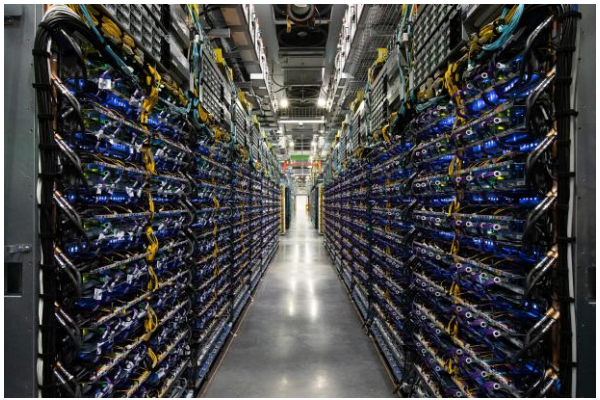
Google

Ironwood: Delivering Best in Class perf, perf/TCO and perf/Watt for Reasoning Model Training and Serving

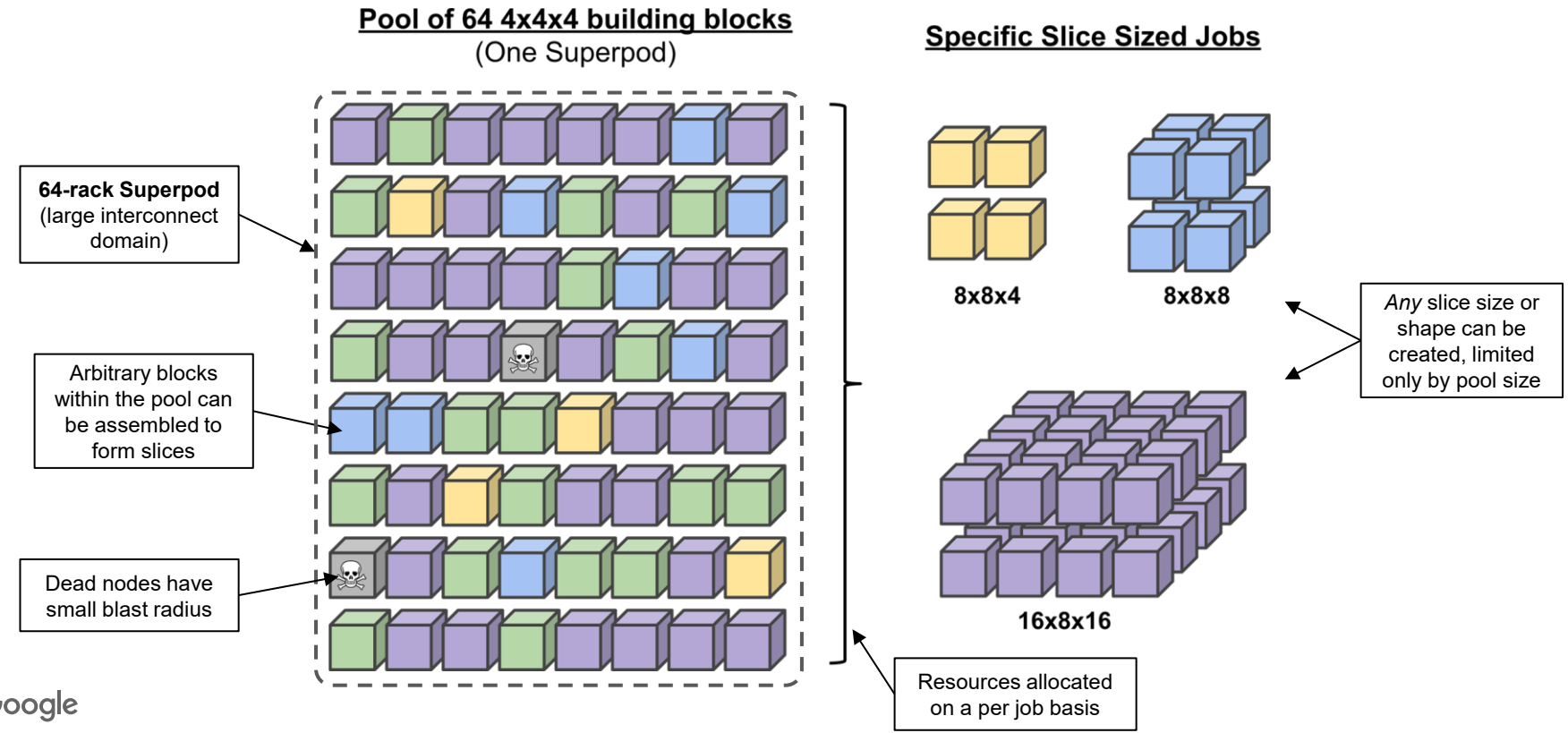
Norman P. Jouppi and Sridhar Lakshmanamurthy
With contributions from many others
Hot Chips August 26, 2025



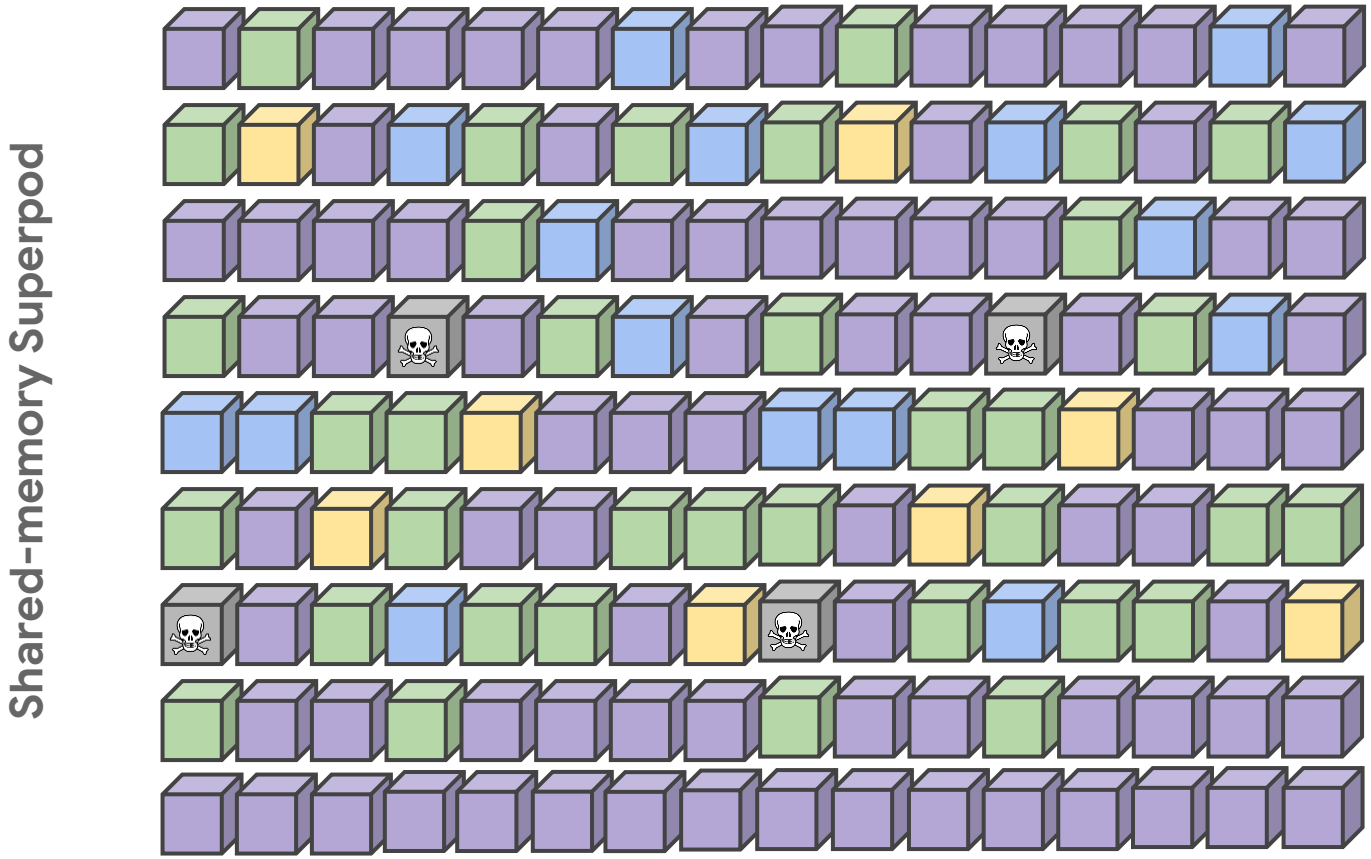
Breakthrough Innovations in Ironwood TPU Systems

- 9216 Ironwood chips use optical circuit switches (OCS) to share memory
- Directly addressable shared HBM memory capacity of 1.77 PB
- 42.5 Exaflops of ML Compute using FP8 precision
- Emphasis on RAS (reliability, availability, and serviceability)
- Industry-leading compute power efficiency, 2x perf/W over previous generation
- 3rd generation of liquid cooling infrastructure
- 4th-generation Sparsecore for embeddings and collectives offload
- Deployment at hyperscale underway

TPUv4: 4096 Chips Share Memory Using OCS

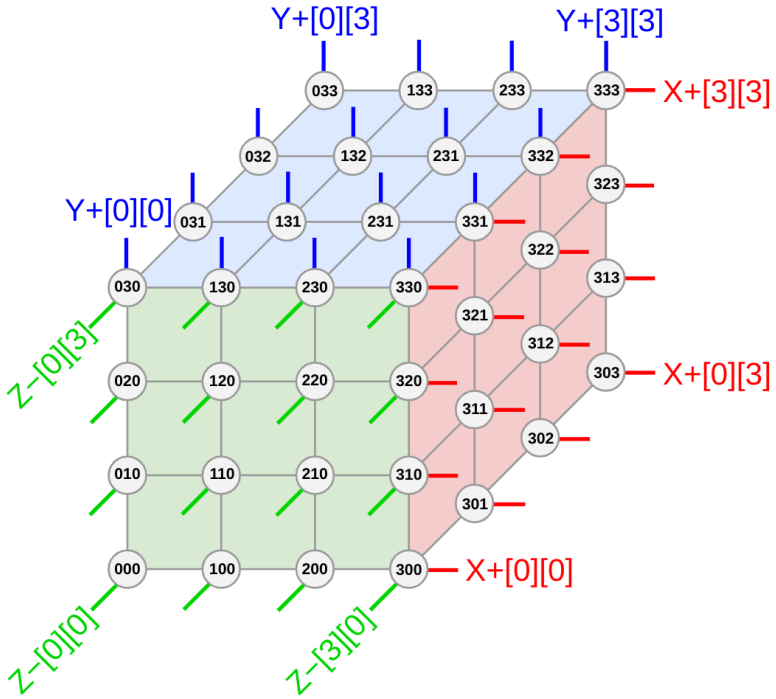


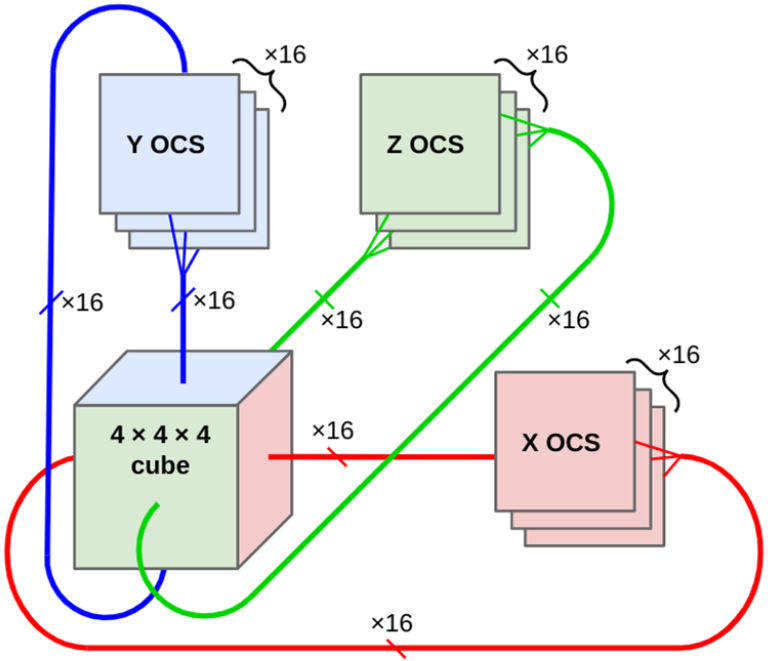
9216 Ironwood Chips Sharing Memory Using OCS



Racks Are Connected With Optical Circuit Switches (OCS)

- Different ranks of OCS connect different dimensions and indices

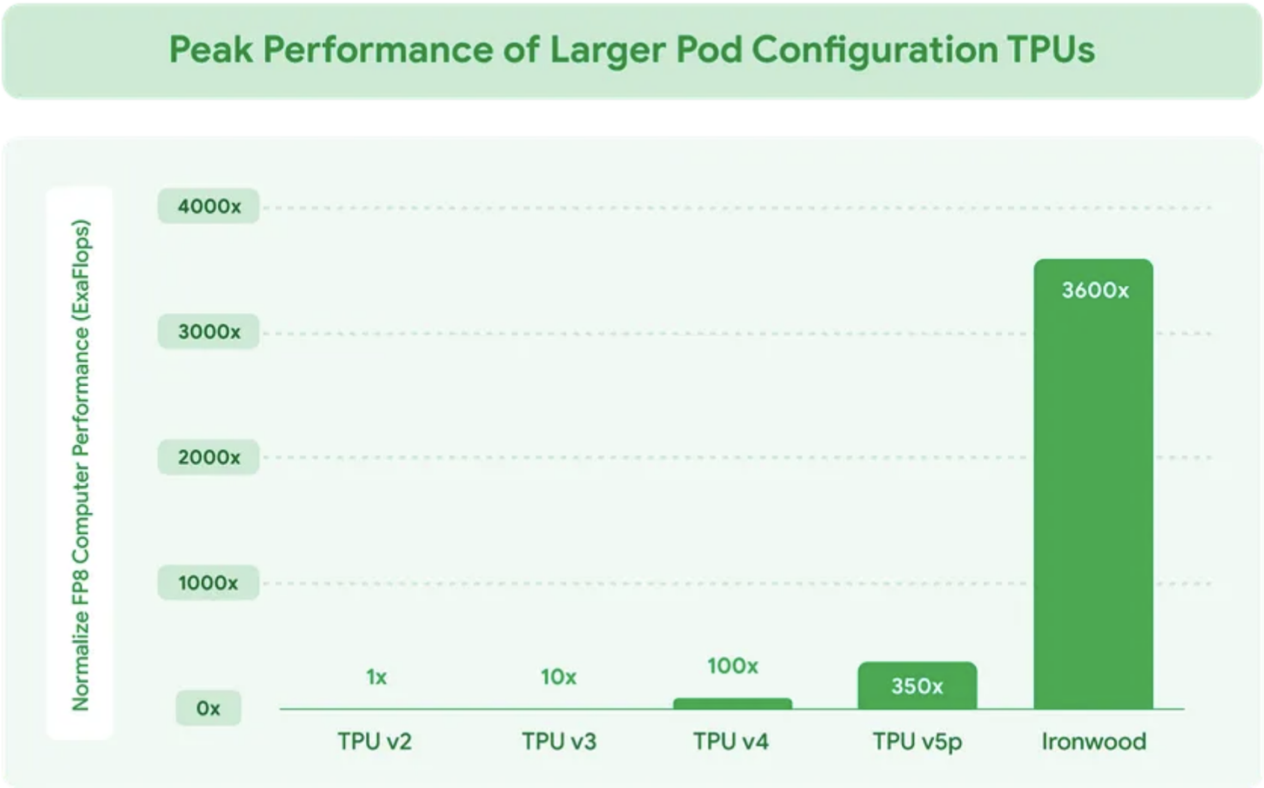




Directly-addressable Shared HBM Memory of 1.77 PB

- Sets a new record for shared-memory multiprocessors
- Enables low-overhead high-bandwidth sharing of data
  - Accelerates fine-grained parallel processing
  - Enables higher degrees of parallelism with high perf/TCO
- Able to support huge models efficiently

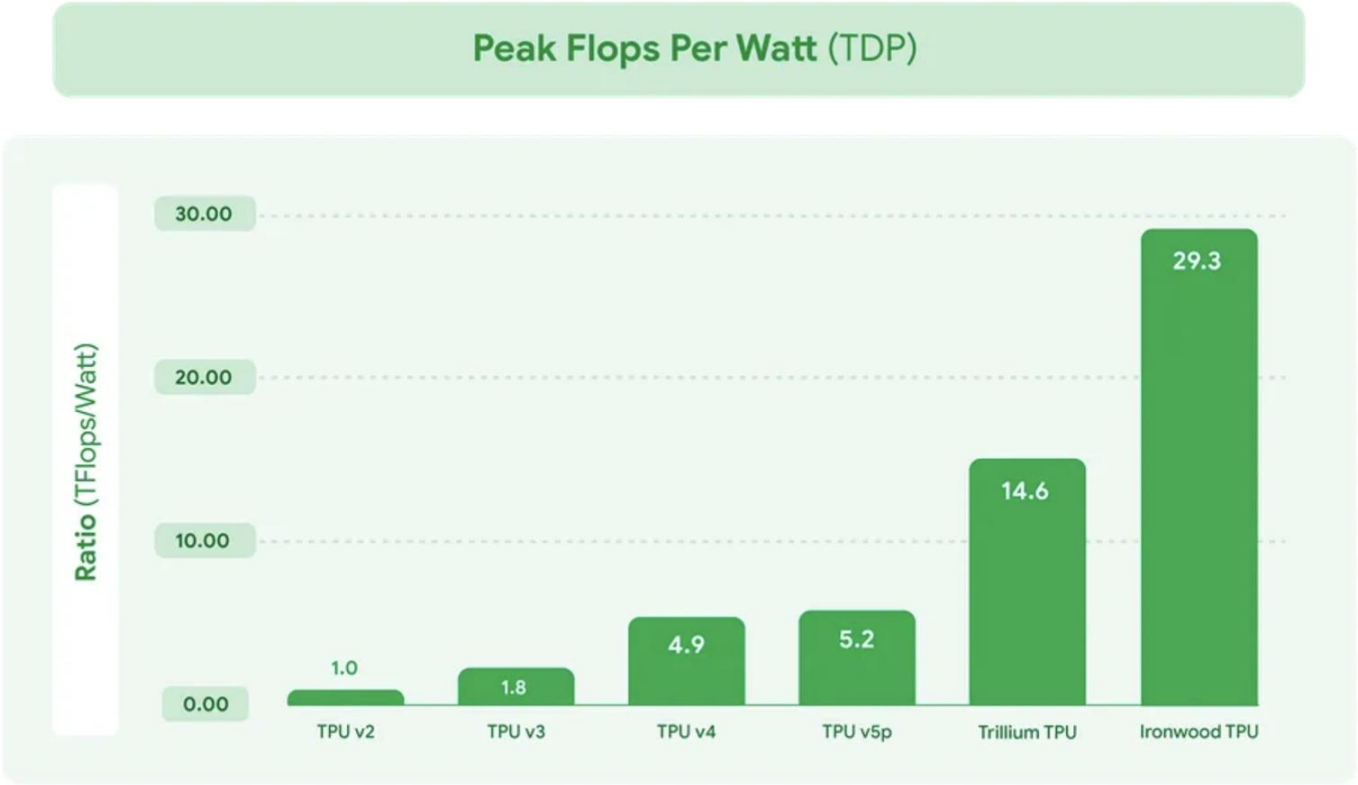
42.5 Exaflops of ML Compute using FP8 Precision



Emphasis on RAS (Reliability, Availability, and Serviceability)

- Enables productive scaling to extreme sizes

Industry-leading Compute Power Efficiency

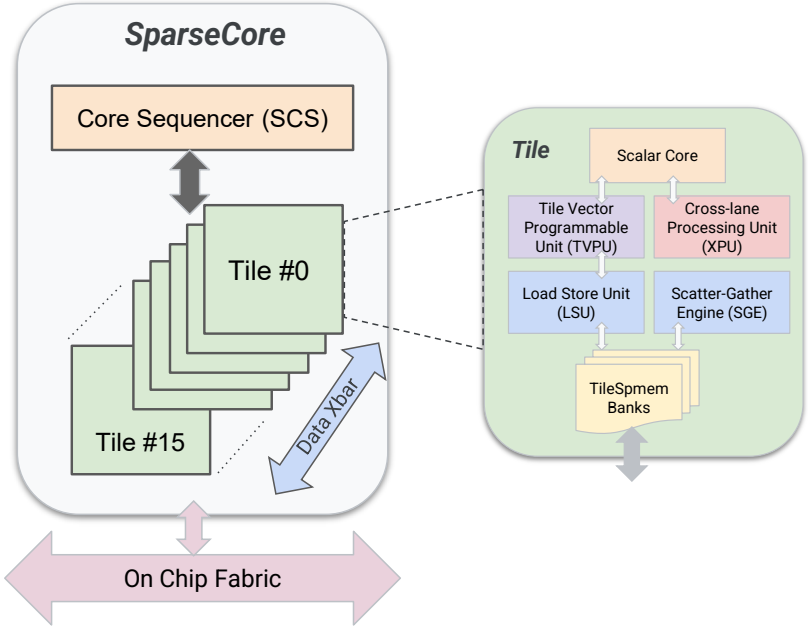


3rd Generation of Liquid Cooling Infrastructure

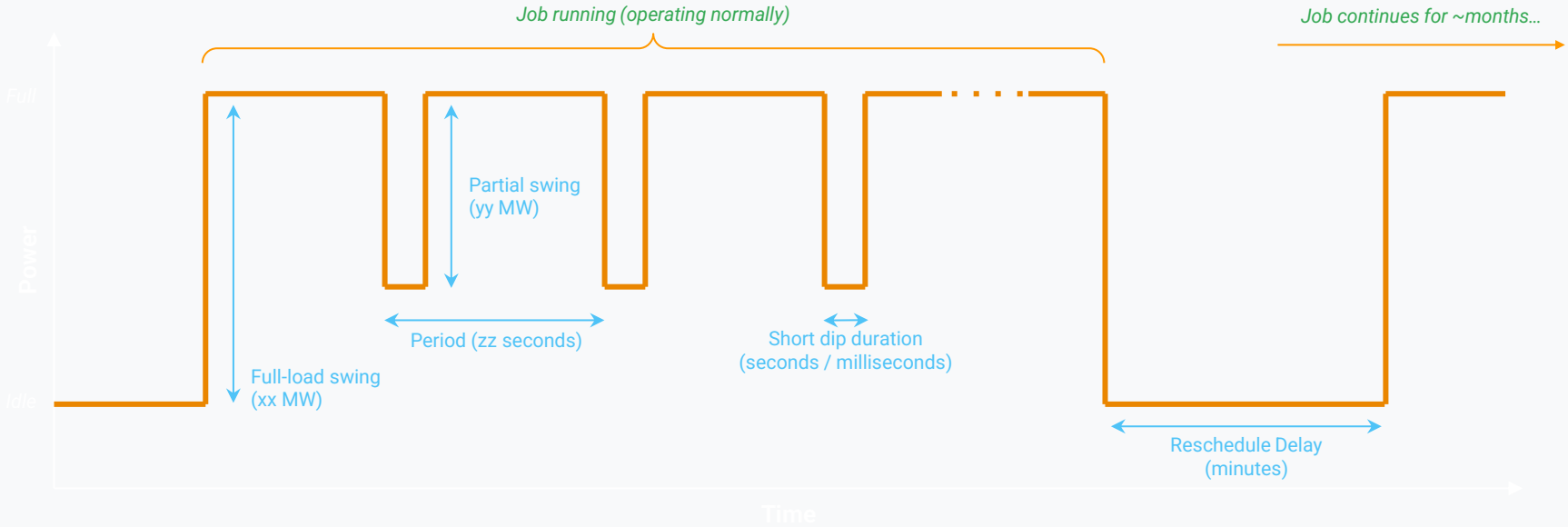
- For details please see Jorge Padilla's 2025 Hot Chips tutorial slides

4th Gen SparseCore: Accelerate Embeddings & Collectives

- 2.4x FLOPS vs. 3rd Gen SparseCore
- Offload collectives for pretraining, RL-based fine tuning
  - Execute in parallel with main TensorCore computation
- Embedding processing for large recommendation models
- SparseCores leverage non-coherent shared memory across a pod
- Massive memory parallelism (millions of outstanding references between nodes in the pod) is exploited with multithreading



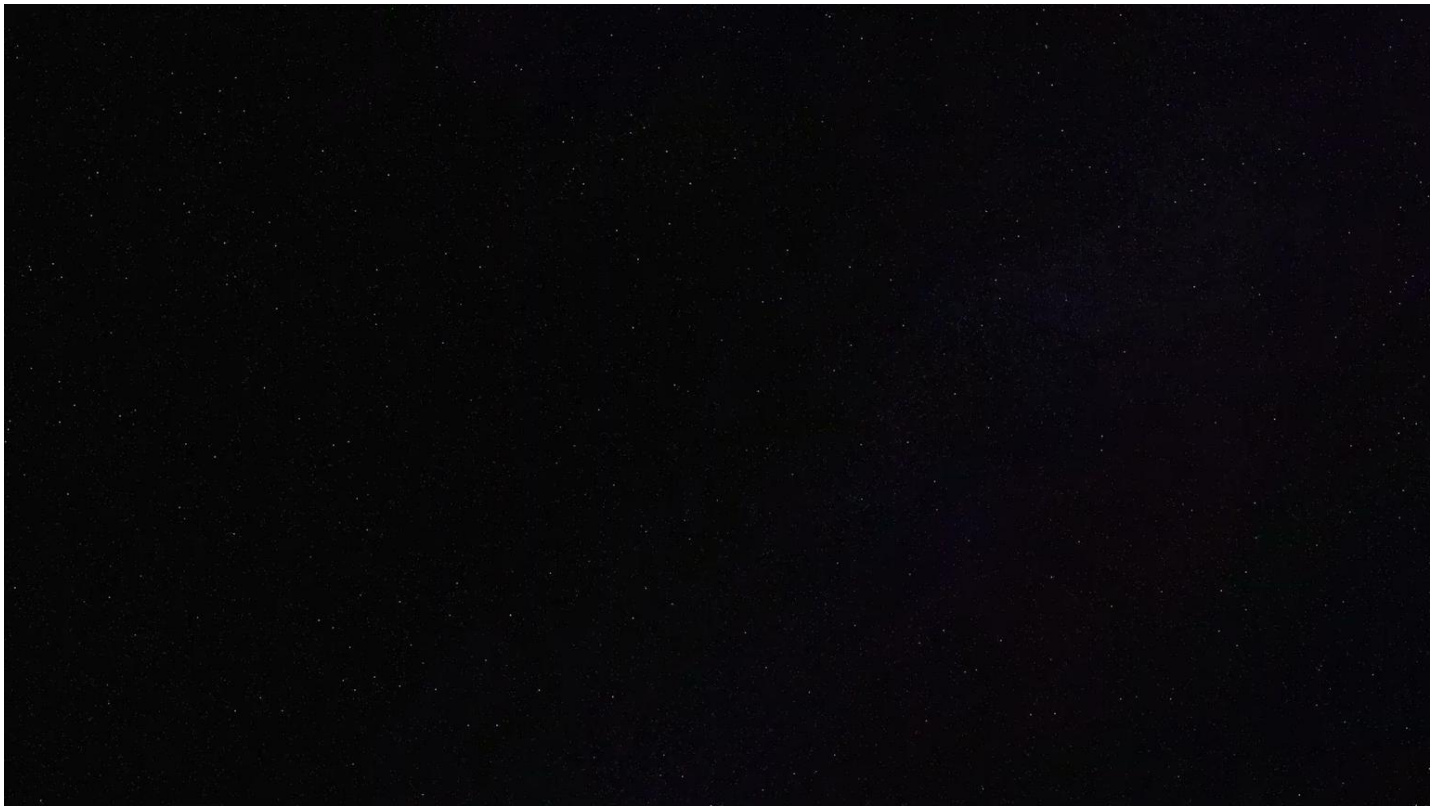
Large Scale Pre-training & MW Scale Load Swing



- Unprecedented load swings (MW scale in sec/ms) real with large-scale pretraining
- Ironwood supports both HW and SW features to smooth out these swings
- Deploys a full-stack approach to proactive power shaping - **Google Project Smoothie**

Google

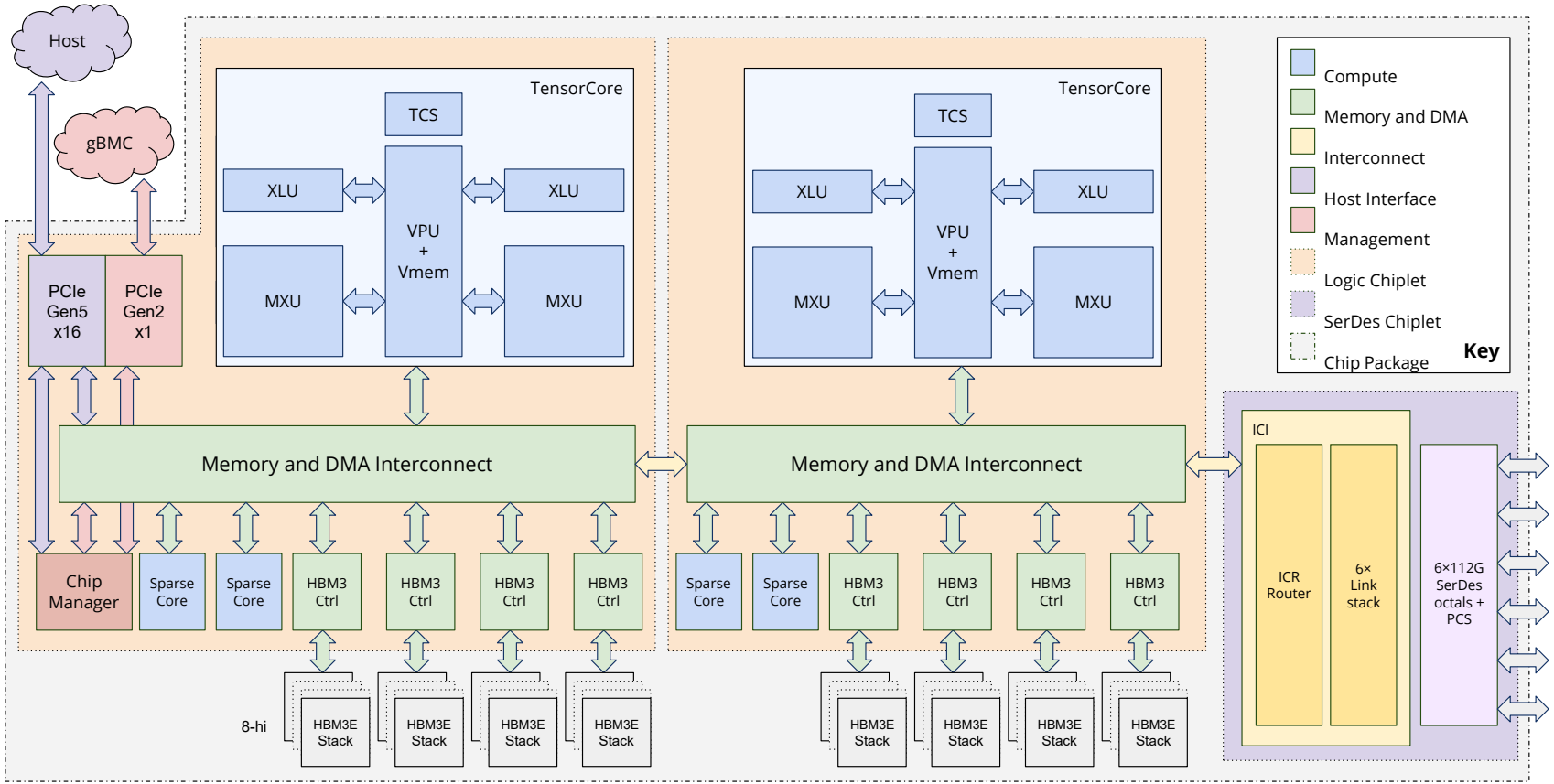
Deployment at Hyperscale Underway



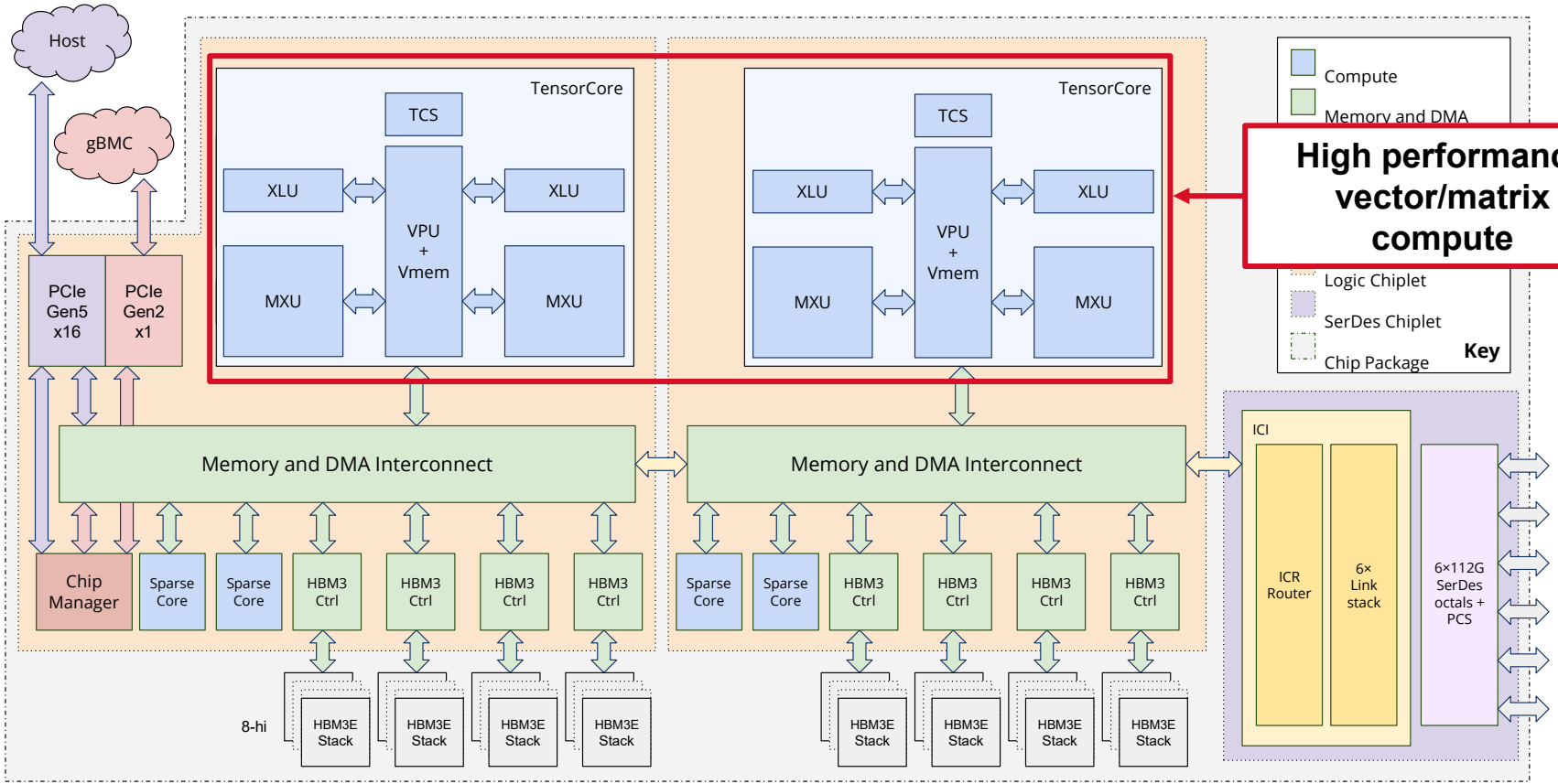
Ironwood Chip

- First dual compute die TPU, 4614 TFLOPS of FP8
  - >10x compute compared to TPU v5p
  - Capable of large scale pretraining of foundation models
- 8 stacks of HBM3E, peak BW 7.3 TB/s, capacity 192 GiB
  - Optimized for serving latest generation thinking models
- 1.2 TBps I/O to gluelessly scale-up to 9216 chips
- Industry leading cold plate thermal solution
- Integrated root-of-trust (iROT) for secure computing
- Functional BIST & silent data corruption (SDC) mitigation
- Logic repair to improve yield
- Dynamic voltage/frequency scaling for efficient perf/W
- AI based ALU circuits, floor plan optimization

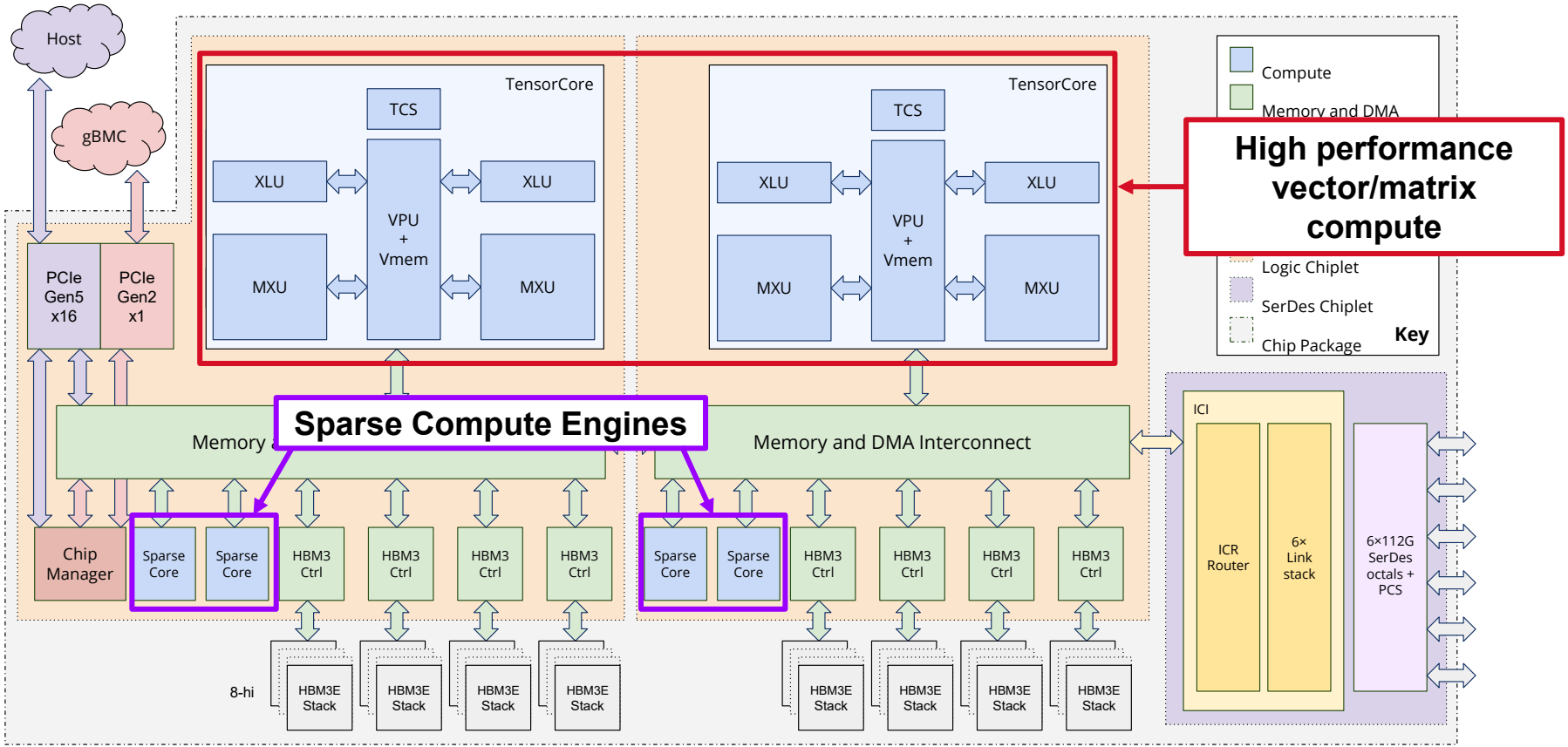
Ironwood Architecture



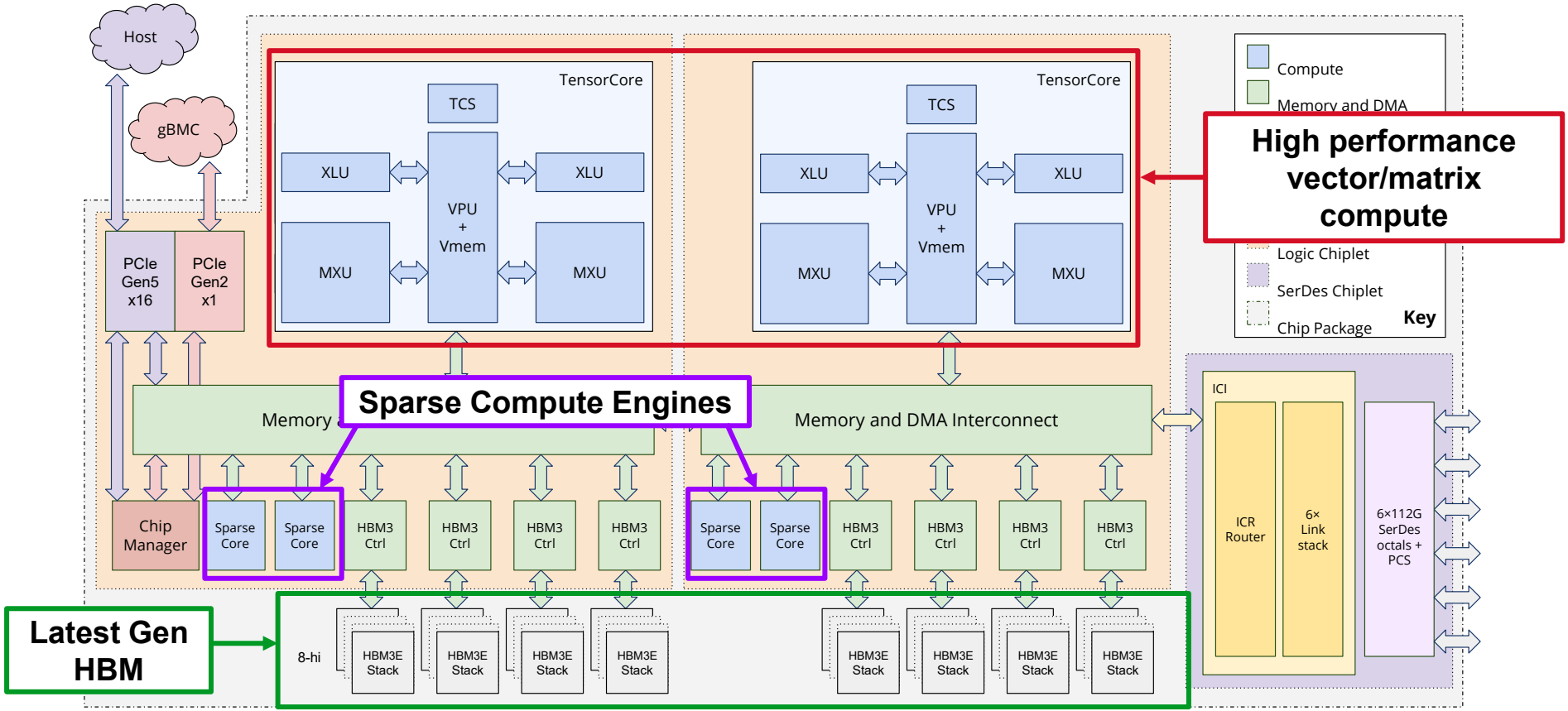
Ironwood Architecture



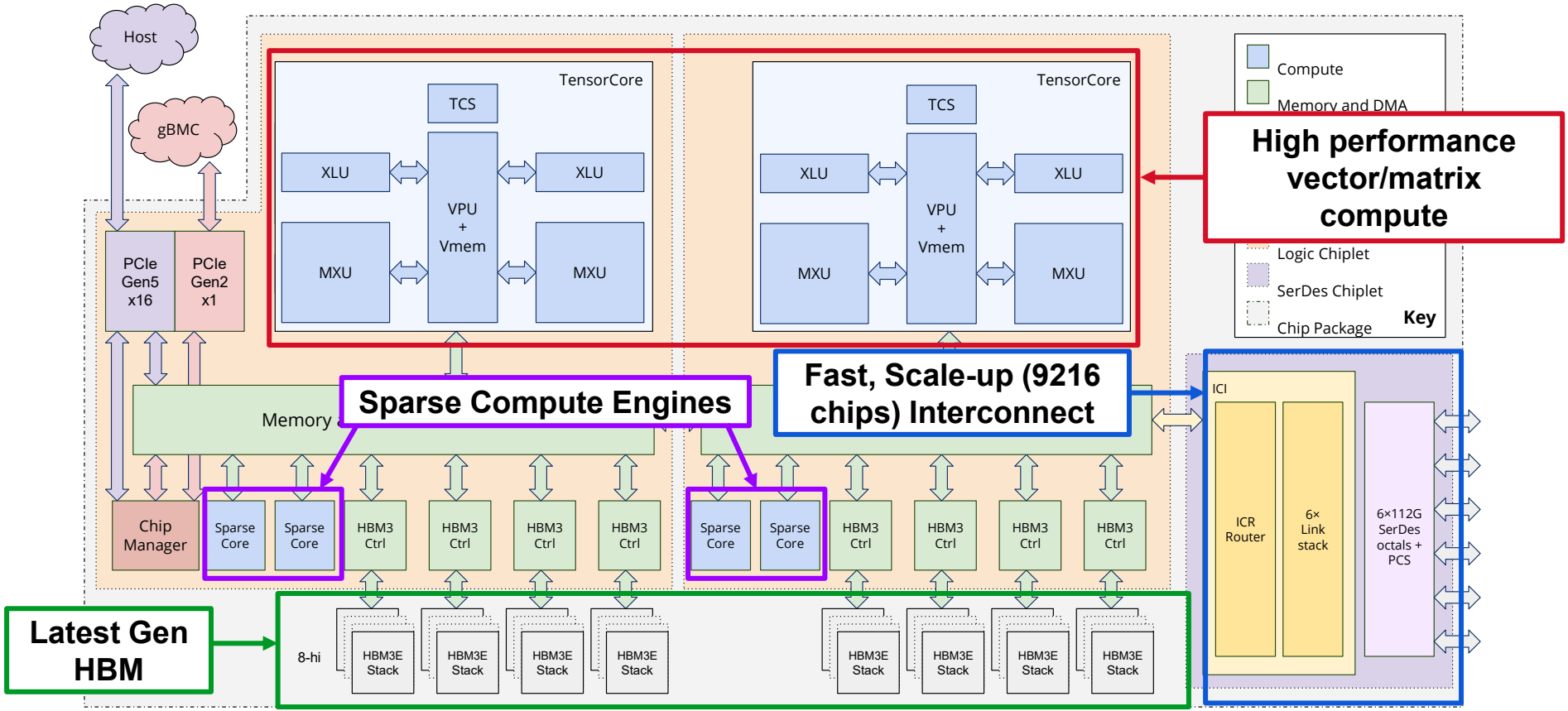
Ironwood Architecture



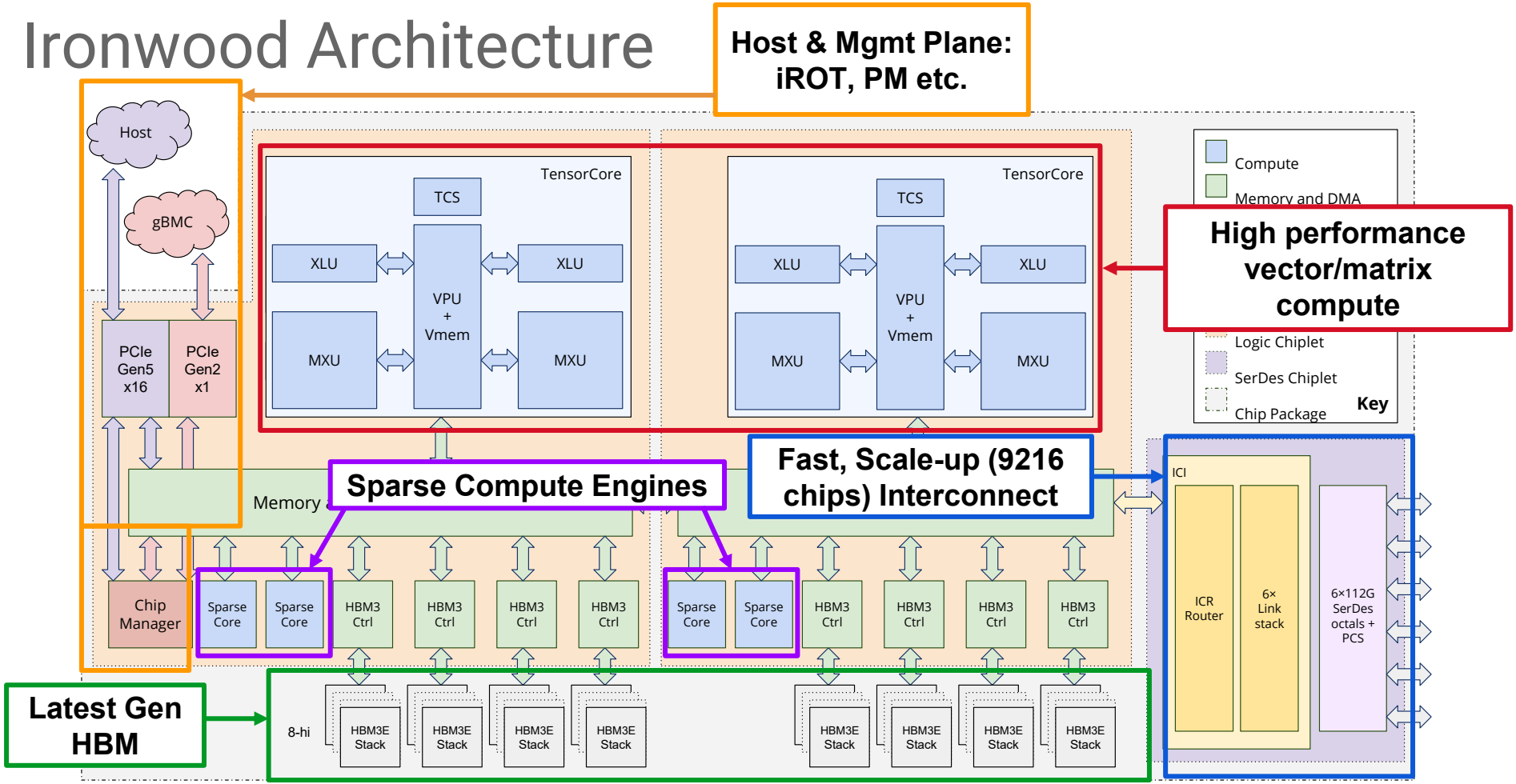
Ironwood Architecture



Ironwood Architecture



Ironwood Architecture

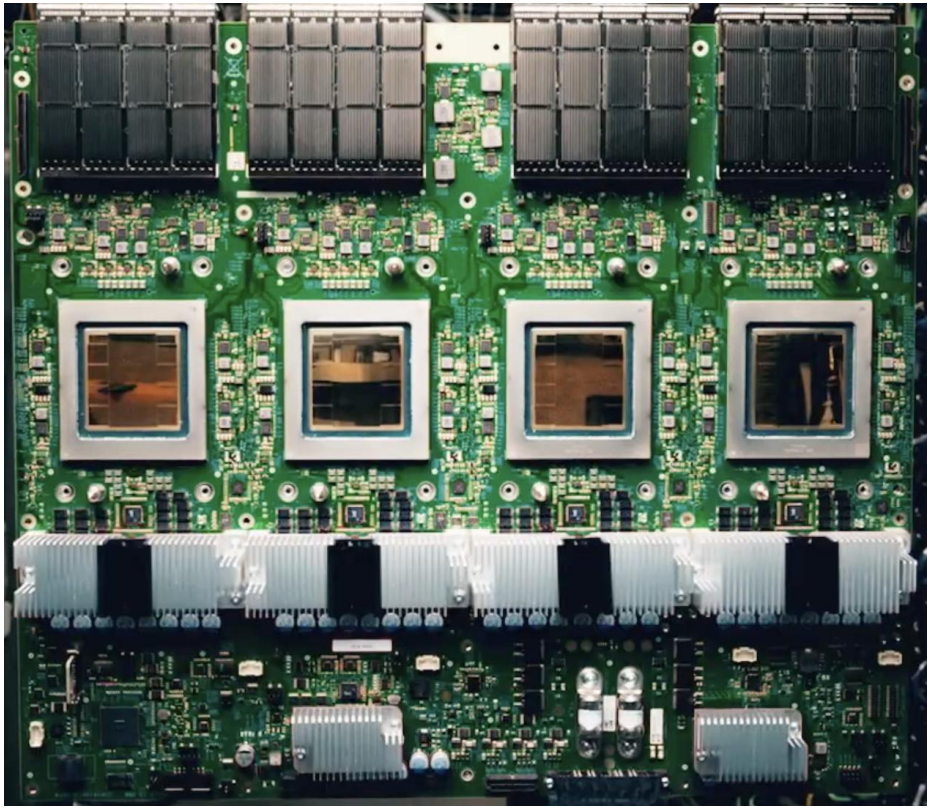


Confidential Computing Support

- Key requirement for Internal & Cloud Customers
  - Training and serving LLMs
- Ironwood supports integrated root-of-trust (iROT) controller
  - Hardware support for secure boot
  - Hardware support for secure test and debug
- Ironwood also adds hardware support for PCIe Data Object Exchange (DOE) and Component Measurement and Authentication (CMA)

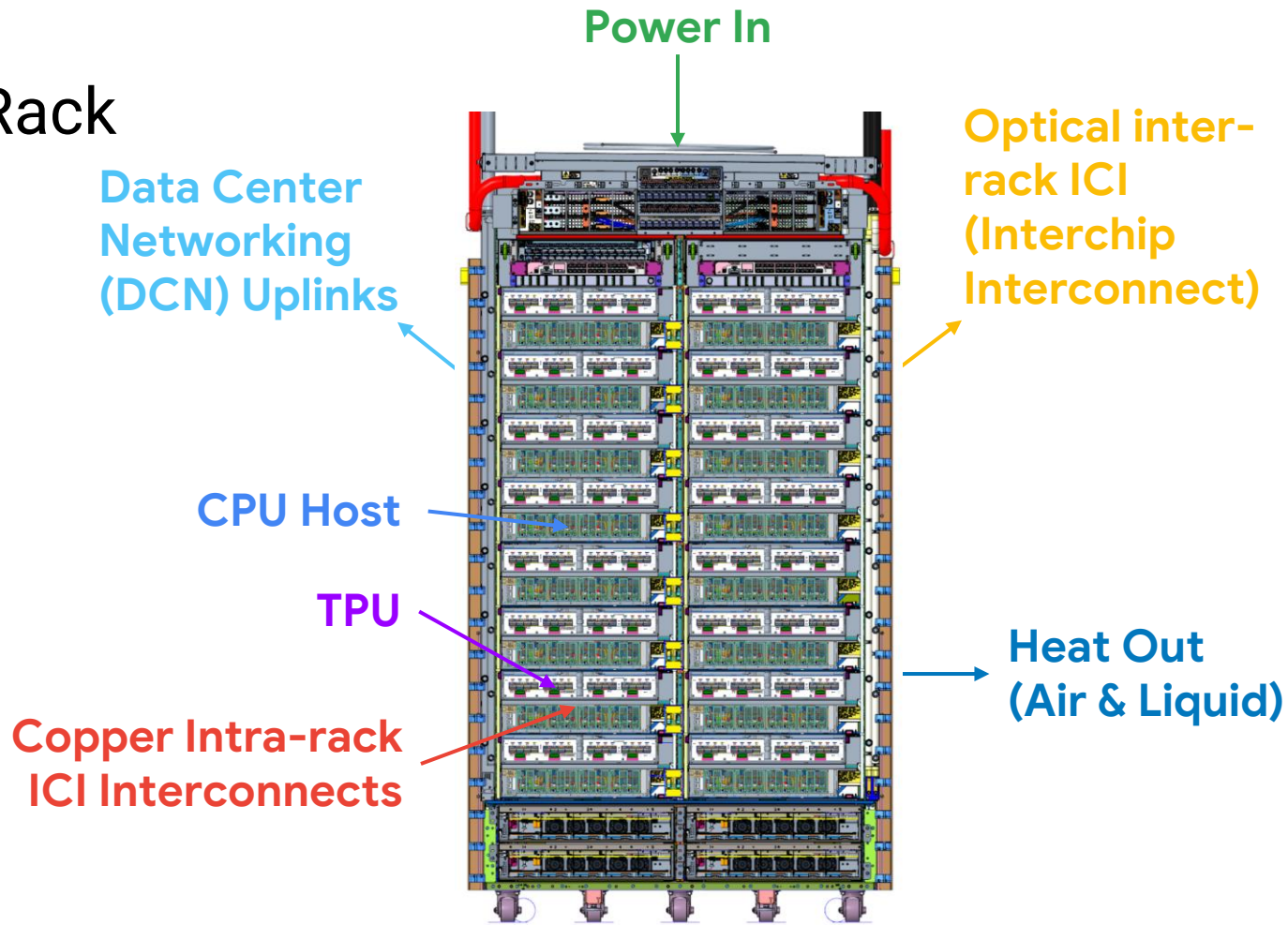
Ironwood Tray

- 4 Ironwood TPUs per tray
- Liquid cooled
  - 4 chips and VRs with parallel water flow
  - Flow rate controlled by valve
  - Similar to variable fan speeds for air
  - Yields more efficient water cooling
- PCIe Gen5x16 per TPU for host I/O
  - On back side of board
- 18 OSFP[1] connectors for off-board interconnect
  - 16 on top side shown in photo



[1]OSFP = “Octal Small Form Factor Pluggable”

The Rack


Power In
Data Center Networking (DCN) Uplinks
Optical inter-rack ICI (Interchip Interconnect)
CPU Host
TPU
Heat Out (Air & Liquid)
Copper Intra-rack ICI Interconnects

Optimal ML Performance: From Chip to Data Center

- Maximize ML throughput under dynamically varying power budgets
- Achieve optimal energy efficiency and Perf/W

Chip/System/Rack Power Management Capabilities + **Datacenter Power Awareness & Controls** = **Optimal Perf/Power within datacenter power capacity**

The “Knobs”

The “Brain”

Optimal ML Performance: From Chip to Data Center

- Maximize ML throughput under dynamically varying power budgets
- Achieve optimal energy efficiency and Perf/W

Chip/System/Rack Power Management Capabilities	+	**Datacenter Power Awareness & Controls**	=	**Optimal Perf/Power within datacenter power capacity**
The “Knobs”		The “Brain”		

Targeting 30% additional throughput per Datacenter in the same power budget

Ironwood Continues Our Lead in Scale-up and Scale-out

- 9216 Ironwood chips share memory per superpod with optical circuit switches (OCS)
  - Huge boost for scale-up
- Directly addressable shared HBM memory capacity of 1.77 PB
  - Enables super low-latency and high BW sharing of data
- 42.5 Exaflops of ML Compute per superpod using FP8 precision
  - Zettaflops across multiple superpods with scale-out
- Emphasis on RAS (reliability, availability, and serviceability)
  - Enables productive scaling to extreme sizes
- Industry-leading compute power efficiency, 2x perf/W over previous generation
  - Who wants nuclear reactors?
- 3rd generation of liquid cooling infrastructure
  - 8+ years of production experience, over 1GW in production
- 4th-generation Sparsecore for embeddings and collectives offload
  - Accelerate recommendation models and overlap collectives with computation
- Available soon in Google Cloud