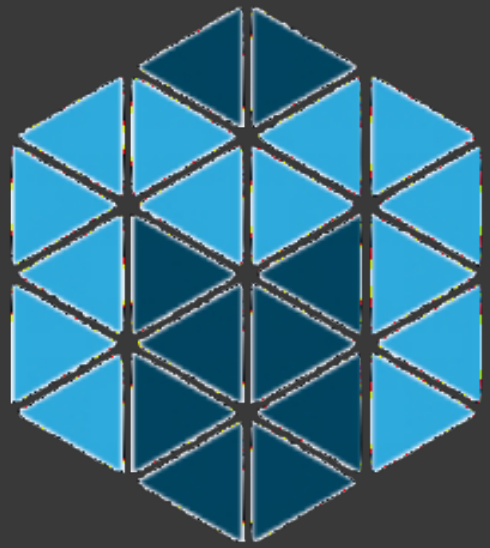
Apache Mesos as an SDK for Building Distributed Frameworks

Strata SC, 2014-02-13

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



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
- because JVM? not so much…
- reality: scheduling batch is simple; scheduling services is hard/expensive

The Modern Kernel: Top Linux Contributors...

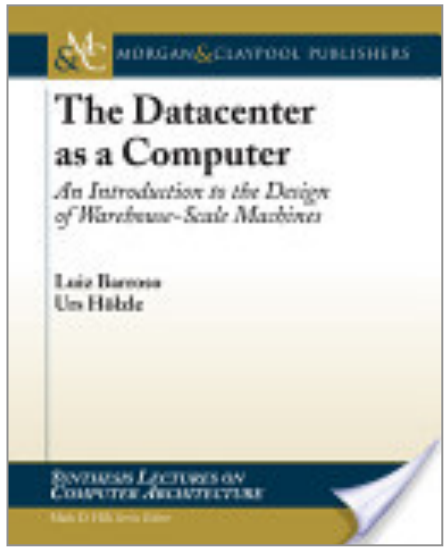
arstechnica.com/information-technology/2013/09/...

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%

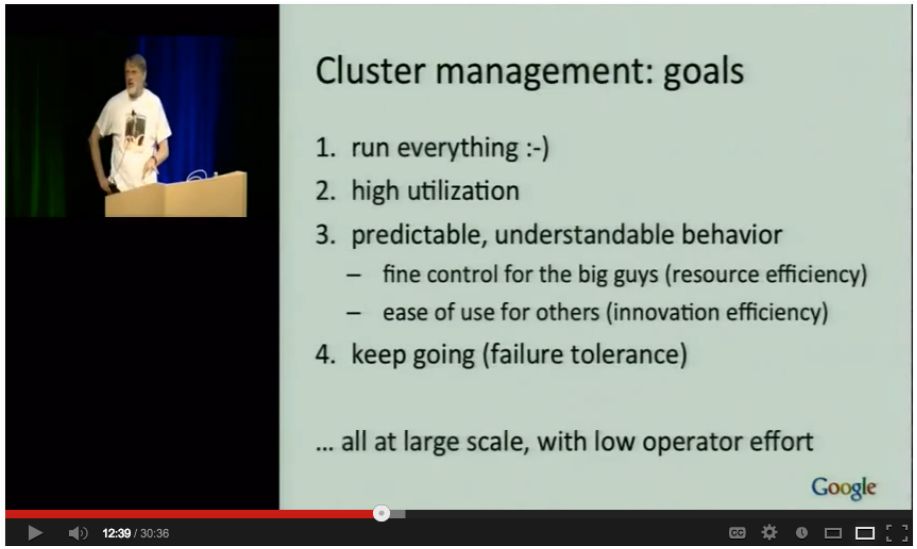
“Return of the Borg”

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



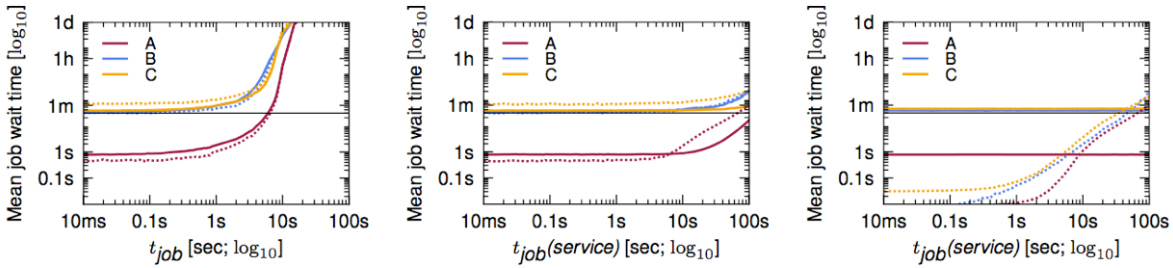
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines
Luiz André Barroso, Urs Hölzle
research.google.com/pubs/pub35290.html

2011 GAFS Omega
John Wilkes, et al.
youtu.be/0ZFMlO98Jkc



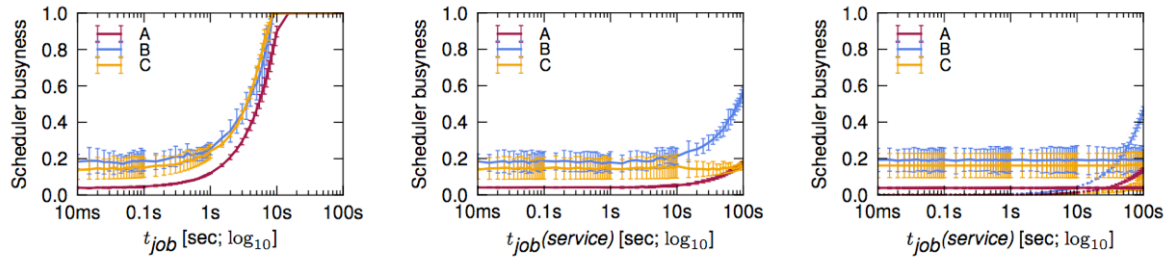
Google describes the technology…

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



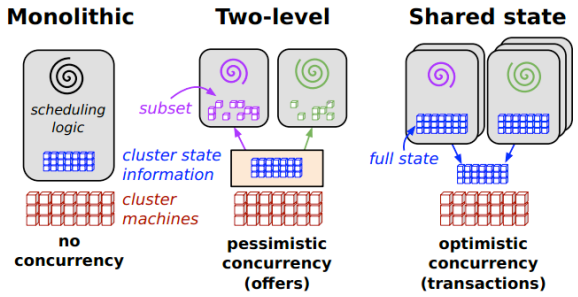
(a) Single-path. (b) Multi-path. (c) Shared state.

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**

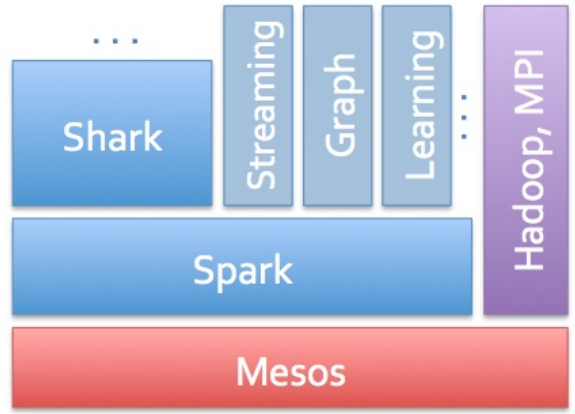
- *etc.*

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

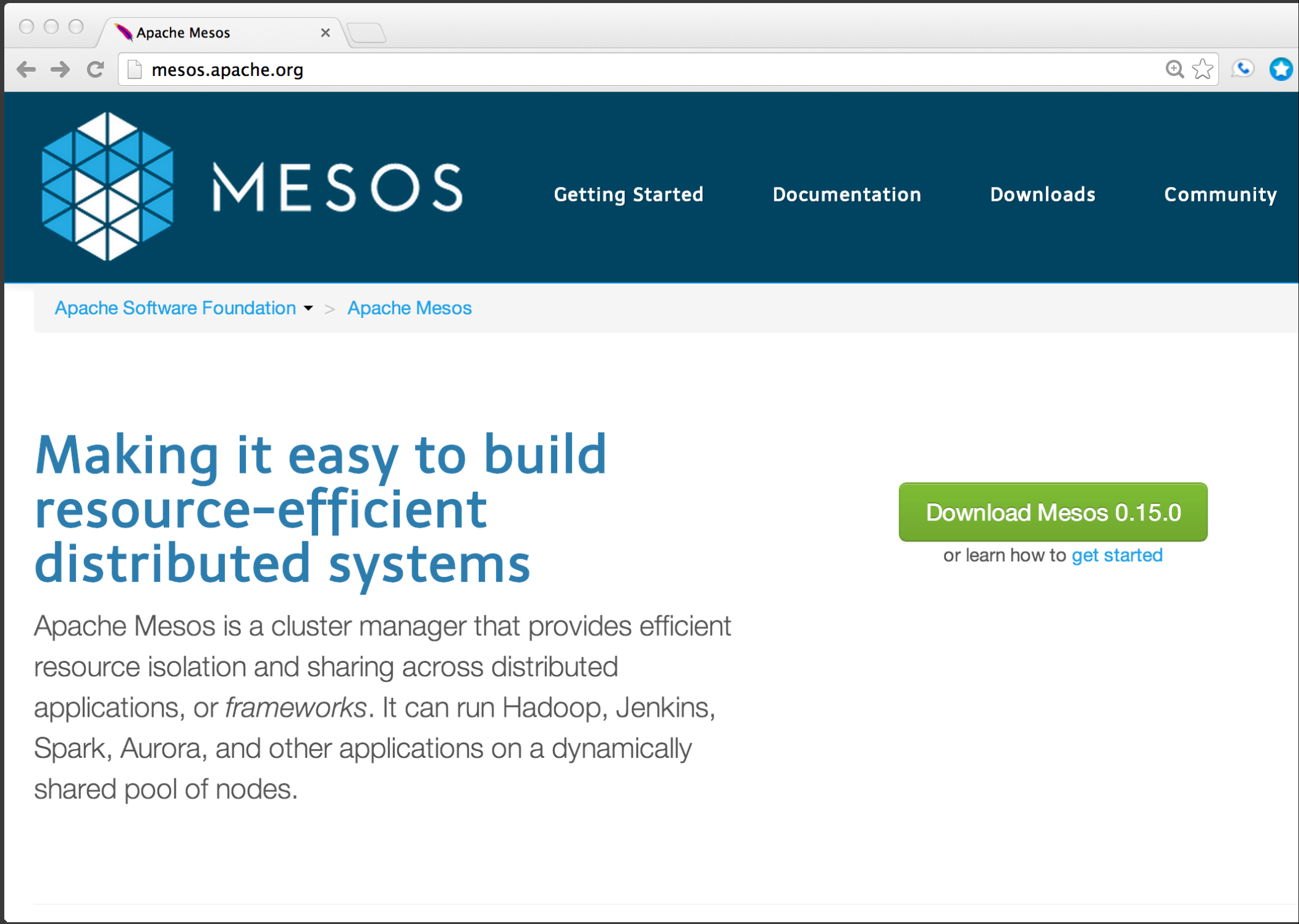


amplab
UC BERKELEY

Beyond Hadoop

keep in mind priorities for interdisciplinary efforts, to break down silos – extending beyond a de facto "priesthood" of data engineering

- higher utilization
- lower latency
- highly-available, long running services
- more than "Just JVM" – e.g., Python growth

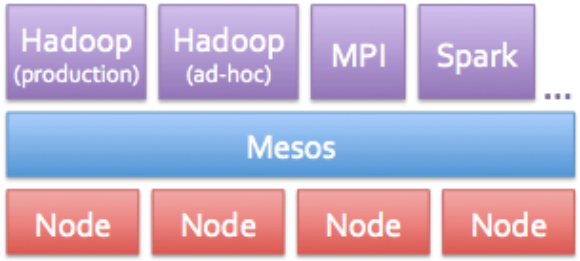

Apache Mesos
mesos.apache.org
MESOS
Getting Started
Documentation
Downloads
Community
Apache Software Foundation
Apache Mesos
Making it easy to build resource-efficient distributed systems
Apache Mesos is a cluster manager that provides efficient resource isolation and sharing across distributed applications, or *frameworks*. It can run Hadoop, Jenkins, Spark, Aurora, and other applications on a dynamically shared pool of nodes.
Download Mesos 0.15.0
or learn how to get started

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

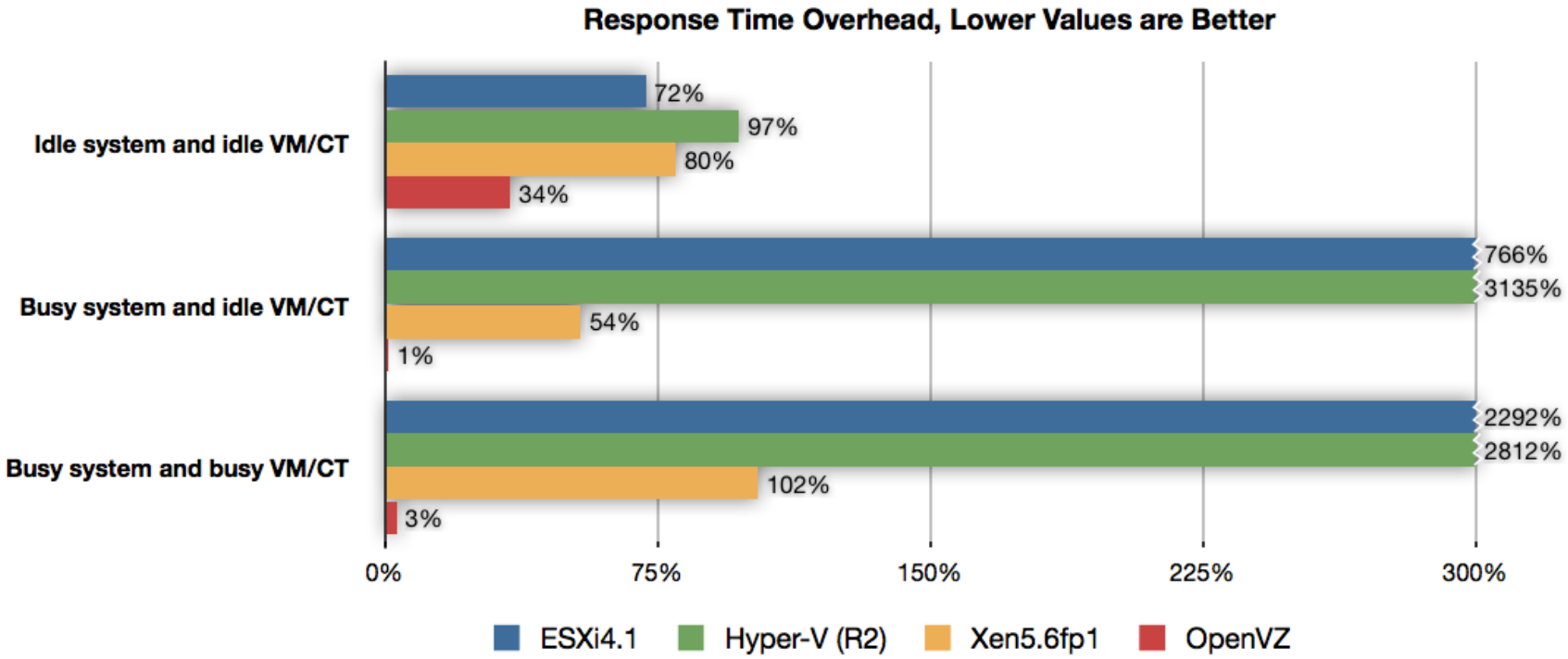


- top-level Apache project
- scalability to 10,000s of nodes
- obviates the need for virtual machines
- isolation (pluggable) for CPU, RAM, I/O, FS, etc.
- fault-tolerant leader election based on Zookeeper
- APIs in C++, Java, Python, Go
- web UI for inspecting cluster state
- available for Linux, OpenSolaris,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

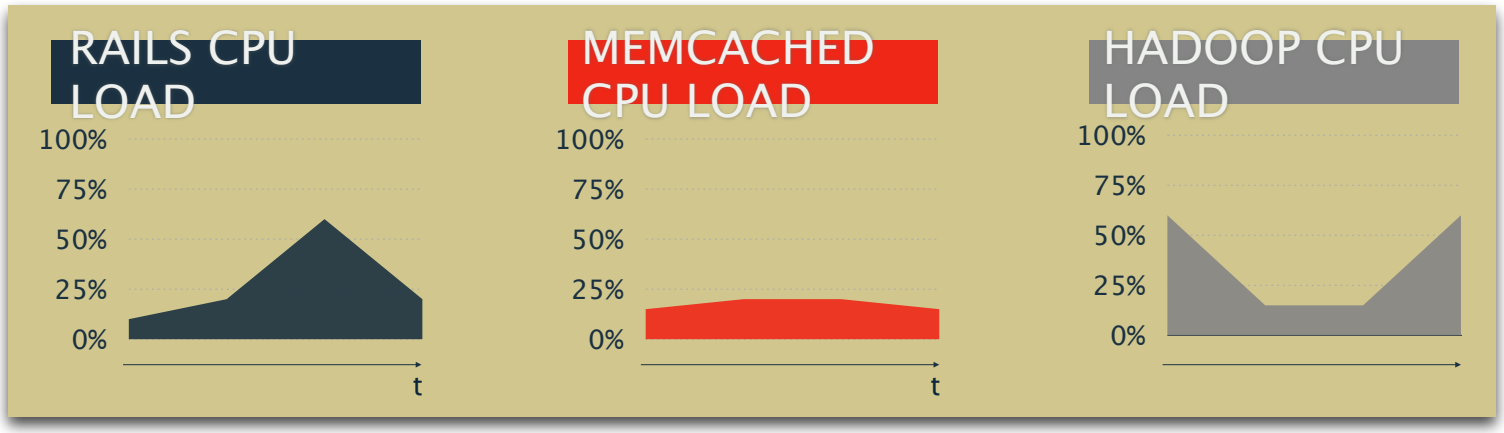
What are the costs of Virtualization?

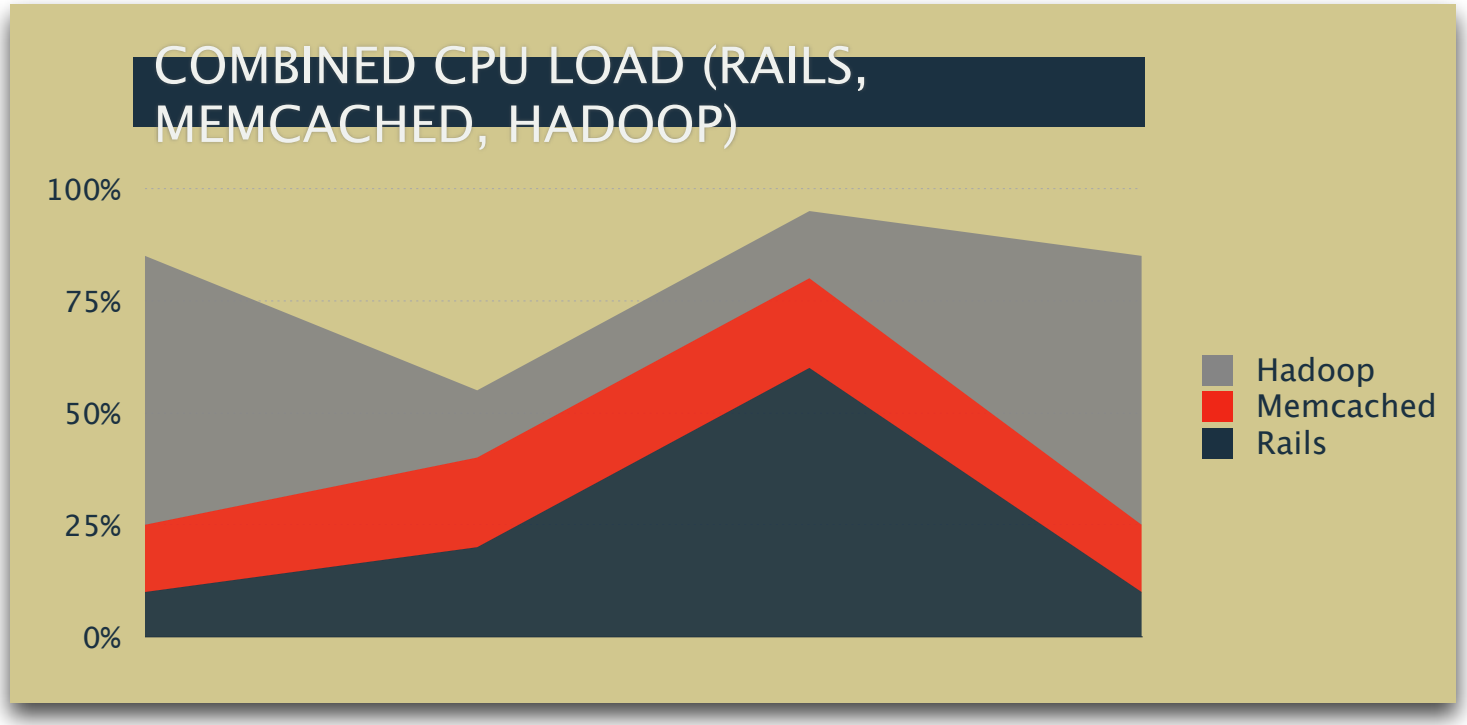
benchmark type	OpenVZ improvement
mixed workloads	210%-300%
LAMP (related)	38%-200%
I/O throughput	200%-500%
response time	order magnitude

more pronounced at higher loads



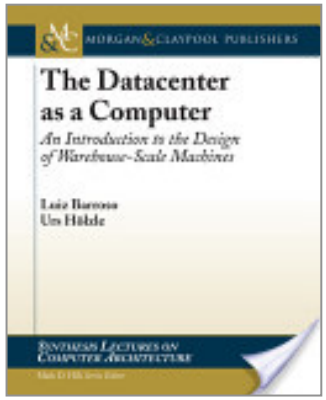
What are the costs of Single Tenancy?





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

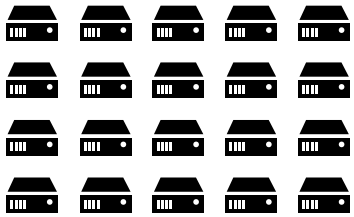
subtle benefits – arguably, more important for Enterprise IT:

- 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

DATACENTER

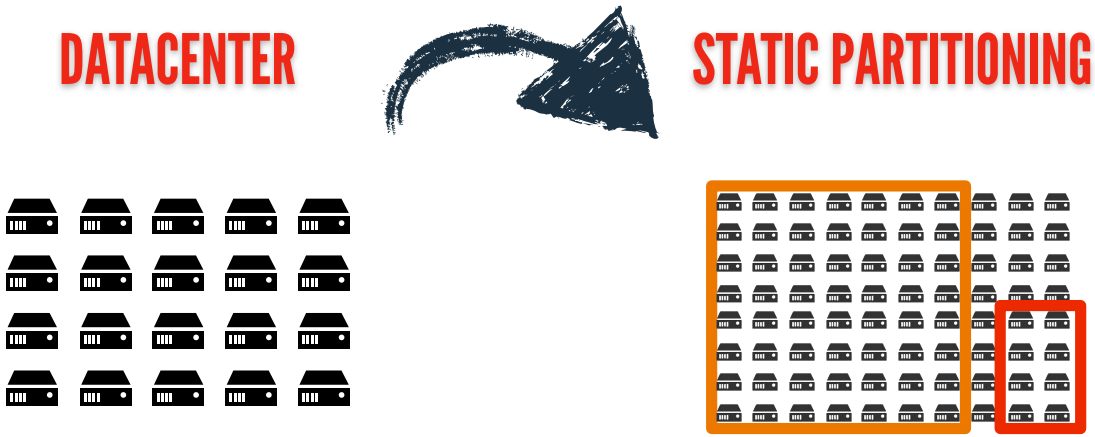


- *low utilization rates*
- *longer time to ramp up new services*

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*
- *static partitioning limits elasticity*

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, Cassandra, Hypertable, MPI

Python workloads:
DPark, Exelixi

Meta-Frameworks / HA Services:
Aurora, Marathon

Distributed Cron:
Chronos

Containers:
Docker

Fault-tolerant distributed systems...

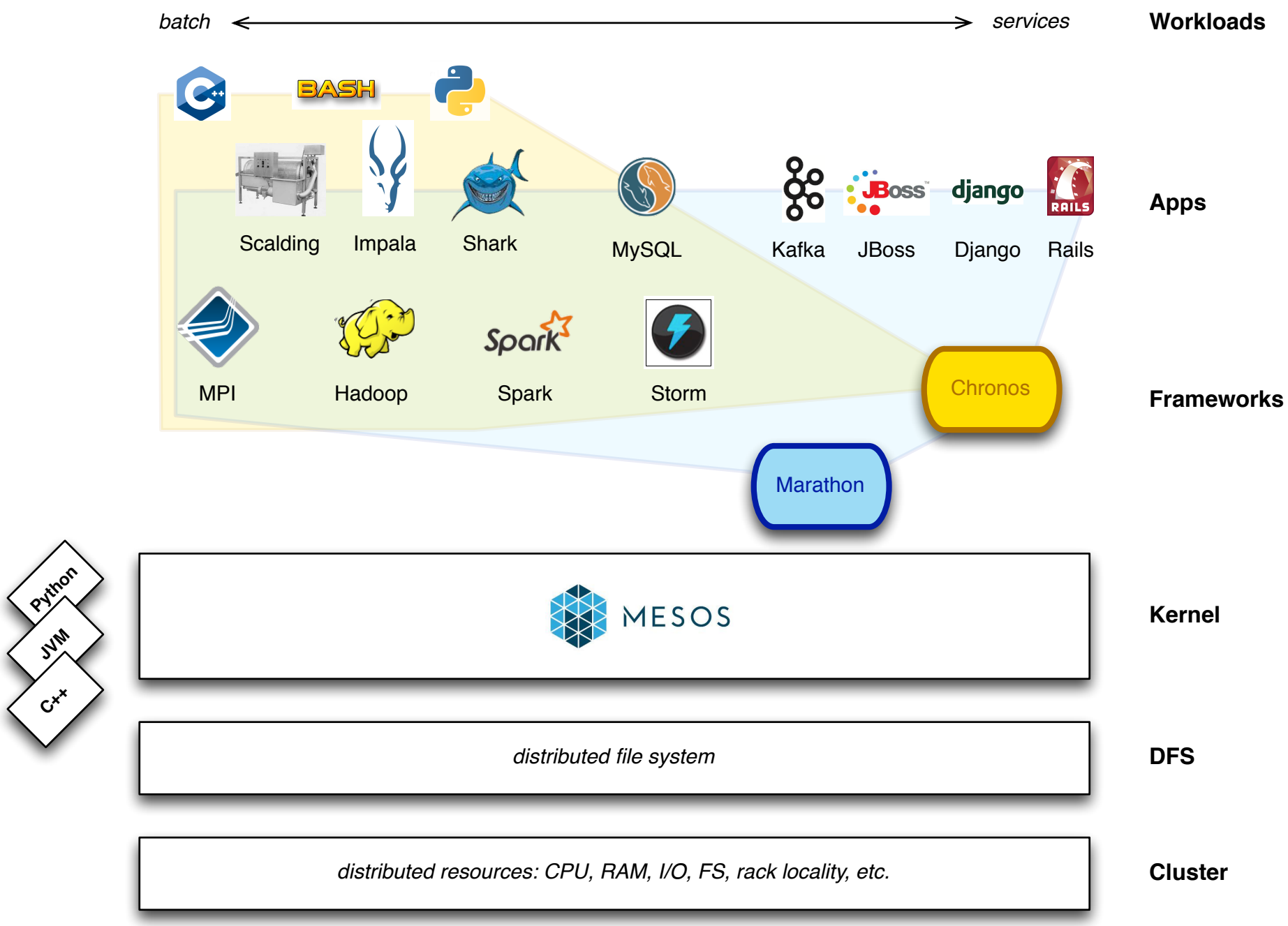
...written in 100-300 lines of C++, Java/Scala, Python, Go, etc.

...building blocks, if you will

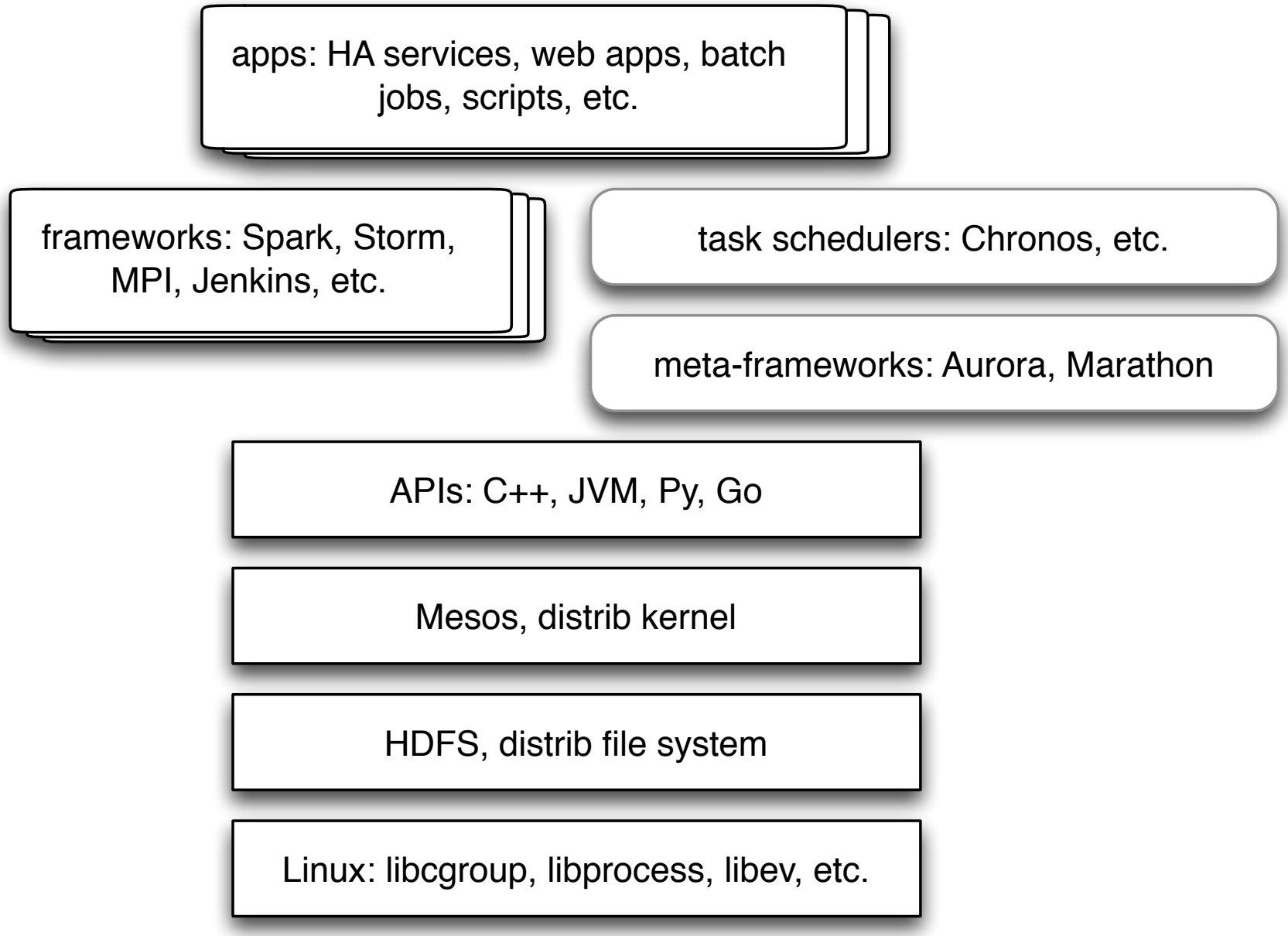
Q: required lines of network code?

A: probably none

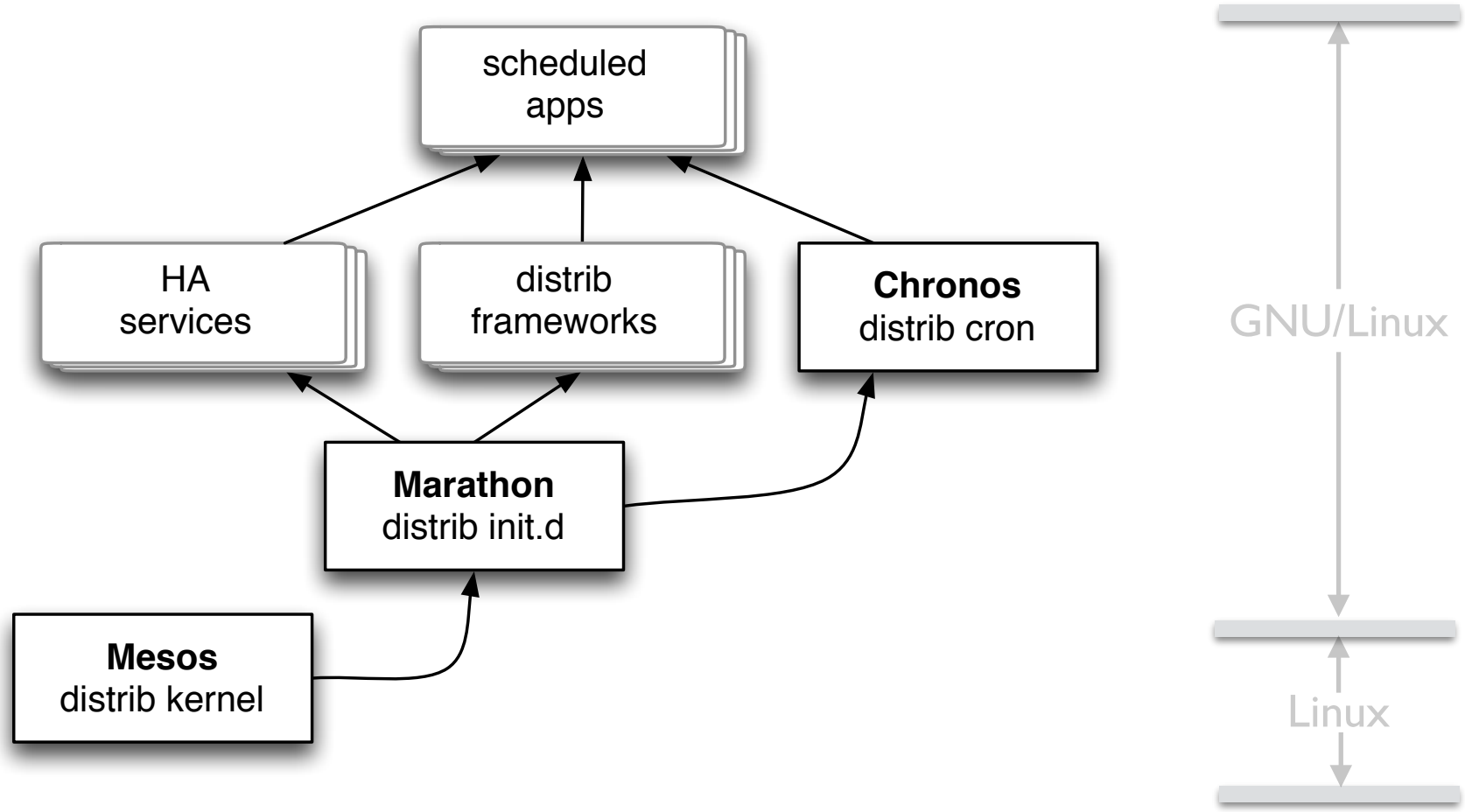
Mesos – architecture



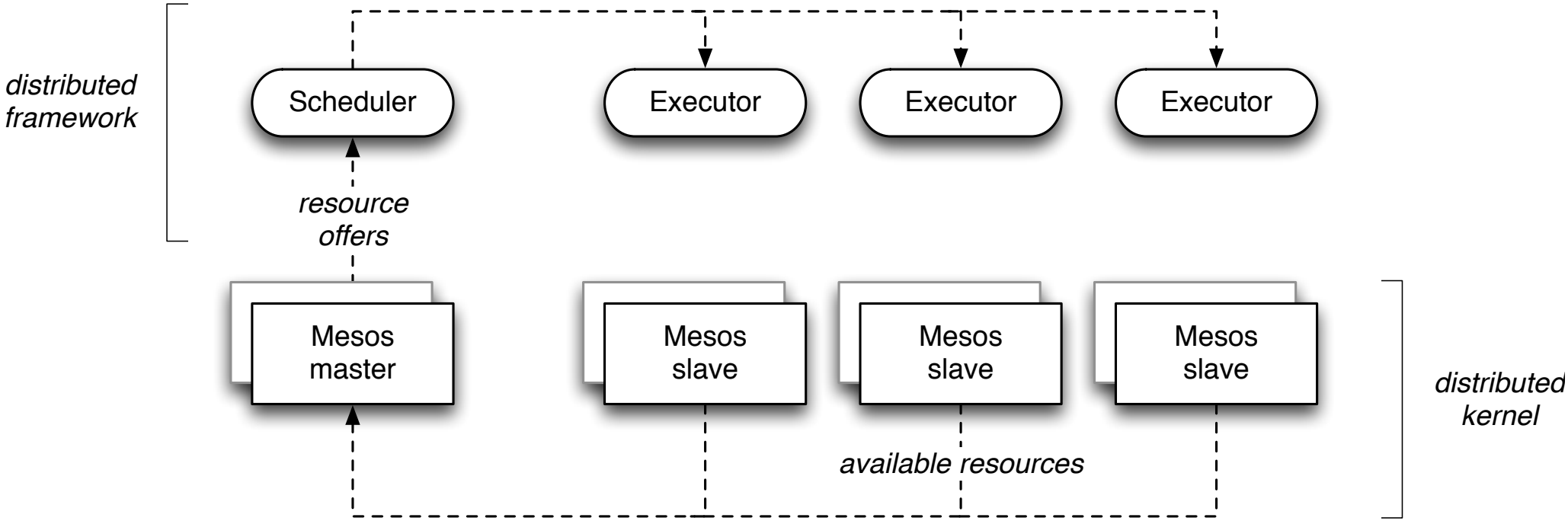
Mesos – architecture



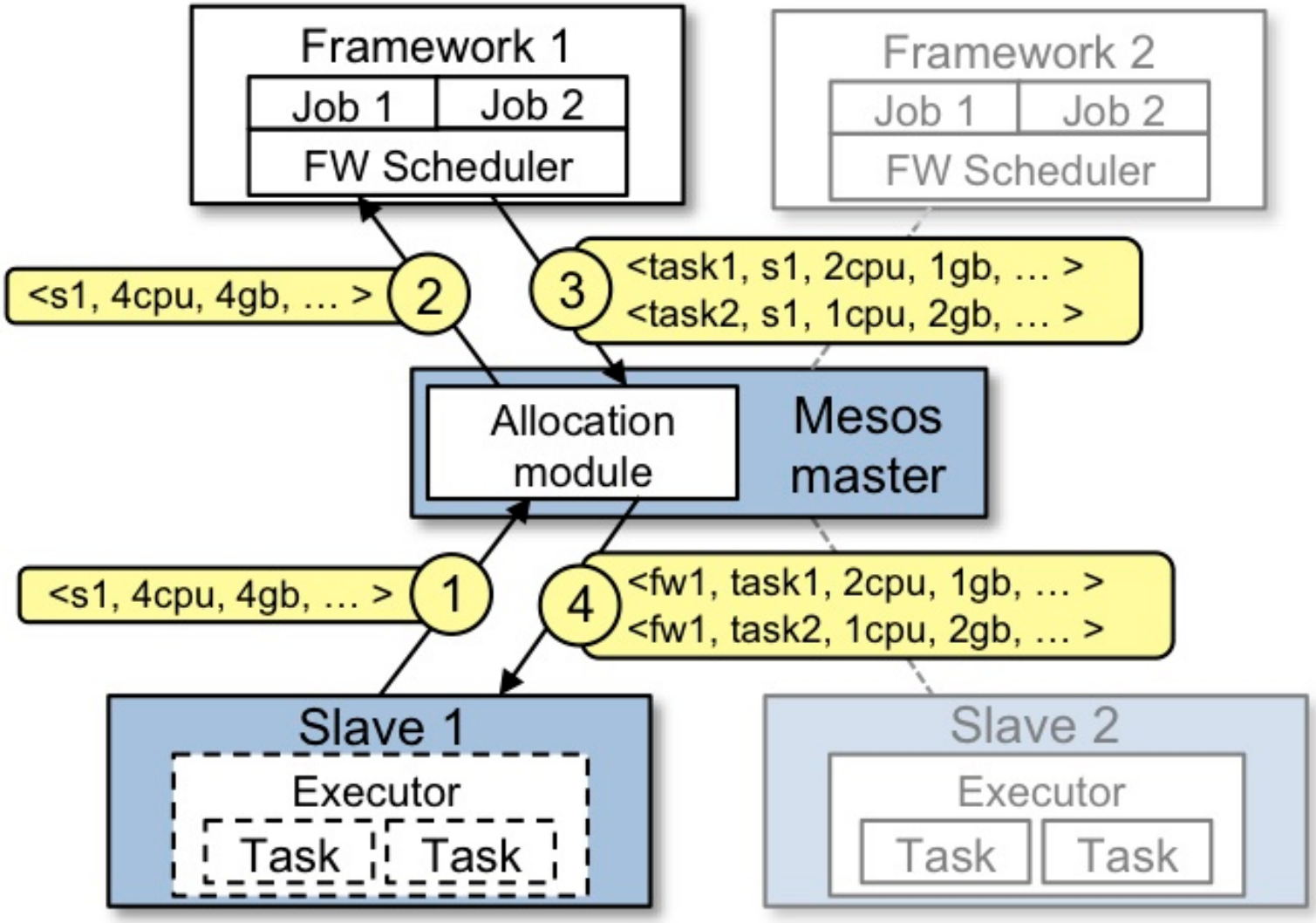
Mesos – dynamics



Mesos – dynamics



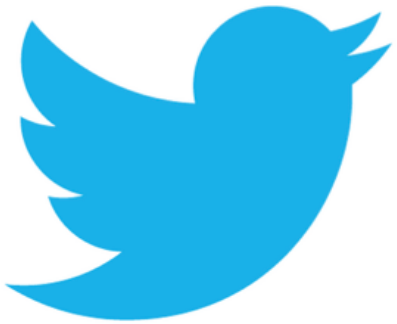
Example: Resource Offer in a Two-Level Scheduler



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

Because...
Use Cases

Production Deployments (public)



HubSpot

MediaCrossing™
BRIDGING THE MARKET™

vimeo

pinkbike.com

sharethrough

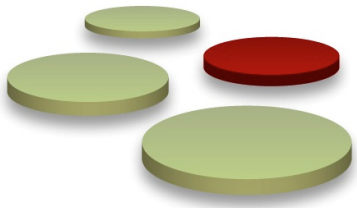
Sigmoid

redhat

xogito
[kä-gi-tō]

device scape™

CLOUD PHYSICS



iQIY 爱奇艺

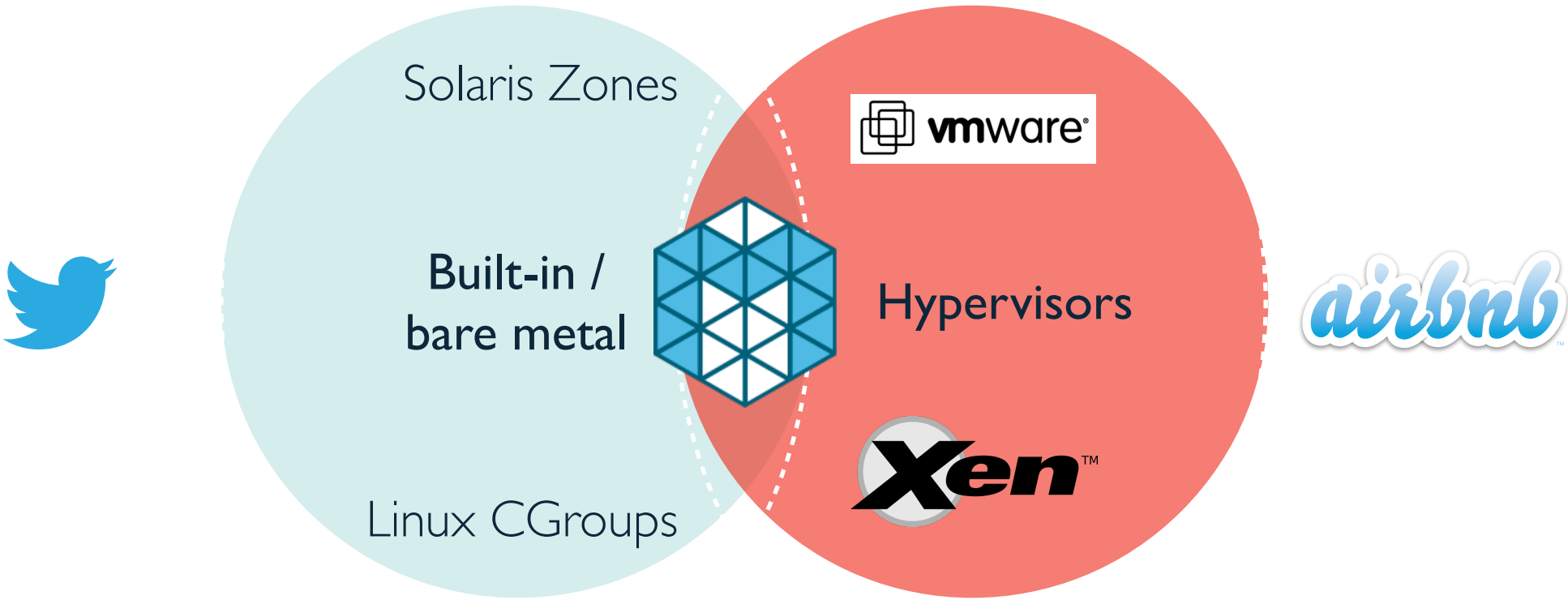
OOYALA®

OpenTable®

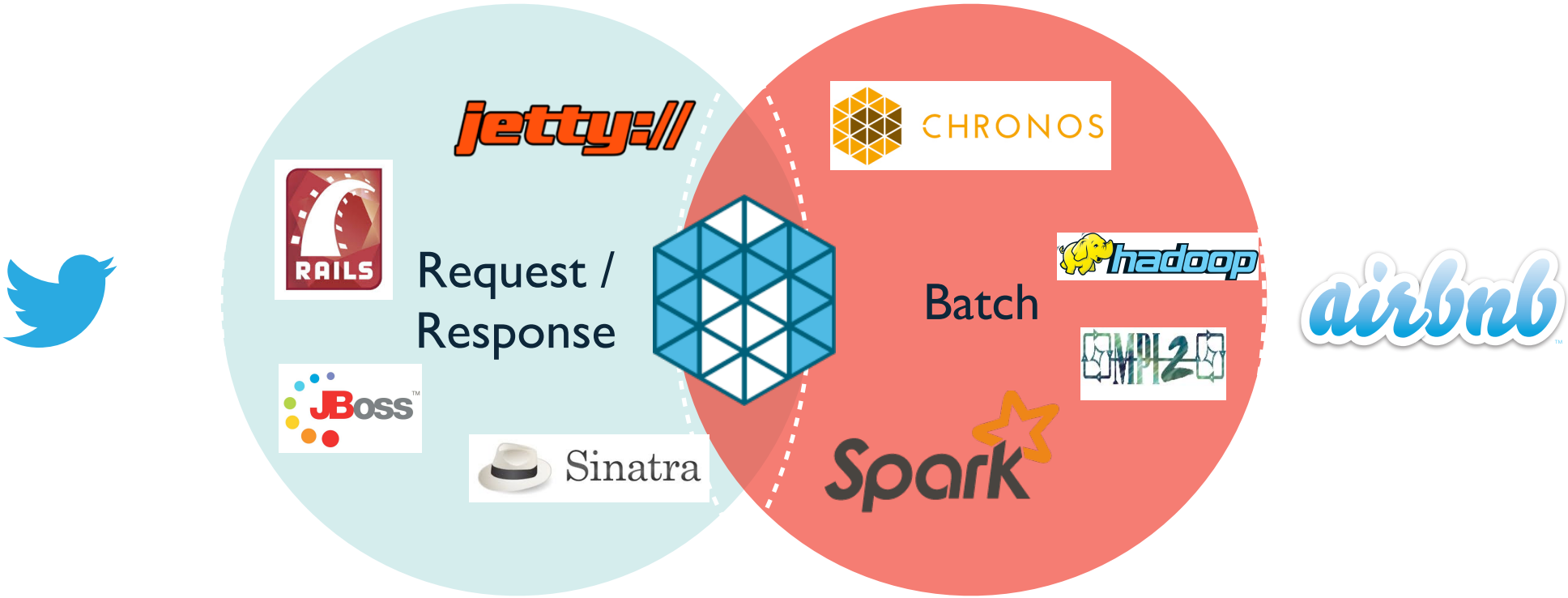
airbnb™
Travel like a human.

CATEGORIZE

Opposite Ends of the Spectrum, One Common Substrate

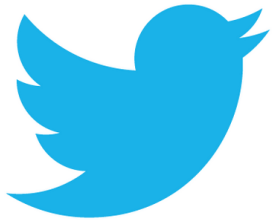


Opposite Ends of the Spectrum, One Common Substrate



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.

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

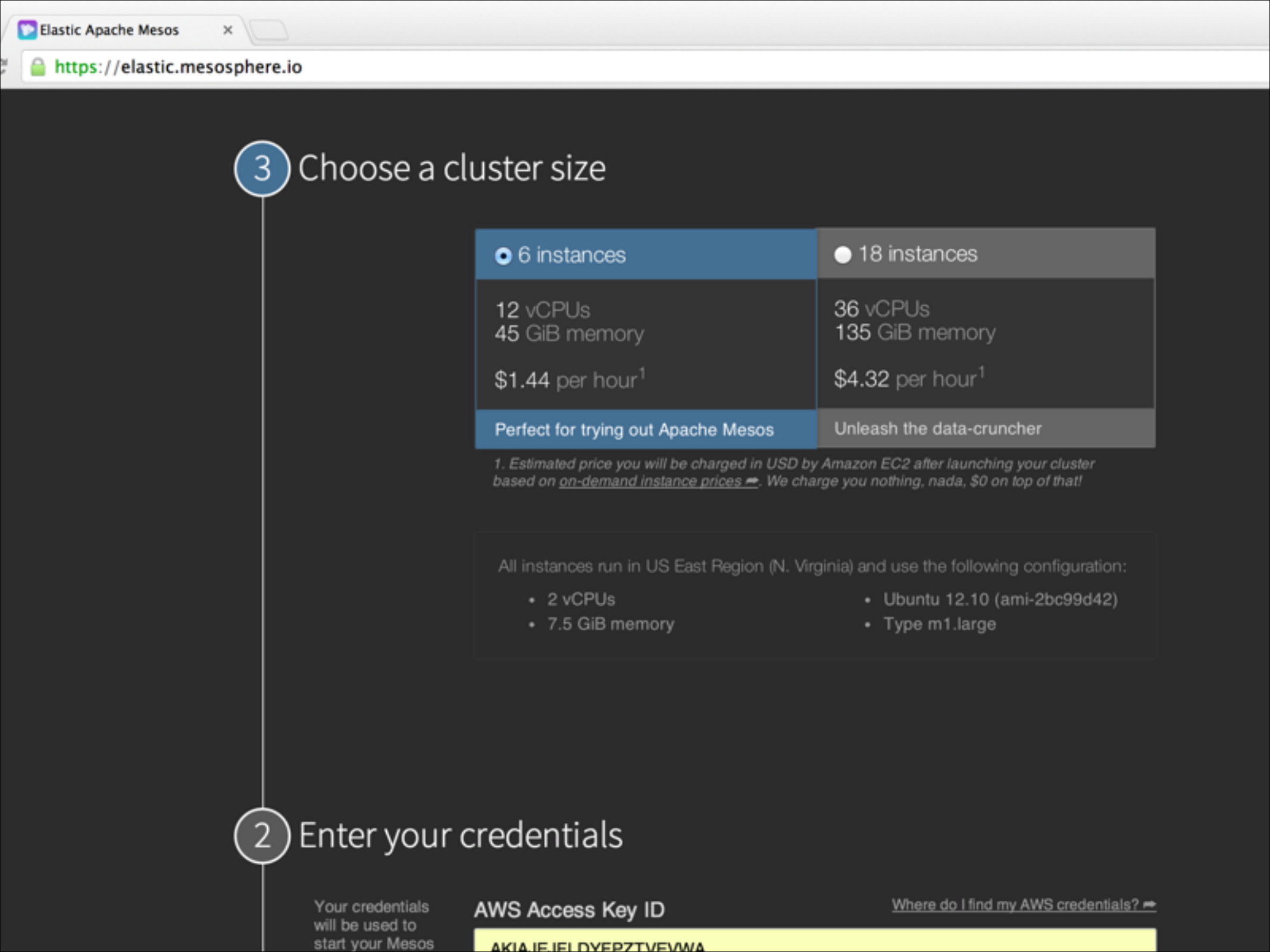
Elastic Apache Mesos is a web service that automates the creation of Apache Mesos clusters on Amazon Elastic Compute Cloud (EC2). It provisions EC2 instances, installs dependencies including Apache ZooKeeper and HDFS, and delivers you a cluster with all the services running.

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

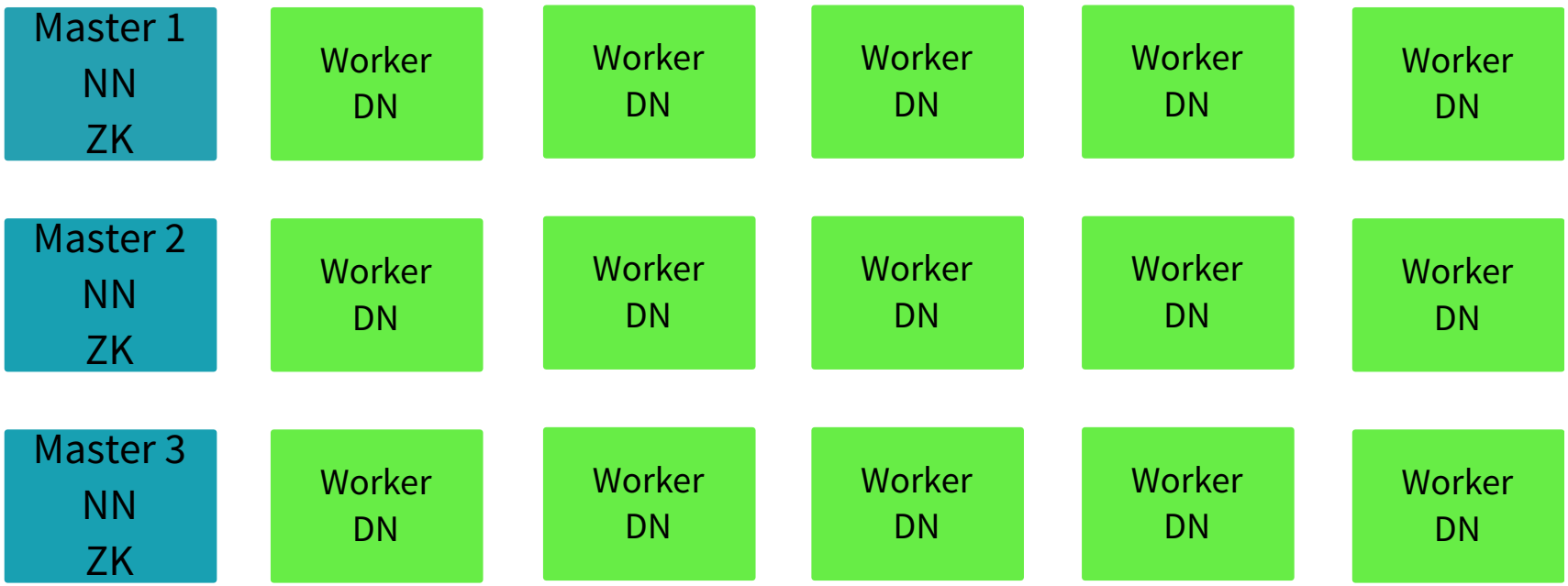
You just pay for your EC2 instances; Elastic Apache Mesos costs you nothing, nada, zilch on top of that.

3 Choose a cluster size

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

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/

Save the Date:
Apr 3, 2014
Mesos Summit
(or something)

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

O'REILLY®

Data

Search

Posts by Topic: Big Data Data Science Health Data Data Journalism Hadoop Visit oreilly.com

Learning Apache Mesos

Print

by Paco Nathan | @pacoid | +Paco Nathan | Comment | January 15, 2014

Listen

Read Later

Tweet 38 +1 10 Like 17 Share 12

In the summer of 2012, Accel Partners hosted an invitation-only Big Data conference at Stanford. Ping Li stood near the exit with a checkbook, ready to invest $1MM in pitches for real-time analytics on clusters. However, *real-time* means many different things. For MetaScale working on the Sears turnaround, real-time means shrinking a 6 hour window on a mainframe to 6 minutes on Hadoop. For a hedge fund, real-time means compiling Python to run on GPUs where milliseconds matter, or running on FPGA hardware for microsecond response.

With much
further fron
scale produ
a related p
cluster is a

Apache Mesos: Open Source Datacenter Computing

Print

by Paco Nathan | @pacoid | +Paco Nathan | Comment | January 8, 2014

Listen

Read Later

Tweet 89 +1 8 Like 45 Share 42

Virtual machines (VMs) have enjoyed a long history, from IBM's CP–40 in the late 1960s on through the rise of VMware in the late 1990s. Widespread VM use nearly became synonymous with "cloud computing" by the late 2000s: public clouds, private clouds, hybrid clouds, etc. One firm, however, bucked the trend: Google.

Google's datacenter computing leverages isolation in lieu of VMs. Public disclosure is limited, but the Omega paper from EuroSys 2013 provides a good overview. See also two YouTube videos: John Wilkes in 2011 GAFS Omega and Jeff Dean in Taming Latency Variability… For the business case, see an earlier Data blog post, that discusses how multi-tenancy and efficient utilization translates into improved ROI.

Former Airbnb engineers simplify Mesos to manage data jobs in the cloud
Jordan Novet
VentureBeat (2013-11-12)
venturebeat.com/2013/11/12/former-airbnb-engineers-simplify...

Mesosphere Adds Docker Support To Its Mesos-Based Operating System For The Data Center
Frederic Lardinois
TechCrunch (2013-09-26)
techcrunch.com/2013/09/26/mesosphere...

Play Framework Grid Deployment with Mesos
James Ward, **Flo Leibert**, et al.
Typesafe blog (2013-09-19)
typesafe.com/blog/play-framework-grid...

Mesosphere Launches Marathon Framework
Adrian Bridgwater
Dr. Dobbs (2013-09-18)
drdobbs.com/open-source/mesosphere...

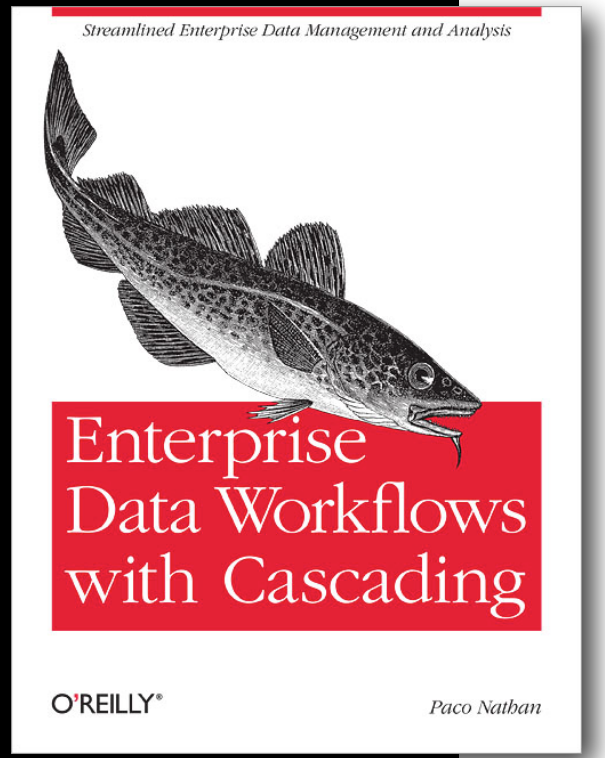
New open source tech Marathon wants to make your data center run like Google's
Derrick Harris
GigaOM (2013-09-04)
gigaom.com/2013/09/04/new-open-source...

Running batch and long-running, highly available service jobs on the same cluster
Ben Lorica
O'Reilly (2013-09-01)
strata.oreilly.com/2013/09/running-batch...

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/