MESOS: SHARING THE CLUSTER

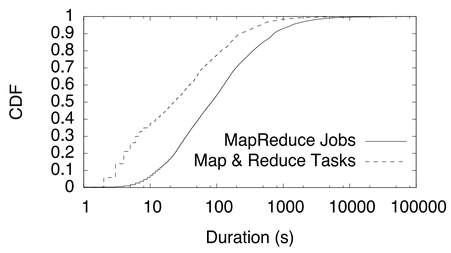
Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony Joseph, Randy Katz, Scott Shenker, Ion Stoica (UC Berkeley)

MOTIVATION

- Rapid innovation in cluster computing applications: MapReduce, Spark, Webapps, Distributed FSs & DBs, …
- Organizations currently statically partition clusters
- Instead, enable *multiple* frameworks to share same cluster

PROBLEM

- Traditional coarse-grained cluster schedulers unsuitable
  - Applications consist of fine-grained tasks and need to scale up and down dynamically
  - Data locality crucial for efficiency
  - Users run queries interactively
  - Apps developed by multiple groups and rapidly evolving



Example: Hadoop job and task durations at Facebook

SOLUTION

- Mesos is a "*cluster operating system*" over which diverse parallel applications can run
  - Provides minimal API for efficient resource sharing, then leaves maximum control to applications

CONTRIBUTIONS

- *Fine-grained sharing*: applications divide work into tasks that are multiplexed in time & space
- *Distributed scheduling model* (resource offers) to support varied application needs
  - Mesos chooses *how many* resources to offer each FW
  - Frameworks chooses *which* resources to accept
  - Deployments at:

IMPLEMENTATION

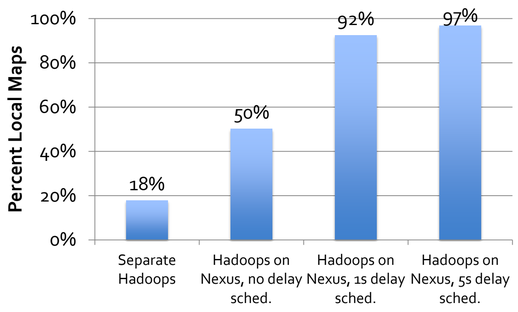
- 20,000 lines of C++
- APIs in C, C++, Java, Python
- Isolation via Linux containers
- Fault tolerance via ZooKeeper

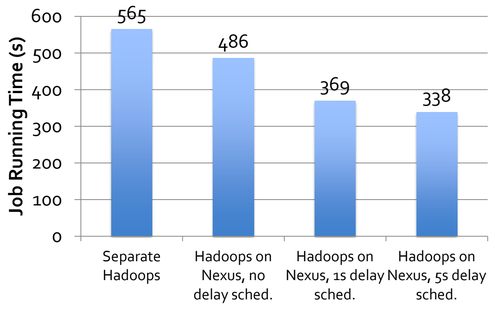
APPLICATIONS

- Hadoop port: 900 line patch
- MPI port: 160 line wrapper
- Torque port
- Spark: 1300 lines
- Elastic web apps

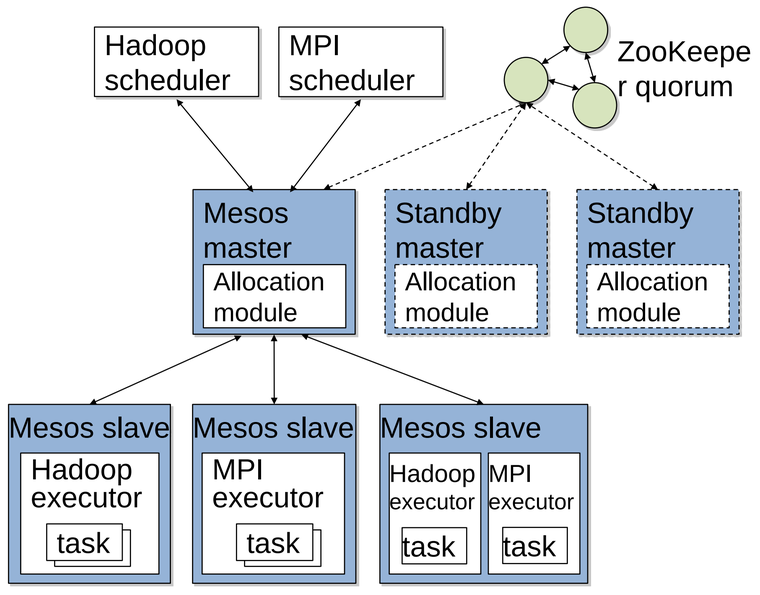
DATA LOCALITY RESULTS

Data Locality Through Delay Scheduling



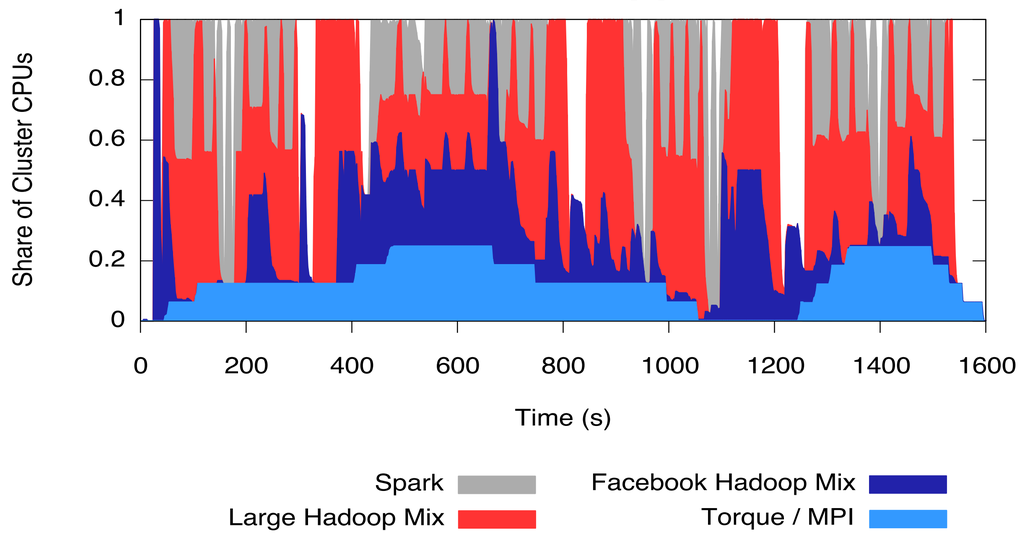


ARCHITECTURE

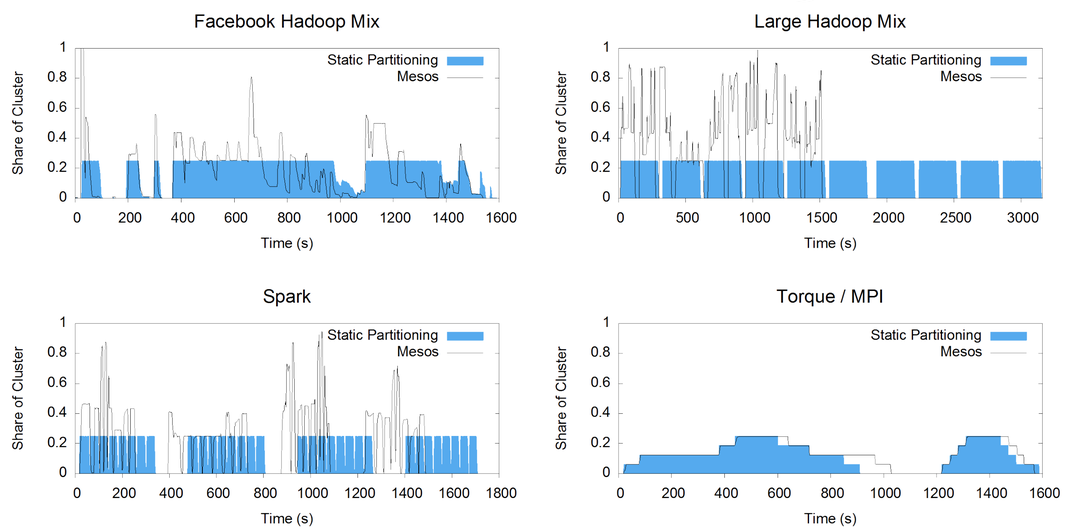


MACROBENCHMARK RESULTS

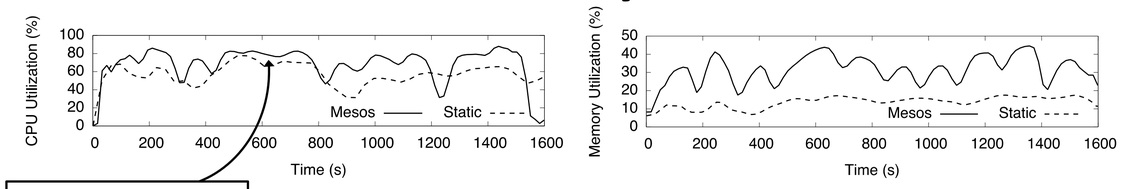
CPU Allocation to Applications



CPU Allocation: Static Partitioning vs. Mesos



CPU and Memory Utilization



~10% Higher Utilization (ave)!

Job Runtimes

Framework	Sum of Exec Times with Static Partitioning (s)	Sum of Exec Times on Mesos (s)	Speedup
Facebook Hadoop Mix	7235	6319	**1.14**
Large Hadoop Mix	3143	1494	**2.10**
Spark	1684	1338	**1.26**
Torque / MPI	3210	3352	**0.96**

~2x speedup!

Intel Science & Technology Center for Cloud Computing