Server Benchmarking with CloudSuite 4.0

Ali Ansari†, Shanqing Lin†, Miguel Peón-Quirós⁺, Arash Pourhabibi†, Mark Sutherland†,

Babak Falsafi†⁺, Michael Ferdman‡

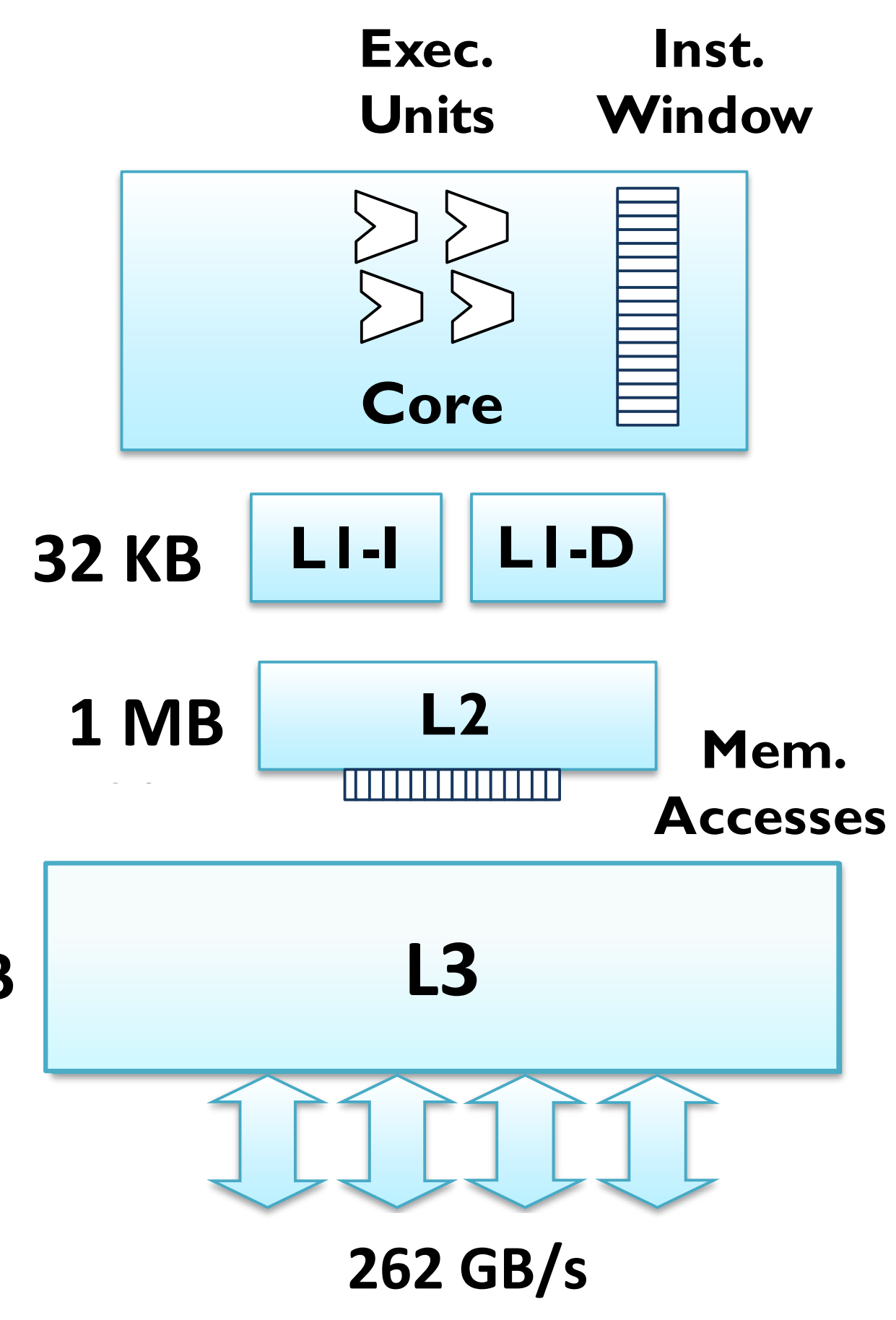
†PARSA, EPFL ⁺ EcoCloud, EPFL ‡Stony Brook University

Cloud Server Efficiency

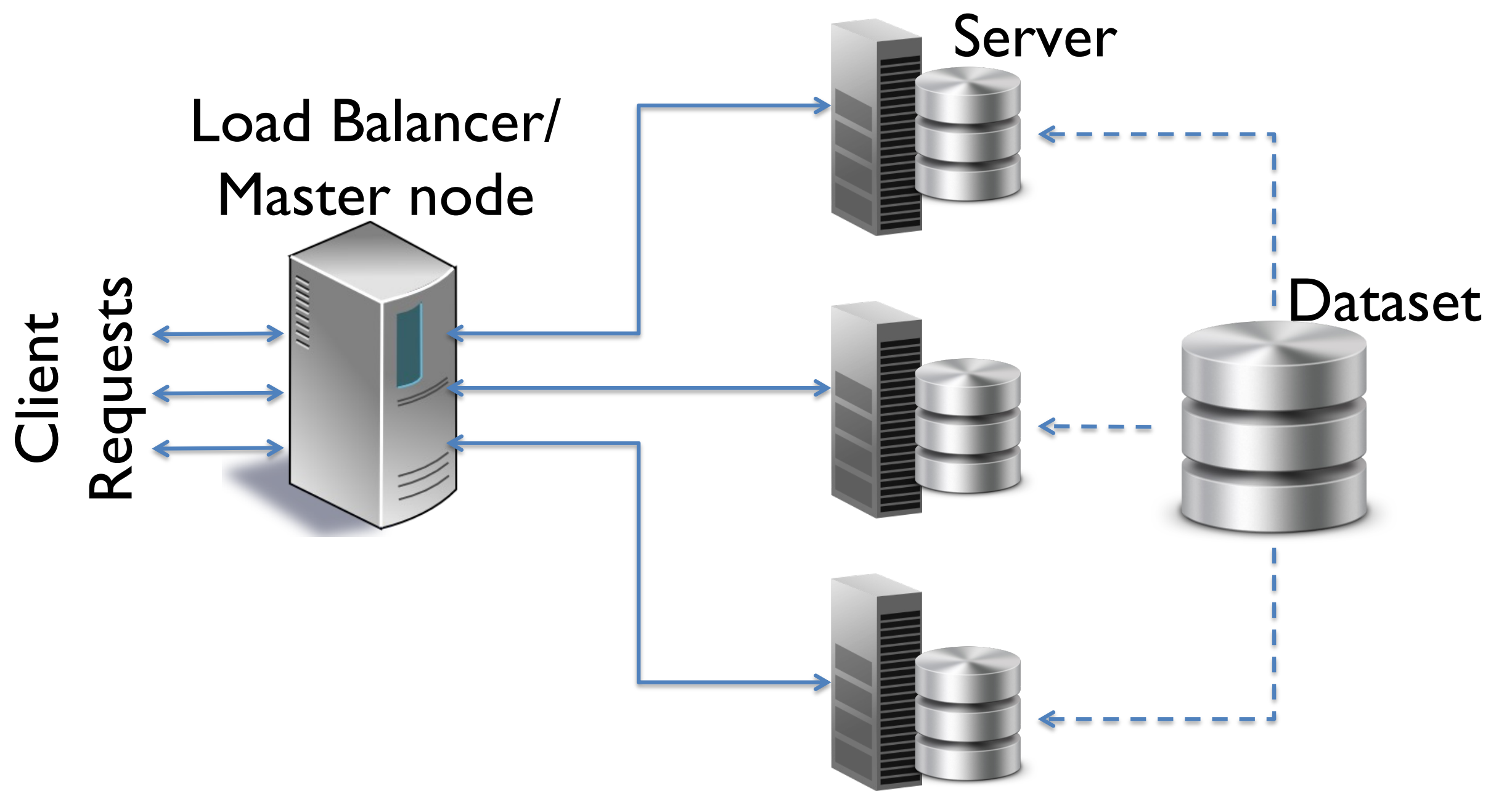
- Constant demand for more servers
- Increasing costs of HW, space & power

Modern Servers are Scale-Up

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



Cloud Applications are Scale-out

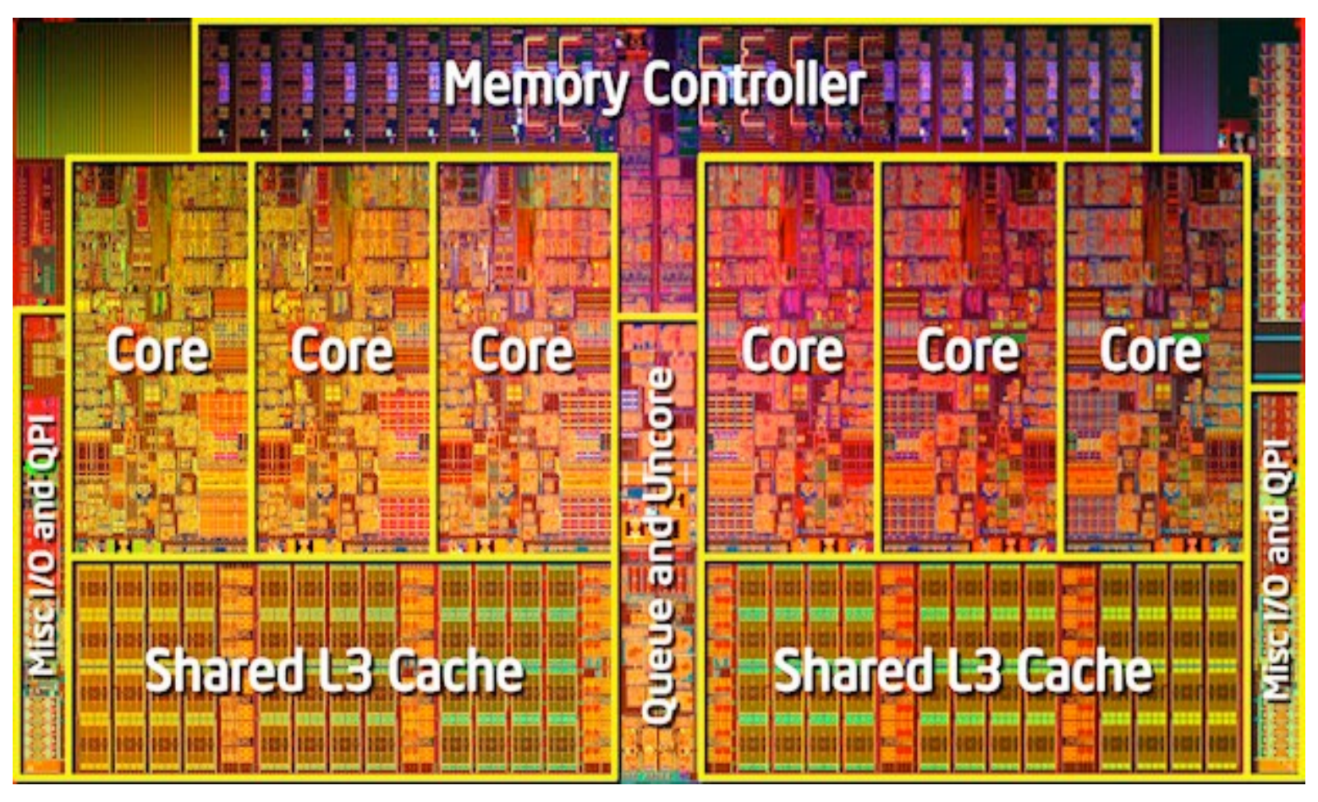


- Serve independent requests/tasks
- Operate on huge dataset split into shards
- Communicate infrequently

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 specialized scale-out processors

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

Graph Analytics
GraphX

In-Memory Analytics
Recommendation System

Web Search
Apache Solr & Nutch

Media Streaming
Nginx, HTTP Server

Web Serving
Nginx, PHP server

Data Caching
Memcached

Data Serving
Cassandra NoSQL

Facebook oss-performance

Django Workload

What is New in CloudSuite 4.0?

- Facebook oss-performance with 7 internal benchmarks based on HHVM
- Django workload by Intel and Instagram that serves large-scale mobile clients
- Images for ARM and RISC-V architectures using Docker multi-arch builds

New benchmarks and multi-architecture support in the latest version

Research Directions

- Identifying mismatches between workloads' characteristics and processors' implementation to propose workload-specific processor design
- Deployment of ARM and RISC-V as emerging server architectures
- Power and energy consumption characteristics of scale-out server workloads
- Industry's response to scale-out workloads' requirements over the past decade

Interesting opportunities for research on scale-out server workloads