

# CS 744: MESOS

Shivaram Venkataraman

Fall 2021

# ADMINISTRIVIA

- Assignment 1: Due tonight!
- Assignment 2 out soon
- Project details
  - Create project groups
  - Bid for projects/Propose your own
  - Work on Introduction
  - Final report / poster presentation

## Applications

Machine Learning

SQL

Streaming

Graph

Computational Engines

Scalable Storage Systems

Resource Management



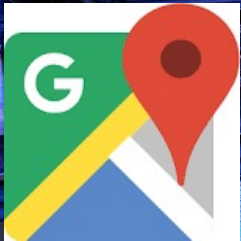
Datacenter Architecture



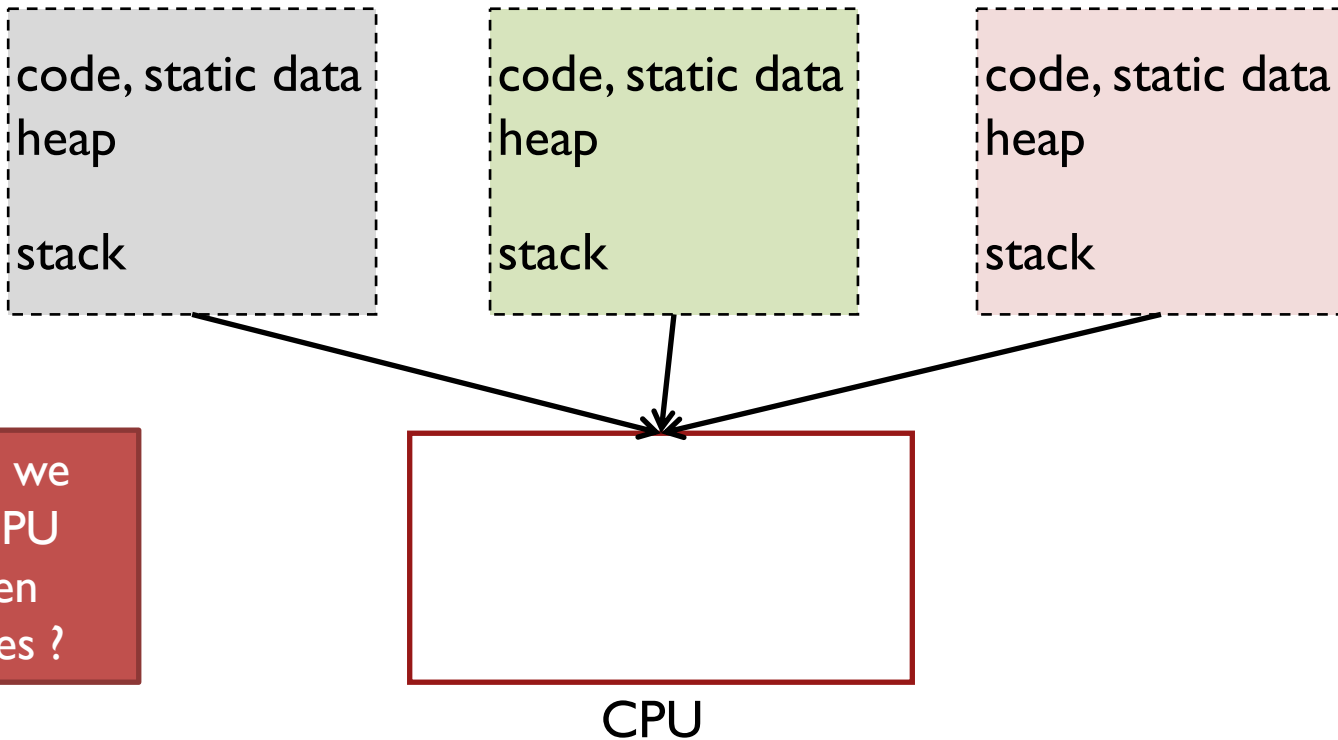
MapReduce

GFS

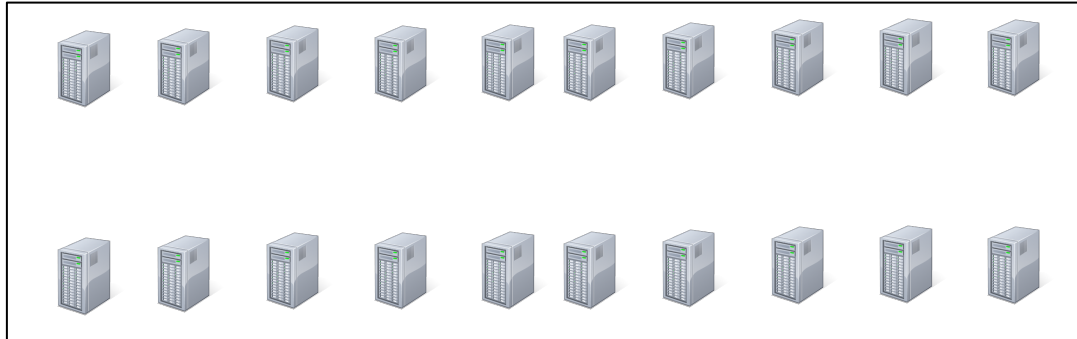
Spark



# BACKGROUND: OS SCHEDULING



# CLUSTER SCHEDULING



# TARGET ENVIRONMENT

Multiple MapReduce versions

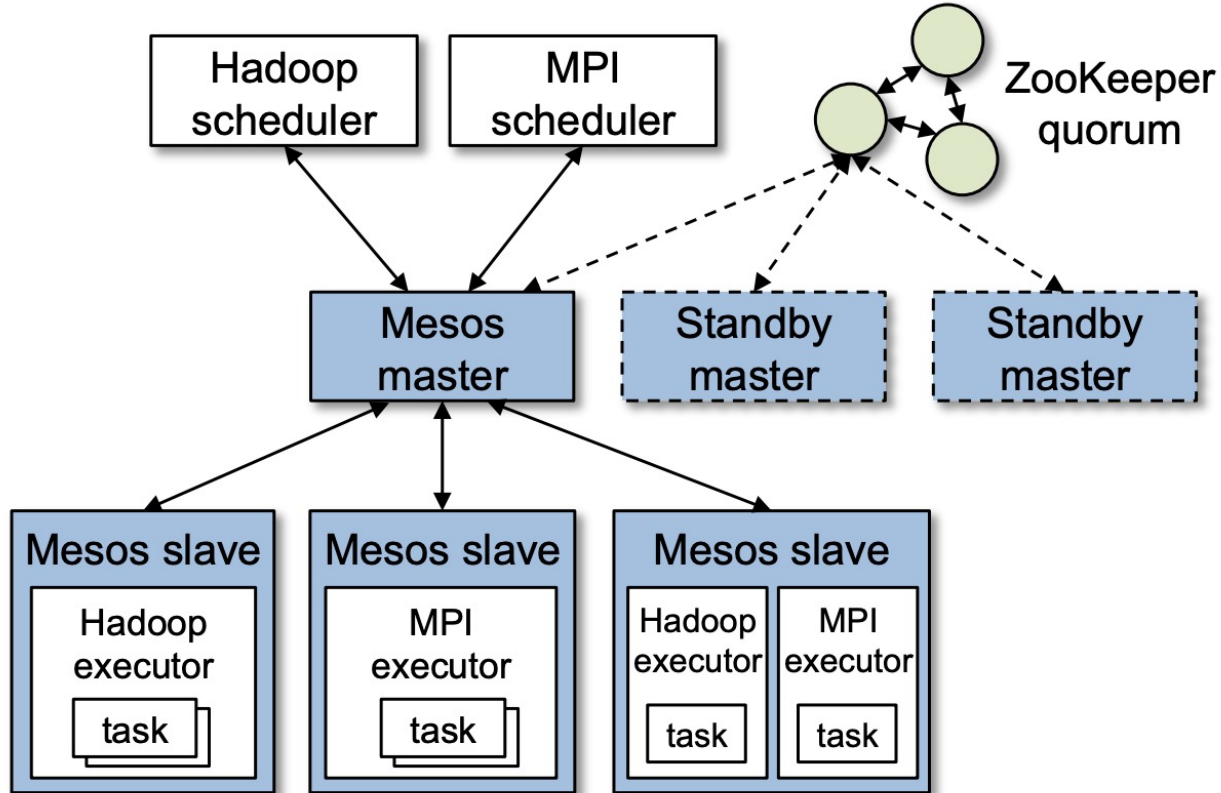
Mix of frameworks: MPI, Spark, MR

Data sharing across frameworks

Avoid per-framework clusters

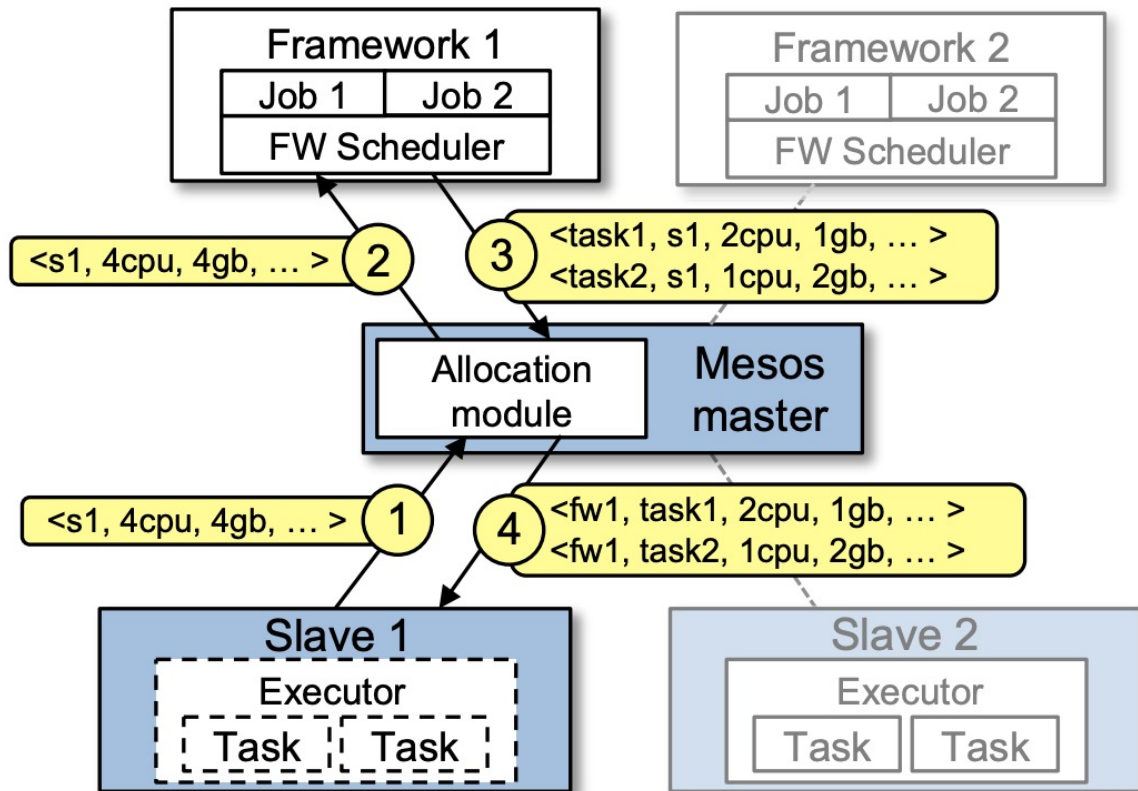


# DESIGN





# RESOURCE OFFERS



# CONSTRAINTS

## Examples of constraints

Data locality → soft constraint

GPU machines → hard constraint

## Constraints in Mesos:

Applications can reject offers

Optimization: Filters

# DESIGN DETAILS

## Allocation:

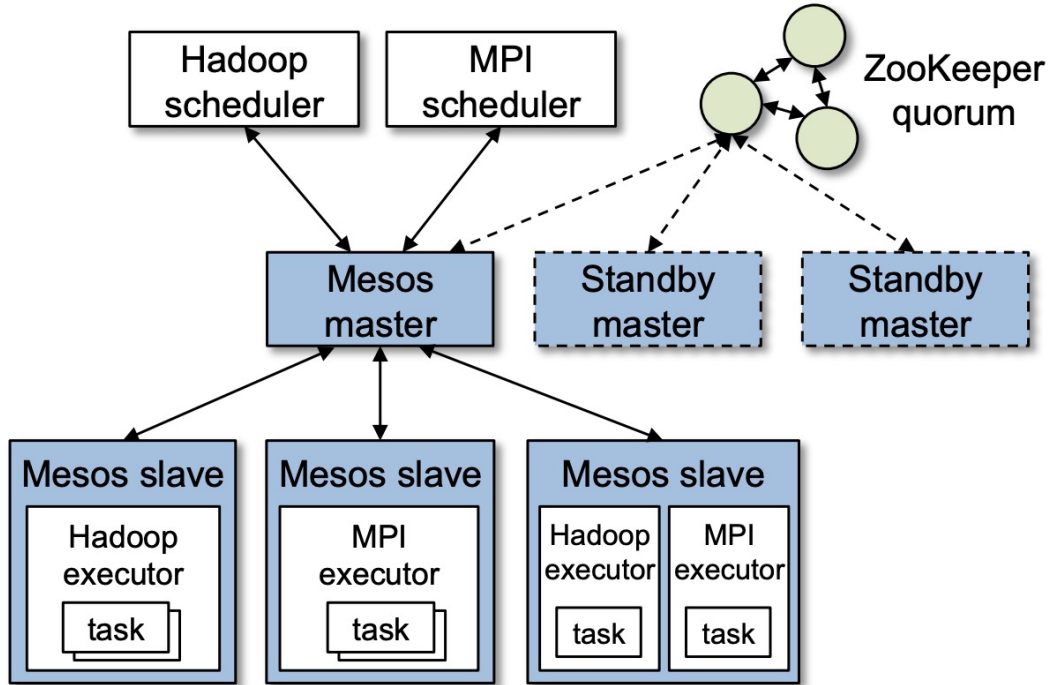
Tasks are short, allocate when they finish

Long tasks? Revocation beyond guaranteed

## Isolation

Containers (Docker)

# FAULT TOLERANCE



# HANDLING PLACEMENT PREFERENCES

What is the problem?

More frameworks have preferred nodes than available

Who gets the offers?

How do we do allocations?

Lottery scheduling – offers weighted by num allocations

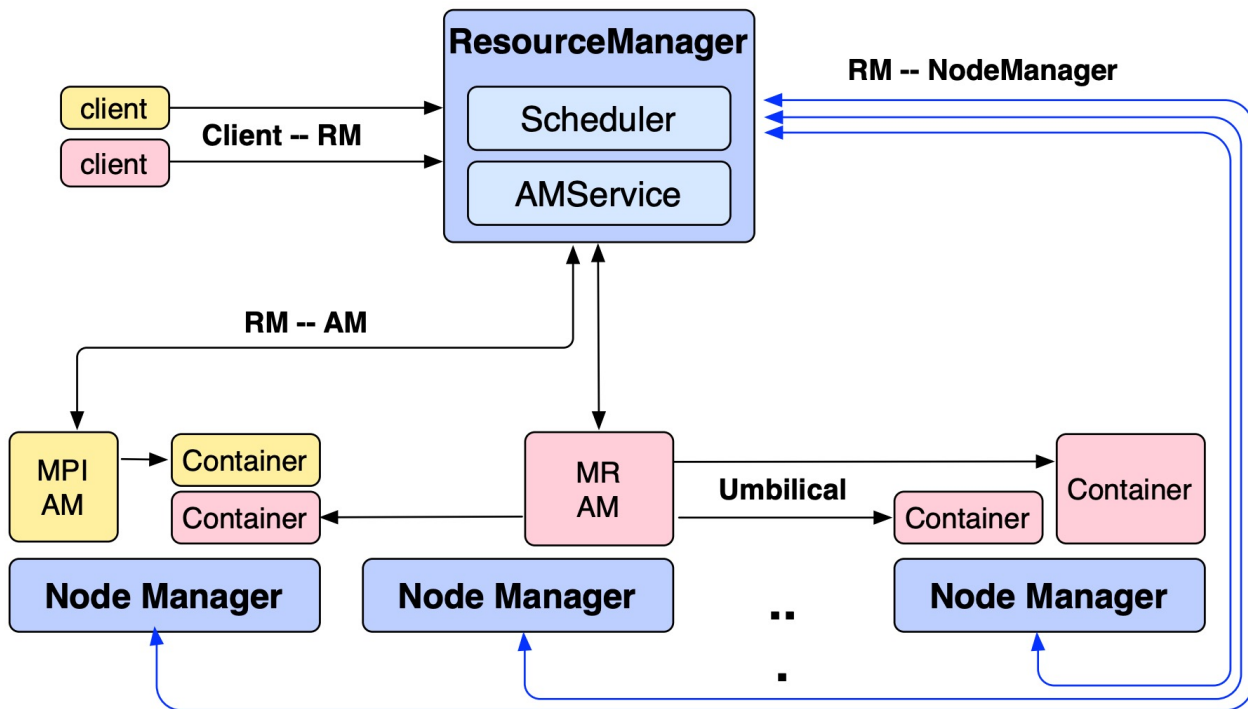
# CENTRALIZED VS DISTRIBUTED

Framework complexity

Fragmentation, Starvation

Inter-dependent framework

# COMPARISON: YARN



Per-job scheduler

AM asks for resource  
RM replies

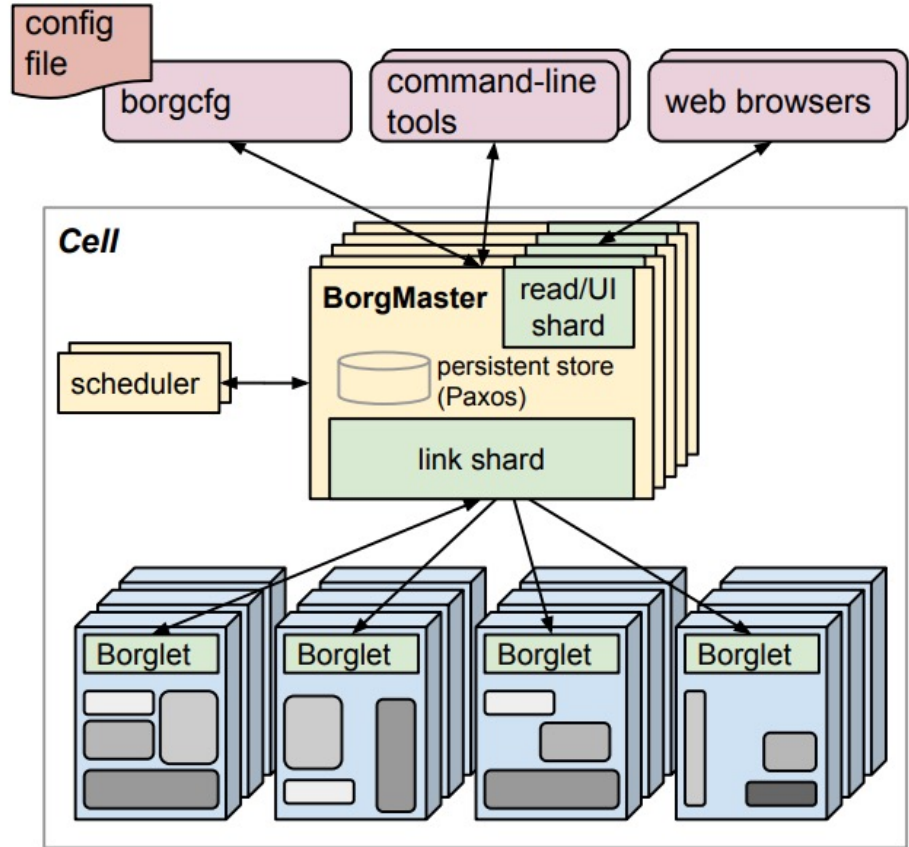


# COMPARISON: BORG

Single centralized scheduler

Requests mem, cpu in cfg  
Priority per user / service

Support for quotas / reservations



# SUMMARY

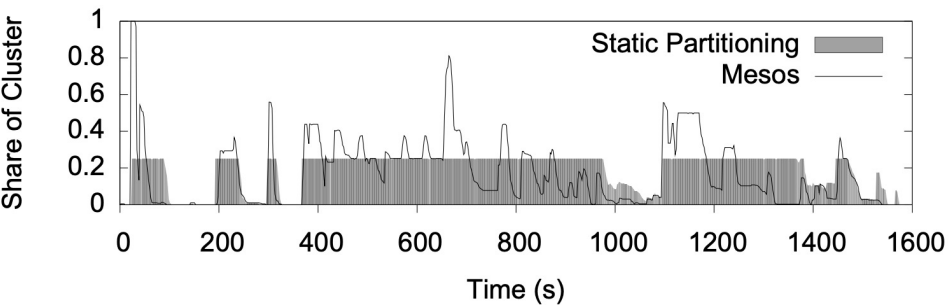
- Mesos: Scheduler to share cluster between Spark, MR, etc.
- Two-level scheduling with app-specific schedulers
- Provides scalable, decentralized scheduling
- Pluggable Policy ? Next class!

# DISCUSSION

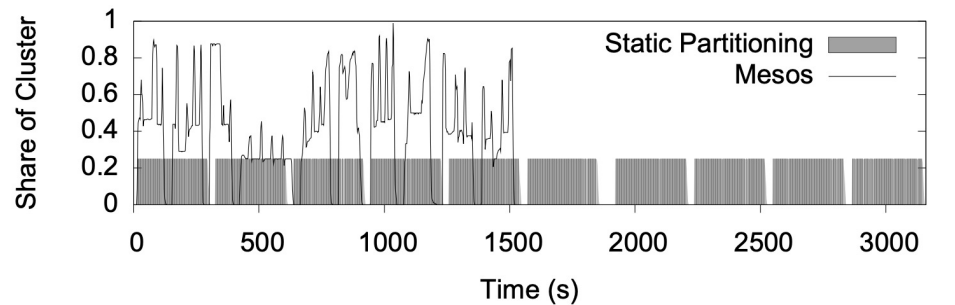
<https://forms.gle/FSkKVbu94nLA4g3v9>

What are some problems that could come up if we scale from 10 frameworks to 1000 frameworks in Mesos?

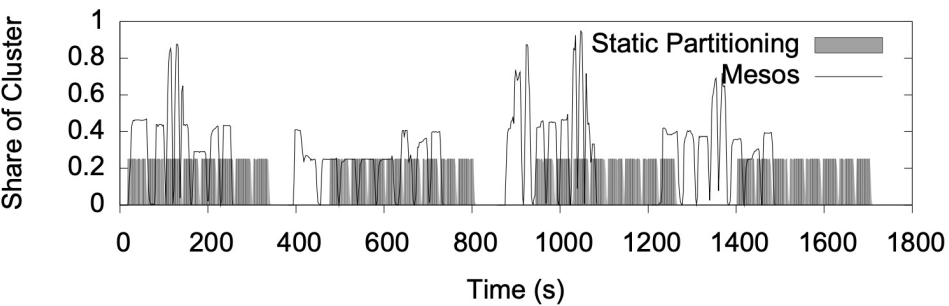
(a) Facebook Hadoop Mix



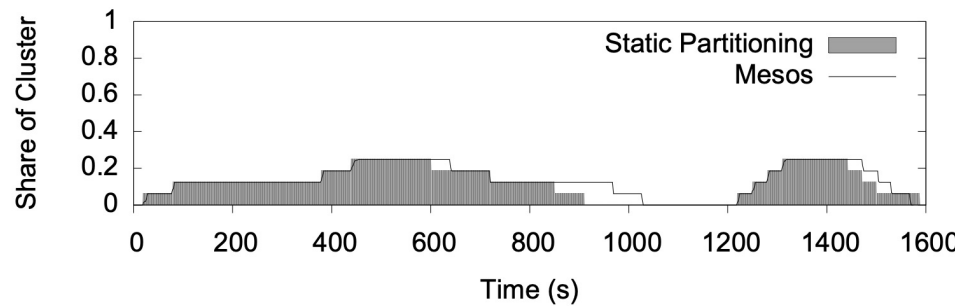
(b) Large Hadoop Mix



(c) Spark



(d) Torque / MPI



# NEXT STEPS

Next class: Scheduling Policy

Further reading

- <https://www.umbrant.com/2015/05/27/mesos-omega-borg-a-survey/>
- <https://queue.acm.org/detail.cfm?id=3173558>