InspireSemi™

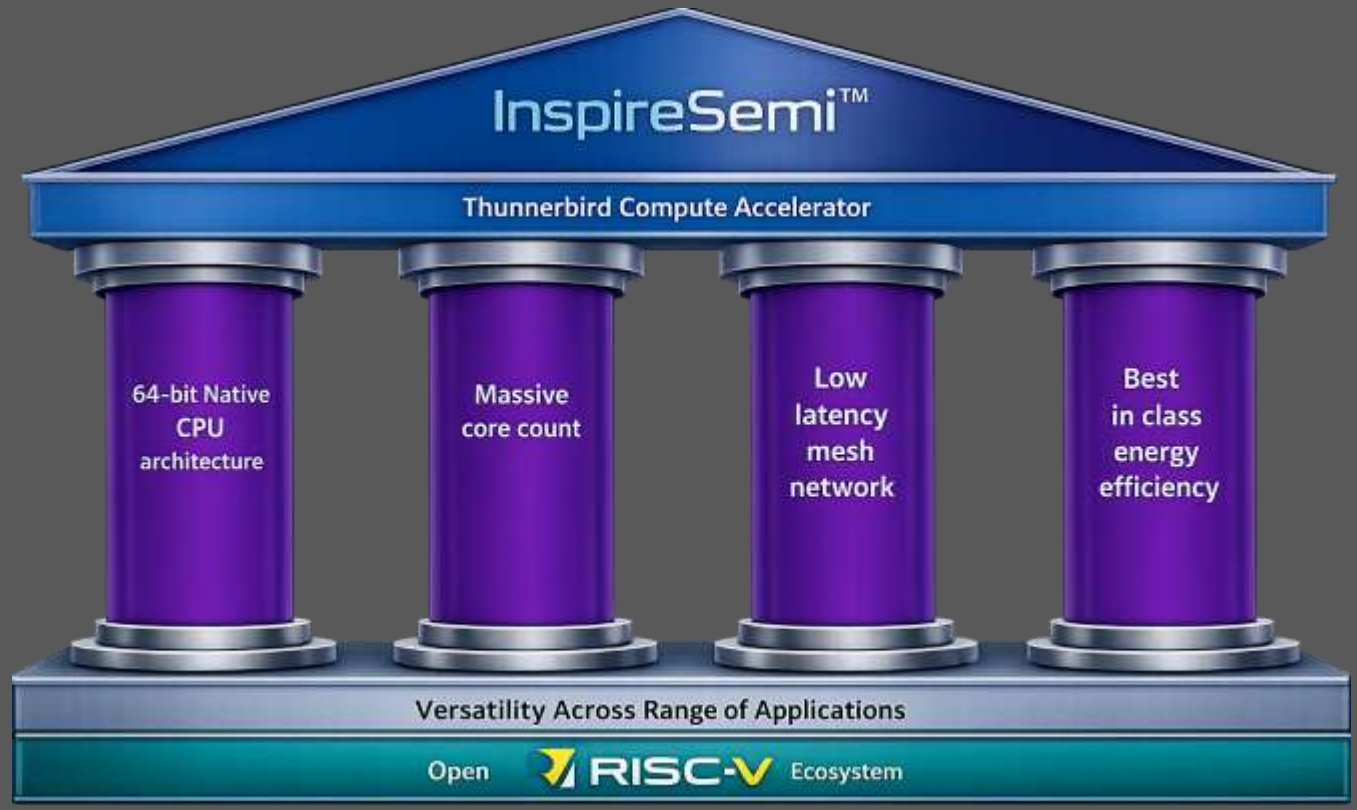
Disruptive Next Generation HPC-AI Accelerated Computing Platform

RISC-V®

HPC Next: Reflections and Progress on the RISC-V Ecosystem

June 2026

ISC High Performance
The HPC Event.



Accomplished Leadership Team

Alexander Gray, Founder, President & CTO

- 20 years experience in tech startups, entrepreneurship
- CryptoCore, SolarBridge, SunPower
- Holds 9 patents
- BSEE, University of Illinois at Urbana-Champaign

SolarBridge Technologies SUNPOWER

James Hickman, Chairman & CEO

- 25+ year extensive business and investment career
- Has been CEO of multiple startups
- Was investor #1 in InspireSemi
- Former US Army intelligence officer
- BS Mathematics & Computer Science, US Military Academy at West Point
- MS Accounting & Information Management, University of Texas at Dallas

Jack Cartwright, CFO

- 20+ years experience in tech startups and public companies
- Datum, Spruce Services, BT Global, Atkins Global, Cisco
- Former US Army Logistics & Operations officer
- MBA, University of Texas; MS Accounting, University of Miami

DATUM spruce BT ATKINS Member of the SNC-Lavalin Group CISCO

Thomas Fedorko, COO

- 35+ years hands-on technical and business leadership in semiconductor operations in both large IDM and startups
- Eta Compute, Uhnder, Bluetechnix, Black Sand (Qualcomm), Luminary Micro (TI), Oak Technology, Motorola SPS
- Technical degree from DeVry University and graduate of the Motorola Management Institute

Eta Compute UHNDER Qualcomm Black Sand Technologies Texas Instruments Luminary Micro OAK MOTOROLA

Doug Norton, CMO

- 35+ years experience; enterprise, startups, Federal
- Viviota, Nimbix, TMT, Virtana, Newisys, CoWare (Synopsys), Cadence, IBM
- President of The HPC-AI Society, Mentor at UT ATI (Austin Technology Incubator at University of Texas)
- RISC-V International: member HPC-SIG & Marketing team
- BSEE, Missouri University of Science & Technology

Viviota NIMBIX tmj Texas Multidore virtana NEWISYS CoWare cadence IBM

Reflections on RISC-V and why we chose it

- RISC-V is a clean sheet design
- RISC-V is a global standard, not dependent on any single vendor
  - Stable long-term platform for hardware and software
- RISC-V enables tightly coupled custom extensions while maintaining compatibility with standard software
- RISC-V enables very efficient standard scalar processing
- RISC-V promised efficient standard vector processing
- RISC-V enables hardware/software codesign
- **RISC-V is a better business model**

Companies and their ISAs come and go

- Digital Equipment Corporation, RIP! (**PDP-11, VAX, Alpha**)
- Intel's dead ISAs (**i960, i860, Itanium, …**)
- **SPARC**
  - Sun opened v8 as IEEE 1754-1994
  - Acquired by Oracle, SPARC development closed down in 2017
- **MIPS**
  - Sold to Imagination, bought by Wave
  - Opened up MIPS R6 ISA in 2018, then made not open in 2019
  - Now a RISC-V IP provider
- **IBM POWER**
  - Initially proprietary, opened up ISA in 2019. Crickets…
- **ARM**
  - Sold to Softbank in 2016
  - Nvidia announced acquiring ARM 2020, deal cancelled 2022
  - IPO September 2023

Why entrust your software investment to a proprietary ISA?

Special thanks to Krste Asanovic
Professor Emeritus at UC Berkeley
Chief Architect, RISC-V International

Movement is happening because *new business model* changes everything

ISA	Chips?	Architecture License?	Commercial Core IP?	Add Own Instructions?	Open-Source Core IP?
x86	Yes, ***two*** vendors	***No***	***No***	***No***	***No***
ARM	Yes, ***many*** vendors	Yes, ***expensive and restrictive***	Yes, ***one*** vendor	***No*** *(Mostly)*	***No***
RISC-V	Yes, ***many*** vendors	Yes, ***free***	Yes, ***many*** vendors	***Yes***	***Yes, many*** *available*

- Pick ISA first, then pick vendor or build own core
- Add your own extension without getting permission
- Commercial, academic, and open-source ecosystems can coalesce around a single open standard

Inherent sustainable ISA advantage

- Simple base + modular extensions (profiles)
  - Much lower area/power at equivalent performance levels over other ISAs
  - Technical reasons: reduced dynamic code size, compare+branch instruction, no complex instructions (e.g., two integer read ports max, no flags), …
- Clean-sheet design
  - Avoids legacy warts
- Designed for extensibility while retaining standard software stack
  - Variable-length ISA part of original design, supports vector extensions, custom additions
- Community-designed
  - Input from leading experts in academia and industry
- No limit to performance levels or application domains – will surpass all other architectures quickly

HPC-AI hardware landscape in modern era

RISC-V everywhere

AI

esperanto.ai

NVIDIA GPU

AMD INSTINCT

Tsavorite SCALABLE INTELLIGENCE

SambaNova SYSTEMS

intel. GAUDI

intel. DATA CENTER GPU MAX SERIES

GRAPHCORE

groq

cerebras

HPC

InspireSemi™

NEXTSILICON

Deterministic, Predictive Perf

XILINX®

ALTERA now part of Intel

RIVOS

tenstorrent

intel.

AMD

ARM

VENTANA

Qualcomm

General Purpose Compute

RISC-V IP companies

SiFive

ANDES

MIPS a GlobalFoundries company

OPENHW FOUNDATION

ARC Synopsys

AheadComputing

Sd

ARTERIS IP

Thunderbird accelerated computing platform

- *"Supercomputer cluster-on-a-chip"*
  - 1,536 RISC-V CPU cores per chip, all 64-bit (>6,000 per PCIe card)
  - Native FP64 supports all important high precision HPC applications
- *Ultra-efficient, ultra-compact custom CPU cores*
  - Based on modern open-standard RISC-V architecture
  - InspireSemi designed high-perf, low-power superscalar CPUs
- *Patent pending low latency interconnect fabric*
  - Key to efficient utilization of so many cores
  - Scales to multiple-chip arrays up to 256 chips
- *Energy efficiency: 75 GFLOPS/watt (FP64)*
  - Focus on performance/watt, fits in current datacenters
- *Supports existing open software ecosystem*
  - Easily adapt your software programs
  - Fast – no big investment or training required
- *Recognized global partners to deliver turnkey solutions*
  - High-volume across multiple markets and geographies

RISC-V®

Lenovo

PENGUIN COMPUTING

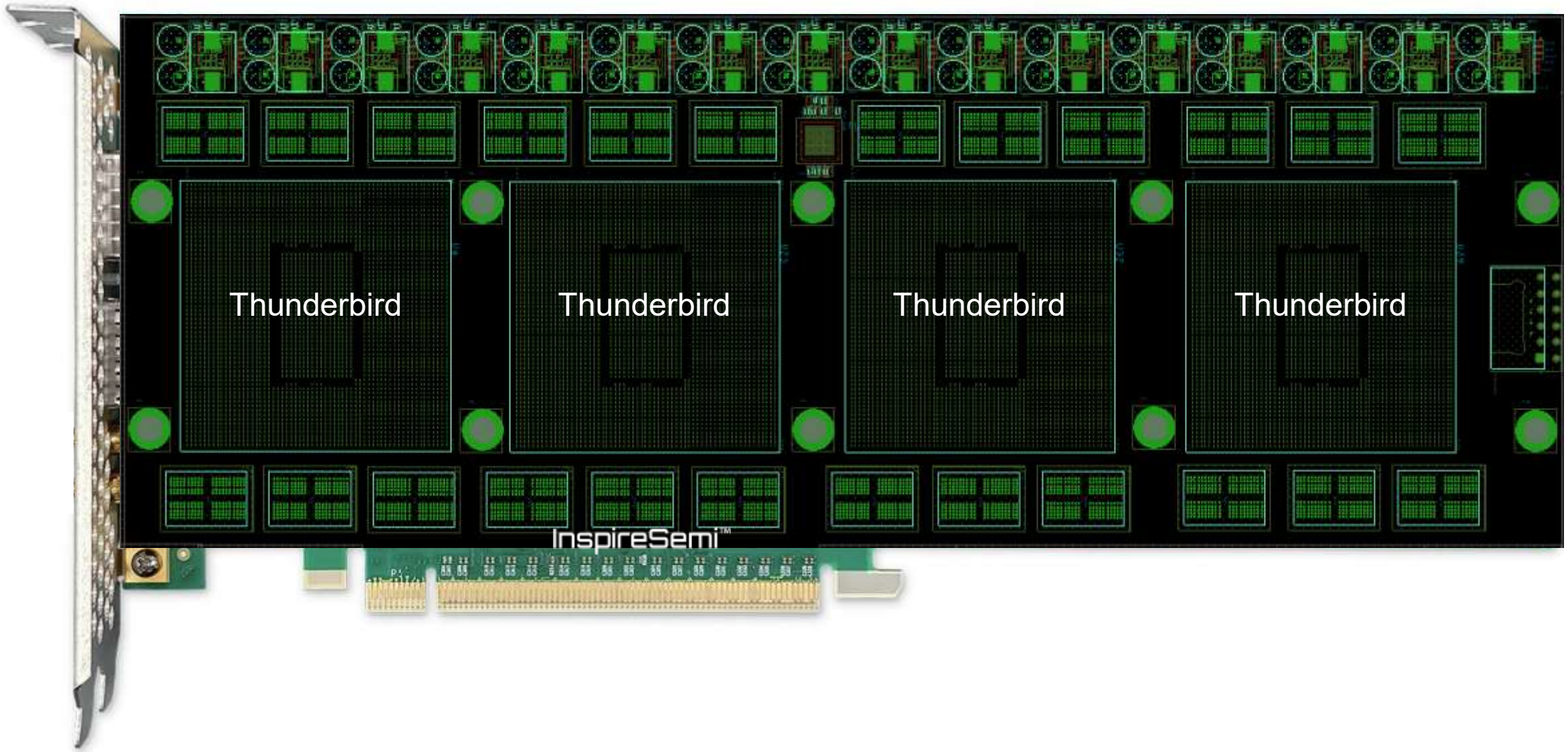
World Wide Technology

POWERED BY DELL Technologies

SUPERMICRO

E4 COMPUTER ENGINEERING

Thunderbird PCIe accelerator raw card



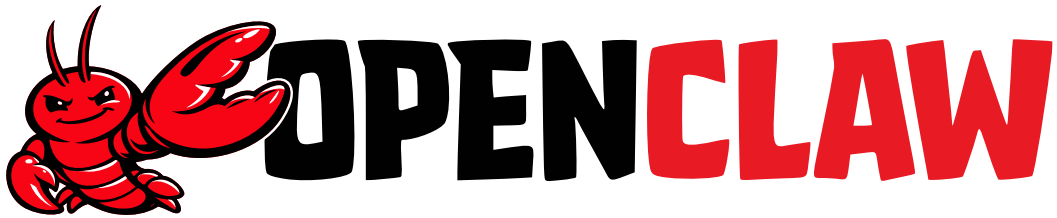
Revolutionizing HPC-AI datacenters

One Thunderbird board has more CPU cores than an ***entire rack of standard Intel, AMD, or ARM servers!***

- Less cost: datacenter footprint, servers, networking, cables, power, cooling
- Less complexity, interconnects, points of failure
- No need to build new datacenters

VS.

Turns out the same math holds true for agentic AI orchestration use cases



Open software ecosystem solves customer porting challenges

- Supports open standard software, eliminates vendor lock-in (unlike competing compute accelerators)
  - **Linux!** provides access to existing HPC-AI software ecosystem
  - Plus Zephyr and FreeRTOS (Real-Time Operating Systems)
  - Eliminates need to learn and use proprietary software stacks
- Also supports OpenMP offload model
- Uses standard CPU-style programming models
  - No need for CUDA, ROCM, etc. that GPUs require
  - No need for disruptive software algorithm rewrites
  - Standard compiler, OpenMP, MPI, etc. approaches
- Leverages key RISC-V software ecosystem
  - Address-accurate QEMU model
  - Standard compilers, e.g.- GCC, Gfortran, GDB toolchains
  - Standard HPC libraries, e.g. – BLAS, LAPACK, FFTW

Do We Still Need GPUs?

Rethinking AI and Scientific Computing on Matrix-Enhanced CPUs

Jack Dongarra,[1] Torsten Hoefler,[2] Satoshi Matsuoka[3]

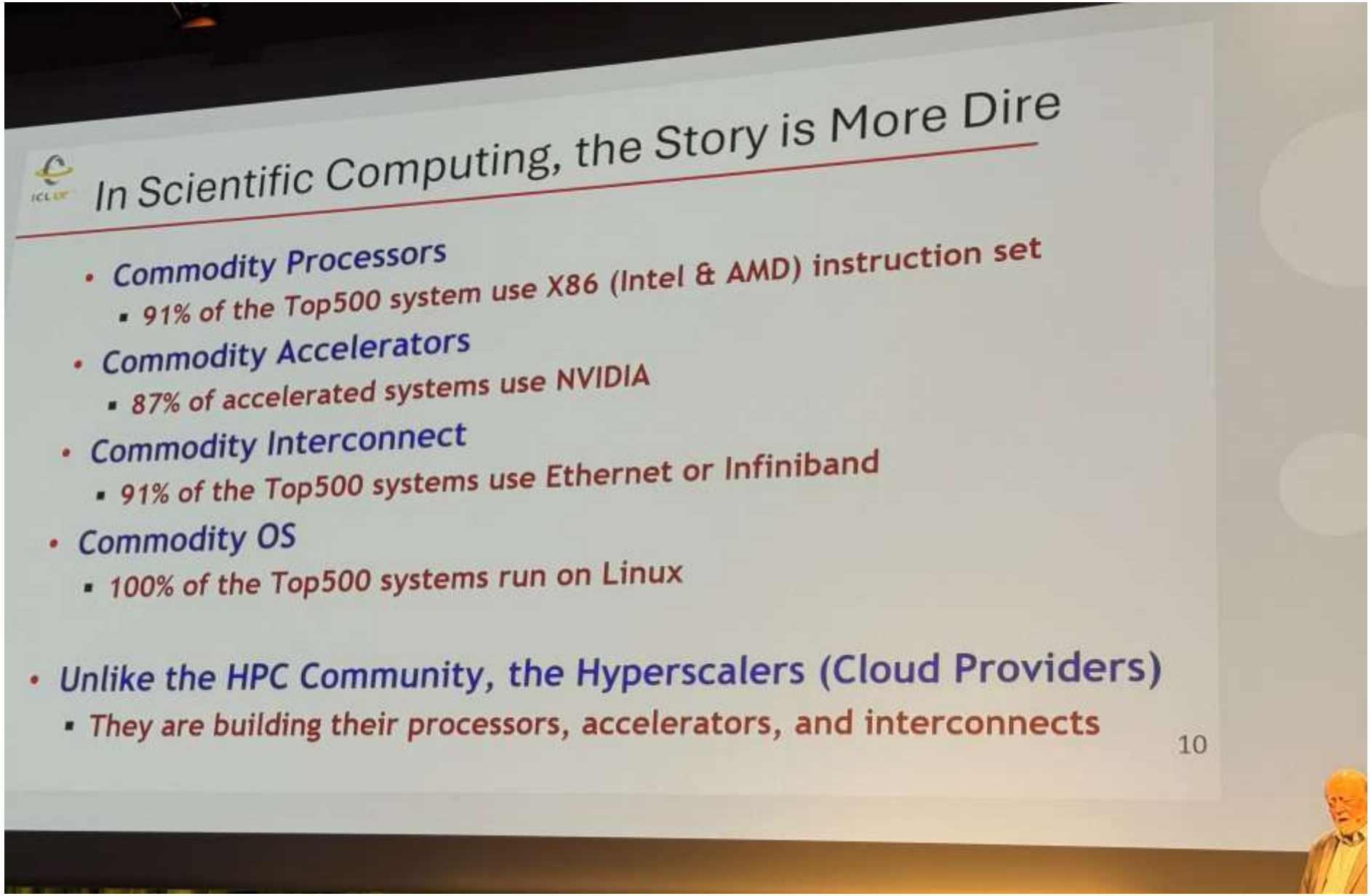
[1]University of Tennessee [2]ETH Zurich [3]RIKEN Center for Computational Science (R-CCS)

CACM revised version — June 19, 2026

http://bit.ly/no-gpus

- LineShine and Fugaku Top500 dominance causing GPU rethink
- CPU-based designs from ARM with SVE/SME and Intel's AVX/AMX, coupled with mixed precision and HBM memory, are changing the game
  - "we may not need GPUs for many AI and scientific workloads if we have a CPU with SVE and SME, large amounts of HBM, native support for FP64, FP32, FP16, and INT8, and a high-throughput matrix-multiply engine.
  - ***Such a processor attacks the very limitations that led to GPU adoption***—insufficient parallelism, insufficient memory bandwidth, and the lack of low-precision tensor performance—while preserving the advantages of CPUs: generality, programmability, strong control flow, mature system software, and suitability for irregular scientific workloads."
- **It is inevitable that RISC-V systems for HPC & AI will be next**

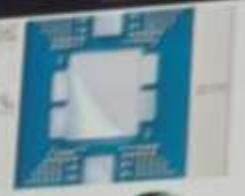
Scientific computing “codesign” - Jack Dongarra



Hyperscaler codesign - Jack Dongarra

Cloud Providers are Designing and Using Their Own Processors

- Alibaba
  - CIPU, 128 core ARM based
  - Alibaba's Elastic Compute Service



- AWS Graviton4
  - 96 ARM cores, 7 chiplet design
  - ~100 billion transistors, DDR5 memory

- Google TPU7
  - 2X TPU3 performance
  - 4096 units per "pod"
  - Reconfigurable optical interconnect

- Microsoft Azure
  - Project Catapult/Brainwave FPGA accelerator
  - Cobalt 100 (128 Neoverse N2 ARMv9 cores)
  - Maia 200 AI accelerator
  - Q1 2026 $32.9B investment in data centers

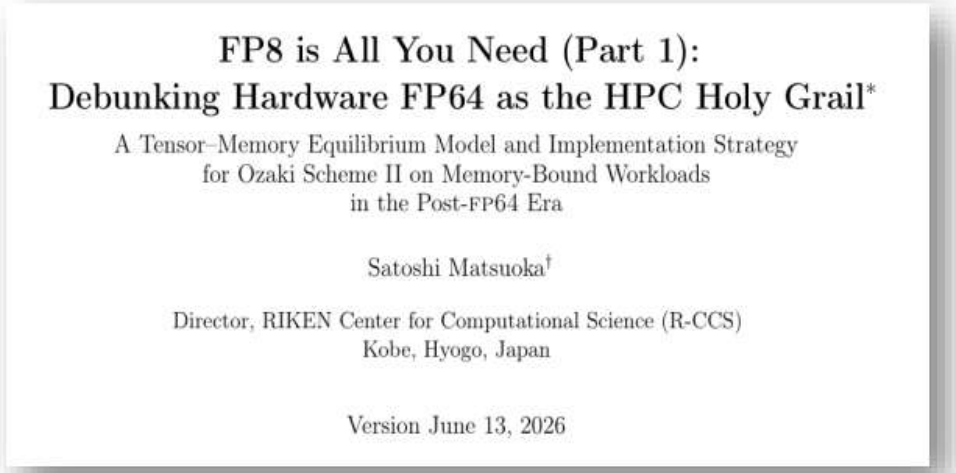
In generative AI and cloud computing, we see genuine co-design successes.

Modern accelerators and AI ASICs are shaped in close dialogue with a small set of dominant workloads: transformer models, recommendation engines, cloud infrastructure accelerators, and related kernels.

Every hyperscaler AI vendor has designed and fabricated custom solutions to improve performance, increase reliability, and decrease costs.

Hyperion Research: FP64 vs. FP4, an evolving debate

FP64 in an FP4 World or Vice Versa



FP8 is All You Need (Part 1):
Debunking Hardware FP64 as the HPC Holy Grail*

A Tensor–Memory Equilibrium Model and Implementation Strategy
for Ozaki Scheme II on Memory-Bound Workloads
in the Post-FP64 Era

Satoshi Matsuoka†

Director, RIKEN Center for Computational Science (R-CCS)
Kobe, Hyogo, Japan

Version June 13, 2026

https://arxiv.org/pdf/2606.06510

From NVIDIA Technical Blog*

- *At the same time, dedicated FP64 vector performance remains critical for scientific applications that are not dominated by matrix kernels*
  - In these cases, performance is constrained by data movement through registers, caches, and high-bandwidth memory (HBM) rather than raw compute.
  - A balanced GPU design therefore provisions sufficient FP64 resources to saturate available memory bandwidth, avoiding over-allocation of compute capacity that cannot be effectively utilized

** https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer/*

- The IEEE 754 FP standard is not just about matrix multiplies that gives high fidelity results
  - It is a way to ensure portability and consistency in calculations across a wide range of numerical rules and practices
- **How long to resolve these issues with FP4 to the satisfaction of the scientific and engineering community? (2036?)**

RISC-V gap analysis (2025)

Still more work to do for world domination

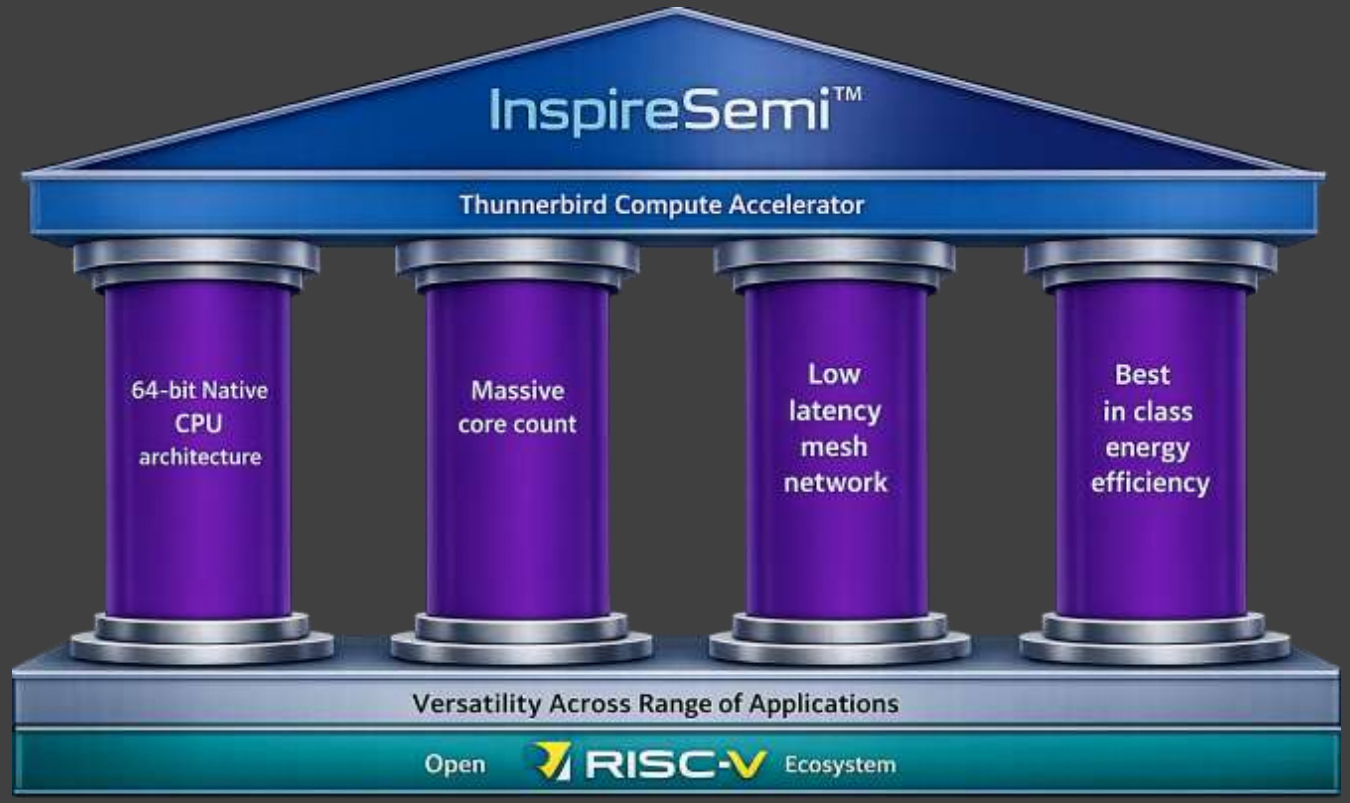
- ***Recommendation 1:*** The HPC community, and HPC vendors, should look to support mature RISC-V networking solutions that can be leveraged in production supercomputers. This also includes experimenting with, and maturing, existing support for OmniPath, Ultra Ethernet and InfiniBand.
- ***Recommendation 2:*** There are several matrix extension standards in progress, members of the HPC community should look to engage with and ensure that they meet the needs of scientific computing.
- ***Recommendation 3:*** We should prioritize optimization and maturity of the Lustre filesystem to RISC-V **Software**
- ***Recommendation 4:*** Working with compiler developers to ensure these better support long vectors, as well as ensuring that frontends such as Fortran
- ***Recommendation 5:*** We should prioritize profiling and debugging tooling on RISC-V and engage with the RISC-V Performance Events Specification task group to ensure that a set of events are standardized that are then useful for performance monitoring tooling.
- ***Recommendation 6:*** While numerous HPC applications have already been ported to RISC-V, the community should priorities encouraging ISVs to port their codes and other applications such as VASP, Nektar++, NEMO and SENGA (in that order) are also desirable. Further tuning these applications and the underlying libraries for the architecture would be beneficial.

RISC-V is inevitable

- Industry wants this business model
  - More vendors competing for business with most cores designed
- RISC-V will have the best processors
  - Inherently better ISA than others
  - Designed to scale from embedded to datacenter
- RISC-V will have the best ecosystem
  - Largest number of players
  - All sockets become RISC-V over time
  - Software wants to run on the best processors available

InspireSemi™

Disruptive Next Generation HPC-AI Accelerated Computing Platform



Thunderbird™

"Supercomputer cluster-on-a-chip"

Raw performance: 24 TFLOPS (FP64)
Energy efficiency: 75 GLOPS/watt (FP64)
Low latency: 10x faster
Determistic + Predictable Performance
Value: exceptional performance/$

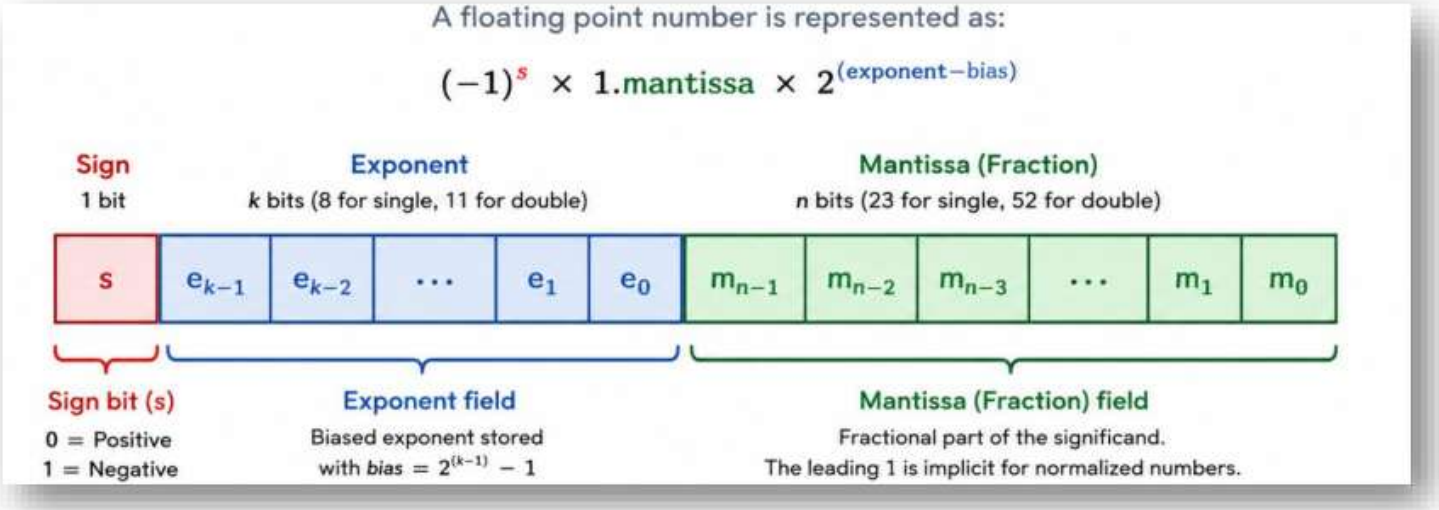
Appendix

In the Beginning: Chaos

- Burrough B5500 (1961)
  - 47 bit words
  - Bit 0 - always 1, Bit 1 – mantissa sign, Bit 2 – exponent sign (radix = 8), Bit 3-8 – unsigned exponent, Bit 9-47 - unsigned mantissa (integer)
- CDC 6000 (1964)
  - 59 bit words
  - Bit 0 – mantissa sign, Bit 1-11 – exponent, Bit 12-69 unsigned mantissa (integer)
  - Two's compliment format for exponent and mantissa
- IBM System 360/370 (1964)
  - 64 bit words (double precision)
  - Bit 0 - mantissa sign, Bit 1-7 – (exponent in excess – 64) (radix = 16), Bit 8-31, unsigned mantissa
  - Bits 32-63, optional add on mantissa
- Vax PDP-11 (1977)
  - 32 bit words
  - Bit 0 - mantissa, Bits 1-8 (exponent in excess – 128) (radix =2), Bits 9-15, 16-31 unsigned mantissa fraction

Let There Be Light:

- In I985, along came IEEE 754, a technical standard developed by the IEEE
- Specified interchange and arithmetic formats, along with methods for binary and decimal floating-point arithmetic in computer programming environments, including handling of exception conditions
- Addressed the need for portability and consistency in floating-point computations by defining precise representations and operations that ensure predictable results
- But just as important, defined binary and decimal formats, arithmetic operations (such as addition, subtraction, multiplication, division, and square root), rounding modes, and exception handling for overflow, underflow, and invalid operations



- E.g. FP64 accommodates values roughly from 2.2×10^{-308} to 1.8×10^{308} with about 15 decimal digits of precision
- **It took over a decade for this standard to move from concept to reality**

A Quick Example of Thoughtful IE^3 754 Standards: Rounding

Or: Why it took a decade

- Round To Nearest (Ties to Even) — *Default*
  - Rounds to the nearest representable value
  - If the exact value falls exactly halfway between two representable values (a tie), it rounds to the value with an even least significant bit (i.e., ending in 0)
- Round To Nearest (Ties Away from Zero)
  - Also rounds to the nearest representable value
  - However, if the value falls exactly halfway, it rounds to the value further from zero (i.e., rounds positive numbers up and negative numbers down)
- Round Toward Zero (Truncation):
  - Drops all extra bits and chooses the representable value closest to, but not greater in magnitude than, the exact value (towards 0)
- Round Toward Positive Infinity (Round Up):
  - Rounds up to the closest representable value that is strictly greater than the exact value (towards +∞)
- Round Toward Negative Infinity (Round Down):
  - Rounds down to the closest representable value that is strictly less than the exact value (towards –∞)

Along Comes FP4

Two main and evolving standards for FP4 targeted for current AI-centric hardware

- NVIDIA FP4: Introduced with NVIDIA Blackwell GPUs
  - Uses an E2M1 layout (1 sign bit, 2 exponent bits, 1 mantissa bit) in 16-element blocks
  - Two-Level Scaling: Employs fine-grained E4M3 (FP8) scaling factors per 16-value block and a second-level FP32 scalar for the entire tensor
- Microscaling FP4: Part of the Open Compute Project standard, supported by AMD (CDNA) and NVIDIA, which uses an E2M1 layout in 32-element blocks
  - Uses Shared Scaling: Groups of 32 elements share a common E8M0 (8-bit exponent, no mantissa) scaling factor
- Sample FP4 Rounding Modes:
  - For inferencing often uses round-to-nearest (RTN)
  - For training: stochastic rounding (SR)

    For example, 2.4: 40% chance of rounding to 3 and a 60% chance of rounding to 2

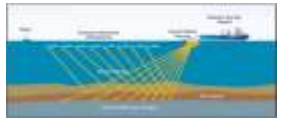
FP64 in an FP4 World

- NVIDIA GPUs use a specialized emulation approach to handle FP64 demands by leveraging high-throughput, low-precision tensor cores
- However, there are concerns with FP64 emulation that include IEEE 754 non-compliance
  - Data-dependent accuracy
  - Potential numerical instability in complex simulations
  - Failure to properly account for *Not a Number* errors, infinite numbers, or specific signed zero scenarios
- Although DGEMM (double-precision matrix multiplication) emulation is deemed effective, there are few production-ready solutions for more complex transcendental function (exponential, logarithmic, etc.) and trigonometric math functions (sin(x), cos(x), etc.)
- For its part, AMD primarily focuses on native FP64 in its MI430X
  - Consumes much more chip real estate and power than emulated counterpart (~16-64X)
  - AMD is 'studying' FP64 emulation but favors delivering the highest native FP64 GPU on the market

Addresses the need to accelerate All HPC-AI software

What customers always wanted…not "yet another GPU"

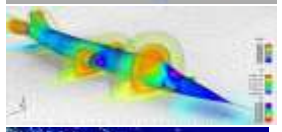

Financial simulations


Geology: Seismic

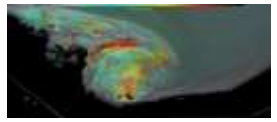
Datacenter GPUs primarily focused on AI


Financial Trading & Graph Analytics


Energy: Reservoir Modeling & Sim


CAE/Computational Fluid Dynamics


Government Lab Simulations


Climate & Weather Modeling


Cybersecurity & Cryptography


Genomics, Pharma, Life Sciences

InspireSemi Thunderbird

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

- Versatile platform delivers unprecedented capability
- 4 chip PCIe card delivers >6,000 64-bit CPU cores (FP64)
- ***Raw performance: 24 TFLOPS (FP64)***
- ***Energy efficiency: 75 GLOPS/watt (FP64)***
- ***Low latency: 10x faster***
- ***Determistic + Predictable Performance***
- ***Value: exceptional performance/$***

InspireSemi™

Sample Thunderbird customer & analyst feedback

BERKELEY LAB

"High-precision calculations are vital for generating reliable scientific data, which serves as the foundation for large language models. These include biomedical research, drug design, medical devices, climate change research, and applications that require deep simulation and modeling."
- Kathy Yelick, Director of Lawrence Berkeley National Laboratory (LBNL)

Argonne NATIONAL LABORATORY

"Double precision calculations are crucial for AI in science and engineering to ensure AI surrogate models are accurate and not hallucinating. The need for physics-based models is not going away soon."
- Charles Edward Catlett, Senior Computer Scientist (and Executive Director of the Trillion Parameter Consortium)

World Wide Technology

"With the momentum of AI and the convergence of AI and HPC, it is time to look outside the status quo and leverage a new technology base, like InspireSemi's Thunderbird product line. Thunderbird is ideal for workflows that require the highest performance and lowest power. It is easy to integrate, making it a valuable addition to the HPC and AI industry."
- Earl J. Dodd, Global HPC Business Practice Leader

Intersect360 RESEARCH

"The pursuit of AI has been a tremendous boon for high-performance architectures across the board. For pure AI investments, most of the attention is on GPUs, but many organizations are seeking a more versatile solution, built on processing elements that are suited to a variety of HPC, AI, and analytics workloads. This is where we see a market opportunity for companies like InspireSemi with its Thunderbird platform."
- Addison Snell, CEO

InspireSemi Thunderbird in the news

techradar.

Supercomputer-on-a-chip goes live: single PCIe card packs more than 6,000 RISC-V cores, with the ability to scale to more than 360,000 cores

- InspireSemi has announced the successful tapeout of the Thunderbird I Accelerated Computing chip for fabrication at TSMC

tom's HARDWARE

Thunderbird packs up to 6,144 CPU cores into a single AI accelerator and scales up to 360,000 cores — InspireSemi's RISC-V 'supercomputer-cluster-on-a-chip' touts higher performance than Nvidia GPUs

- The Holy Grail of supercomputing chip design is an architecture that combines the versatility and programmability of CPUs with the explicit parallelism of GPUs, and InspireSemi strives to achieve just that

TECHSPOT

Move over GPUs, with 1,536 cores the Thunderbird RISC-V CPU is ready to eat your lunch

- Open source enables small industries to participate in the accelerator boom

ALL ABOUT CIRCUITS

InspireSemi Announces Tapeout of Thunderbird Accelerated Computing Chip

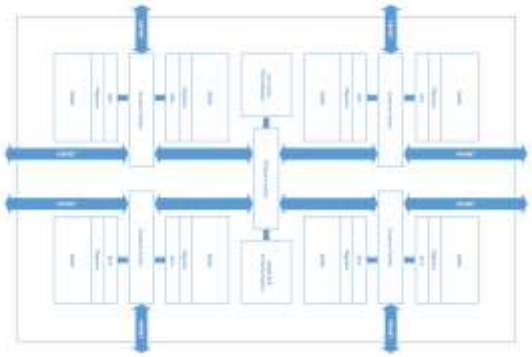
- InspireSemi's new compute chip couples the parallel processing of GPUs with the versatility of CPUs

JPR Jon Peddie Research

InspireSemi announces tape-out of RISC-V HPC chip

- InspireSemi's Thunderbird I RISC-V chip offers high-performance computing for underserved applications, with an emphasis on energy efficiency and competitive pricing

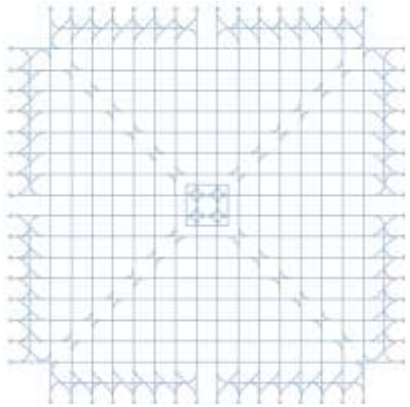
Thunderbird technology overview



Core

64-bit superscalar RISC-V CPU cores:

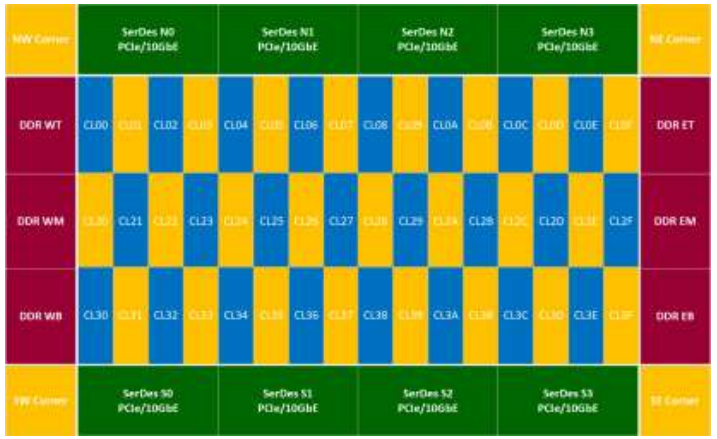
- Custom InspireSemi hi-perf design
- RV64IMAFDC, Privilege
- Multiple-issue, out-of-order, variable instruction width
- Mixed-precision floating point
- Cryptography & network extensions
- Plans to add vector, matrix extensions
- Tightly integrated memory and core-core network fabric
- Standard programming model



Interconnect Fabric

Manhattan street grid of 32-lane superhighways:

- Full utilization of precious routing area -> extreme bandwidth
- Flyover interchanges -> low congestion
- Express bypass lanes -> low latency
- Multiple onramps/offramps to each core
- 240TB/s local, 40TB/s global
- Uniform cellular layout
- Protocol supports AMO, RDMA, virtual channels



Top

- 1,536 CPU cores
- 64KB of SRAM per core
- SMP or HPC cluster-on-a-chip
- Network fabric extensible up 256 chips
- Six LP/DDR4 memory controllers
- 128 lanes PCIe Gen4 / chip-to-chip
- 10Gb Ethernet
- Algorithm-specific accelerators (e.g. – AES encryption)

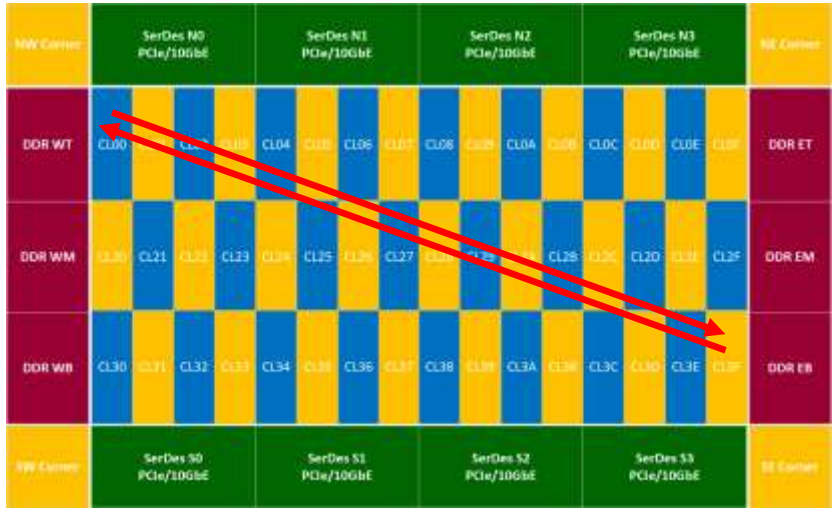
InspireSemi™

High speed, low latency Network-on-Chip (NoC) enables capabilities beyond conventional architectures

- Thunderbird designed to deliver real world application benefits
  - Software friendly, all CPU architecture (double precision FP64 RISC-V cores) will work with all HPC & AI software
  - High speed, low latency core-to-core communications for predictable performance
  - MIMD architecture (vs. high latency GPU SIMD)
  - Large memory – can address larger problems than fit in many GPUs
  - Distributed memory – each core has its own 64KB local fast memory
- *Result = Greater application performance with less power consumption*
- *Deterministic + predictable performance for workloads where GPUs fall short*

Example – Thunderbird vs. leading GPU latency

1-40 clock cycles

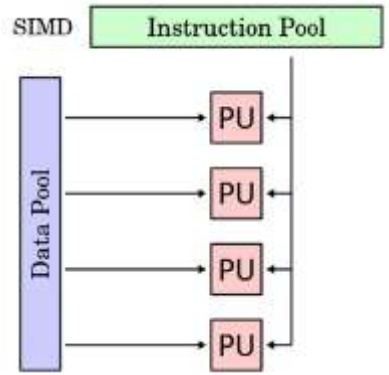


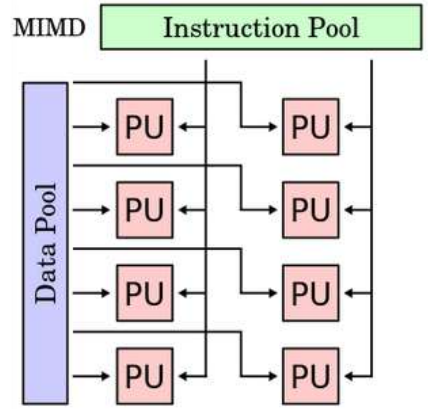
10x
greater efficiency

100-400 clock cycles

Thunderbird is deterministic, GPUs are not

- **Reproducibility of results is a must for many key applications, disqualifying GPUs**
  - E.g- financial trading, cryptography, robotics, healthcare imaging, self-driving cars, smart weapons, …
- Definition:
  - A deterministic system is one that always produces the same output for a given input
  - There is no randomness or chance involved in the system's behavior
  - Instead, the system follows predictable rules that determine how the inputs will be transformed into outputs
- GPU non-determinism problems
  - GPU's use very complex hardware/software task schedulers to try to hide their latency problems (switching between tasks while waiting for others)
  - These schedulers are not deterministic, meaning users cannot know when their tasks will finish
  - They're also proprietary, obscure, and change often, frustrating any attempt to predict their behavior (as with many aspects of GPU architecture)
  - Output results can also vary, due to varying order of sub-task completion and rounding errors
- **Thunderbird solves this**
  - Each core is controlled independently by its own program thread
  - Real-time operating systems provide deterministic software scheduling
  - Easy bare-metal programming gives developers absolute control down to single clock cycles
  - Atomic instruction set provides rich, straightforward task synchronization options
  - Delivering the same results, the same time, every time





Thunderbird performance specifications

6 – 8 TFLOPS
per chip (FP64)
24 – 32 TFLOPS
per card (FP64)

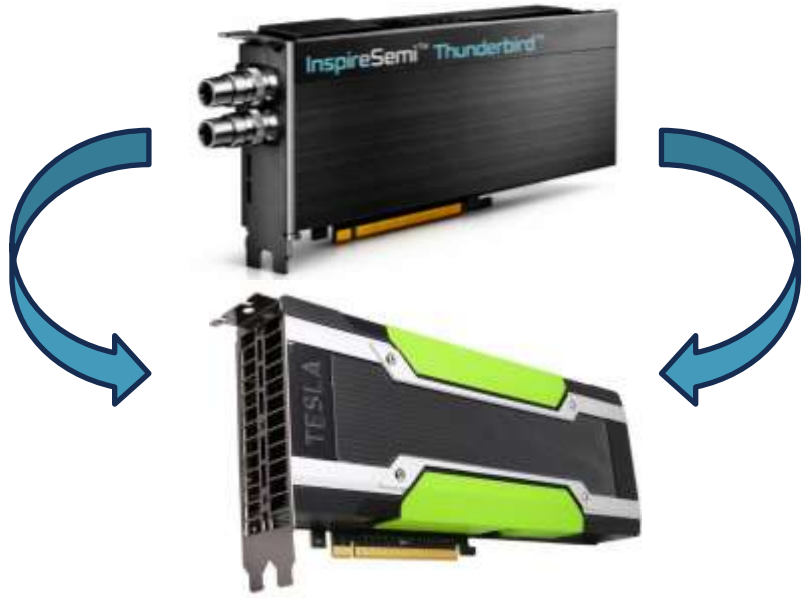
75 GFLOPS/ watt
(FP64)

10x
reduction in latency

Early customer feedback is unlike GPUs, Thunderbird is likely to hit its peak performance numbers due to its architecture and interconnect technology

Providing the “ground truth” for AI in science & engineering

Accelerating the datacenter HPC-AI market



- Versatile all-CPU architecture applicable to all HPC-AI software, much of which does not benefit from GPUs
- High precision, high performance, low power 64-bit native processors to solve “big math” problems required by most HPC software

- Thunderbird runs the highly accurate physics-based HPC simulations needed to train AI surrogate models for science and engineering
- GPUs run fast AI surrogate models to approximate long-running physics-based HPC simulations

World class supply chain partners

- TSMC - Wafer fab
  - World's largest semiconductor foundry
  - Developing the most advanced process nodes
  - Secured 12nm wafer capacity for 2026+
  - Secured 3nm wafer capacity for 2027+

- ASE – Chip package & test
  - Worlds largest and highest quality OSAT (Outsourced Semiconductor Assembly & Test)
  - Leading edge package design

- Imec: Value Chain Aggregator (VCA)
  - Enable early access to tier-1 supply chain
  - Support engineering and early-prod volumes

Thunderbird addresses *all* HPC-AI customer needs

Disrupting on everything that matters

	InspireSemi Thunderbird	CPU - Servers	GPU Accelerators	AI Accelerators	FPGA
Applicability / Versatility	Applicable to all HPC-AI applications since CPU based	Applicable to all HPC-AI applications since CPU based	Applicable to most AI applications but only some HPC apps	Applicable only to AI applications	Very limited, for prototyping and special cases
Performance	High for broad range of HPC apps	Slow, need h/w accelerators	High for AI and some HPC apps	High for AI only	Medium
Energy consumption	Low ~50W/chip (150W max)	High 650W+/chip (+ many servers)	Very High ~1400W+	High ~600W to Very High 20kW	Med ~300W
Scalability	256 chips	1-4 chips/server	2-8 chips	Varies	1 chip
Cost & Complexity	Low: one PCIe card vs. full rack of connected servers	High: many servers, InfiniBand, switches, cables, cooling …	Med to High: depends on vendor, form factor	Med to $Millions: depends on vendor, form factor	Med $8K-$10K
Programming model	Standard CPU-like, Any language, Full instruction set	Standard CPU, Any language, Full instruction set	Specialized C variant (CUDA, ROCM, SYCL)	Proprietary, obscure	Hardware description language (Verilog)
Software ecosystem	Open-source, Linux, compilers, libraries, AI frameworks, existing applications	Robust	Limited, proprietary	AI frameworks and proprietary software stacks	None

Disclaimer

This presentation contains statements which constitute “forward-looking information” within the meaning of applicable securities laws, including statements regarding the plans, intentions, beliefs and current expectations of InspireSemi with respect to future business activities and operating performance.

Often, but not always, forward-looking information can be identified by the use of words such as “plans”, “expects”, “is expected”, “budget”, “scheduled”, “estimates”, “forecasts”, “intends”, “anticipates”, or “believes” or variations (including negative variations) of such words and phrases, or statements formed in the future tense or indicating that certain actions, events or results “may”, “could”, “would”, “might” or “will” (or other variations of the forgoing) be taken, occur, be achieved, or come to pass. Forward-looking information includes, but is not limited to, information regarding: (i) the business plans and expectations of the Company including expectations with respect to production and development; and (ii) expectations for other economic, business, and/or competitive factors. Forward-looking information is based on currently available competitive, financial and economic data and operating plans, strategies or beliefs as of the date of this presentation, but involve known and unknown risks, uncertainties, assumptions and other factors that may cause the actual results, performance or achievements of InspireSemi, to be materially different from any future results, performance or achievements expressed or implied by the forward-looking information. Such factors may be based on information currently available to InspireSemi, including information obtained from third-party industry analysts and other third-party sources, and are based on management’s current expectations or beliefs. Any and all forward-looking information contained in this presentation is expressly qualified by this cautionary statement.

Investors are cautioned that forward-looking information is not based on historical facts but instead reflect InspireSemi’s management’s expectations, estimates or projections concerning future results or events based on the opinions, assumptions and estimates of management considered reasonable at the date the statements are made. Forward-looking information reflects InspireSemi’s current beliefs and is based on information currently available to it and on assumptions it believes to be not unreasonable in light of all of the circumstances. In some instances, material factors or assumptions are discussed in this presentation in connection with statements containing forward-looking information. Such material factors and assumptions include, but are not limited to: the impact of the COVID-19 pandemic on the Transaction or the company; the ongoing conflict between Russia and Ukraine and any actions taken by other countries in response thereto, such as sanctions or export controls; and anticipated and unanticipated costs and other factors referenced in this presentation and the Filing Statement, including, but not limited to, those set forth in the Filing Statement under the caption “Risk Factors”. Although the Company has attempted to identify important factors that could cause actual actions, events or results to differ materially from those described in forward-looking information, there may be other factors that cause actions, events or results to differ from those anticipated, estimated or intended. Forward-looking information contained herein is made as of the date of this presentation and, other than as required by law, the Company disclaims any obligation to update any forward-looking information, whether as a result of new information, future events or results or otherwise. There can be no assurance that forward-looking information will prove to be accurate, as actual results and future events could differ materially from those anticipated in such statements. Accordingly, readers should not place undue reliance on forward-looking information. Should one or more of these risks or uncertainties materialize, or should assumptions underlying the forward-looking information prove incorrect, actual results may vary materially from those described herein as intended, planned, anticipated, believed, estimated or expected.