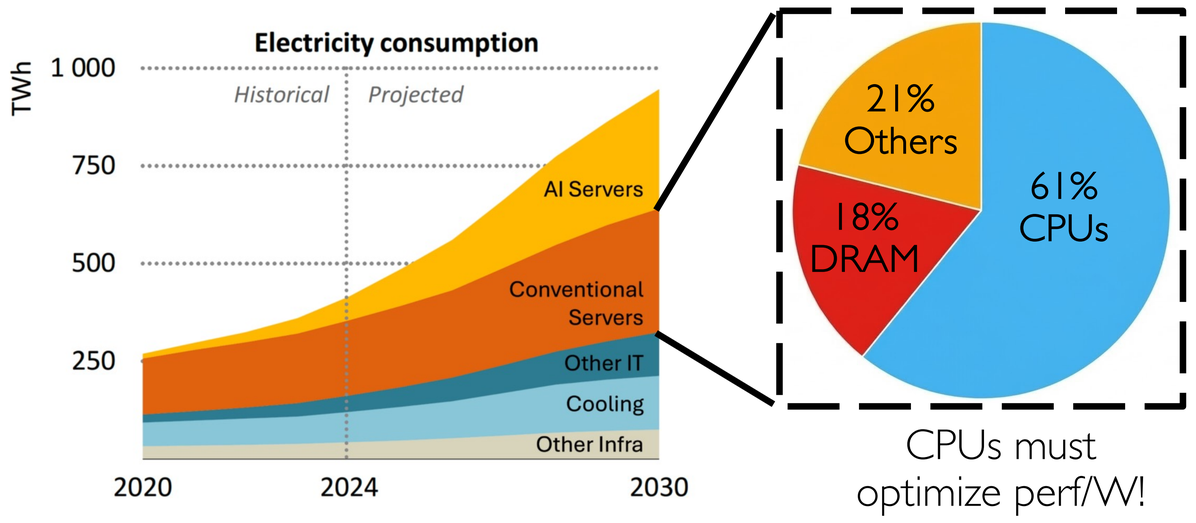
Optimal Silicon Efficiency in Servers

Ayan Chakraborty †, Ali Ansari †, Shanqing Lin †, Yuanlong Li †,
August Ning †, Mohammad Alian ✣, Michael Ferdman ★, Pejman Lotfi-Kamran ‡, Babak Falsafi †

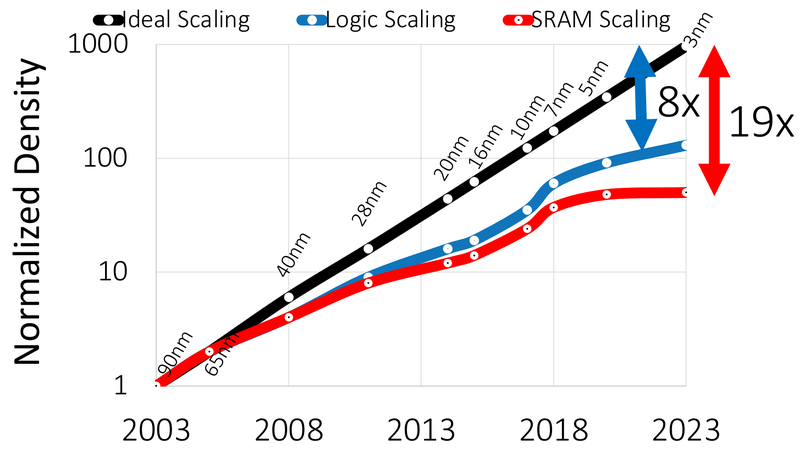
† EcoCloud, EPFL ✣ Cornell University ★ Stony Brook University ‡ EPFL/IPM

Datacenters Are Running Out Of Power



CPUs must optimize perf/W!

Moore's Law Is Dead

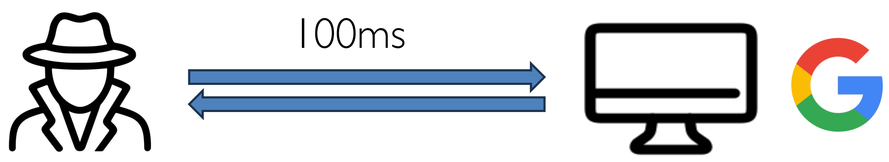


- Cannot have more transistors within the same area budget
- Bigger CPUs increase CAPEX

CPUs must optimize perf/mm²!

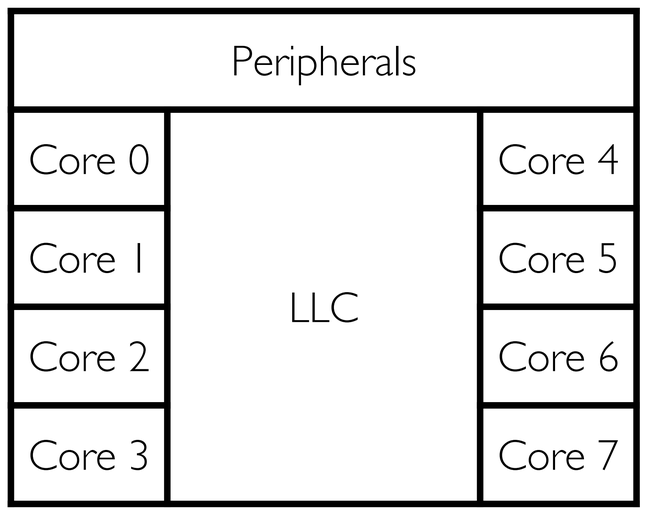
What Is Important For Server Workloads?

- Server workloads primarily care about throughput (reqs/sec)
- However, online services must also feel responsive



- Workloads' service-level objectives (SLO) must be met
- Performance metric: Throughput under SLO

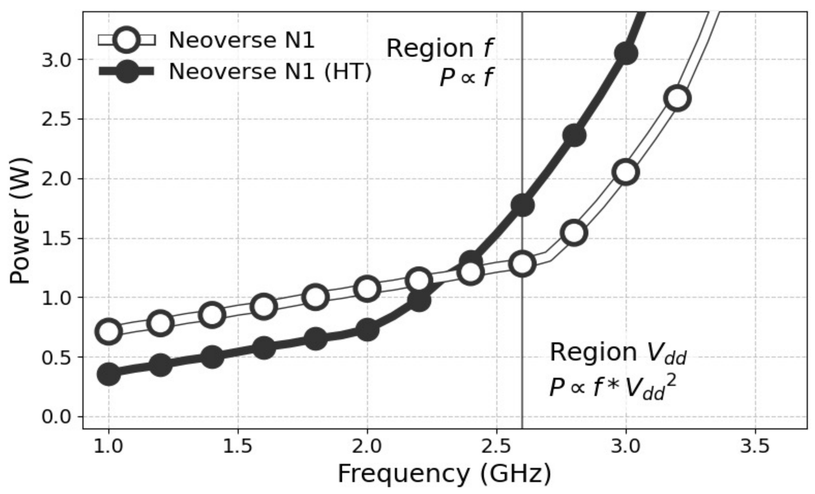
Modern CPUs Optimize For Single Core Performance



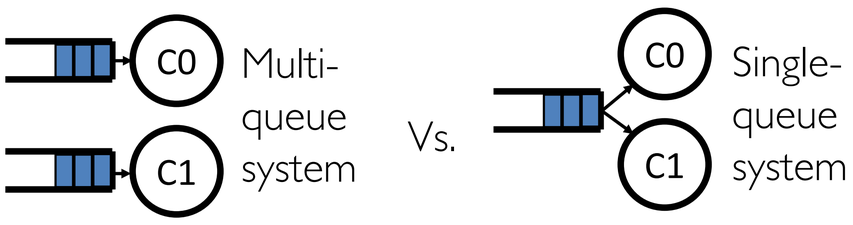
- A single Zen 3 chiplet: 8x 4-way OoO @ 3 GHz
- LLC is over-provisioned and acts as dim silicon
- Meets SLO ✓
- Single-core perf ↑
- Number of cores ↓

What is optimal silicon efficiency w/ SLO & fixed power density?

Perf. Tradeoffs With Power



- Power reduces super-linear with frequency
- But service rate is sub-linear with frequency
- Queues essential to effectively use extra cores



✓ Low Sync Overhead ✗ Load Balancing

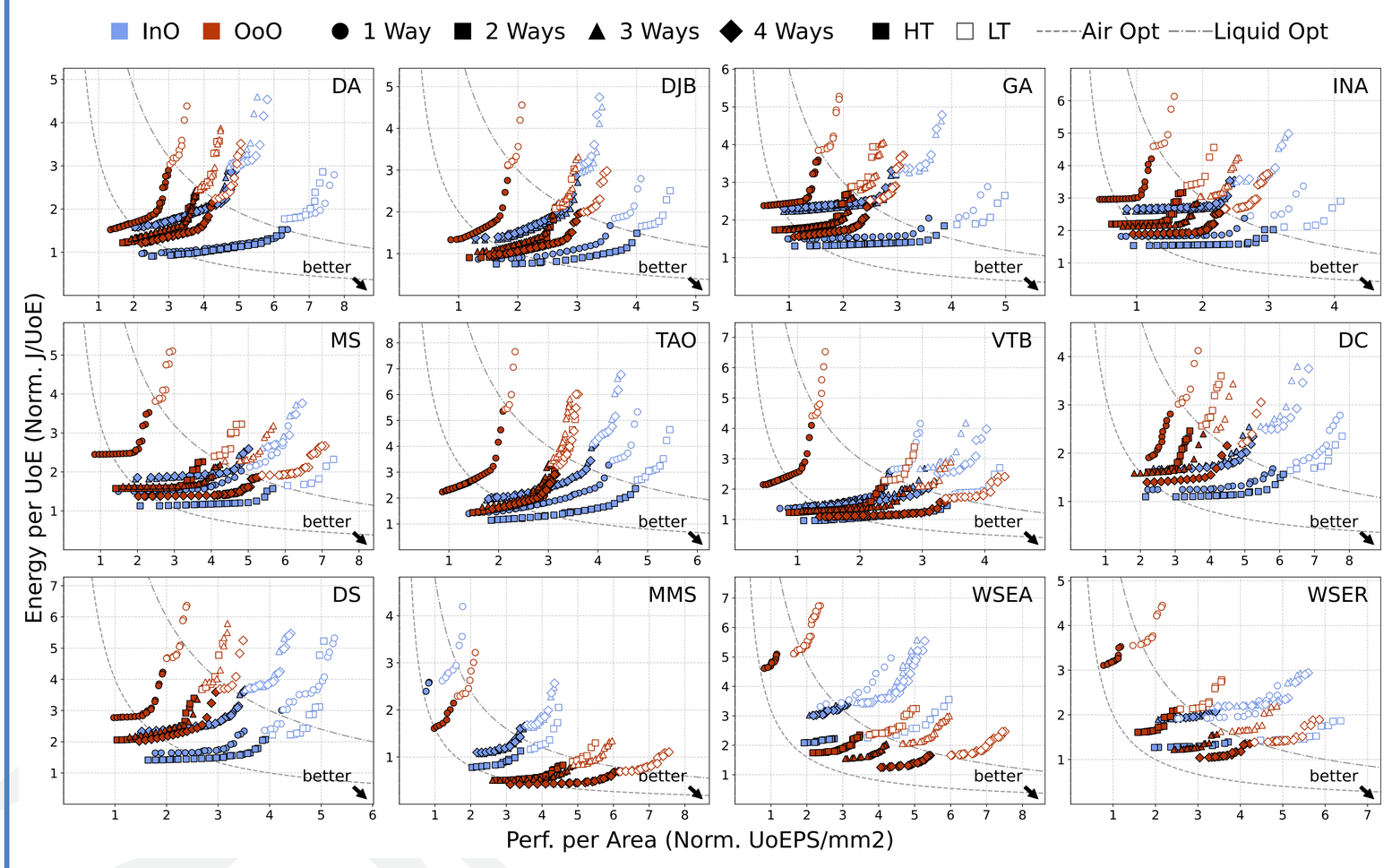
✗ Low Sync Overhead ✓ Load Balancing

Design Space

Parameter	Range
Core Microarch.	InO, OoO
Pipeline Width	1, 2, 3, 4
L1 (4-way) Size (KB)	8, 16, 32, 64
BTB (4-way) Entries	1K, 2K, 4K, 8K, 16K
SMS (16-way) entries	N/A, 1K, 2K, 4K, 8K, 16K
Clock Frequency	1GHz – 3GHz
Transistor Type	LT, HT

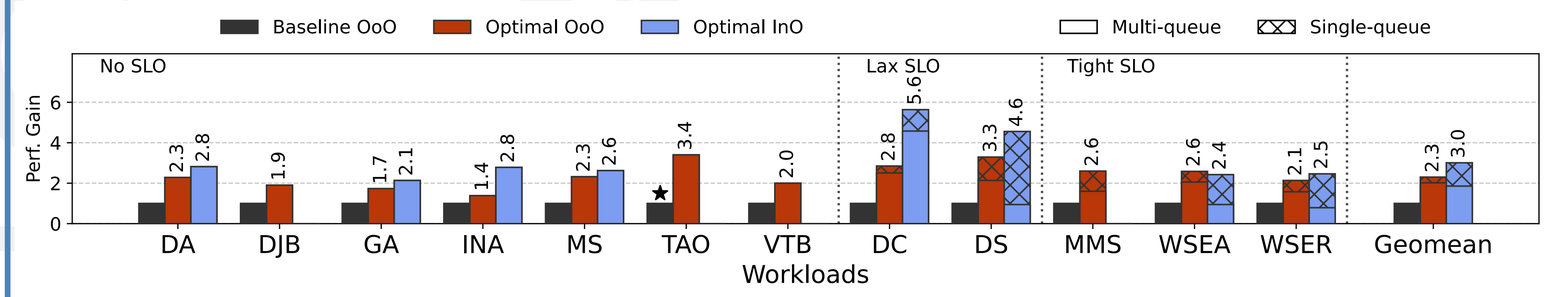
Workloads: CloudSuite, DCPerf, Microservices

Design Space Exploration For Pareto Optimal Configurations



Single core pareto optimal configurations

- Used to find optimal microarch. and frequencies that tradeoff perf/mm² and perf/W



Evaluation of optimal air-cooled OoO and InO chiplets

- Reducing frequency to 1.5-2GHz with HT transistors leads to 2.3× geomean perf. gain
- Single-queue systems allow InO to be competitive even for tight SLO workloads