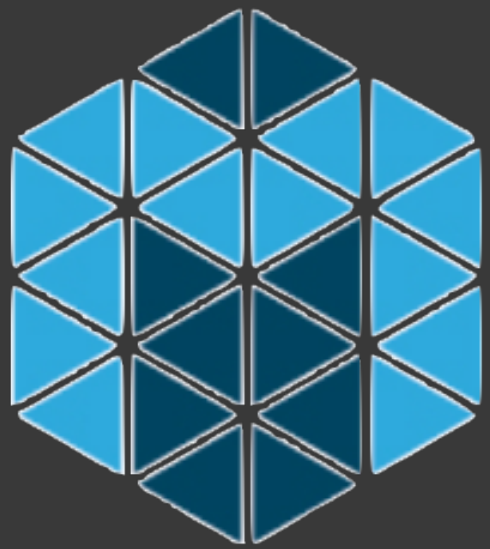
Datacenter Computing with Apache Mesos

BigData DC Meetup @AddThis
2014-04-15

Paco Nathan
http://liber118.com/pxn/
@pacoid



meetup.com/bigdatadc/events/172610652/

A Big Idea

Have you heard about
“data democratization” ? ? ?

⇒ making data available
throughout more of the organization

Have you heard about
“data democratization” ? ? ?

⇒ making data available
throughout more of the organization

Then how would you handle
“cluster democratization” ? ? ?

⇒ making data+resources available
throughout more of the organization

Have you heard about
“data democratization” ? ? ?

⇒ making data available
throughout more of the organization

In other words, how to remove silos...

Then how would you handle
“cluster democratization” ? ? ?

⇒ making data+resources available
throughout more of the organization

Lessons from Google

Datacenter Computing

Google has been doing *datacenter computing* for years,
to address the complexities of large-scale data workflows:

- leveraging the modern kernel: isolation in lieu of VMs
- "most (>80%) jobs are batch jobs, but the majority of resources (55–80%) are allocated to service jobs"
- mixed workloads, multi-tenancy
- relatively high utilization rates
- JVM? not so much…
- reality: scheduling batch is simple; scheduling services is hard/expensive

The Modern Kernel: Top Linux Contributors...

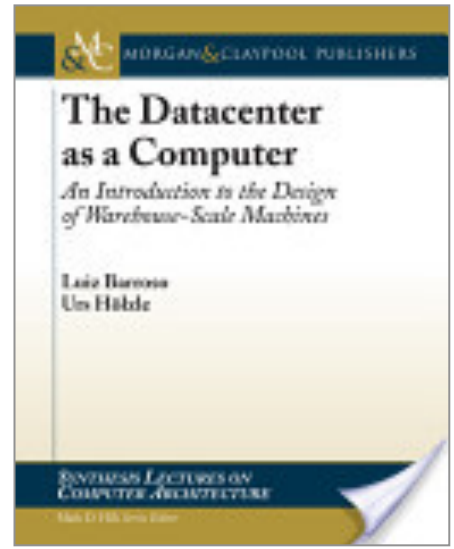
arstechnica.com/information-technology/2013/09/...

Company	Changes	Total
None	12,550	13.6%
Red Hat	9,483	10.2%
Intel	8,108	8.8%
Texas Instruments	3,814	4.1%
Linaro	3,791	4.1%
SUSE	3,212	3.5%
Unknown	3,032	3.3%
IBM	2,858	3.1%
Samsung	2,415	2.6%
Google	2,255	2.4%
Vision Engraving Systems	2,107	2.3%
Consultants	1,529	1.7%
Wolfson Microelectronics	1,516	1.6%
Oracle	1,248	1.3%
Broadcom	1,205	1.3%

Company	Changes	Total
NVidia	1,192	1.3%
Freescale	1,127	1.2%
Ingics Technology	1,075	1.2%
Renesas Electronics	1,010	1.1%
Qualcomm	965	1.0%
Cisco	871	0.9%
The Linux Foundation	840	0.9%
Academics	831	0.9%
AMD	820	0.9%
Inktank Storage	709	0.8%
NetApp	707	0.8%
LINBIT	705	0.8%
Fujitsu	694	0.7%
Parallels	684	0.7%
ARM	664	0.7%

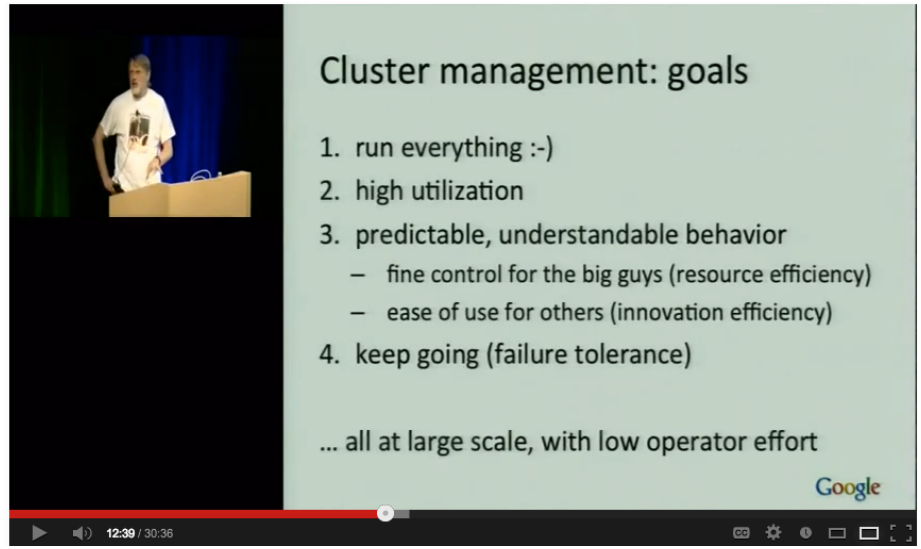
“Return of the Borg”

Return of the Borg: How Twitter Rebuilt Google’s Secret Weapon
Cade Metz
wired.com/wiredenterprise/2013/03/google-borg-twitter-mesos



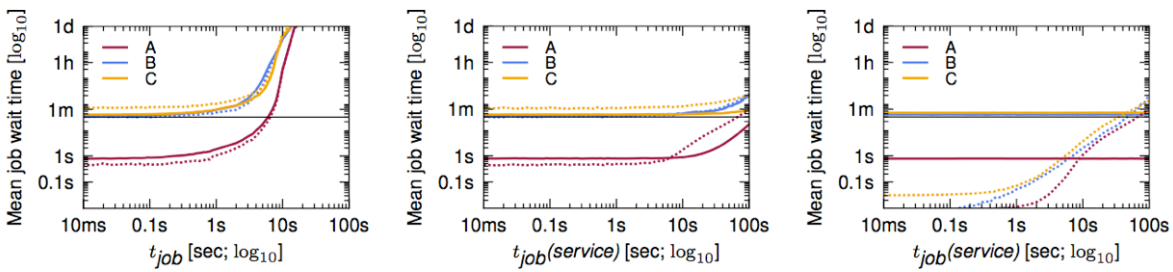
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines
Luiz André Barroso, Urs Hölzle
research.google.com/pubs/pub35290.html

2011 GAFS Omega
John Wilkes, et al.
youtu.be/0ZFMIO98Jkc



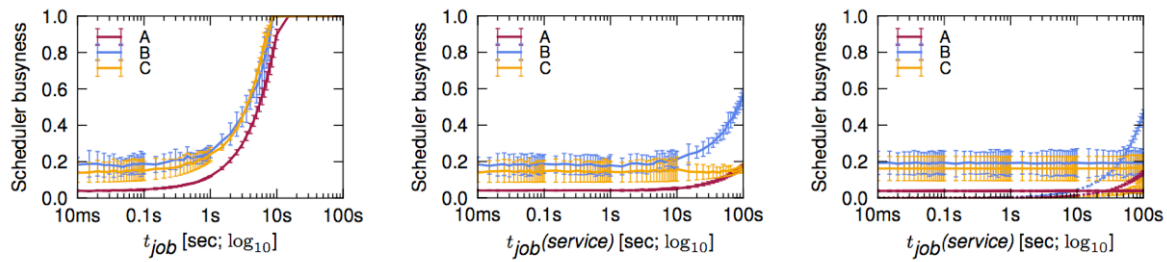
Google describes the technology…

Omega: flexible, scalable schedulers for large compute clusters
Malte Schwarzkopf, Andy Konwinski, Michael Abd-El-Malek, John Wilkes
eurosys2013.tudos.org/wp-content/uploads/2013/paper/Schwarzkopf.pdf



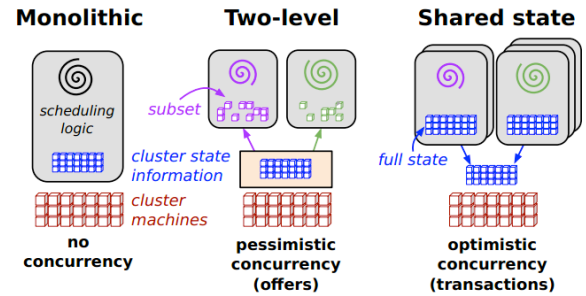
(a) Single-path. (b) Multi-path. (c) Shared state.

Figure 5: Schedulers' job wait time, as a function of t_{job} in the monolithic single-path case, $t_{job}(service)$ in the monolithic multi-path and shared-state cases. The SLO (horizontal bar) is 30s.



(a) Single-path. (b) Multi-path. (c) Shared state.

Figure 6: Schedulers' busyness, as a function of t_{job} in the monolithic single-path case, $t_{job}(service)$ in the monolithic multi-path and shared-state cases. The value is the median daily busyness over the 7-day experiment, and error bars are one ± median absolute deviation (MAD), i.e. the median deviation from the median value, a robust estimator of typical value dispersion.



Google describes the business case...

Taming Latency Variability
Jeff Dean
plus.google.com/u/0/+ResearchatGoogle/posts/C1dPhQhcDRv

Commercial OS Cluster Schedulers

- **IBM Platform Symphony**

- **Microsoft Autopilot**

Arguably, some grid controllers are quite notable in-category:

- **Univa Grid Engine** *(formerly SGE)*

- **Condor**

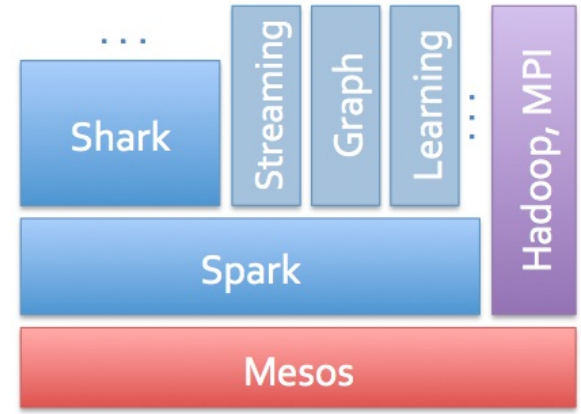
- *etc.*

Emerging at Berkeley

Beyond Hadoop

Hadoop – an open source solution for fault-tolerant parallel processing of batch jobs at scale, based on commodity hardware… however, other priorities have emerged for the **analytics lifecycle**:

- apps require integration beyond Hadoop
- multiple topologies, mixed workloads, multi-tenancy
- significant disruptions in h/w cost/performance curves
- higher utilization
- lower latency
- highly-available, long running services
- more than "Just JVM" – e.g., Python growth



amplab UC BERKELEY

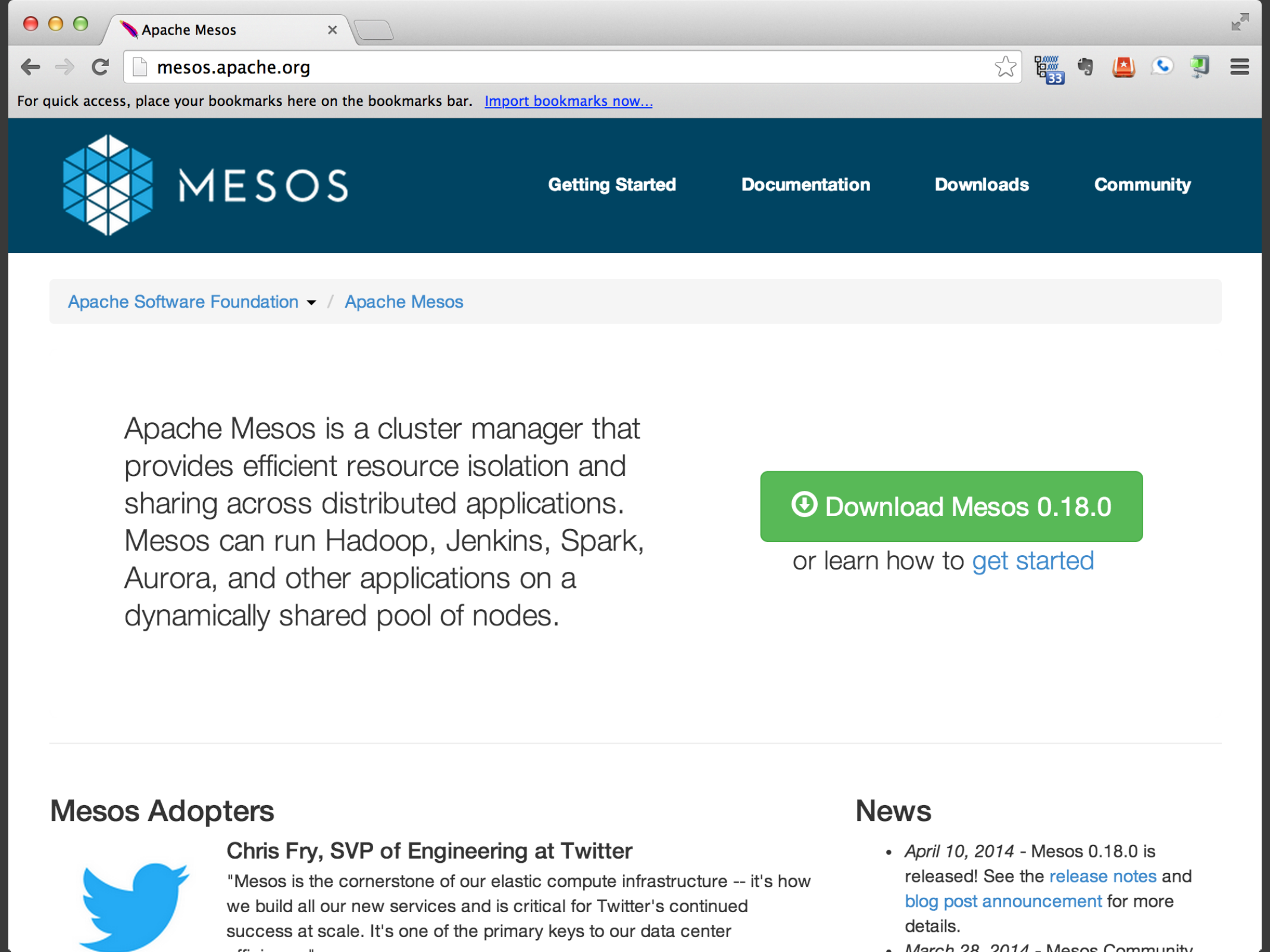
Just No Getting Around It

"There's Just No Getting Around It: You're Building a Distributed System"
Mark Cavage
ACM Queue (2013-05-03)
queue.acm.org/detail.cfm?id=2482856

key takeaways on architecture:

- decompose the business application into discrete services on the boundaries of fault domains, scaling, and data workload
- make as many things as possible stateless
- when dealing with state, deeply understand CAP, latency, throughput, and durability requirements

"Without practical experience working on successful—and failed—systems, most engineers take a "hopefully it works" approach and attempt to string together off-the-shelf software, whether open source or commercial, and often are unsuccessful at building a resilient, performant system. In reality, building a distributed system requires a methodical approach to requirements along the boundaries of failure domains, latency, throughput, durability, consistency, and desired SLAs for the business application at all aspects of the application."

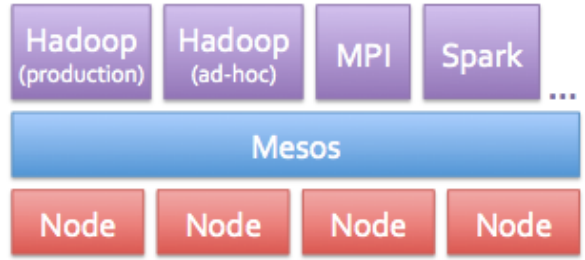

Apache Mesos
mesos.apache.org
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
MESOS
Getting Started
Documentation
Downloads
Community
Apache Software Foundation / Apache Mesos
Apache Mesos is a cluster manager that provides efficient resource isolation and sharing across distributed applications. Mesos can run Hadoop, Jenkins, Spark, Aurora, and other applications on a dynamically shared pool of nodes.
Download Mesos 0.18.0
or learn how to get started
Mesos Adopters
Chris Fry, SVP of Engineering at Twitter
"Mesos is the cornerstone of our elastic compute infrastructure -- it's how we build all our new services and is critical for Twitter's continued success at scale. It's one of the primary keys to our data center
News
April 10, 2014 - Mesos 0.18.0 is released! See the release notes and blog post announcement for more details.

Mesos – open source datacenter computing

a common substrate for cluster computing

mesos.apache.org

heterogenous assets in your datacenter or cloud
made available as a homogenous set of resources

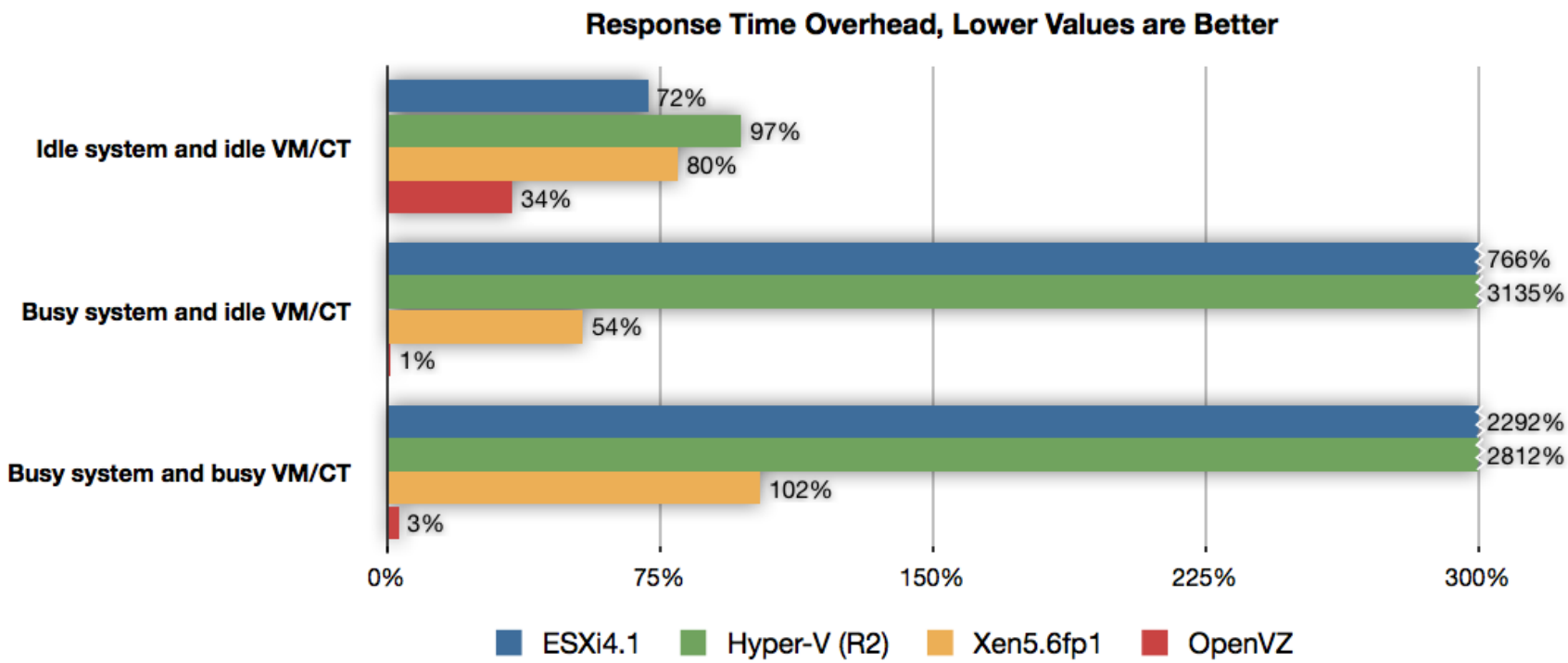


- top-level Apache project
- scalability to 10,000s of nodes
- obviates the need for virtual machines
- isolation (pluggable) for CPU, RAM, I/O, FS, etc.
- fault-tolerant leader election based on Zookeeper
- APIs in **C++**, **Java/Scala**, **Python**, **Go**, **Erlang**, **Haskell**
- web UI for inspecting cluster state
- available for Linux, OpenSolaris, Mac OSX

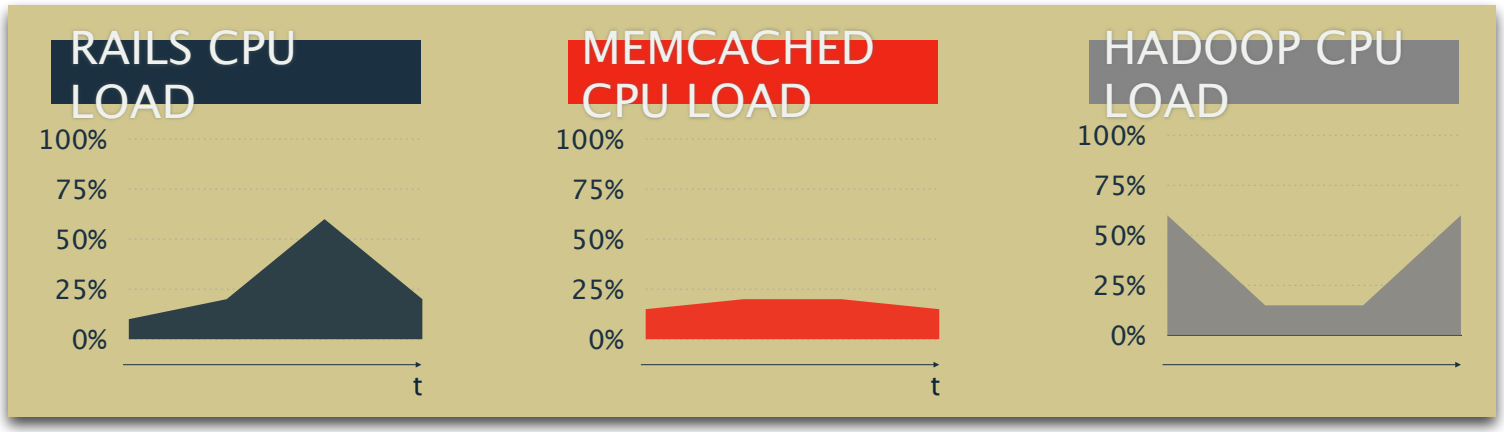
What are the costs of Virtualization?

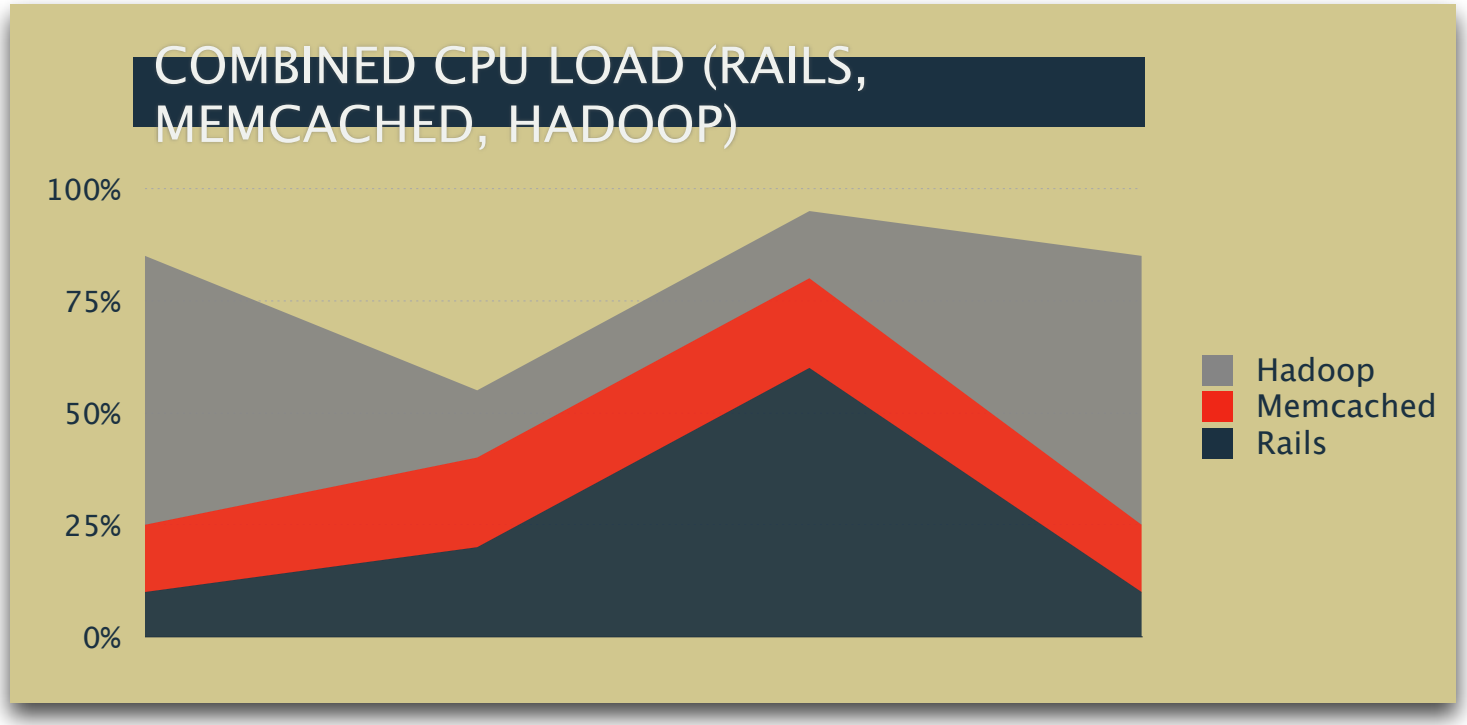
benchmark type	OpenVZ improvement
mixed workloads	210%-300%
LAMP (related)	38%-200%
I/O throughput	200%-500%
response time	order magnitude

more pronounced at higher loads



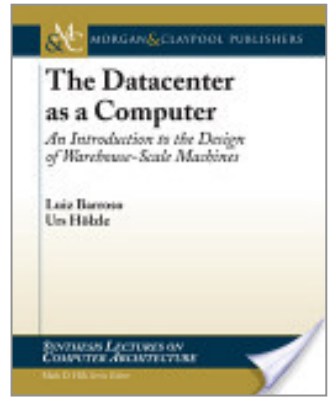
What are the costs of Single Tenancy?





Arguments for Datacenter Computing

rather than running several specialized clusters, each at relatively low utilization rates, instead run many mixed workloads



obvious benefits are realized in terms of:

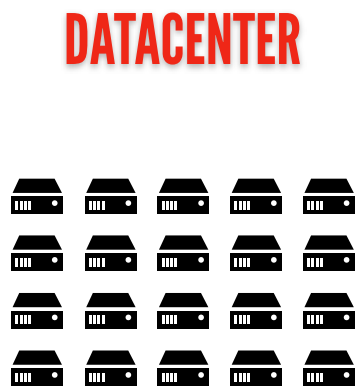
- scalability, elasticity, fault tolerance, performance, utilization
- reduced equipment capex, Ops overhead, etc.
- reduced licensing, eliminating need for VMs or potential vendor lock-in

subtle benefits – arguably, more important for Enterprise IT:

- reduced time for engineers to ramp up new services at scale
- reduced latency between batch and services, enabling new high ROI use cases
- enables Dev/Test apps to run safely on a Production cluster

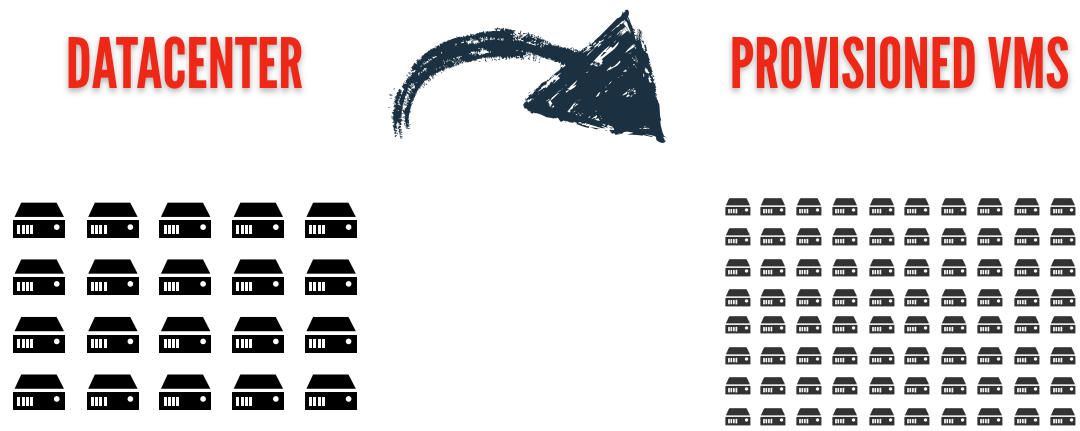
Analogies and Architecture

Prior Practice: Dedicated Servers



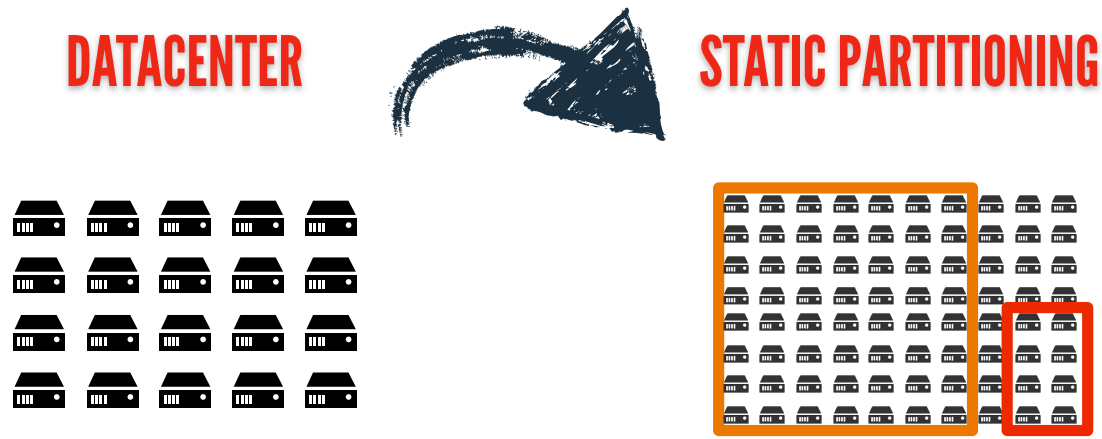
- *low utilization rates*
- *longer time to ramp up new services*

Prior Practice: Virtualization



- *even more machines to manage*
- *substantial performance decrease due to virtualization*
- *VM licensing costs*

Prior Practice: Static Partitioning



- *even more machines to manage*
- *substantial performance decrease due to virtualization*
- *VM licensing costs*
- *failures make static partitioning more complex to manage*

Mesos: One Large Pool of Resources

"We wanted people to be able to program for the datacenter just like they program for their laptop."

Ben Hindman

Frameworks Integrated with Mesos

Continuous Integration:
Jenkins, GitLab

Big Data:
Hadoop, Spark, Storm, Kafka

Python workloads:
DPark, Exelixi

Meta-Frameworks / HA Services:
Aurora, Marathon

Orchestration:
Singularity

Distributed Cron:
Chronos

Data Storage:
ElasticSearch, Cassandra, Hypertable

Containers:
Docker, Deimos

Parallel Processing:
Chapel, MPI

Fault-tolerant distributed systems...

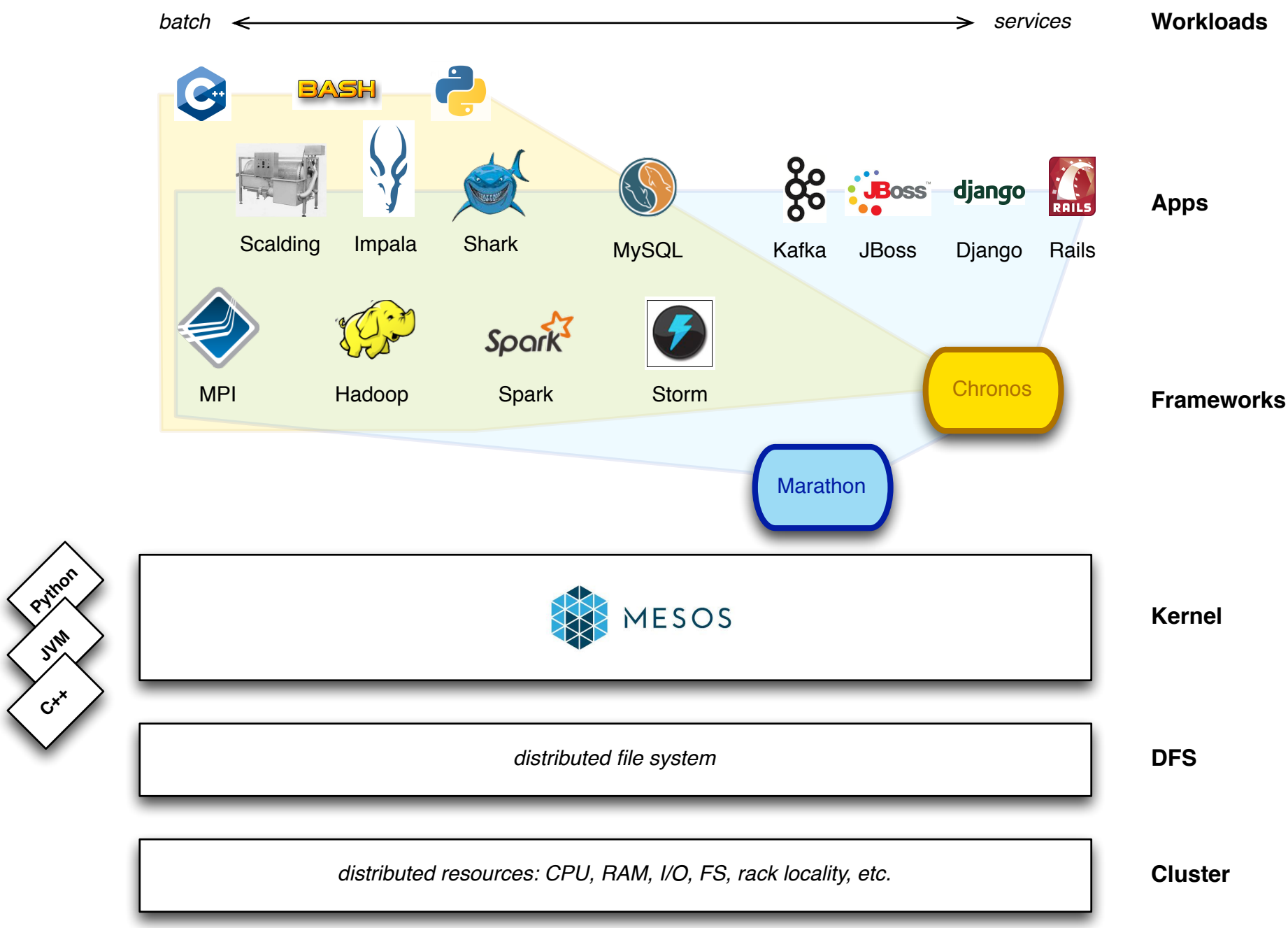
...written in 100-300 lines of
C++, Java/Scala, Python, Go, etc.

...building blocks, if you will

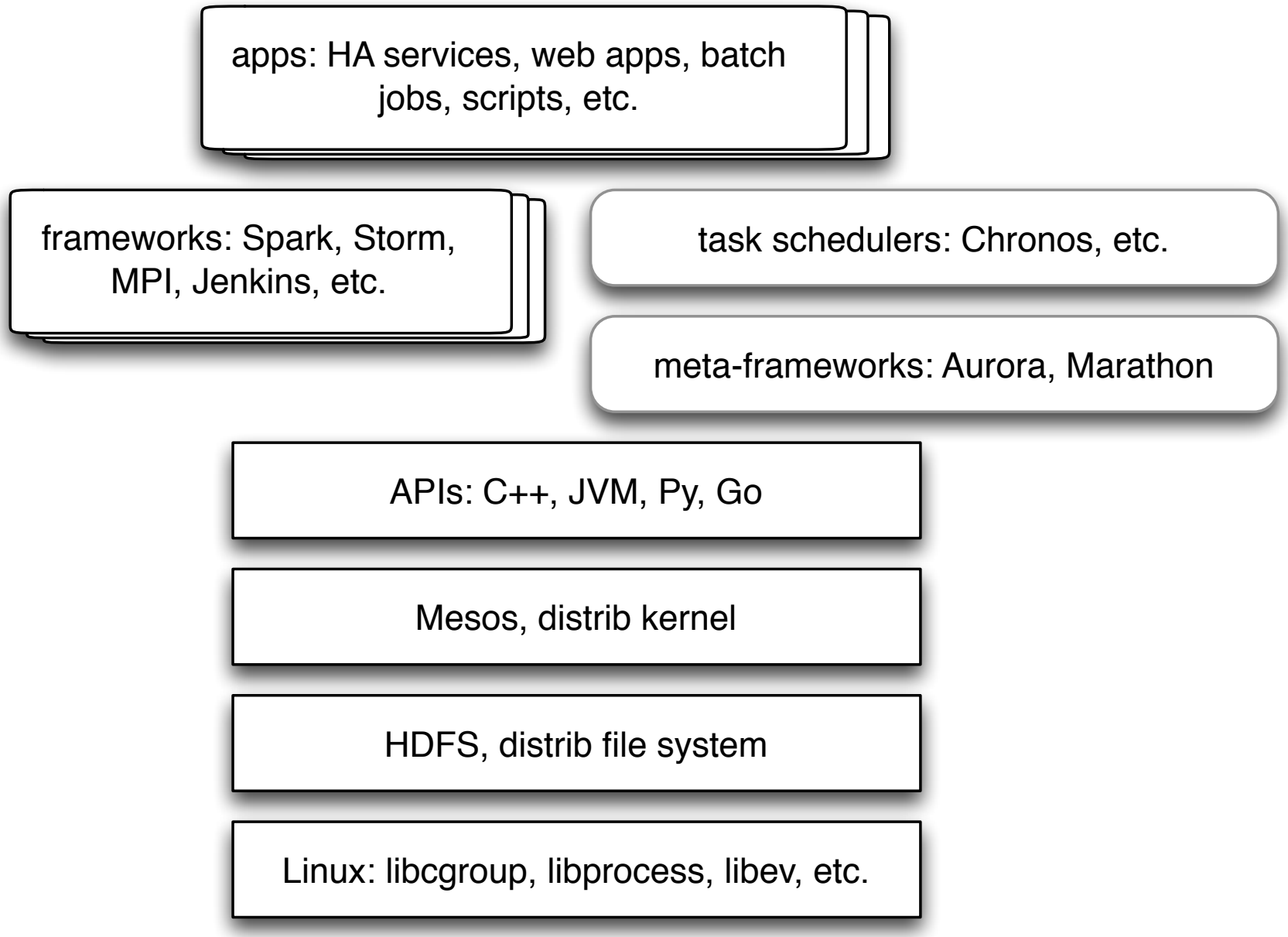
Q: required lines of network code?

A: probably none

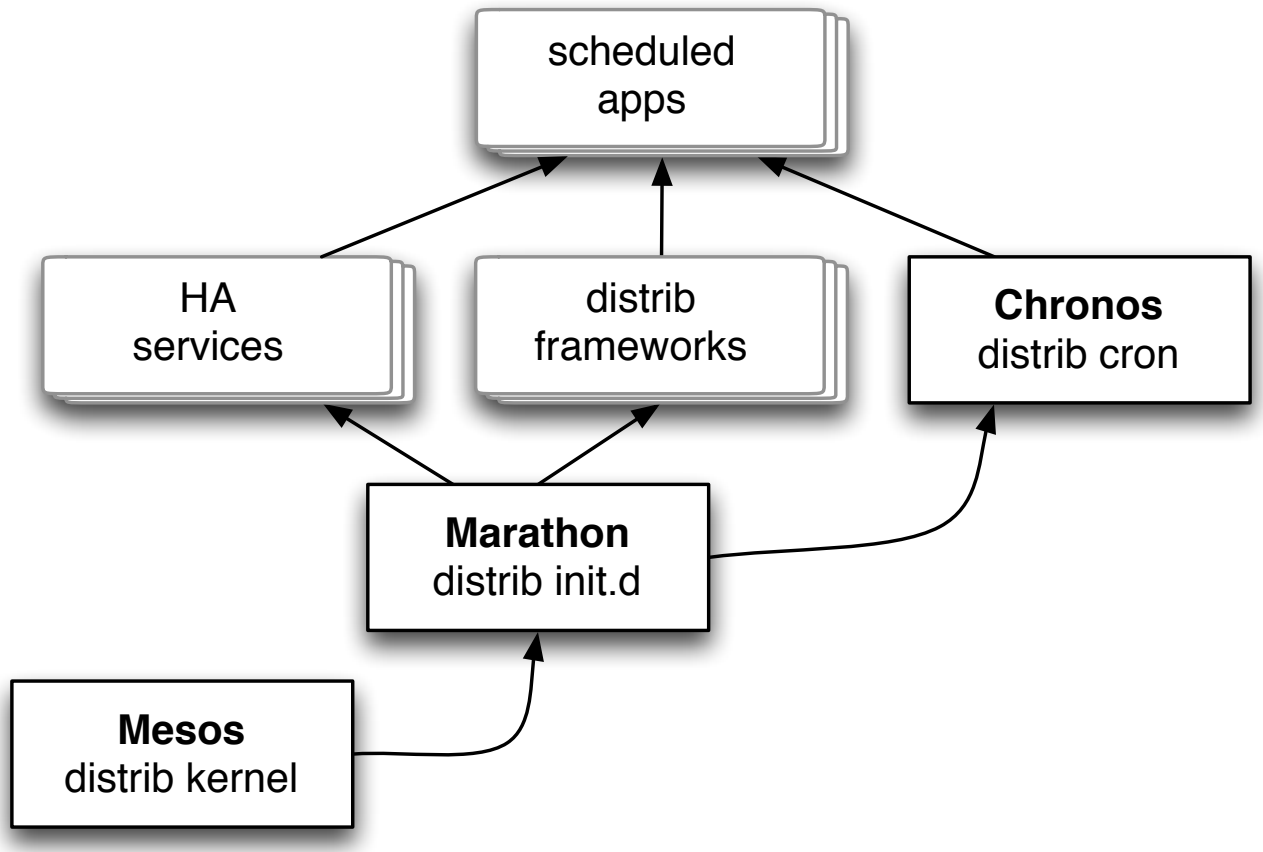
Mesos – architecture



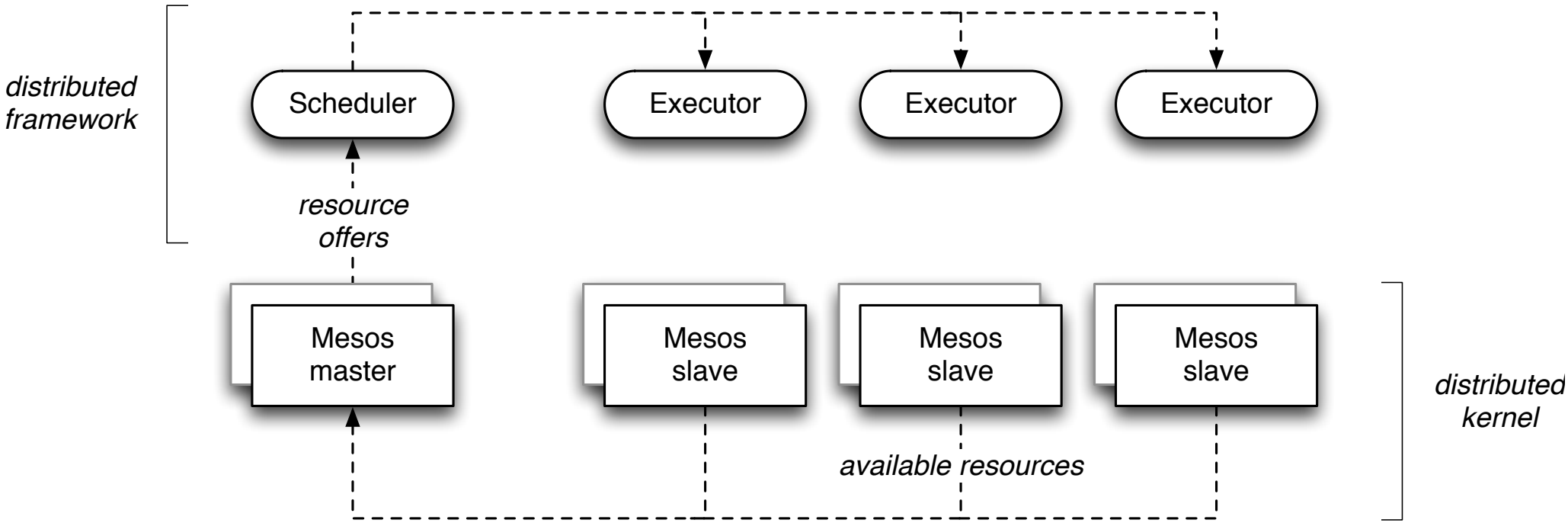
Mesos – architecture



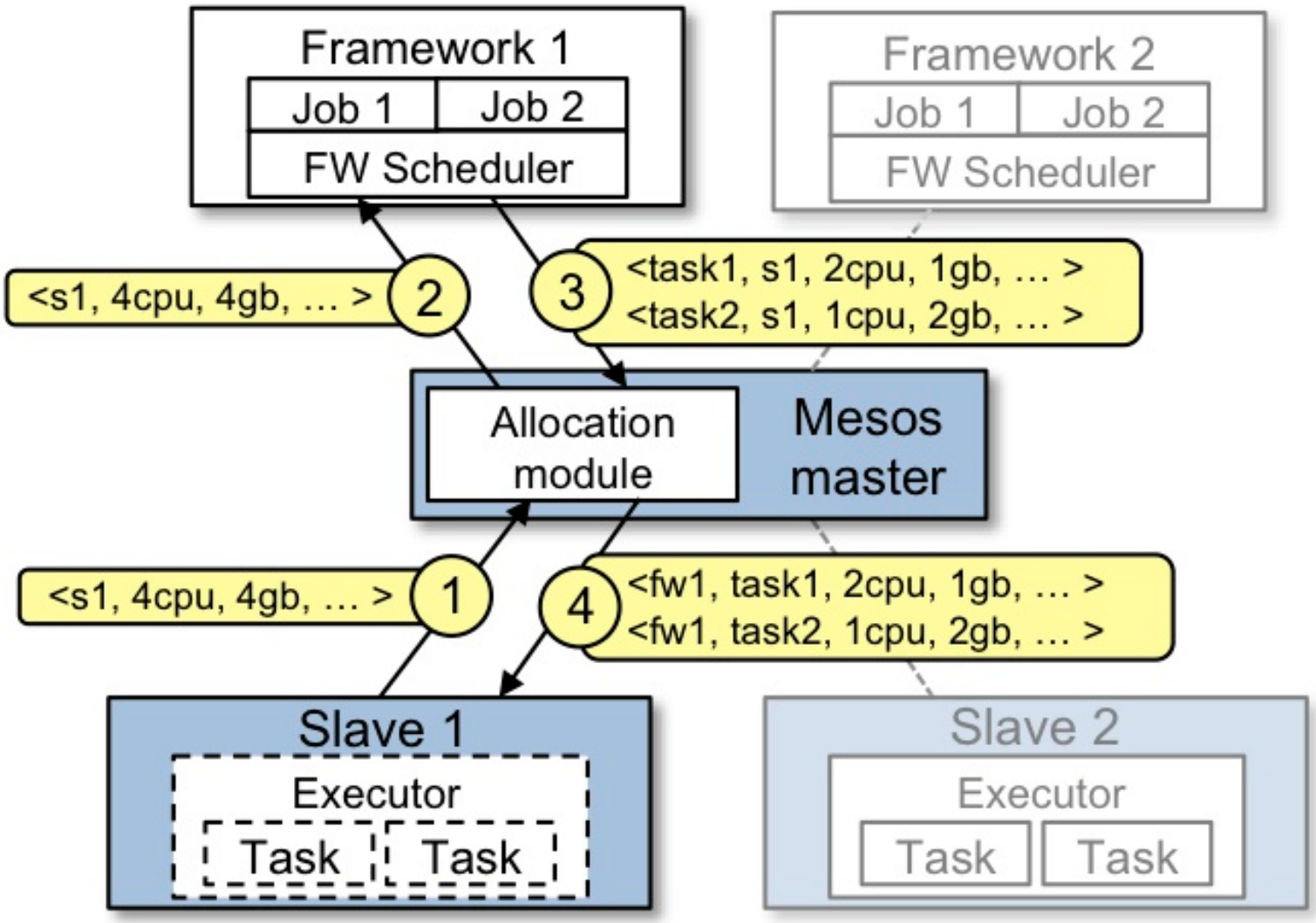
Mesos – dynamics



Mesos – dynamics

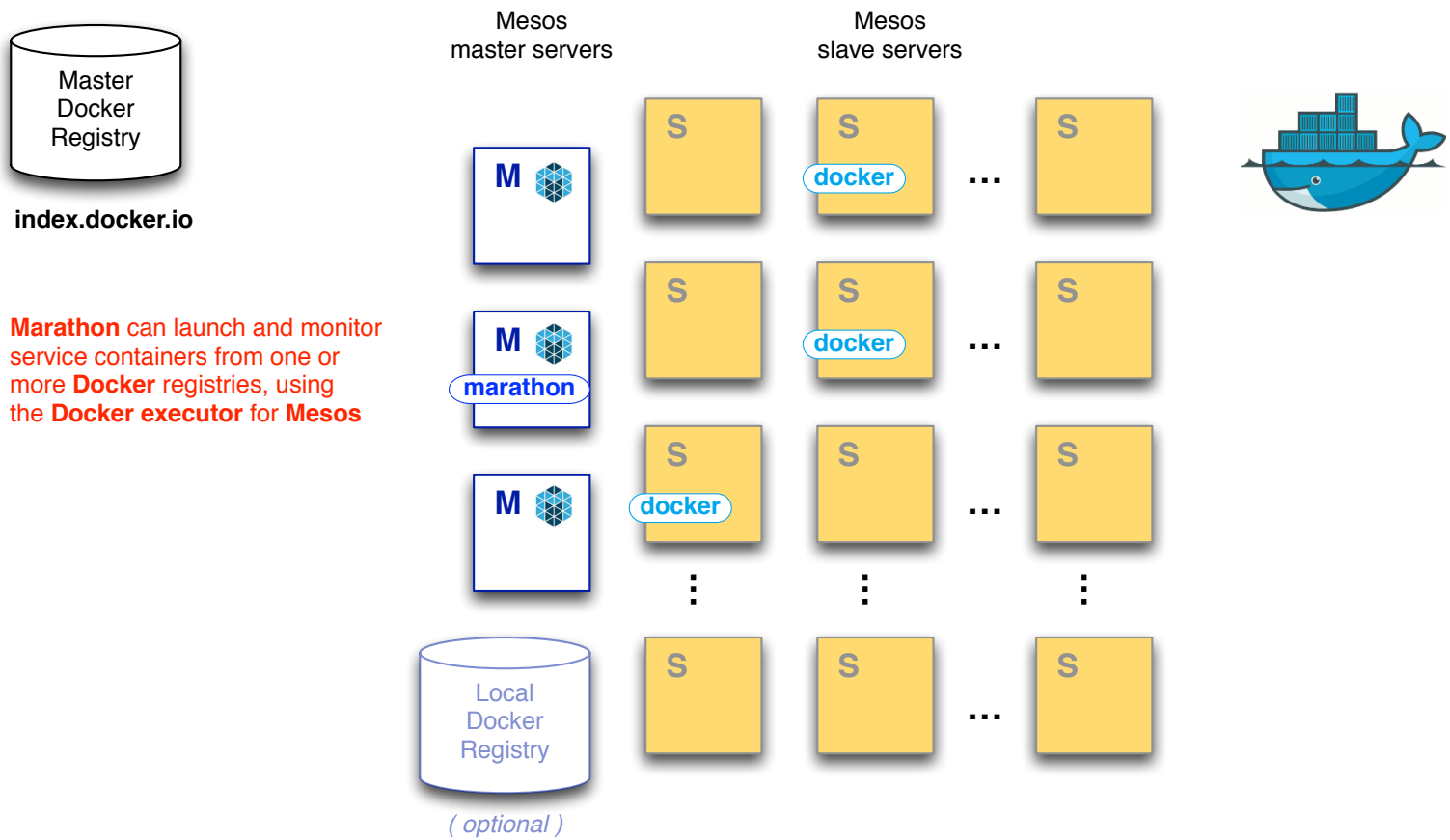


Example: Resource Offer in a Two-Level Scheduler



mesos.apache.org/documentation/latest/mesos-architecture/

Example: Docker on Mesos



mesosphere.io/2013/09/26/docker-on-mesos/

Example: Docker on Mesos

Mesos Master Server

```
init
 |
 + mesos-master
 |
 + marathon
 |
```

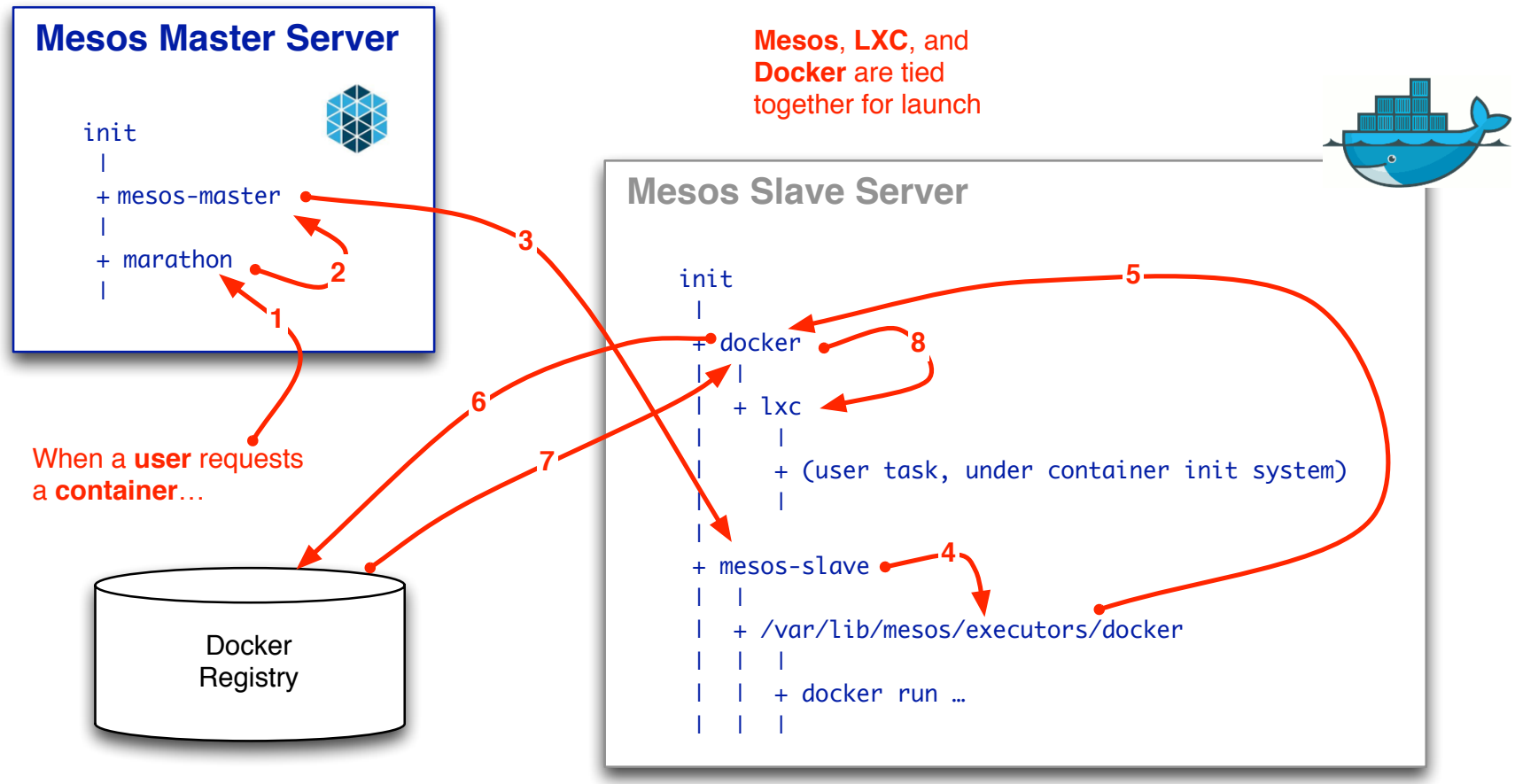
The **executor**, monitored by the **Mesos slave**, delegates to the local **Docker daemon** for image discovery and management. The **executor** communicates with **Marathon** via the **Mesos master** and ensures that **Docker** enforces the specified resource limitations.

Mesos Slave Server

```
init
 |
 + docker
 |  |
 |  + lxc
 |     |
 |     + (user task, under container init system)
 |     |
 |
 + mesos-slave
 |  |
 |  + /var/lib/mesos/executors/docker
 |  |  |
 |  |  + docker run …
 |  |  |
```

mesosphere.io/2013/09/26/docker-on-mesos/

Example: Docker on Mesos

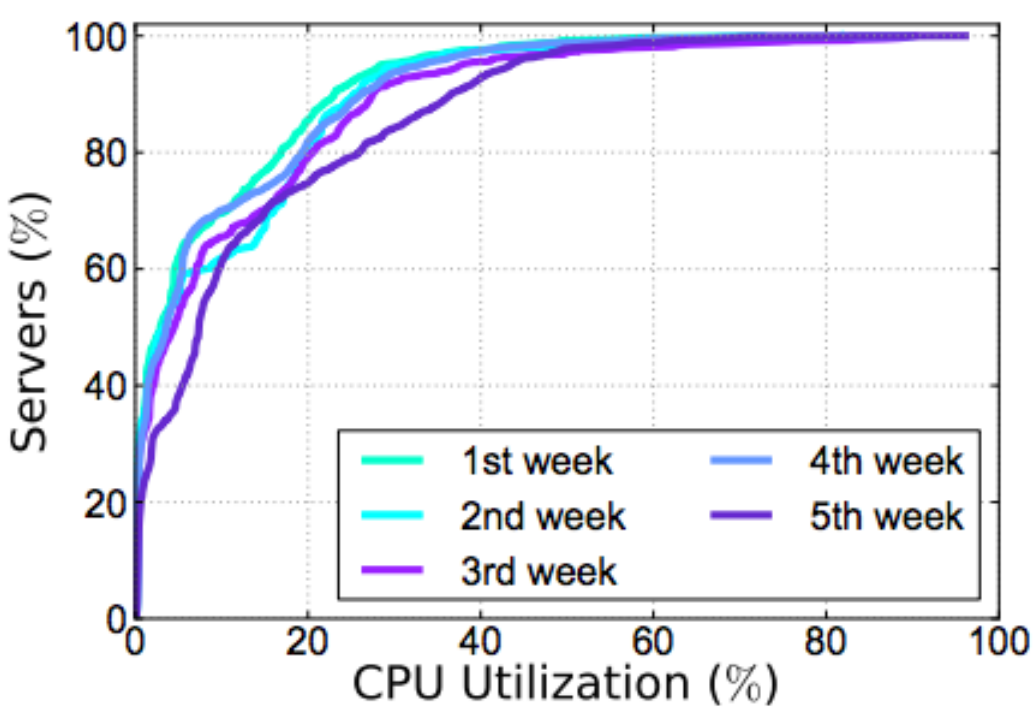


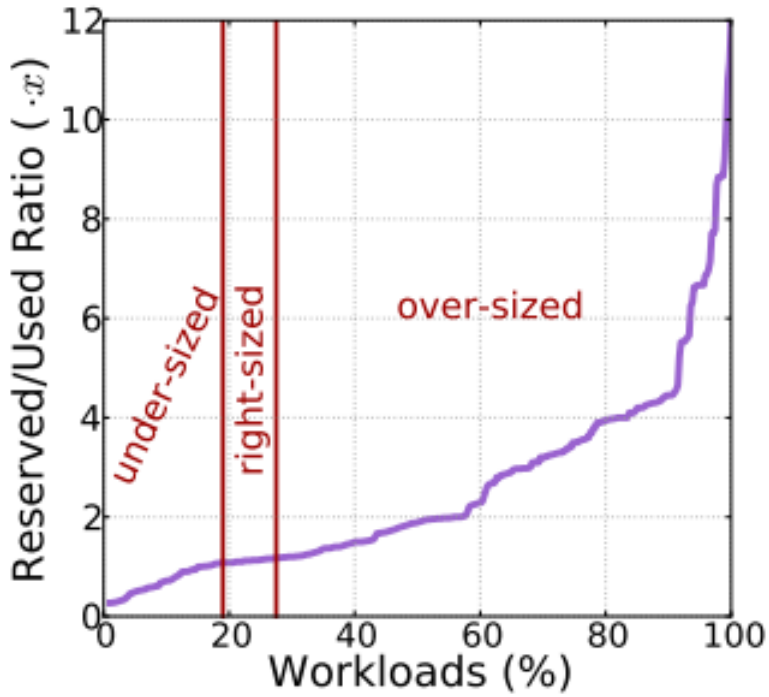
mesosphere.io/2013/09/26/docker-on-mesos/

Looking Ahead...

Quasar+Mesos @ Stanford, Twitter, etc....

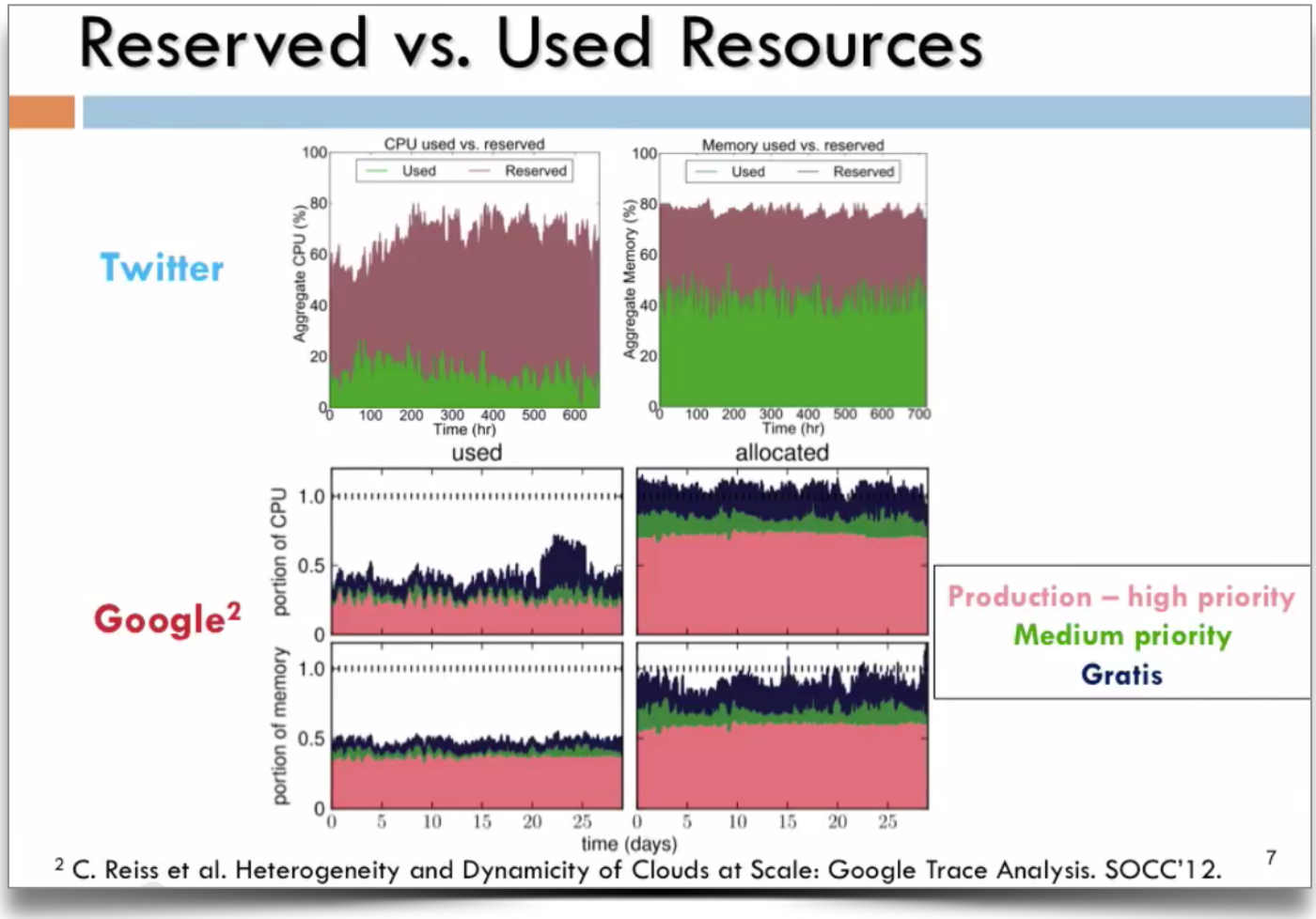
Quasar: Resource-Efficient and QoS-Aware Cluster Management
Christina Delimitrou, **Christos Kozyrakis**
stanford.edu/~cdel/2014.asplos.quasar.pdf





Quasar+Mesos @ Stanford, Twitter, etc....

Improving Resource Efficiency with Apache Mesos
Christina Delimitrou
youtu.be/YpmElyi94AA



Quasar+Mesos @ Stanford, Twitter, etc....

Consider that for datacenter computing at scale, surge in workloads implies:

- large cap-ex investment, long lead-time to build
- utilities cannot supply the power requirements

Even for large players that achieve 2x beyond typical industry DC util rates, those factors become show-stoppers. Even so, high rates of over-provisioning are typical, so there's much room to improve.

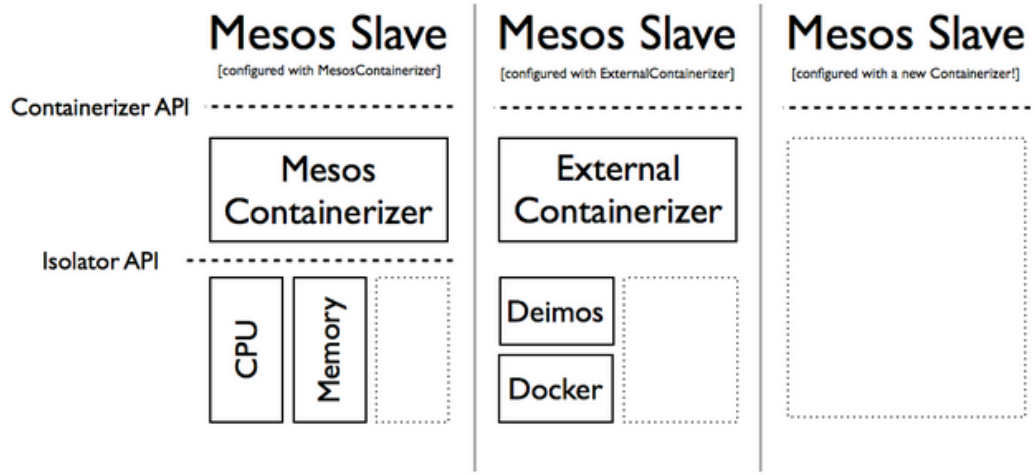
Experiences with Quasar+Mesos showed:

- 88% apps get >95% performance
- ~10% overprovisioning instead of 500%
- up to 70% cluster util at steady state
- 23% shorter scenario completion

What's New in Mesos 0.18 ?

mesos.apache.org/blog/mesos-0-18-0-released/

- Containerizer API
- Isolator API
- Cgroup layout change



Containerization and Docker:

> *"The first change is in terminology. The Mesos Slave now uses a Containerizer to provide an environment for each executor and its tasks to run in. Containerization includes resource isolation but is a more general concept that can encompass such things as packaging."*

Because...
Use Cases

Production Deployments (public)

HubSpot

PayPal™

vimeo

ebay®

sharethrough

Atigeo™

Atlassian

ignidata

igniting business with data

medidata

DueDil

Sigmoid

xogito

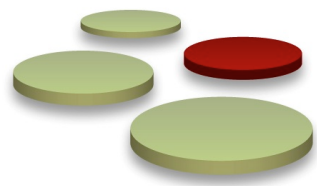
[kä-gi-tō]

CLOUD PHYSICS

redhat

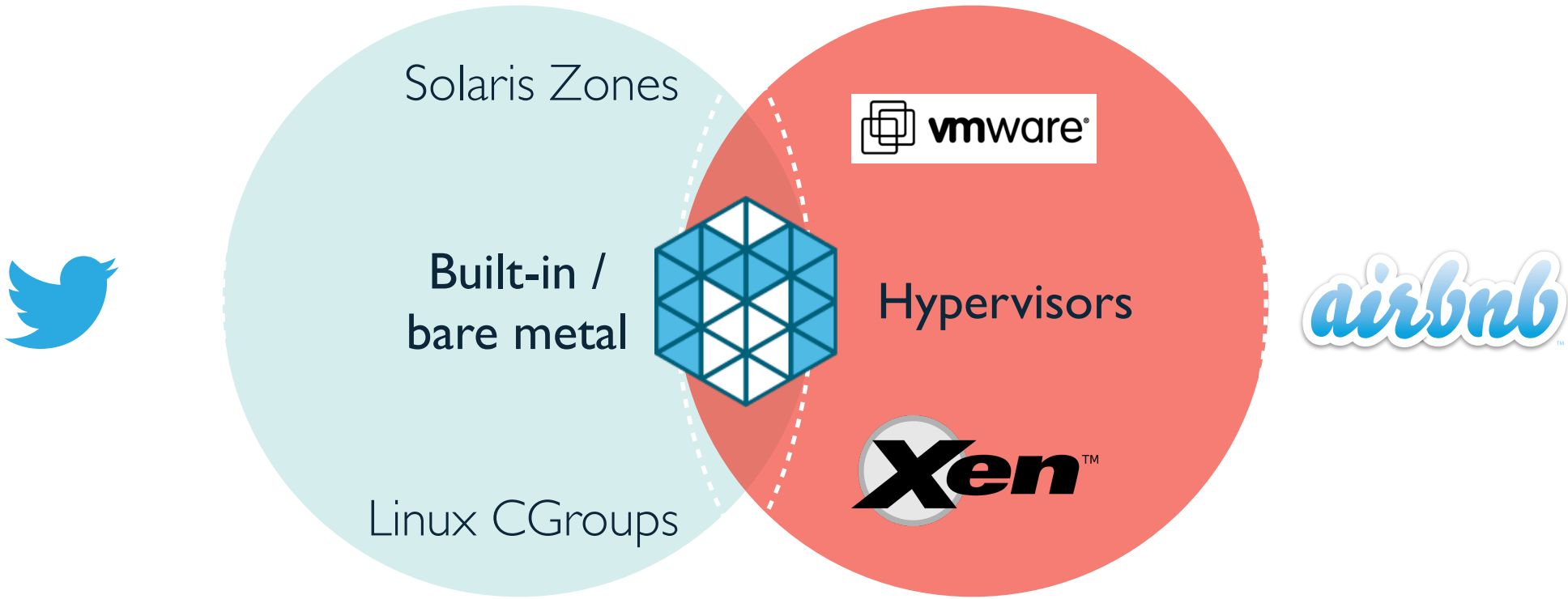
device scape™

OOYALA®

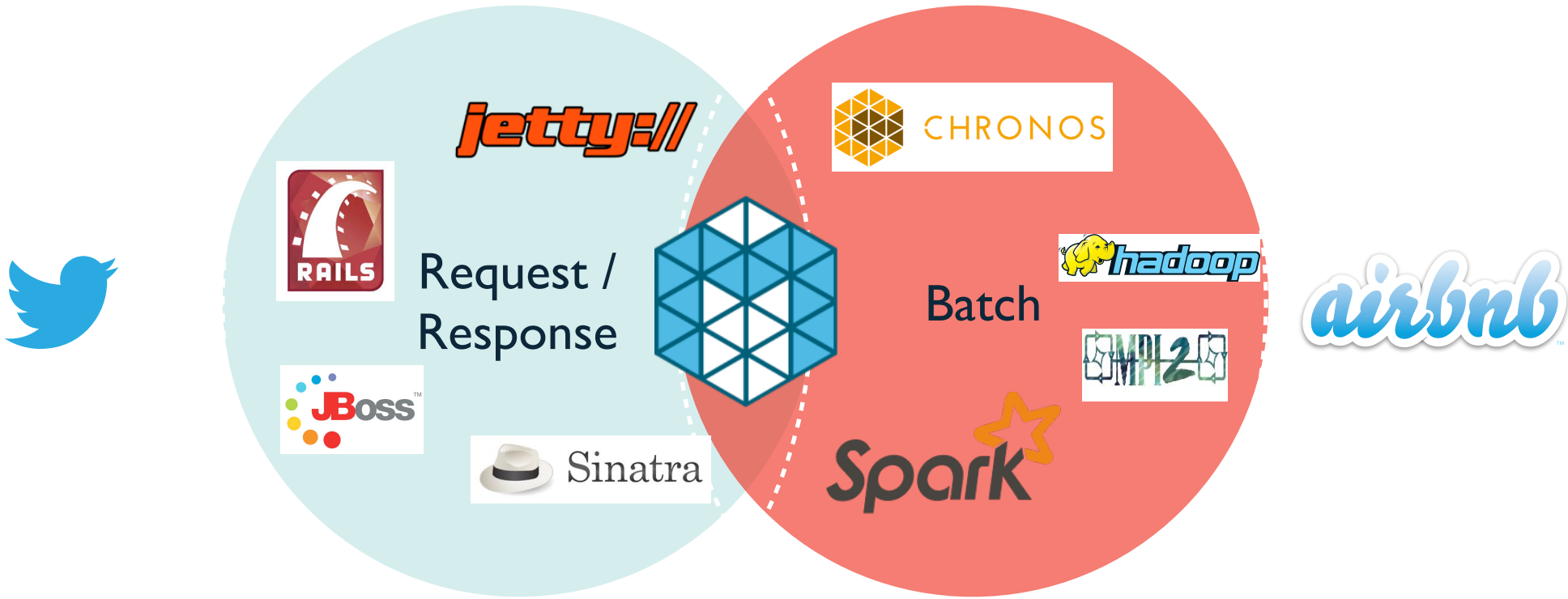


OpenTable®

Opposite Ends of the Spectrum, One Common Substrate

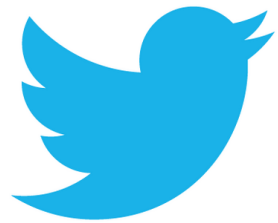


Opposite Ends of the Spectrum, One Common Substrate



Case Study: Twitter (bare metal / on premise)

"Mesos is the cornerstone of our elastic compute infrastructure – it's how we build all our new services and is critical for Twitter's continued success at scale. It's one of the primary keys to our data center efficiency."



Chris Fry, SVP Engineering

blog.twitter.com/2013/mesos-graduates-from-apache-incubation

wired.com/gadgetlab/2013/11/qa-with-chris-fry/

- key services run in production: analytics, typeahead, ads
- Twitter engineers rely on Mesos to build all new services
- instead of thinking about static machines, engineers think about resources like CPU, memory and disk
- allows services to scale and leverage a shared pool of servers across datacenters efficiently
- reduces the time between prototyping and launching

Case Study: Airbnb (fungible cloud infrastructure)

"We think we might be pushing data science in the field of travel more so than anyone has ever done before… a smaller number of engineers can have higher impact through automation on Mesos."

Mike Curtis, VP Engineering
gigaom.com/2013/07/29/airbnb-is-engineering-itself-into-a-data...

- improves resource management and efficiency
- helps advance engineering strategy of building small teams that can move fast
- key to letting engineers make the most of AWS-based infrastructure beyond just Hadoop
- allowed company to migrate off Elastic MapReduce
- enables use of Hadoop along with Chronos, Spark, Storm, etc.

Case Study: eBay (continuous integration)

eBay PaaS Team
ebaytechblog.com/2014/04/04/delivering-ebays-ci-solution-with-apache-mesos-part-i/

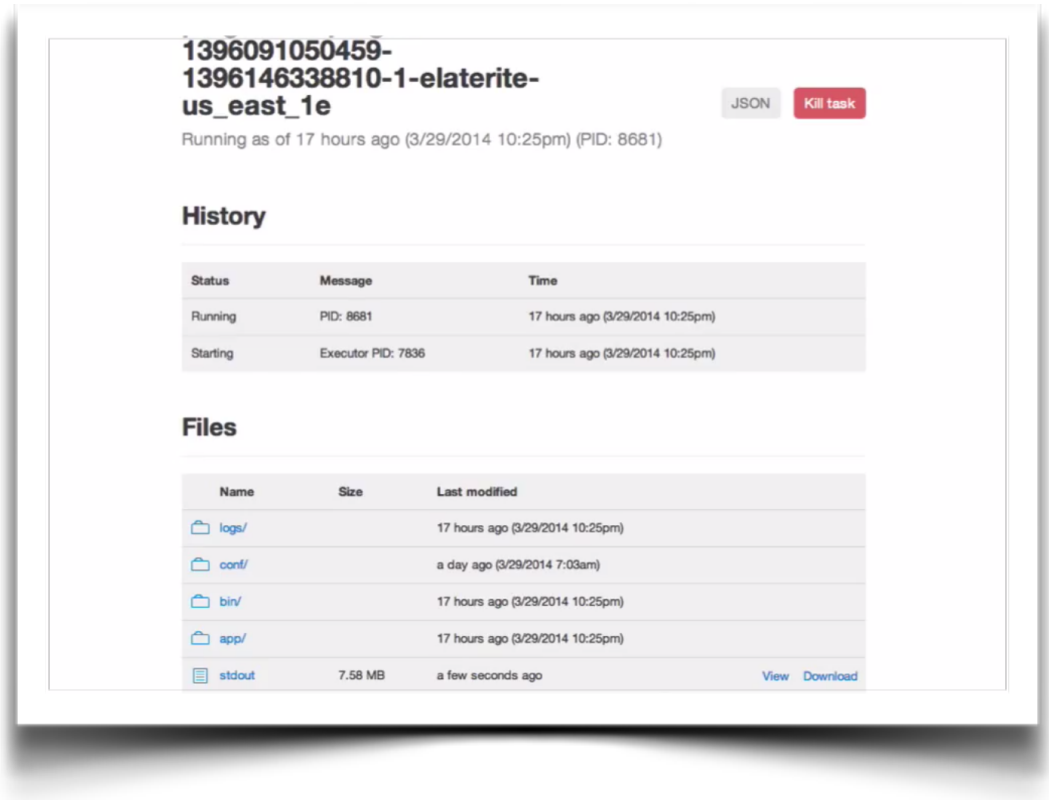
- cluster management (PaaS core framework services) for CI
- integration of: OpenStack, Jenkins, Zookeeper, Mesos, Marathon, Ansible

In eBay's existing CI model, each developer gets a personal CI/Jenkins Master instance. This Jenkins instance runs within a dedicated VM, and over time the result has been VM sprawl and poor resource utilization. We started looking at solutions to maximize our resource utilization and reduce the VM footprint while still preserving the individual CI instance model. After much deliberation, we chose Apache Mesos for a POC. This post shares the journey of how we approached this challenge and accomplished our goal.

Case Study: HubSpot (cluster management)

Tom Petr
youtu.be/ROn14csiikw

- 500 deployable objects; 100 deploys/day to production; 90 engineers; 3 devops on Mesos cluster
- "Our QA cluster is now a fixed $10K/month — that used to fluctuate"



Resources

Apache Mesos Project
mesos.apache.org

Twitter
@ApacheMesos

Mesosphere
mesosphere.io

Tutorials
mesosphere.io/learn

Documentation
mesos.apache.org/documentation

2011 USENIX Research Paper
usenix.org/legacy/event/nsdi11/tech/full_papers/Hindman_new.pdf

Collected Notes/Archives
goo.gl/jPtTP

DIY

http://elastic.mesosphere.io

http://mesosphere.io/learn

Elastic Apache Mesos

https://elastic.mesosphere.io

Launch an Apache Mesos Cluster in 3 2 1

Elastic Apache Mesos is a web service that automates the creation of Apache Mesos clusters on Amazon Elastic Compute Cloud (EC2). It provisions EC2 instances, installs dependencies including Apache ZooKeeper and HDFS, and delivers you a cluster with all the services running.

Mesos allows you to easily share compute resources and data between frameworks like Apache Hadoop and Apache Spark.

You just pay for your EC2 instances; Elastic Apache Mesos costs you nothing, nada, zilch on top of that.

3 Choose a cluster size

● 6 instances	● 18 instances
12 vCPUs 45 GiB memory $1.44 per hour[1]	36 vCPUs 135 GiB memory $4.32 per hour[1]

Elastic Apache Mesos

https://elastic.mesosphere.io

3 Choose a cluster size

6 instances	18 instances
12 vCPUs 45 GiB memory	36 vCPUs 135 GiB memory
$1.44 per hour[1]	$4.32 per hour[1]
Perfect for trying out Apache Mesos	Unleash the data-cruncher

1. Estimated price you will be charged in USD by Amazon EC2 after launching your cluster based on on-demand instance prices. We charge you nothing, nada, $0 on top of that!

All instances run in US East Region (N. Virginia) and use the following configuration:

- 2 vCPUs
- 7.5 GiB memory
- Ubuntu 12.10 (ami-2bc99d42)
- Type m1.large

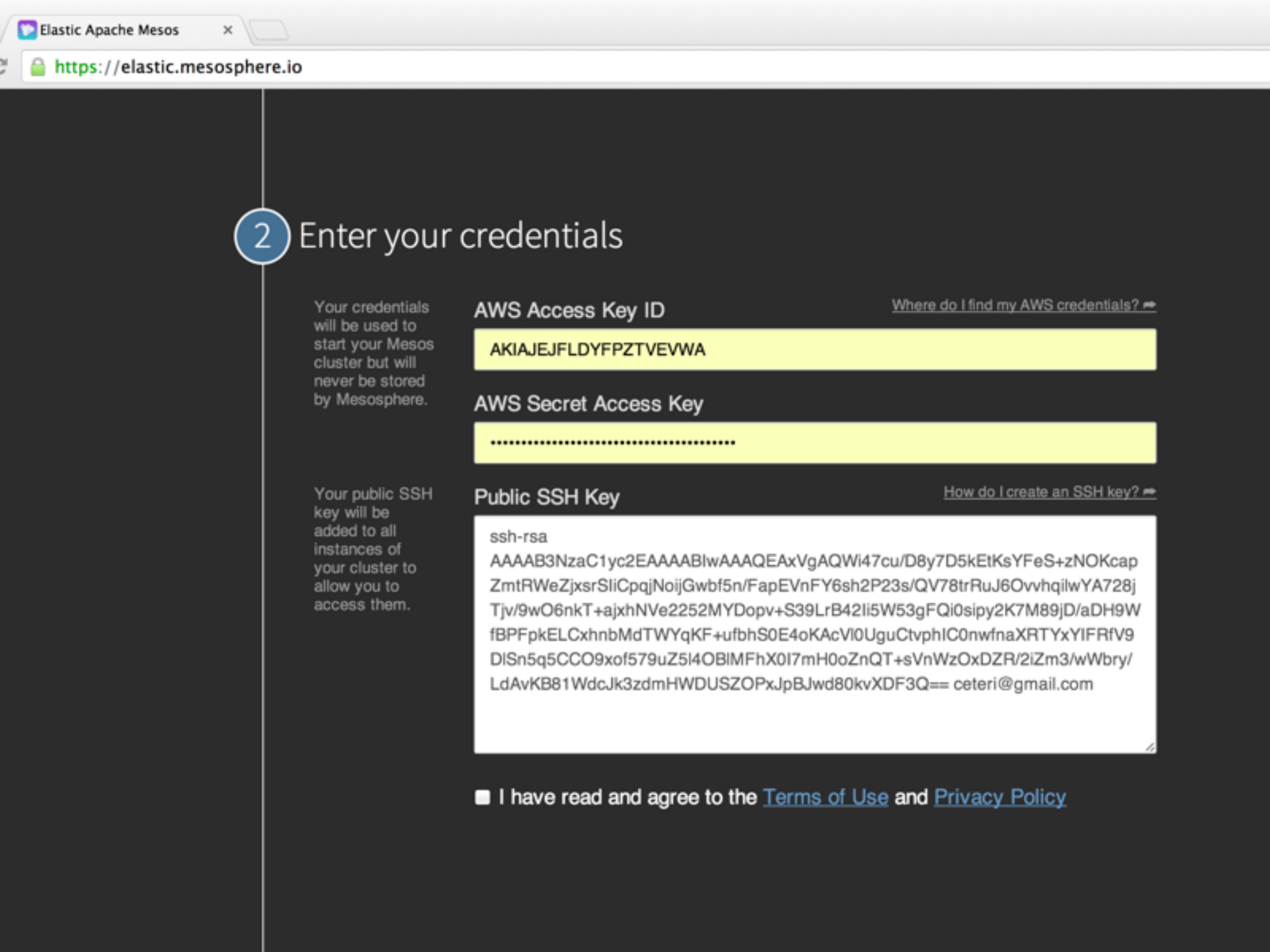
2 Enter your credentials

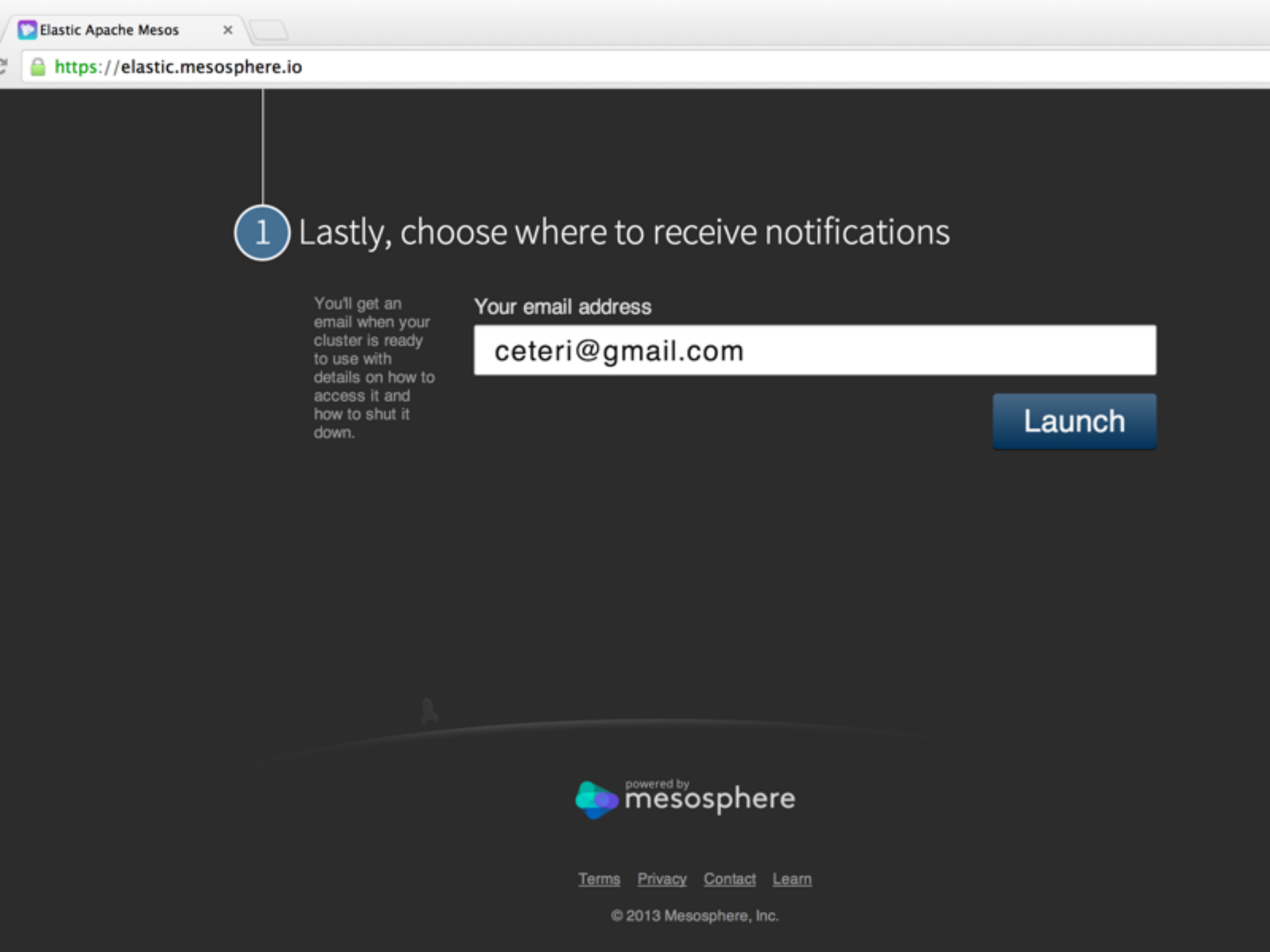
Your credentials will be used to start your Mesos

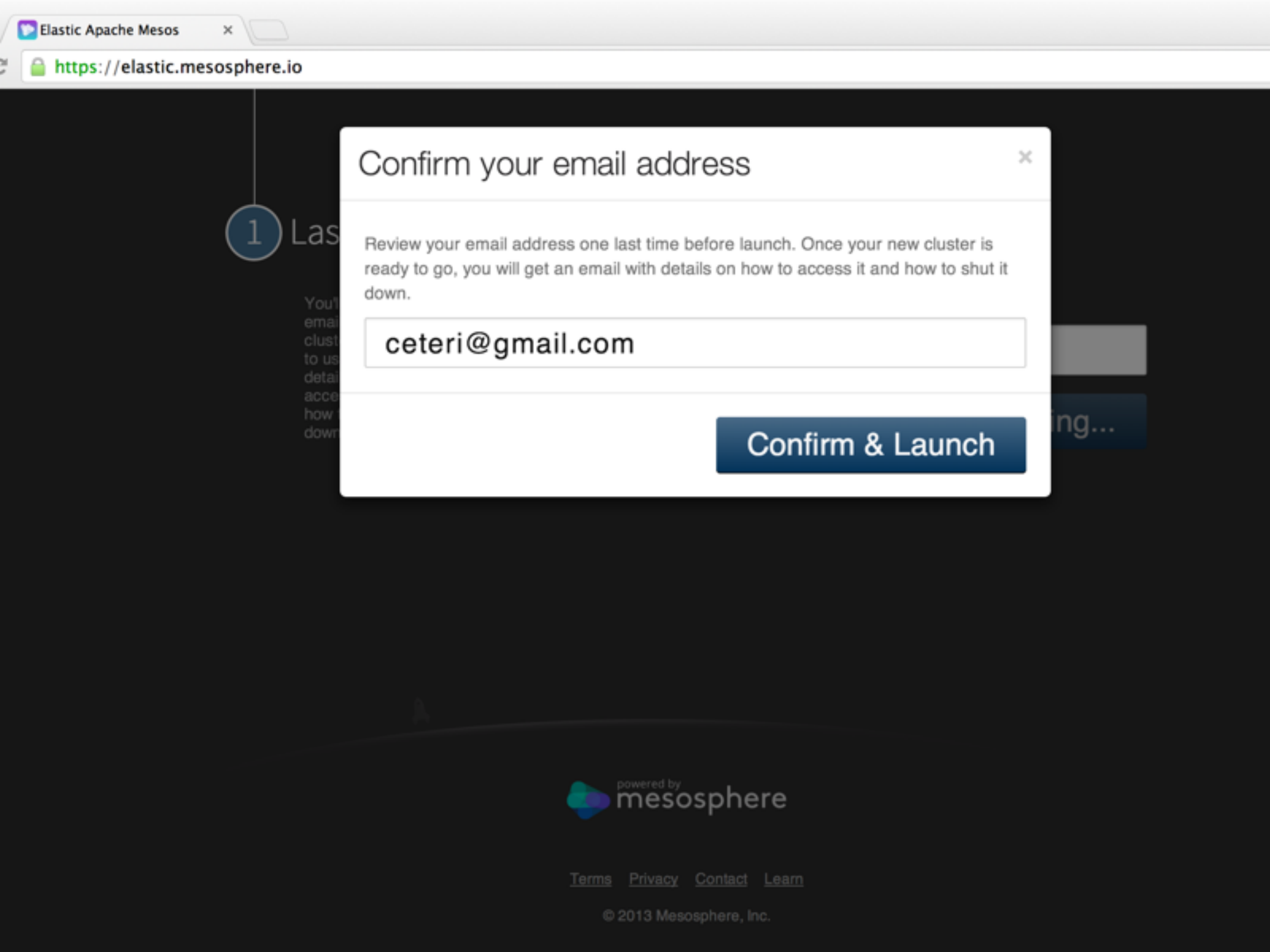
AWS Access Key ID

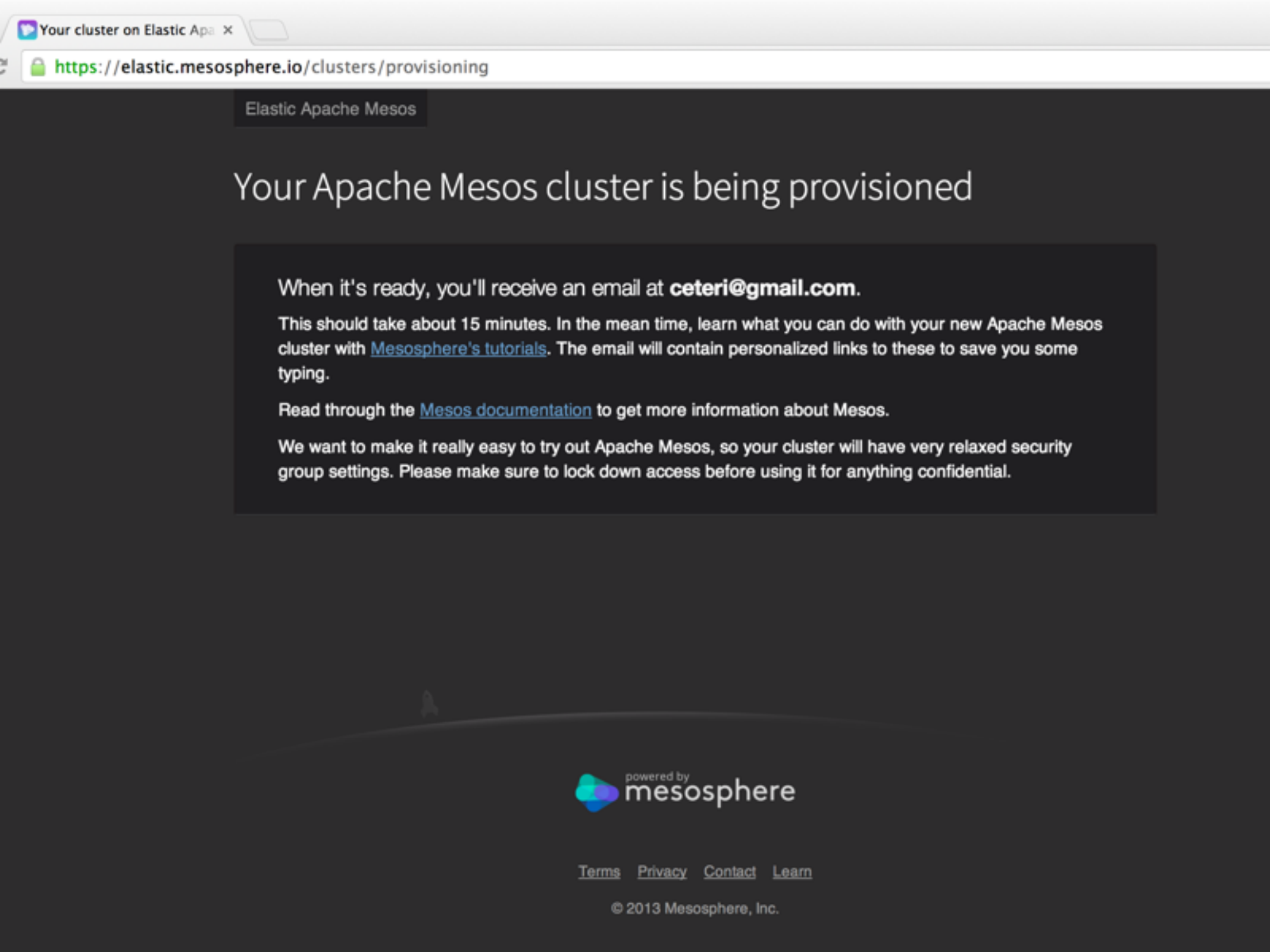
Where do I find my AWS credentials?

AKIAJEJELDYEPZTVEVWA

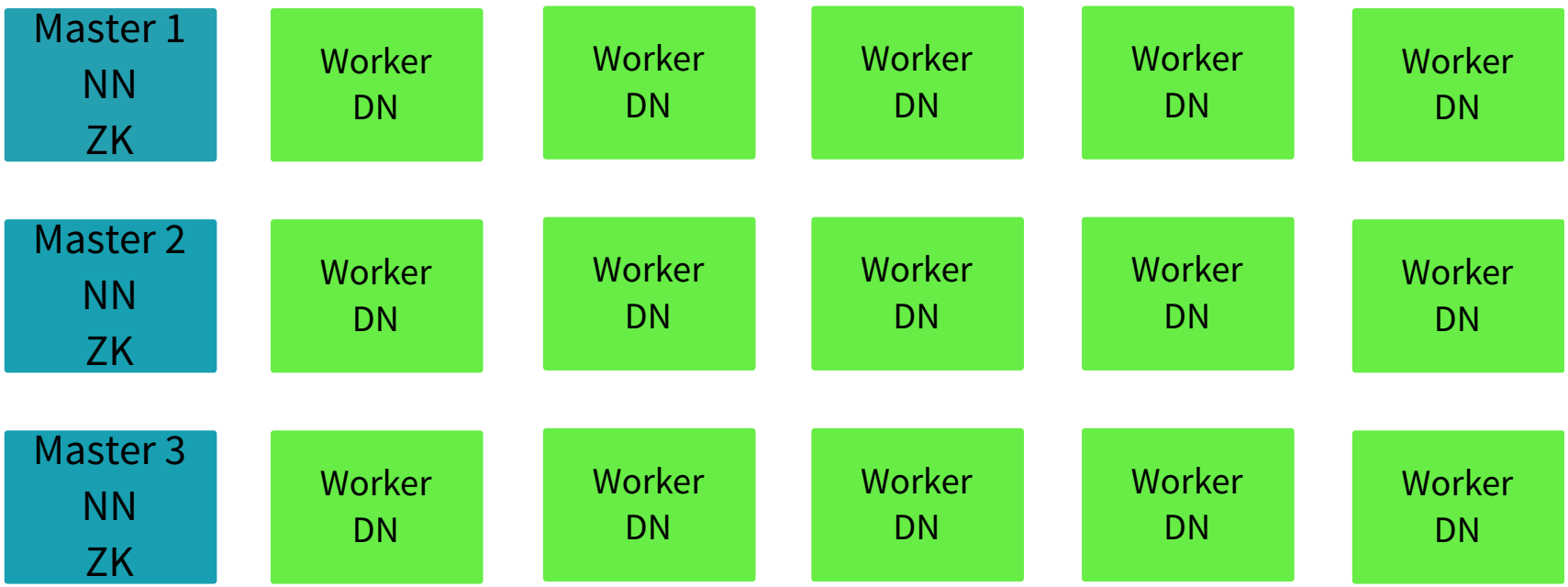

Elastic Apache Mesos
https://elastic.mesosphere.io
2 Enter your credentials
Your credentials will be used to start your Mesos cluster but will never be stored by Mesosphere.
AWS Access Key ID
Where do I find my AWS credentials?
AKIAJEJFLDYFPZTVEVWA
AWS Secret Access Key
Your public SSH key will be added to all instances of your cluster to allow you to access them.
Public SSH Key
How do I create an SSH key?
ssh-rsa
AAAAB3NzaC1yc2EAAAABIwAAAQEAxVgAQWi47cu/D8y7D5kEtKsYFeS+zNOKcap
ZmtRWeZjxsrSIiCpqjNoijGwbf5n/FapEVnFY6sh2P23s/QV78trRuJ6OvvhqilwYA728j
Tjv/9wO6nkT+ajxhNVe2252MYDopv+S39LrB42Ii5W53gFQi0sipy2K7M89jD/aDH9W
fBPFpkELCxhnbMdTWYqKF+ufbhS0E4oKAcVl0UguCtvphIC0nwfnaXRTYxYIFRfV9
DlSn5q5CCO9xof579uZ5l4OBlMFhX0I7mH0oZnQT+sVnWzOxDZR/2iZm3/wWbry/
LdAvKB81WdcJk3zdmHWDUSZOPxJpBJwd80kvXDF3Q== ceteri@gmail.com
I have read and agree to the Terms of Use and Privacy Policy


Elastic Apache Mesos
https://elastic.mesosphere.io
1
Lastly, choose where to receive notifications
You'll get an email when your cluster is ready to use with details on how to access it and how to shut it down.
Your email address
ceteri@gmail.com
Launch
powered by mesosphere
Terms
Privacy
Contact
Learn
© 2013 Mesosphere, Inc.


Elastic Apache Mesos
https://elastic.mesosphere.io
Confirm your email address
Review your email address one last time before launch. Once your new cluster is ready to go, you will get an email with details on how to access it and how to shut it down.
ceteri@gmail.com
Confirm & Launch
powered by mesosphere
Terms
Privacy
Contact
Learn
© 2013 Mesosphere, Inc.


Your cluster on Elastic Apa
https://elastic.mesosphere.io/clusters/provisioning
Elastic Apache Mesos
Your Apache Mesos cluster is being provisioned
When it's ready, you'll receive an email at ceteri@gmail.com.
This should take about 15 minutes. In the mean time, learn what you can do with your new Apache Mesos cluster with Mesosphere's tutorials. The email will contain personalized links to these to save you some typing.
Read through the Mesos documentation to get more information about Mesos.
We want to make it really easy to try out Apache Mesos, so your cluster will have very relaxed security group settings. Please make sure to lock down access before using it for anything confidential.
powered by mesosphere
Terms Privacy Contact Learn
© 2013 Mesosphere, Inc.

Elastic Mesos



More

Success! Your Apache Mesos cluster is ready

Inbox x

The Mesosphere Team support@mesosphere.io via mail4.wdc04.manc 3:20 PM (2 minutes ago)

to me

Great news,

Your Apache Mesos cluster is up and running!

View running frameworks and tasks in your Mesos UI:
http://54.235.3.97:5050

To get started with Hadoop on Mesos, visit the Hadoop on Mesos quickstart tutorial.
For more advanced Hadoop use, visit the Package Hadoop for Mesos tutorial.
For fast in-memory number crunching, try the Apache Spark on Mesos tutorial.
To run repeating jobs via Chronos, visit the Chronos on Mesos tutorial.
To learn how to build ETL pipelines, try ETL with Chronos and Hadoop.
And to run your long-lived services, visit the Marathon tutorial.

The IP addresses of your Mesos master nodes are: 54.235.3.97 54.204.193.157 54.204.133.148
Your Mesos slave nodes are: 54.234.5.237 54.221.134.235 54.227.20.164

You can connect to your instances via ssh using the "ubuntu" user.

Shutdown your cluster on Elastic Apache Mesos:
https://elastic.mesosphere.io/clusters/54.235.3.97

View running instances on AWS:
https://console.aws.amazon.com/ec2/v2/home?region=us-east-1

Have fun,
The Mesosphere Team
http://mesosphere.io/

More

Success! Your Apache Mesos cluster is ready

Inbox x

The Mesosphere Team support@mesosphere.io via mail4.wdc04.manc 3:20 PM (2 minutes ago)

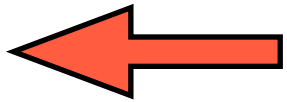
to me

Great news,

Your Apache Mesos cluster is up and running!

View running frameworks and tasks in your Mesos UI:
http://54.235.3.97:5050

To get started with Hadoop on Mesos, visit the Hadoop on Mesos quickstart tutorial.
For more advanced Hadoop use, visit the Package Hadoop for Mesos tutorial.
For fast in-memory number crunching, try the Apache Spark on Mesos tutorial.
To run repeating jobs via Chronos, visit the Chronos on Mesos tutorial.
To learn how to build ETL pipelines, try ETL with Chronos and Hadoop.
And to run your long-lived services, visit the Marathon tutorial.



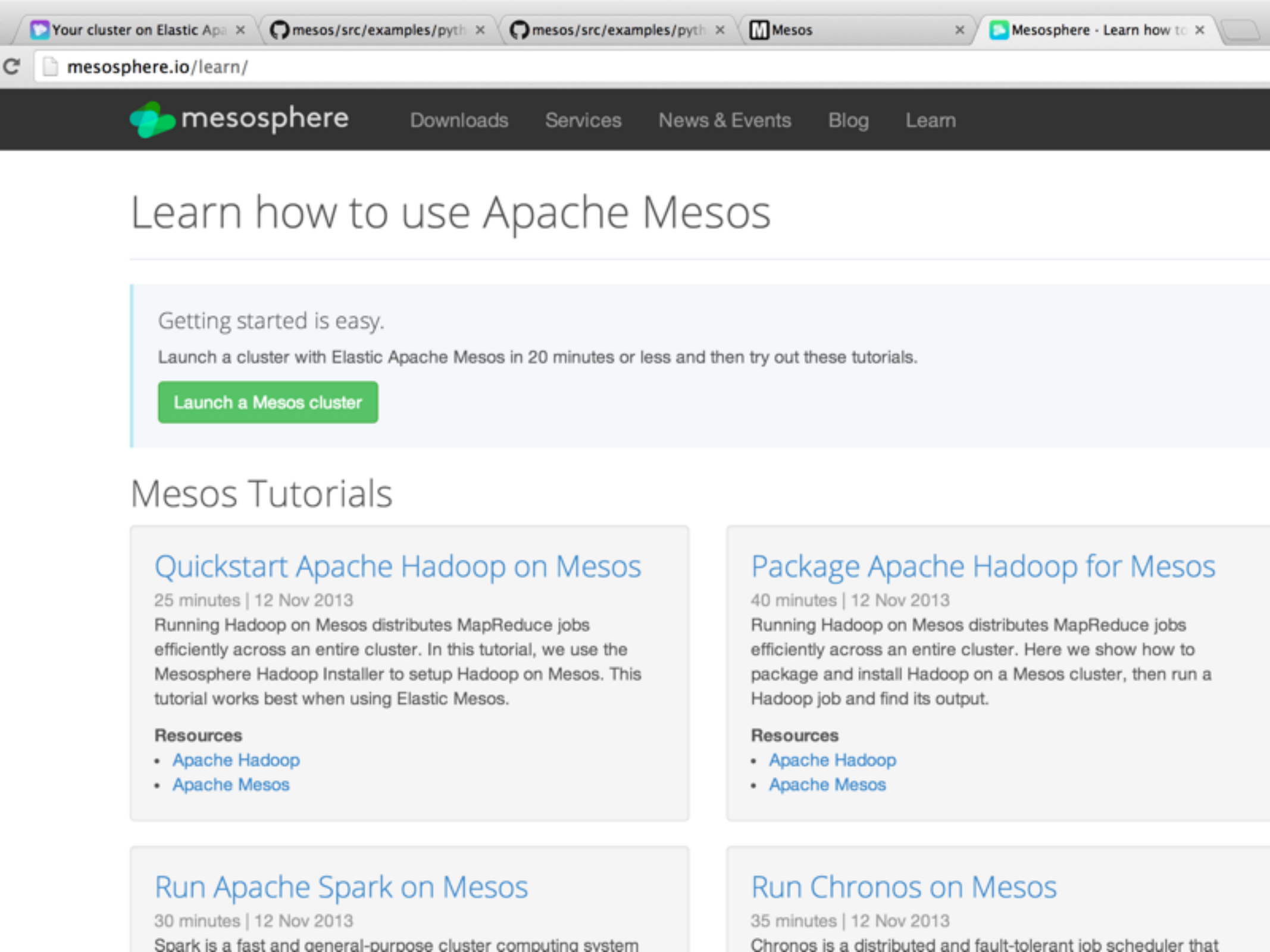
The IP addresses of your Mesos master nodes are: 54.235.3.97 54.204.193.157 54.204.133.148
Your Mesos slave nodes are: 54.234.5.237 54.221.134.235 54.227.20.164

You can connect to your instances via ssh using the "ubuntu" user.

Shutdown your cluster on Elastic Apache Mesos:
https://elastic.mesosphere.io/clusters/54.235.3.97

View running instances on AWS:
https://console.aws.amazon.com/ec2/v2/home?region=us-east-1

Have fun,
The Mesosphere Team
http://mesosphere.io/


Your cluster on Elastic Apa
mesos/src/examples/pyth
mesos/src/examples/pyth
Mesos
Mesosphere - Learn how to
mesosphere.io/learn/
mesosphere
Downloads
Services
News & Events
Blog
Learn
Learn how to use Apache Mesos
Getting started is easy.
Launch a cluster with Elastic Apache Mesos in 20 minutes or less and then try out these tutorials.
Launch a Mesos cluster
Mesos Tutorials
Quickstart Apache Hadoop on Mesos
25 minutes | 12 Nov 2013
Running Hadoop on Mesos distributes MapReduce jobs efficiently across an entire cluster. In this tutorial, we use the Mesosphere Hadoop Installer to setup Hadoop on Mesos. This tutorial works best when using Elastic Mesos.
Resources
Apache Hadoop
Apache Mesos
Package Apache Hadoop for Mesos
40 minutes | 12 Nov 2013
Running Hadoop on Mesos distributes MapReduce jobs efficiently across an entire cluster. Here we show how to package and install Hadoop on a Mesos cluster, then run a Hadoop job and find its output.
Resources
Apache Hadoop
Apache Mesos
Run Apache Spark on Mesos
30 minutes | 12 Nov 2013
Spark is a fast and general-purpose cluster computing system
Run Chronos on Mesos
35 minutes | 12 Nov 2013
Chronos is a distributed and fault-tolerant job scheduler that

Resources
- Apache Hadoop
- Apache Mesos

Resources
- Apache Hadoop
- Apache Mesos

Run Apache Spark on Mesos

30 minutes | 12 Nov 2013

Spark is a fast and general-purpose cluster computing system which makes parallel jobs easy to write. This tutorial shows you how to run Spark on Mesos.

Resources
- Apache Mesos
- Apache Spark

Run Chronos on Mesos

35 minutes | 12 Nov 2013

Chronos is a distributed and fault-tolerant job scheduler that supports complex job topologies. This walkthrough shows how to install Chronos on a Mesos cluster, how to use Chronos' web UI to schedule a job, and how to navigate the Mesos web UI to find the job's output.

Resources
- Chronos
- Apache Mesos

ETL with Chronos and Hadoop

35 minutes | 12 Nov 2013

This tutorial shows you how to use Chronos to schedule a typical ETL pipeline that downloads some data, runs a Hadoop job, and prints the output, all on the same cluster.

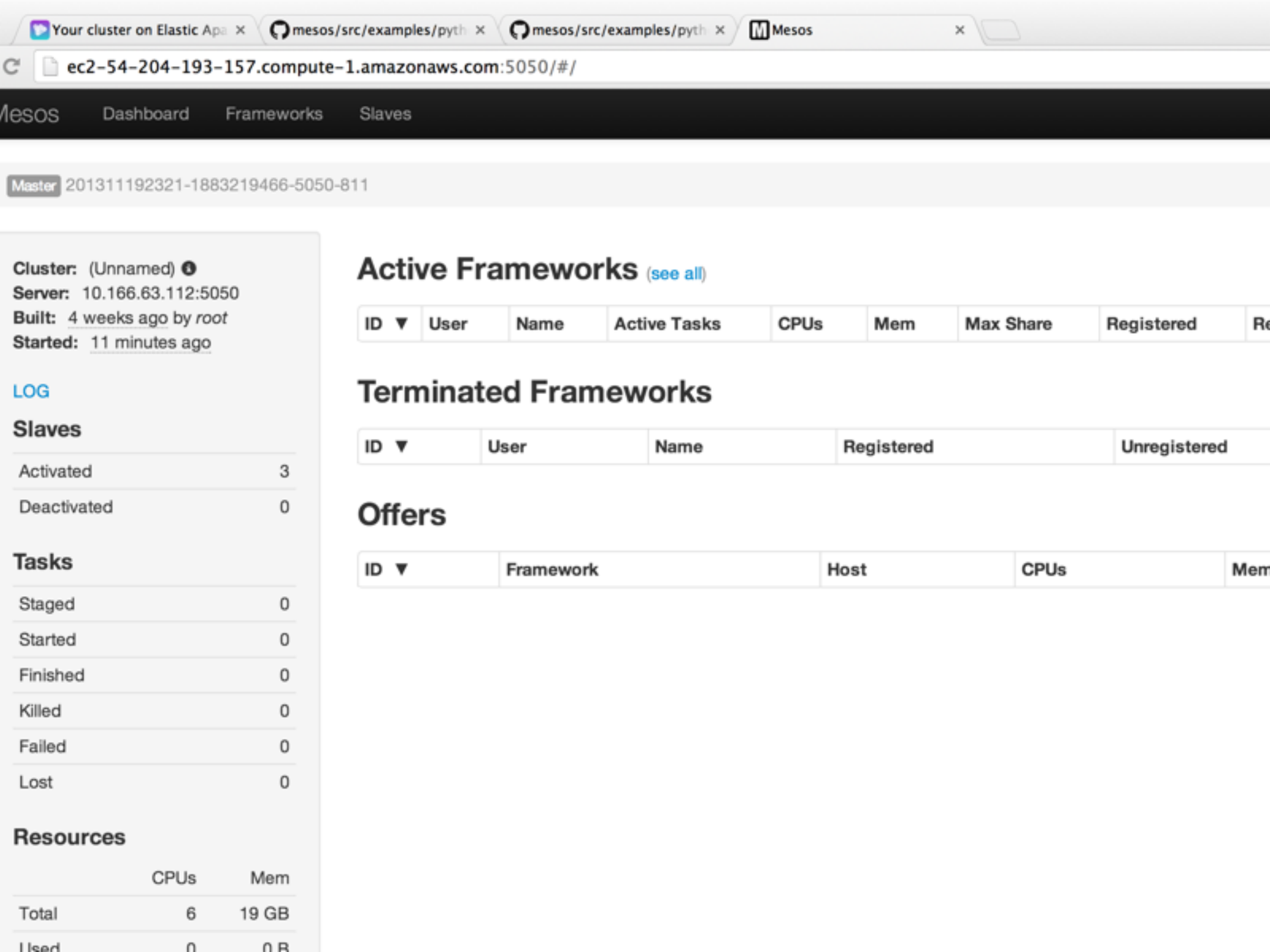
Resources
- Chronos
- Apache Mesos
- Apache Hadoop

Run services with Marathon

30 minutes | 12 Nov 2013

Marathon is a Mesos framework for long-running services. It ensures that a service stays up even when machines or entire racks fail. This tutorial shows you how to install Marathon on a Mesos cluster and run an example web app with it.

Resources
- Marathon
- Apache Mesos


Your cluster on Elastic Apa
mesos/src/examples/pyth
mesos/src/examples/pyth
Mesos
ec2-54-204-193-157.compute-1.amazonaws.com:5050/#/
Mesos
Dashboard
Frameworks
Slaves
Master 201311192321-1883219466-5050-811
Cluster: (Unnamed)
Server: 10.166.63.112:5050
Built: 4 weeks ago by root
Started: 11 minutes ago
LOG
Slaves
Activated 3
Deactivated 0
Tasks
Staged 0
Started 0
Finished 0
Killed 0
Failed 0
Lost 0
Resources
CPUs Mem
Total 6 19 GB
Active Frameworks (see all)
ID User Name Active Tasks CPUs Mem Max Share Registered
Terminated Frameworks
ID User Name Registered Unregistered
Offers
ID Framework Host CPUs

ありがとう
ございました

Calendar:

GlueCon
Broomfield, May 21
gluecon.com/2014/

DockerCon
SF, Jun 9
dockercon.com

Spark Summit
SF, Jun 30
spark-summit.org/2014

OSCON
PDX, Jul 20
oscon.com/oscon2014/

Strata NYC + Hadoop World
NYC, Oct 15
strataconf.com/stratany2014

New Book:

“Just Enough Math”,
with Allen Day @MapR Asia

advanced math for business people,
to leverage open source for Big Data

galleys: July 2014 @ OSCON
oscon.com/oscon2014/public/schedule/detail/34873

Enterprise Data Workflows with Cascading

O'Reilly, 2013

shop.oreilly.com/product/0636920028536.do

monthly newsletter for updates, events, conference summaries, etc.:

liber118.com/pxn/