Niklas Nielson, Greg Mann & Sunil Shah

POWERING THE INTERNET WITH APACHE MESOS

MESOSPHERE

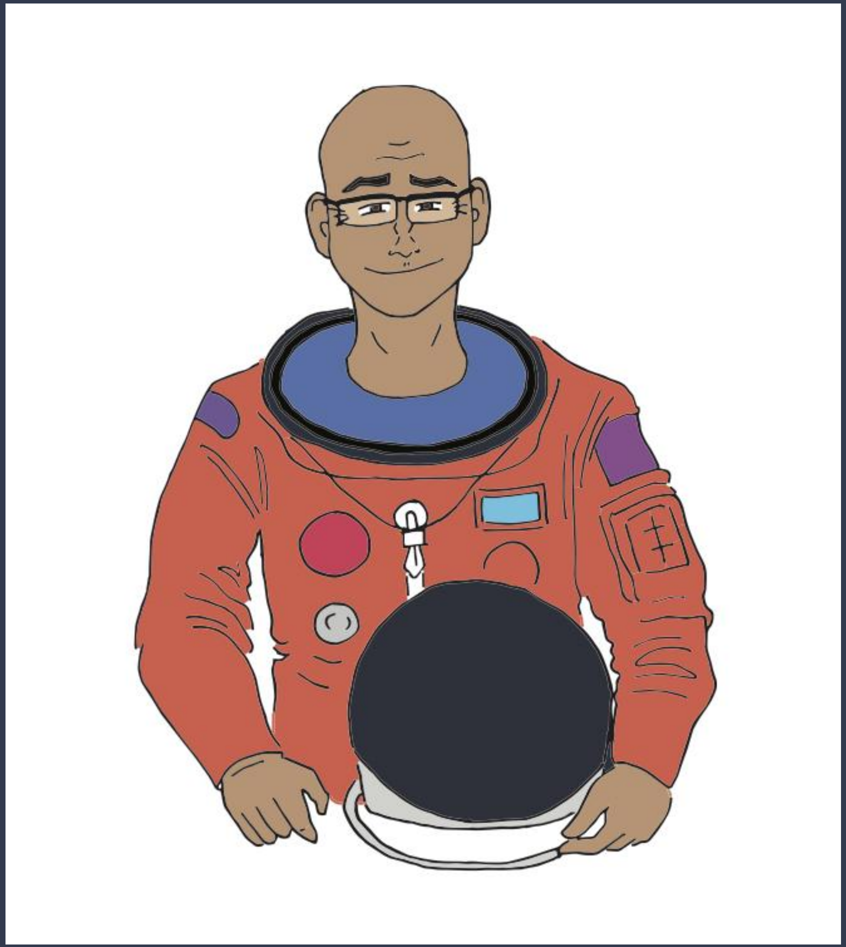
LIFE WITHOUT MESOS

COMPLEX WORKLOADS

Operating a large datacenter is a hard job!

Modern clusters must accommodate heterogeneous workloads. Tasks may vary with respect to:

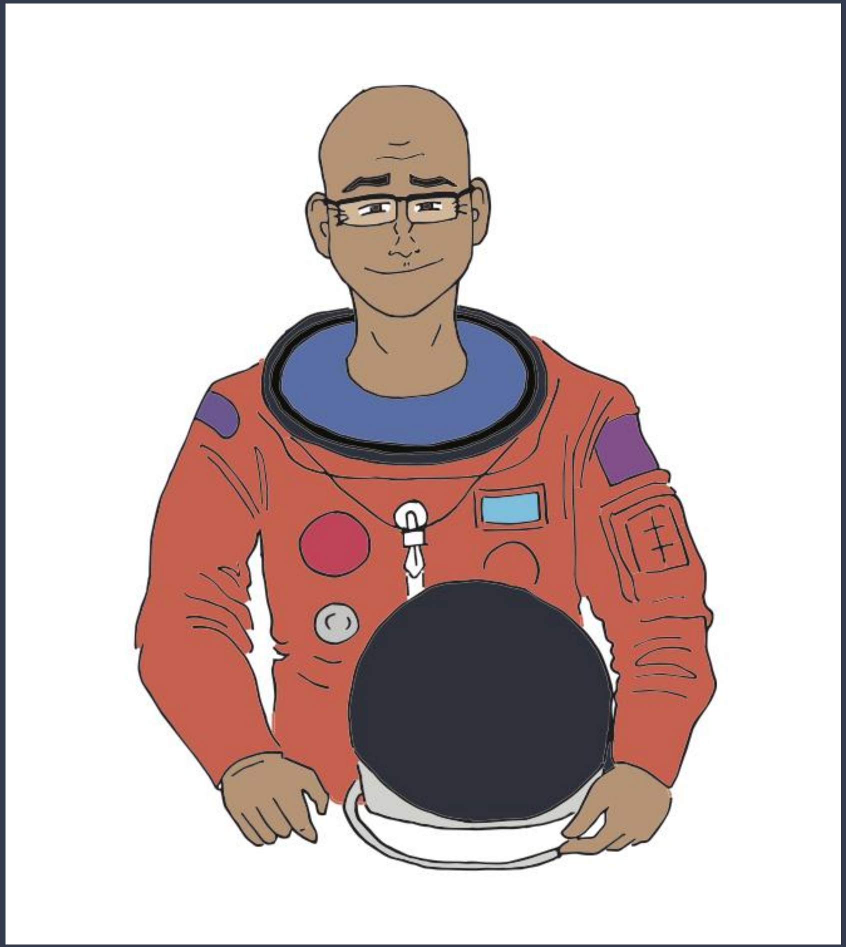
- Runtime
- Resource needs
- Priority
- Communication patterns



COMPLEX WORKLOADS

Dan, our operator, is forced to partition the datacenter to accommodate these demands.

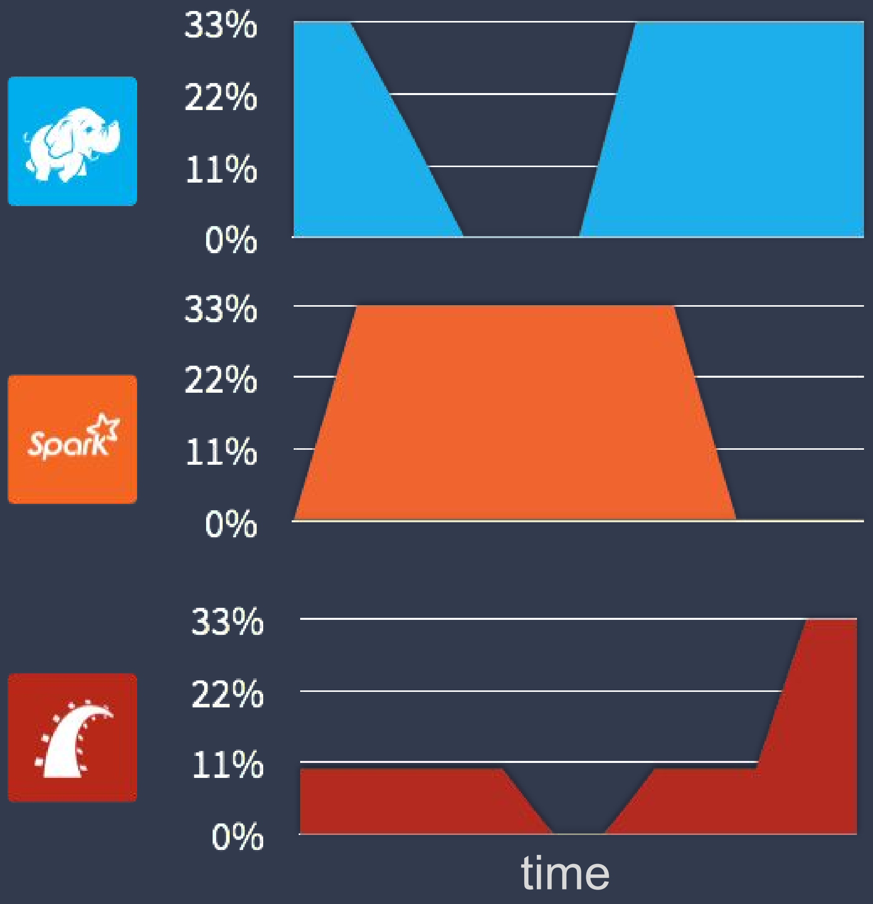
He must address errors and failures **manually**.



KEEP IT STATIC

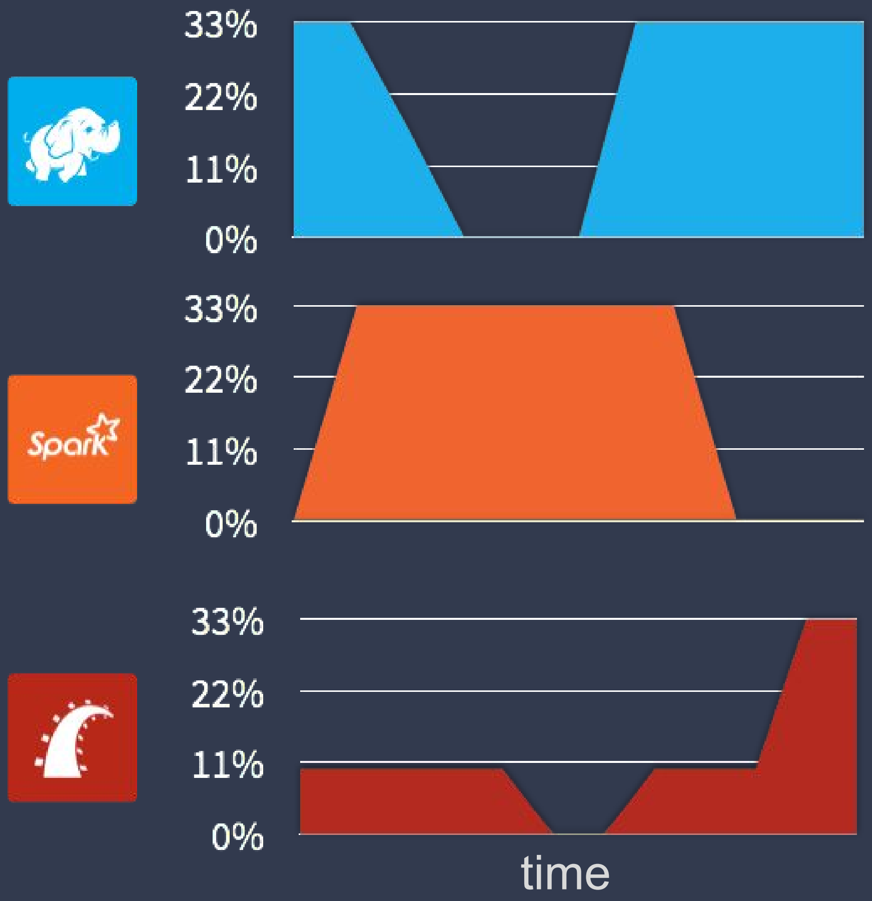
A naive approach to handling varied app requirements: **static partitioning**.

This can cope with heterogeneity, but is very expensive.



KEEP IT STATIC

Maintaining sufficient headroom to handle peak workloads on all partitions leads to **poor utilization** overall.



CROSS-FRAMEWORK HEADACHES

Running multiple frameworks makes things difficult:

- Real-time performance metrics
- Tracing for post-hoc analysis
- Inter-framework cooperation

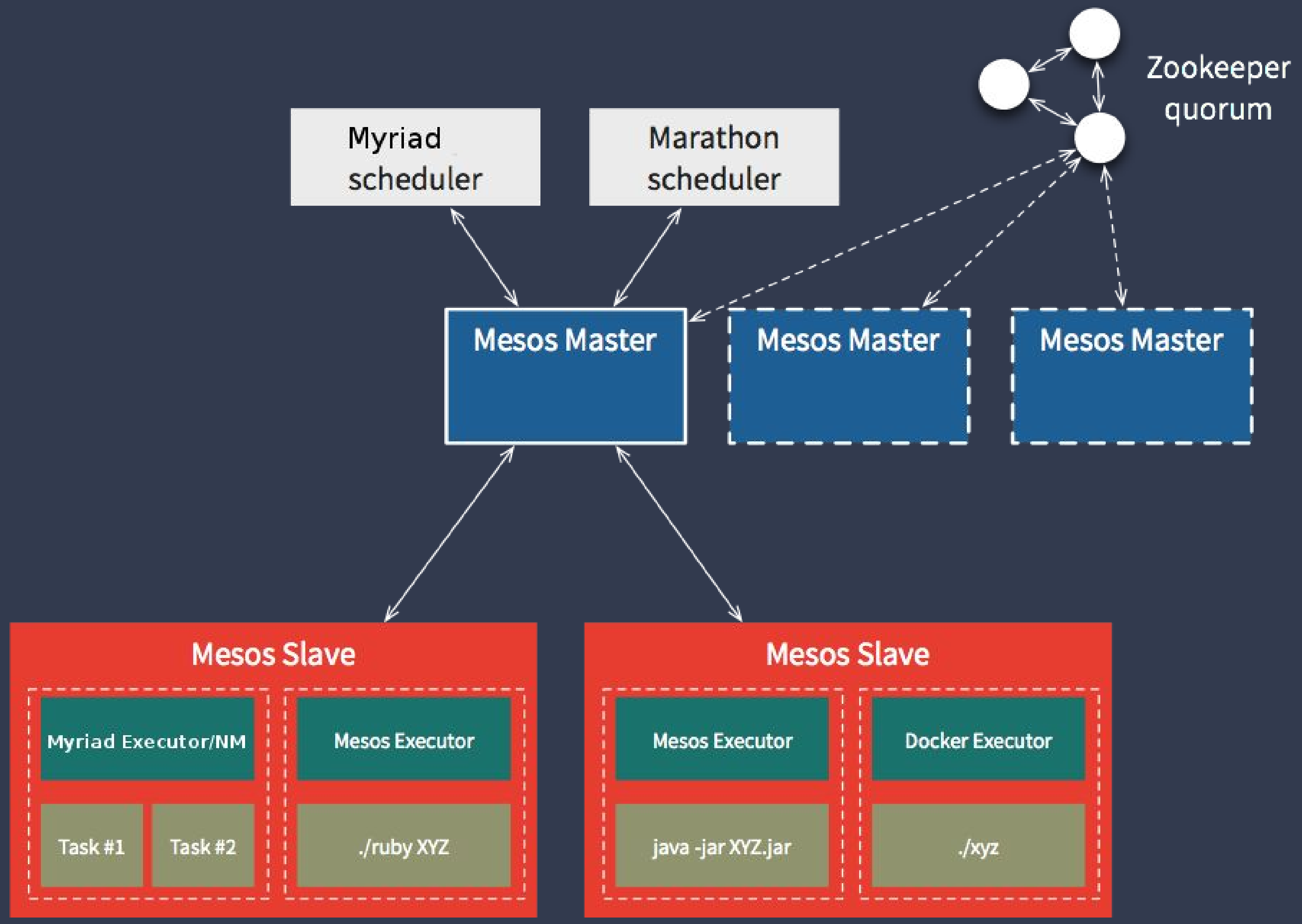
HISTORY OF MESOS

TIMELINE

SLIDE TITLE

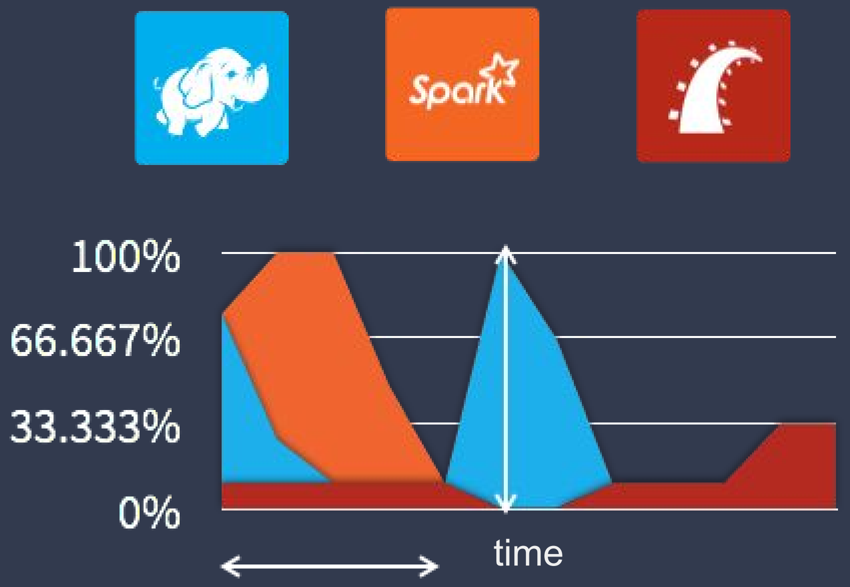
MESOS FUNDAMENTALS

ARCHITECTURE



SHARED RESOURCES

Multiple frameworks can use the same cluster resources, with their share adjusting dynamically.



RECENT DEVELOPMENTS

OVERSUBSCRIPTION

NETWORK ISOLATION

Why isolate containers on the network?

- Improved security
- Reduced latency
- Port conflicts
- Service discovery
- External communication

NETWORK ISOLATION

To isolate tasks, we isolate their containers:

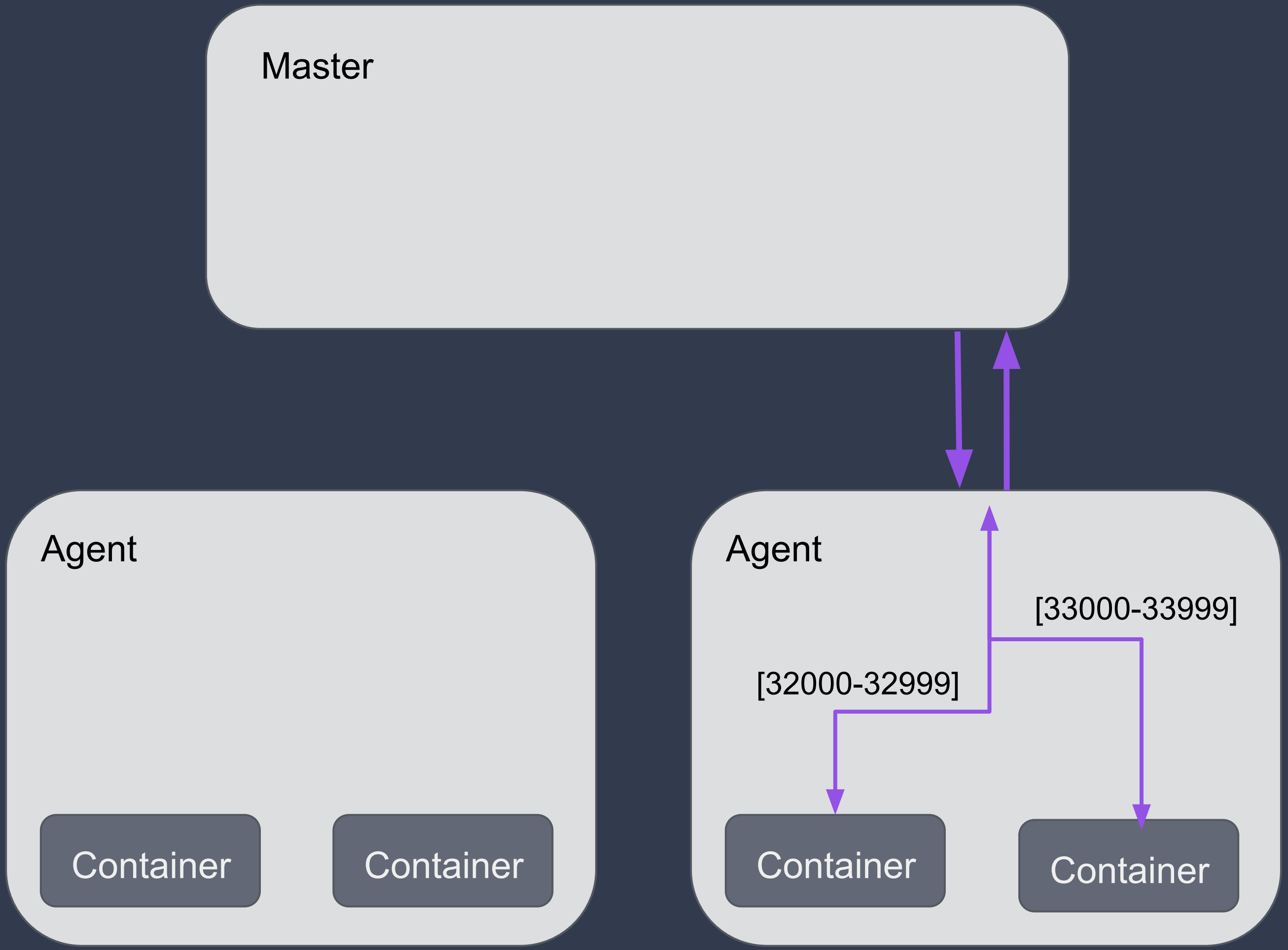
- Mesos containerizer
- Docker
- Universal containerizer (in progress)

Implemented as modules:

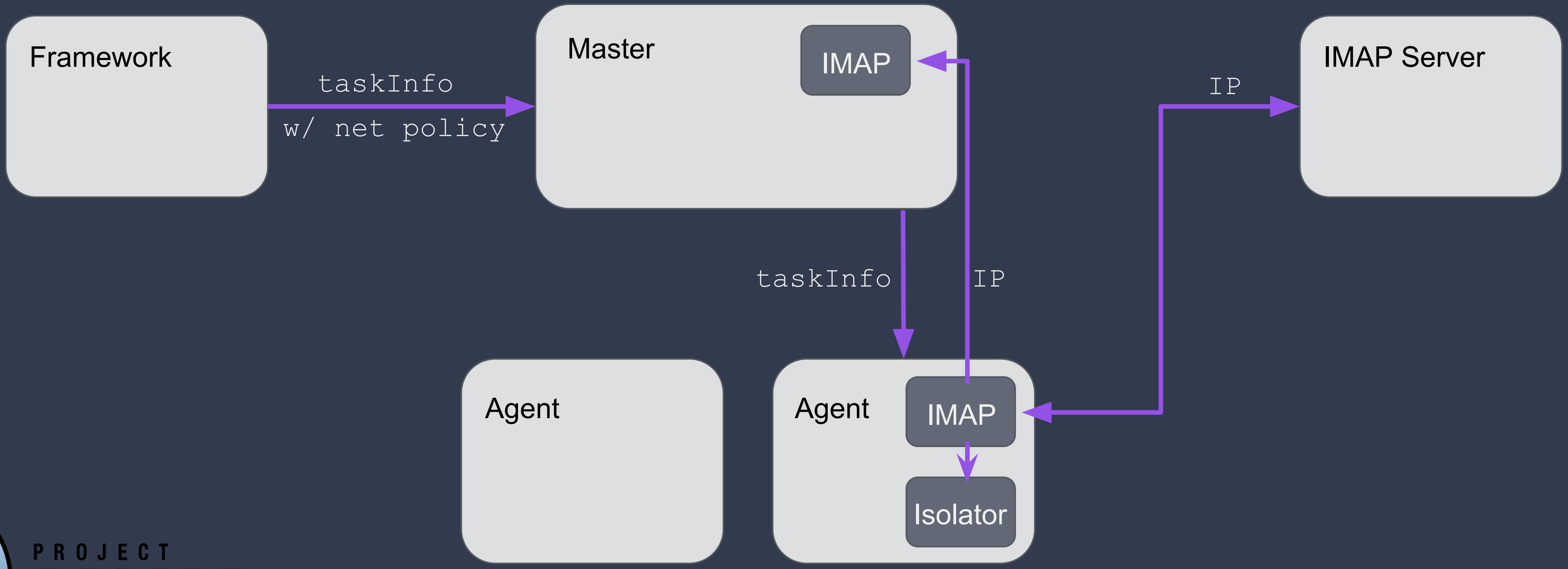
- Project Calico
- Port-mapping isolation
- …

PORT-MAPPING ISOLATOR

- Standard linux netfilter policies bypassed on agent
- Custom queueing disciplines instead
- Ports assigned and tracked via scheduler (ex: Aurora)



CALICO NETWORK ISOLATION



© 2015 Mesosphere, Inc. All Rights Reserved.

CALICO NETWORK ISOLATION

Calico Network Virtualizer:

- Pure Layer-3 solution
- Modifies the agent's Forwarding Information Base to route container traffic
- ACLs defined to enforce security
- Advertises routes to local containers via BGP
- Can assign IP-per-container

THE ROAD AHEAD

EXTERNAL VOLUMES

Mesos will provide…

- facilities to connect agents to external storage volumes
- primitives for task/agent reconciliation with respect to external volumes
- access to volumes from multiple providers

CLUSTER-WIDE RESOURCES

Some resources are not node-local:

- External storage volumes
- Software licenses
- IP addresses
- External network bandwidth

CLUSTER-WIDE RESOURCES

Currently, all resources are advertised by agent nodes.

Cluster-wide resources must be offered by a separate entity, either a Mesos **framework** or a Mesos **module**.

THE DATACENTER OPERATING SYSTEM

DCOS aims to make developing & deploying distributed apps easier.

Short-term:

- Software installation/removal
- Seamless upgrades
- Automatic failure detection, reconciliation

THE DATACENTER OPERATING SYSTEM

Long-term: the distributed SDK