Neil Conway, Niklas Nielsen, Greg Mann & Sunil Shah

POWERING THE INTERNET WITH APACHE MESOS

MESOSPHERE

© 2015 Mesosphere, Inc. All Rights Reserved.

MESOS: ORIGINS

© 2015 Mesosphere, Inc. All Rights Reserved.

THE BIRTH OF MESOS

Spring 2009

CS262B

Ben Hindman, Andy Konwinski and Matei Zaharia create “Nexus” as their CS262B class project.

March 2010

TWITTER TECH TALK

The grad students working on Mesos give a tech talk at Twitter.

September 2010

MESOS PUBLISHED

Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center is published as a technical report.

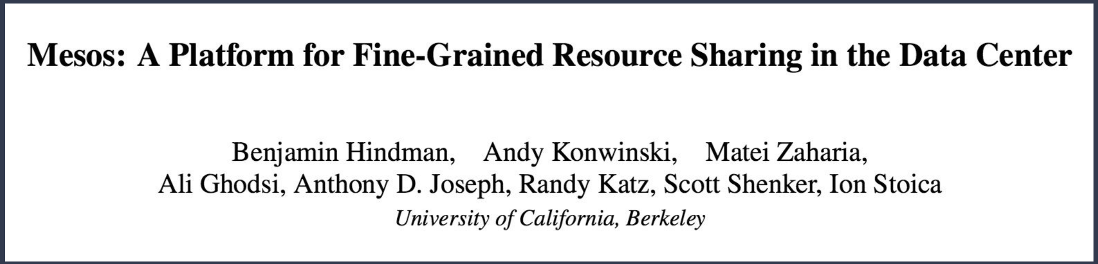
December 2010

APACHE INCUBATION

Mesos enters the Apache Incubator.

© 2015 Mesosphere, Inc. All Rights Reserved.

TECHNOLOGY

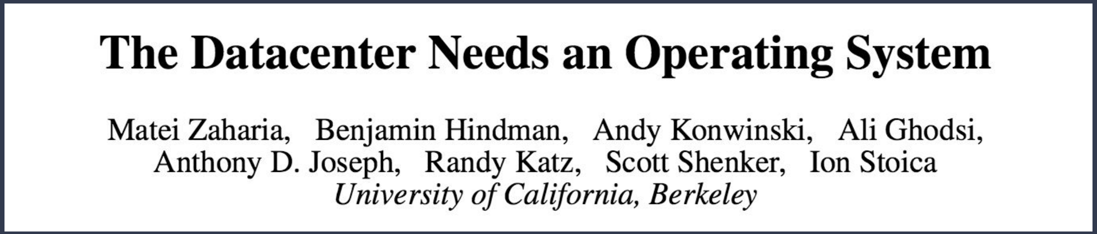

Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center

Benjamin Hindman, Andy Konwinski, Matei Zaharia,
Ali Ghodsi, Anthony D. Joseph, Randy Katz, Scott Shenker, Ion Stoica
University of California, Berkeley

Sharing resources between batch processing frameworks

- Hadoop
- MPI
- Spark

VISION


The Datacenter Needs an Operating System

Matei Zaharia, Benjamin Hindman, Andy Konwinski, Ali Ghodsi,
Anthony D. Joseph, Randy Katz, Scott Shenker, Ion Stoica
University of California, Berkeley

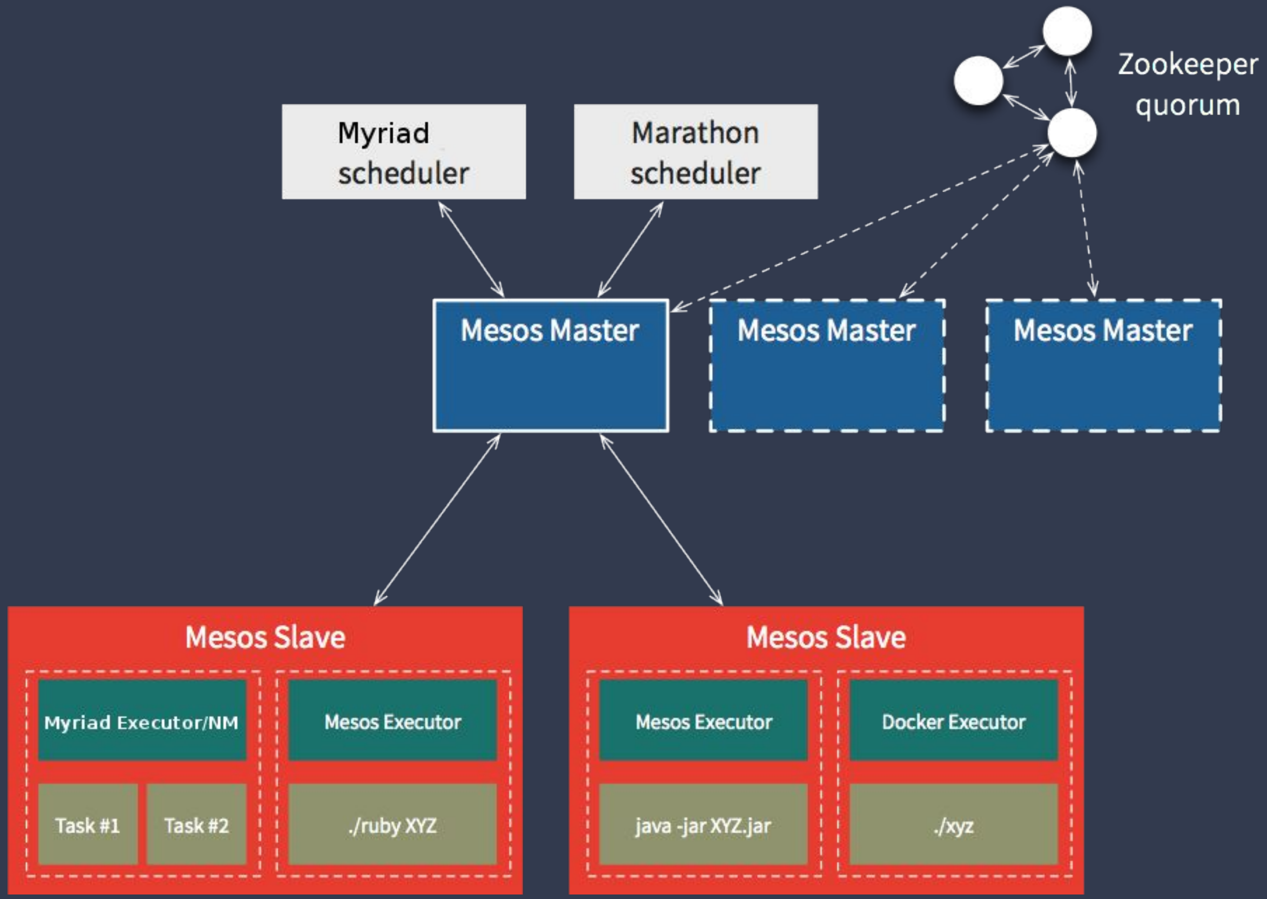
What does an operating system provide?

- Resource management
- Programming abstractions
- Security
- Monitoring, debugging, logging

© 2015 Mesosphere, Inc. All Rights Reserved.

MESOS FUNDAMENTALS

ARCHITECTURE

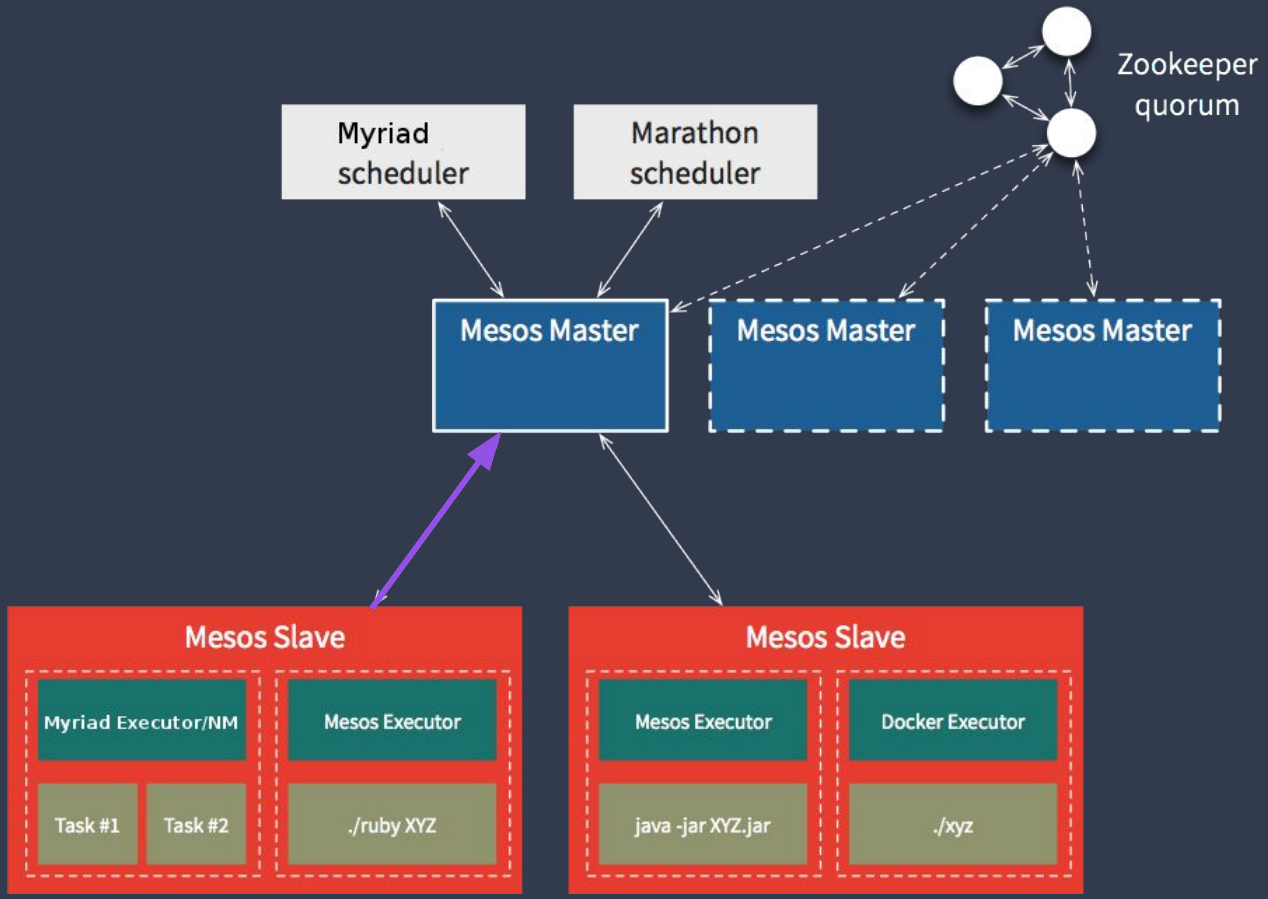


© 2015 Mesosphere, Inc. All Rights Reserved.

MESOS FUNDAMENTALS

ARCHITECTURE

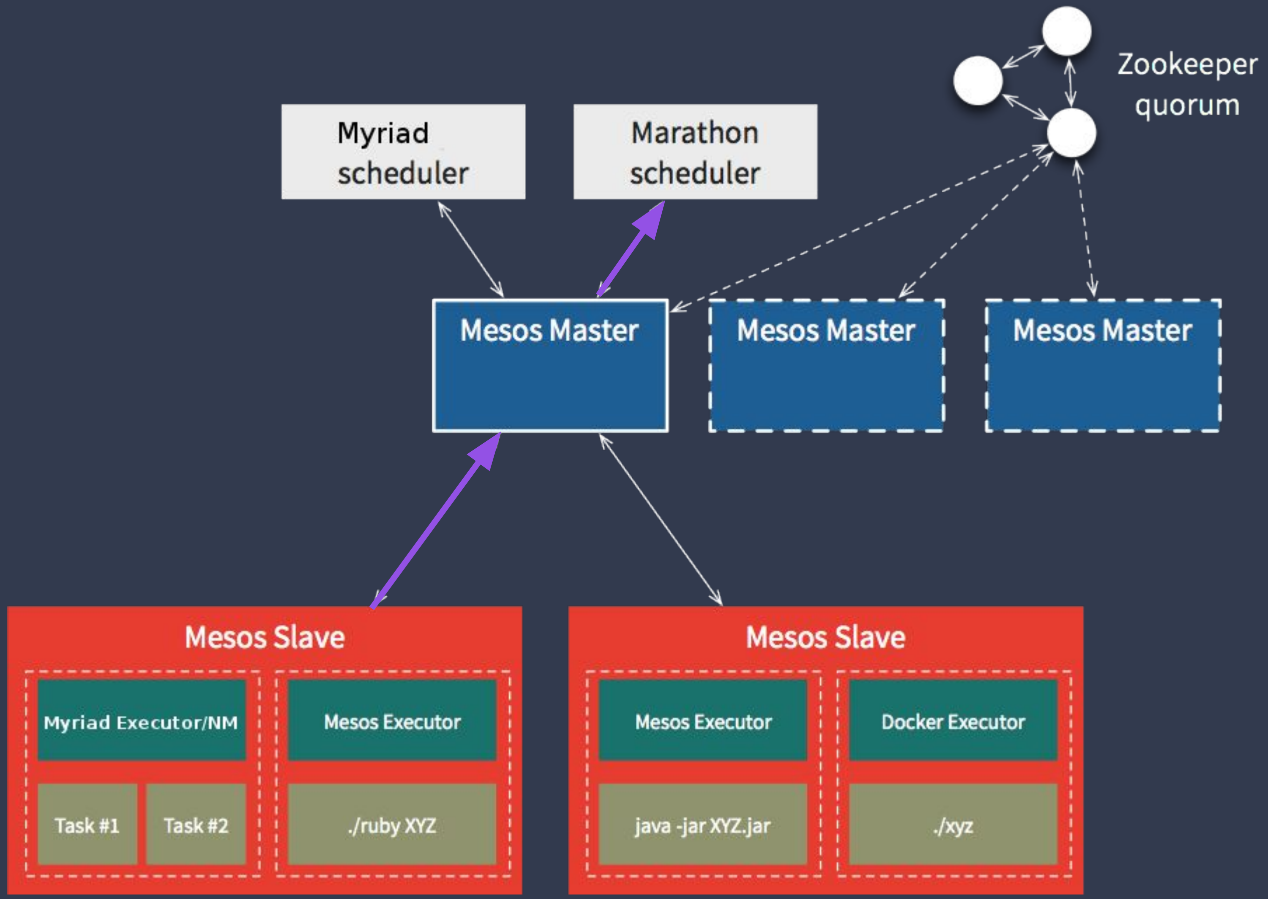
- Agents advertise resources to Master
- Master offers resources to Framework
- Framework rejects/uses resources
- Agents report task status to Master



MESOS FUNDAMENTALS

ARCHITECTURE

- Agents advertise resources to Master
- Master offers resources to Framework
- Framework rejects/uses resources
- Agents report task status to Master

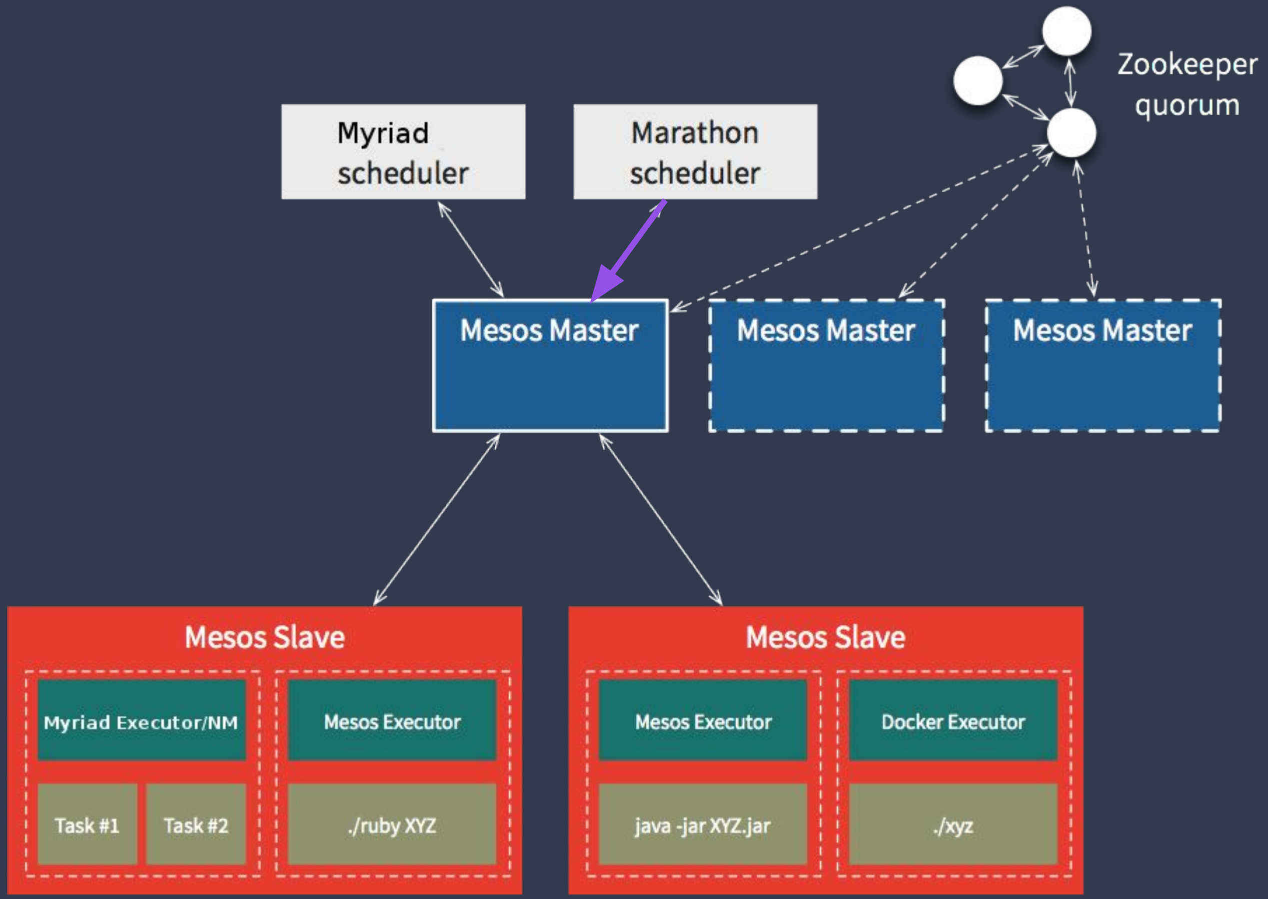


© 2015 Mesosphere, Inc. All Rights Reserved.

MESOS FUNDAMENTALS

ARCHITECTURE

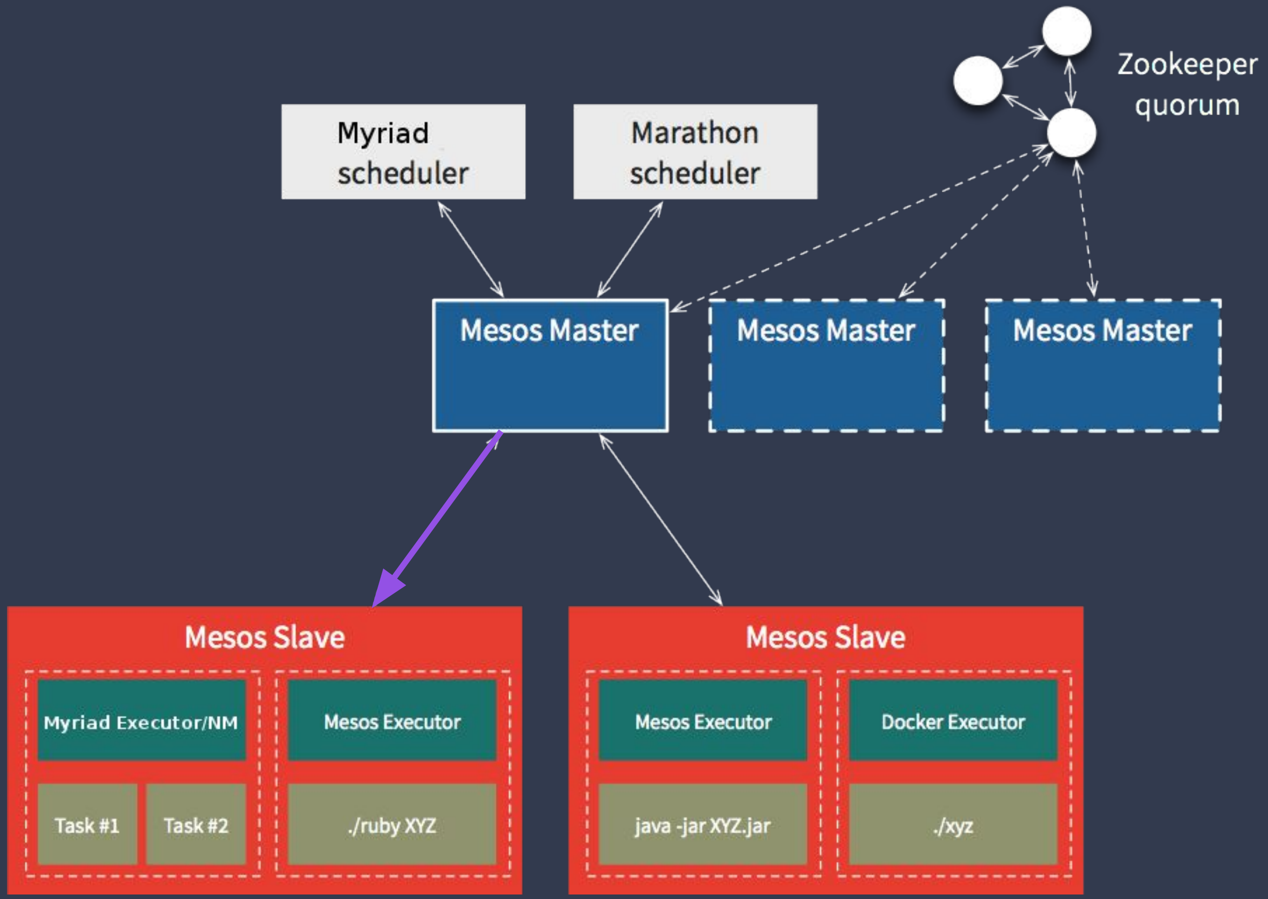
- Agents advertise resources to Master
- Master offers resources to Framework
- Framework rejects/uses resources
- Agents report task status to Master



MESOS FUNDAMENTALS

ARCHITECTURE

- Agents advertise resources to Master
- Master offers resources to Framework
- Framework rejects/uses resources
- Agents report task status to Master

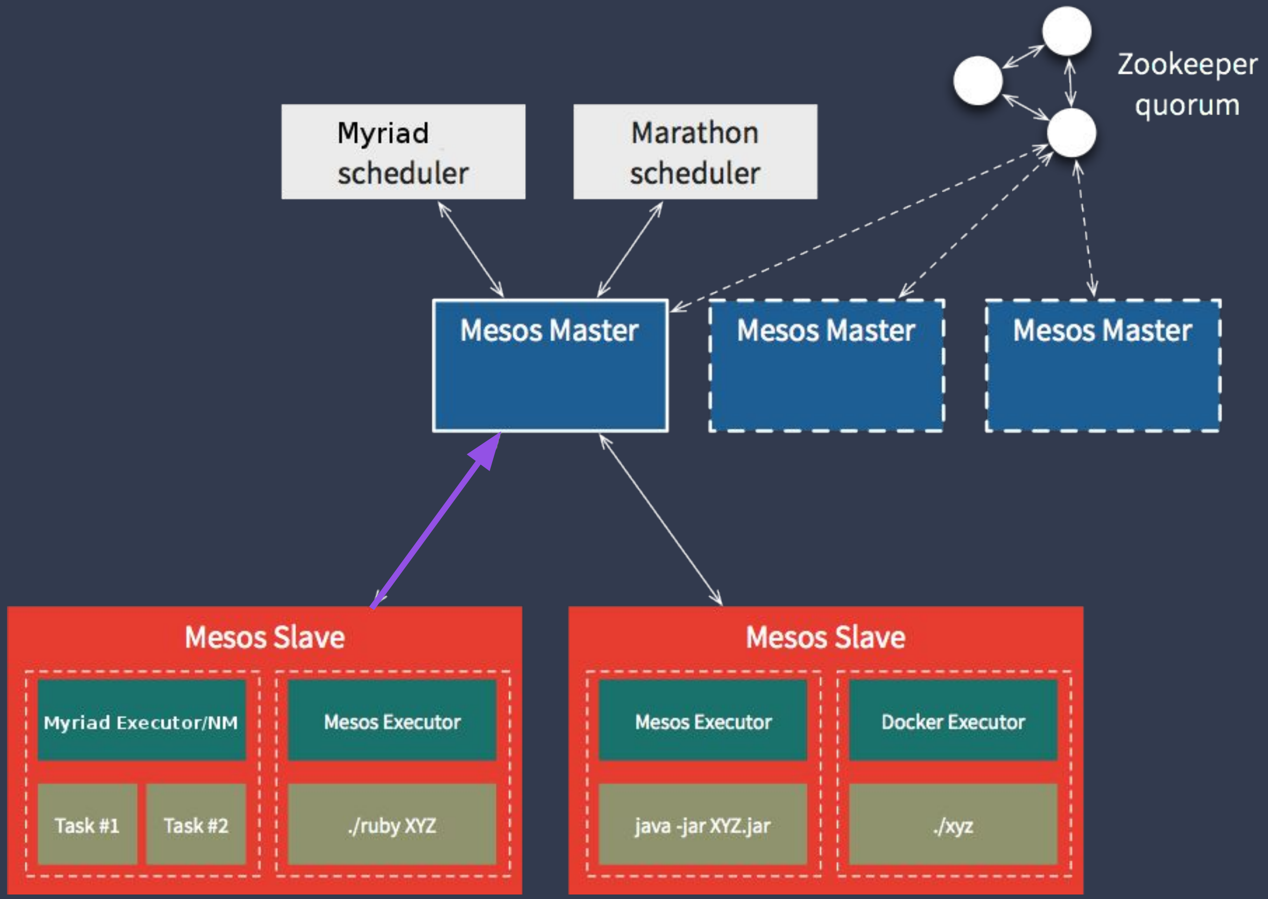


© 2015 Mesosphere, Inc. All Rights Reserved.

MESOS FUNDAMENTALS

ARCHITECTURE

- Agents advertise resources to Master
- Master offers resources to Framework
- Framework rejects/uses resources
- Agents report task status to Master

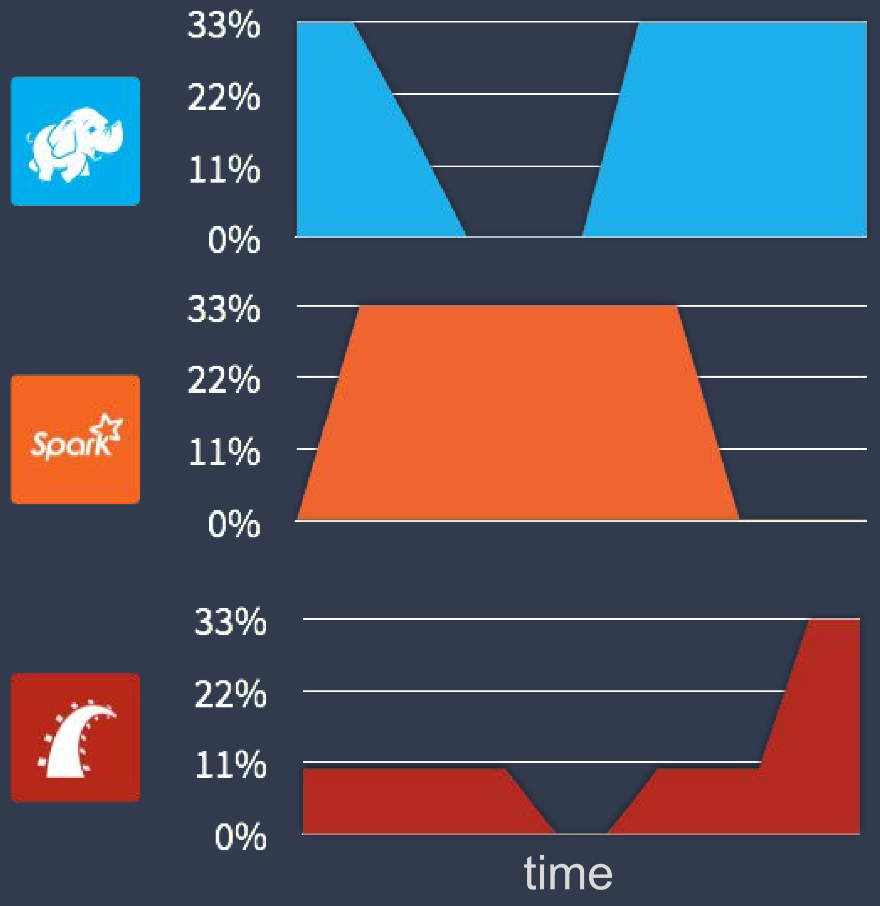


© 2015 Mesosphere, Inc. All Rights Reserved.

KEEP IT STATIC

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

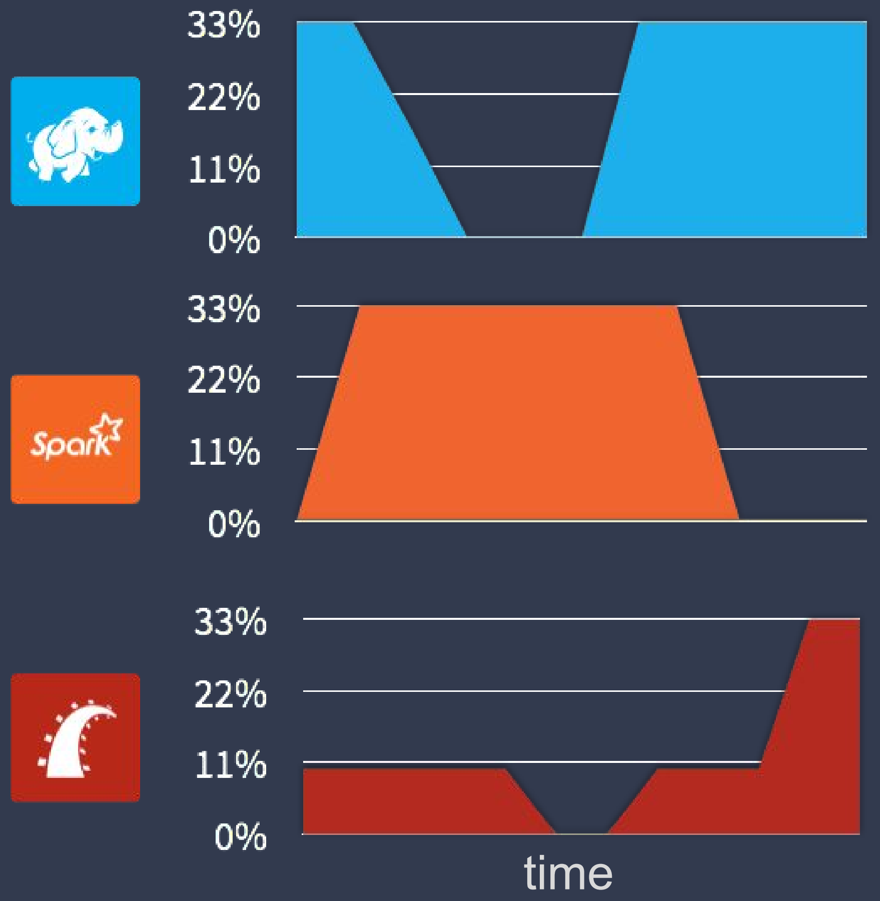
This can cope with heterogeneity, but is very expensive.



© 2015 Mesosphere, Inc. All Rights Reserved.

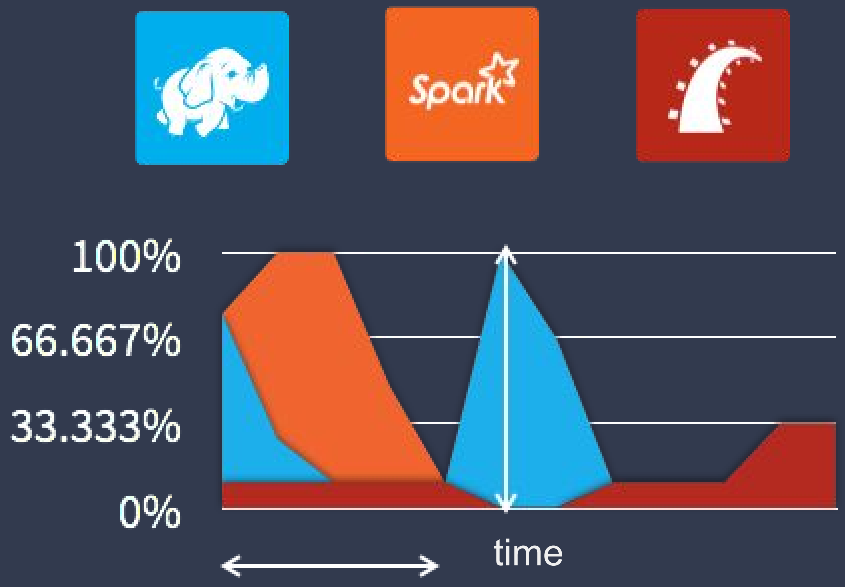
KEEP IT STATIC

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



SHARED RESOURCES

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



TWITTER & MESOS

THE BIRTH OF MESOS

Spring 2009

CS262B

Ben Hindman, Andy Konwinski and Matei Zaharia create Nexus as their CS262B class project.

March 2010

TWITTER TECH TALK

The grad students working on Mesos give a tech talk at Twitter.

September 2010

MESOS PUBLISHED

Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center is published as a technical report.

December 2010

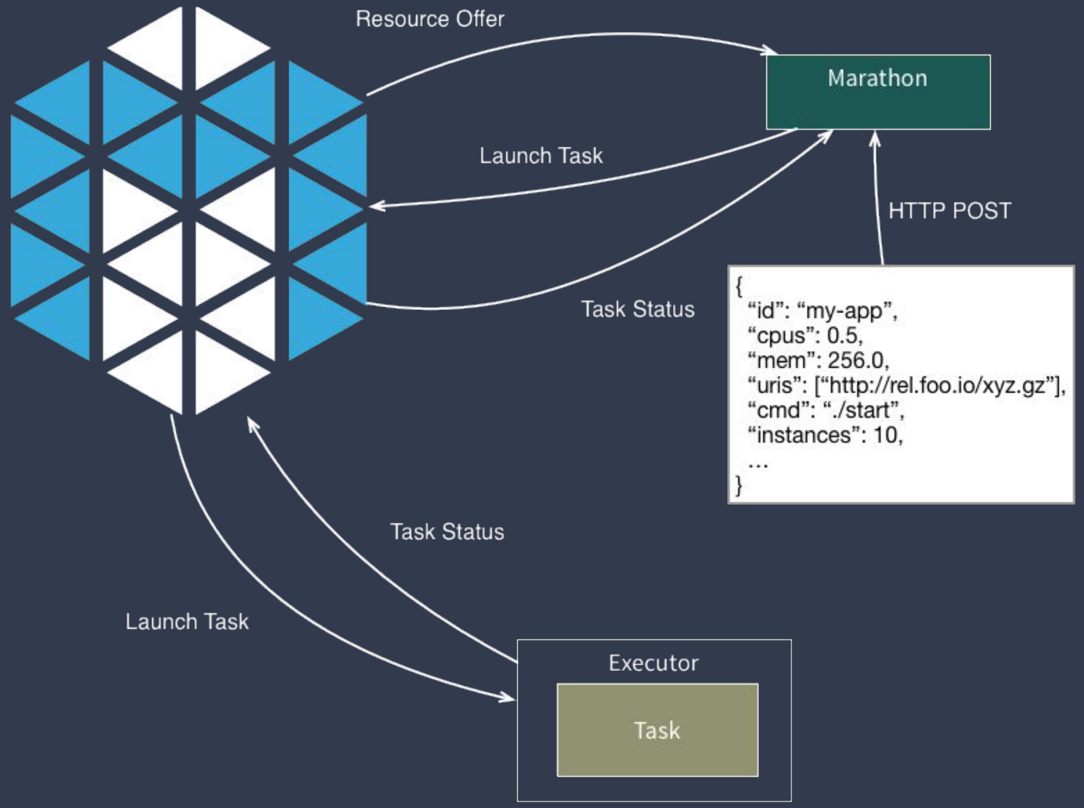
APACHE INCUBATION

Mesos enters the Apache incubator.

© 2015 Mesosphere, Inc. All Rights Reserved.

MESOS REALLY HELPS

- Former Google engineers at Twitter thought Mesos could provide the same functionality as Borg.
- Mesos actually works pretty well for long running services.



```
{
  "id": "my-app",
  "cpus": 0.5,
  "mem": 256.0,
  "uris": ["http://rel.foo.io/xyz.gz"],
  "cmd": "./start",
  "instances": 10,
  ...
}
```

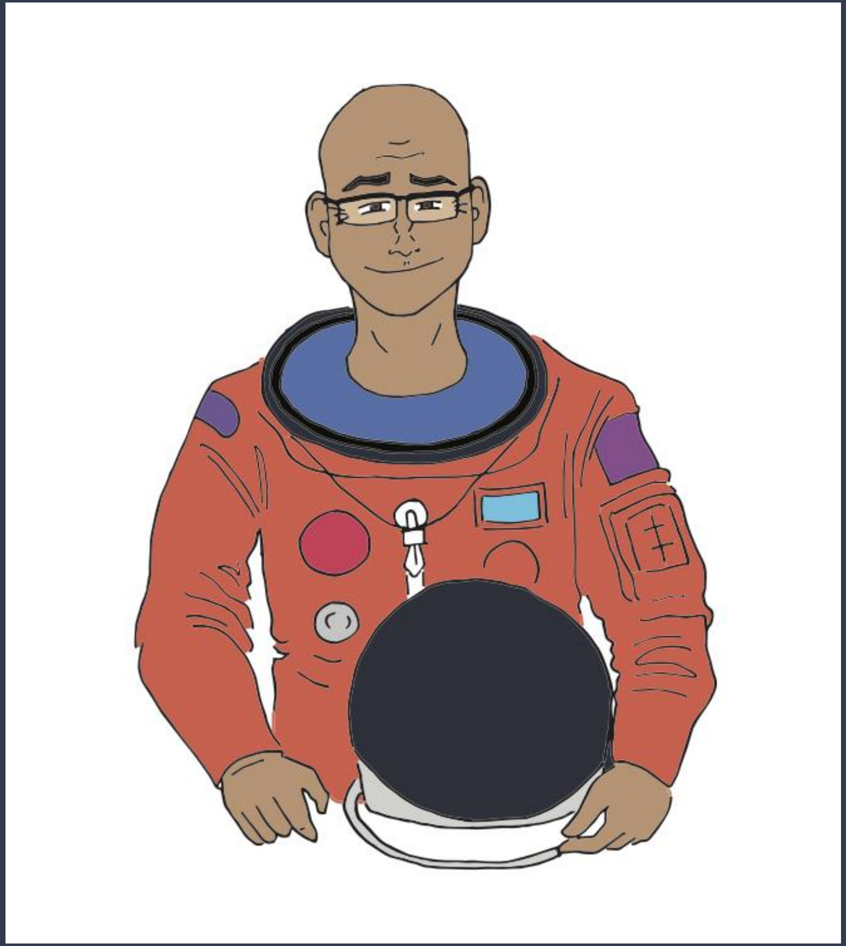
© 2015 Mesosphere, Inc. All Rights Reserved.

LIFE WITHOUT MESOS

© 2015 Mesosphere, Inc. All Rights Reserved.

SAY HI TO DAN

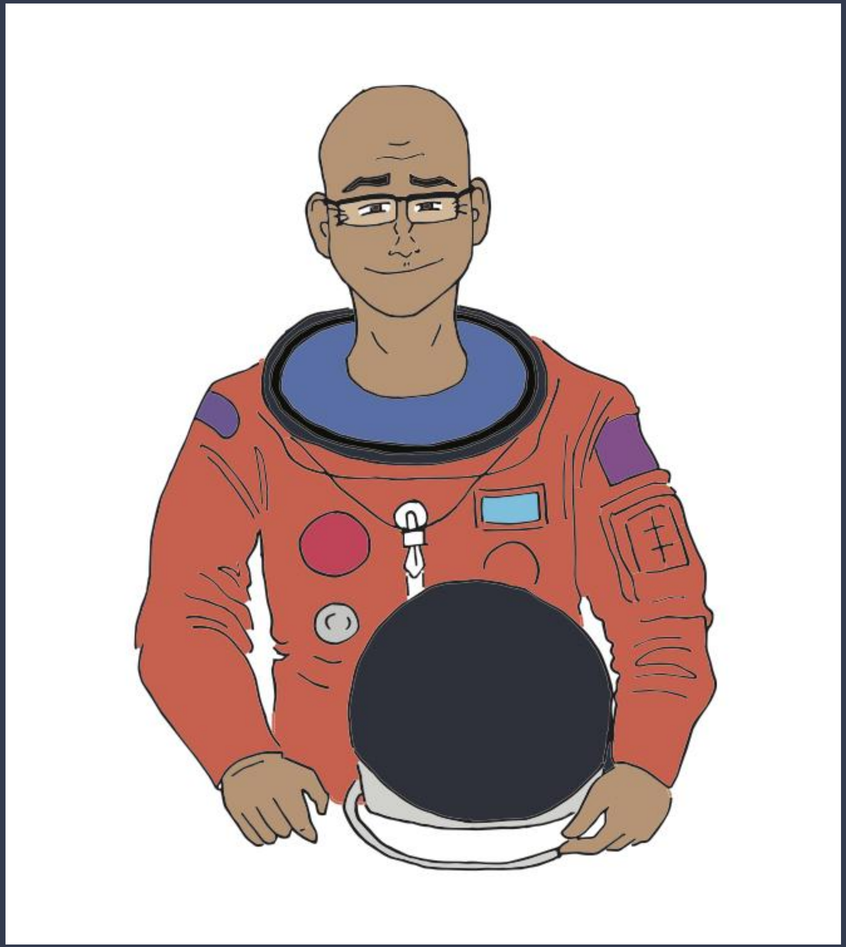
- Dan is a member of operations staff in a non-Google, non-Facebook company with large and growing users and workloads.



SO MUCH PROCESS

You’re a developer wanting to deploy a new service.

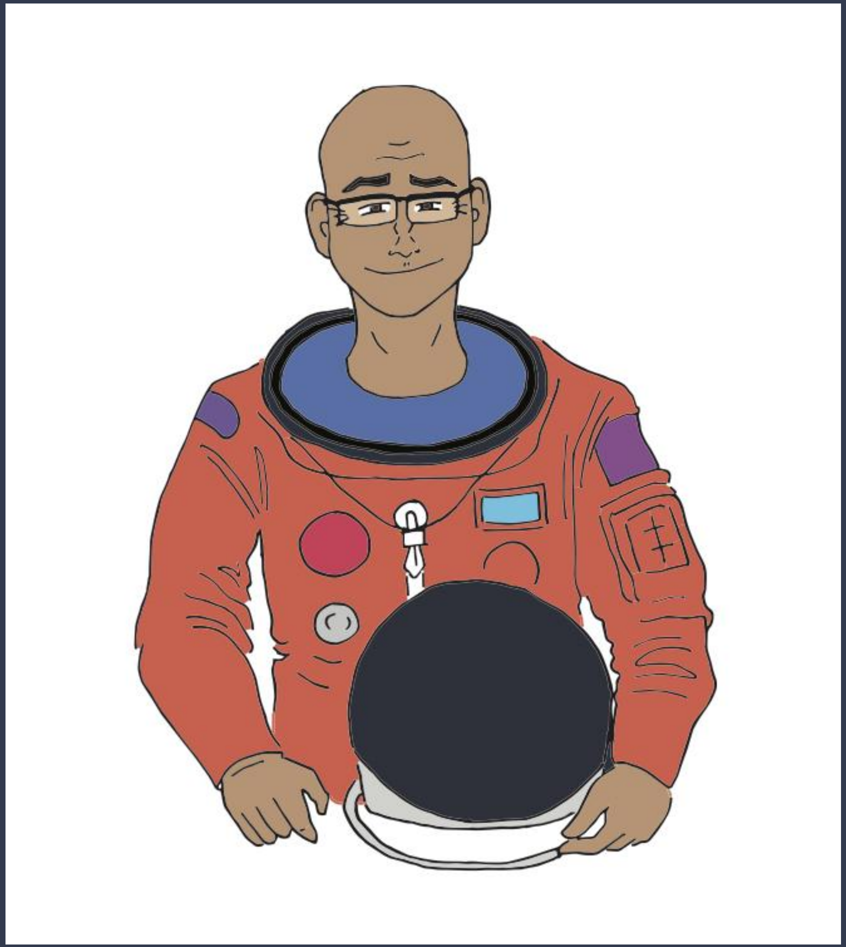
1. How many resources do you need?
   (Better overestimate, it usually takes a while to provision these.)
2. What dependencies does your application have?
3. Who monitors your applications and handles it falling over?



CHANGE IS PAINFUL

You have more users and/or you want to upgrade your application.

1. Submit another resource request.
2. Provision new machines.
3. How do we get any upgraded binaries/dependencies to existing machines?



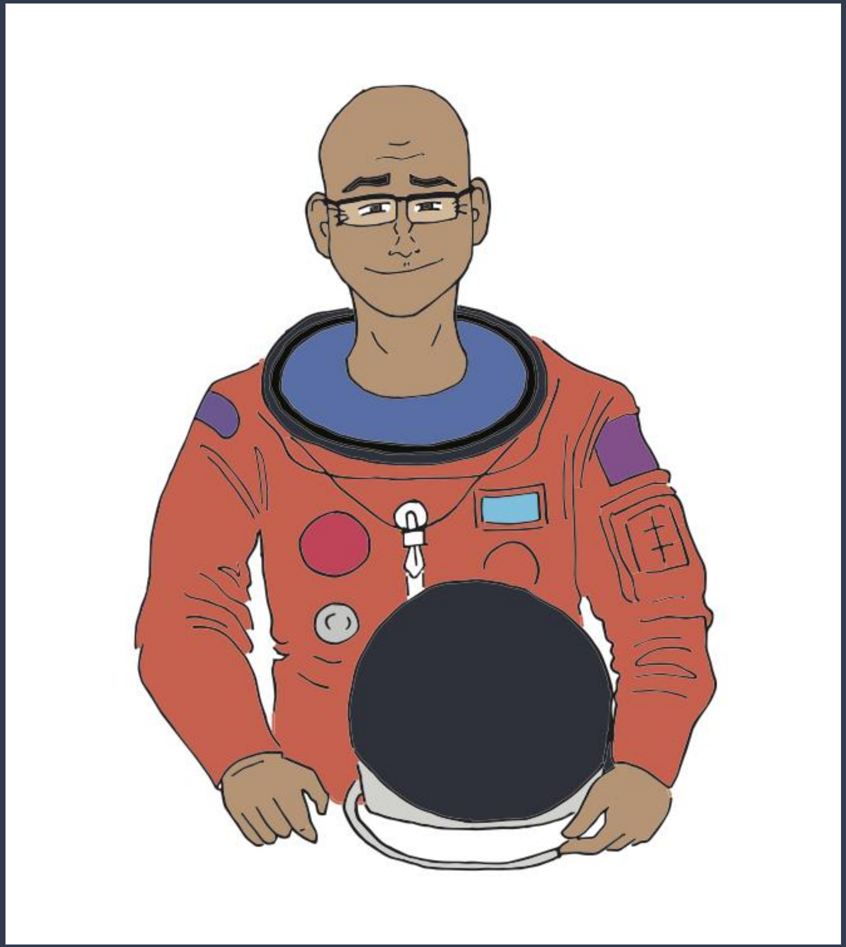
© 2015 Mesosphere, Inc. All Rights Reserved.

COMPLEX WORKLOADS

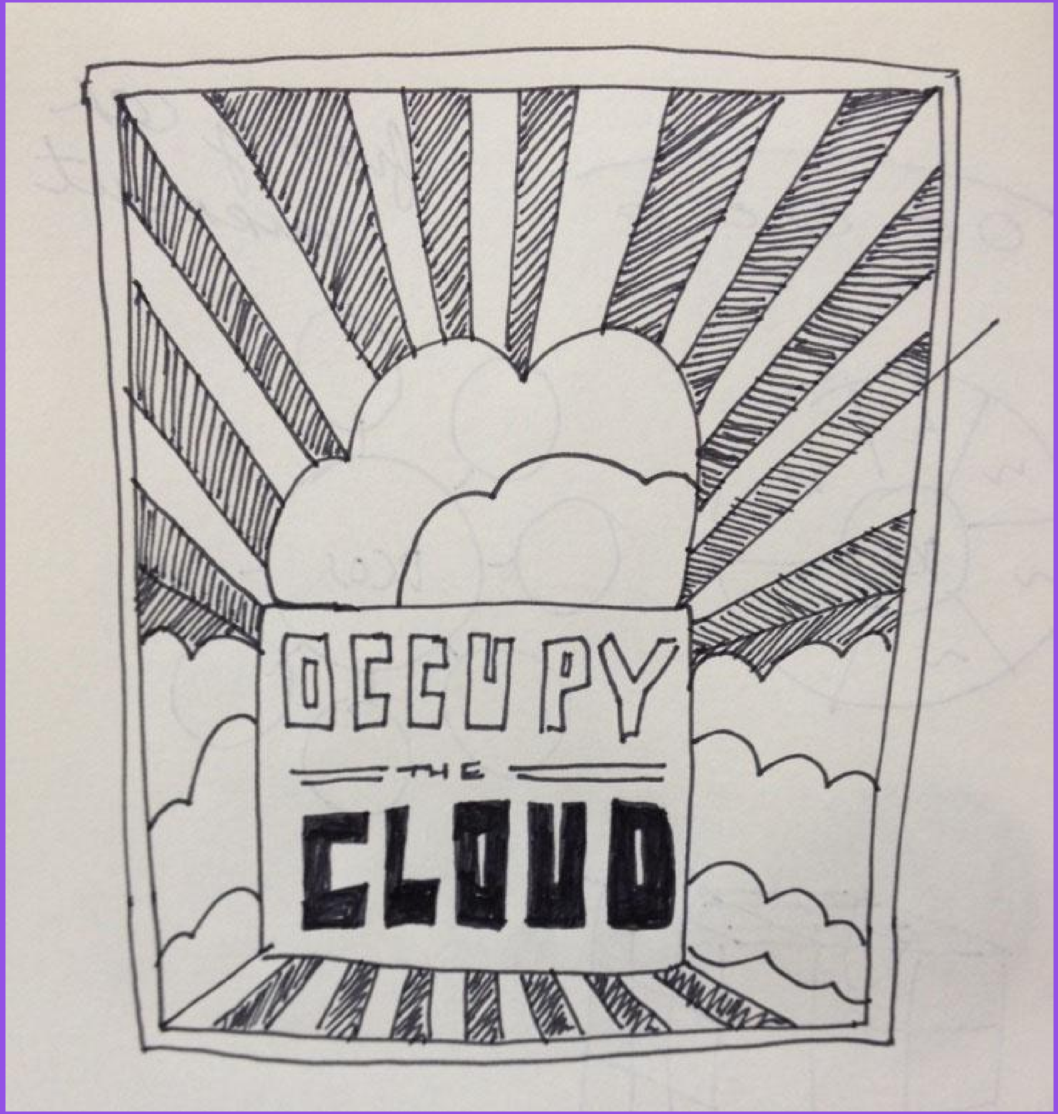
Dan, our operator, is forced to partition the datacenter to accommodate these demands. **Utilisation suffers**.

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

He has to deal with dependencies **on a one-off basis** for each of his developers' applications.



© 2015 Mesosphere, Inc. All Rights Reserved.



MESOSPHERE & THE DCOS

MISSION TO THE MESOSPHERE

April 2013

MESOSPHERE

Mesosphere is formed by engineers who have been using Mesos at Twitter and AirBnB.

June 2013

MESOS GRADUATES

Mesos graduates from the Apache Incubator to become a top level project.

April 2015

APPLE ANNOUNCES J.A.R.V.I.S.

Apple announces that the Siri infrastructure now runs on Mesos, atop “thousands” of nodes.

August 2015

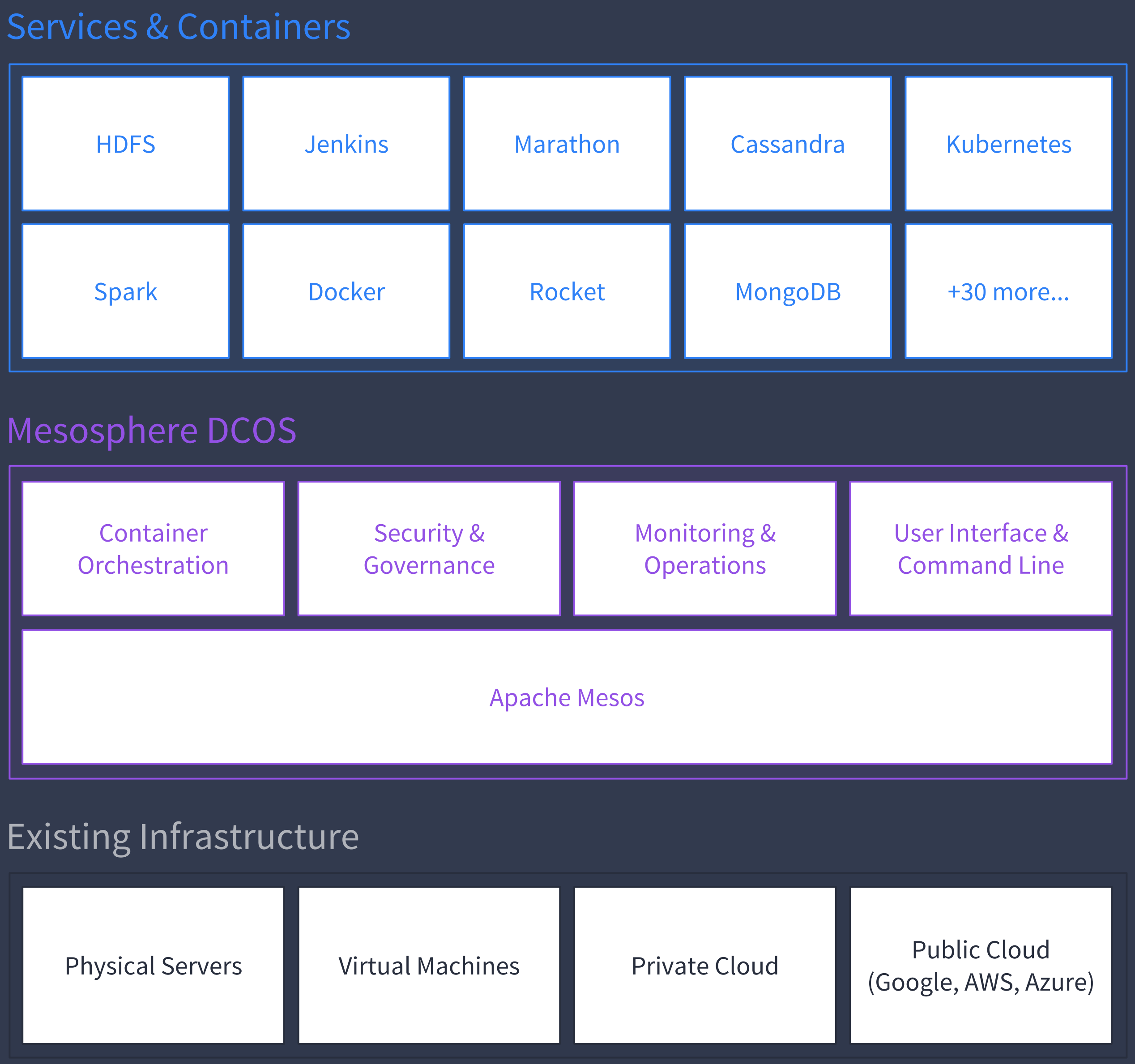
VERIZON SCALE DEMO

Verizon demonstrates launching 50,000 containers in less than 90 seconds using Mesos and Mesosphere’s Marathon scheduler.

© 2015 Mesosphere, Inc. All Rights Reserved.

The Vision

MESOSPHERE DCOS



© 2015 Mesosphere, Inc. All Rights Reserved.

THE DATACENTER OPERATING SYSTEM

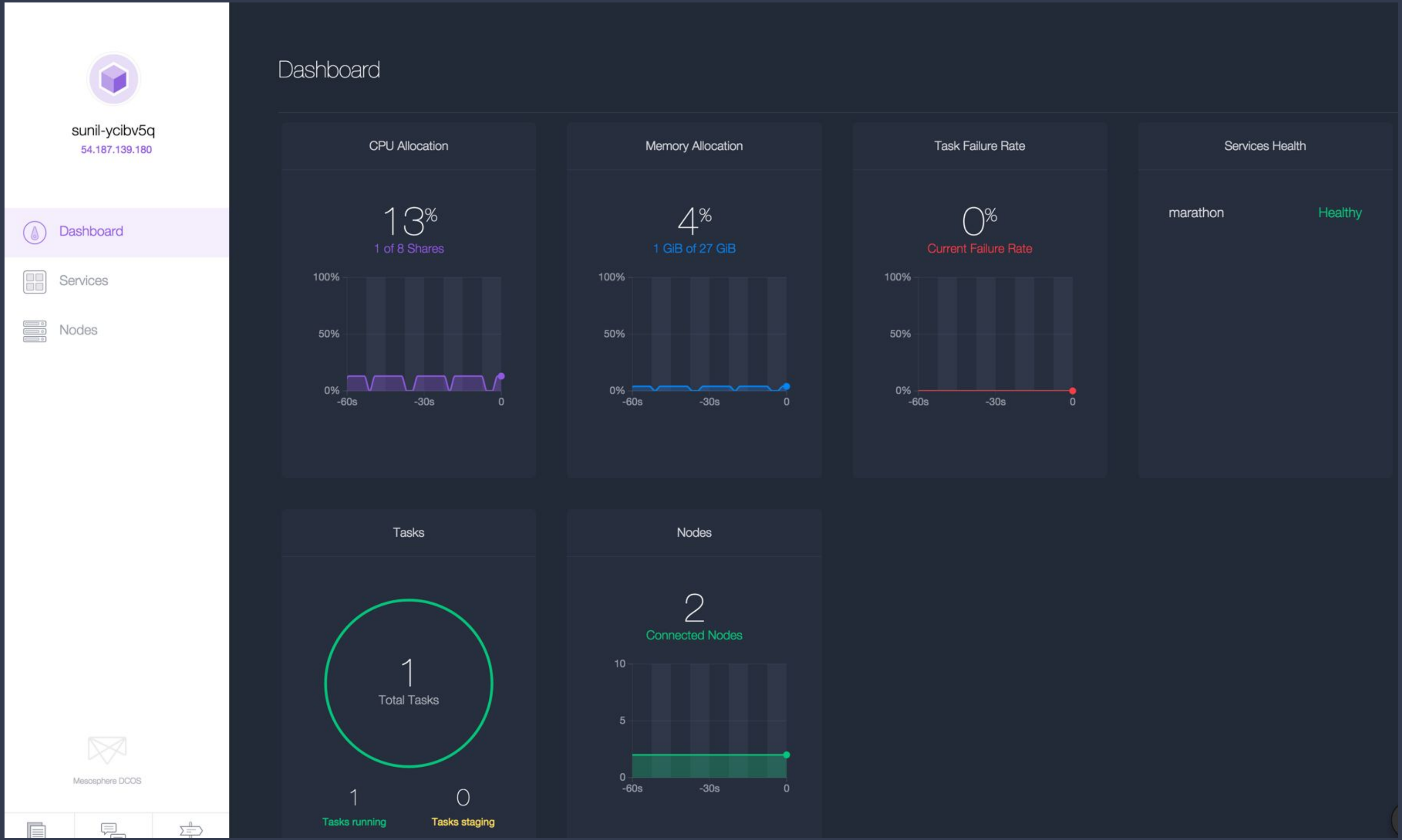
DCOS aims to make developing & deploying distributed apps easier.

Short-term:

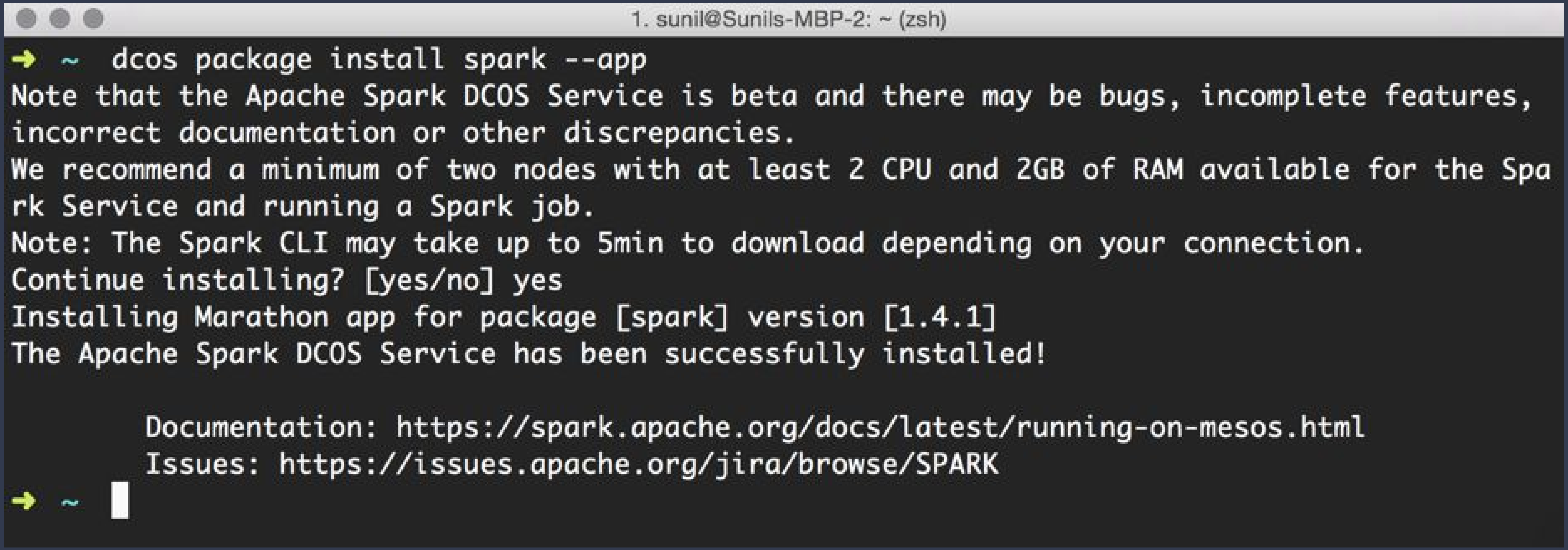
- Software installation/removal
- Seamless upgrades
- Automatic failure detection, reconciliation

© 2015 Mesosphere, Inc. All Rights Reserved.

A UNIFIED INTERFACE TO THE DATACENTER



THE COMMAND LINE TO THE DATACENTER


1. sunil@Sunils-MBP-2: ~ (zsh)

```
➜  ~  dcos package install spark --app
Note that the Apache Spark DCOS Service is beta and there may be bugs, incomplete features,
incorrect documentation or other discrepancies.
We recommend a minimum of two nodes with at least 2 CPU and 2GB of RAM available for the Spa
rk Service and running a Spark job.
Note: The Spark CLI may take up to 5min to download depending on your connection.
Continue installing? [yes/no] yes
Installing Marathon app for package [spark] version [1.4.1]
The Apache Spark DCOS Service has been successfully installed!

        Documentation: https://spark.apache.org/docs/latest/running-on-mesos.html
        Issues: https://issues.apache.org/jira/browse/SPARK
➜  ~ █
```

© 2015 Mesosphere, Inc. All Rights Reserved.

PRODUCTION CUSTOMERS AND MESOS USERS

Bloomberg

verizon

NETFLIX

2σ TWO SIGMA

PayPal

ebay

airbnb

HubSpot

Government Agencies

yelp

© 2015 Mesosphere, Inc. All Rights Reserved.

WHAT WILL IT TAKE TO MAKE DAN HAPPY?

© 2015 Mesosphere, Inc. All Rights Reserved.

CONTAINERS EVERYWHERE

Many Mesos tasks run in **containers**:

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

© 2015 Mesosphere, Inc. All Rights Reserved.

CONTAINERS EVERYWHERE

Many Mesos tasks run in **containers**:

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

Containers use standard linux features to create an isolated execution environment:

- kernel namespaces
  - process isolation
- control groups (cgroups)
  - resource isolation
- chroot
  - filesystem isolation
- seccomp
  - restricted kernel access

© 2015 Mesosphere, Inc. All Rights Reserved.

CONTAINERS EVERYWHERE

Containers also help Dan solve his dependency problem by giving tasks everything they need to run.

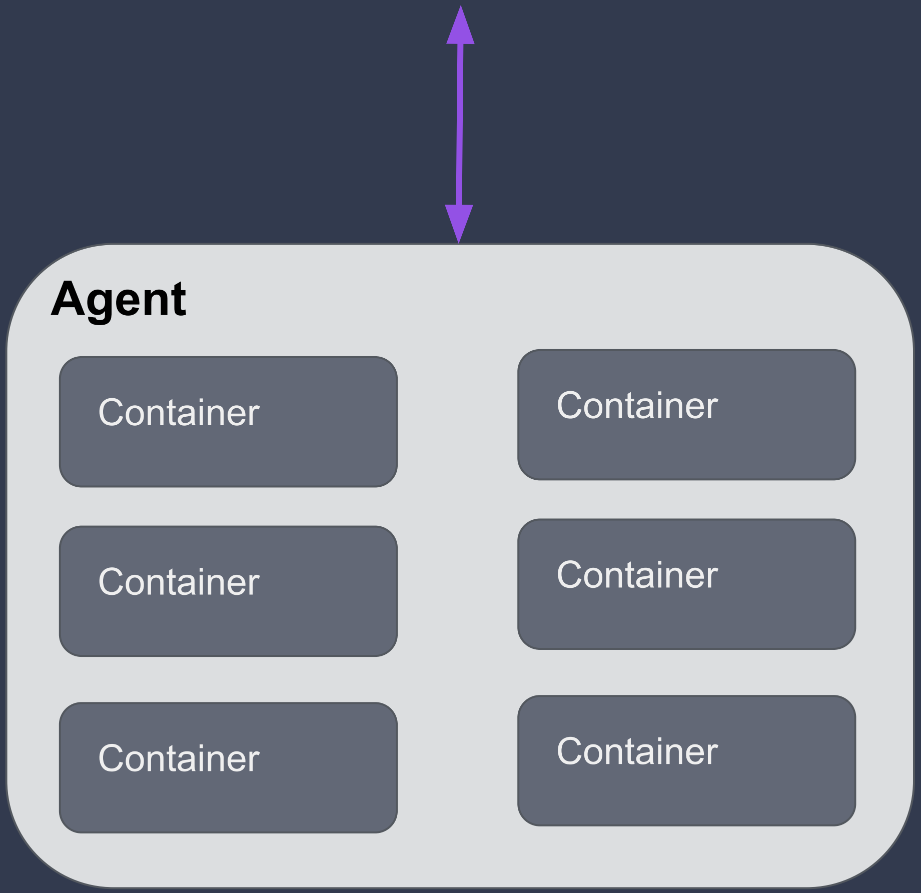
© 2015 Mesosphere, Inc. All Rights Reserved.

CONTAINER NETWORKING

Containers isolate tasks on the agent, but what about their communication?

The status quo in a Mesos cluster: one IP per agent.

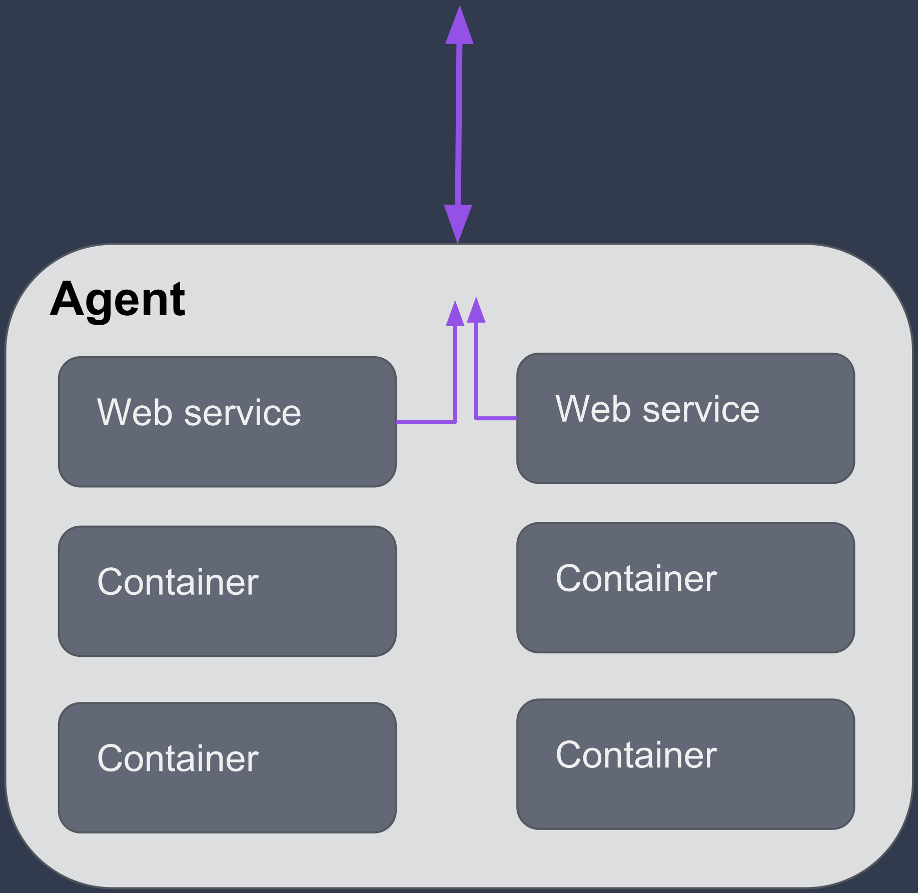
Many containers per agent: they must share a single IP.



CONTAINER NETWORKING

This causes headaches:

- Port conflicts
- Security compromises
- Performance
- Service discovery

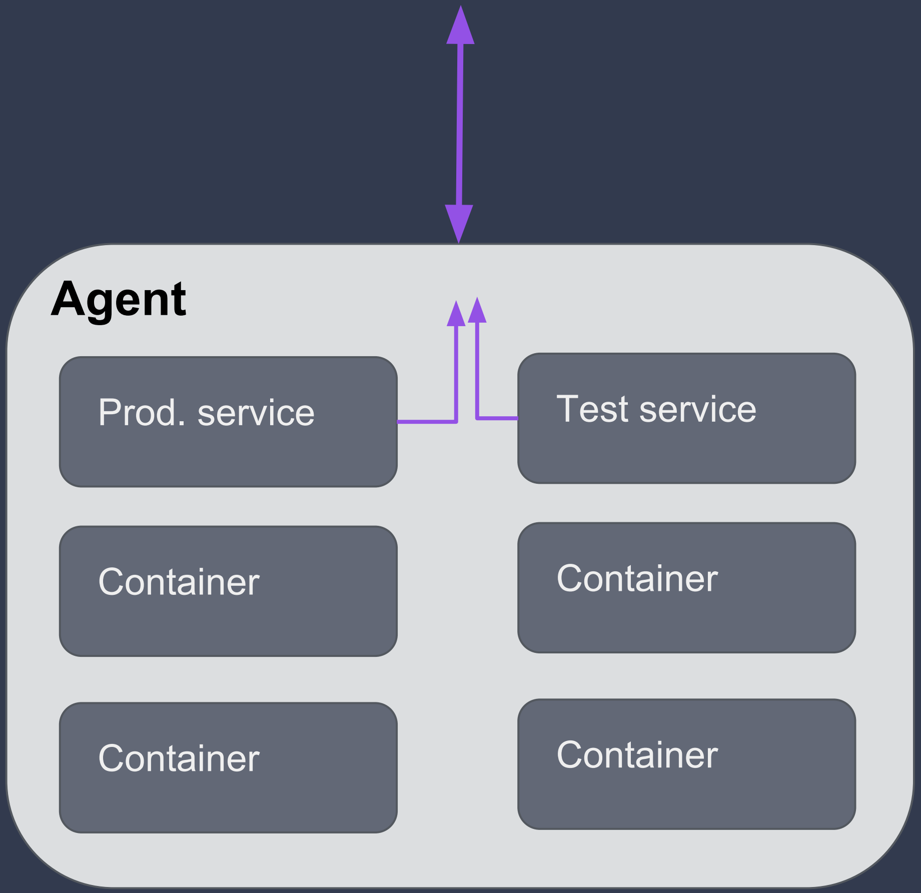


© 2015 Mesosphere, Inc. All Rights Reserved.

CONTAINER NETWORKING

This causes headaches:

- Port conflicts
- Security compromises
- Performance
- Service discovery

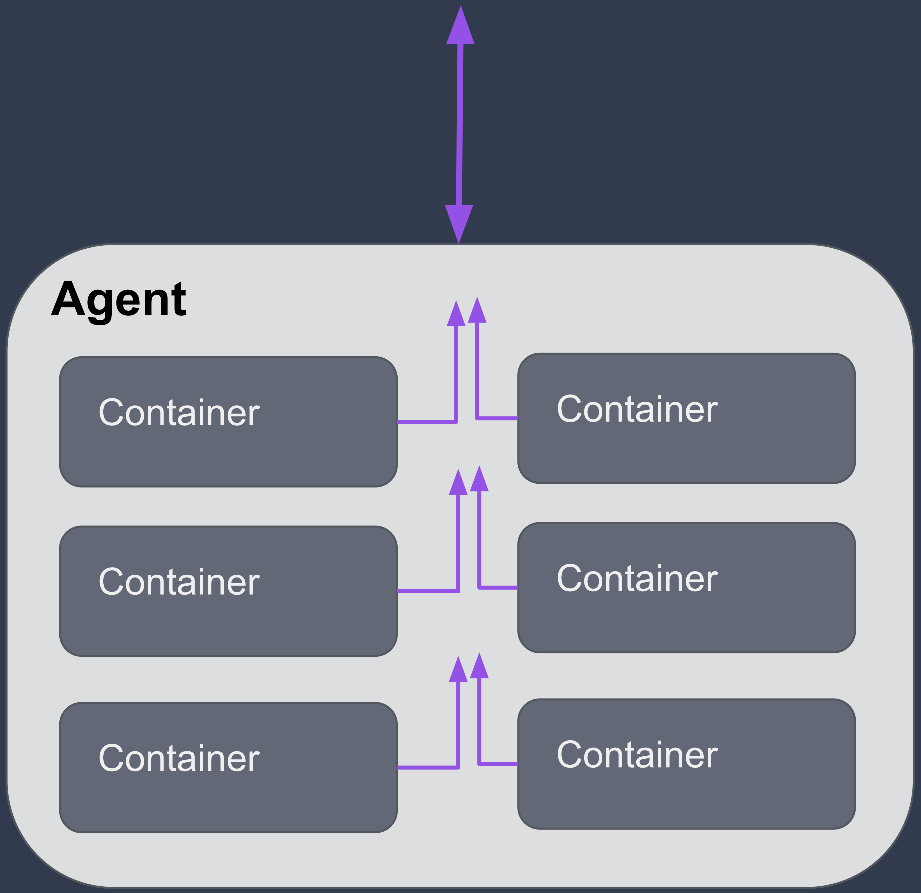


© 2015 Mesosphere, Inc. All Rights Reserved.

CONTAINER NETWORKING

This causes headaches:

- Port conflicts
- Security compromises
- Performance
- Service discovery

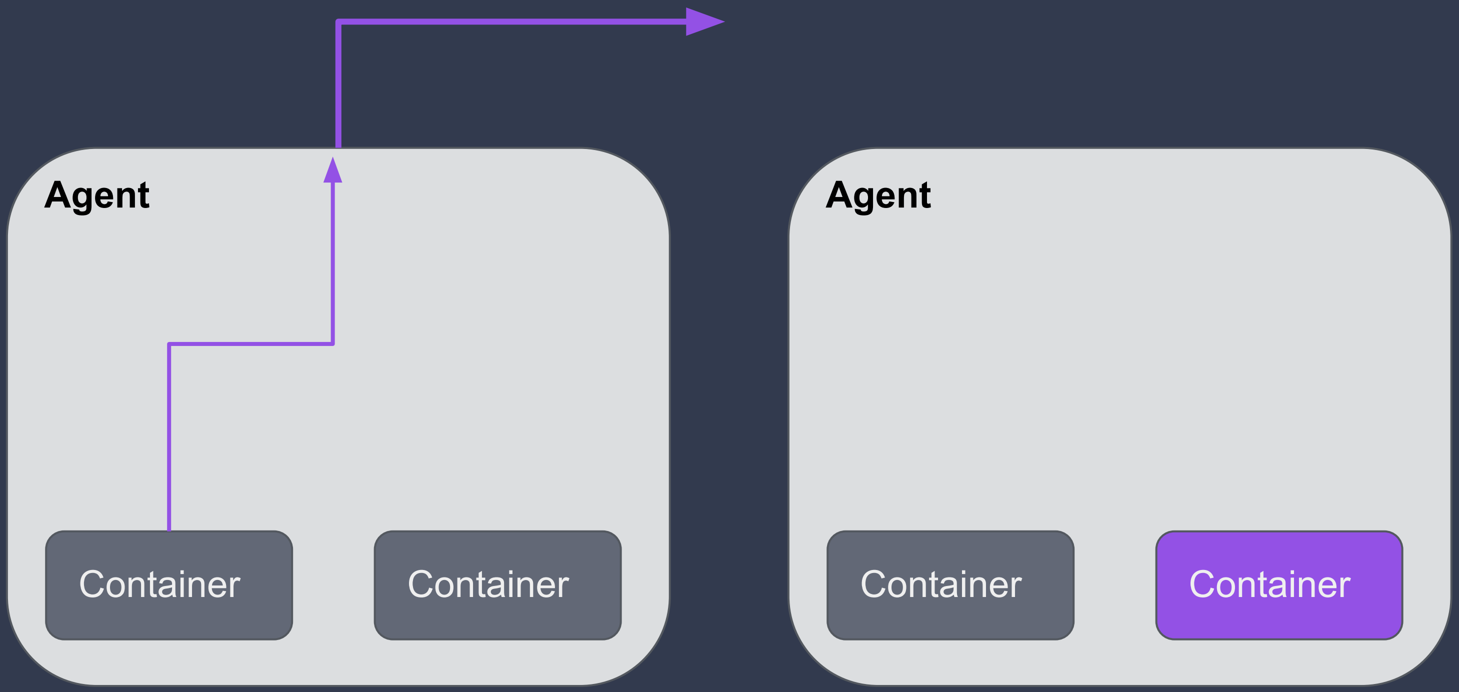


© 2015 Mesosphere, Inc. All Rights Reserved.

CONTAINER NETWORKING

This causes headaches:

- Port conflicts
- Security compromises
- Performance
- Service discovery



© 2015 Mesosphere, Inc. All Rights Reserved.

NETWORK ISOLATION

Segregating containers’ network traffic can solve these problems in an elegant, maintainable way.

© 2015 Mesosphere, Inc. All Rights Reserved.

NETWORK ISOLATION

Segregating containers’ network traffic can solve these problems in an elegant, maintainable way.

Implemented as Mesos modules:

- Project Calico
- Port-mapping isolation
- …

© 2015 Mesosphere, Inc. All Rights Reserved.

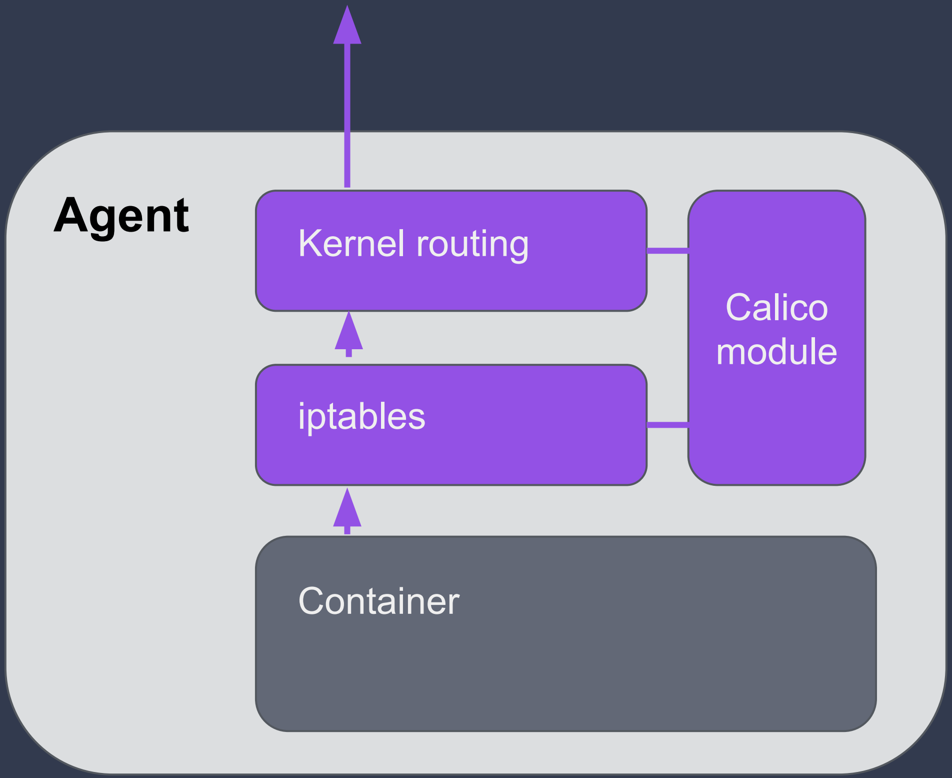
CALICO NETWORK ISOLATION

Calico Network Virtualizer & IP Address Manager:

- Pure Layer-3 solution
- Uses linux features to route container traffic
- Provides security policies
- Advertises routes to local containers via BGP
- Can assign IP-per-container

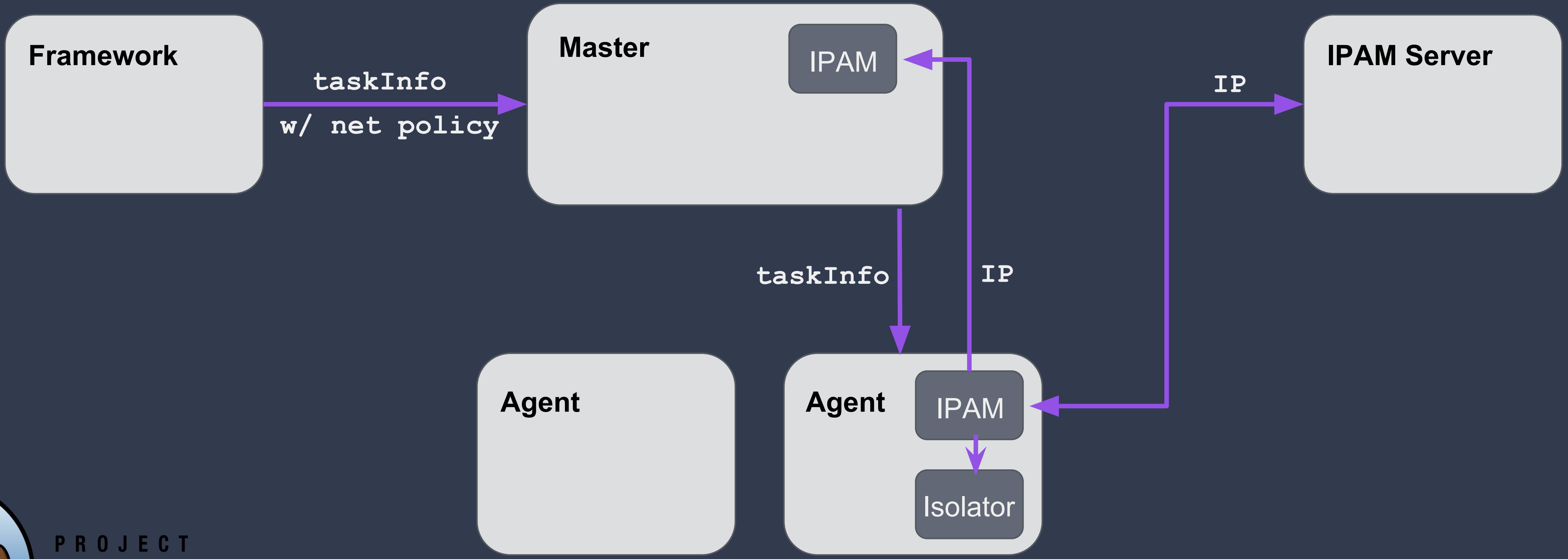
© 2015 Mesosphere, Inc. All Rights Reserved.

CALICO NETWORK ISOLATION

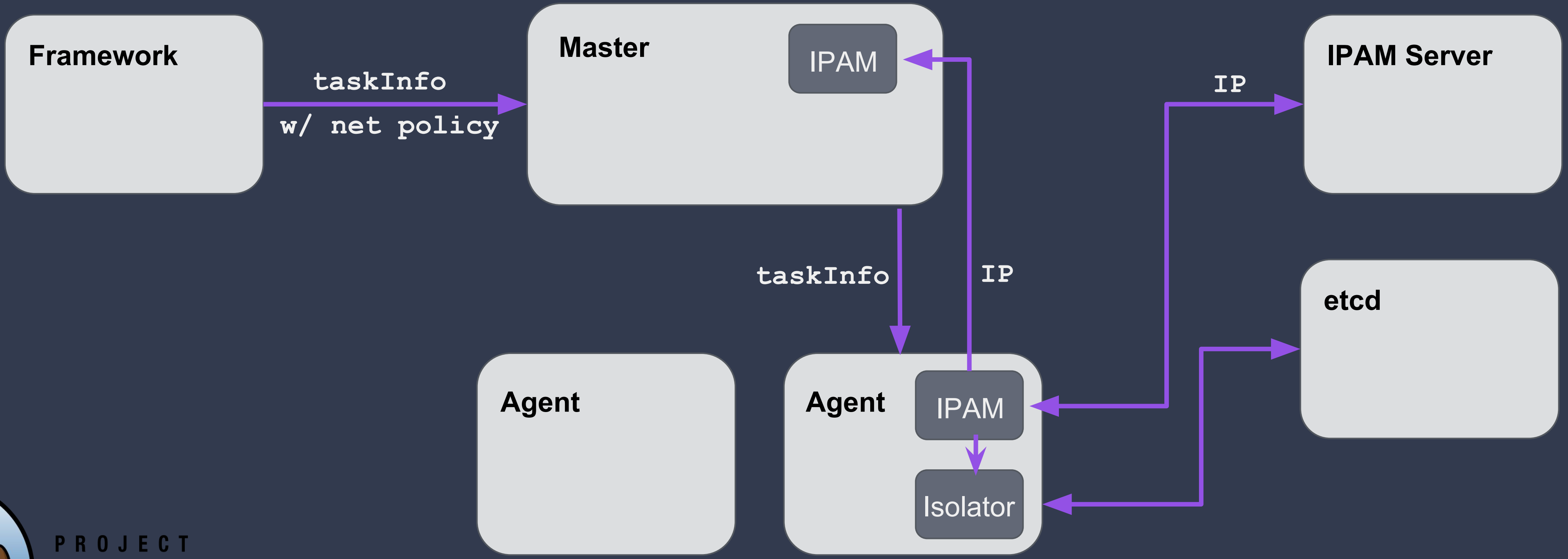


© 2015 Mesosphere, Inc. All Rights Reserved.

CALICO NETWORK ISOLATION



CALICO NETWORK ISOLATION



© 2015 Mesosphere, Inc. All Rights Reserved.

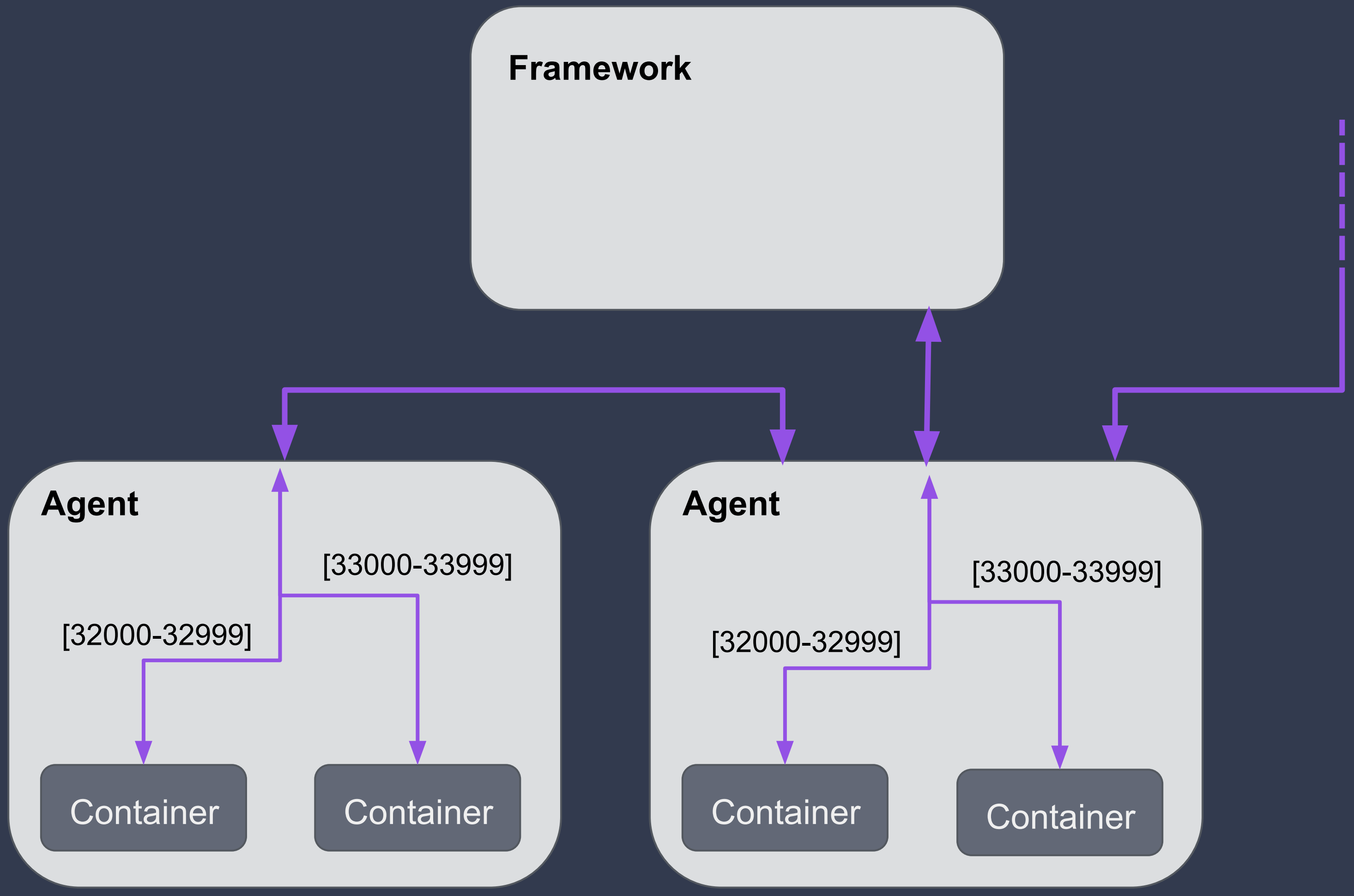
NETWORK ISOLATION

What if I don’t have enough IPs to go around?

© 2015 Mesosphere, Inc. All Rights Reserved.

PORT-MAPPING ISOLATOR

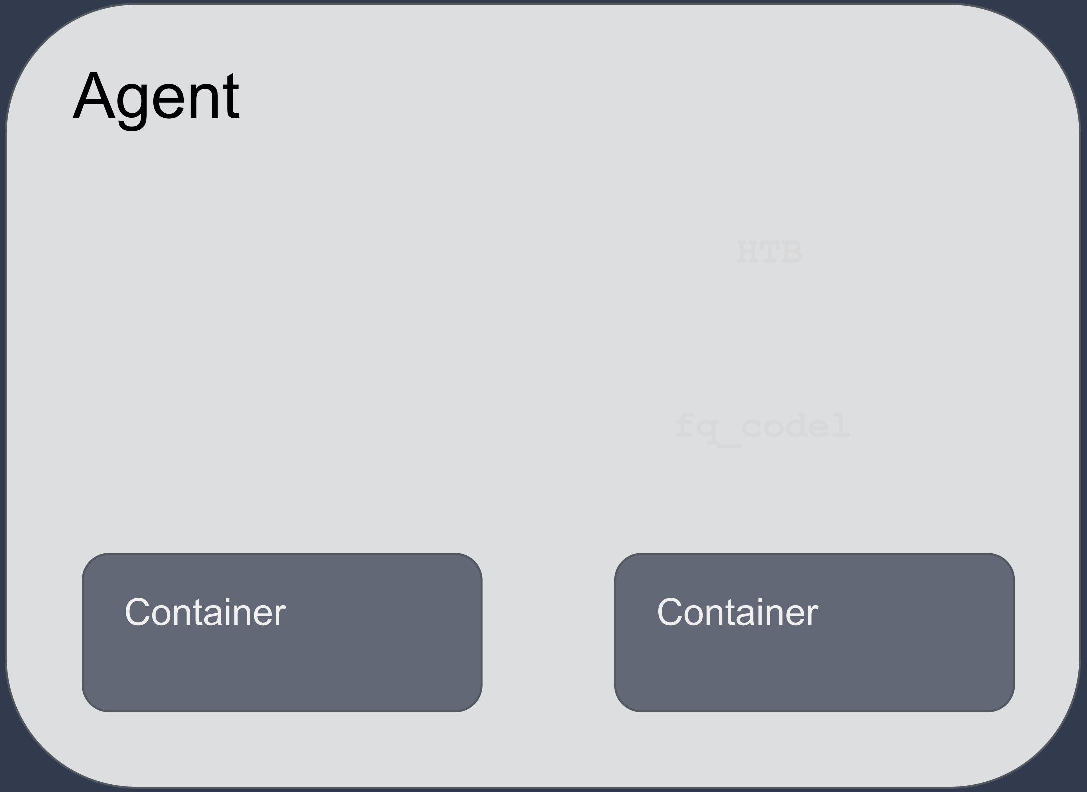
- Ports distributed amongst containers on each agent
- Network traffic routed by port using TC rules
- Implemented with `libnl` (via netlink messages)
- Ports assigned and tracked via scheduler (ex: Aurora)



© 2015 Mesosphere, Inc. All Rights Reserved.

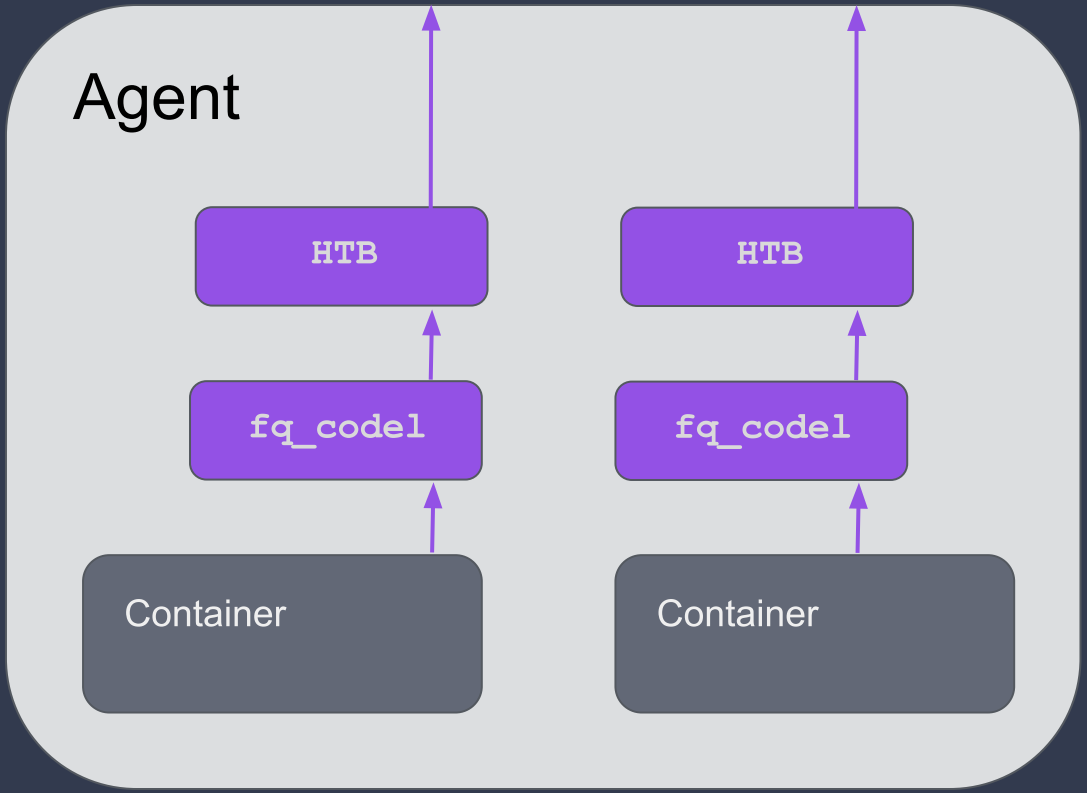
PORT-MAPPING ISOLATOR

What about performance?



PORT-MAPPING ISOLATOR

- `fq_codel` defines discrete network flows for containers
- Separate flows prevent buffer bloat
- Hierarchical token bucket (HTB) employed to limit bandwidth

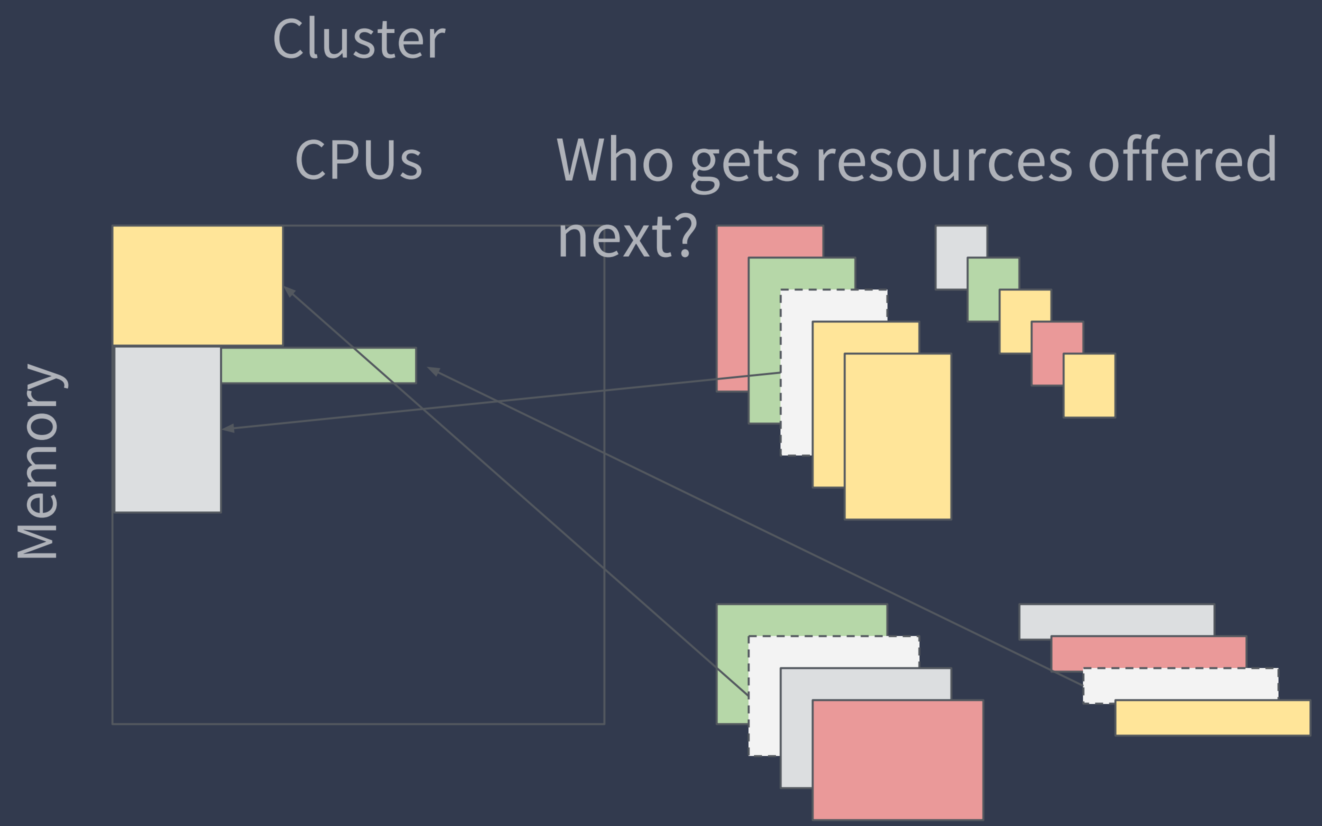


© 2015 Mesosphere, Inc. All Rights Reserved.

WORKLOADS CHANGED SINCE 2009

© 2015 Mesosphere, Inc. All Rights Reserved.

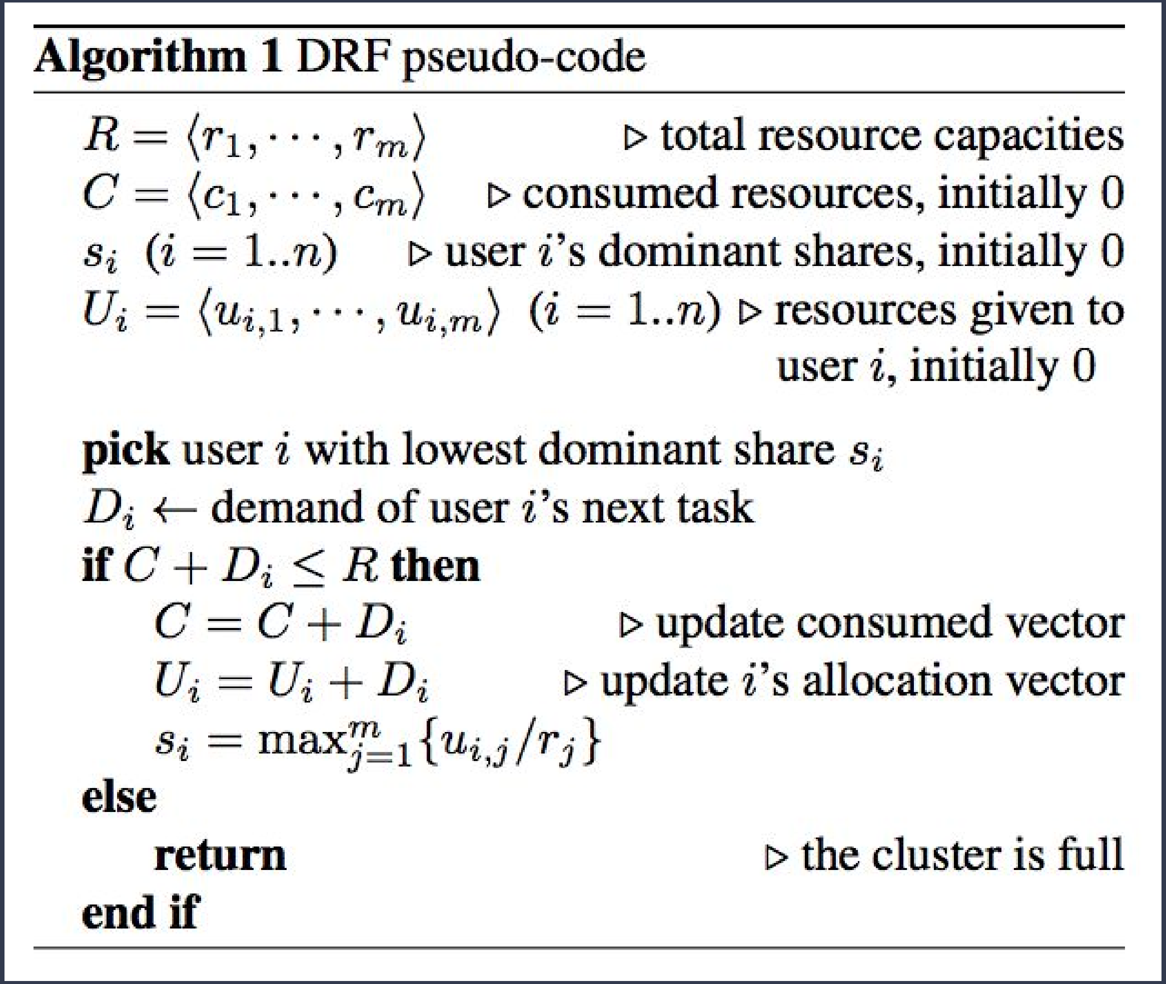
FAIRNESS FOR MULTI-DIMENSIONAL RESOURCES?



Who gets resources offered next?

While we can do fitting, how do we express fairness over different units and dimensions?

DOMINANT RESOURCE FAIRNESS



Algorithm 1 DRF pseudo-code

```
$R = \langle r_1, \cdots, r_m \rangle$    ▷ total resource capacities
$C = \langle c_1, \cdots, c_m \rangle$    ▷ consumed resources, initially 0
$s_i \ (i = 1..n)$    ▷ user $i$'s dominant shares, initially 0
$U_i = \langle u_{i,1}, \cdots, u_{i,m} \rangle \ (i = 1..n)$    ▷ resources given to user $i$, initially 0

pick user $i$ with lowest dominant share $s_i$
$D_i \leftarrow$ demand of user $i$'s next task
if $C + D_i \leq R$ then
    $C = C + D_i$    ▷ update consumed vector
    $U_i = U_i + D_i$    ▷ update $i$'s allocation vector
    $s_i = \max_{j=1}^{m}\{u_{i,j}/r_j\}$
else
    return    ▷ the cluster is full
end if
```

TIME DIMENSIONALITY HAS CHANGED!

Multitenancy now expands domains of multiple batch schedulers with a mix of:

- Long lived services
- Storage services
- Short lived analytics tasks

Is extreme fairness what you really want?

© 2015 Mesosphere, Inc. All Rights Reserved.

SEVERAL POs AT CUSTOMER SITES

“My framework is starved! Why isn’t my framework receiving any resources?”

- Some frameworks has a lot of work to do, others less. All gets a fair share by default.
- Configuration is hard with weights, static reservations, etc

© 2015 Mesosphere, Inc. All Rights Reserved.

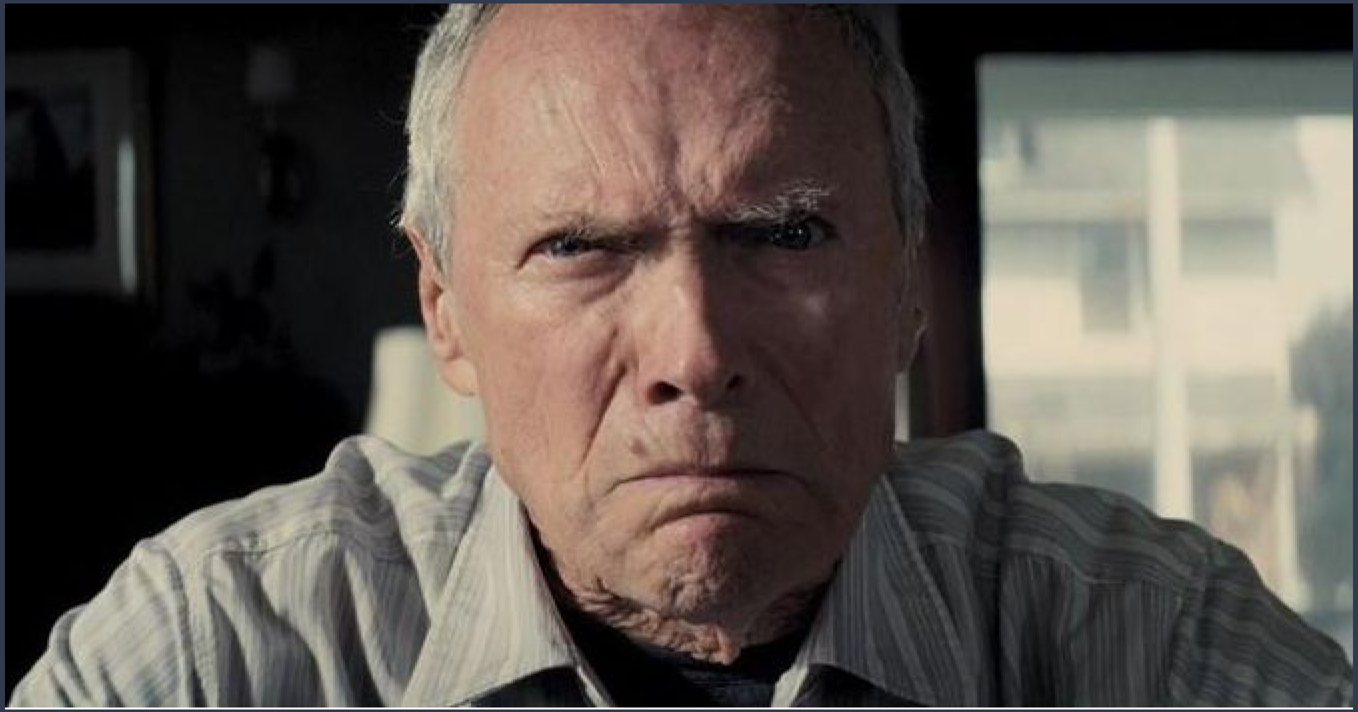
EASY TO MAKE MISTAKES IN SCHEDULER IMPLEMENTATIONS

```
resourceOffers(offers) {
  ...
  if (is_ok(offer)) {
    launchTasks(offer);
  }
}
```

```
resourceOffers(offers) {
  ...
  if (is_ok(offer)) {
    launchTasks(offer);
  } else {
    declineOffer(offer);
  }
}
```

© 2015 Mesosphere, Inc. All Rights Reserved.

POs MAKE DAN UNHAPPY



WHAT DID TWITTER DO?

- Uses Apache Aurora for most of its operations
- It implemented preemption assuming it was the only scheduler available

MULTI-TENANCY BECOMES TOO RISKY FOR CRITICAL SYSTEMS

- Companies partition Mesos cluster into many smaller Mesos clusters
- Run multiple copies of the same framework on top of Mesos
- Avoid running multiple frameworks all together
- That was surely not the intent

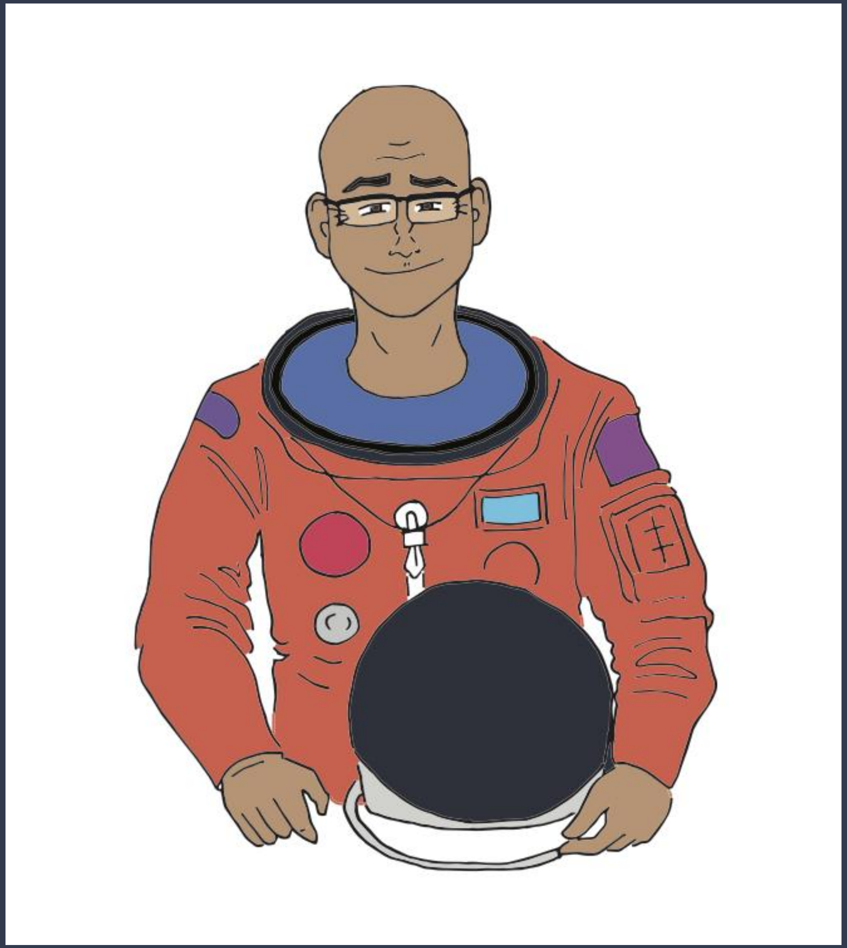
© 2015 Mesosphere, Inc. All Rights Reserved.

IN THE WORKS

- Quotas ensure minimum set of resources for frameworks
- Optimistic offers enables resource parallelism
- Cooperative preemption through Inverse offers

© 2015 Mesosphere, Inc. All Rights Reserved.

LET’S ASSUME DAN IS HAPPY WITH HIS CLUSTER

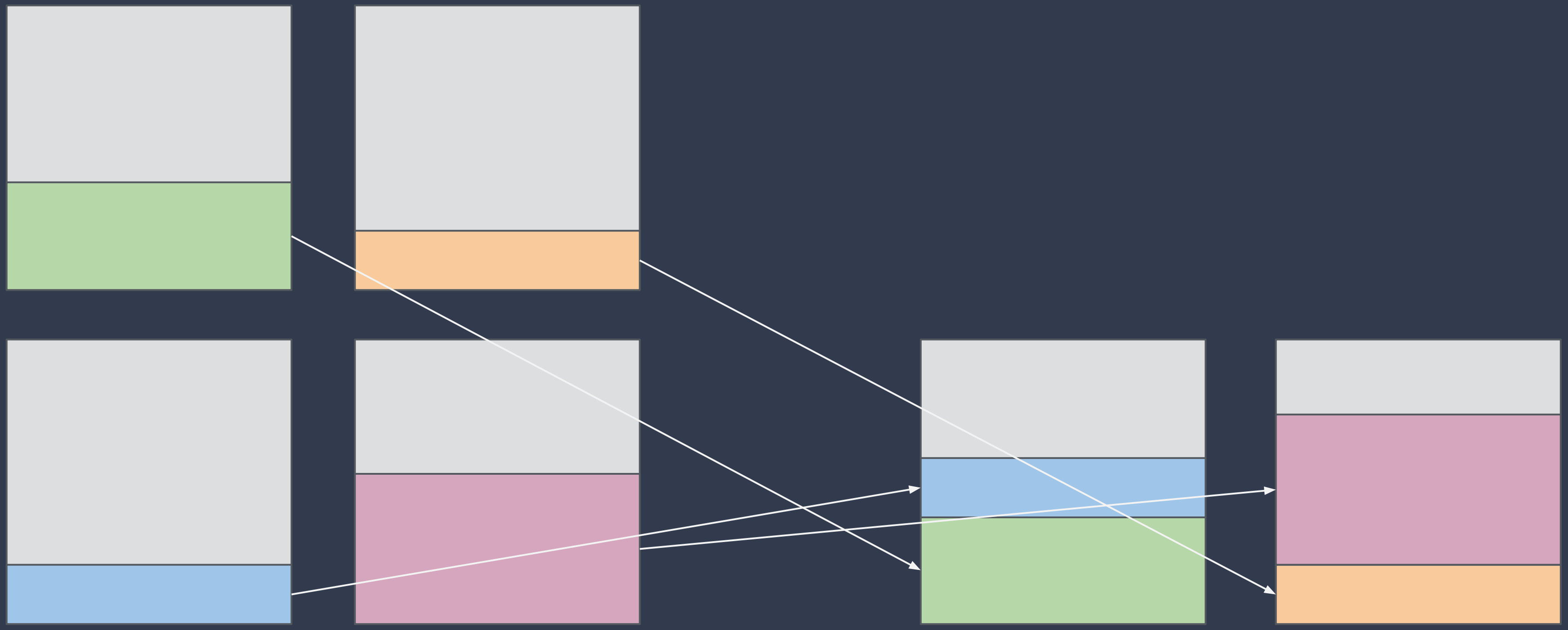


THAT MAKES HIS BOSS HAPPY

© 2015 Mesosphere, Inc. All Rights Reserved.

AND THEIR CFO IS HAPPY TOO

MESOS HELPS REDUCE WASTED RESOURCES

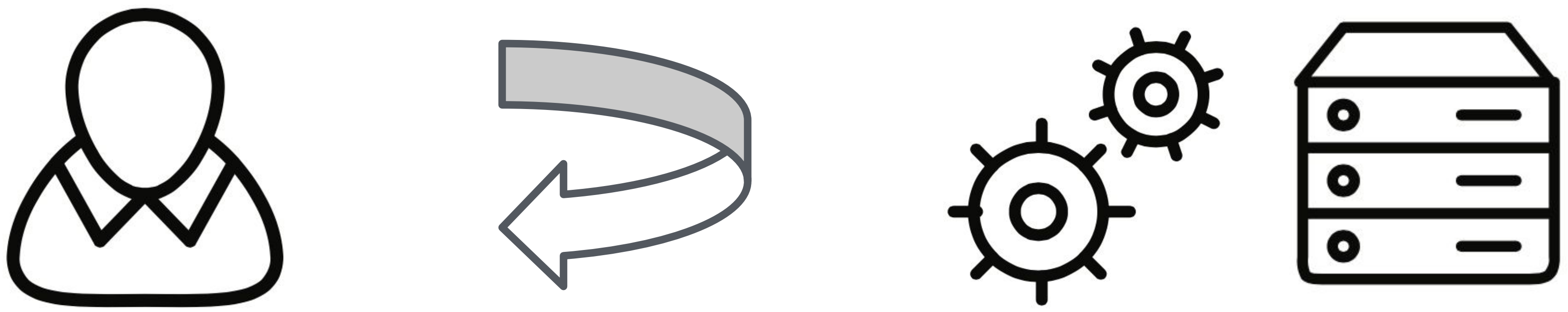


© 2015 Mesosphere, Inc. All Rights Reserved.

ESTIMATING RESOURCES IS HARD

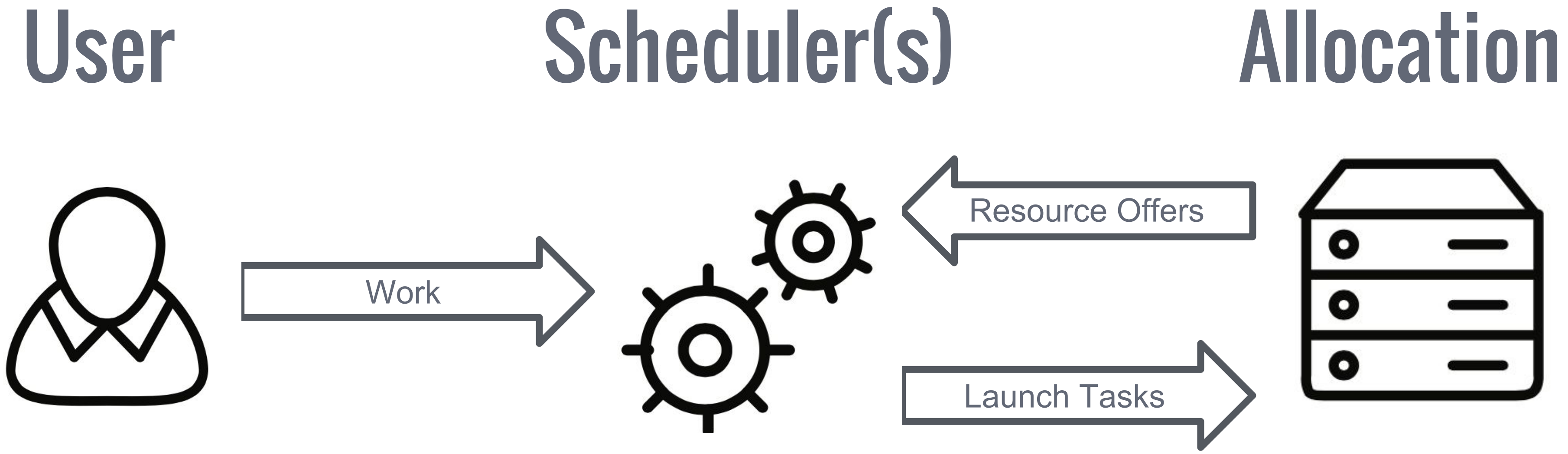
© 2015 Mesosphere, Inc. All Rights Reserved.

MESOS ENABLES MULTIPLE SCHEDULER ALGORITHMS



“Please run container X on Y resources”

MESOS ENABLES MULTIPLE SCHEDULER ALGORITHMS



© 2015 Mesosphere, Inc. All Rights Reserved.

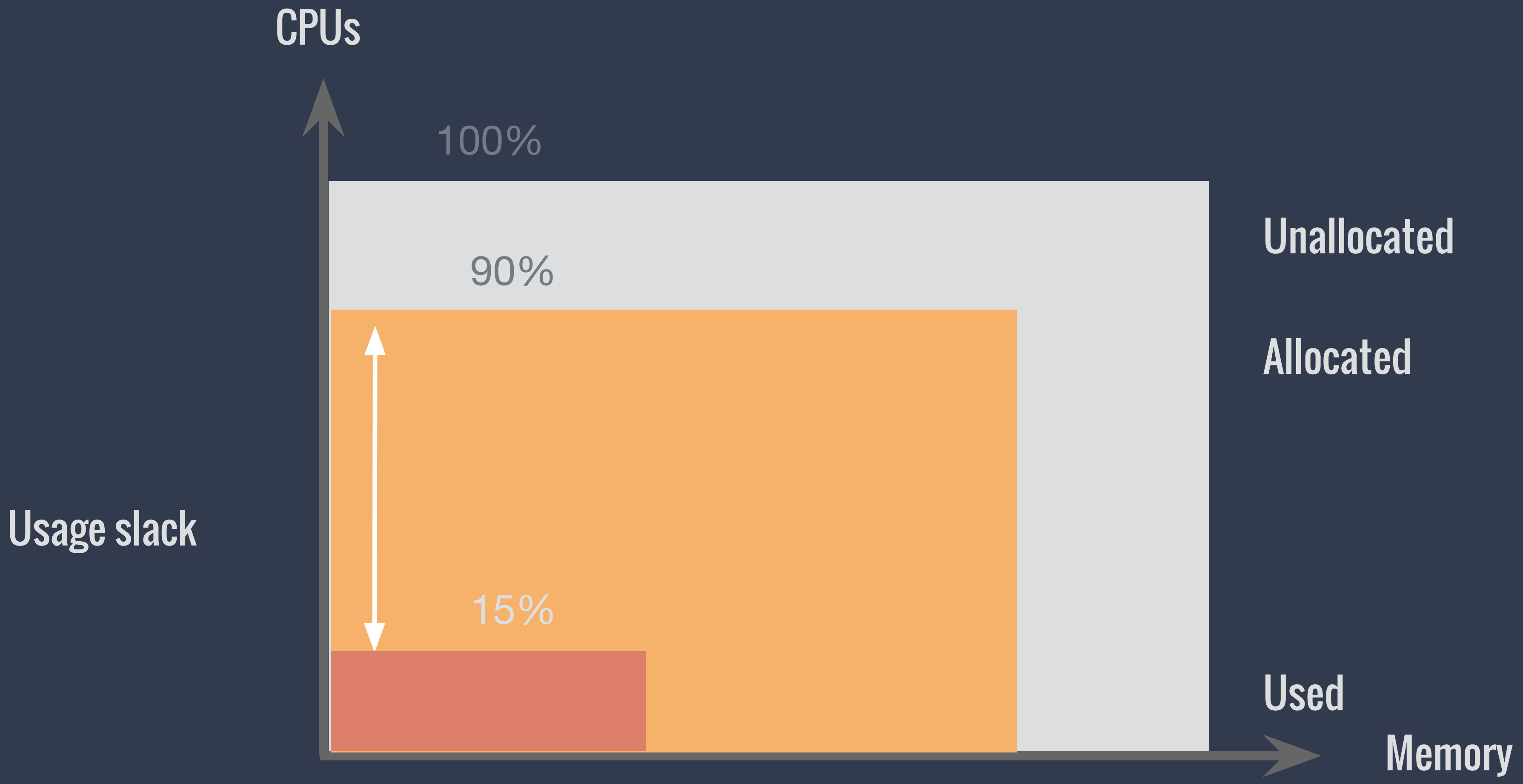
RESOURCES REPRESENT ALLOCATION

Resources represent allocation

How are users supposed to know how many resources their workload requires?

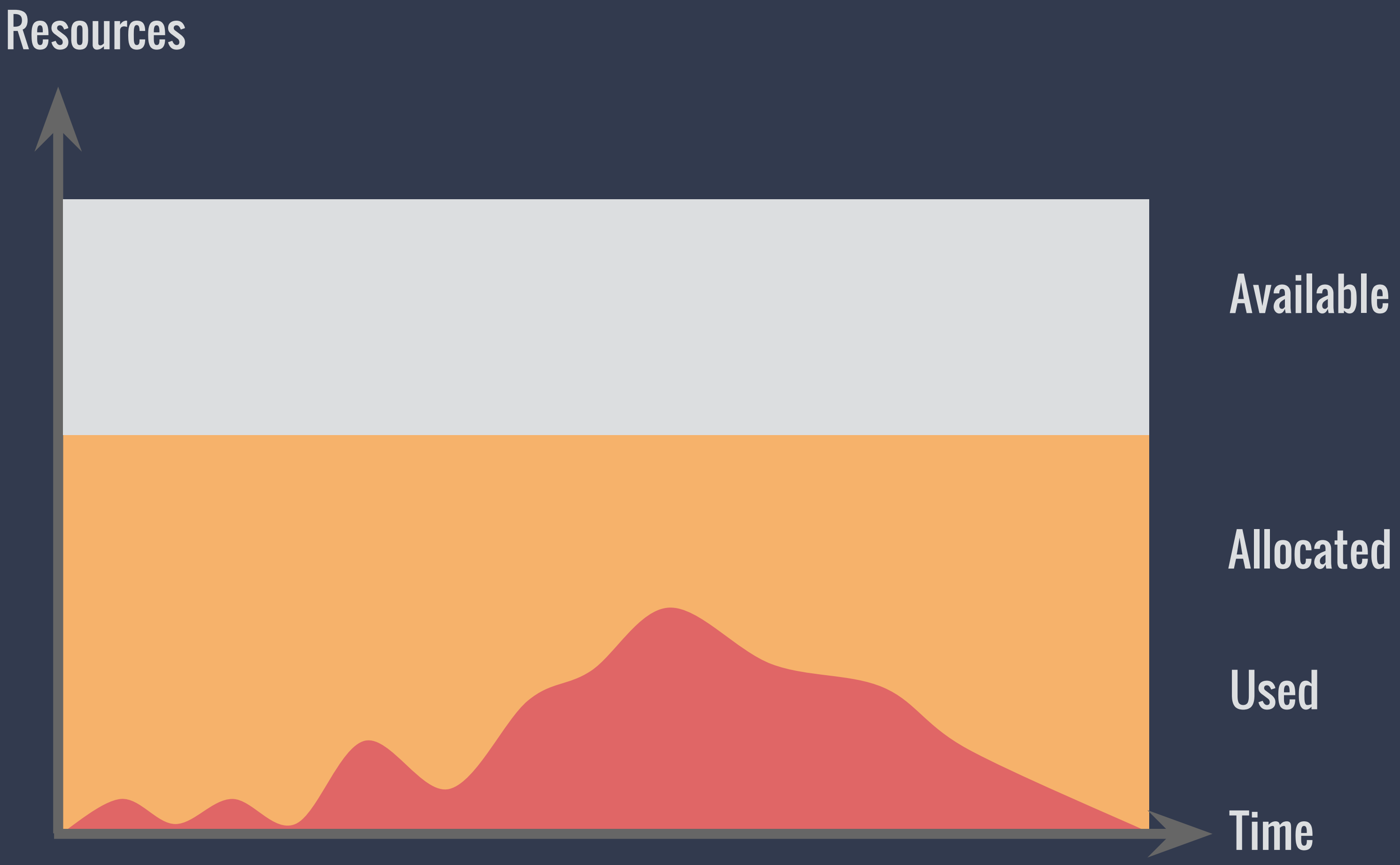
© 2015 Mesosphere, Inc. All Rights Reserved.

RESOURCES REPRESENT ALLOCATION



© 2015 Mesosphere, Inc. All Rights Reserved.

USAGE SLACK HURTS UTILISATION

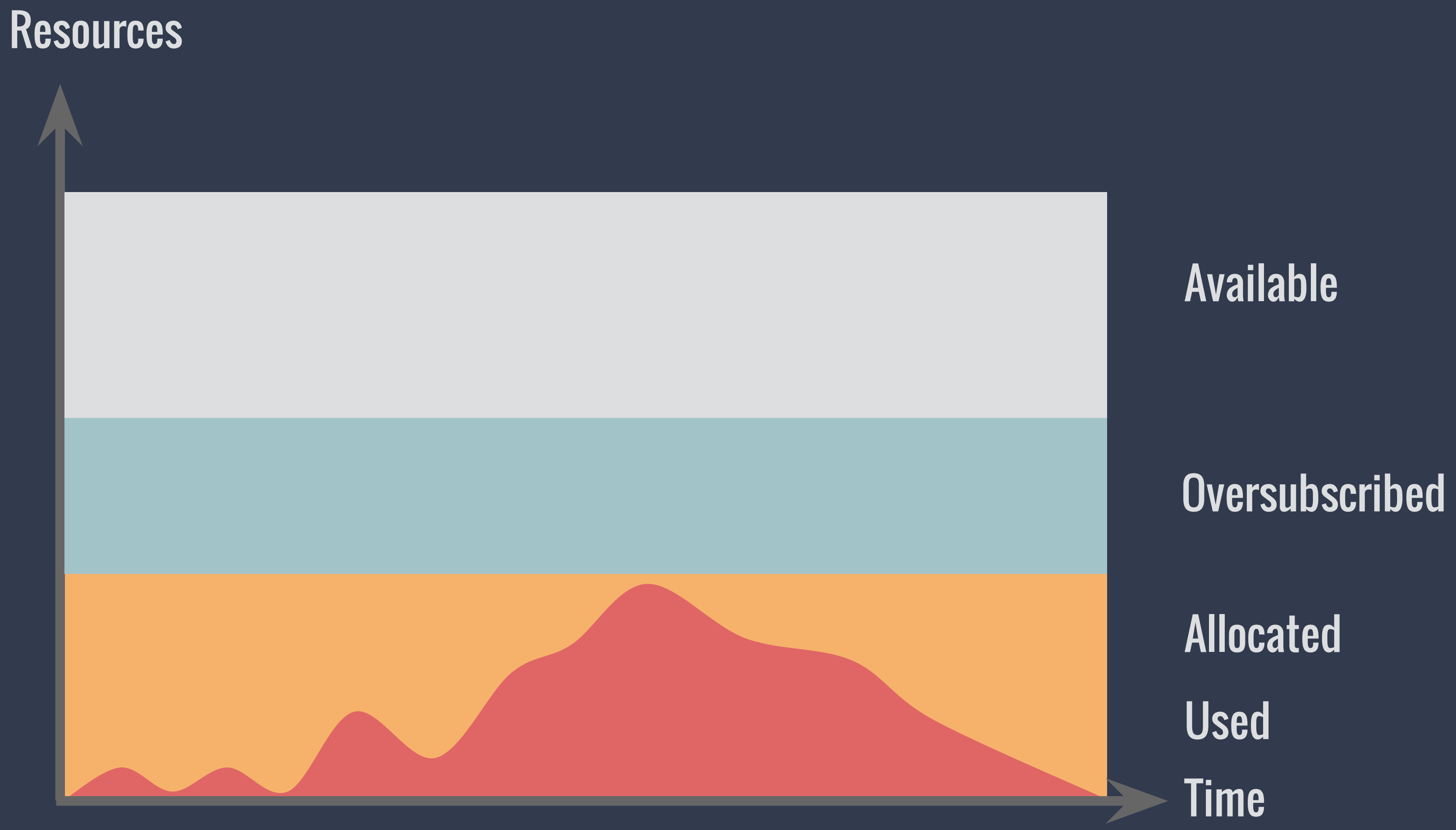


© 2015 Mesosphere, Inc. All Rights Reserved.

FIRST STEPS TOWARDS IMPROVED UTILISATION

© 2015 Mesosphere, Inc. All Rights Reserved.

OVERSUBSCRIPTION ENABLES TASKS TO RUN ON SLACK



© 2015 Mesosphere, Inc. All Rights Reserved.

TWO COMPONENTS ENABLE OVERSUBSCRIPTION

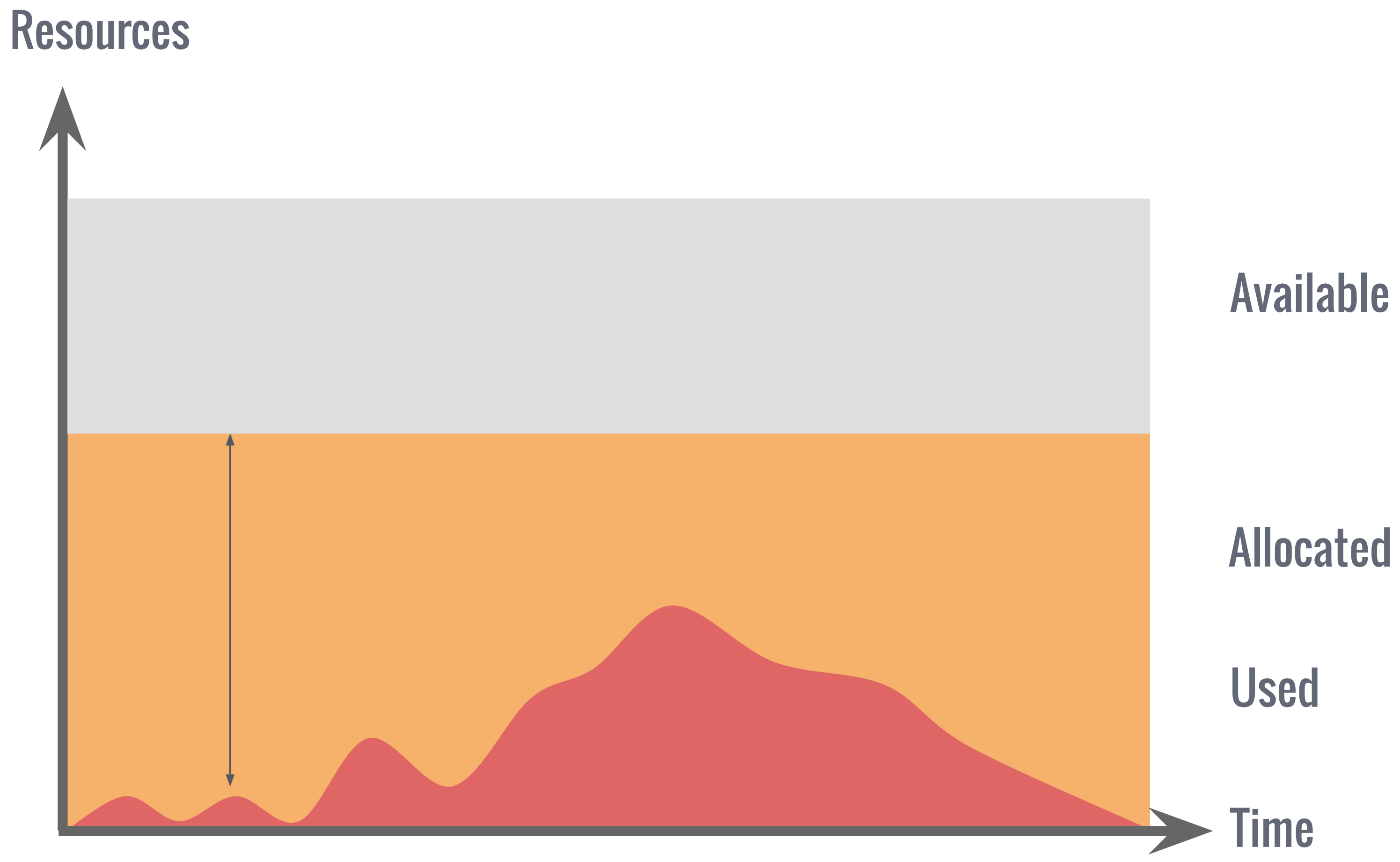
Resource Estimator

Quality of Service Controller

© 2015 Mesosphere, Inc. All Rights Reserved.

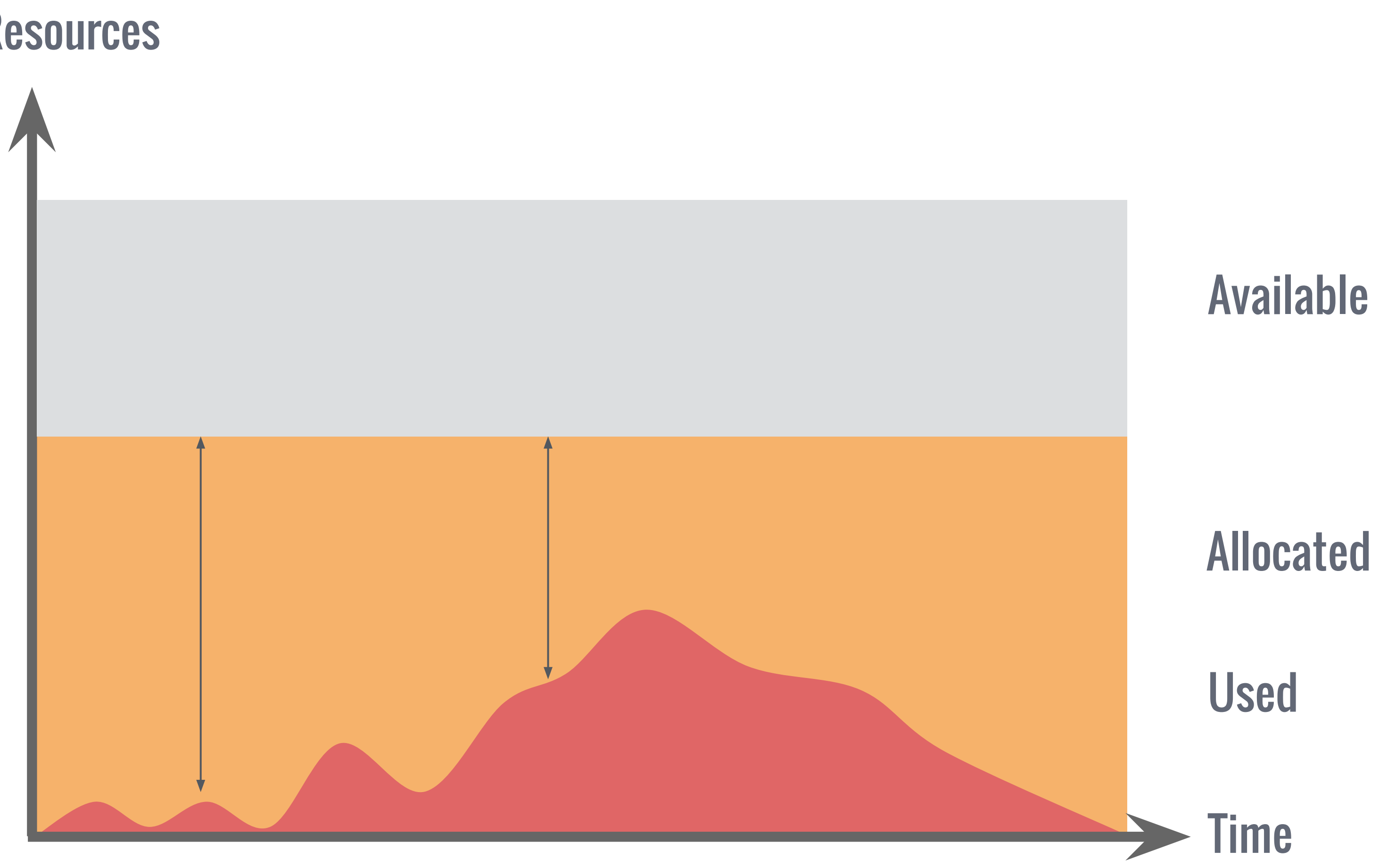
ESTIMATING OVERSUBSCRIBABLE RESOURCES

How many resources should be oversubscribed?



WHAT DO WE DO ABOUT MISPREDICTIONS?

Now, what happens when things change?



© 2015 Mesosphere, Inc. All Rights Reserved.

THE QoS CONTROLLER

- Can shut down best effort containers
- In the future, it will be able to correct by
  - Freezing
  - Throttling
  - Resizing
  - Cooperating with the framework

© 2015 Mesosphere, Inc. All Rights Reserved.

MANY RESOURCES CANNOT BE ISOLATED

- Logical units on the chip
- Last level caches
- Memory bandwidth
- I/O
- Chip power supply

© 2015 Mesosphere, Inc. All Rights Reserved.

OVERSUBSCRIPTION WITH INTEL: SERENITY

https://goo.gl/jWtu7V

WRAPPING UP

- Mesos is being used in production at huge scale
- It forms the core of an operating system for the datacenter
- Lots of exciting work yet to do!

Slides at http://mesosphere.github.io/presentations

(P.S., we’re currently hiring interns!)

© 2015 Mesosphere, Inc. All Rights Reserved.