Mesos and Borg and Kubernetes
Lecture 12, cs262a

Ion Stoica & Ali Ghodsi
UC Berkeley
October 7, 2020

Today's Papers

Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center,

Benjamin Hindman, Andy Konwinski, Matei Zaharia,

Ali Ghodsi, Anthony D. Joseph, Randy Katz, Scott Shenker, Ion Stoica, NSDI'11

https://people.eecs.berkeley.edu/~alig/papers/mesos.pdf

Large-scale cluster management at Google with Borg,

Abhishek Verma, Luis Pedrosa, Madhukar R. Korupolu, David Oppenheimer, Eric Tune, John Wilkes, EuroSys'15

http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43438.pdf

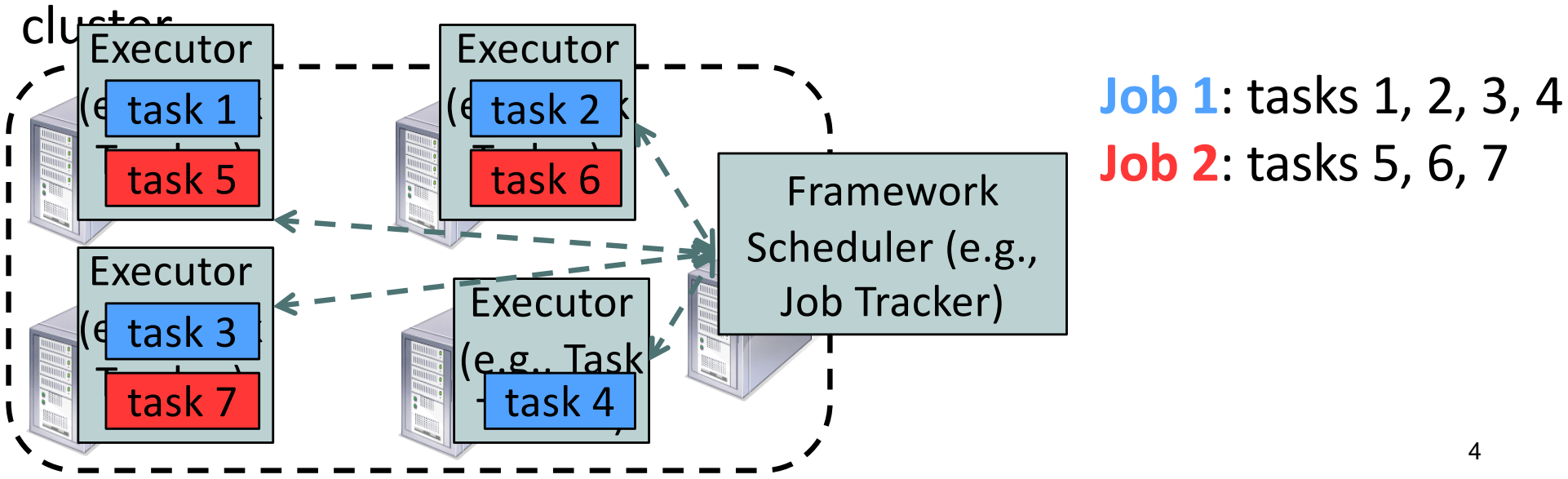
Motivation

- Rapid innovation in cloud computing

- Today
  - No single framework optimal for all applications
  - Each framework runs on its dedicated cluster or cluster partition

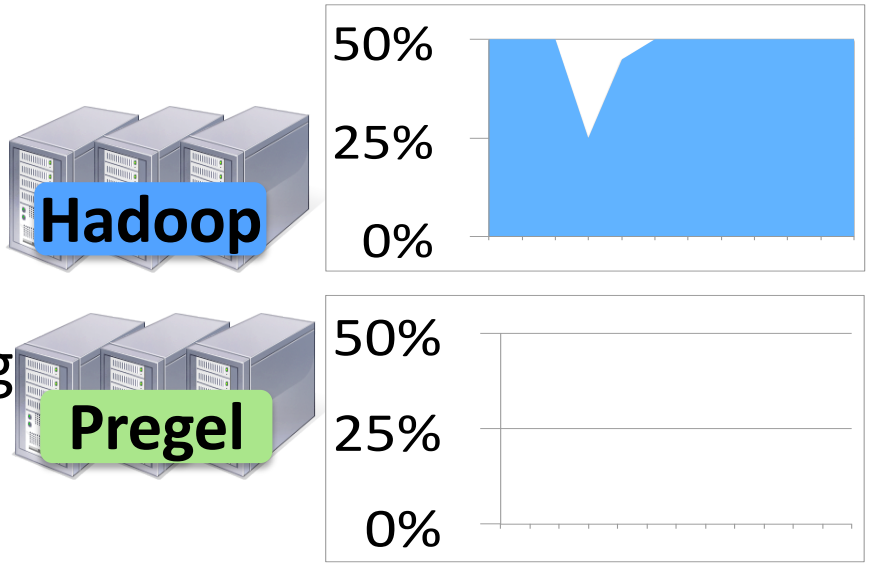
Computation Model: Frameworks

- A ***framework*** (e.g., Hadoop, MPI) manages one or more ***jobs*** in a computer cluster
- A ***job*** consists of one or more ***tasks***
- A ***task*** (e.g., map, reduce) is implemented by one or more processes running on a single machine



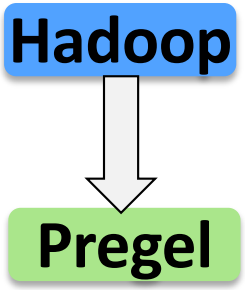
One Framework Per Cluster Challenges

- Inefficient resource usage
  - E.g., Hadoop cannot use available resources from Pregel's cluster
  - No opportunity for stat. multiplexing



- Hard to share data
  - Copy or access remotely, expensive

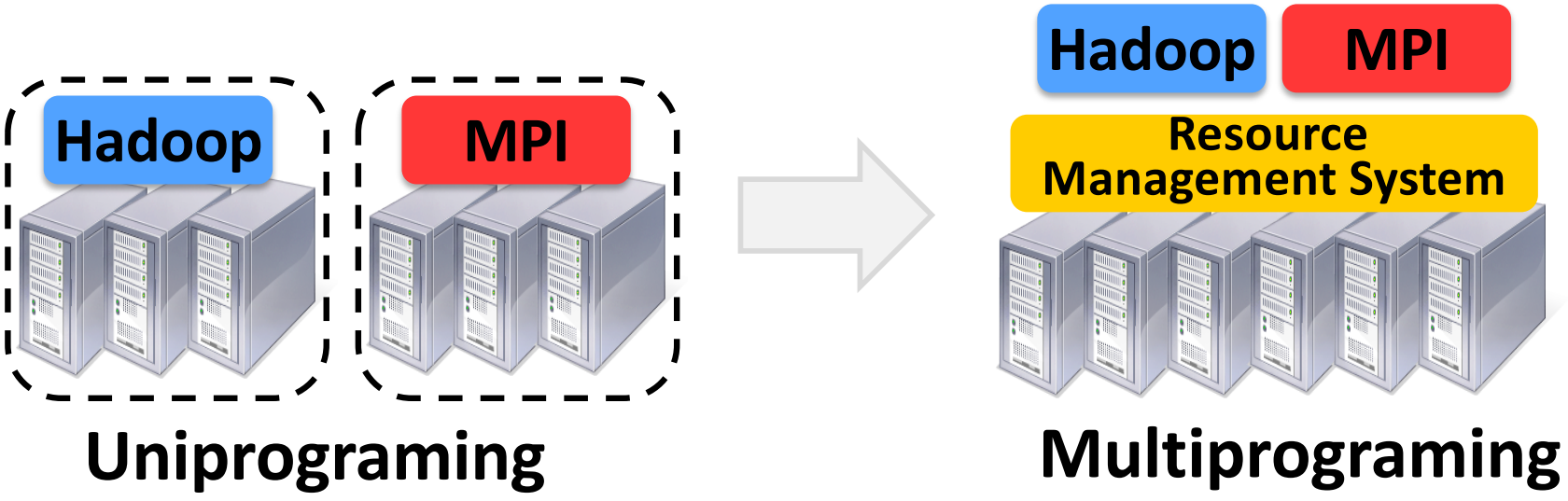
- Hard to cooperate
  - E.g., Not easy for Pregel to use graphs generated by Hadoop



Need to run multiple frameworks on same cluster

What do we want?

- Common resource sharing layer
  - Abstracts ("virtualizes") resources to frameworks
  - Enable diverse frameworks to share cluster
  - Make it easier to develop and deploy new frameworks (e.g., Spark)



Fine Grained Resource Sharing

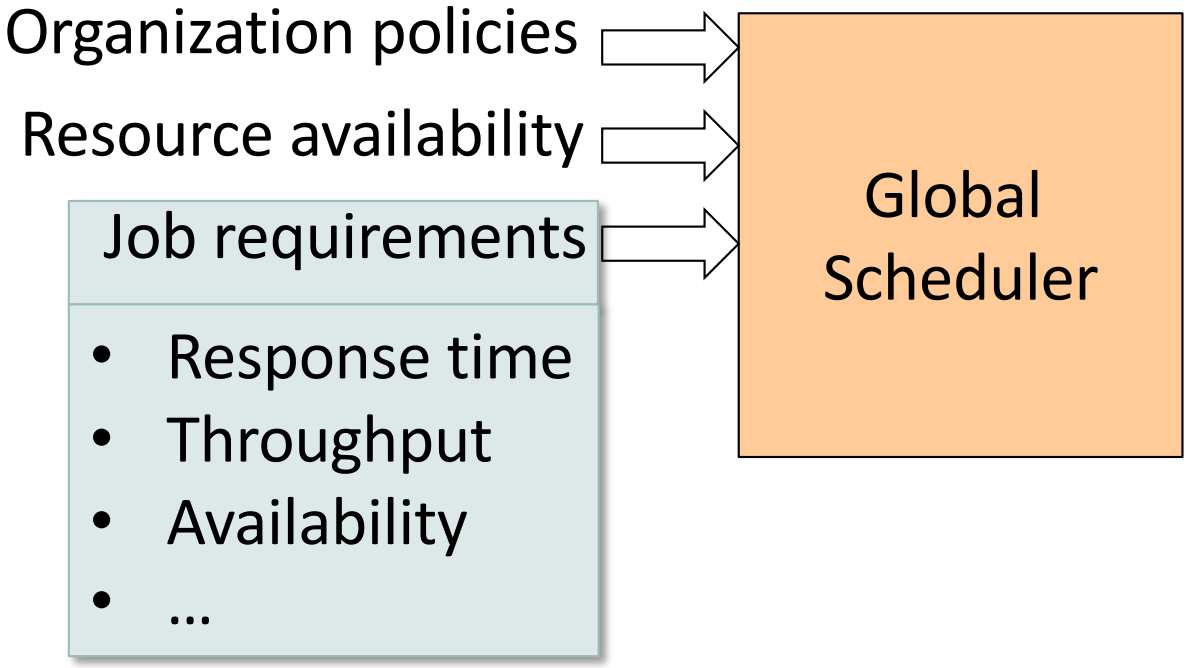
- Task granularity both in **time** & **space**
  - Multiplex node/time between tasks belonging to different jobs/frameworks
- Tasks typically short; median ~= 10 sec, minutes

- Why fine grained?
  - Improve data locality
  - Easier to handle node failures

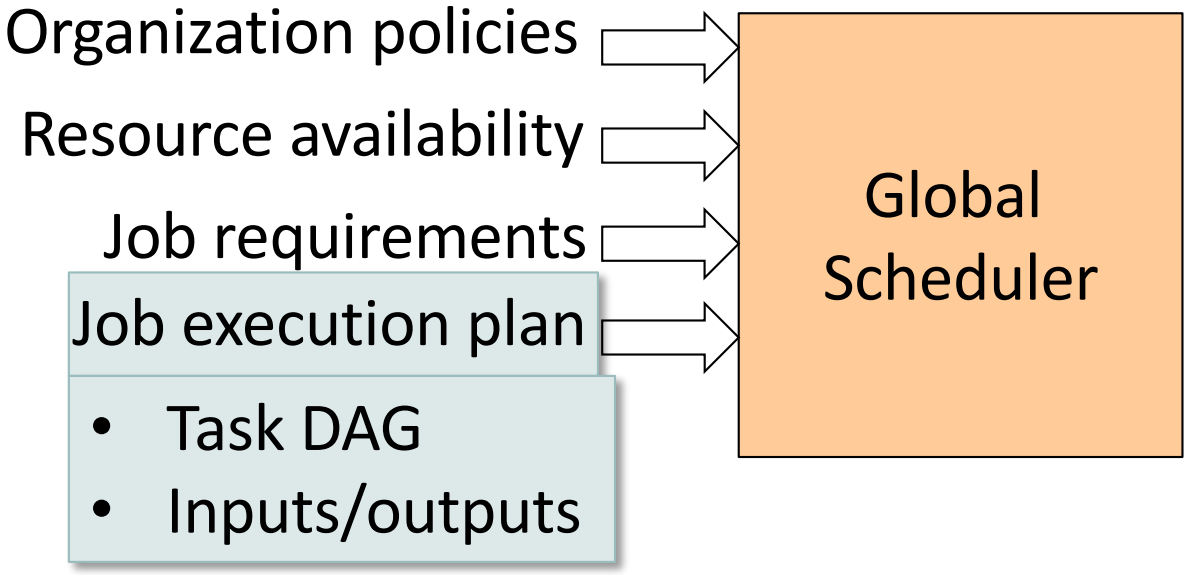
Goals

- **Efficient utilization** of resources
- **Support diverse frameworks** (existing & future)
- **Scalability** to 10,000's of nodes
- **Reliability** in face of node failures

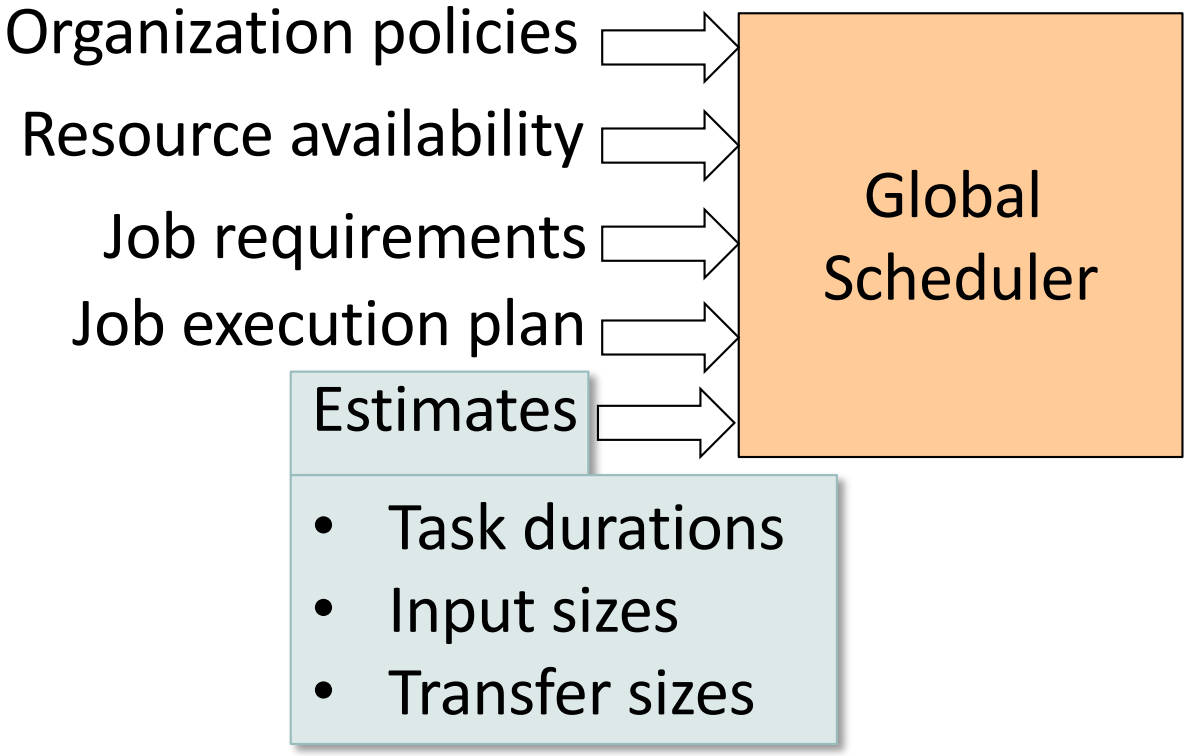
Approach: Global Scheduler



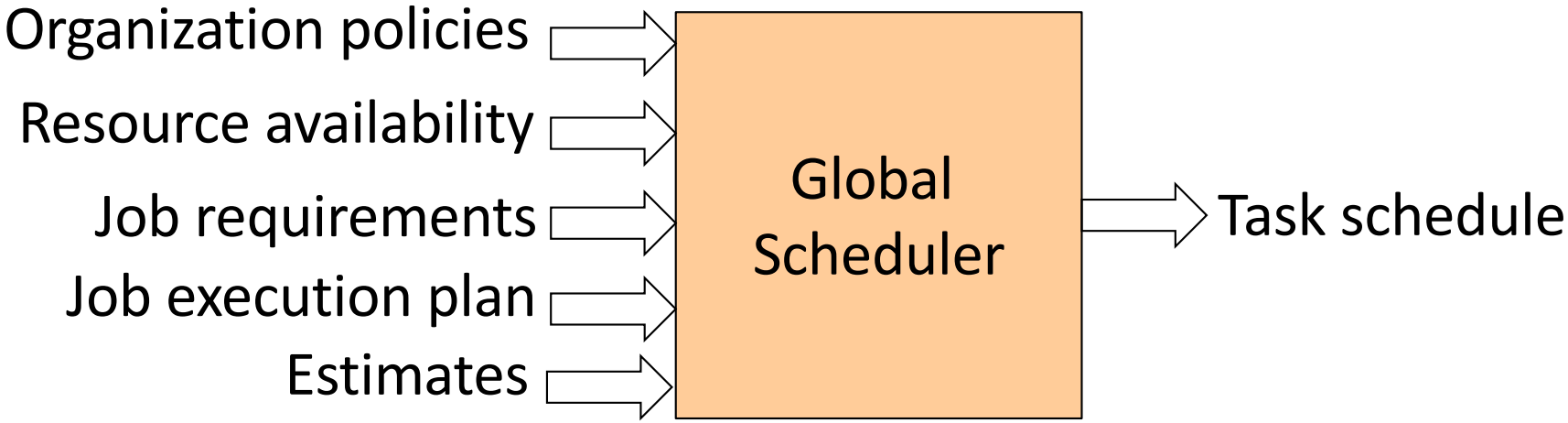
Approach: Global Scheduler



Approach: Global Scheduler



Approach: Global Scheduler



- Advantages: can achieve optimal schedule
- Disadvantages:
  - Complexity → hard to scale and ensure resilience
  - Hard to anticipate future frameworks' requirements
  - Need to refactor existing frameworks

Two Berkeley Nobel Prize

Reinhard Genzel

For showing that a massive black hole at the center of the Milky Way existed.

Jeniffer Doudna

For her research in developing CRISPR-Cas9.