MixApart: Decoupled Analytics for Shared Storage Systems

Madalin Mihailescu, *Gokul Soundararajan*, Cristiana Amza
University of Toronto, *NetApp*

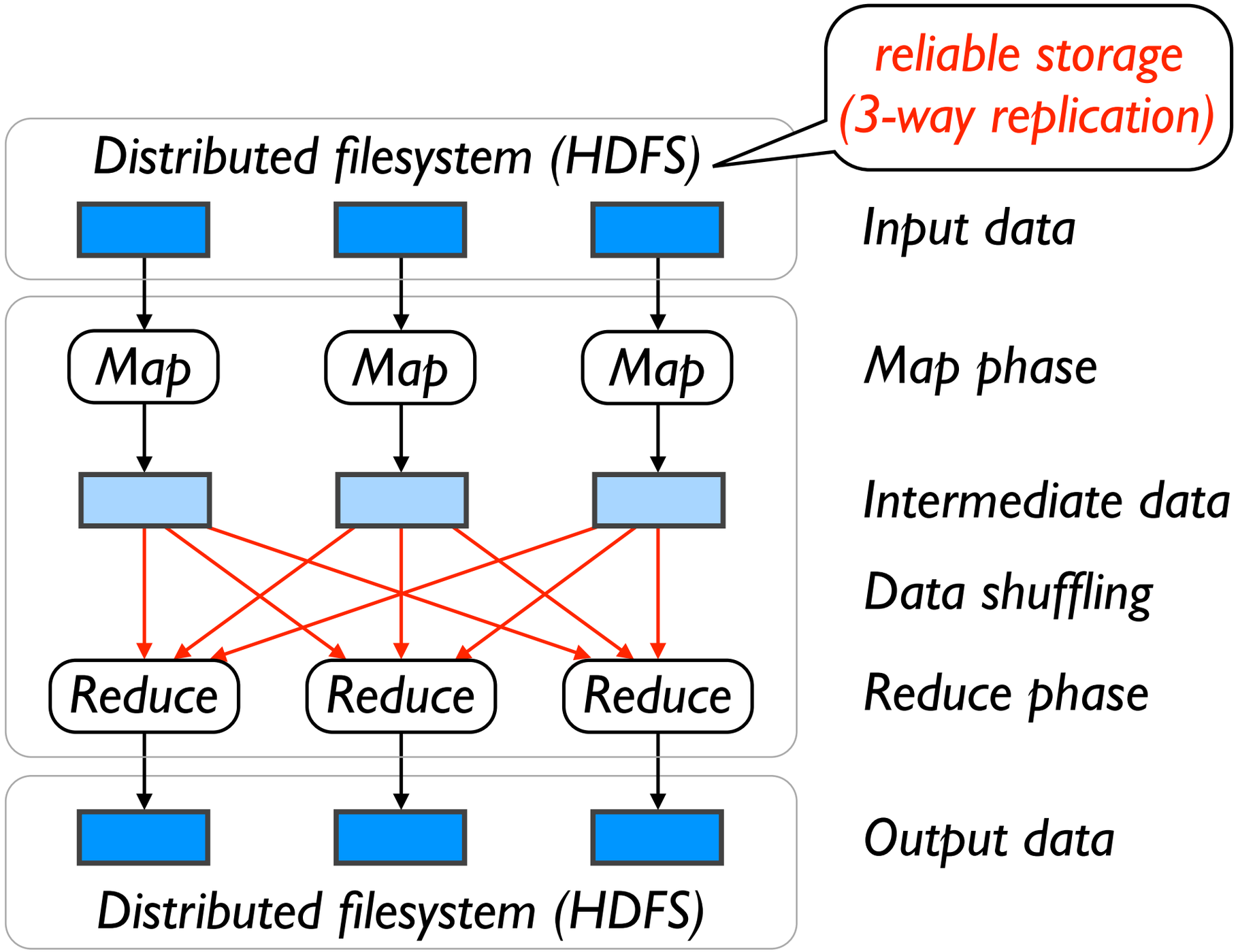
Data Analytics with Hadoop

Easy parallel processing of large datasets

Key concepts

- Leverage the *MapReduce* paradigm
- Data stored in *commodity distributed filesystem* (HDFS)
- Obtain performance through
  - Job/data *partitioning*
  - *Co-locating* processing with data

Flow in Hadoop



Analytics File Systems

Built for *low-value* data – e.g., web logs

- *Minimal* feature set for *data protection, storage efficiency*

Workload *characteristics*

- *Large* files, *appends*, *relaxed* consistency

High aggregate throughput

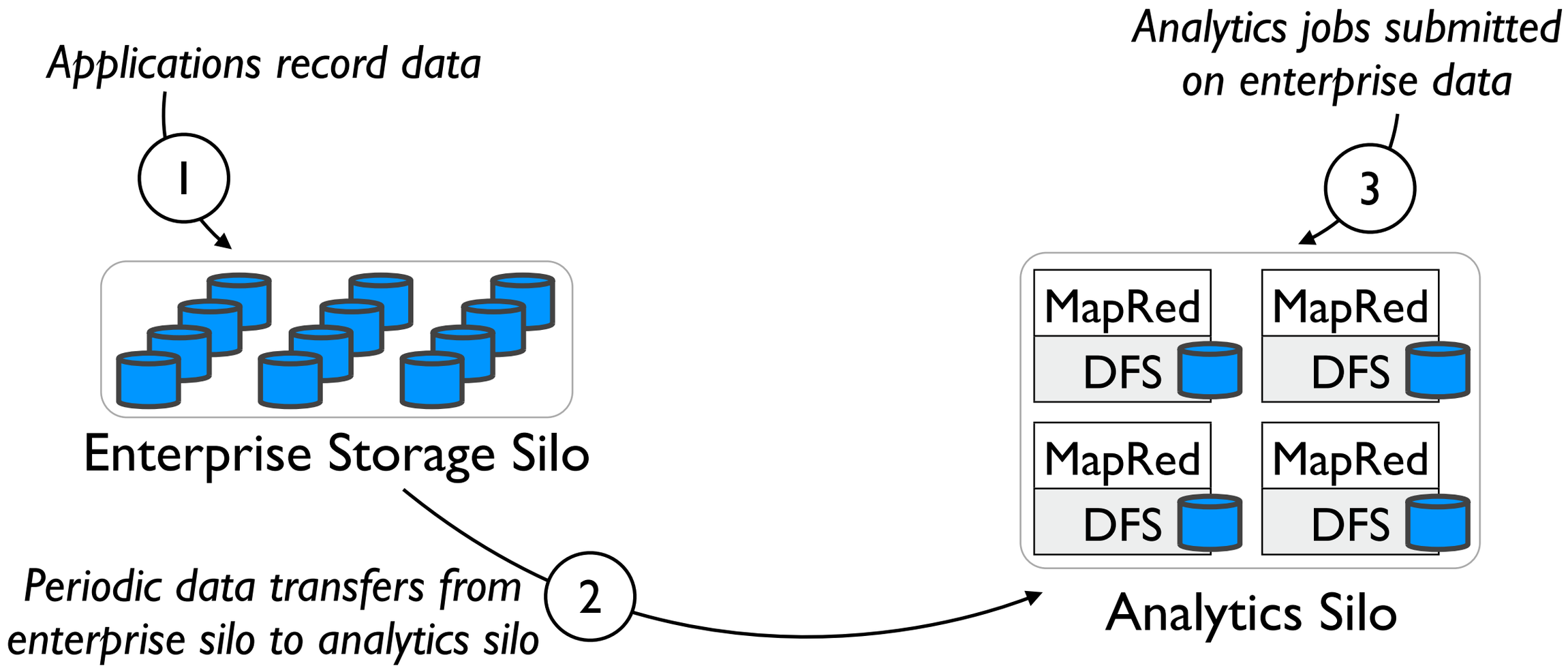
Analytics on Enterprise Data

Enterprise Storage Systems

- Manage *high-value* data – e.g., corporate e-mails
  - Rich *data management* capabilities
- Workload *characteristics*
  - *Small* files, *overwrites*, *strong* consistency

Disparate design points lead to multiple custom-built storage silos!

Enterprise-level Analytics

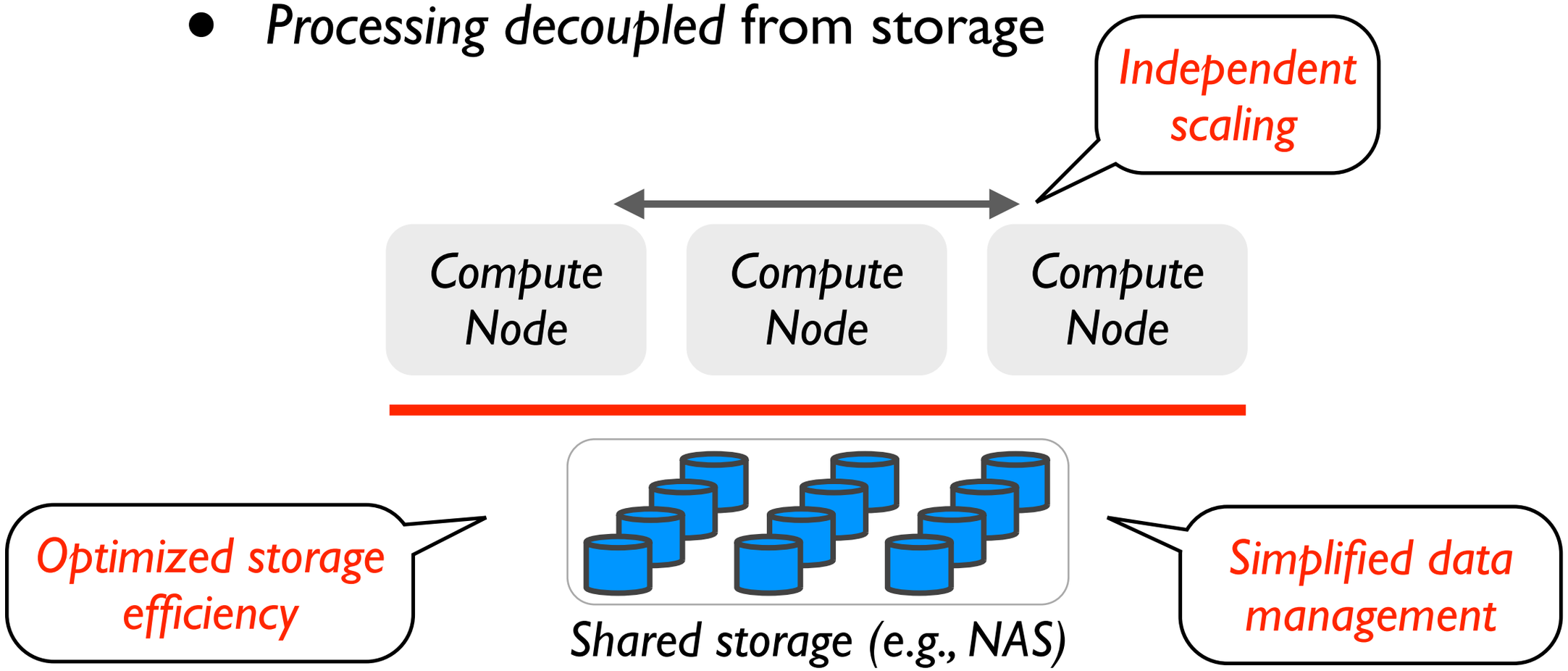


Multiple silos ➡ excessive hardware costs & high-maintenance cross-silo data management

MixApart

Integrate analytics with *existing* enterprise storage

- *Data* stored in *enterprise storage*
- *Processing decoupled* from storage



But Will It Scale?

MapReduce *workload analysis*

- Extrapolate from recent studies*
- Production traces from Facebook, Bing, Yahoo!

---

* Berkeley papers – NSDI '12, EuroSys '12

Workload Analysis

High *data reuse* across jobs

- *11%, 7%, 6%* of jobs at Facebook, Bing, Yahoo! read *singly accessed input*
- *60%* estimated optimal *reuse rates*
- *Iterative processing* (e.g., machine learning, job pipelines) benefits reuse rates

➡ *Large inexpensive disk-based cache for performance*

Workload Analysis

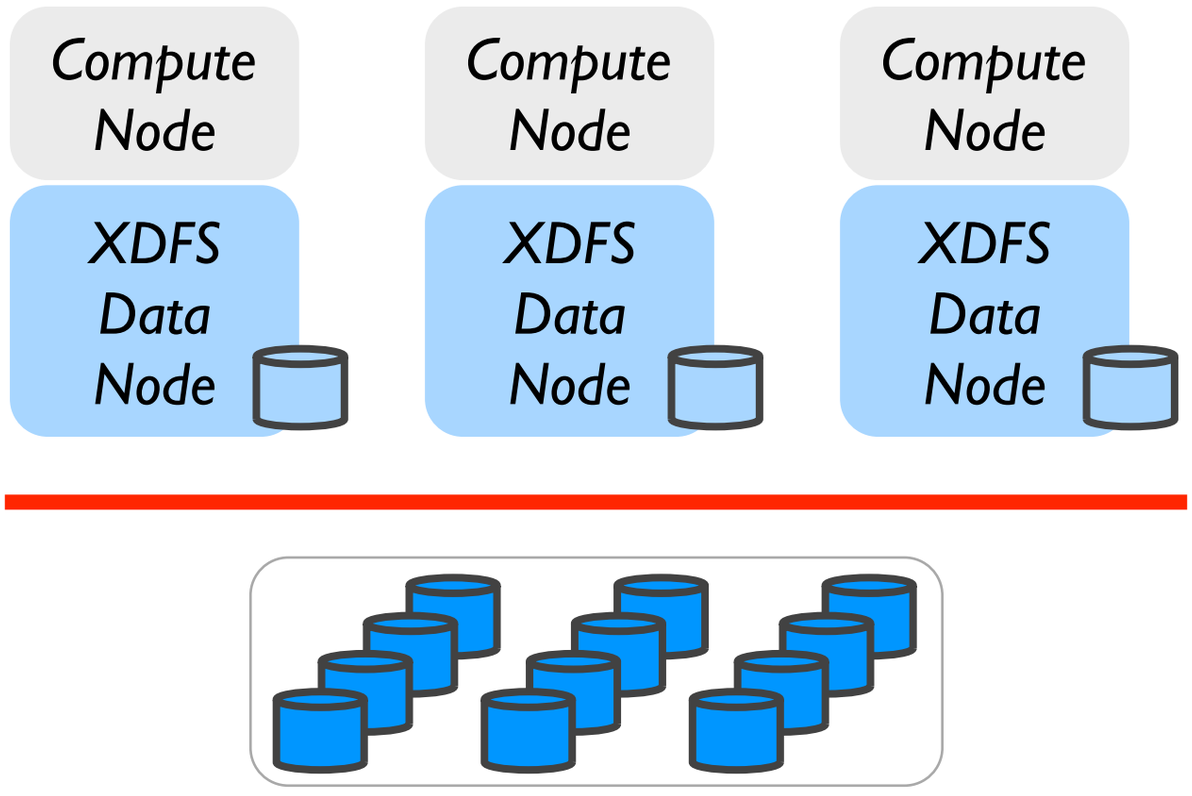
CPU-intensive Input Phases

- Compute, compression, serialization, setup, etc.
- Median *map task durations*: *19s* at Facebook, *26s* at Yahoo!
- Low task *I/O rates* – e.g., *25Mbps* for *64MB* of input
  - 1Gbps storage bandwidth sustains 40 tasks

➡ *Well-provisioned enterprise storage can sustain low to moderate compute clusters*

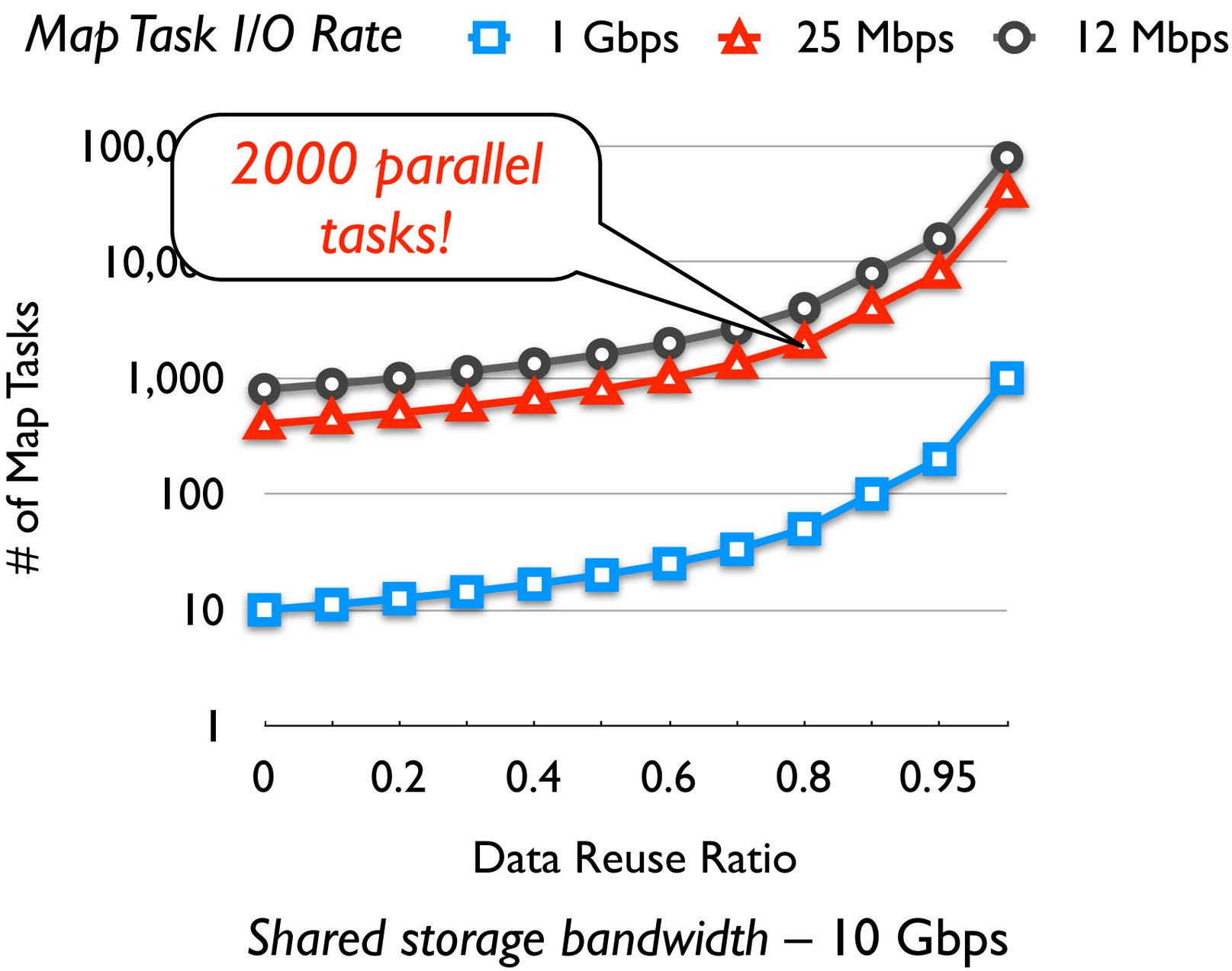
MixApart Architecture

XDFS disk cache tier (*stateless*) for *performance/scalability*



Shared storage (e.g., NAS) for *reliability/data management/storage efficiency*

Scalability Estimates



Shared storage bandwidth – 10 Gbps

Workload Analysis

Predictable I/O Demands

- *Homogeneous* jobs (6 classes at Facebook)
- *Homogeneous* per-job tasks
  - *Task I/O rates* derived at *job submission*

➡ *Just-in-time parallel data transfers from shared storage into cache*

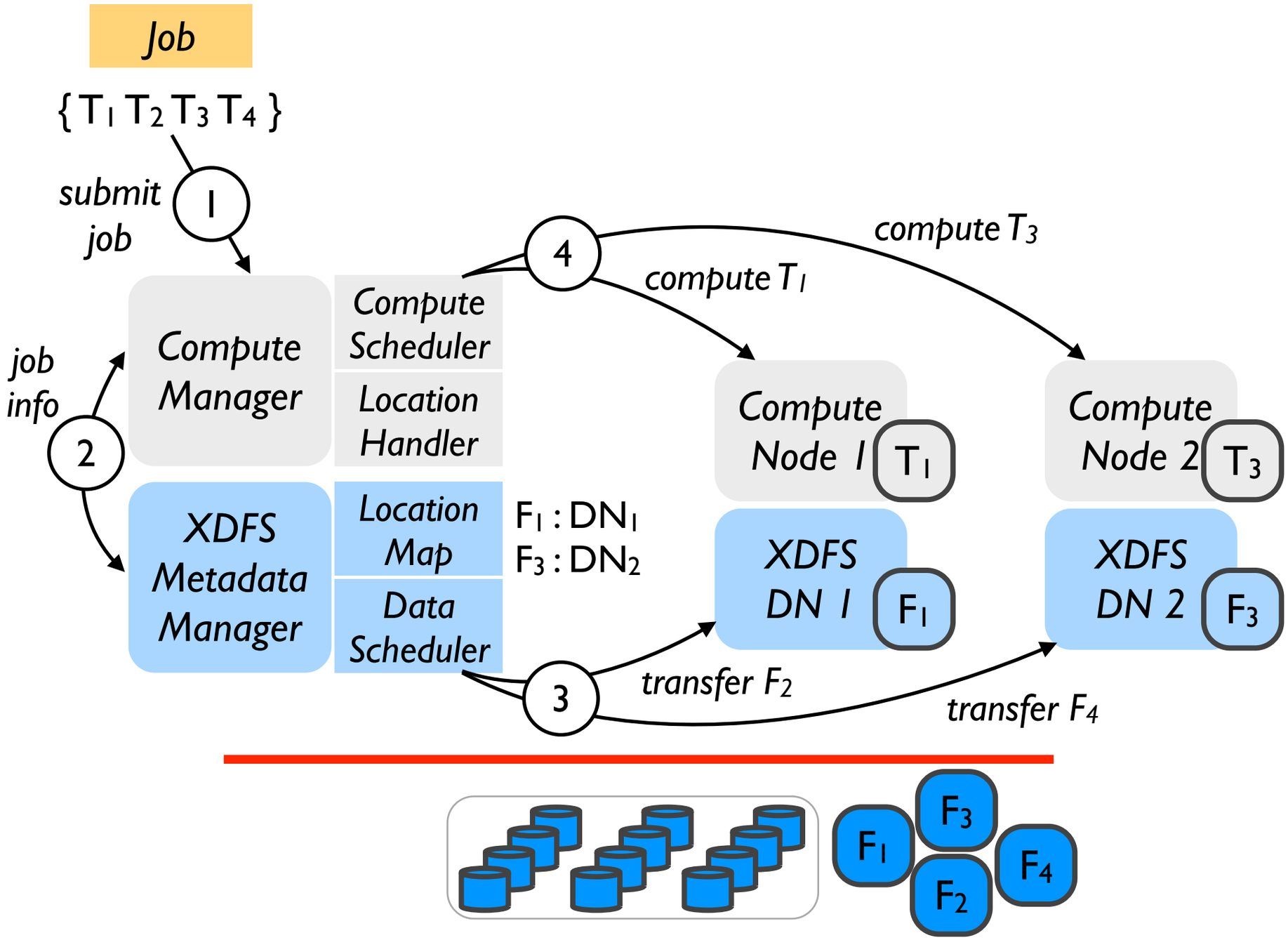
Coordination

Synchronized compute and storage transfers

Key components

- *Data-aware Compute Scheduler*
  - Schedules tasks using *policy* (e.g., FIFO) and *cache contents*
- *Compute-aware Data Scheduler*
  - Schedules *in-parallel transfers* from storage to cache using *available bandwidth* and *I/O rates*

Coordination



Evaluation

MixApart *prototype* based on Hadoop

Testbed

- *100-core cluster* on Amazon EC2
- *Local EC2 instance storage* for XDFS cache/HDFS
- *NFS server*
  - 4 EBS volumes in RAID-0 setting
  - *1Gbps* bandwidth

Evaluation

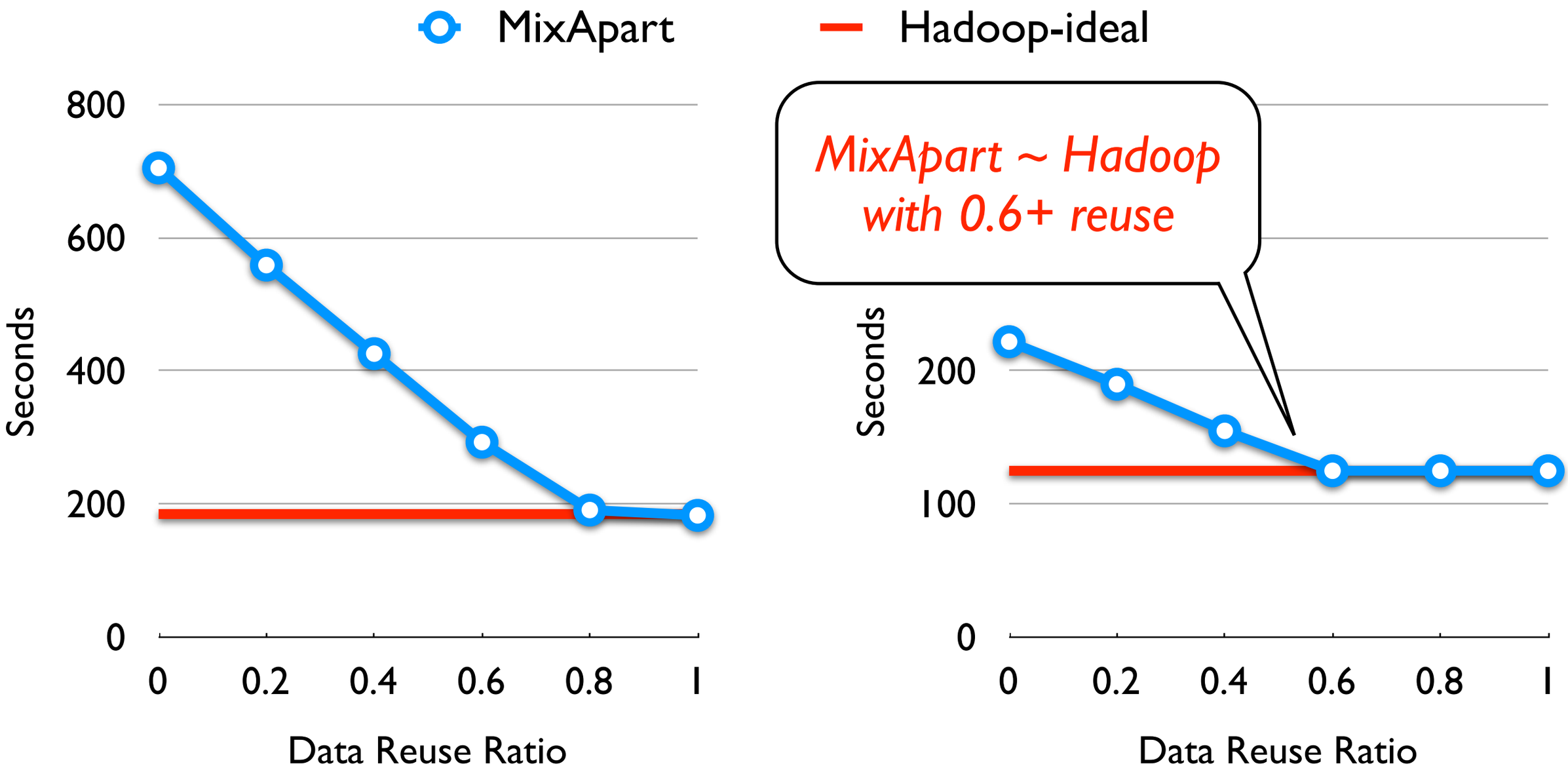
Goal – comparable performance to *ideal Hadoop with no ingest*

Dataset – 12 days of *Wikipedia statistics*

Workload – job to *aggregate page views* for regex

- *I/O intensive – uncompressed* input (*I/O rate* 50 Mbps)
- *CPU intensive – compressed* input (*I/O rate* 20 Mbps)

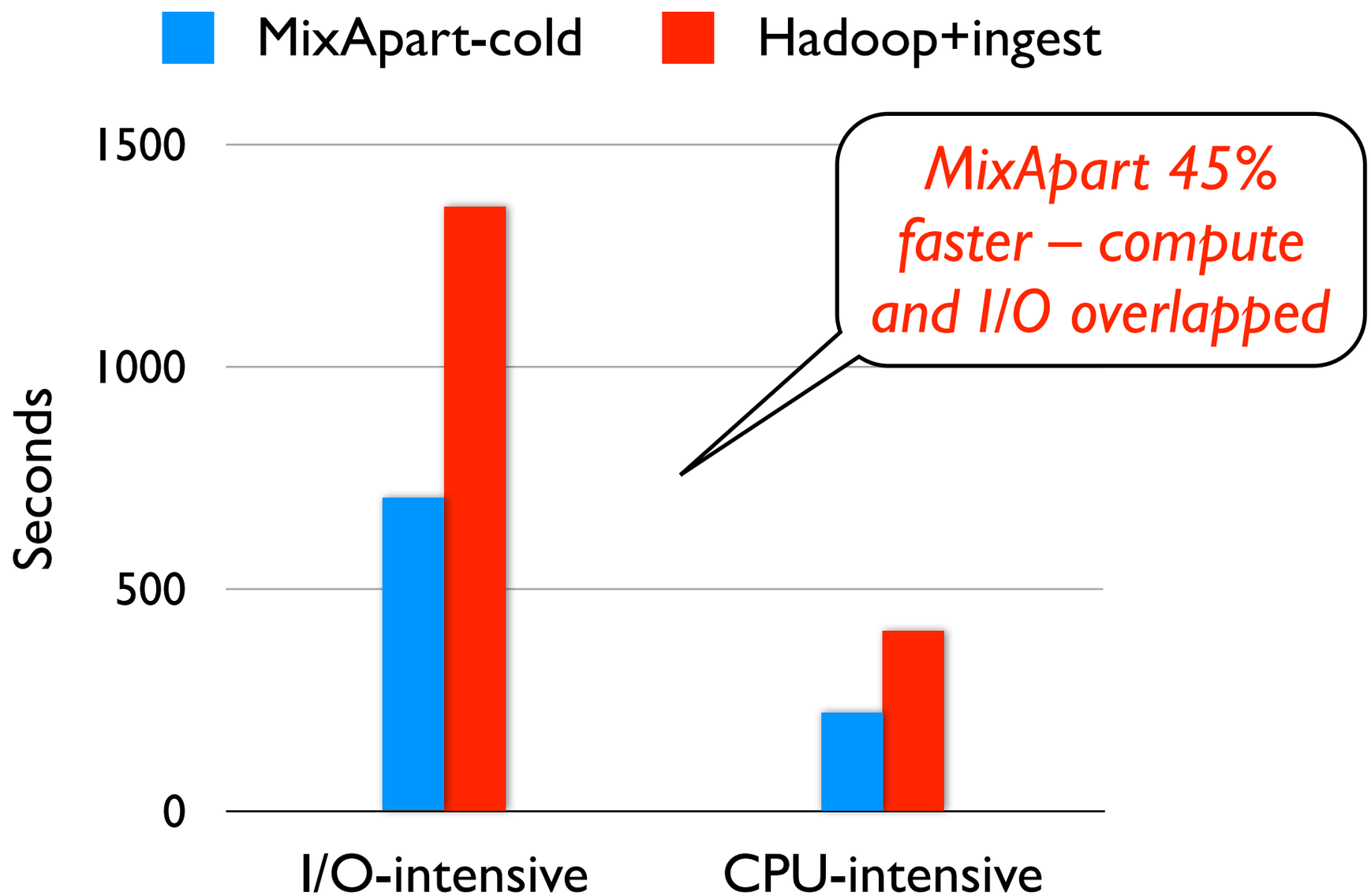
Caching and Scheduling



a) I/O intensive

b) CPU intensive

Impact of Data Ingest



MixApart: Decoupled Analytics for *Shared* Storage Systems