MESOS

DESIGN

- Resource Allocation
  - Resource Offer: Data locality, Scalability, Robustness.
  - Filters
  - Revoke Resource
- Isolation
  - OS Containers
- Scalability and Robustness
  - Set filters
  - Make frameworks respond to offers fast
  - Rescinds offers
- Fault tolerance
  - Standby masters

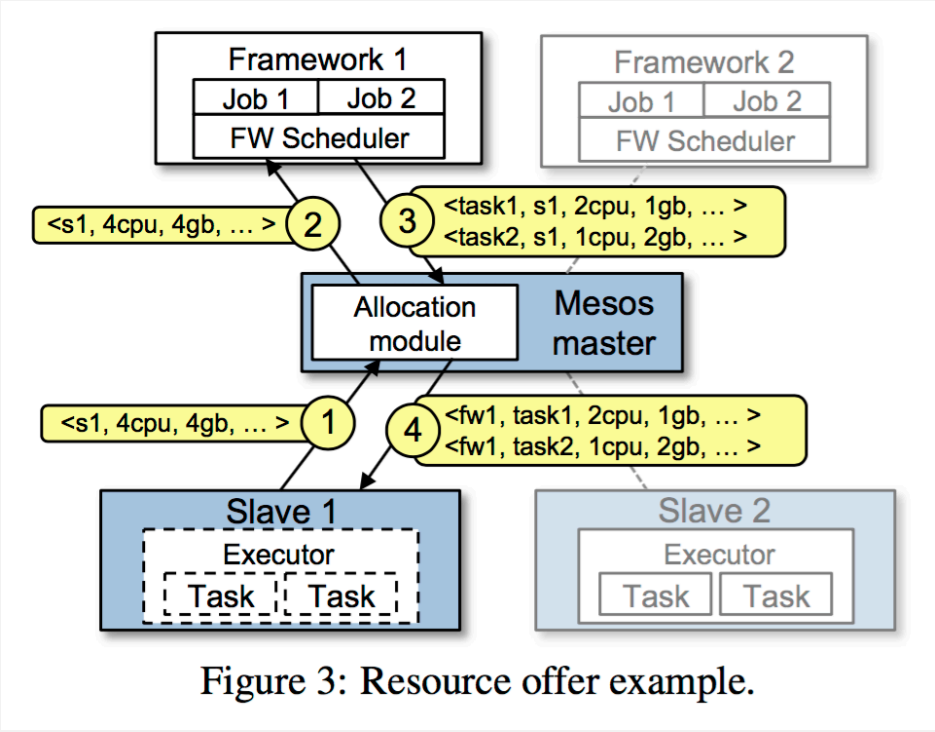


Figure 3: Resource offer example.

BEHAVIOR

- Def:
  - Ramp-up time: time to wait
  - Elastic: e.g. Mapreduce
- Homogeneous Tasks
  - The following table

	Elastic Framework		Rigid Framework	
	Constant dist.	Exponential dist.	Constant dist.	Exponential dist.
Ramp-up time	T	$T \ln k$	T	$T \ln k$
Completion time	$(1/2 + \beta)T$	$(1 + \beta)T$	$(1 + \beta)T$	$(\ln k + \beta)T$
Utilization	1	1	$\beta/(1/2 + \beta)$	$\beta/(\ln k - 1 + \beta)$

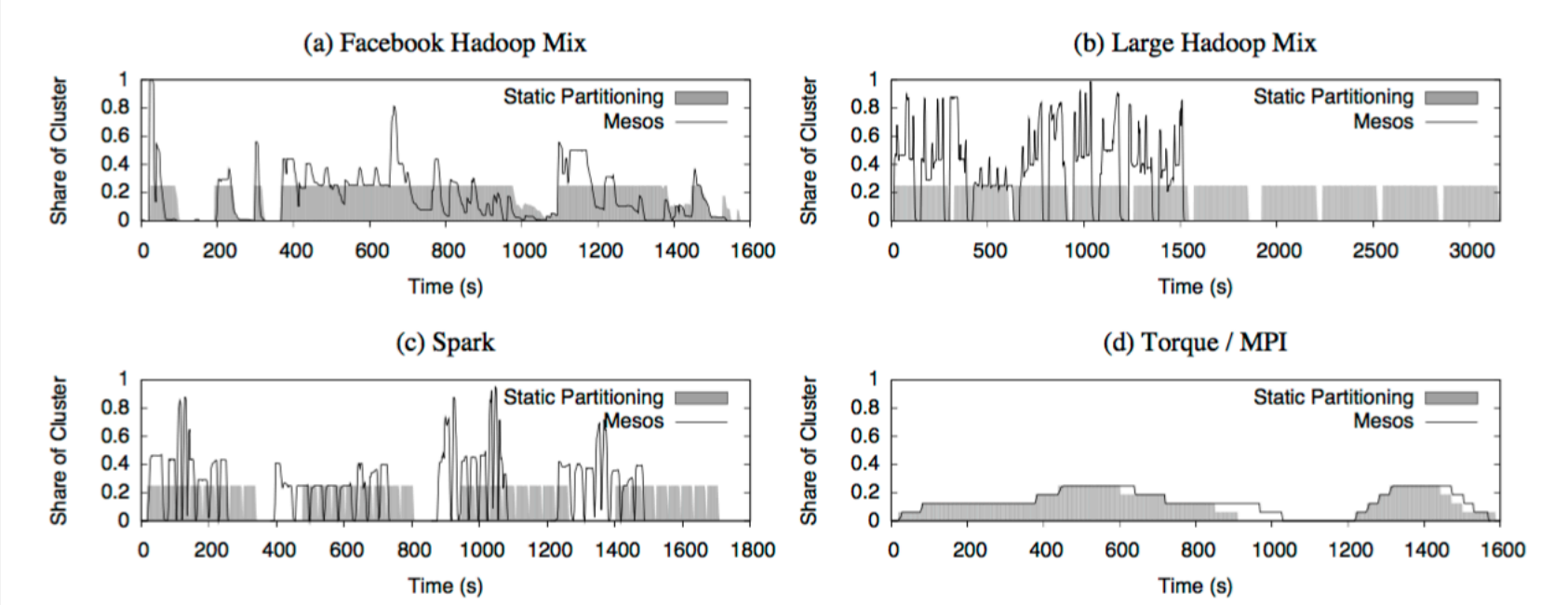
- Placement Preferences
  - Allocate with probability proportional to the total number of slots. The table still holds for ramp-up time and completion time
- Heterogeneous Tasks
  - Fraction of long tasks not close to 1. Nodes support multiple slots. -> Still performs well.
  - Alleviate to impact of long tasks: Set maximum duration for some resources on nodes.

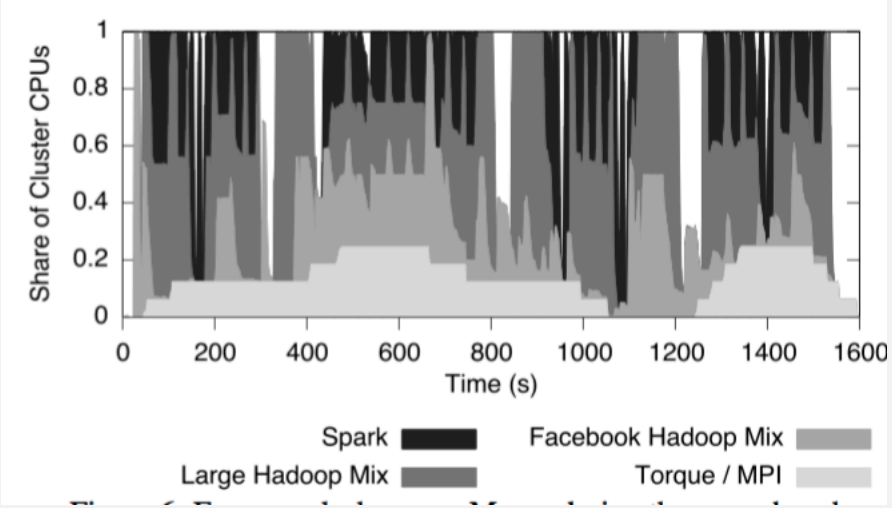
FRAMEWORK INCENTIVES

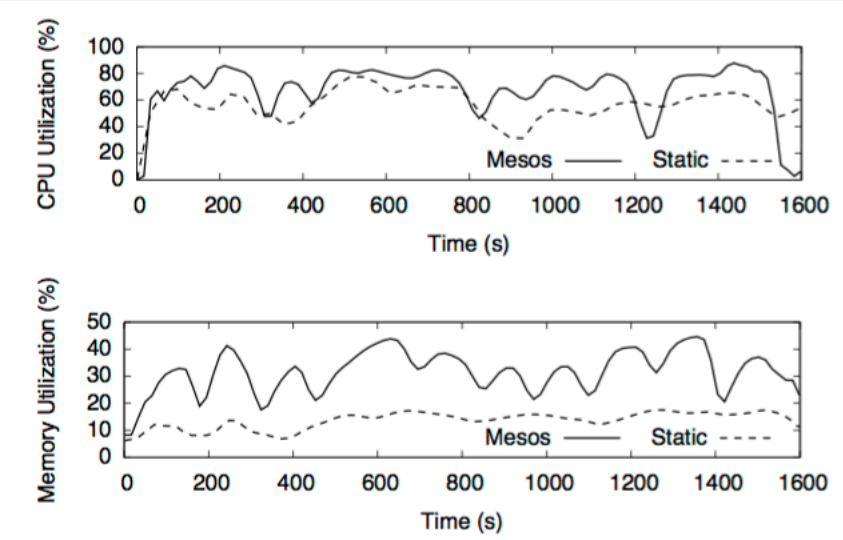
- Short Tasks
- Scale Elastically
- Do not accept unknown resources

EVALUATION COMPARED TO STATIC PARTITIONING

- Hadoop small jobs
- Hadoop large jobs
- Spark
- Torque/MPI (Rigid framework)







EVALUATION

- Hadoop small jobs
- Hadoop large jobs
- Spark
- Torque/MPI (Rigid framework)

Framework	Sum of Exec Times w/ Static Partitioning (s)	Sum of Exec Times with 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**

- Small Jobs: Resource offering
- Large Jobs: Fill in the gaps
- Torque:
  - Wait for all the resources
  - Wait for the last task

Framework	Job Type	Exec Time w/ Static Partitioning (s)	Avg. Speedup on Mesos
Facebook Hadoop Mix	selection (1)	24	**0.84**
	text search (2)	31	**0.90**
	aggregation (3)	82	**0.94**
	selection (4)	65	**1.40**
	aggregation (5)	192	**1.26**
	selection (6)	136	**1.71**
	text search (7)	137	**2.14**
	join (8)	662	**1.35**
Large Hadoop Mix	text search	314	**2.21**
Spark	ALS	337	**1.36**
Torque / MPI	small tachyon	261	**0.91**
	large tachyon	822	**0.88**

EVALUATION

- Scalability
- Failure Recovery
- Performance Isolation
- Little Overhead.
- Data locality
  - Resource offering
  - In contrast, running the Hadoop instances on Mesos improves data locality, even without delay scheduling, because each Hadoop in- stance has tasks on more nodes of the cluster (there are 4 tasks per node), and can therefore access more blocks locally.
  - Delay scheduling