Server Benchmarking with CloudSuite 4.0

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

Babak Falsafi†✢, Michael Ferdman‡

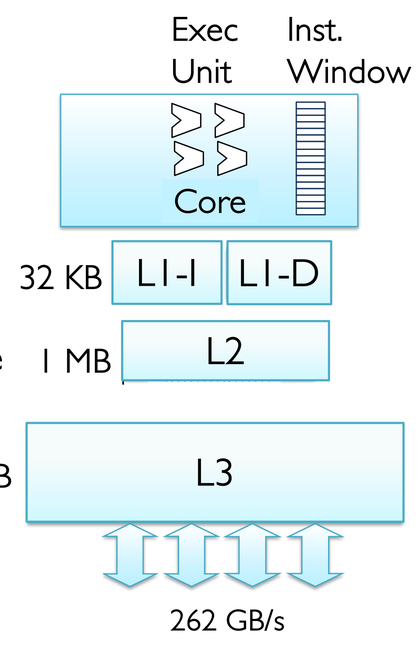
†PARSA, EPFL ✢ EcoCloud, EPFL ‡Stony Brook University

Cloud Server Efficiency

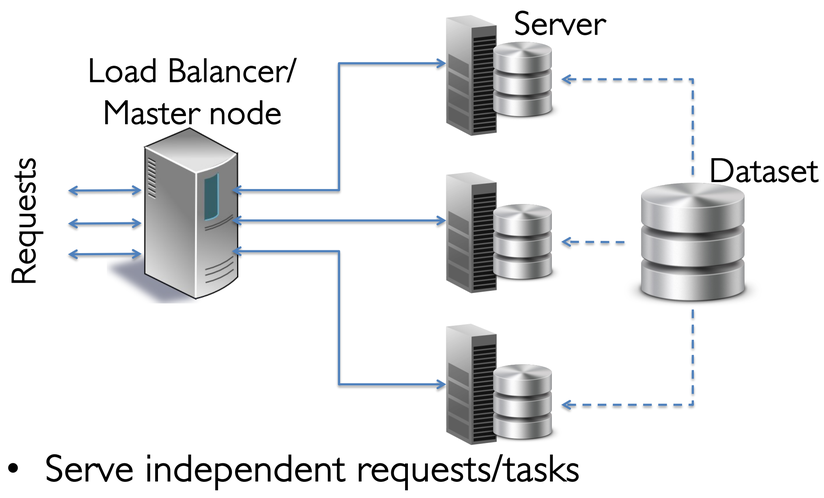
- Constant demand for more servers
- Increasing costs of HW, space & power

Modern Servers are Scale-Up

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



Cloud Applications are Scale-out

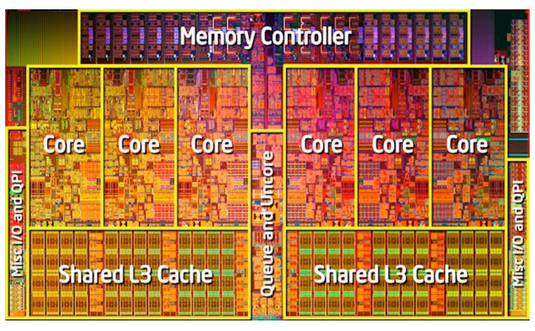


- Serve independent requests/tasks
- Operate on huge dataset split into shards
- Communicate infrequently

How efficient are scale-up servers for scale-out applications?

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 a cloud-native CPU design

Processors for Scale-Out Workloads

- Many non-aggressive OoO cores
- Smaller & faster LLC
- Fast instruction-supply path
- Correctly provisioned off-chip B/W
- Special accelerators for a workload
- Leverage repetitions to simplify HW

There is plenty of room for improving processors running scale-out workloads.

CloudSuite 4.0

- Data Analytics — Machine learning — MAHOUT
- Graph Analytics — GraphX
- In-Memory Analytics — Recommendation System — movielens, Spark
- Web Search — Apache Solr, nutch
- Media Streaming — Nginx, HTTP Server
- Web Serving — Nginx, PHP server
- Data Caching — Memcached
- Data Serving — Cassandra NoSQL, yahoo!, Cassandra

What is New in CloudSuite 4.0?

- Updated software stacks
- Supporting the ARM ISA
- Updated datasets and documentations
- Proper tuning guidelines
- Tested on real modern x86 and ARM servers

arm

Updated software stacks and multi-architecture support in the latest version

Research Directions

- Software stack maturity in x86 and ARM ISAs
- Cloud-native CPU design for scale-out server workloads
- Power and energy consumption characteristics of scale-out server workloads
- Properly utilizing a socket with 100 cores and beyond (e.g., AmpereOne)
- Novel core frontend design leveraging instruction commonality among the cores

Interesting opportunities for research on scale-out server workloads

COMPAS Computer Architecture Stony Brook

Stony Brook University