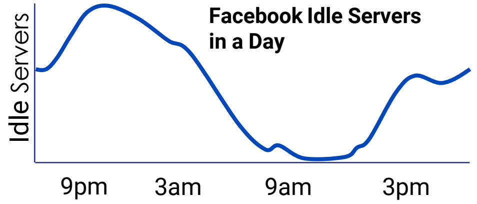
Data Center Power Consumption

- 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



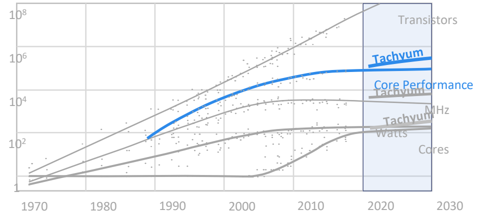
Low Server Utilization

- 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



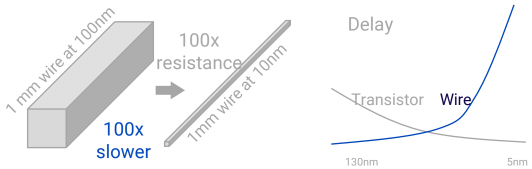
Performance Plateau and Moore's Law

- 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

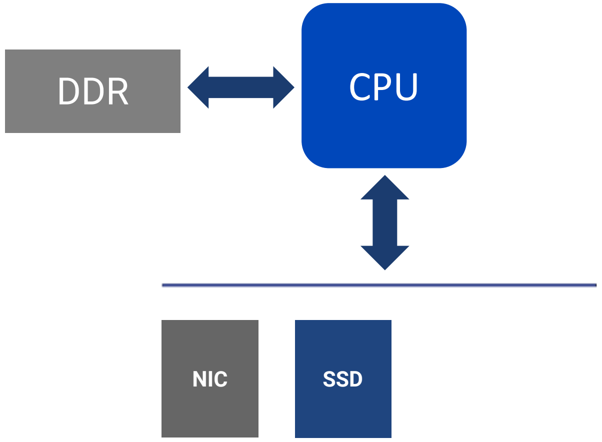


HPC vs. AI

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	

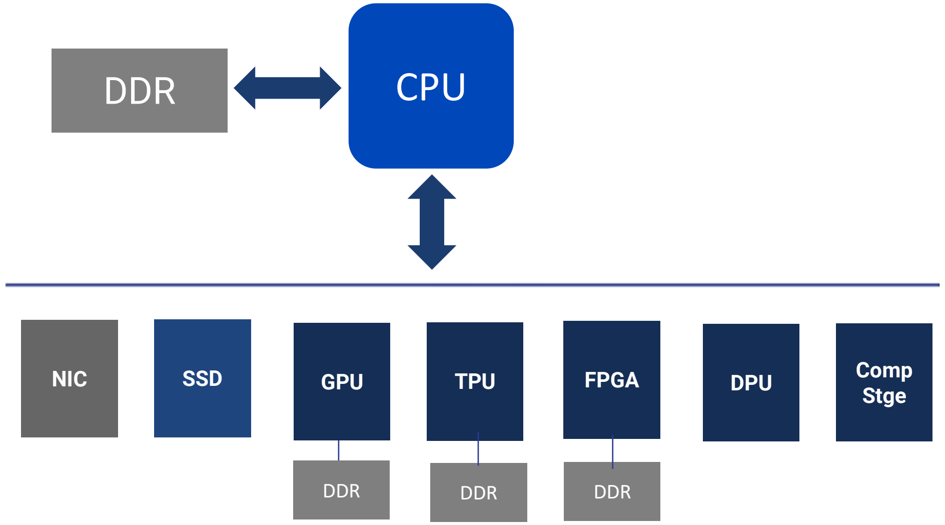
Homogeneous vs. Heterogeneous Systems

Homogeneous



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

Heterogeneous

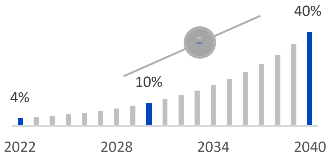


Pros	Cons
• Accelerates specific workloads, including HPC and AI • Scalable	• Needs special programming • Expensive, power-hungry • Under-utilized – contrary to software-defined data center

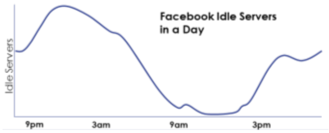
Tachyum Prodigy – The World's First Universal Processor

Problems

Data Center Pain Points

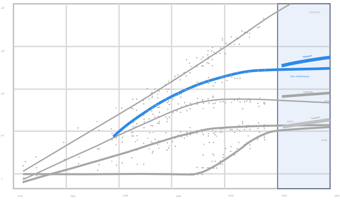


Data Center Power Consumption

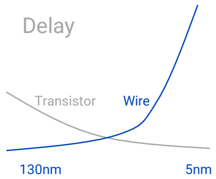


Low Server Utilization

Industry Transformation

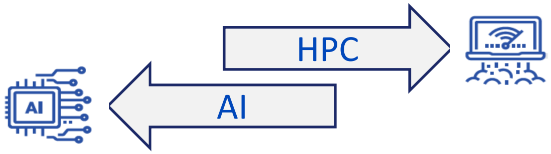


Performance Plateau

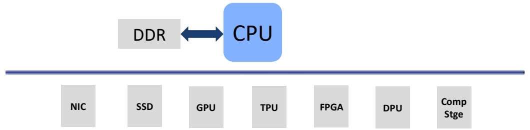


Slow Wires

HPC/AI Divergence



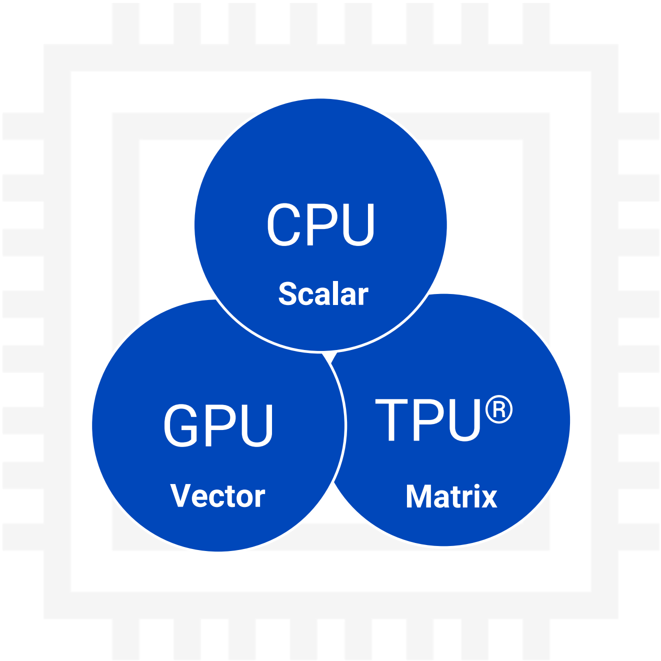
Accelerator Sprawl



Solution

Tachyum Prodigy Cloud / AI / HPC Supercomputer Chip

Unifies the Functionality of CPU, GPU, and TPU®

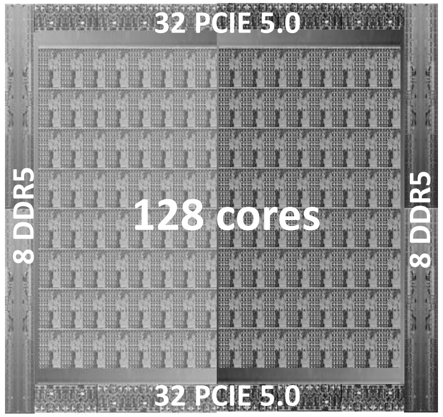


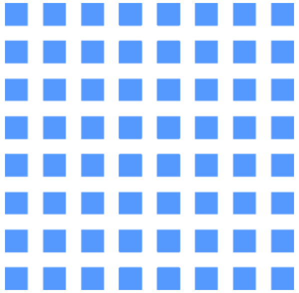
- Over 3x performance of Xeon
- Up to 10x performance at same power
- Faster than NVIDIA H100 in HPC and AI

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 Data Types Optimizes Efficiency
	• Quantization Support Using Low Precision Data Types
	• Scatter/Gather for efficient storing and loading matrices





Sampling End of 2022

* Technology to be disclosed at a future date

Tachyum Prodigy Software Ecosystem

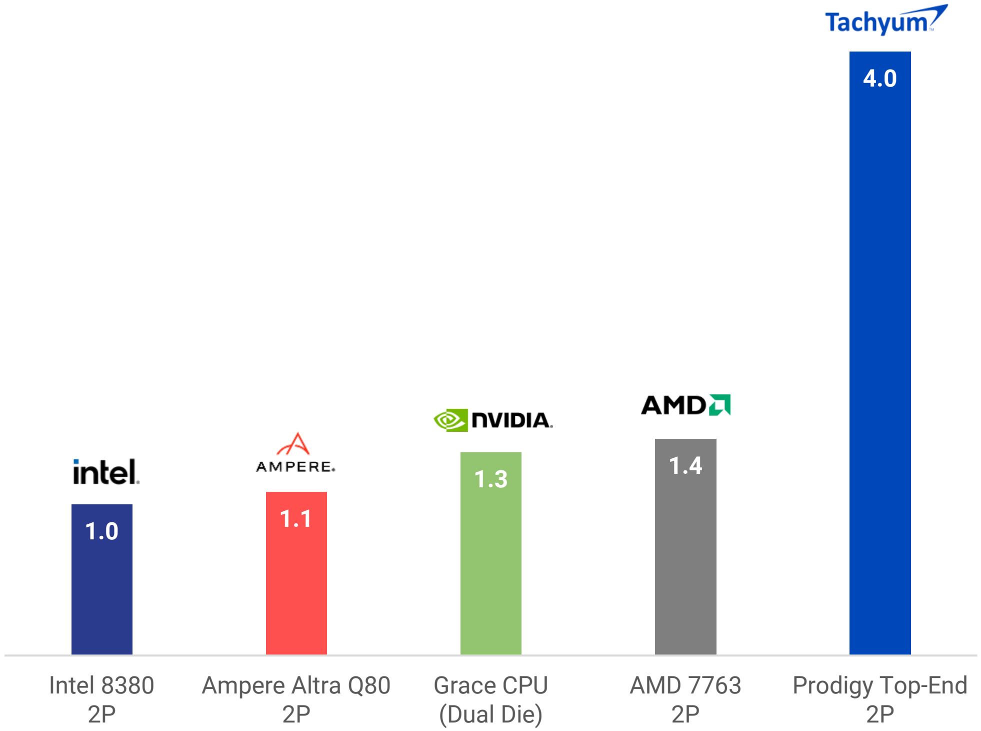
Applications	- Broad range of applications compiled to run natively on Prodigy	APACHE RabbitMQ Vim Java POSTFIX python Perl MariaDB Lua gRPC SQLite php mongoDB git DOVECOT
Frameworks & Libraries	- Support for major AI frameworks and scientific libraries for cutting-edge matrix and vector performance	TensorFlow Google BERT QUANTUMESPRESSO YOLO PyTorch TensorFlow Lite LAPACK OpenMP LAMMPS
System Software	- GCC, Linux and FreeBSD are ported to Prodigy along with the GNU libraries	KVM HPCG FreeBSD GCC XJTAG OpenBMC
Emulation	- SW Emulation with QEMU and C-model - Prodigy Hardware FPGA Emulation - Prodigy Runs x86, Arm, & RISC-V binaries	QEMU x86 Arm RISC-V
Software Roadmap	- Tachyum's roadmap adds key applications for big data, containers, and virtualization	docker ceph U-Boot LLVM COMPILER INFRASTRUCTURE Xen Project APACHE Spark HPCG slurm workload manager coreboot LinuxBoot

Prodigy Advantages

	Workload Requirements	Prodigy Differentiation
General Server	High DRAM and I/O Bandwidth	• Industry-leading 16 DDR5-7200+ Memory Controllers • 64 lanes of PCIe from 2 x 16 w/ bifurcation down to x2
	Scalable Platforms for Maximum Flexibility	• Hardware Coherency Supports 2 and 4 socket Platforms
HPC/AI	Highly Parallel	• 2 x 1024 Vector Units • 4 Kb Matrix AI Unit Supporting 16x16, 8x8, and 4x4 Matrixes
	Range of Data Types	• FP64, FP32, TFloat32, BFloat16, FP8, Int8, and TAI • Sparsity and Super-Sparsity

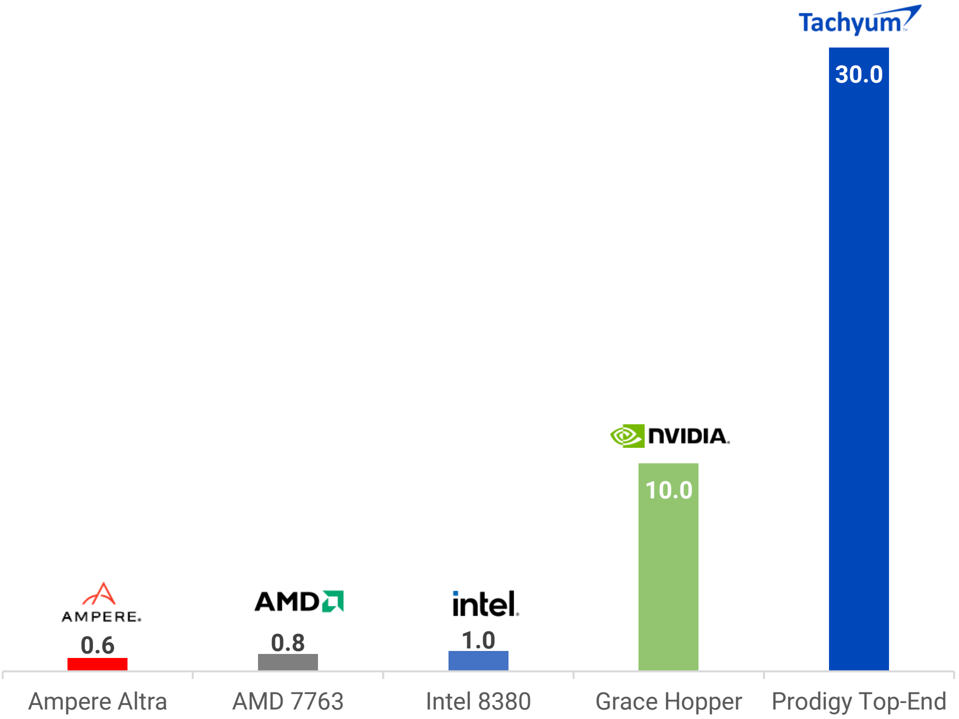
Prodigy vs. x86 and Arm

SPECrate 2017 Integer



Prodigy SPECrate 2017 Integer Performance
up to 4x Higher than Competition

Floating Point Raw Performance (FP64)



Prodigy Floating Point Raw Performance
up to >30x Higher than Competition

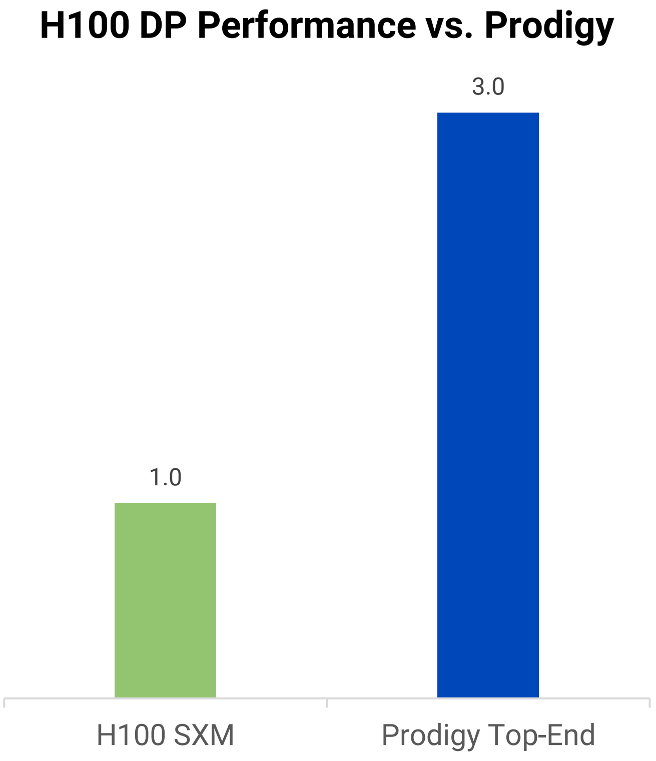
Sources: Spec.org, Nvidia GTC22 Keynote, Prodigy compiled using gcc 11.2

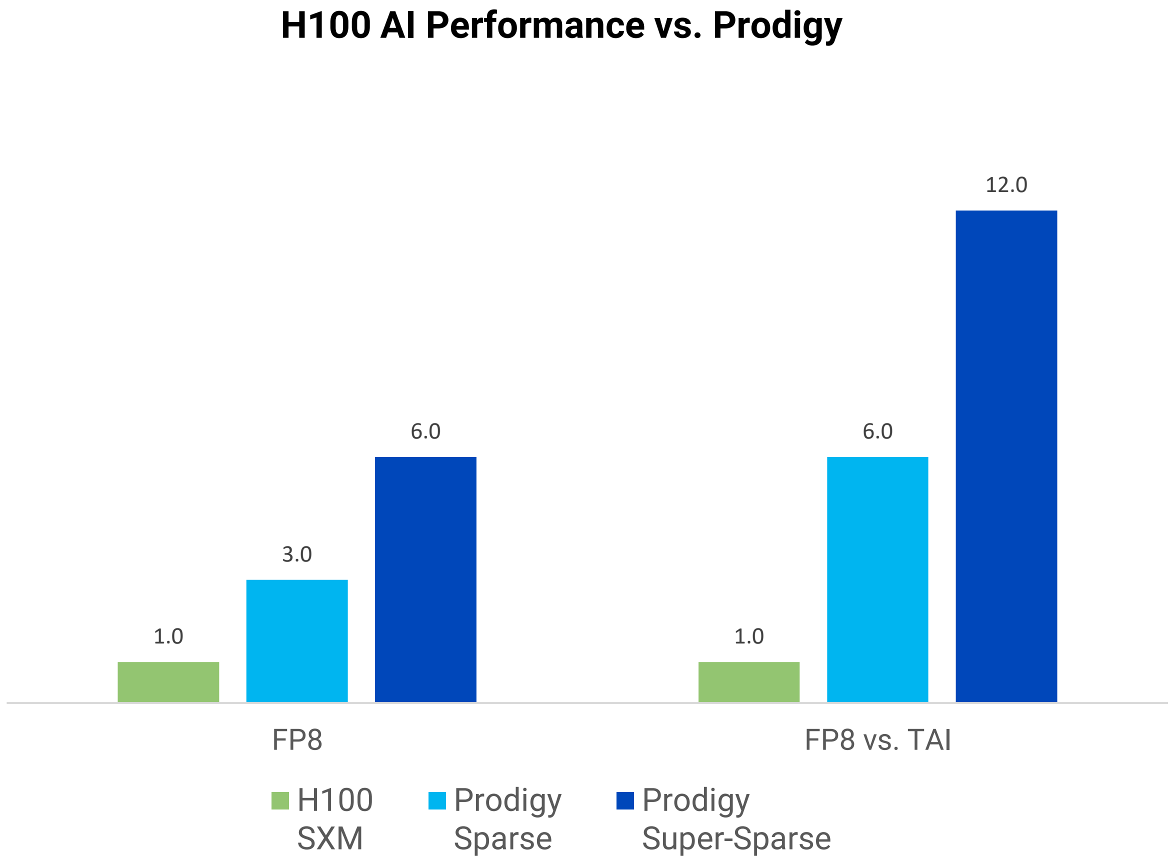
Matrix / Vector Processing Built from the Ground Up - ***Not Bolted On***

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

Prodigy vs. Nvidia H100 GPU – HPC and AI





Prodigy Delivers Up to **12x Higher AI Performance** and **3x Higher HPC Performance** than H100

Prodigy vs. Nvidia H100 – Rack-Level Comparison

H100 DGX POD

- 4 x H100 DGX
- 32 x H100 SMX

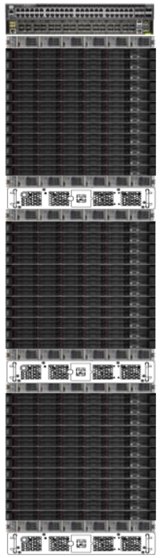
960 TF HPC FP64
128 PF AI FP8 Sparse

Air-Cooled Prodigy Rack

- 16 4P 3U Servers
- 64 Prodigy Mid-Range Chips

4.6 PF HPC FP64
1.2 EF AI TAI Sparse

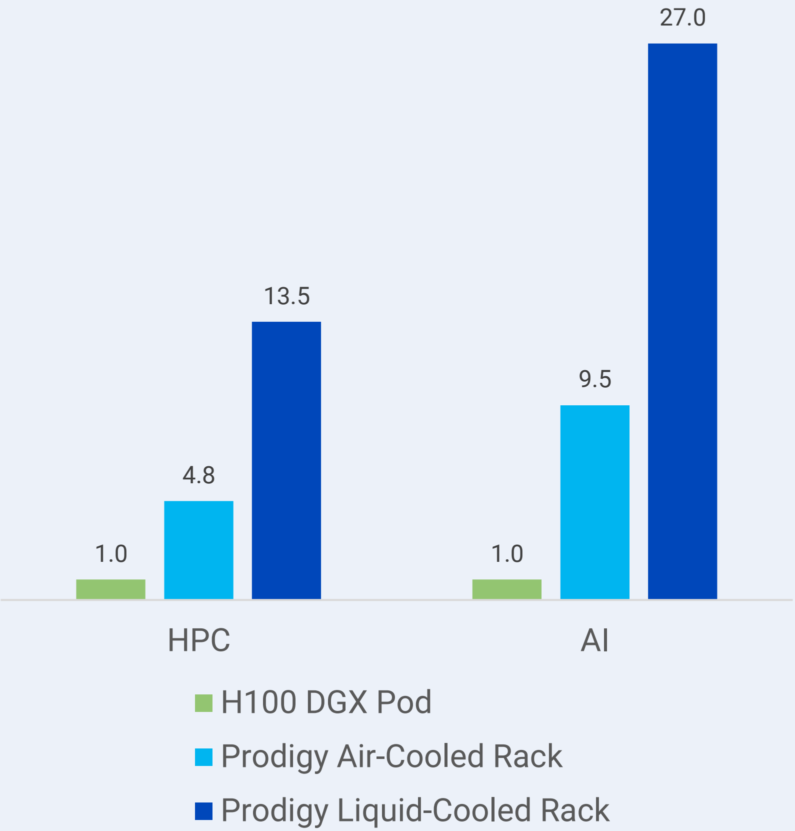
Liquid-Cooled Prodigy Rack



- 36 4P 1U Servers
- 144 Prodigy Top-End Chips

12.9 PF HPC FP64
3.5 EF AI TAI Sparse

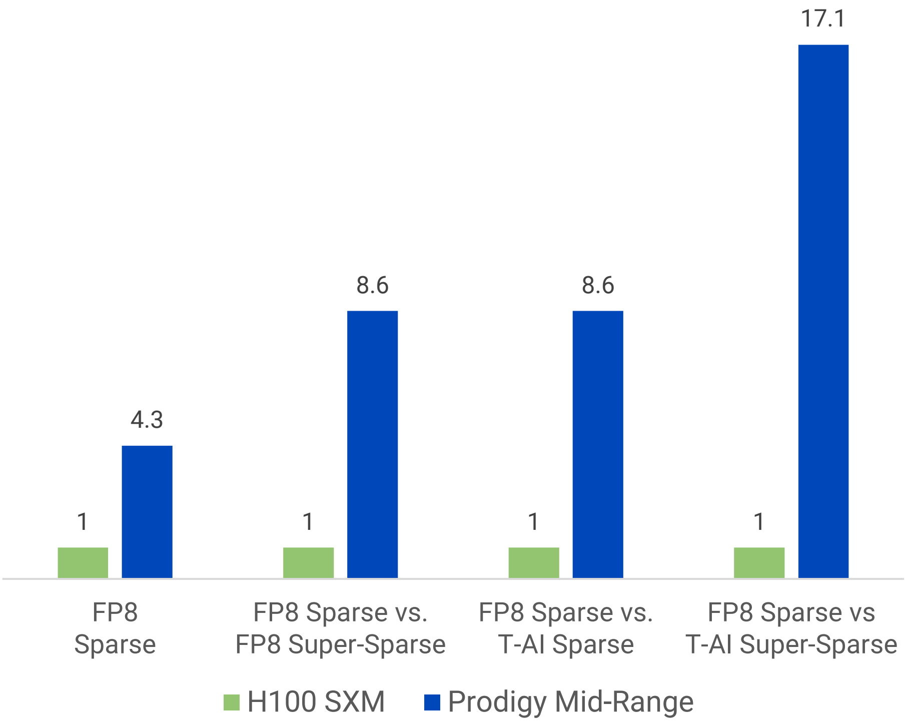
Prodigy Rack Performance Normalized to H100 DGX Pod



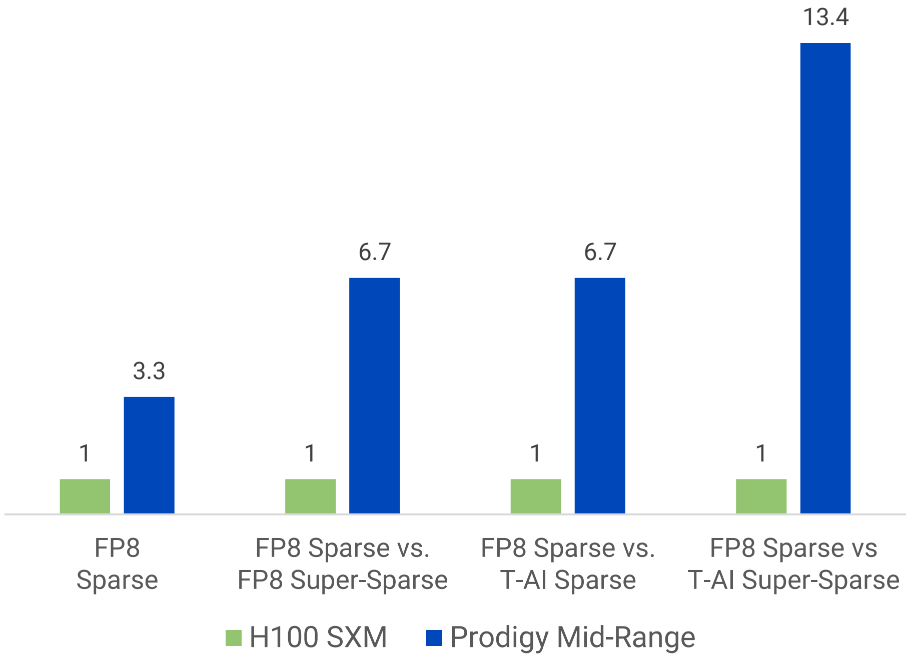
Prodigy vs. Nvidia H100

Rack Performance/TCO and Performance/W

H100 Rack Performance/TCO vs. Prodigy



H100 Rack Performance/W vs. Prodigy



Prodigy Rack Solutions Deliver **>17x Higher Performance/TCO and >13x Higher Performance/W** than H100 SXM

Summary

Prodigy Feature	HPC	AI/ML
High Performance Parallel Processing	✓	✓
Range of Floating-Point Precision	✓	✓
High Performance Vector and Matrix Operations	✓	✓
Sparsity and Quantization 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	✓	✓

Thank You

visit

www.tachyum.com