THE SOCIETY OF HPC PROFESSIONALS

# InspireSemi™

## Disruptive Next Generation Accelerated Computing Platform

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

## Thunderbird Compute Accelerator for HPC, AI, & Graph Analytics

May 30, 2024

# Accomplished Leadership Team

## Alexander Gray, Founder, President & CTO

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

SolarBridge Technologies SUNPOWER

## Ron Van Dell, CEO

- 40 years experience and an exceptional track record of success and proven leadership skills in early-stage, turn-around and established businesses
- Former CEO of Primarion (Infineon), SolarBridge, and several other semiconductor and hardware startups
- GM Dell, VP-GM of Communication Products at Harris Semi (Intersil/Renesas)
- BSEE Michigan Technological University

SolarBridge Technologies primarion HARRIS DELL GE

## John B. Kennedy, CFO

- 30+ years experience in tech startups and public companies
- Trilumina, SolarBridge, Primarion, KPMG
- BS Accounting & Finance, Elmira College, NY

TRILUMINA SolarBridge Technologies primarion KPMG

## 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
- Nimbix, Newisys (Sanmina), CoWare, Cadence, IBM
- President of Society of HPC Professionals, Technology Advisors Group Austin, TEXGHS Innovation Consortium
- RISC-V International: member SIG-HPC & Marketing team
- BSEE, Missouri University of Science and Technology

NIMBIX VIRTUAL INSTRUMENTS NEWISYS IBM
CoWare cadence

# 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.

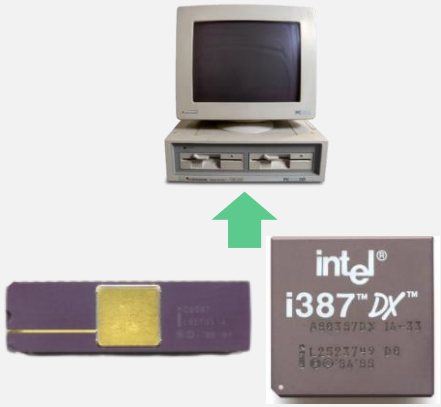
Often, but not always, forward-looking information can be identified by the use of words such as “plans”, “expects”, “is expected”, “budget”, “scheduled”, “estimates”, “forecasts”, “intends”, “anticipates”, or “believes” or variations (including negative variations) of such words and phrases, or statements 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.

# The Third Wave of Accelerated Computing is Here

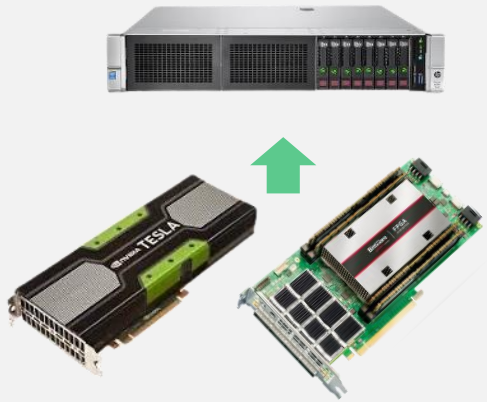
*Thunderbird for HPC, AI, Graph Analytics*

## 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

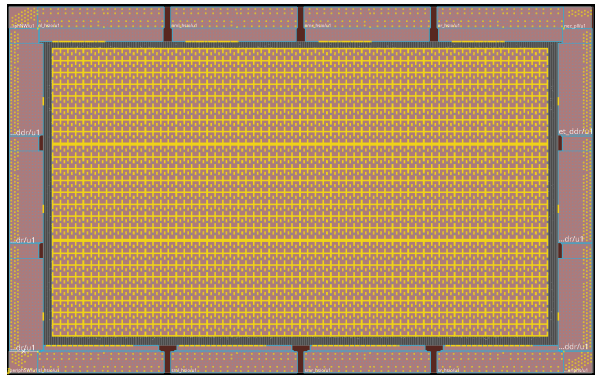
## 2024+
## Thunderbird


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

# Thunderbird Accelerated Computing Solution

- ***Ultra-Efficient, Ultra-compact custom CPU cores***
  - Based on RISC-V instruction set (like ARM but open standard)
  - Modern superscalar, out-of-order, vector-capable cores
  - Ability to add custom instructions
- ***A supercomputer cluster-on-a-chip***
  - 1,536 high-perf 64-bit CPU cores per chip (>6,000 per PCIe card)
  - Comparable to GPU shader count but are independent CPUs
- ***Innovative high speed interconnect fabric***
  - Key to efficient utilization of so many cores
  - Seamlessly spans multiple-chip arrays up to 256 chips!
- ***Energy efficiency: 30-60% power reduction***
  - Focus on performance/watt
  - Higher energy efficiency, fits in current datacenters
- ***Supporting existing open RISC-V software ecosystem***
  - Enables customers to easily adapt their 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

World Wide Technology

E4 COMPUTER ENGINEERING

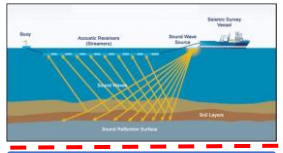
PENGUIN COMPUTING
A subsidiary of SMART Global Holdings, Inc.

2crsi

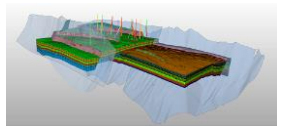
# Addressing the Need to Accelerate All HPC & AI Software

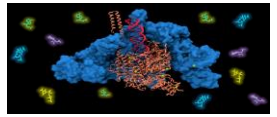
*What customers always wanted…Not "yet another GPU"*

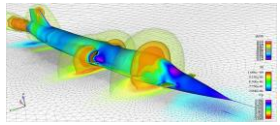

Financial simulations

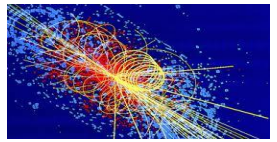

Geology: Seismic

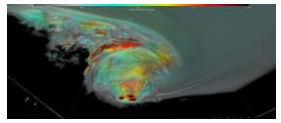

Financial Trading & Graph Analytics


Energy: Reservoir Modeling & Sim

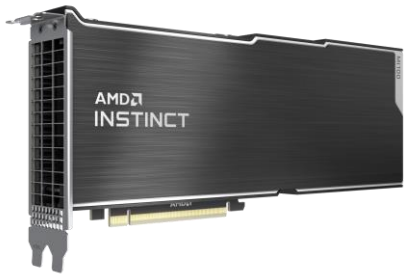

Genomics, Pharma, Life Sciences


CAE/Computational Fluid Dynamics


Government Lab Simulations


Climate & Weather Modeling

Datacenter GPUs

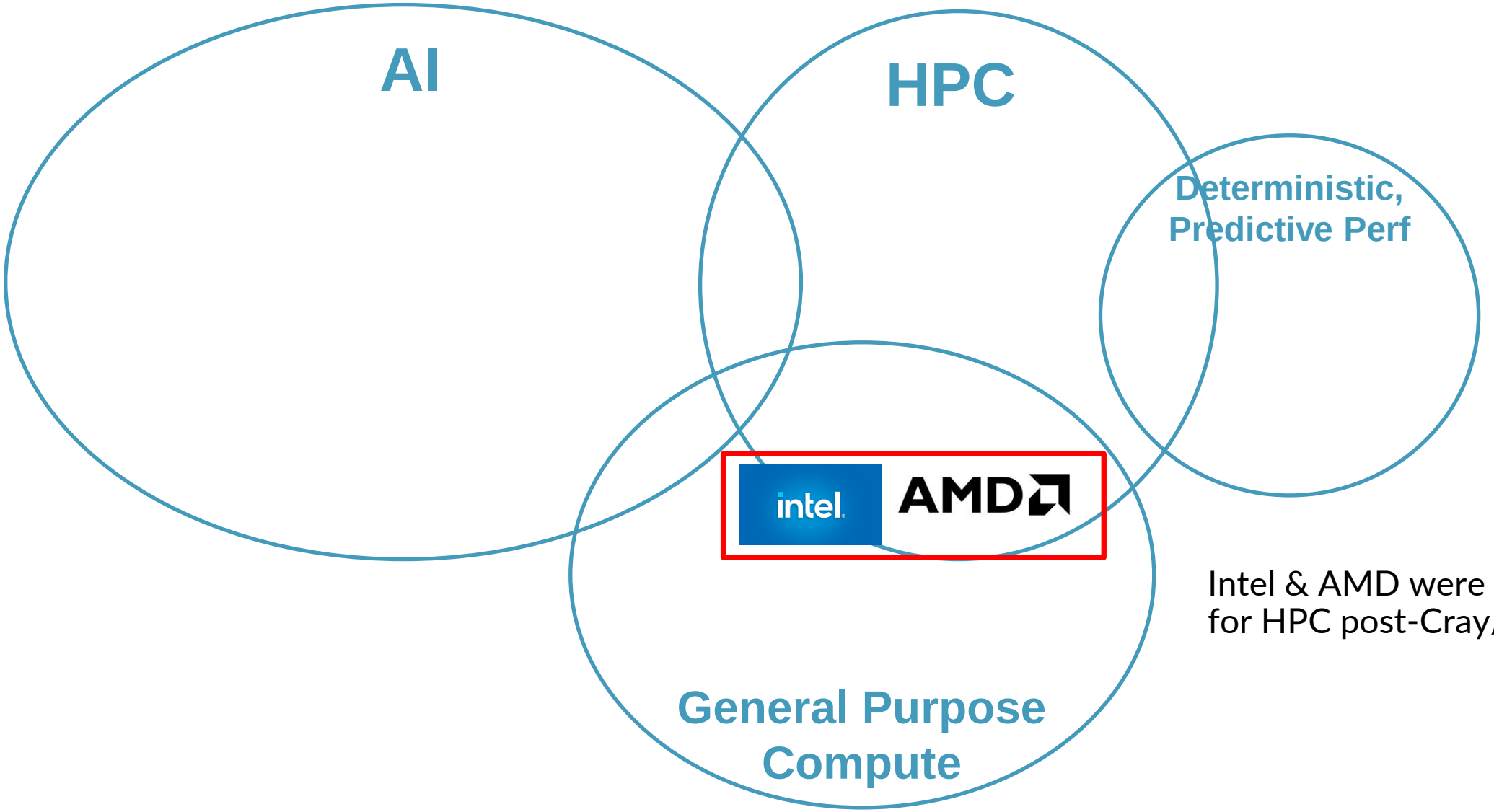




InspireSemi Thunderbird


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

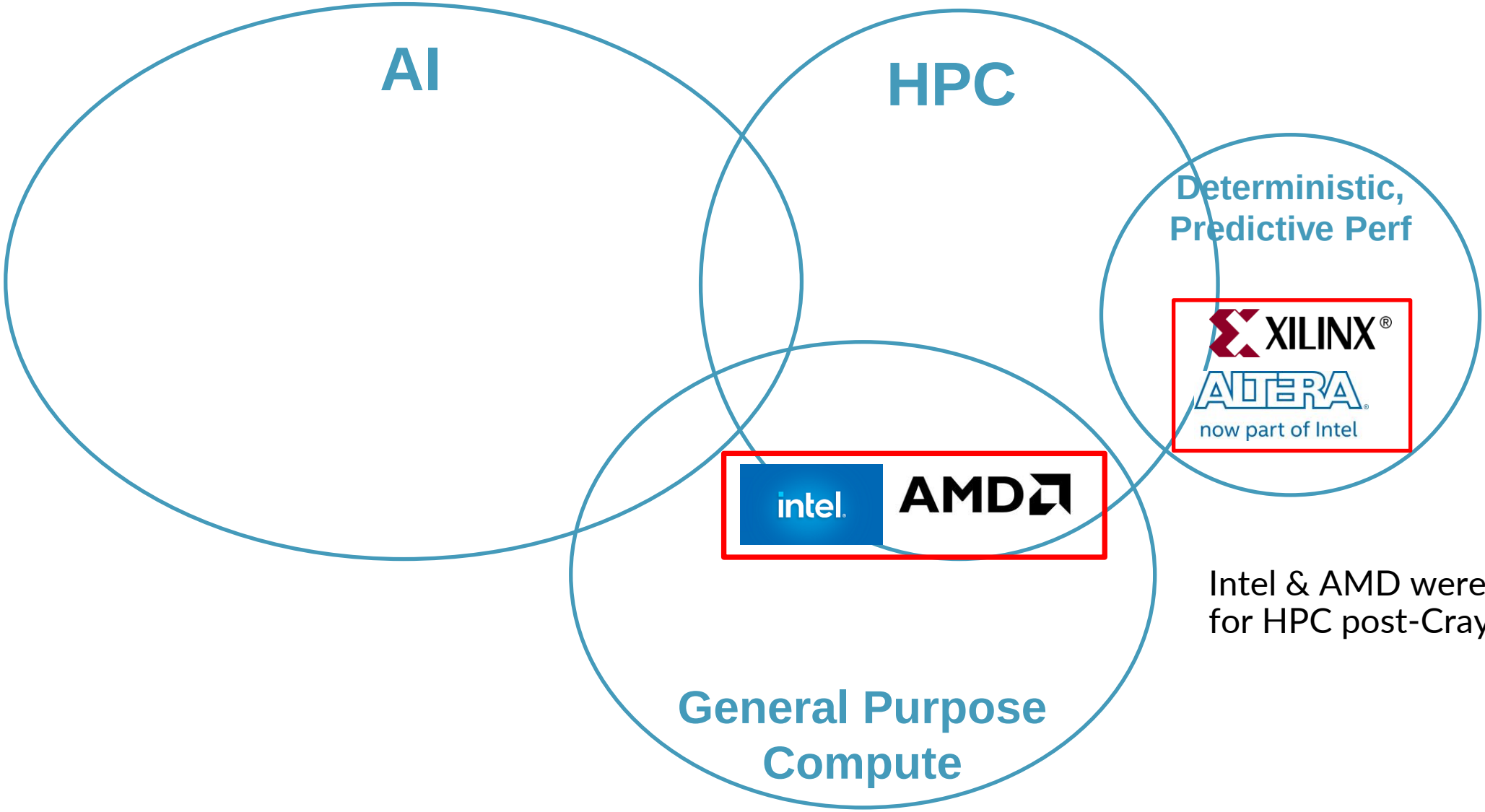
- Versatility as a platform across wide range of applications
- Determistic, with predictable performance, enables use in applications where GPUs fundamentally do not work
- 4 chip PCIe card delivers >6,000 interconnected 64-bit CPU cores
- Large scale computing power, scales up to 256 chips
- Best-in-class for both Performance/$ and Performance/Watt
- Delivers unprecedented capability within an established open software ecosystem

# Competitive Landscape




Intel & AMD were only real option for HPC post-Cray/DEC/Sun/SGI

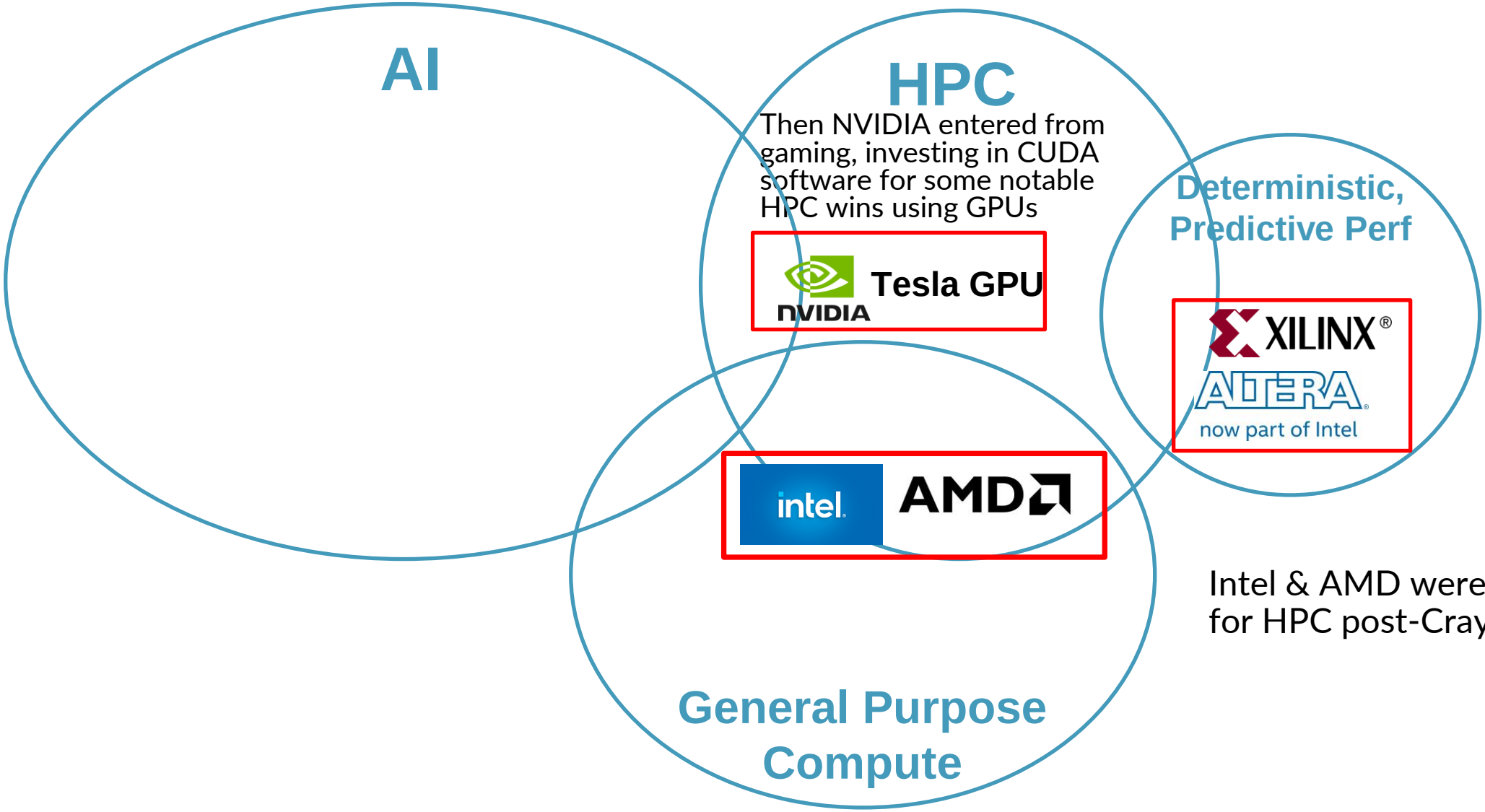
# Competitive Landscape




FPGAs had some niche HPC success

Intel & AMD were only real option for HPC post-Cray/DEC/Sun/SGI

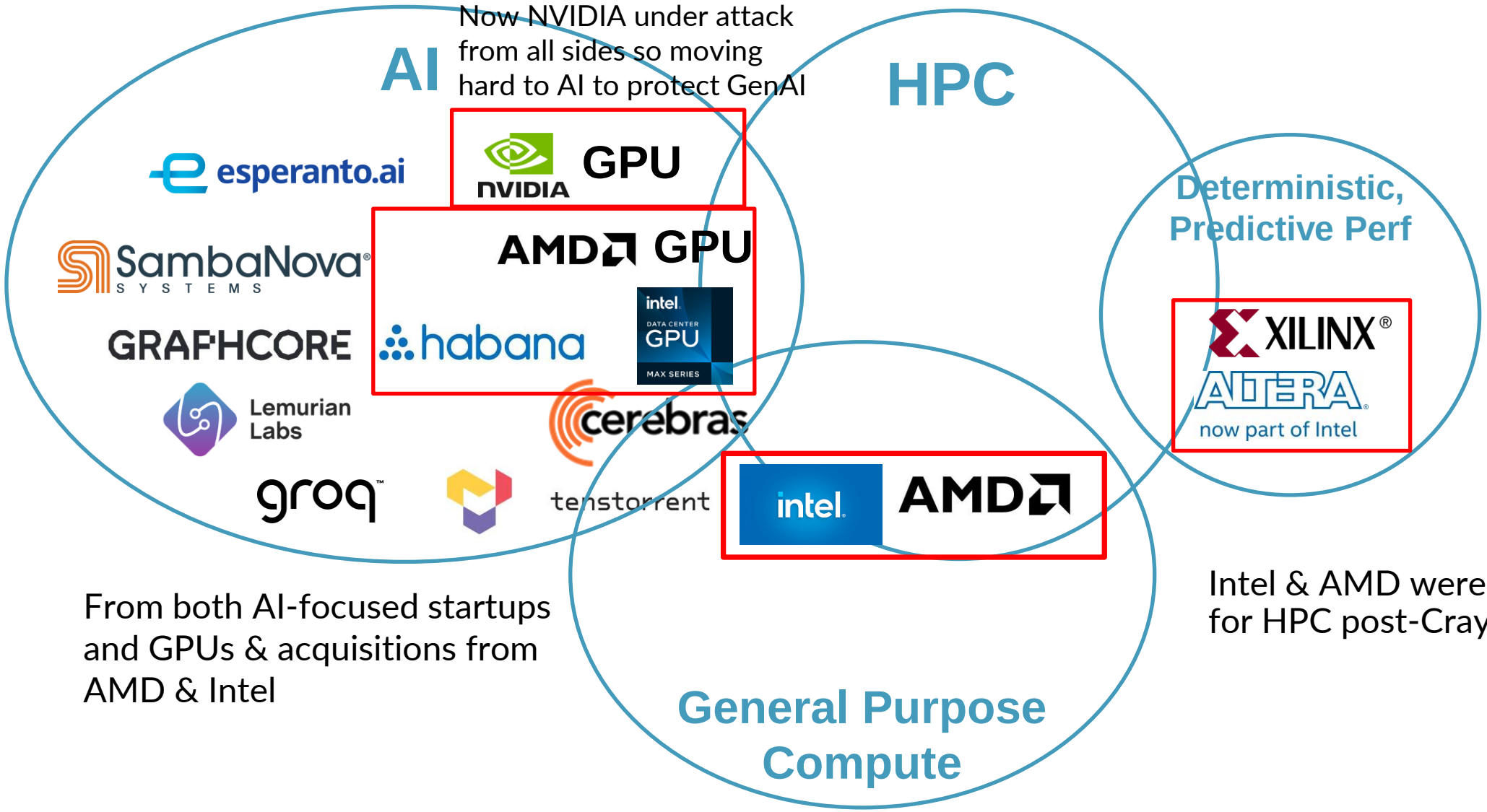
# Competitive Landscape




FPGAs had some niche HPC success

Intel & AMD were only real option for HPC post-Cray/DEC/Sun/SGI

# Competitive Landscape

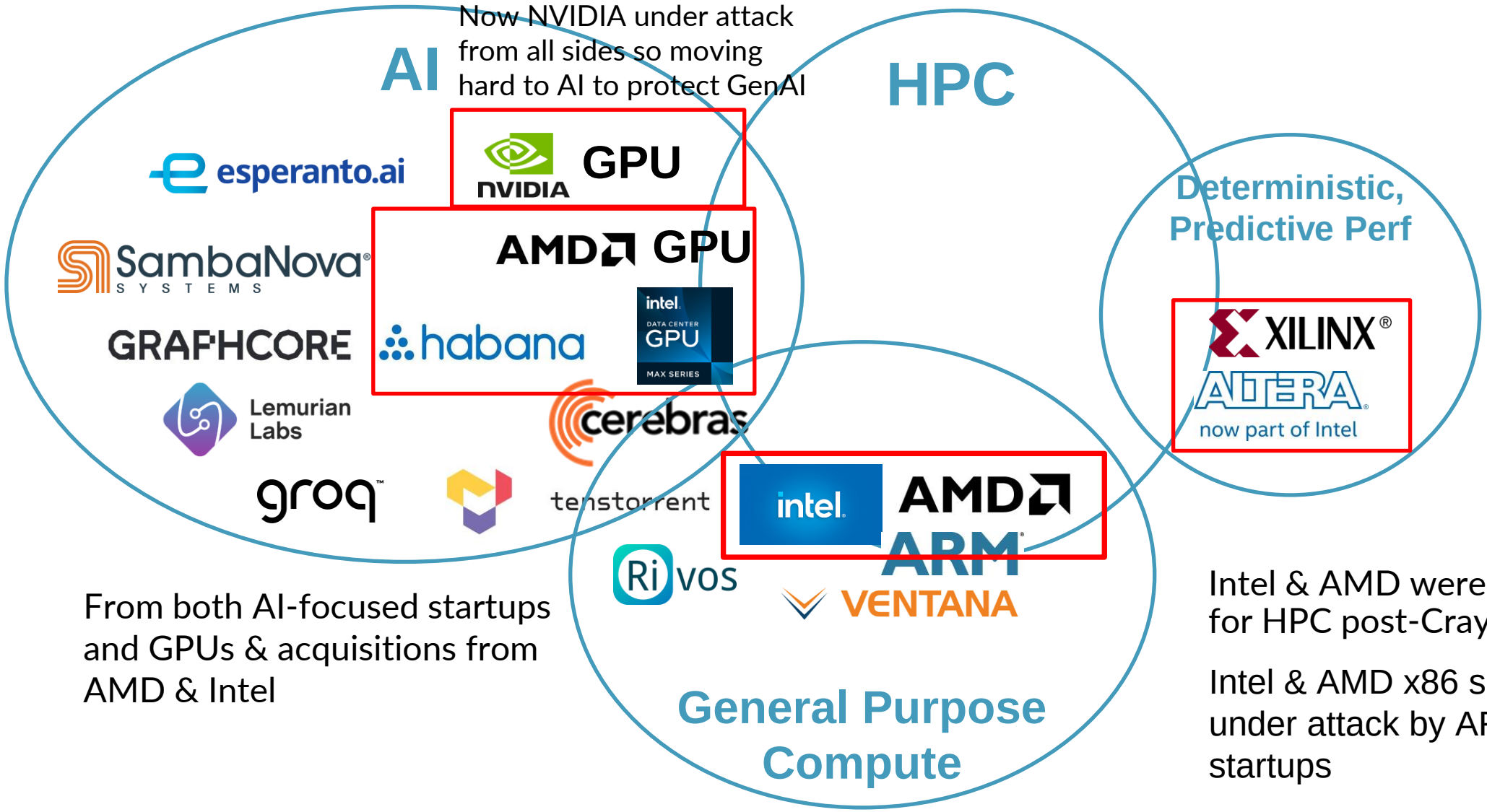



FPGAs had some niche HPC success

From both AI-focused startups and GPUs & acquisitions from AMD & Intel

Intel & AMD were only real option for HPC post-Cray/DEC/Sun/SGI

# Competitive Landscape



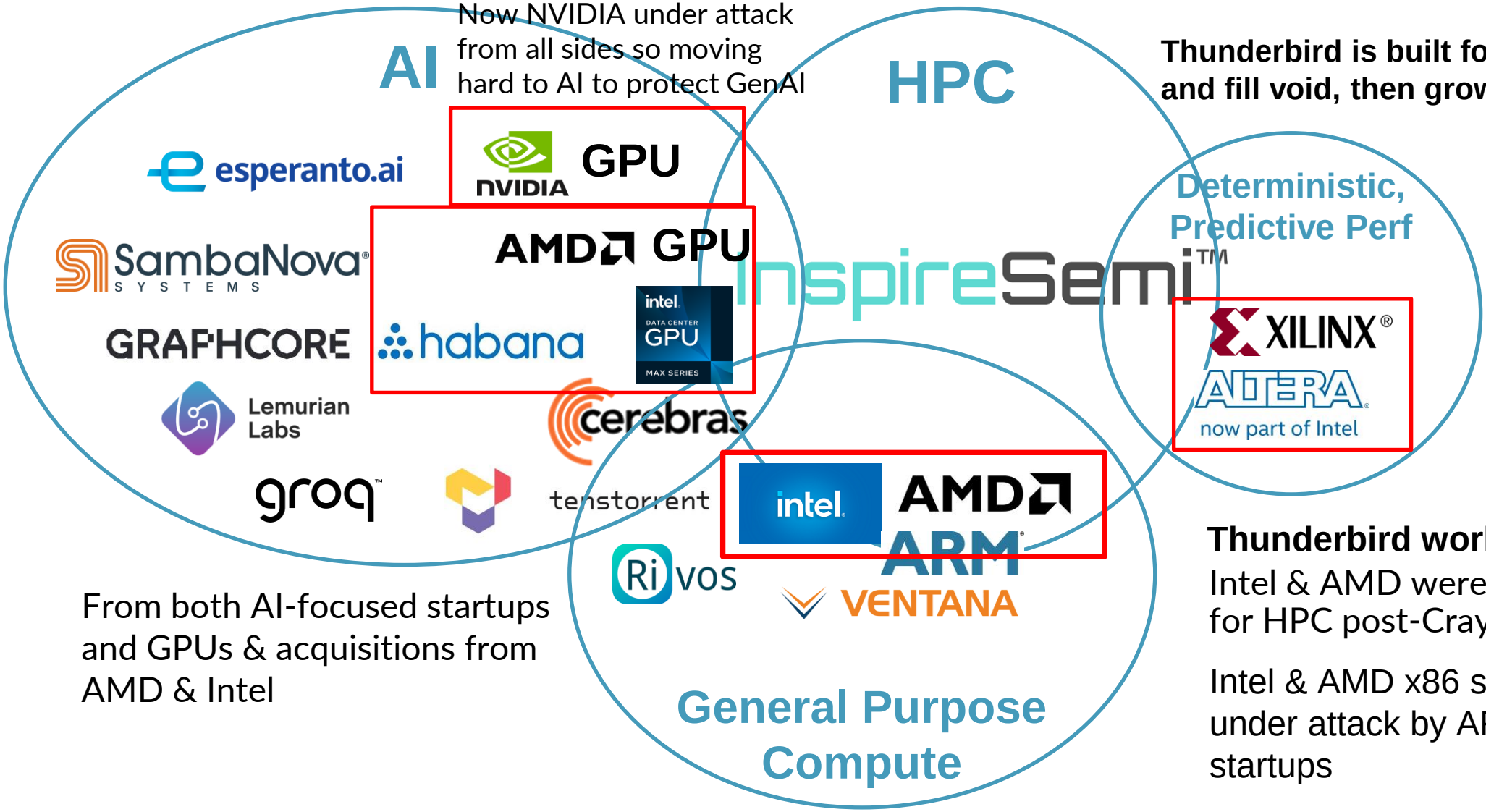

From both AI-focused startups and GPUs & acquisitions from AMD & Intel

FPGAs had some niche HPC success

Intel & AMD were only real option for HPC post-Cray/DEC/Sun/SGI

Intel & AMD x86 server sockets also under attack by ARM and RISC-V startups

# Competitive Landscape




**Thunderbird is built for HPC to disrupt and fill void, then grow in all dimensions**

FPGAs had some niche HPC success

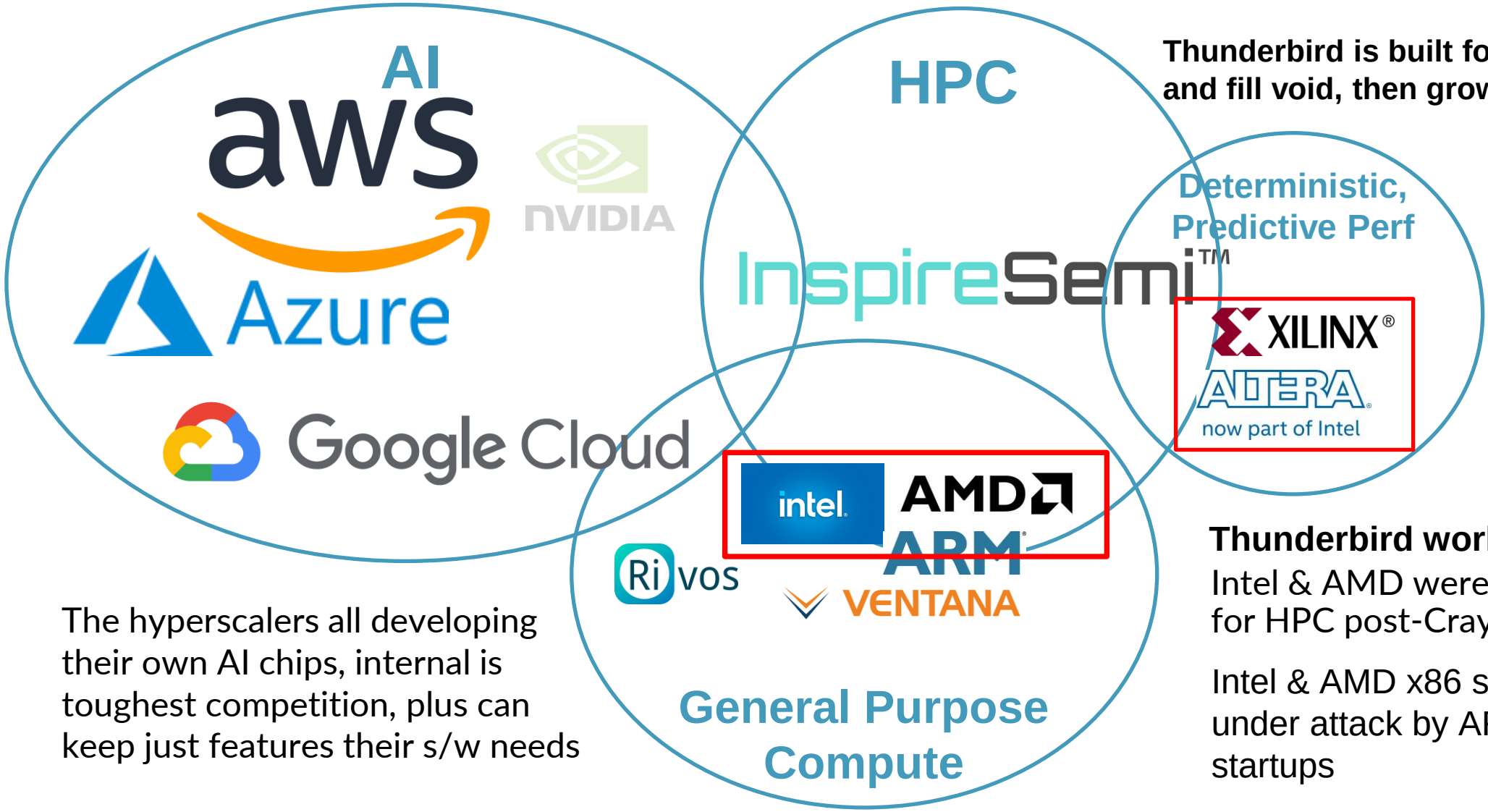
From both AI-focused startups and GPUs & acquisitions from AMD & Intel

**Thunderbird works in all servers**

Intel & AMD were only real option for HPC post-Cray/DEC/Sun/SGI

Intel & AMD x86 server sockets also under attack by ARM and RISC-V startups

# Competitive Landscape




**Thunderbird is built for HPC to disrupt and fill void, then grow in all dimensions**

FPGAs had some niche HPC success

The hyperscalers all developing their own AI chips, internal is toughest competition, plus can keep just features their s/w needs

**Thunderbird works in all servers**

Intel & AMD were only real option for HPC post-Cray/DEC/Sun/SGI

Intel & AMD x86 server sockets also under attack by ARM and RISC-V startups

# Sample Thunderbird Customer & Partner Feedback

"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."
*- Adison Snell, CEO*

"Sandia is pleased to have joined InspireSemi's early access program for their upcoming Thunderbird processor. This is enabling us to engage early by evaluating our challenging application workloads on Thunderbird's massively parallel interconnected CPU architecture, helping to mature the overall RISC-V HPC software stack, and providing our input for future versions of Thunderbird."
*- Kevin Pedretti, Principal Member of Technical Staff, Scalable System Software*

"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*

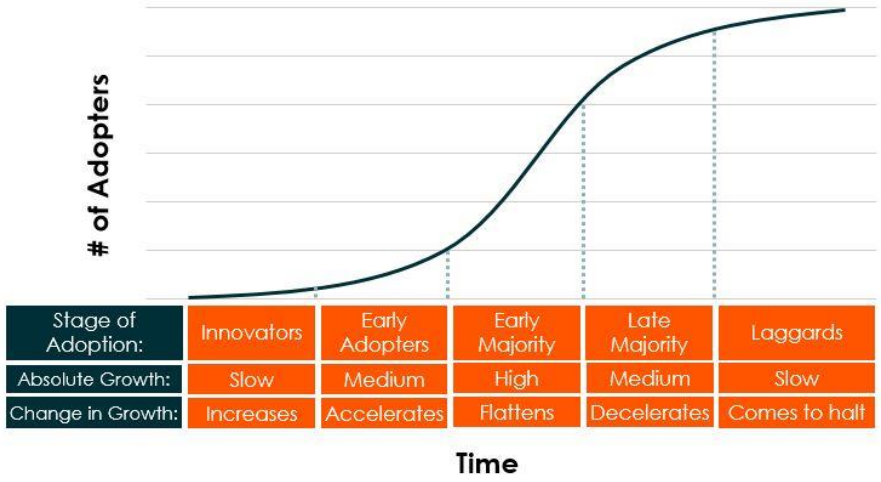
"The combination of your custom designed RISC-V 'sea of cores' plus high-speed interconnect fabric is very smart, and your decision to focus on HPC (and blockchain opportunistically) rather than AI/ML was likewise very smart."
*- Amit Nanda, Advanced Tech Sourcing*

"This will let us run large memory simulations with AI workflows. We will take anything you can give sooner."
*- Rajeev Thakur, Deputy Director, Data Science and Learning Division*

# Thunderbird Early Access Program

*Thunder and Lightning*

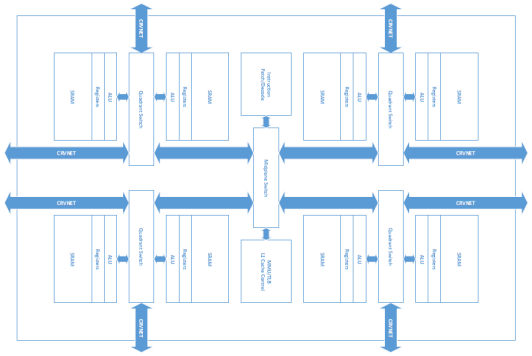
- Objectives
  - Jumpstart Thunderbird adoption and software applications
  - Allocate early hardware for key customers & partners
  - Key customer feedback for Thunderbird 2
- Overview
  - Program capped to provide good support to EAP partners
  - Must identify relevant application (esp. key open source codes)
  - Must dedicate lead and technical resource(s)
- Deliverables
  - Address-accurate QEMU model
  - FPGA emulator dev board w/limited number of CPU cores (“thunder”)
  - Early Thunderbird PCIe board once available (“lightning”)
  - Relevant software for FPGA and PCIe boards
  - Support, regular review meetings
- Status
  - Strong interest, already receiving PO’s and commitments
  - Shipping emulator board to early customers

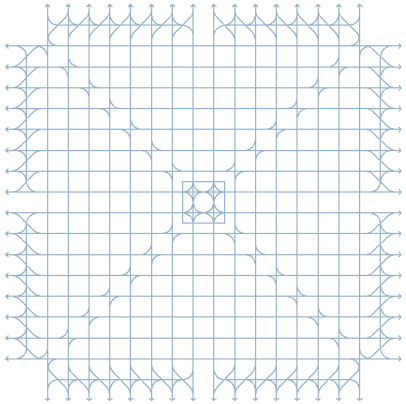


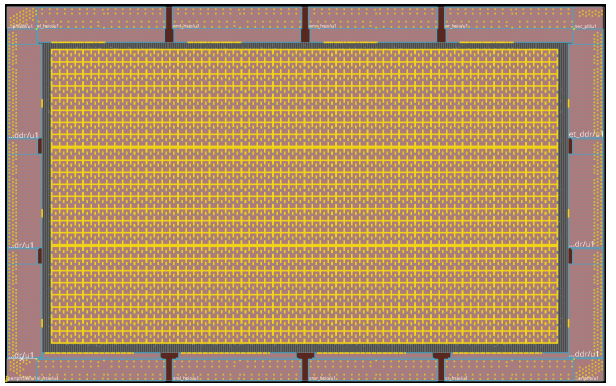
# What is Thunderbird and Why is it Better ?

- Unprecedented performance for HPC applications
  - RAW horsepower per chip
  - Greater effective processing power due to 5x better use of "nameplate" rating
  - 4 chips on a standard format server add-in PCIe card
- Best in class energy efficiency
- Disruptive price point
- Versatile support of wide range of leading HPC sectors within an open-source software ecosystem
- Confident porting of existing programs to the Thunderbird architecture
  (CPU code on a CPU-based system vs. re-architect)
- Double precision math – required for many HPC workloads
- Fully deterministic processing, which GPU-based competition cannot do
  - Required for financial trading, cryptography, robotics, smart weapons, healthcare imaging, self-driving cars, …

# 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 AI, Vector extensions
- Tightly integrated memory and core-core network fabric
- Simple 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

## Top

- 1,536 CPU cores, SMP or HPC cluster-on-a-chip
- Network fabric extensible up 256 chips
- Six DDR4 memory controllers
- 128 lanes PCIe Gen4 / chip-to-chip
- Algorithm-specific accelerators
- Fully routed
- DRC clean

# Thunderbird Architecture – Top Level

*High performance, robust, and scalable*

**1,536 independent CPU cores**

- Organized in 24 clusters of 64 cores
- Symmetric multiprocessor or HPC cluster-on-a-chip models

**Chip-level algorithm-specific accelerators**

**Network fabric extensible to multichip arrays up to a million cores**

- ~240 TB/s local bandwidth
- ~40 TB/s global long-reach, random-traffic bandwidth

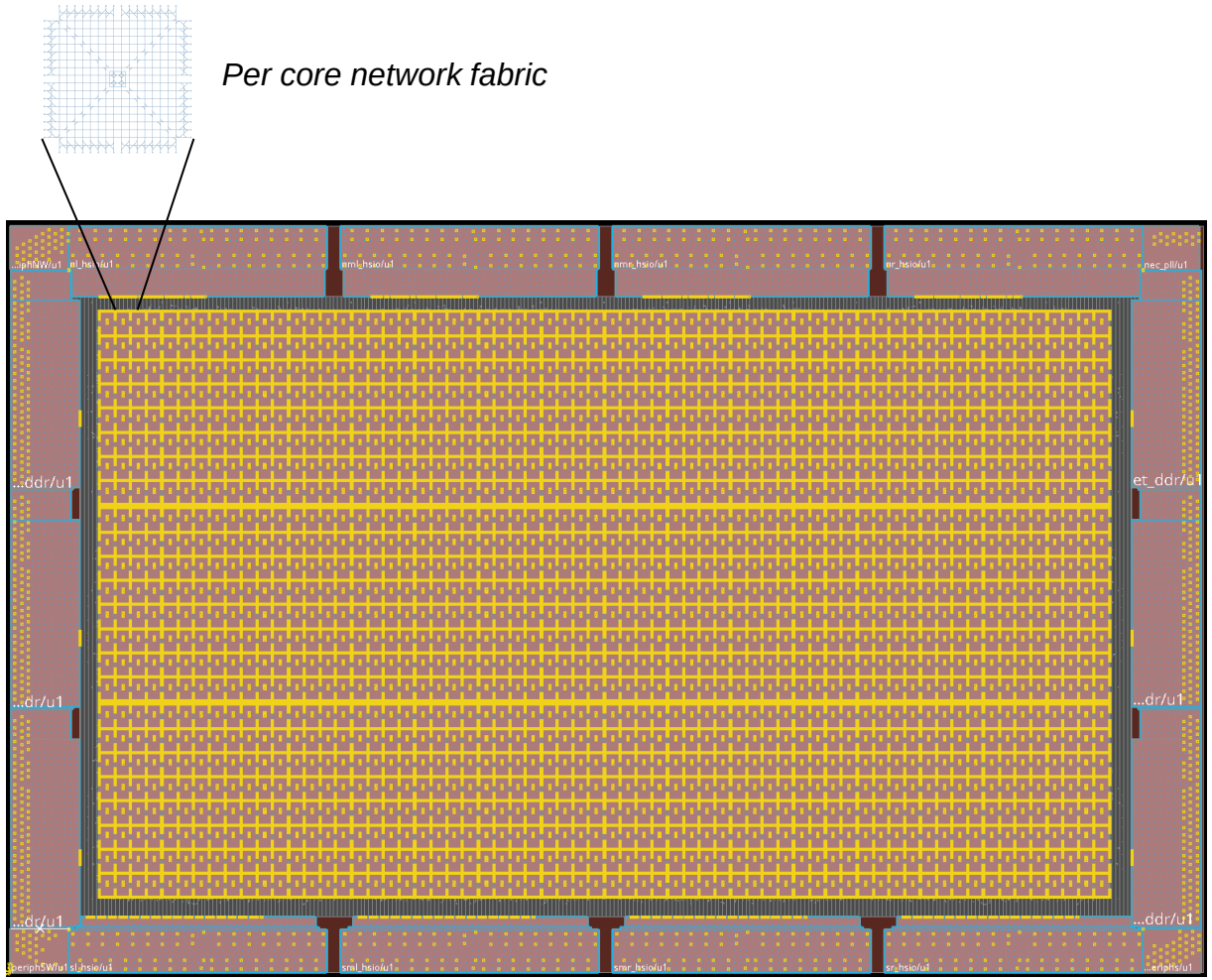
**Six channels DDR4 DRAM memory controllers:**

- Commodity memory chips keep system cost low
- Many narrower channels for high random-access performance

**128 lanes 16 Gbps SerDes**

- 64 lanes PCI express Gen4
- Ethernet 10Gb
- Multi-chip array interconnect

**UARTs, QSPI, SD card**



# Thunderbird Architecture – Cores

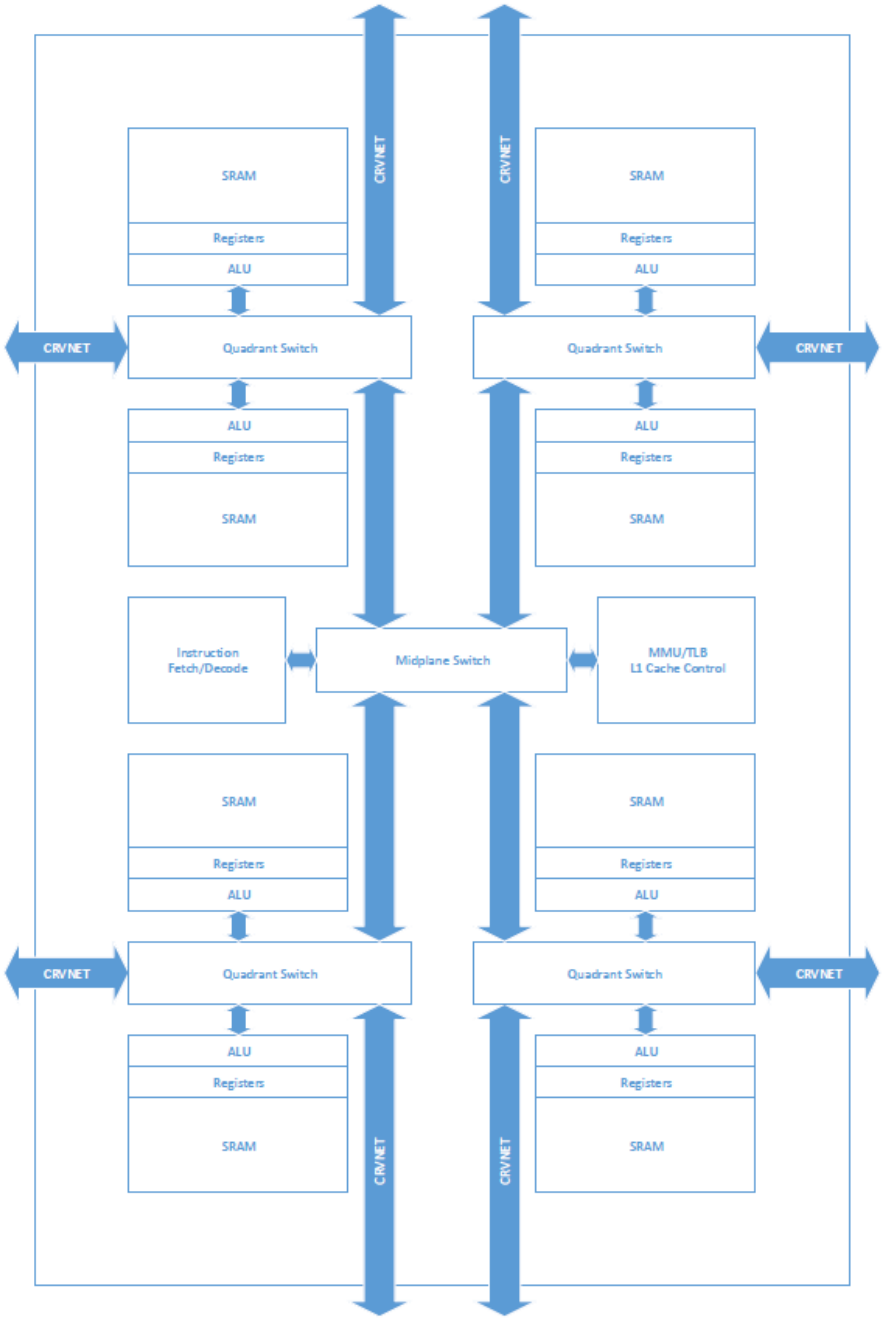
## *Ultra-compact CPU cores with performance and efficiency of an ASIC*

InspireSemi-designed hi-perf 64-bit superscalar CPU cores:

- Multiple-issue, out-of-order execution, variable instruction width
- Short/variable pipeline permits no speculation, simple branch prediction
- 1GHz clock balances high throughput with low power
- Compact unified ALU's for integer, floating point, and SIMD/vector
- 8-lane (512b) vector/SIMD operations
- Custom AI and cryptography extensions
- Tightly integrated SRAM memory and inter-core network fabric
- Efficient instruction fetch, decode, and reordering

RISC-V open-source instruction set:

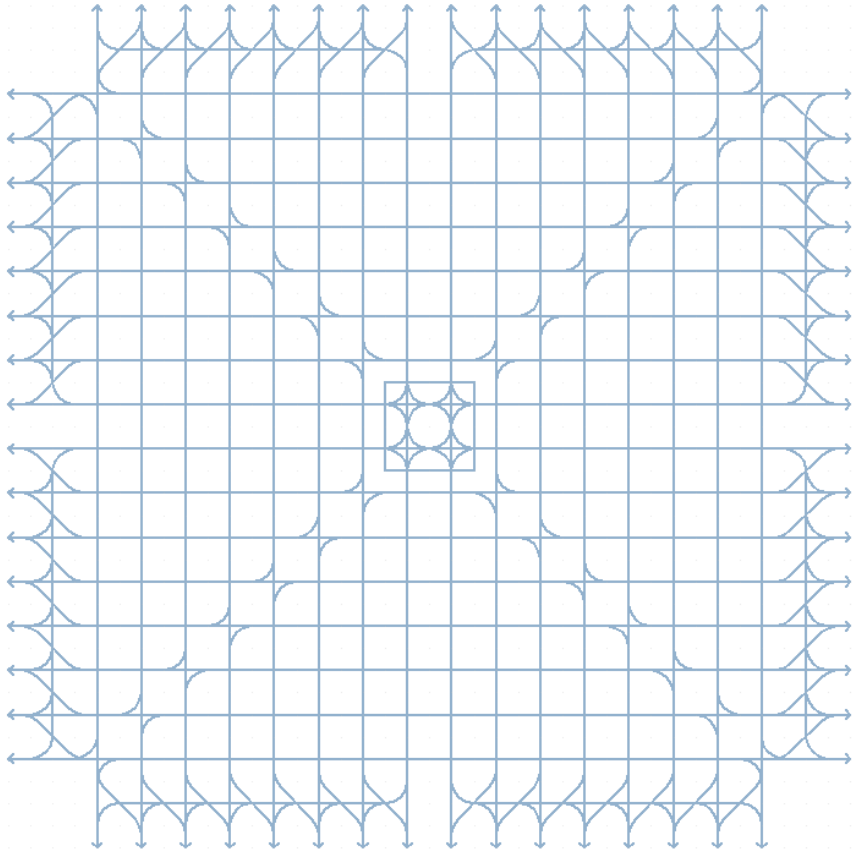
- Royalty free, freely extensible, free of vendor dependency
- Thriving open-source software ecosystem
- Simple programming environment, software porting
- Adopted by AliBaba, Western Digital, MIPS, many others



# Thunderbird Architecture – Network Fabric

*High bandwidth with low latency saves idle cycles and power*

- Extensible to multichip arrays up to a million cores
- Packet network avoids redundant wires of common busses
- Travel over multiple cores per cycle
- Travel in multiple directions per cycle (loop-free)
- Constant distance between repeaters
- 240 TB/s local bandwidth
- 40 TB/s long-reach global bandwidth
- Single-cycle latency within 64-core clusters
- Protocol supports advanced features: AMO, RDMA
- Uniform core/net cells, design once and tile up
- Abuttable with no/minimal routing channel gaps
- Opportunity for distributed virtual cache hierarchy

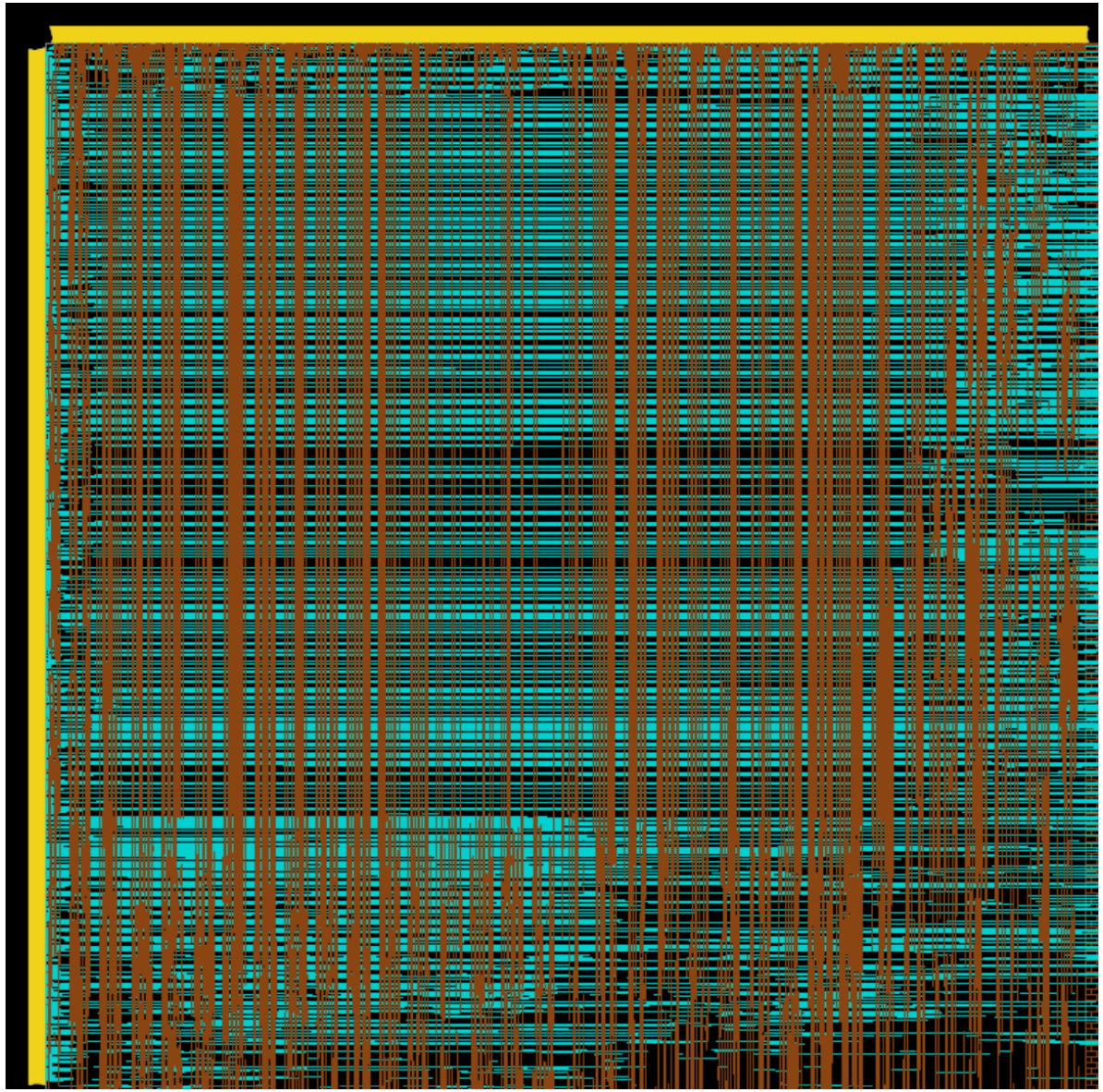


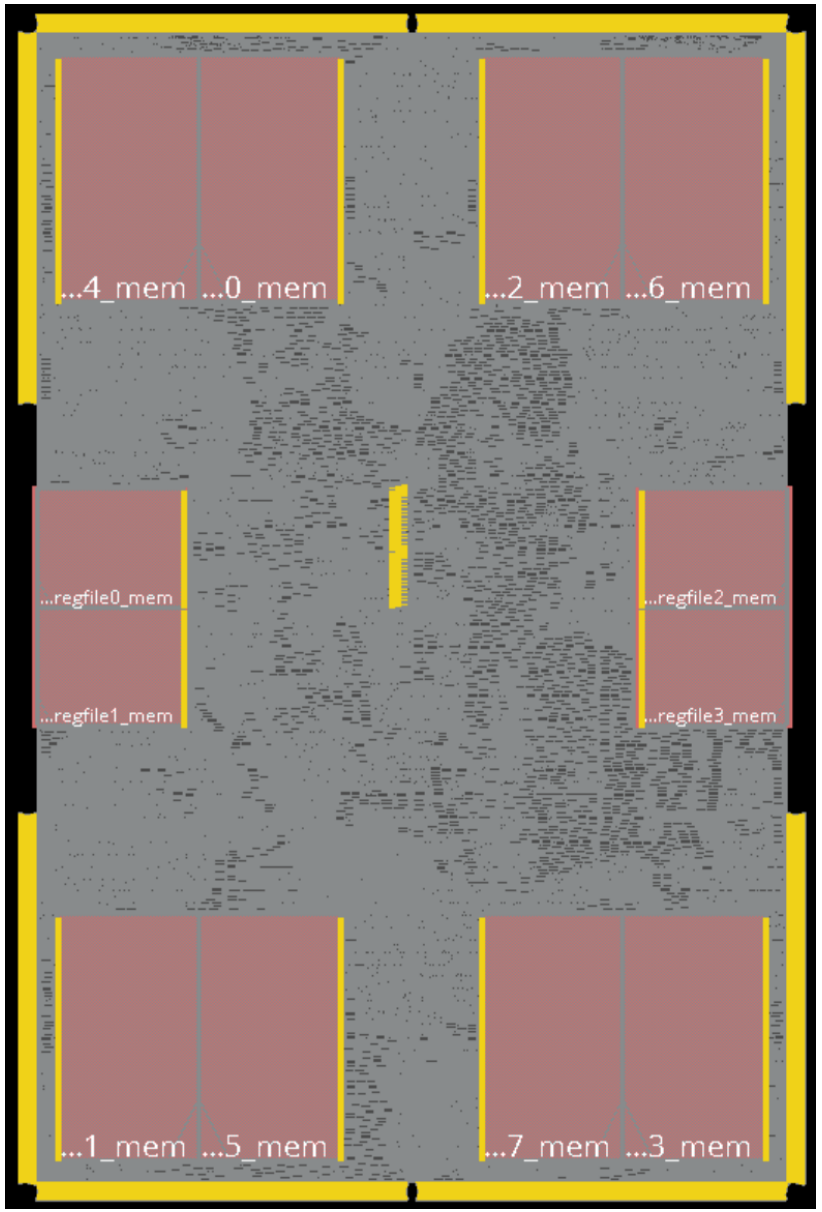
*Per core network fabric*

# Physical Implementation

- Absolutely essential for large, high performance systems.
  Anyone could infer 1M wires in RTL
- Our core tile features 16 CRVnet lanes in each direction
  (64 inbound and 64 outbound total)
  - 340um x 340-600um
  - 8k data wires, only ~400 strobe/ready overhead signals
- Lanes cross 4 cores per cycle, staggered to limit maximum path length
  - Network synchronous with core clocks, ~1GHz
- Cores can inject traffic on any lane, only need to receive on 8
- Proper pin placement in floorplan enables core macro abutment
  - Near 100% utilization of 2 Y-metal layers, with 80% of beachfront tracks as pins for max bandwidth
  - Y metals ideal for 1-2mm routes with a few buffers; same principles apply to coarser/finer metals
  - Within density rules and with room for power grid, vias, etc.

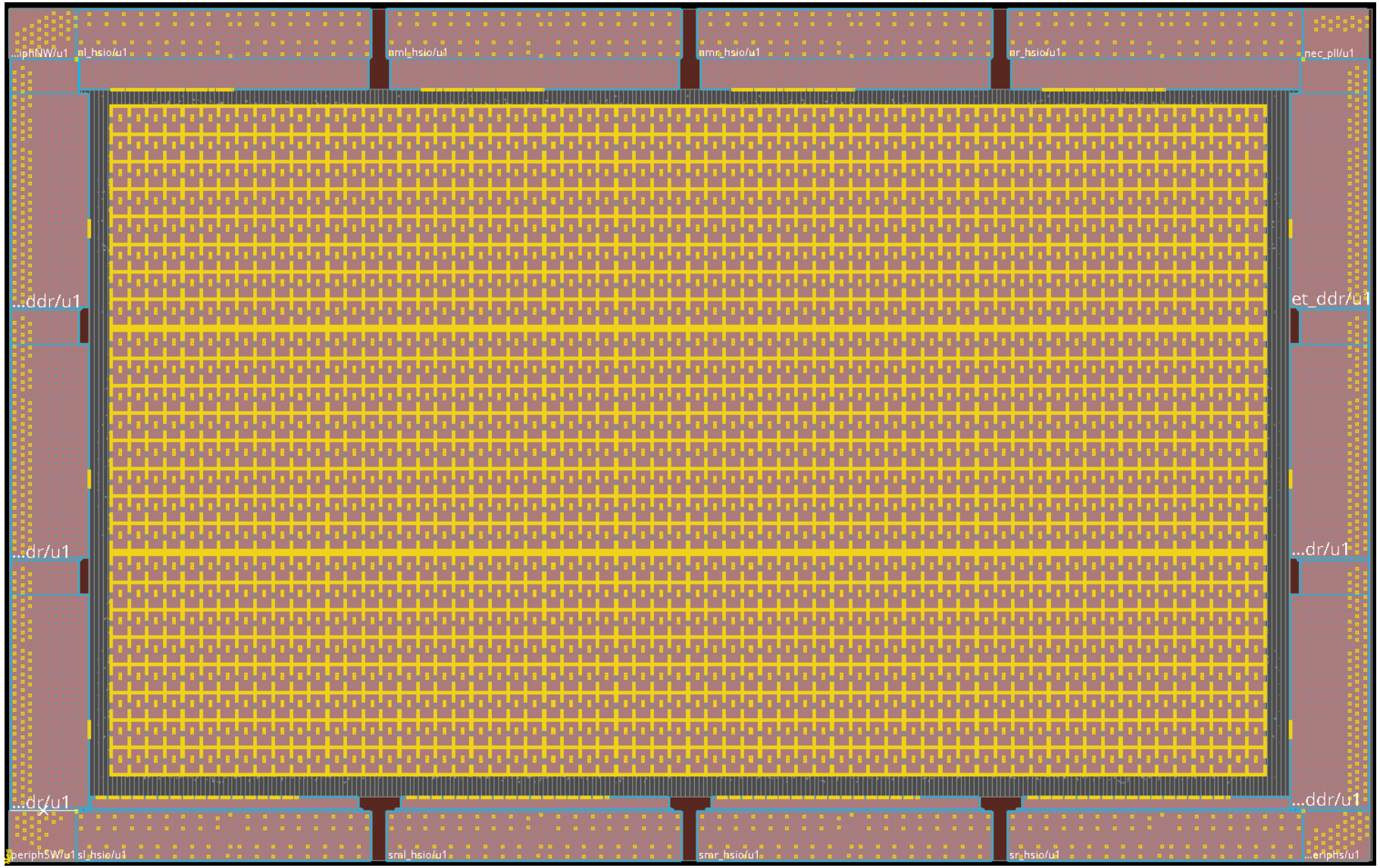
# Core Tile Floorplan & CRVnet Pins/Routes





# 12nm Dry Run Tapeout Complete

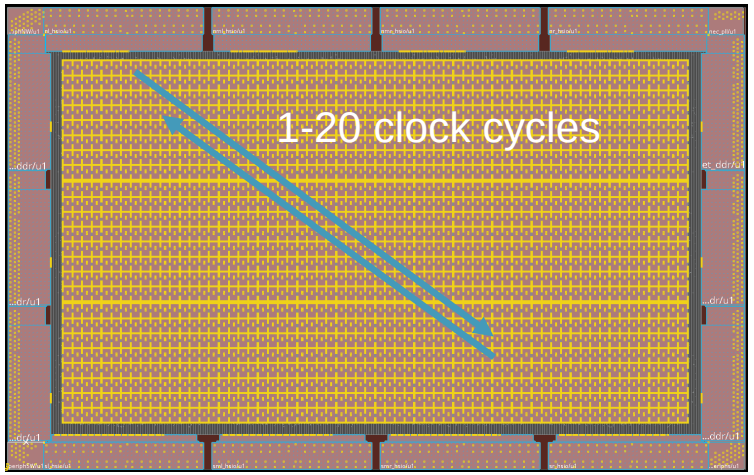
- 1536 core tiles
- 128 CRVNet lanes (~8k pins) per core
- 6x DDR4
- 128 x 16G SerDes
- 26x16mm die
- Fully routed
- DRC clean



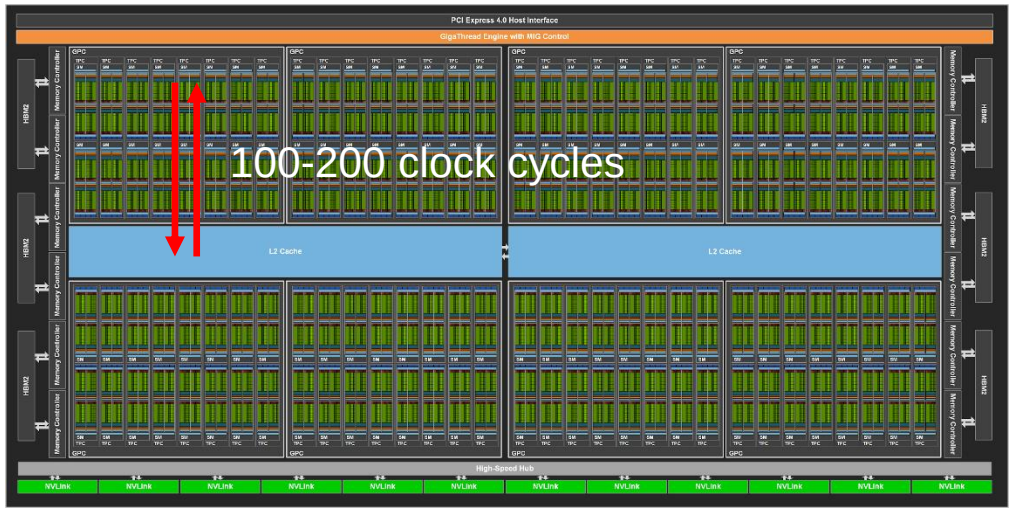
# More Applications, High Utilization, and Low Latency

- 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 GPUs
  - Distributed memory – each core has its own 64KB local fast memory
- ***Result = Greater application performance with less power consumption***
- ***Deterministic + Predictable Performance addresses applications GPUs cannot***

Example – Thunderbird vs. leading GPU latency

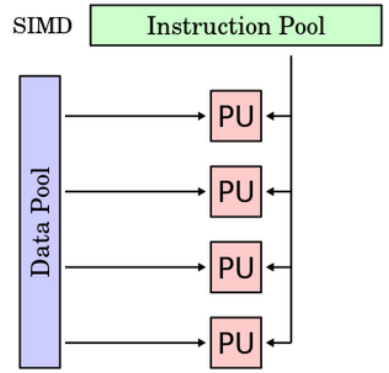


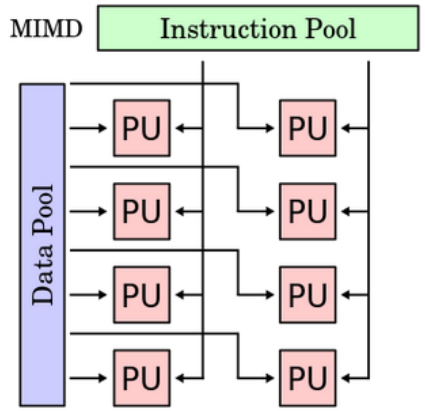

**20x**
*greater efficiency*



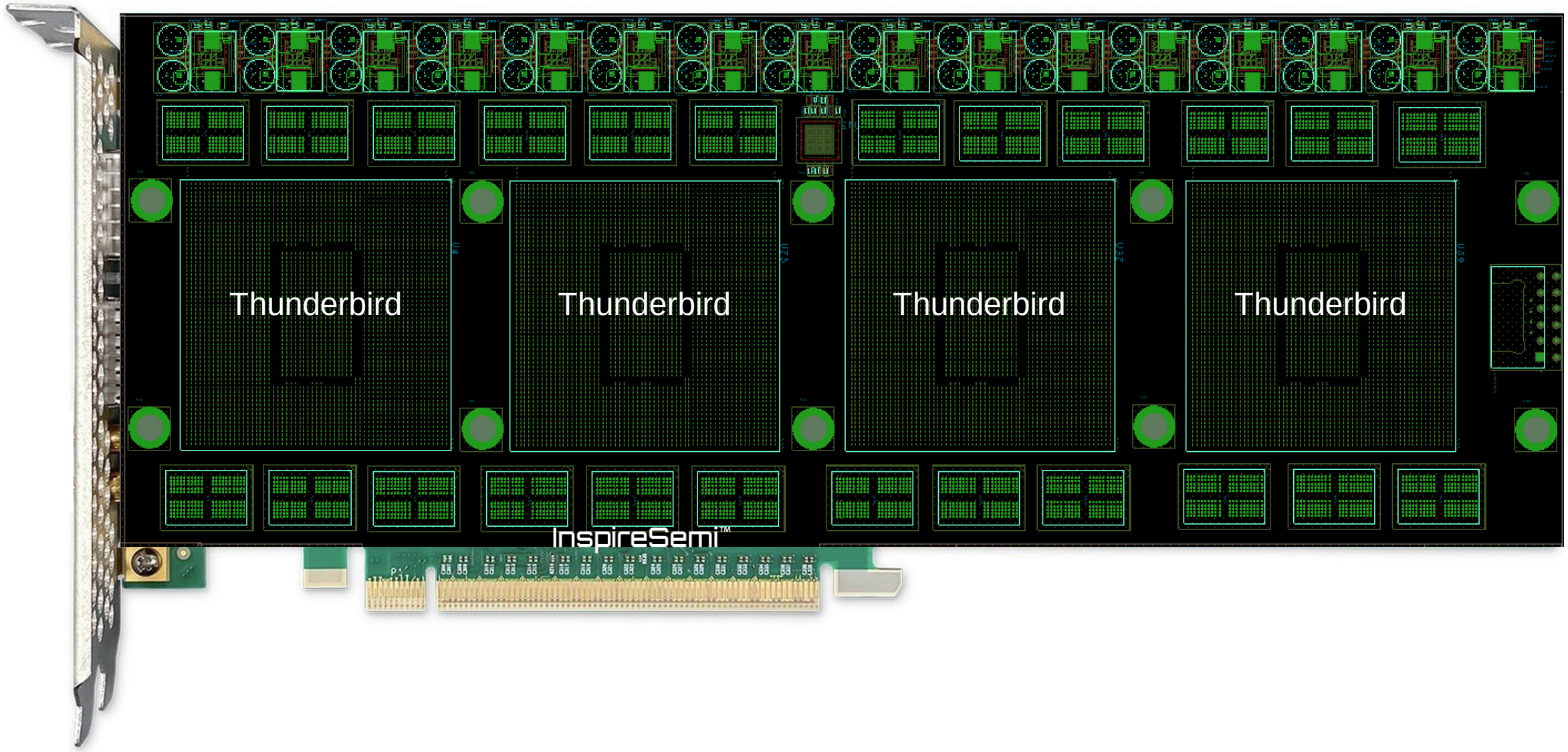
# Thunderbird is Deterministic, GPUs are Not

- **Reproducibility of results is a must for many key applications, disqualifying GPUs**
  - E.g- high-frequency trading, self-driving cars, cryptography, healthcare imaging, 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 PCIe Accelerator Card



# Thunderbird Performance Specifications

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

**50 GFLOPS/ Watt**
*(FP64)*

**20x**
*reduction in latency*

**30 – 60%**
*lower power consumptions*

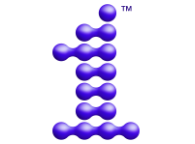
Early customer feedback is unlike GPUs, Thunderbird is likely to hit its peak performance numbers due to its architecture and interconnect technology

# Open Software Ecosystem Solves Customer Porting Challenges

- Leverages established RISC-V software ecosystem
  - Eliminates need for proprietary software stacks
- 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
- Leveraging key frameworks, compilers, & tools for RISC-V
  - Address-accurate QEMU model
  - Standard GCC, Gfortran, GDB toolchains
  - Standard HPC libraries (e.g. – BLAS, LAPACK, FFTW)
- Key Operating Systems
  - Linux
  - Kitten lightweight kernel (LWK)
  - Real-time kernels (RTOS)

















# World Class Supply Chain Partners

- TSMC - Wafer fab
  - World's largest semiconductor foundry
  - Developing the most advanced process nodes
  - Secured 12nm wafer capacity for 2024

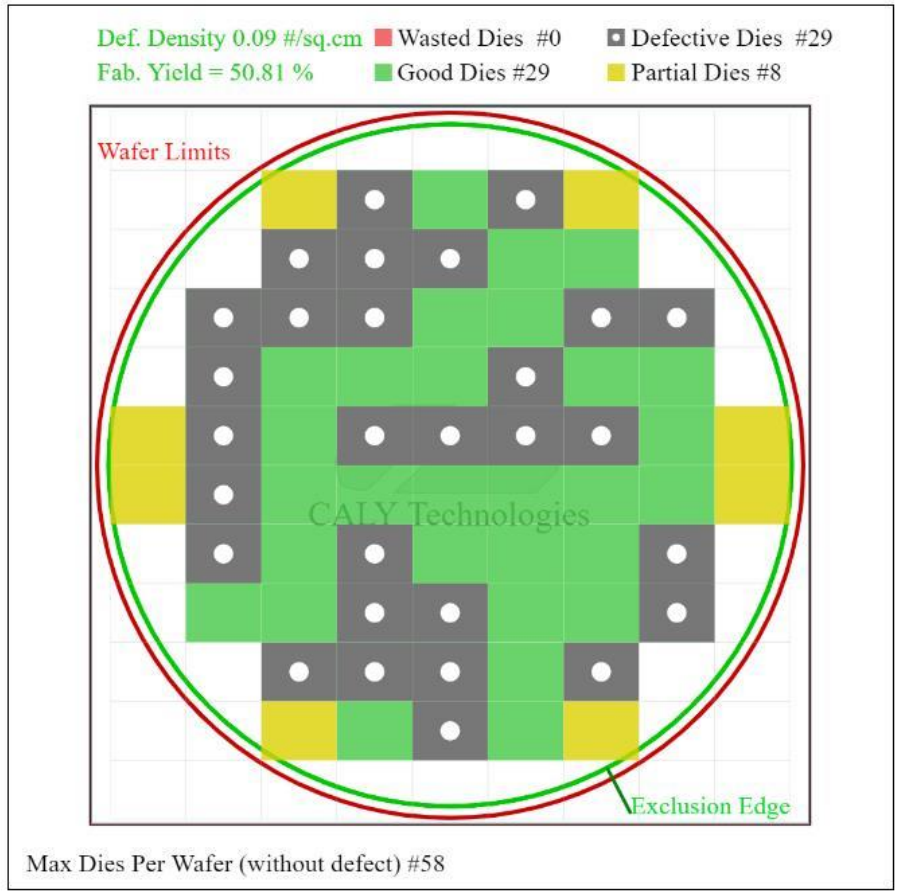
- ASE – Chip package & test
  - Worlds largest and highest quality OSAT (Outsourced Semiconductor Assembly and Test)
  - Leading edge package design

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

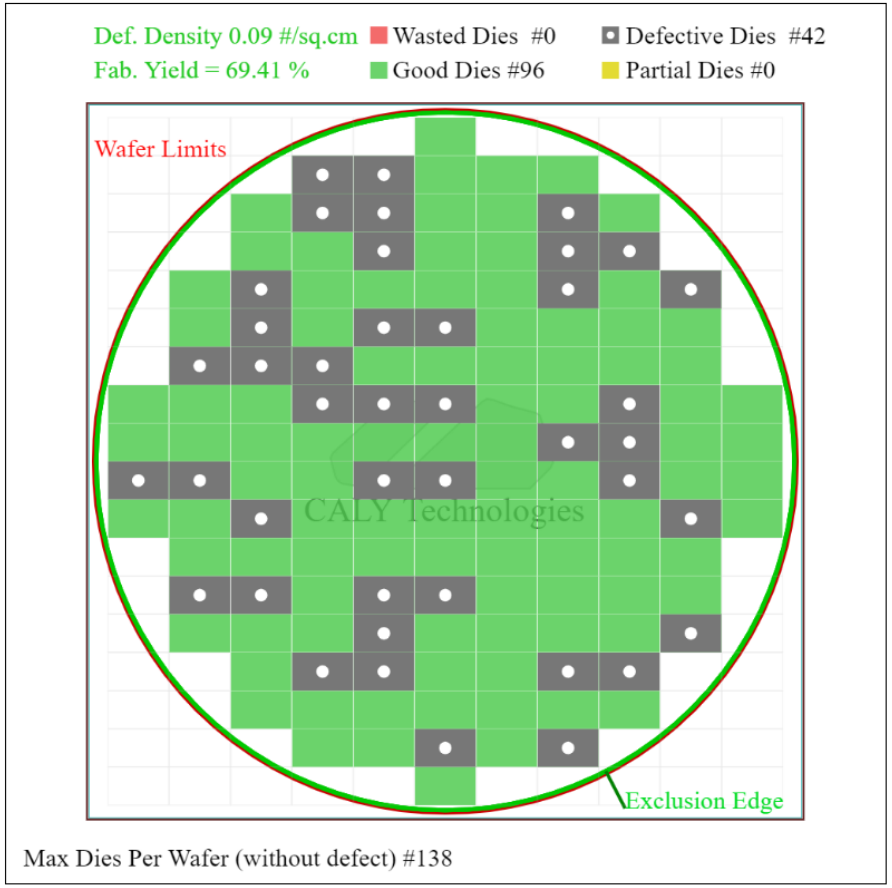
# Thunderbird Smaller Die Size Benefits

*Half competitors' size = 3X Advantage in Yield*

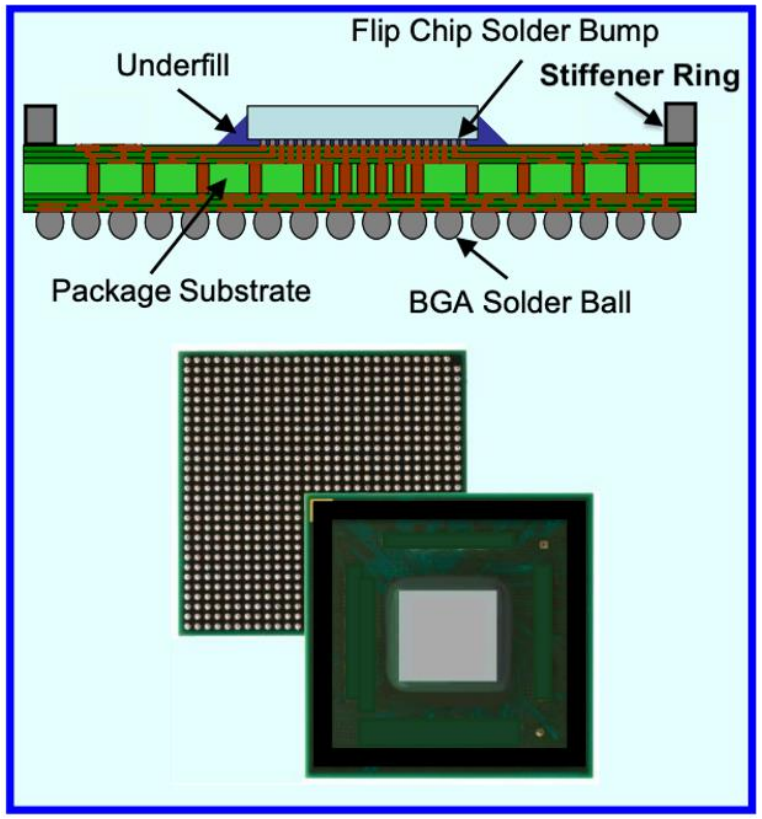
Competitive chips are enormous
- At manufacturable size limits
- Exotic $$ packaging complications
- Gruesome for yield and cost

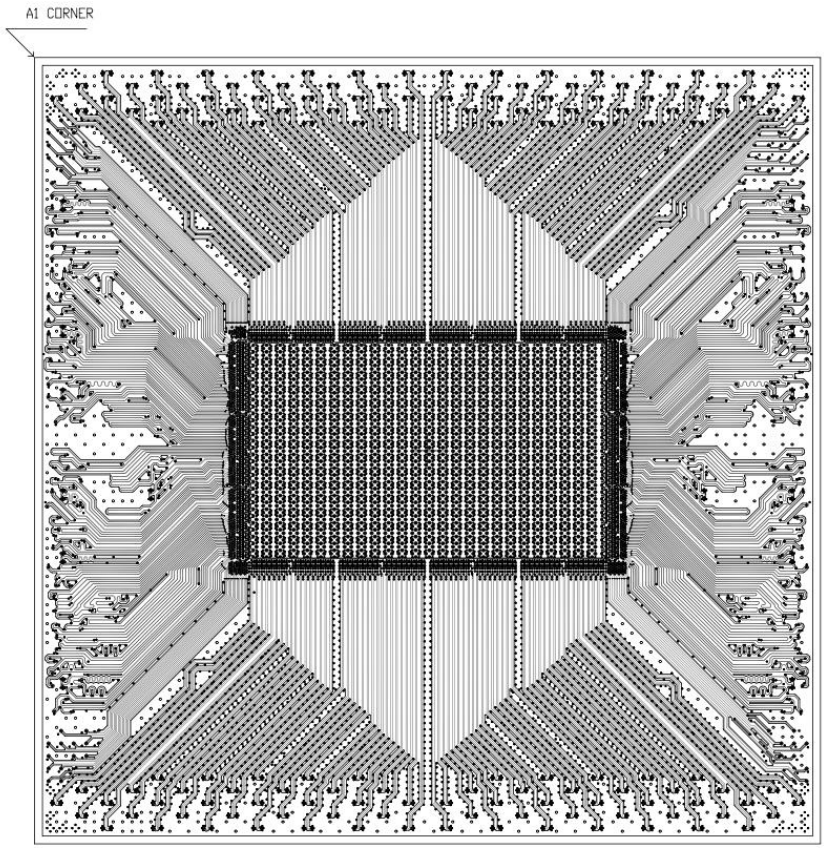


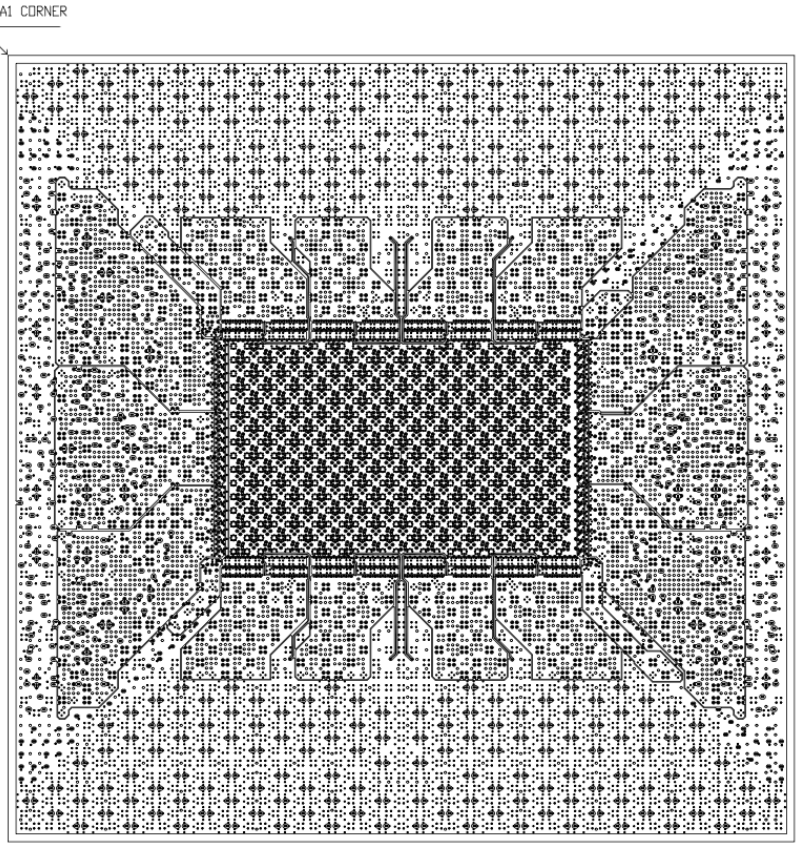

We pack comparable capability into half the area
- Higher yield, better wafer utilization
- Net 70% cost reduction
- Reduced power/latency for on-chip comms



# Thunderbird Substrate Design

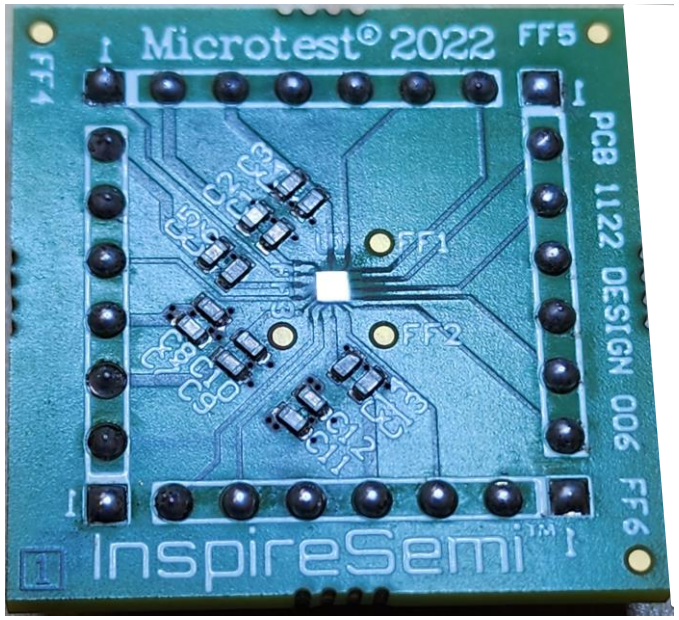


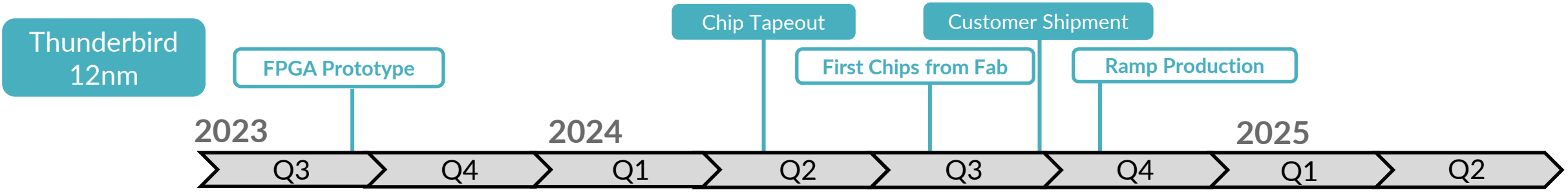



# Fundamental Capabilities Proven with TSMC 5nm Test Chip

- De-risks plan for follow-on 4nm Thunderbird 2
  - Validated team's ability to deliver designs on leading-edge TSMC process node
  - Worked first time, met performance and power targets
  - Including full-custom layout optimized at every level
  - Something not many companies can do, perhaps none this size



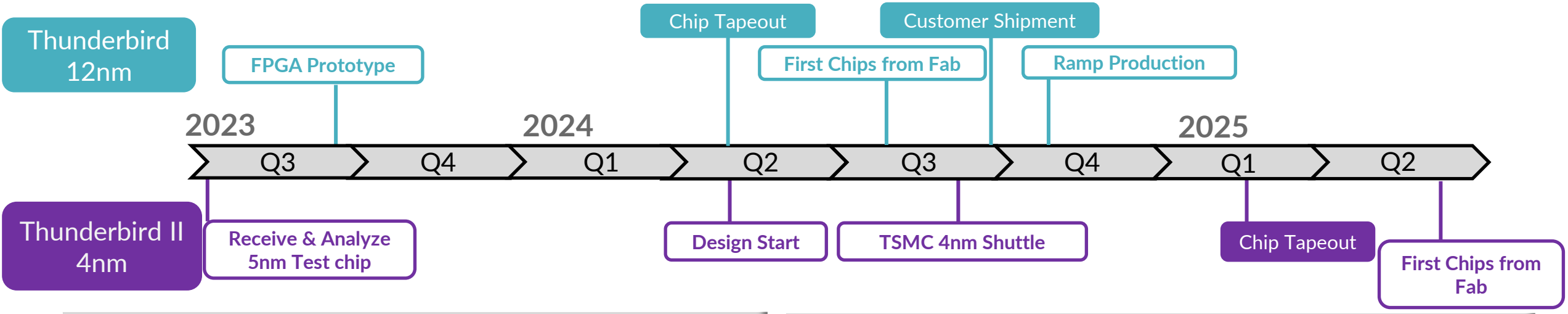
# Robust HPC & AI Roadmap




## Thunderbird

- Addresses complex HPC applications
  - e.g.- CAE/CFD, energy/reservoir modeling & sim, weather, pharma/genomics, finance, fraud detection
  - Low-cost DDR memory for large HPC jobs
- Applicable for AI-augmented HPC
- Ideal for graph analytics, including GenAI “Afterburner Program” for GenAI LLMs

# Robust HPC & AI Roadmap




## Thunderbird

- Addresses complex HPC applications
  - e.g.- CAE/CFD, energy/reservoir modeling & sim, weather, pharma/genomics, finance, fraud detection
  - Low-cost DDR memory for large HPC jobs
- Applicable for AI-augmented HPC
- Ideal for graph analytics, including GenAI "Afterburner Program" for GenAI LLMs

## Thunderbird II

- TSMC 4nm – higher performance, lower power
  - Quadruples core count, up to 10,000/chip
- Additional features for AI
  - High Bandwidth Memory (HBM), CXL
  - AI-specific instructions
- Enhanced vector instructions for HPC

# 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 |