ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

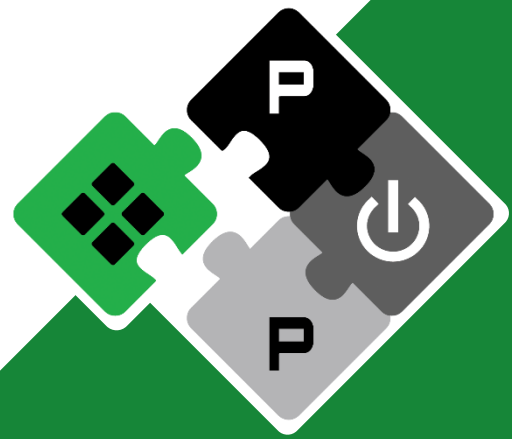
Energy efficiency and Power management for AI accelerators

Tackling the MW rack challenge

Luca Benini

lbenini@iis.ee.ethz.ch

luca.benini@unibo.it



@pulp_platform

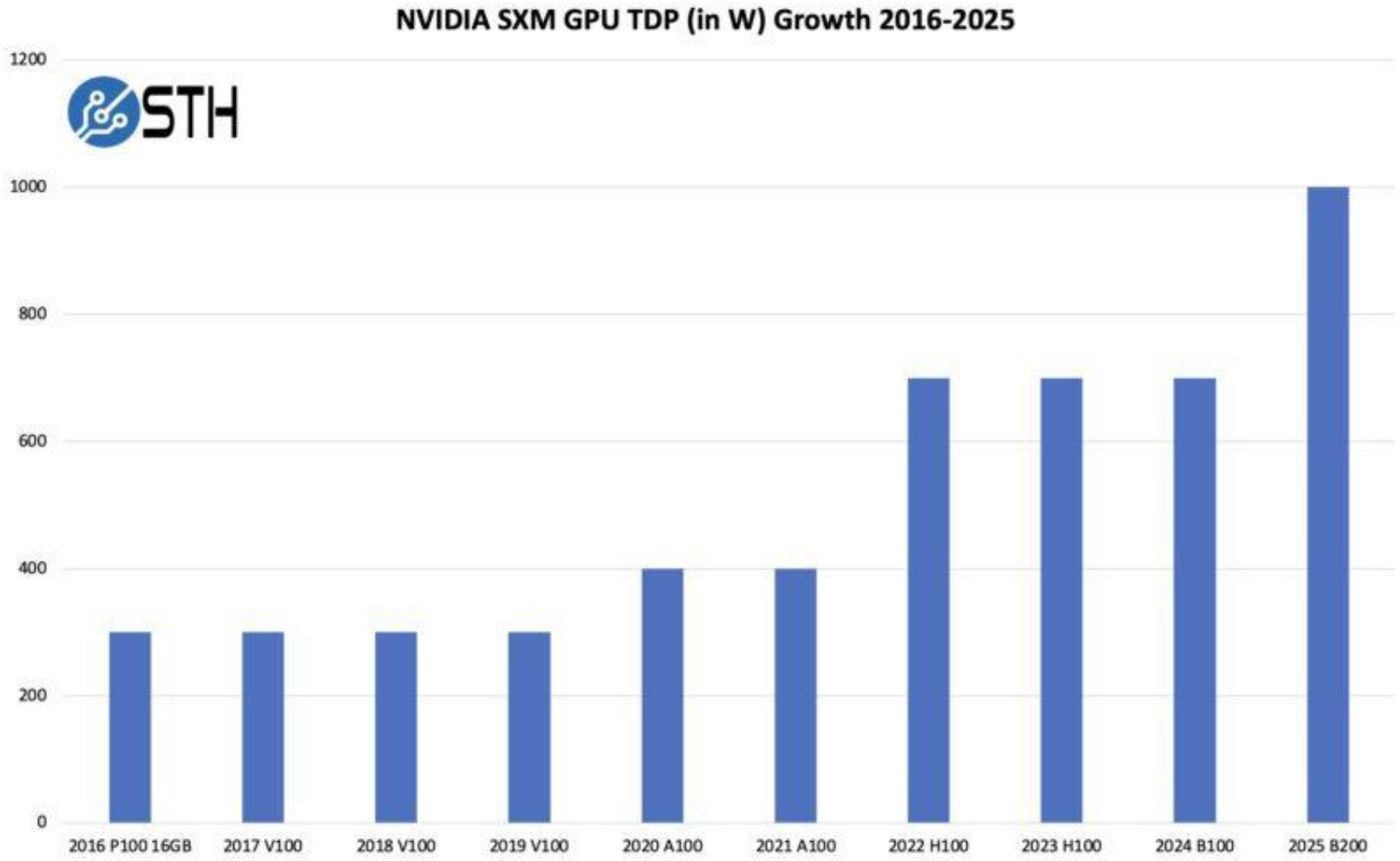
pulp-platform.org

youtube.com/pulp_platform

PULP Platform

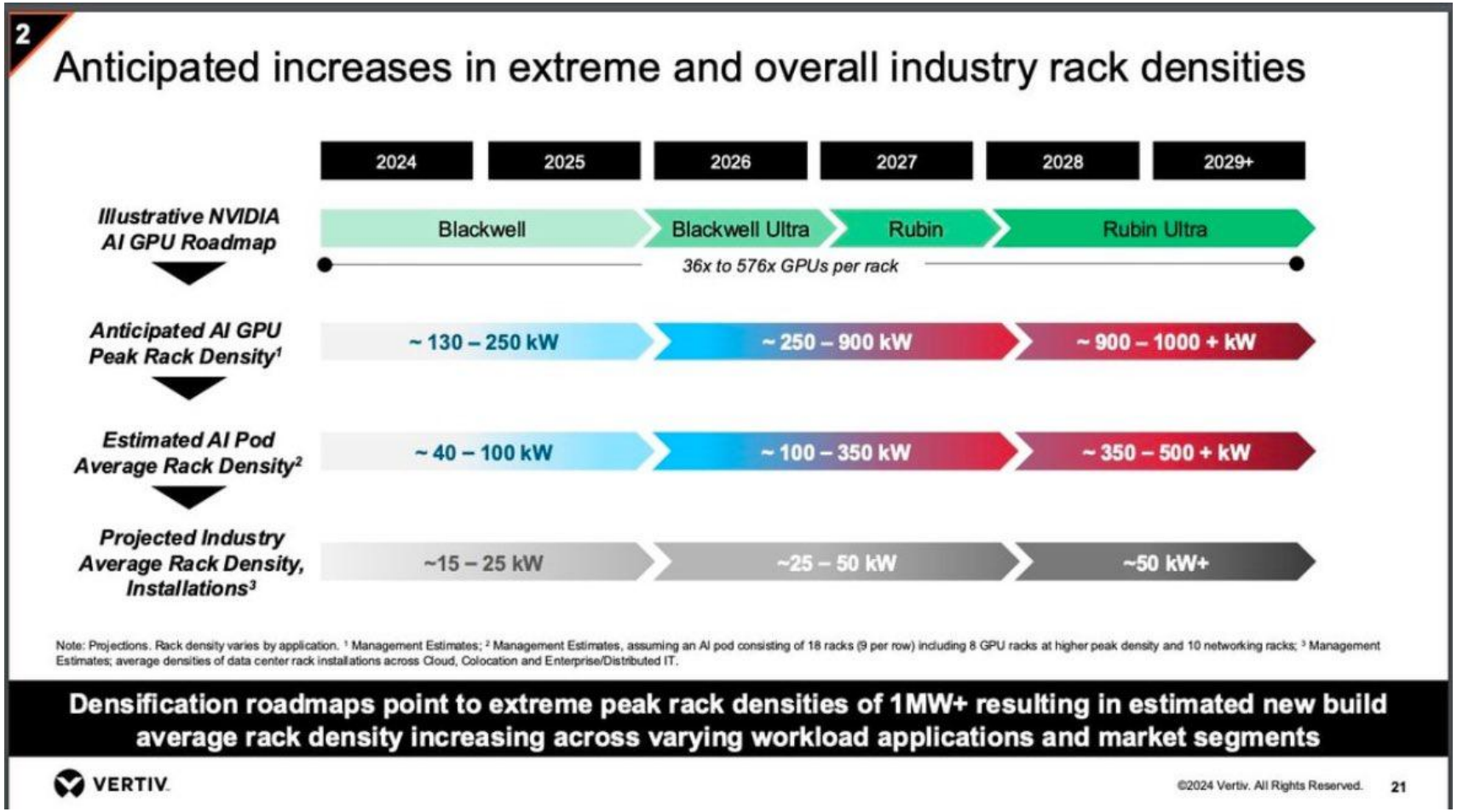
Open Source Hardware, the way it should be!

TDP of server AI Chips is growing Fast



https://www.servethehome.com/why-servers-are-using-so-much-power-tdp-growth-over-time-supermicro-vertiv-intel-amd-nvidia

The MW AI Rack (in 2029)



- GPU Density: larger #MAC/Area at high avg. utilization → more power, heat than CPUs
- Continuous Workloads: AI training and inference → sustained high-power
- Chip Proximity: To achieve maximum performance and low latency → components are packed closer together

https://wccftech.com/ai-servers-can-reach-rack-density-of-1000-kw-likely-with-nvidias-next-gen-rubin-ultra-architecture/?utm_source=chatgpt.com

Outline

- **Boosting efficiency for AI workloads**
- **Managing idleness and temperatue in heterogeneity in accelerated systems**
- **Using AI for managing AI**
- **Conclusions, future perspectives**

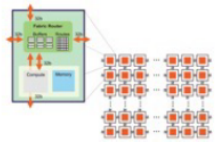
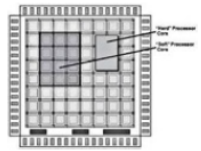
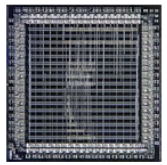
Booster 1: Domain-Specific Architecture









	CPU Core	CPU Core Vector Engine	Parallel Thread Accelerator (GPUs)	Tensor Array Accelerator (TPU, Groq, TensorCore)	Microcore Mesh Accelerator (Cerebras, Tenstorrent)	Computational Block Accelerator (FPGA)	Custom Dataflow Accelerator (ASIC)
Technology label	CPU	Vector	GPU	Tensor	Manycore	FPGA	ASIC
ALUs per core	1-4	8-32	32-64	8x8 to 256x256	1 to 4	Various	App. specific
Cores per processor	4-64	4-64	8-128	1 to 4	100s to 1M	Various	App. specific
Parallel performance	Low	Medium Low	Medium High	High	High	High	Very High
Comp. efficiency (Ops/W)	Low	Medium Low	Medium	High	High	High	Very High
Comp. flexibility	Very High	Medium Low	Medium	Medium Low	Medium	Low	Very Low
Computation scheduling	Dynamic by instruction	Dynamic by instruction	Dynamic by kernel	Static by kernel	Static by kernel	Fixed	Fixed
Code redesign	Seconds	Seconds	Seconds/Minutes	Minutes	Minutes/Hours	Hours	Months

Increasing application specificity -> greater parallel performance -> narrower application/computational kernel enablement

Reuther25 https://arxiv.org/abs/2510.20931v1

Flexible Domain-Specific Architecture (DSA)

Multiple dimensions for Specialization

Extensions to processor cores

ISA extension

Same L0 (RF)

Tensor Engines

ISA extension or fast offload

Decoupled L0

Decoupled Accelerators

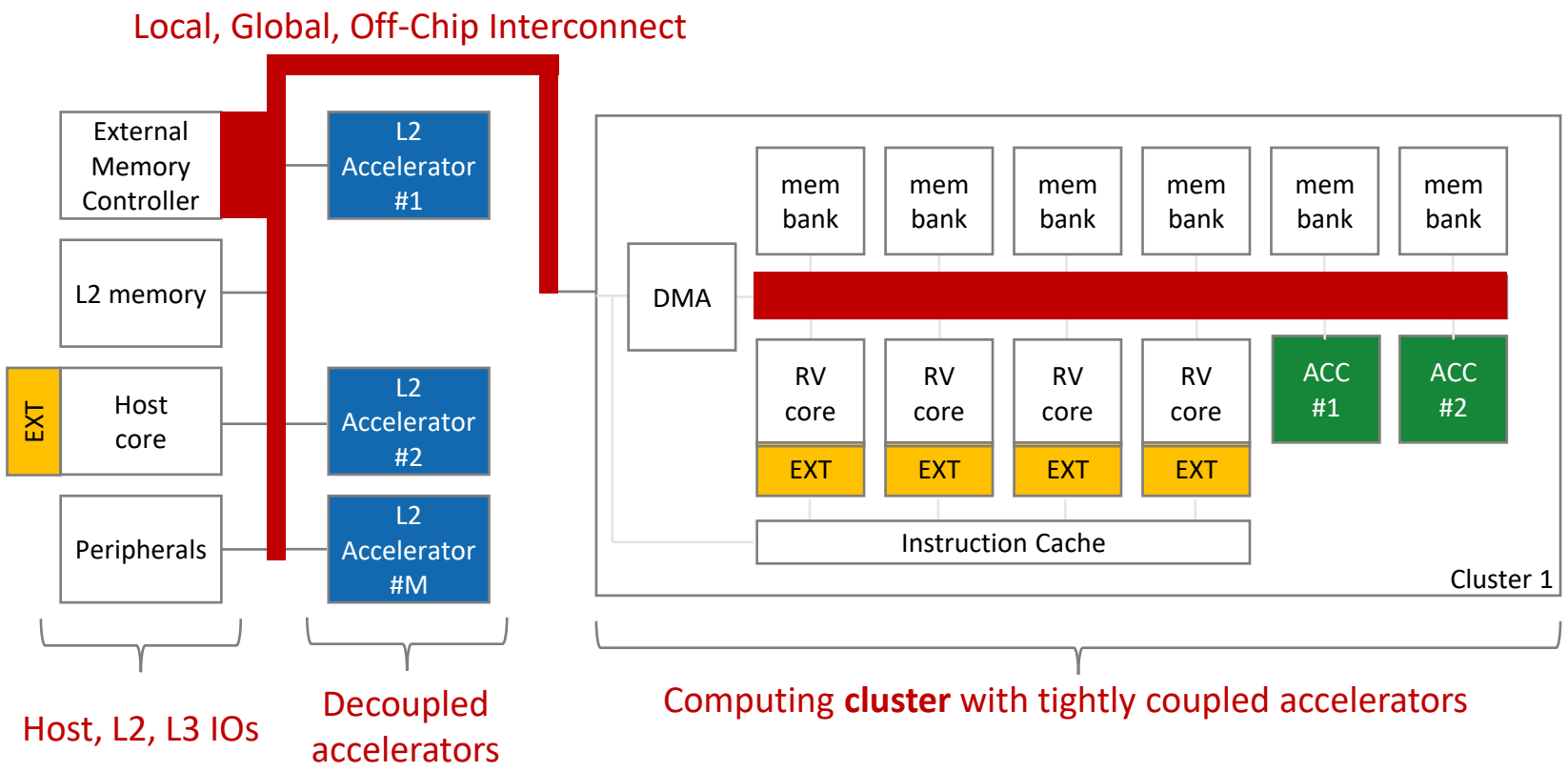
Explicit Coarse Grained offoad

Decoupled L1

Specialized Mem and Interco

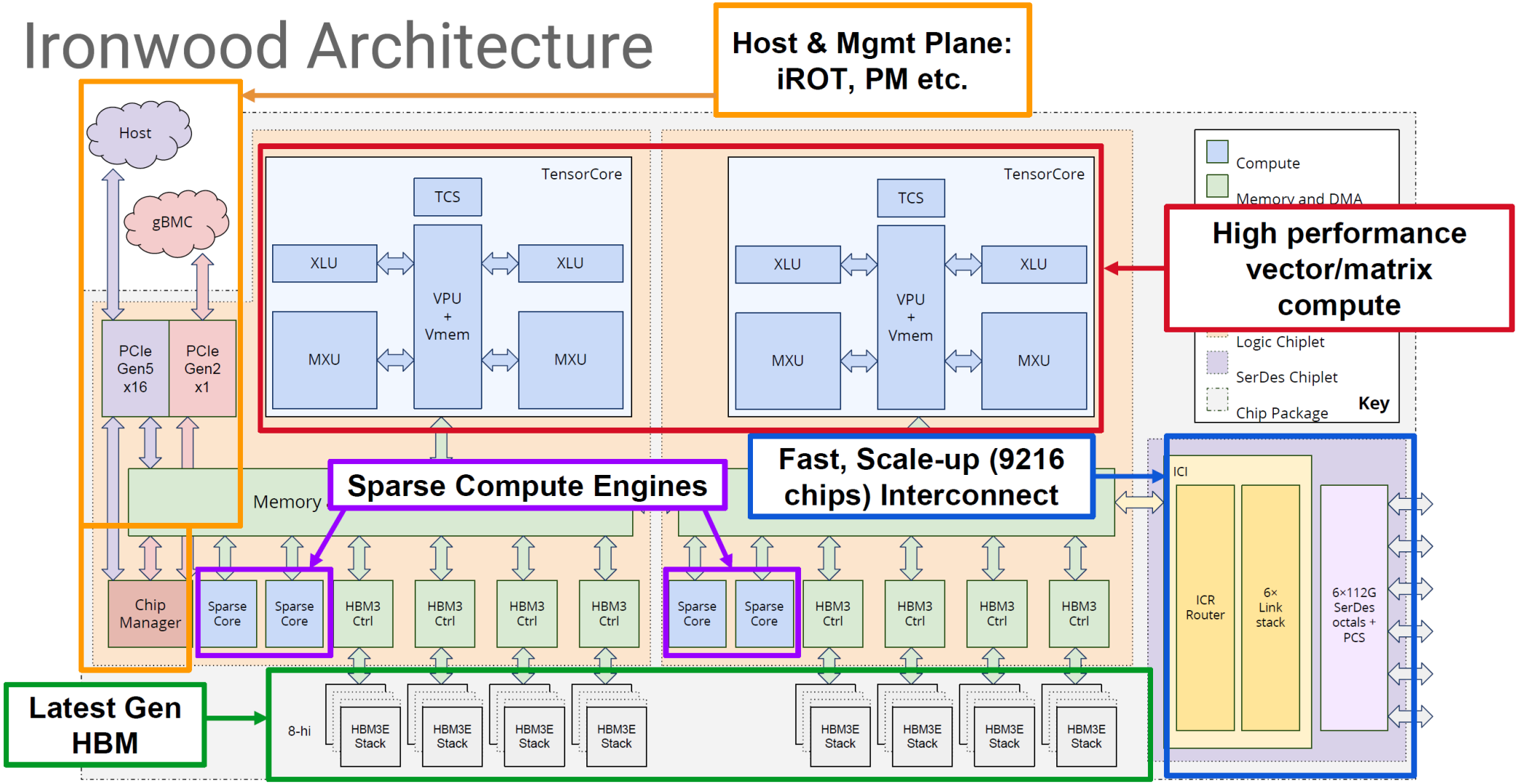
Multicast/reductions/compression

Compute near/in mem & NoC



***Specialization + Scalability,* the two keys!**

A notable Example: Google's Ironwood



Specialization in Perspective: Amdahl's Law

Using 22FDX tech, NT@0.6V, High utilization, minimal IO & overhead

Energy-Efficient RV Core → **20pJ (8bit)**

↓

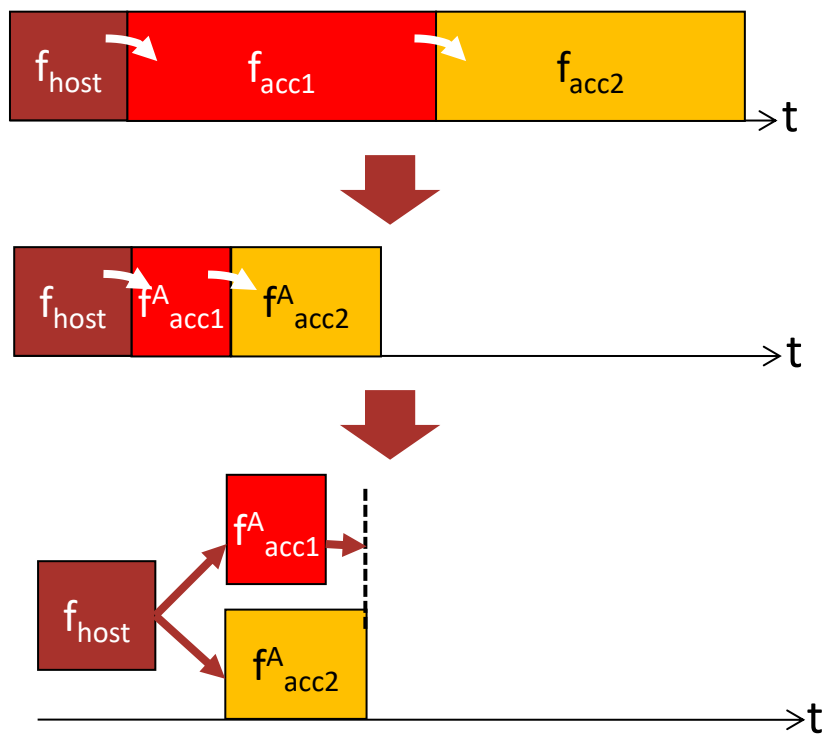
ISA-based 10-20x →**1pJ (4bit)**

↓

Tightly coupled 10-20x → **100fJ (4bit)**

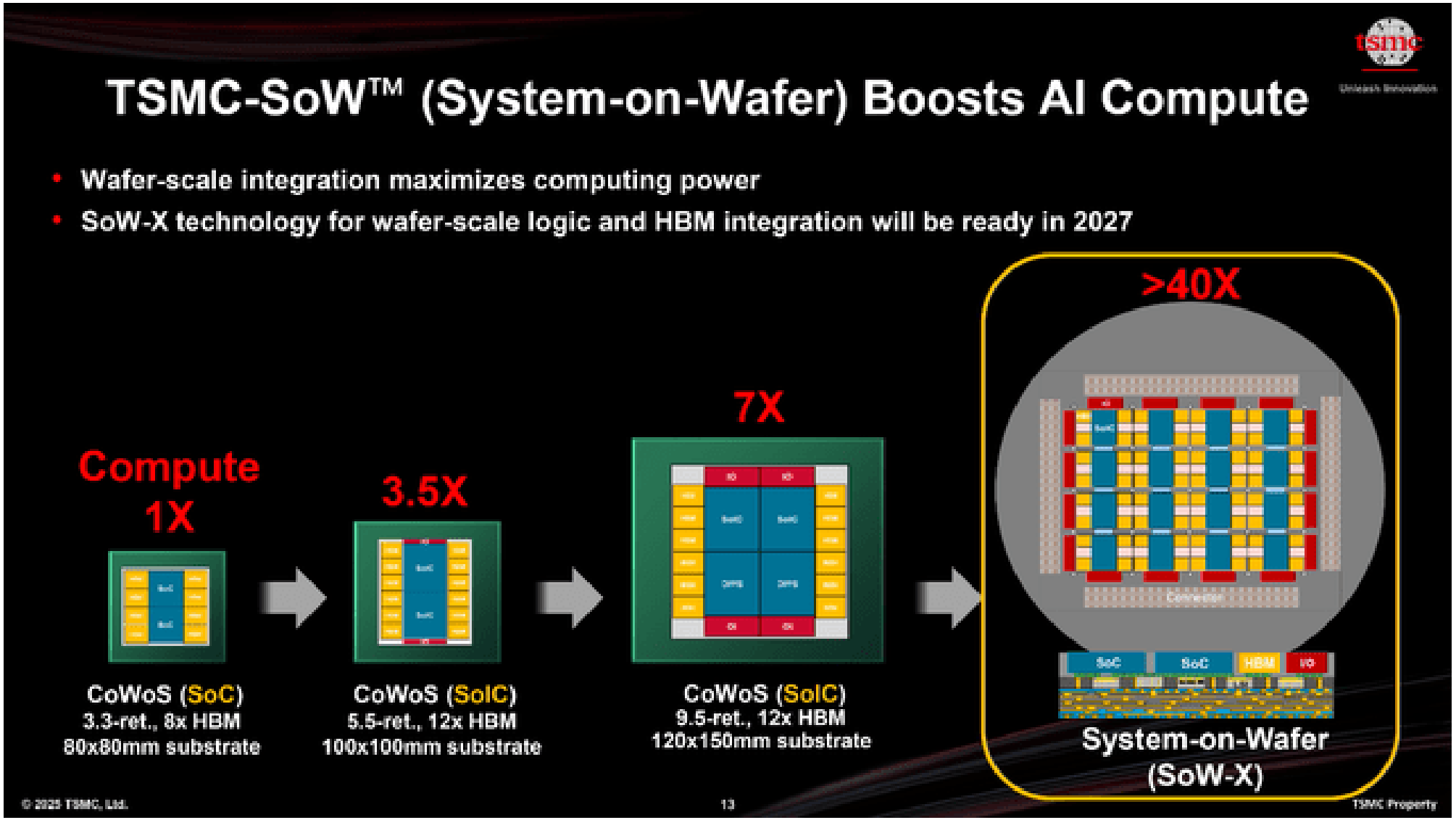
↓

Decoupled 100x →**1fJ (ternary)**

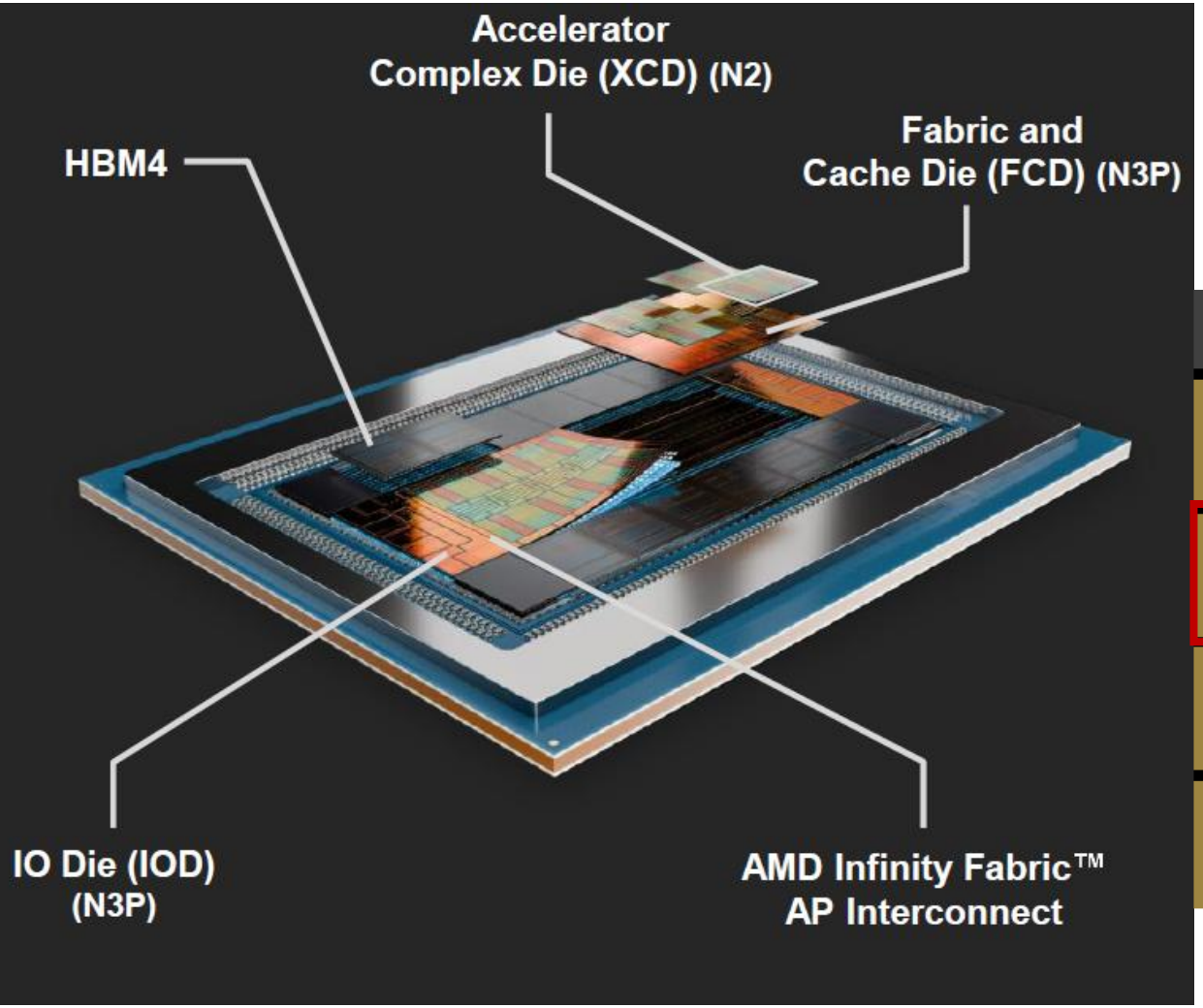


Efficiency vs. utilization Tradeoff → PM to manage idleness

Booster 2: Ultra-high density, 2.5→3D, WS integration



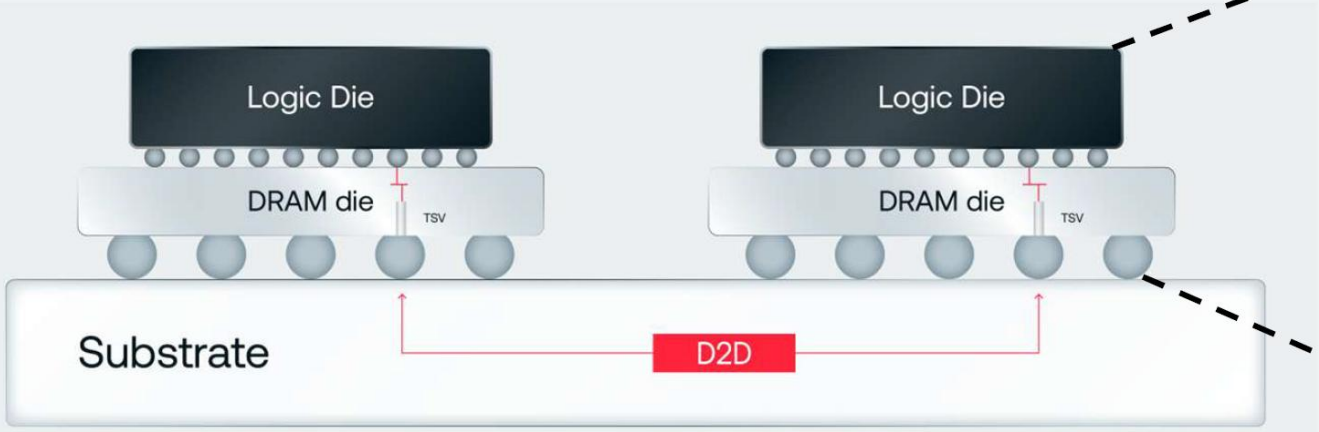
3D for Scale-out: Already in Products!

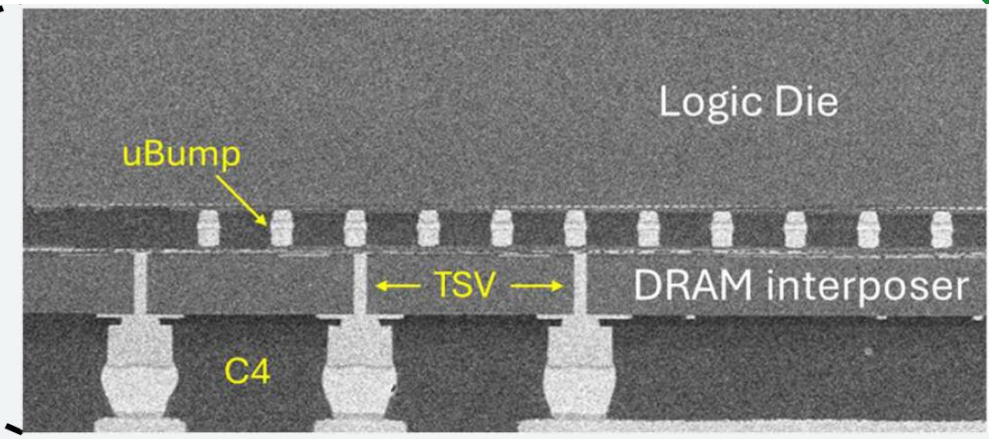


Innovation	What's New	AI Workload Impact
Chiplet Architecture	Compute, memory, cache, and I/O functions are partitioned across specialized dies	Optimizes each function independently for performance, power, and manufacturability
3D Hybrid Bonded XCDs	Compute dies stacked and connected through high-density die-to-die interconnects	Higher compute density and improved performance-per-watt
HBM4 Integration	12 HBM4 stacks 432 GB capacity 23.3 TB/s bandwidth	Larger models, KV caches, and activations remain resident in local memory
CoWoS-L Advanced Packaging	High-bandwidth on-package communication between compute, memory, and I/O dies	Low latency and high bandwidth enable both GPU-wide single operators and sharding techniques like KV cache parallelism, TP, and EP

AMD – Instinct MI455x Hot Chips 2026

Looking forward: more (3D) Heterogeneity + higher stacks







- Top die: TSMC N5 logic
- Bottom die: 3D DRAM
- Integration: 36µ Face to Face (F2F) stacking
- Proven low cost, high volume, high yield process

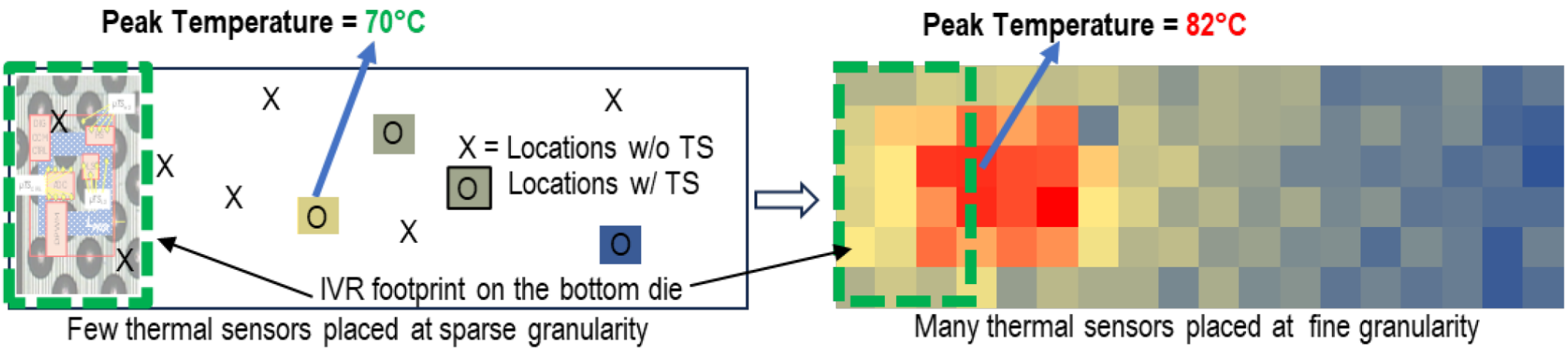
Efficiency vs. Density Tradeoff → PM to manage Temperature

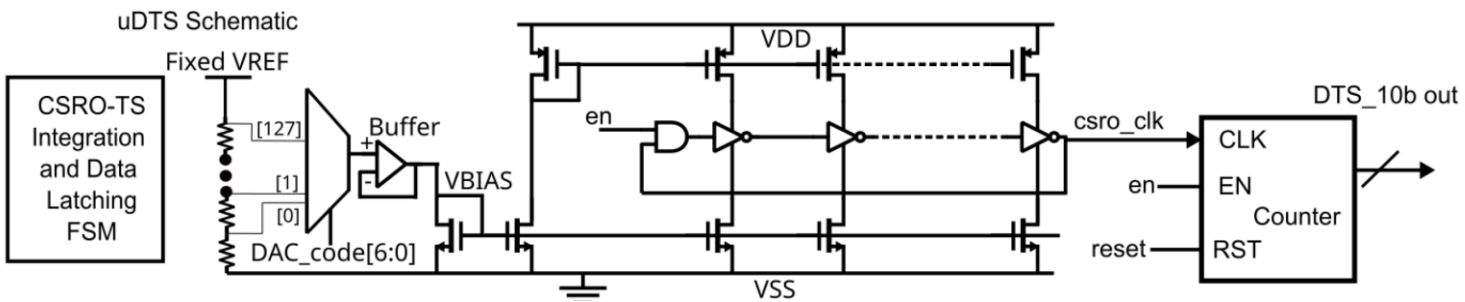
Outline

- **Boosting efficiency for AI workloads**
- **Managing idleness and temperatue in heterogeneity in accelerated systems**
- **Using AI for managing AI**
- **Conclusions, future perspectives**

Ultra-fine grained (digital) sensing (Temp)

Micro-scale Sensors placed in a fine-grained array improve resolution and enable better thermal decisions

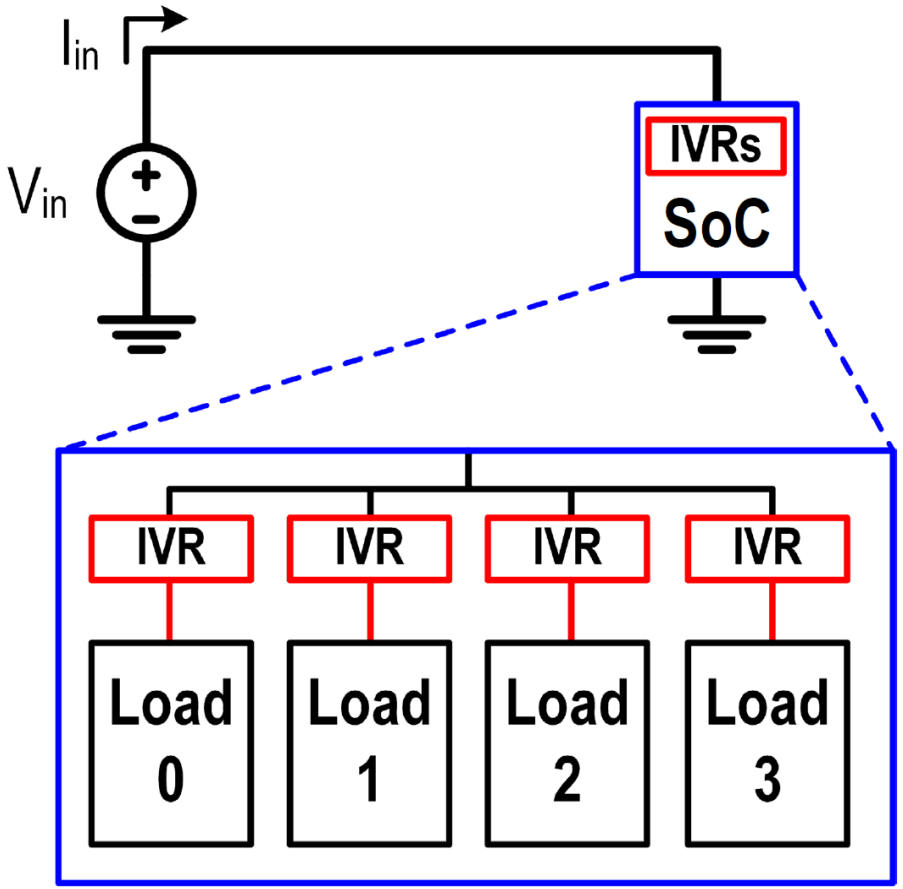




- **Linearity and sensitivity of the CSRO-TS are programmed through VBIAS**
- **Programmable integration time enables noise robustness**

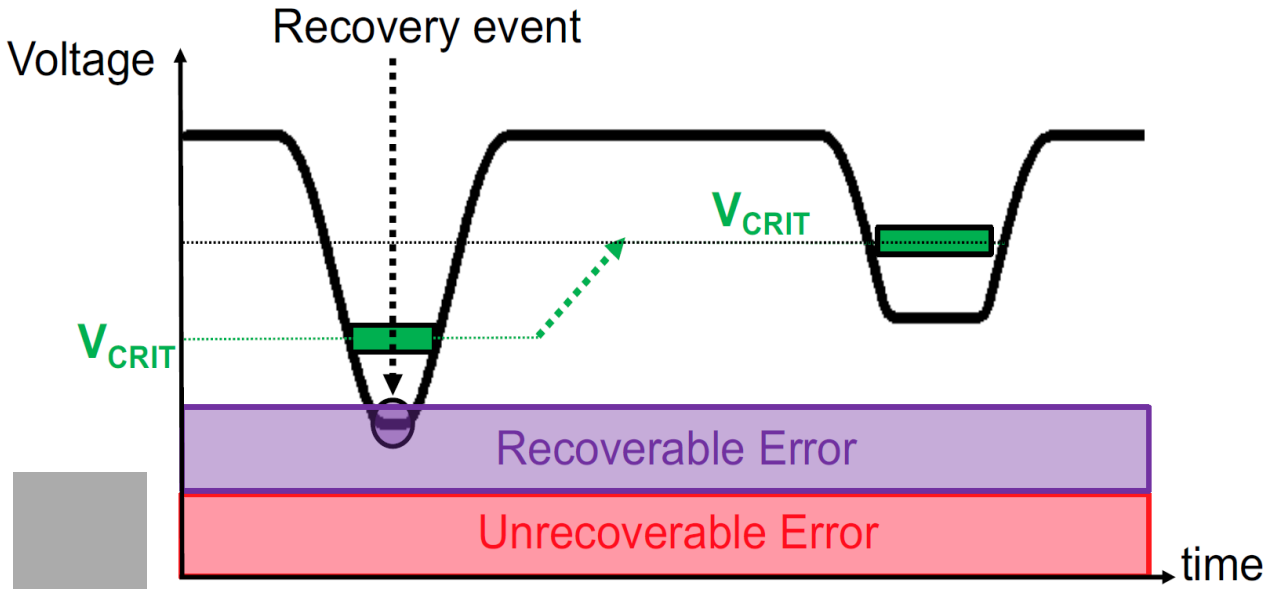
INTEL ISSCC25 8.8

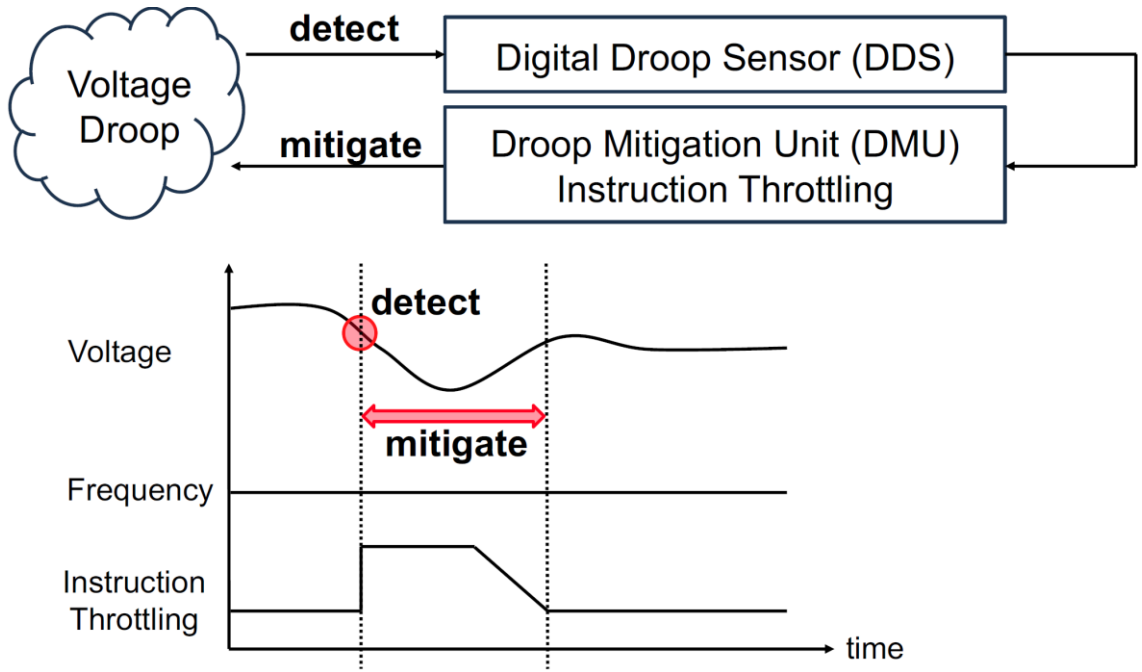
Fine-grained Regulation (Vdd)



- **Minimize <u>total</u> energy dissipation**
  - Voltage Regulator (VR)
  - Load (SoC):
- **Fine-grained DVFS**
  - Improved η_{soc}
  - Needs multiple VRs
- **Power Distribution Network ➔ IVR**
  - Scalability vs η_{VR} tradeoffs

MIMO Control: IBM Telum2 (Z-systems)

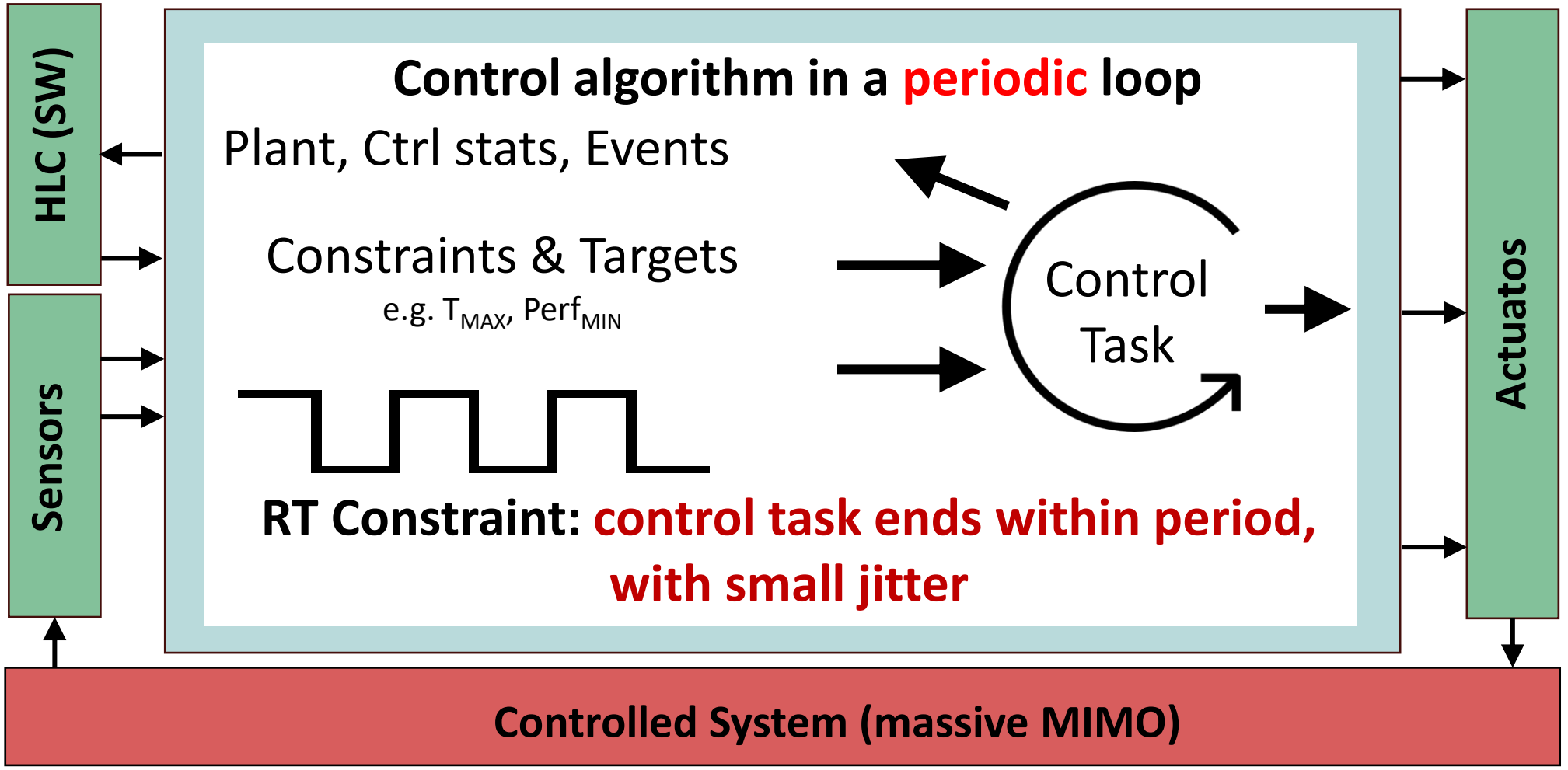




IBM ISSCC25 8.1

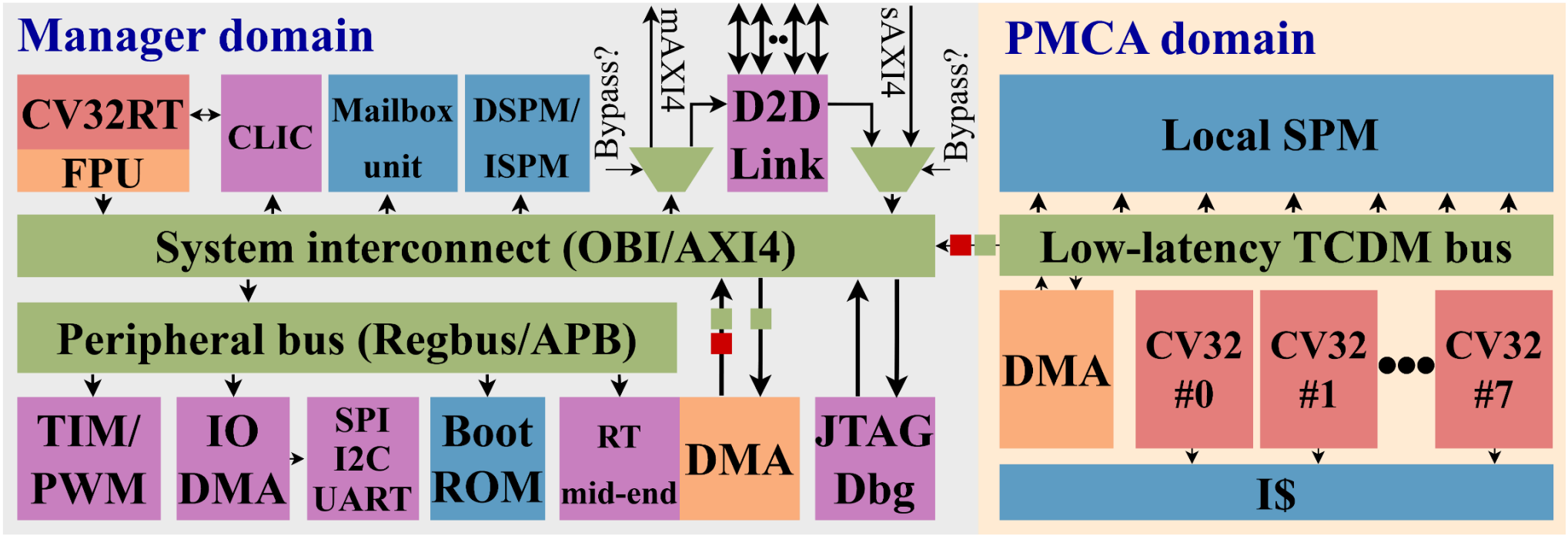
- Multiple interacting control loops
  - Performance loop
  - Voltage droop loop
  - Recoverable error loop (exploiting RAS feature – incl. robust core recovery)

Power Management: Massive Multi-objective MIMO Control



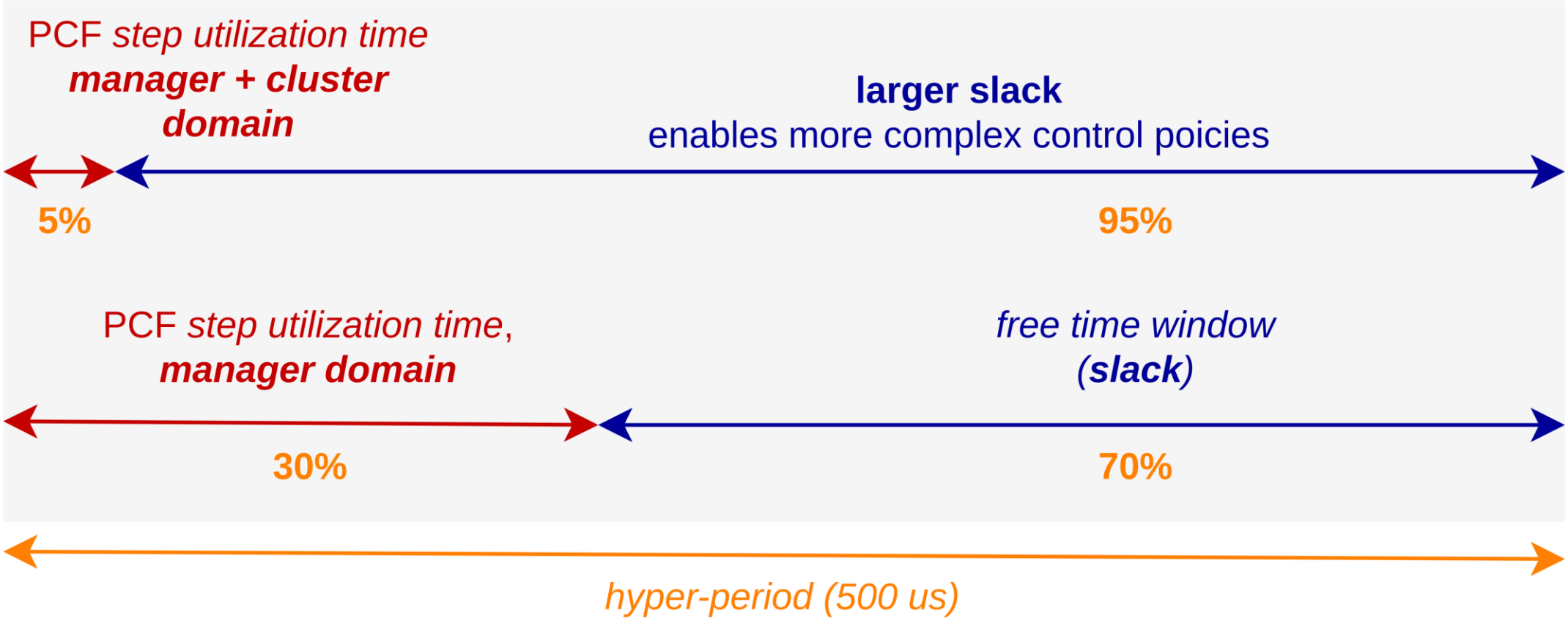
Control-PULP-let: A RT Accelerated PM Engine

- CV32RT: fast interrupt handling in RISC-V
- D2D (die-to-die) connection for chiplet-based power control
- Scatter-gather DMA for N-dimensional periodic sensor reading + 64x mailboxes for fast message exchange (e.g., SCMI)
- Programmable PM policy Accelerator (PMCA)



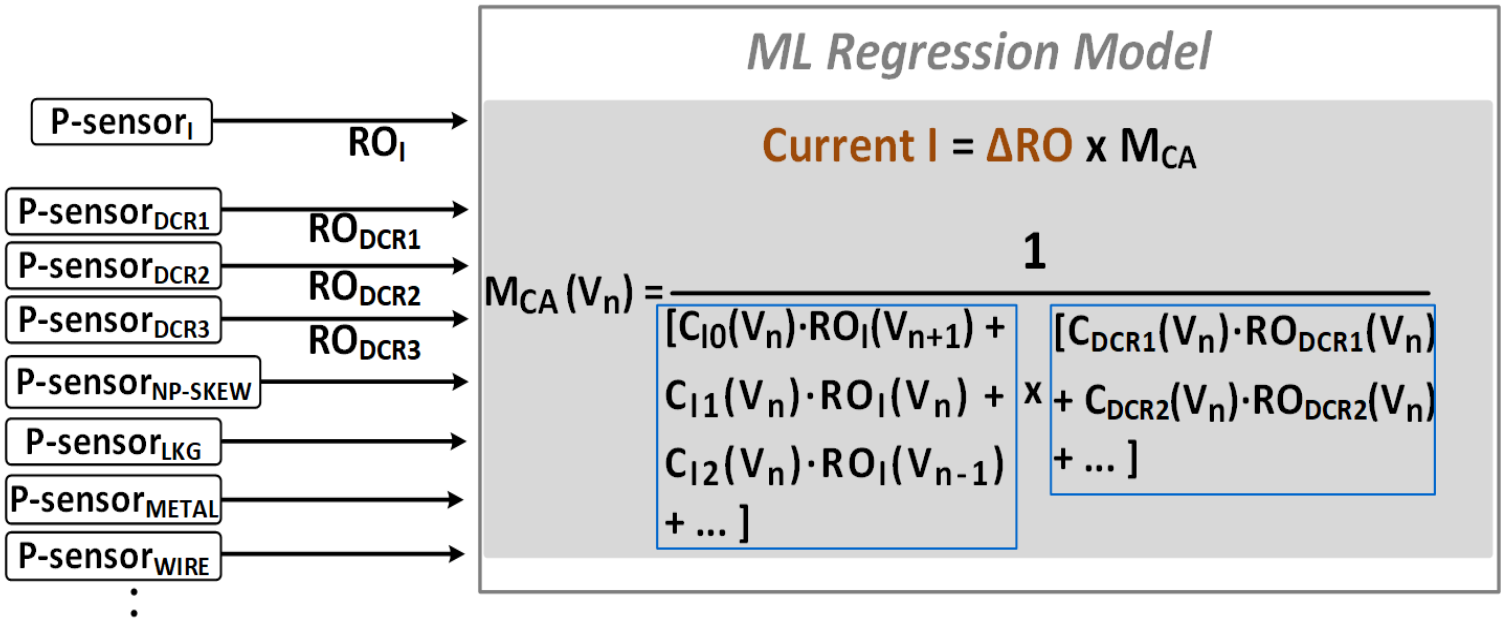
Why Programmable PM Accelerarator?

Programmable accelerator (cluster) advantage for real-time scenarios



ML Based learning for Non-linear Sensors

Current sensing using Count-to-Ampere Models (MCA) based on Ring-oscillator sensors

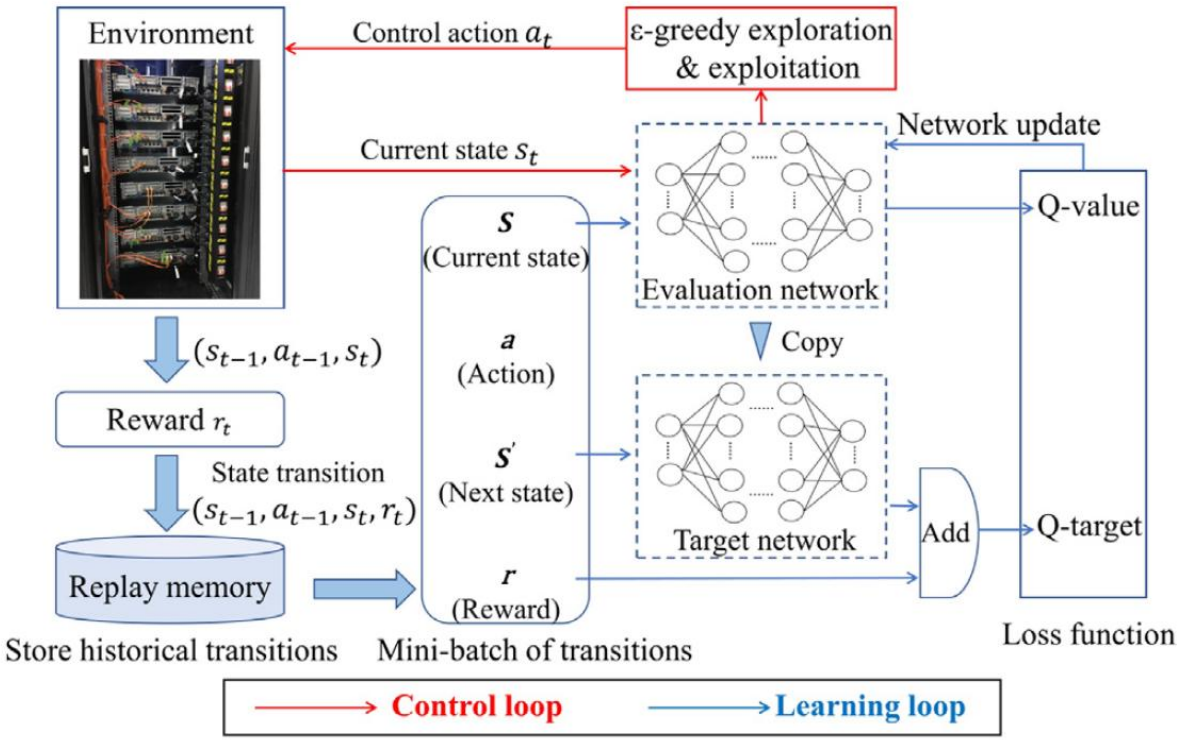


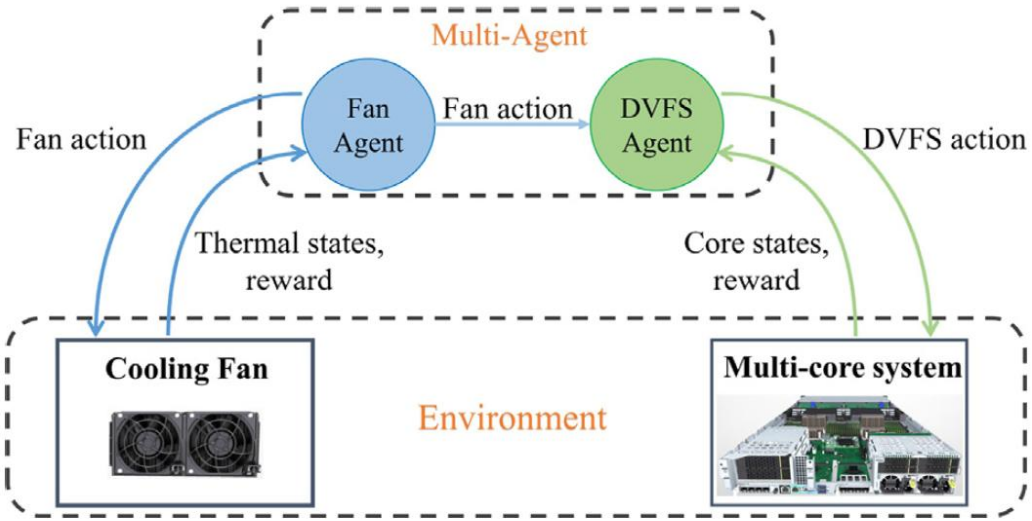
Mediatek ISSCC25 8.1

- ML calibration model leverages silicon data to refine the MCA
  - Scope-measured current I as golden to train the coefficients of ΔRO (RO) from the I-sensor (P-sensors)
  - Multi-scenarios, including generic benchmarks like Geekbenchv6, Antutu, and SpecInt2k6 etc.
- Good tolerances against voltage and temperature fluctuations
- Statistical ML used to eelect the most relevant RO terms (i.e., P-sensors) to simplify the coefficient terms of the model
- Device-specific fine tuning also possible

Advanced AI-based Policies

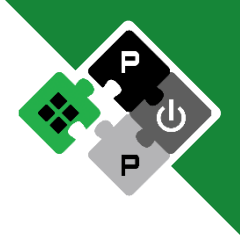
Control law can be learned (e.g. RL) offline and fine-tuned online





www.sciencedirect.com/science/article/abs/pii/S2210537924000222

Summing Up & Future Perspectives



- **PM is more critical than ever for AI hardware**
  - Thermal bottleneck
  - Power delivery bottleneck
  - Impacts the AI bottom line (cost per token)
- **Key requirements for PM in AI hardware**
  - Massive MIMO control (fine-grained & scaled-up)
  - Heterogeneous (processors, accelerators, PIM, PIN…)
  - Multi-die (2.5D → 3.5D)
- **Real-time, (multi-die) scalable and programmable PM needed (open platform is advisable)**
- **AI for power managing AI is a major trend**
  - Learn complex non-linear functions for sensing and actuation
  - Learn optimal policies (e.g. RL)
  - Online learning for fine-tuning device-specific calibration





Thank You!