Server Benchmarking with CloudSuite 4.0

Ali Ansari†, Shanqing Lin†, Rafael Pizarro Solar ✣, Ayan Chakraborty†, Bugra Eryilmaz†

Babak Falsafi†✣, Michael Ferdman‡

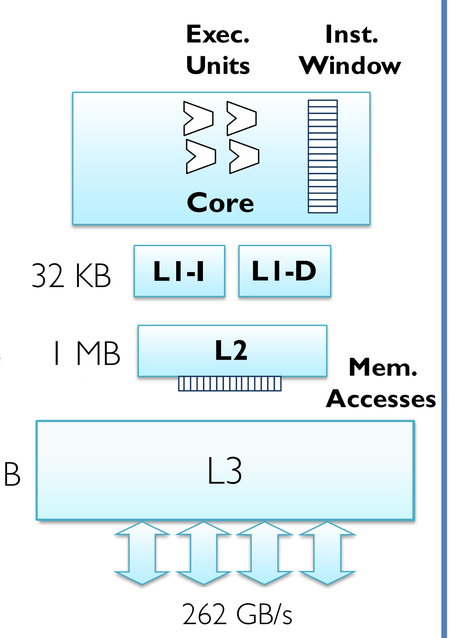
†PARSA, EPFL ✣ EcoCloud, EPFL ‡Stony Brook University

Cloud Server Efficiency

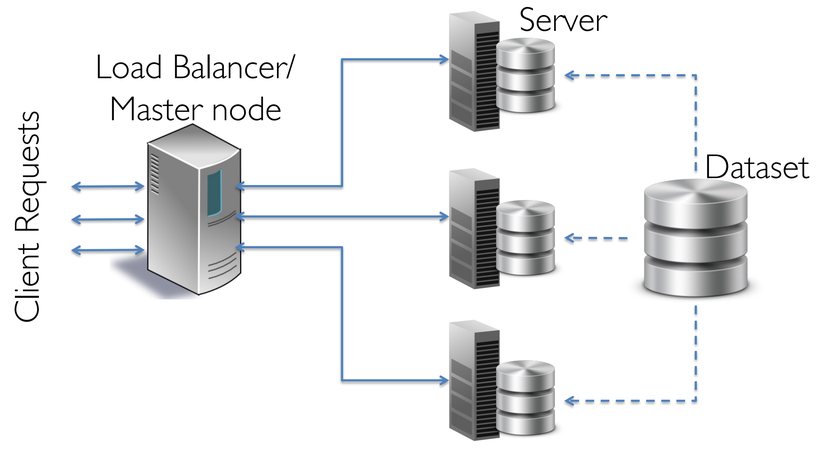
- Constant demand for more servers
- Increasing costs of HW, space & power

Modern Servers are Scale-Up

- Aggressive cores
- Large instruction window
- L2 & large L3 cache
- Vast Bandwidth



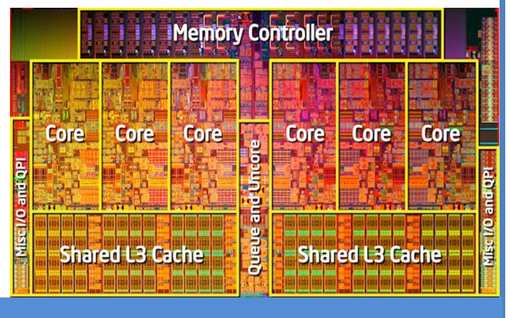
Cloud Applications are Scale-out



- Serve independent requests/tasks
- Operate on huge dataset split into shards
- Communicate infrequently

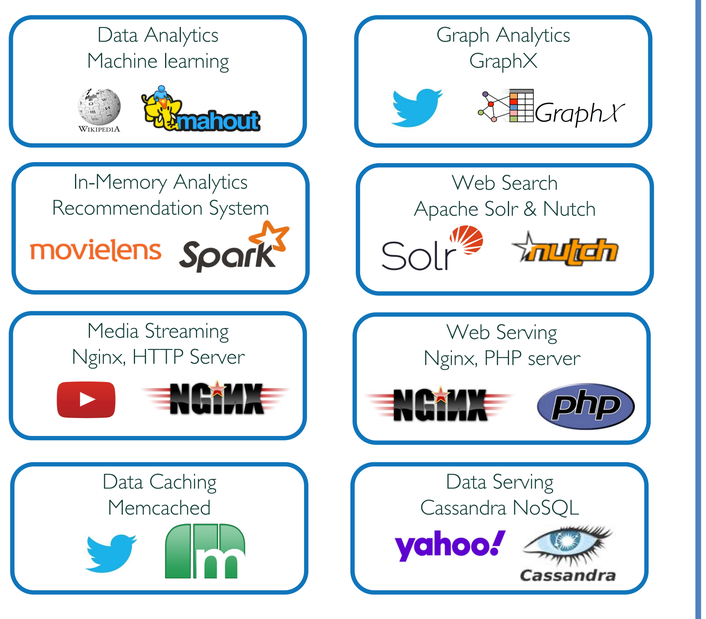
Why not Conventional Scale-Up Processors?

- Developed based on general purpose applications' needs
- One size does not fit all: need for workload-specific hardware specialization
- Missing notion of repetitive request handling
- Clearing the Clouds [Ferdman, ASPLOS'12] already highlighted:
  - Too fat cores: Low power efficiency
  - Too few cores: Low parallelism
  - Too much cache: Slow, waste of silicon



Need for specialized scale-out processors

CloudSuite 4.0



Available on x86, ARM, and RISC-V

Prior Research Using CloudSuite

- [Ferdman, ASPLOS'12]: Clearing the Clouds, examining the scale-out server workloads' execution requirement mismatches with hardware
- [Lotfi-Kamran, ISCA'12]: Scale-out processors, a cloud-native CPU microarchitecture which used in designing Cavium ThunderX CPUs
- [Karakostas, IISWC'14]: Performance Analysis of the Memory Management Unit under Scale-out Workloads
- [Bakhshalipour, HPCA'19]: Bingo Spatial Data Prefetcher
- [Gupta, HPCA'23]: A flash-based system for online services

CloudSuite enabled various architectural research in the community

Research Directions with CloudSuite

- Revisiting the workloads' requirements from hardware
- The difference between x86, ARM, and RISC-V platforms' characteristics for running scale-out server workloads
- Customizing the instruction set for the server workloads' semantics
- Designing accelerators for post-Moore era datacenters
- Improving silicon efficiency for future processors' microarchitecture

Countless and interesting research opportunities with CloudSuite!

PARSA PARALLEL SYSTEMS ARCHITECTURE LAB · EPFL · COMPAS Computer Architecture Stony Brook · Stony Brook University · EcoCloud