OPEN Compute Project®

Roofline analysis and profiling of AI / HPC apps

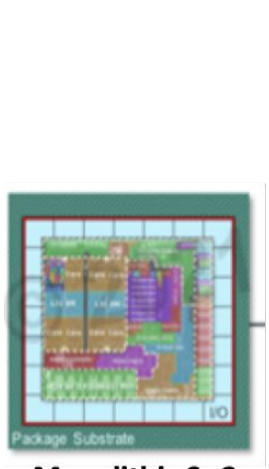
Guide open systems for AI

Georgios Michelogiannakis, Raghu Shankar, and Boon Chong Ang

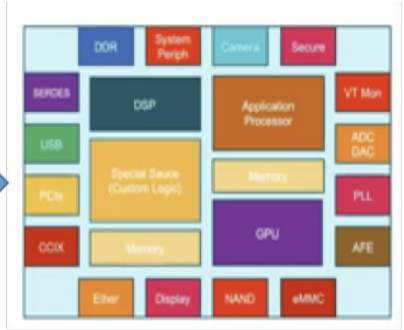
OCP AI/ML Infrastructure Workshop

June 2025

Chiplets Modularity is a promising solution to make HPC&AI specialization accessible



Monolithic SoC



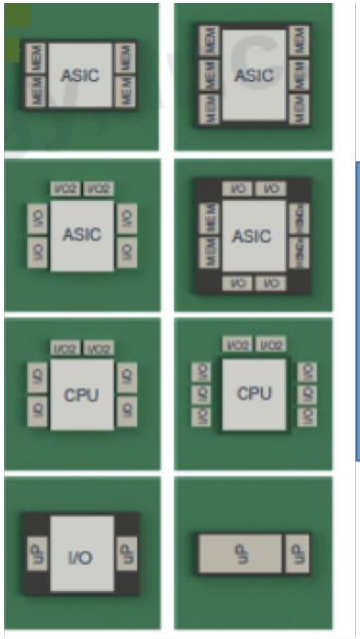
Modularized SoC (Chiplets)

AI/ML

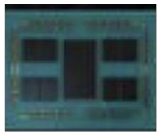
NW Switch

Server

PC



Large FCBGA /MCM



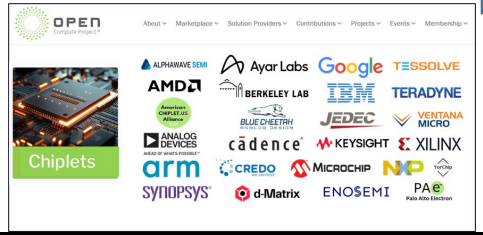
2.5D Si TSV

HD FO RDL /Bridge

3D ubump/ Hybrid bond

Advanced packaging Heterogeneous Integration

Chiplets Marketplace



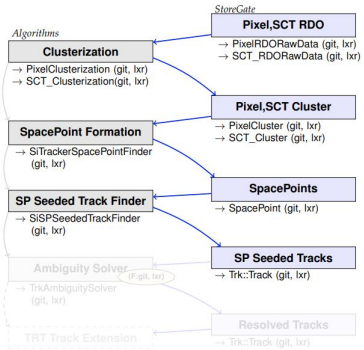
Please attend Patricia Gonzalez-Guerrero's talk later

Application Analysis (Profiling) Guides Chiplet Design

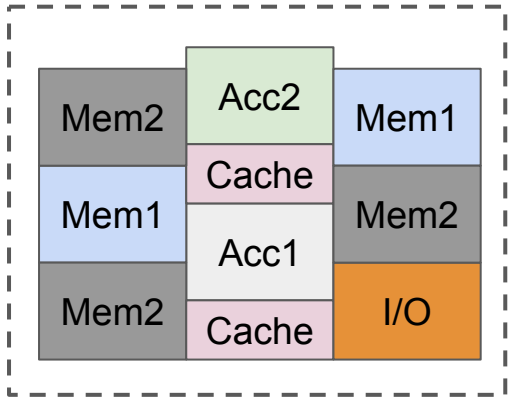
Profiling helps us pick a collection of chiplets (newly-developed or available from the community) for a set of applications

This way, we can serve the mix of applications found in open-science AI & HPC systems and maintain modularity

Application characterization

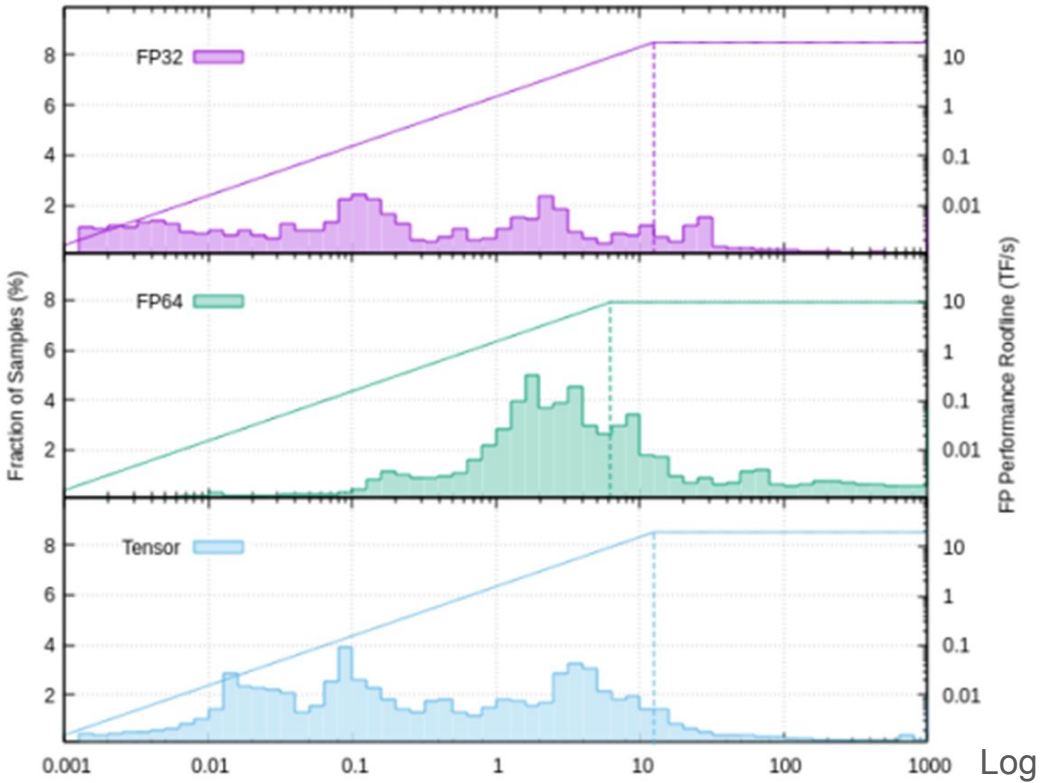


Collection of chiplets that maximize metric of interest



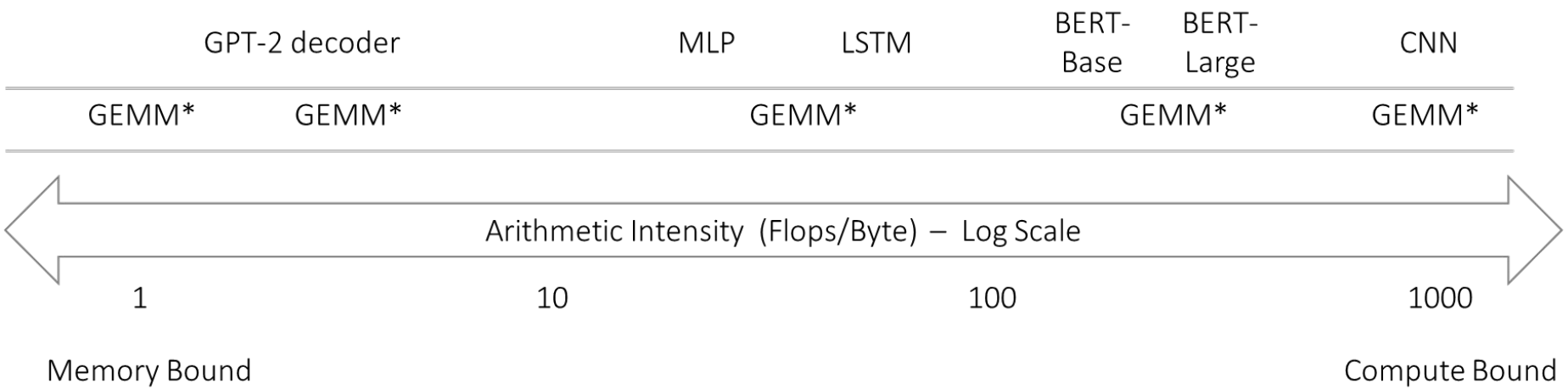
Application Requirements: Literature Search

Compute to Data IO intensity ranges in orders of magnitude – Opportunity for modular & tuned systems optimizing KPIs



- **GPU nodes** – AMD EPYC 7763 “Milan”, 256GB DDR4, 4x Nvidia A100, 40GB HBM2,
- **135,000** jobs, **650** distinct apps during Jul 2024 (avg. ~200 jobs/app)
- **Majority of jobs out of** machine balance
  - Jobs ~0.1 TF/s
  - Systems: ~10 TF/s
- .
- Source: “System-Wide Roofline Profiling - a Case Study on NERSC’s Perlmutter Supercomputer,” Perf Modeling, Benchmarking, and Simulation (PMBS) Nov 2024, Brian Austin, Dhruva Kulkarni, et. al. LBNL

AI – Compute to Memory (Data) ranges in orders of magnitude
Opportunity for modular & tuned systems optimizing KPIs in open systems for AI



- * Arithmetic intensity varies with model size, model, sequence length, data set size, # parameters, #batches, layer
- Adapted from sources:
  (1) "HBM Roadmap ver 1.7 Workshop", KAIST Teralab, June 2025, Terabyte Interconnection and Package Lab
  (2) "Full stack optimization of Transformer Inference: A Survey", Sehoon Kim, et. al, UC Berkeley, arXiv, Feb 2023

Fugaku: HPC jobs characterizations shows wide range

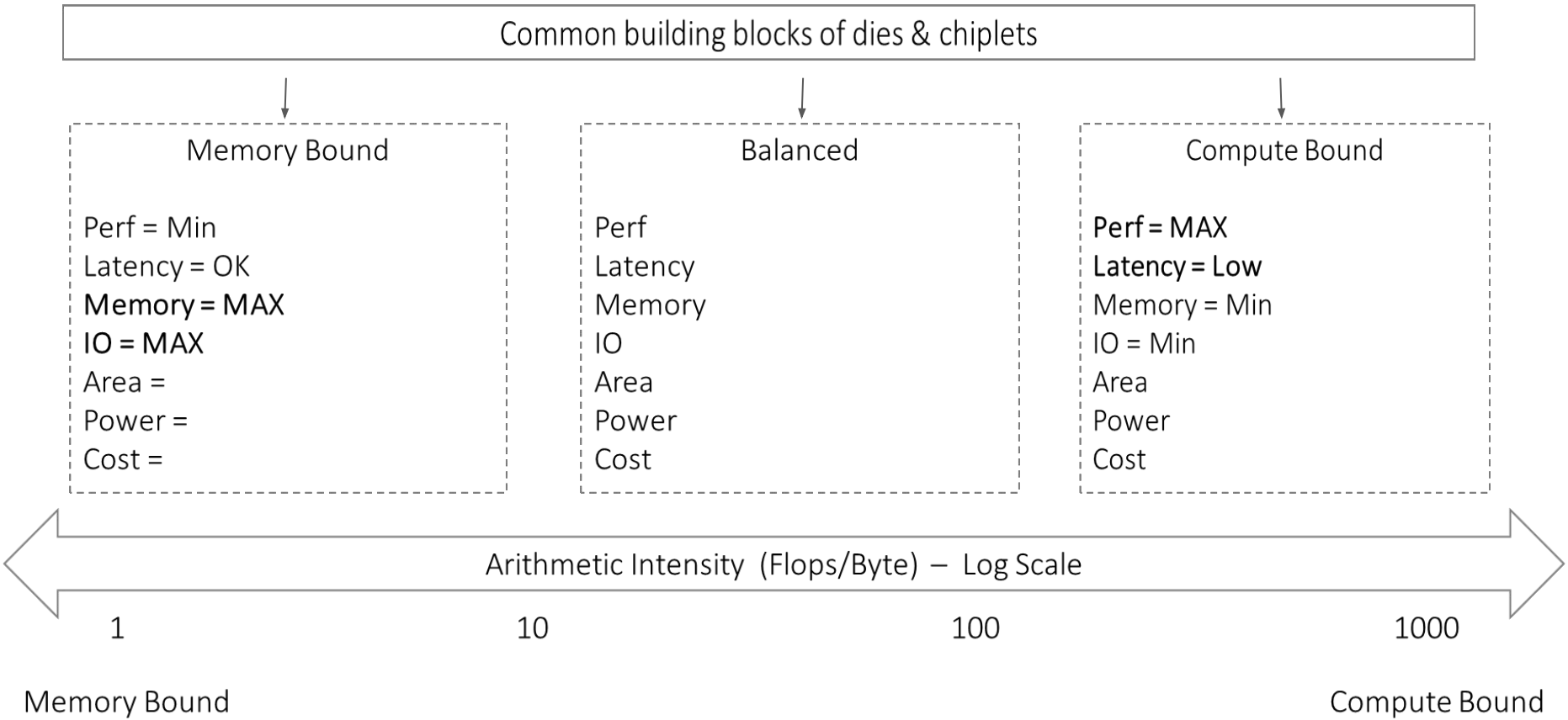
Opportunity: Modularity & IP built for AI leverage for HPC

- 2.1 M jobs
  - Memory bound 1.6M
  - Compute Bound 0.5M

HPC ranges more to the left .. 0.001 to 1

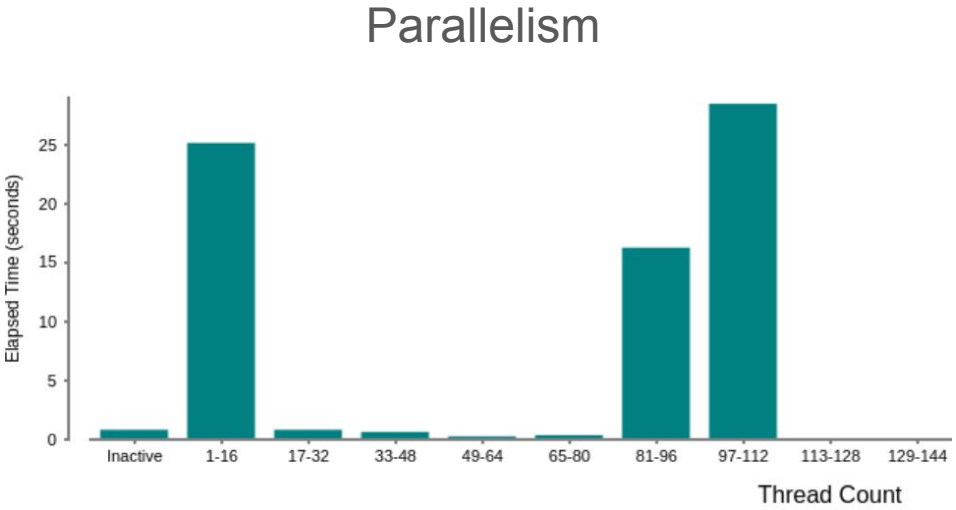
- Source: “MCBound: An Online Framework to Characterize and Classify Memory/Compute-bound HPC Jobs”, Antici, et. al., SC24, Nov 2024

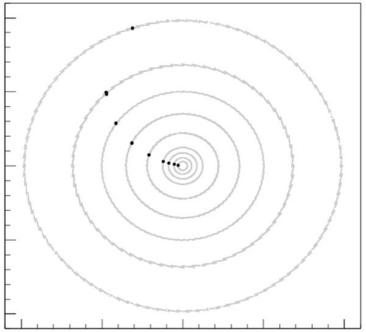
Opportunity for modular & tuned systems optimizing KPIs in open systems for AI



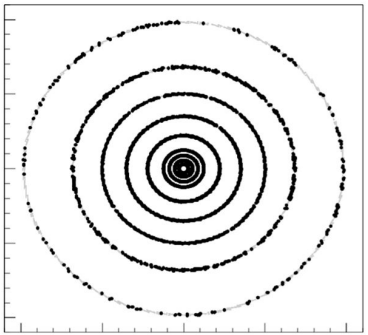
Profiling Results Collected by HPC/AI OCP Workstream Members

Particle Tracking For High Energy Physics (CPU Version)





Hits from one particle

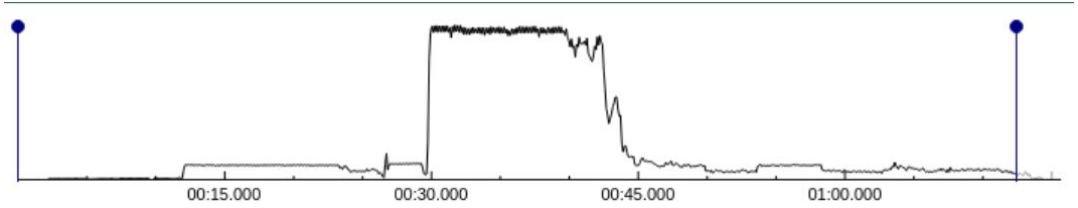


Hits from a fraction of particles
in HL-LHC environment

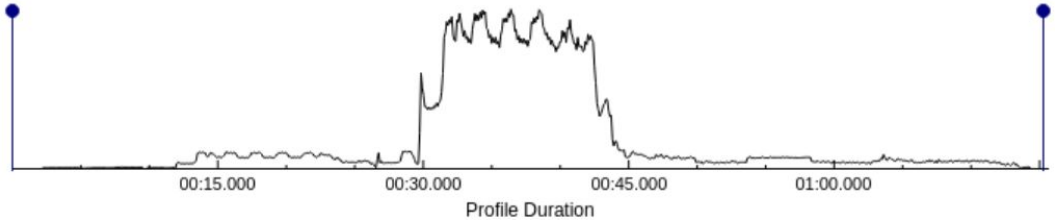
A Combinatorial Explosion in Computational Complexity
With Increased Luminosity for Future Experiments

Impact of Parallelism to Memory

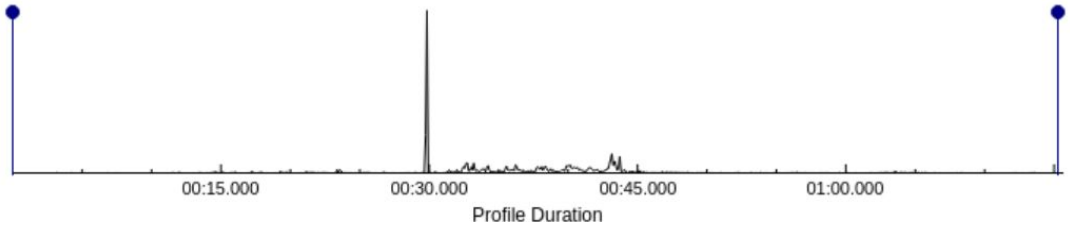
All operations (instructions)



Load/Store histogram



DRAM hit latency



Canneal: Simulated Annealing Approximates Global Optimum

Memory bandwidth

Runtime 3s

Uncore Events

Uncore Event Short Name / Uncore Event Type	Uncore Event Count
UNC_IMC_DRAM_DATA_READS	2,139,123,811
UNC_IMC_DRAM_DATA_READS	2,139,123,811
UNC_IMC_DRAM_DATA_WRITES	265,424,926
UNC_IMC_DRAM_DATA_WRITES	265,424,926

**N/A is applied to non-summable metrics.*

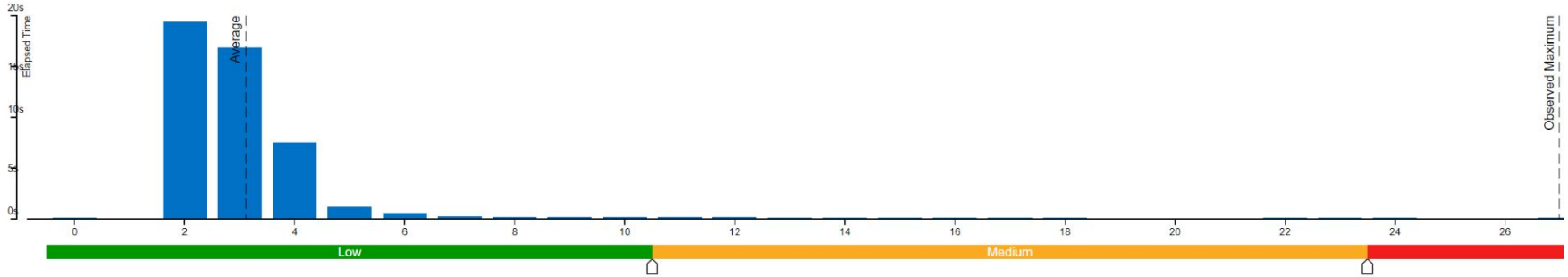
Bandwidth Utilization Histogram

Explore bandwidth utilization over time using the histogram and identify memory objects or functions with maximum contribution to the high bandwidth utilization.

Bandwidth Domain: DRAM, GB/sec

Bandwidth Utilization Histogram

This histogram displays the wall time the bandwidth was utilized by certain value. Use sliders at the bottom of the histogram to define thresholds for Low, Medium and High utilization levels. You can use these bandwidth utilization types in the Bottom-up view t utilization type. To learn bandwidth capabilities, refer to your system specifications or run appropriate benchmarks to measure them; for example, Intel Memory Latency Checker can provide maximum achievable DRAM and Interconnect bandwidth.



Microarchitecture Usage: 27.3% of Pipeline Slots

Retiring:	27.3%	of Pipeline Slots
Front-End Bound:	30.6%	of Pipeline Slots
Bad Speculation:	7.7%	of Pipeline Slots
Back-End Bound:	**34.4%**	**of Pipeline Slots**
Memory Bound:	**21.4%**	**of Pipeline Slots**
L1 Bound:	**16.9%**	**of Clockticks**
DTLB Overhead:	**100.0%**	**of Clockticks**
Load STLB Hit:	100.0%	of Clockticks
Load STLB Miss:	8.2%	of Clockticks
Loads Blocked by Store Forwarding:	0.2%	of Clockticks
Lock Latency:	1.3%	of Clockticks
Split Loads:	0.0%	of Clockticks
4K Aliasing:	0.1%	of Clockticks
FB Full:	11.3%	of Clockticks
L2 Bound:	0.9%	of Clockticks
L3 Bound:	**5.4%**	**of Clockticks**
Contested Accesses:	0.3%	of Clockticks
Data Sharing:	0.1%	of Clockticks
L3 Latency:	7.0%	of Clockticks
SQ Full:	0.0%	of Clockticks
DRAM Bound:	**18.8%**	**of Clockticks**
Memory Bandwidth:	23.0%	of Clockticks
Memory Latency:	20.1%	of Clockticks
Store Bound:	**0.6%**	**of Clockticks**
Core Bound:	13.0%	of Pipeline Slots

Memory Bound: 21.4% of Pipeline Slots

Cache Bound:	23.3%	of Clockticks
DRAM Bound:	**18.8%**	**of Clockticks**
Average DRAM Bandwidth, GB/s:	4.169	

mJ/sec

What’s Next

Request for profiling and prototyping resources

Results critical for recommending next-gen specialized modular chiplet systems

1. Compute/GPU resource for profiling from cloud service providers
   - 2-socket x86 servers nodes + 2x GPUs – 6 months
   - 256 GB DIMM +
   - 2x Nvidia H100, 80 GB, HBM3 or equivalent
   - SSD 300GB +
   - Ubuntu 22.04
2. Need trial licenses for virtual prototyping:
   - Synopsys (platform architect, zebu for FPGA prototyping, design compiler, etc)
   - Cadence
   - Siemens EDA
3. Keysight CloudBuilder, Keysight AI Data Center Builder,
   - Software suite designed to emulate real-world AI workloads, enabling validation and optimization of AI infrastructure components like networks, hosts, and accelerators
4. Grow team of contributors!