InspireSemi™

Disruptive Next Generation Accelerated Computing Platform

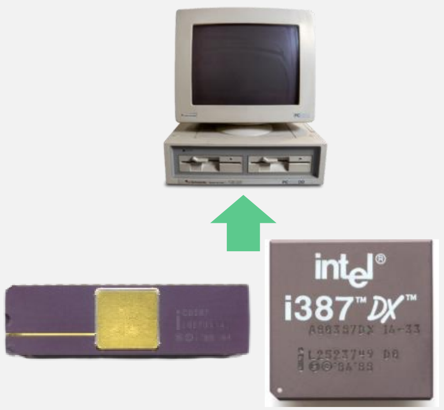
Blistering speed, energy efficiency, versatility, and affordability for HPC, AI and graph analytics applications

SC23 RISC-V Workshop
November 2023

The Third Wave of Accelerated Computing is Here

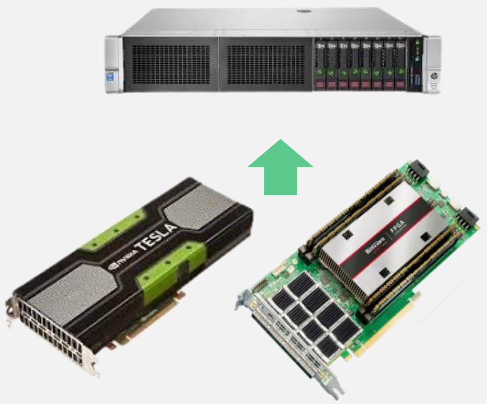
Thunderbird for HPC, AI, Graph Analytics

1980
Math Coprocessor



- Purpose-built widely applicable
- Open software ecosystem
- Plugs into existing computers

2007
GPU, FPGA



- Limited applications benefit
- Proprietary software model
- Plugs into existing servers

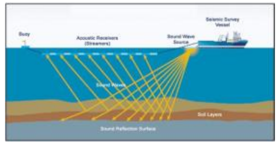
2023+
Thunderbird

- Built for HPC
- Versatile & open software ecosystem
- Plugs into existing servers

Addressing the Need to Accelerate All HPC & AI Software

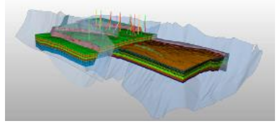
What customers always wanted…Not "yet another GPU"

Financial simulations

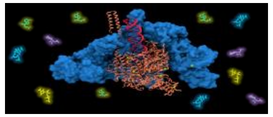


Geology: Seismic

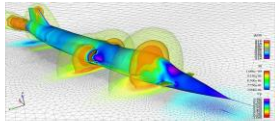
Financial Trading & Graph Analytics



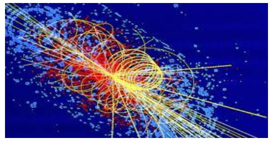
Energy: Reservoir Modeling & Sim



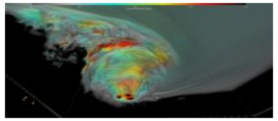
Genomics, Pharma, Life Sciences



CAE/Computational Fluid Dynamics



Nuclear Simulations Fission & Fusion



Climate & Weather Modeling

Datacenter GPUs

InspireSemi Thunderbird

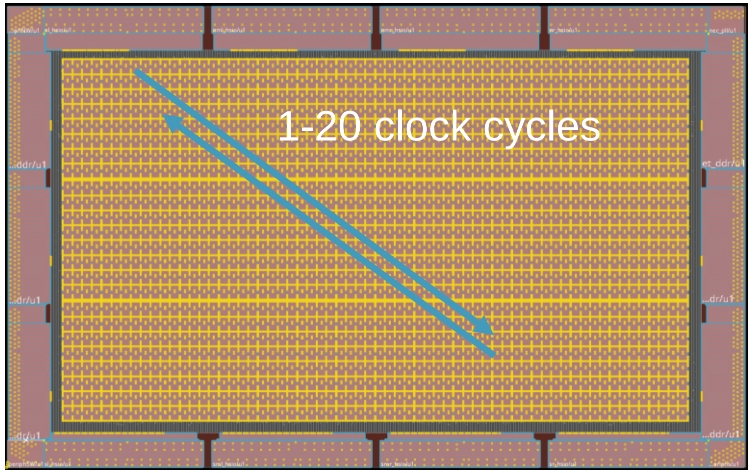
Highly differentiated "supercomputer-cluster-on-a-chip"

- Versatility as a platform across wide range of applications
- Each chip has 1,792 CPU cores connected via high-speed network
- 4 chip PCIe card delivers >7,000 interconnected 64-bit CPU cores
- Large scale computing power, supports up to 256 chips
- Best-in-class for both Performance/$ and Performance/Watt
- Delivers unprecedented capability within an established open software ecosystem

Thunderbird Addresses Key Industry Pain Points

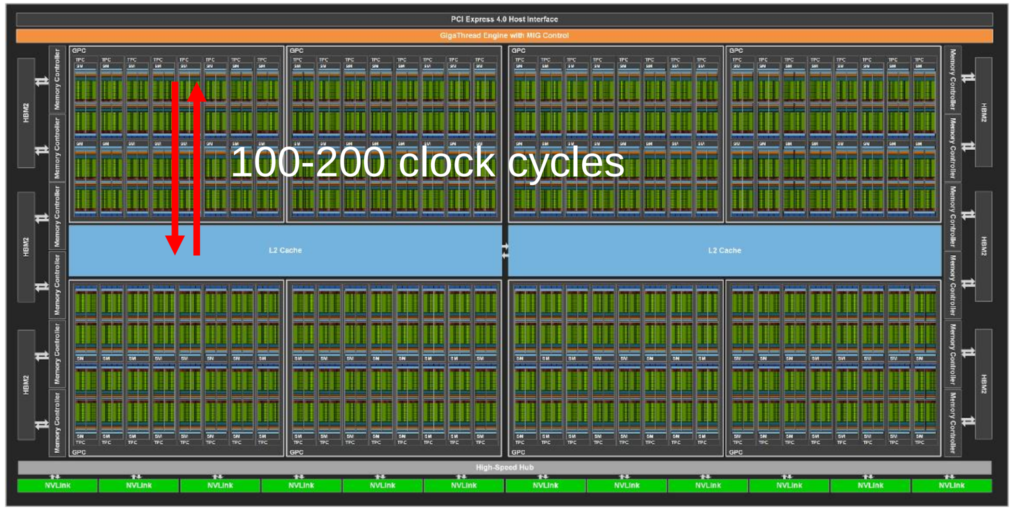
- Customers excited about key Thunderbird architectural advantages vs. competition
  - Greater utilization and real-world application performance
  - Predictable performance, known timing behavior
  - Lower power consumption
- Determinism: Thunderbird addresses applications where GPUs do not work
  - FinTech customer "ah-ha moment" insight – Latency, MIMD vs. SIMD
  - Repeatability of results is a must for many key applications: high-frequency trading, cryptography, healthcare imaging, smart weapons, self-driving cars, …

<u>Latency example – Thunderbird (MIMD) vs. leading GPU (SIMD)</u>



10x

greater efficiency



InspireSemi™

Thunderbird Addresses ALL HPC & AI Customer Needs

	InspireSemi Thunderbird	CPU	GPU	FPGA	AI Accelerators
Architecture	Many programs, many data streams	Few programs, few data streams	Few programs, many data streams	Programmable logic elements	Single program, many data streams
Performance	High for broad range of HPC apps	Slow, need h/w accelerators	High for AI and some HPC apps	Medium	High for AI only
Cost	Low $6,500 for 2 chip PCIe card	High ~$1K-8K (+ more servers)	High ~$7K-48K	High $8K-$10K	High ~$10K - $2.2M
Energy consumption	Low ~150W/chip	Med 240W+/chip (+ more servers)	High ~700W	High ~300W	High ~300W – 20kW
Scalability	256 chips	1-4 chips	2-8 chips	1 chip	1-2 chips
Programming model	Standard CPU-like, Any language, Full instruction set	Standard CPU, Any language, Full instruction set	Specialized C variant (CUDA, ROCM, SYCL)	Hardware description language	Proprietary, obscure
Software ecosystem	Open-source, Linux, compilers, libraries, AI frameworks, existing applications	Robust	Limited, proprietary	None	AI frameworks and proprietary software stacks