FlinkForward 2017 - San Francisco

Flink meet DC/OS

Deploying Apache Flink at Scale

Elizabeth K. Joseph, @pleia2
Ravi Yadav, @RaaveYadav

MESOSPHERE

© 2017 Mesosphere, Inc. All Rights Reserved.

Talk Outline

Part 1

Introduction to Apache Mesos, Marathon, and DC/OS

Part 2

Demonstration of demo data pipeline + Installing Flink on DC/OS

Part 3

DC/OS 1.9 key features for data services and beyond

© 2017 Mesosphere, Inc. All Rights Reserved.

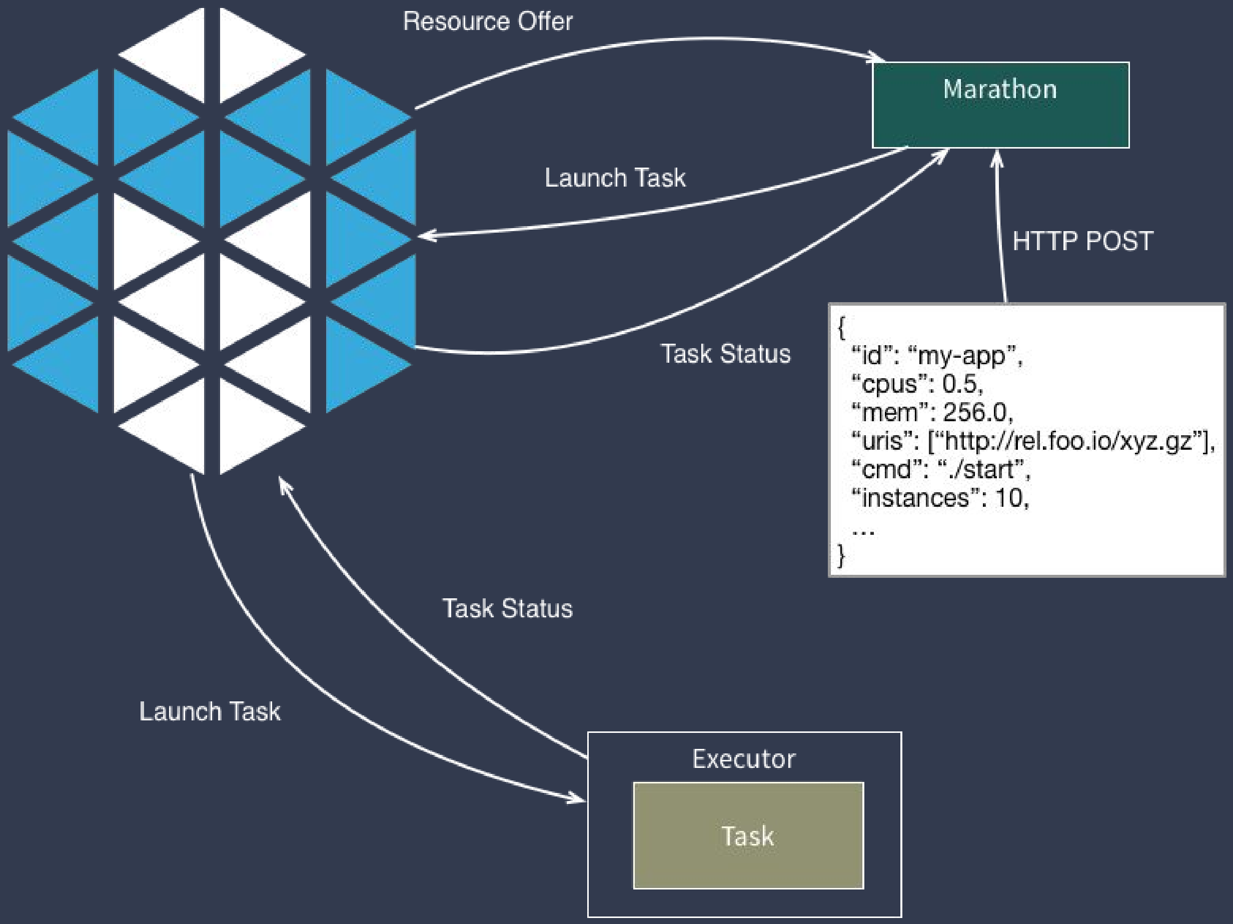
Apache Mesos:
The datacenter kernel

http://mesos.apache.org/

© 2017 Mesosphere, Inc. All Rights Reserved.

Marathon

- Mesos can’t run applications on its own.
- A Mesos framework is a distributed system that has a scheduler.
- Schedulers like Marathon start and keep your applications running. A bit like a distributed init system.
- Mesos mechanics are fair and HA
- Learn more at https://mesosphere.github.io/marathon/



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

© 2017 Mesosphere, Inc. All Rights Reserved.

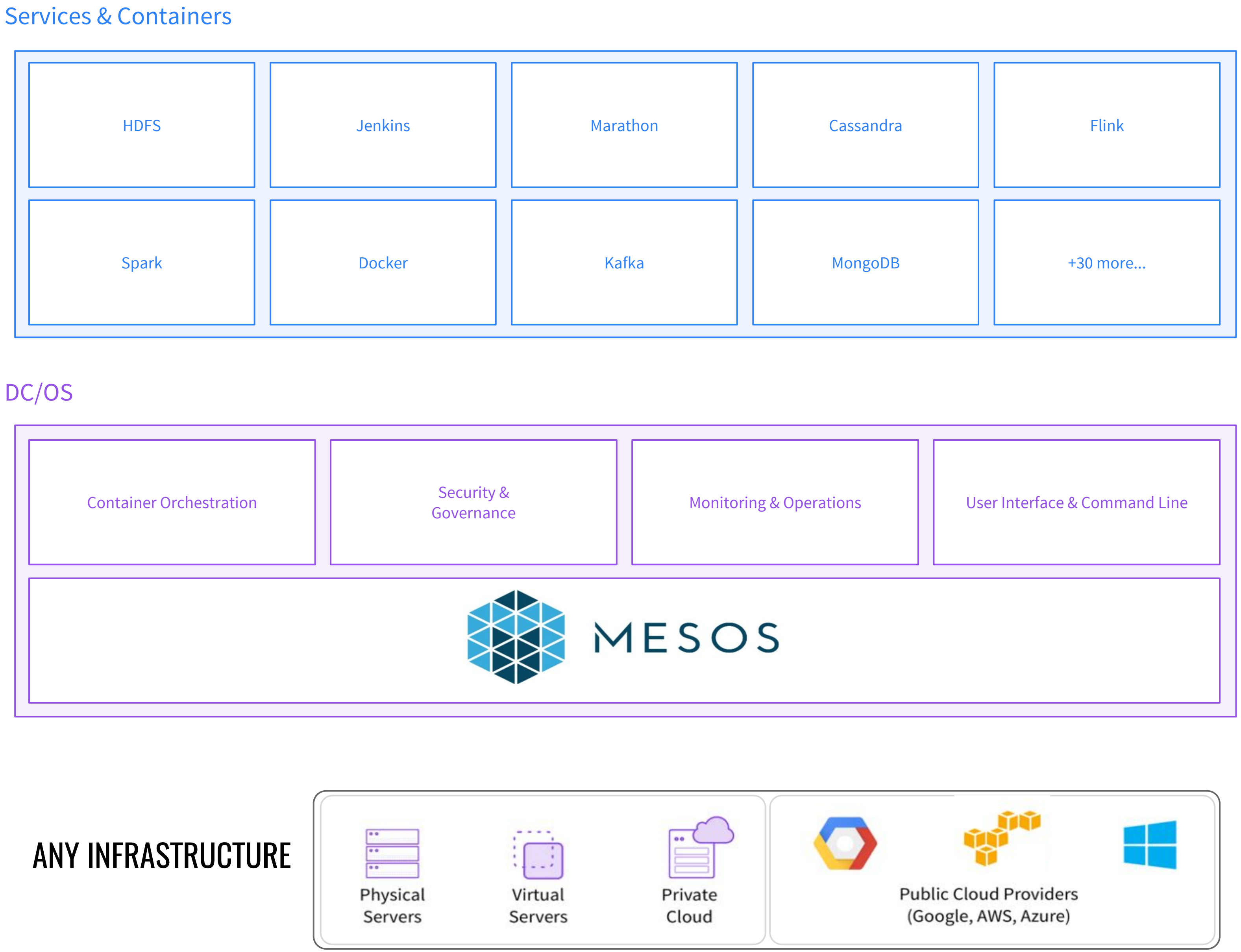
Introducing DC/OS

Solves common problems

- Resource management
- Task scheduling
- Container orchestration
- Self-healing infrastructure
- Logging and metrics
- Network management
- “Universe” of pre-configured apps (including Flink, Kafka…)
- Learn more and contribute at https://dcos.io/

© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS Architecture Overview



Services & Containers

HDFS
Jenkins
Marathon
Cassandra
Flink
Spark
Docker
Kafka
MongoDB
+30 more...

DC/OS

Container Orchestration
Security & Governance
Monitoring & Operations
User Interface & Command Line

MESOS

ANY INFRASTRUCTURE

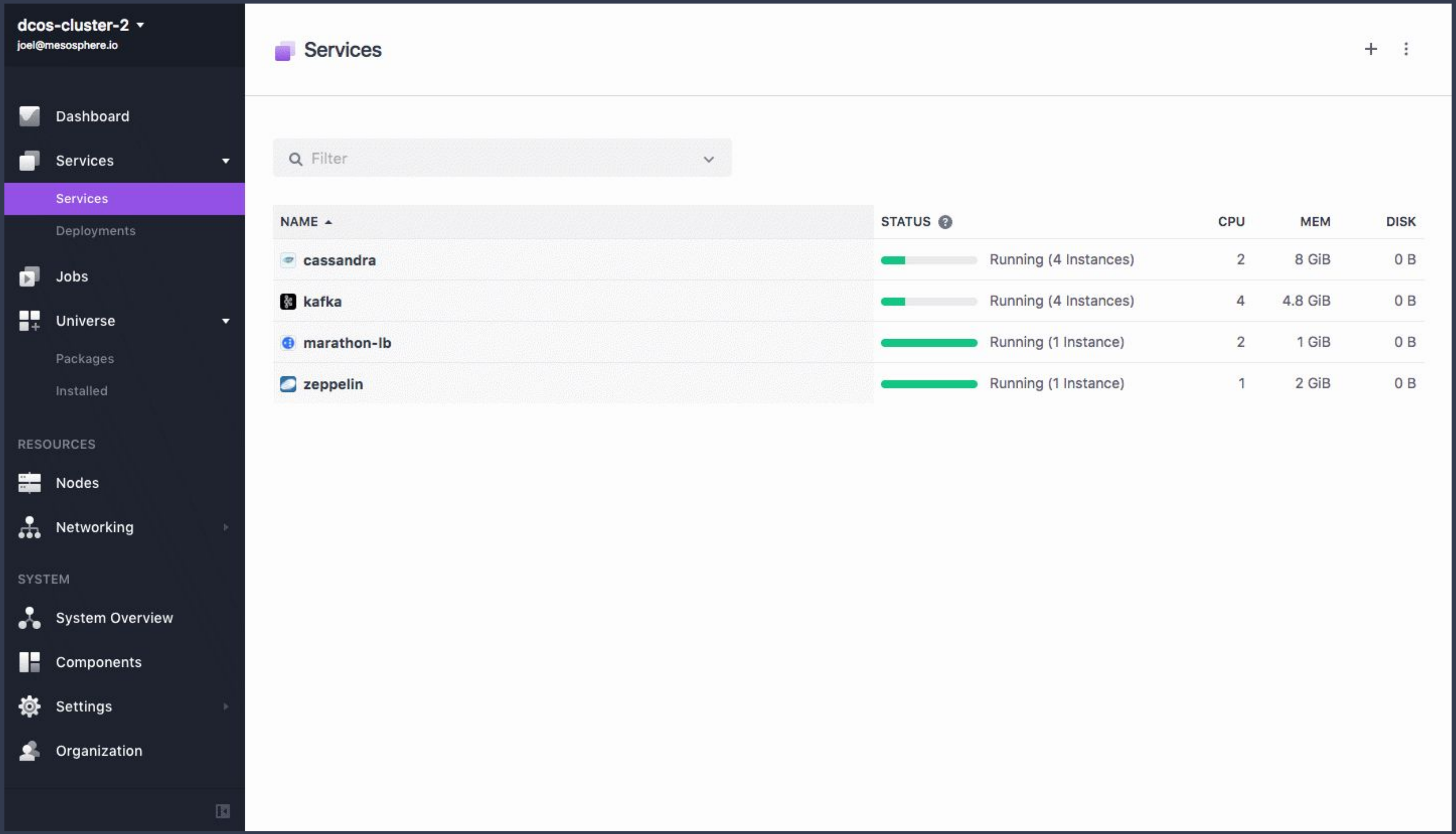
Physical Servers
Virtual Servers
Private Cloud
Public Cloud Providers (Google, AWS, Azure)

© 2017 Mesosphere, Inc. All Rights Reserved.

Interact with DC/OS (1/2)

Web-based GUI

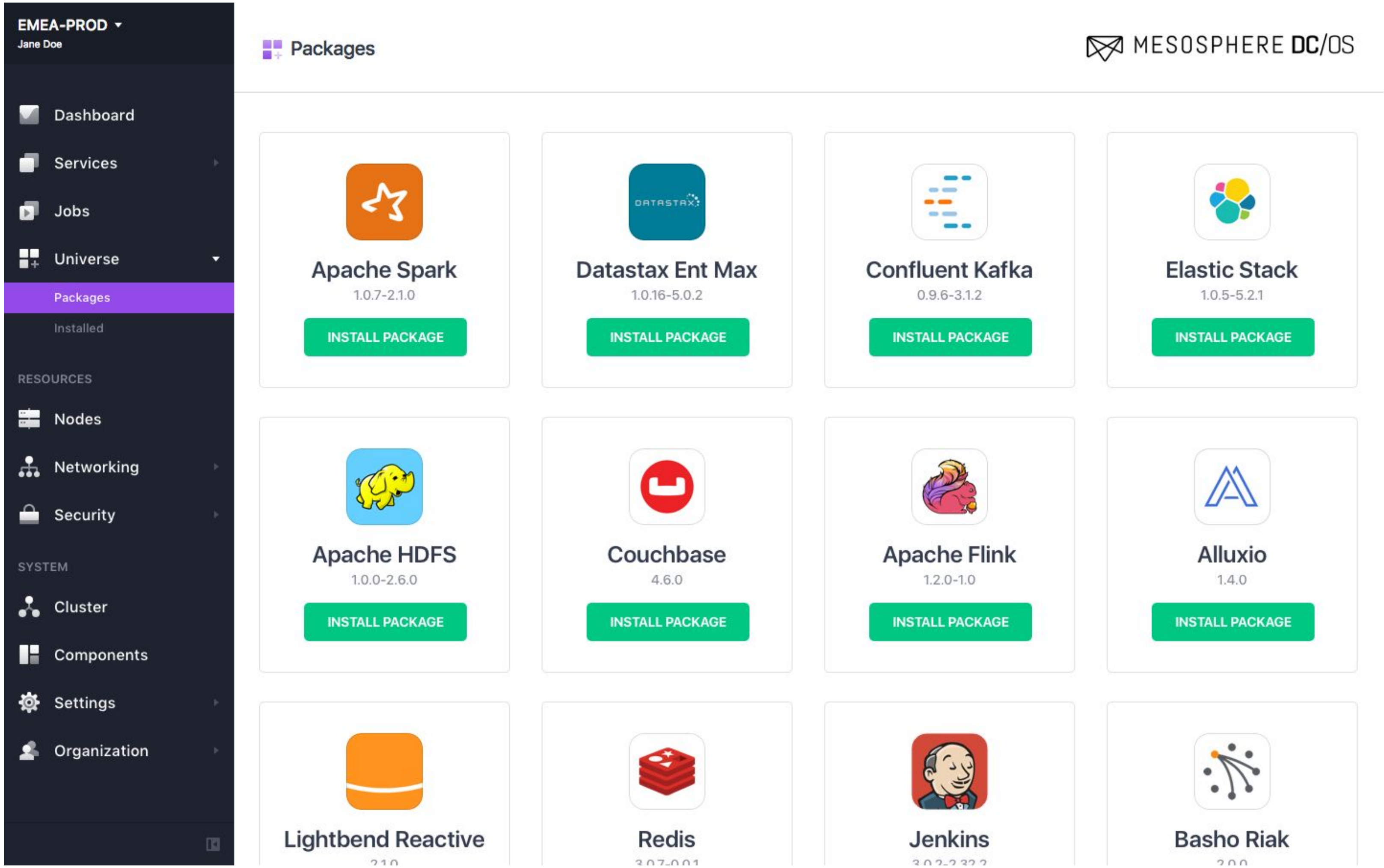
https://dcos.io/docs/latest/usage/webinterface/



NAME	STATUS		CPU	MEM	DISK
cassandra		Running (4 Instances)	2	8 GiB	0 B
kafka		Running (4 Instances)	4	4.8 GiB	0 B
marathon-lb		Running (1 Instance)	2	1 GiB	0 B
zeppelin		Running (1 Instance)	1	2 GiB	0 B

© 2017 Mesosphere, Inc. All Rights Reserved.

Universe



Packages

Apache Spark
1.0.7-2.1.0
INSTALL PACKAGE

Datastax Ent Max
1.0.16-5.0.2
INSTALL PACKAGE

Confluent Kafka
0.9.6-3.1.2
INSTALL PACKAGE

Elastic Stack
1.0.5-5.2.1
INSTALL PACKAGE

Apache HDFS
1.0.0-2.6.0
INSTALL PACKAGE

Couchbase
4.6.0
INSTALL PACKAGE

Apache Flink
1.2.0-1.0
INSTALL PACKAGE

Alluxio
1.4.0
INSTALL PACKAGE

Lightbend Reactive
2.1.0

Redis
3.0.7-0.0.1

Jenkins
3.0.2-2.32.2

Basho Riak
2.0.0

© 2017 Mesosphere, Inc. All Rights Reserved.

Interact with DC/OS (2/2)

CLI tool

https://dcos.io/docs/latest/usage/cli/

API

https://dcos.io/docs/latest/api/

© 2017 Mesosphere, Inc. All Rights Reserved.

Flink on Apache Mesos and DC/OS

According to the December 2016 data Artisans-organized Apache Flink user survey **just under 30% of respondents were running Flink on Apache Mesos**

https://dcos.io/blog/2017/apache-flink-on-dc-os-and-apache-mesos/

You may already be using Apache Mesos!

Version 1.2 of Flink includes support for Apache Mesos and DC/OS, *"it is now possible to run an highly available Flink cluster on Mesos"*
https://flink.apache.org/news/2017/02/06/release-1.2.0.html#run-flink-with-apache-mesos & https://ci.apache.org/projects/flink/flink-docs-release-1.2/setup/mesos.html

© 2017 Mesosphere, Inc. All Rights Reserved.

DEMOS

Demo data pipeline + Installing Flink on DC/OS

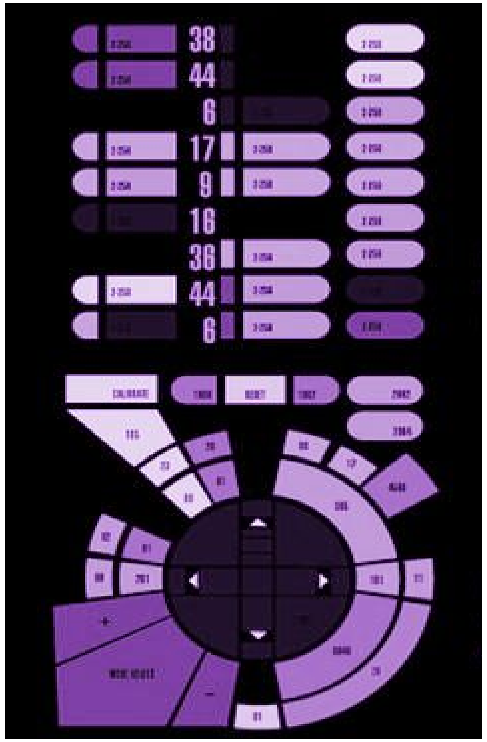
© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9 - Data Services Ecosystem

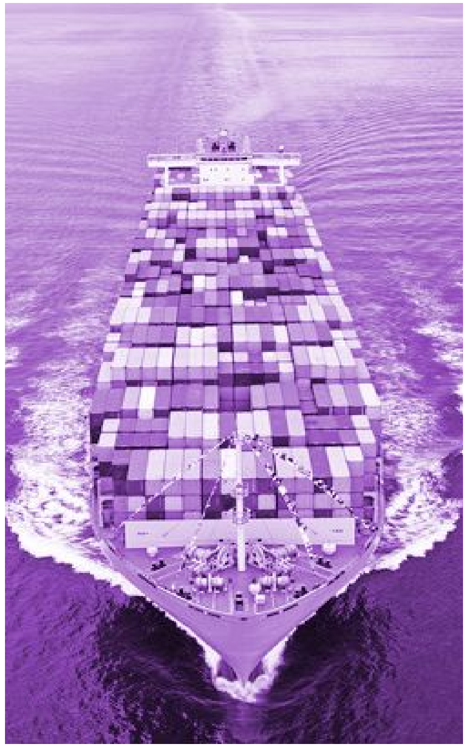
DATA SERVICES ECOSYSTEM

- *Alluxio*
- *Couchbase*
- *Datastax DSE*
- *Elastic (ELK)*
- *Redis*
- *Apache Flink*

OPERATIONS



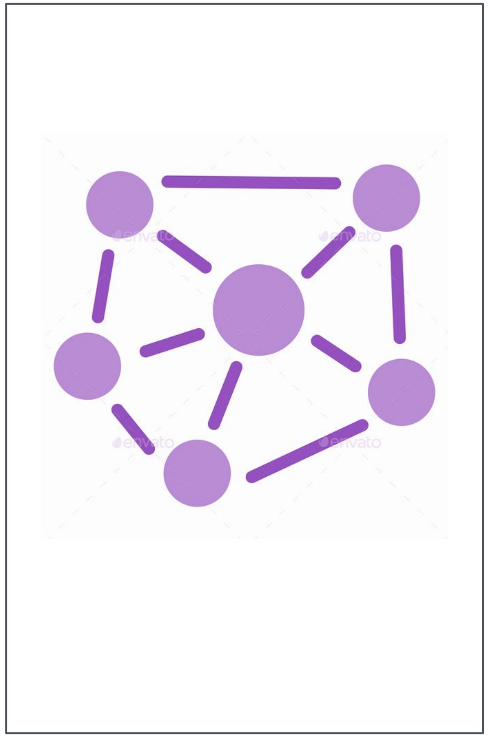
WORKLOADS



© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9 - Operations

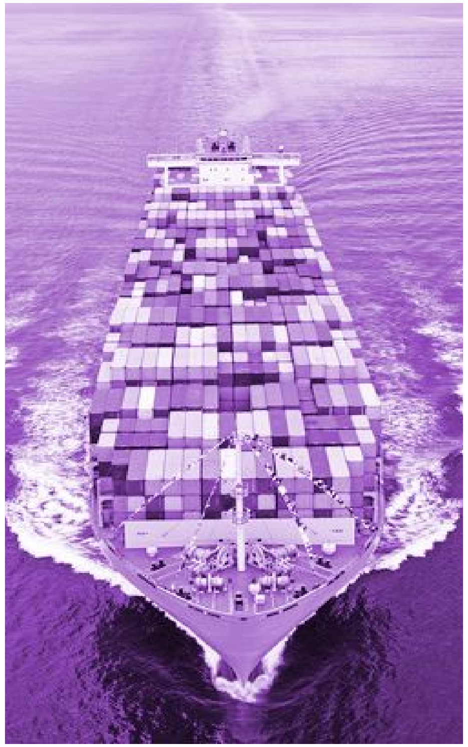
DATA SERVICES ECOSYSTEM



OPERATIONS

- Remote Container Shell
- Unified Metrics
- Unified Logging
- Deployment Failure Debugging
- Upgrades & Configuration updates

WORKLOADS



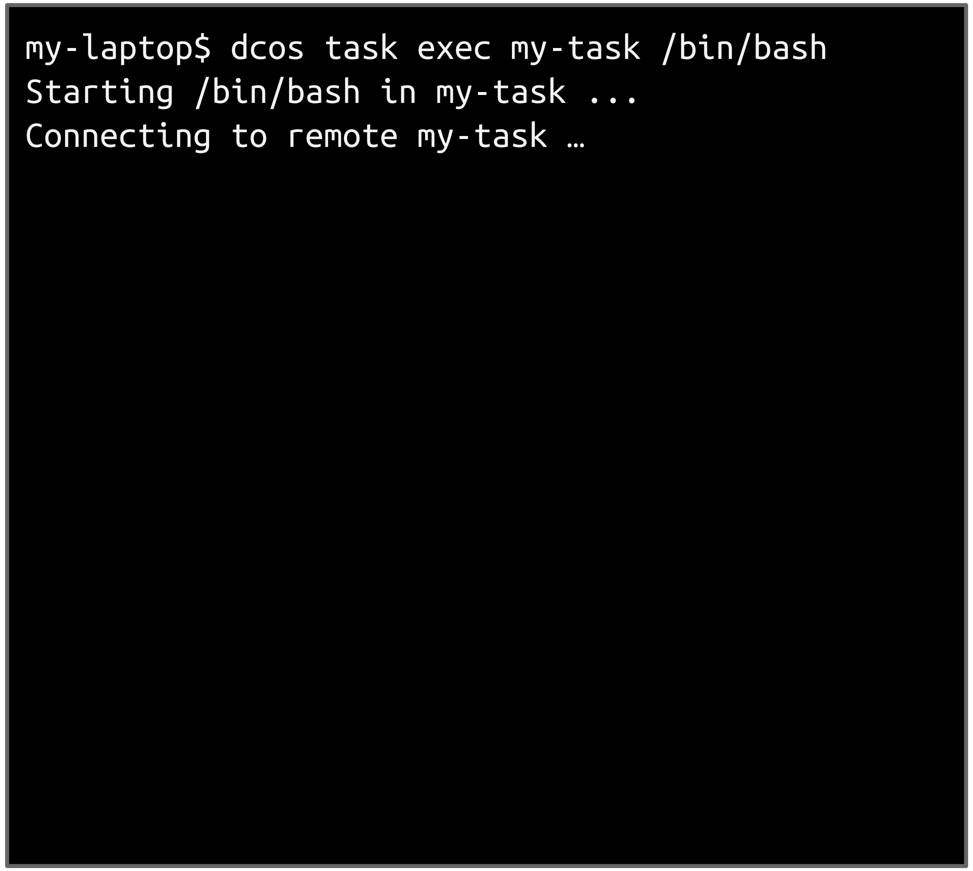
© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9

DC/OS: OPERATIONS

REMOTE CONTAINER SHELL

- Open encrypted, interactive, remote session to your containers
- Remotely execute commands for real time app troubleshooting
- Provide developers access to their own applications, not the entire host or cluster



```
my-laptop$ dcos task exec my-task /bin/bash
Starting /bin/bash in my-task ...
Connecting to remote my-task …
```

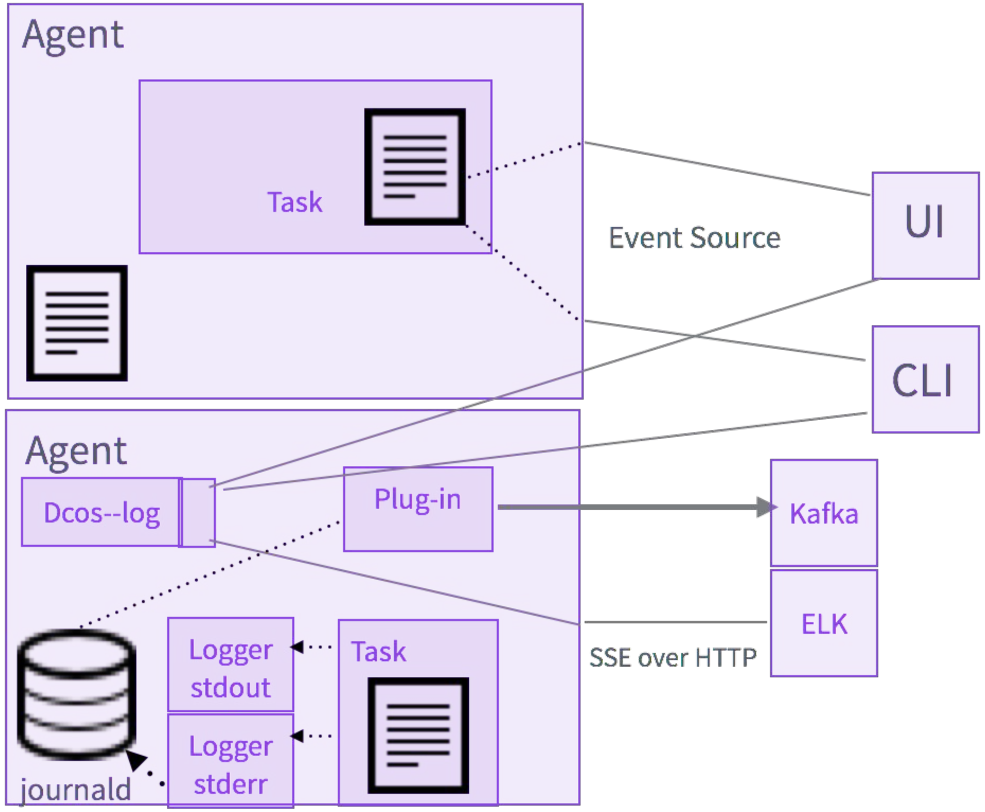
© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9

DC/OS: OPERATIONS

UNIFIED LOGGING

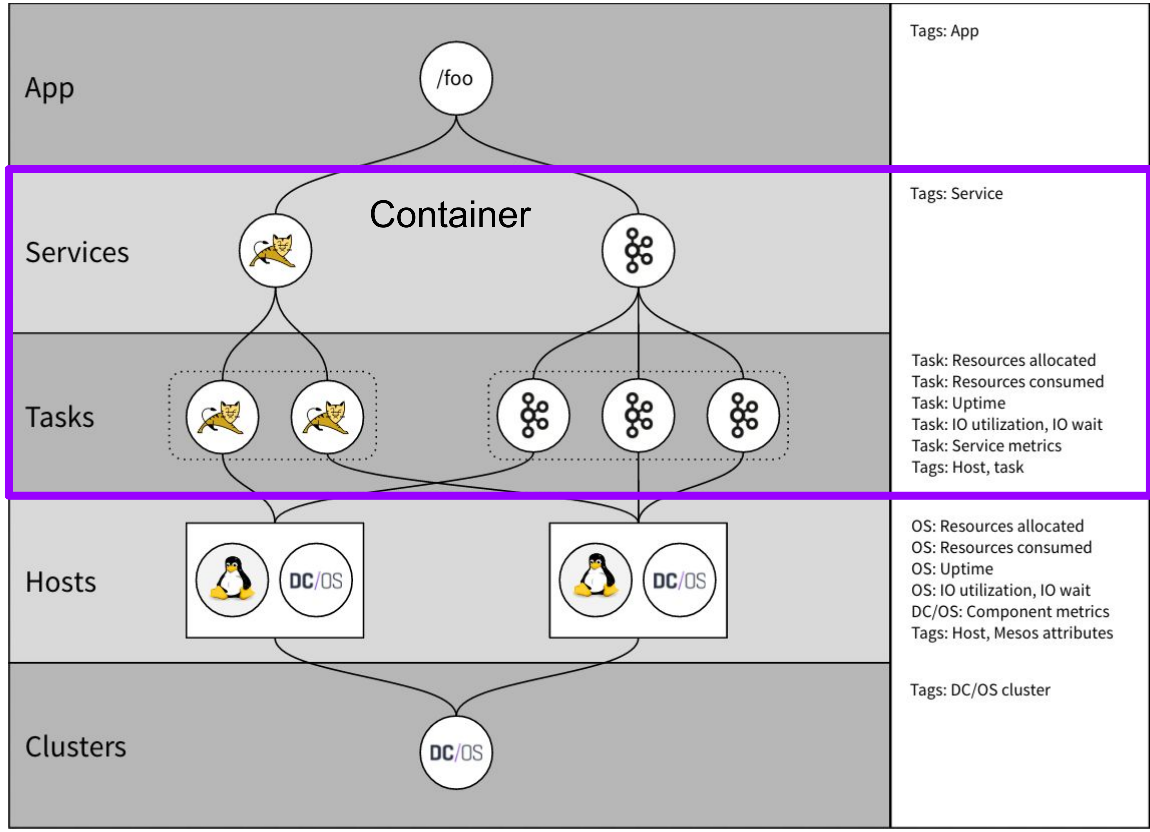
- Access application, DC/OS and OS logs
- Easily troubleshoot applications with critical metadata such as container id and app id
- Integrate easily with existing logging systems



© 2017 Mesosphere, Inc. All Rights Reserved.

UNIFIED METRICS

- Single API for system, container and application metrics
- Metadata such as host id and container id are automatically added to assist in debugging
- Integrate easily with existing metrics systems



Tags: App

Tags: Service

Task: Resources allocated
Task: Resources consumed
Task: Uptime
Task: IO utilization, IO wait
Task: Service metrics
Tags: Host, task

OS: Resources allocated
OS: Resources consumed
OS: Uptime
OS: IO utilization, IO wait
DC/OS: Component metrics
Tags: Host, Mesos attributes

Tags: DC/OS cluster

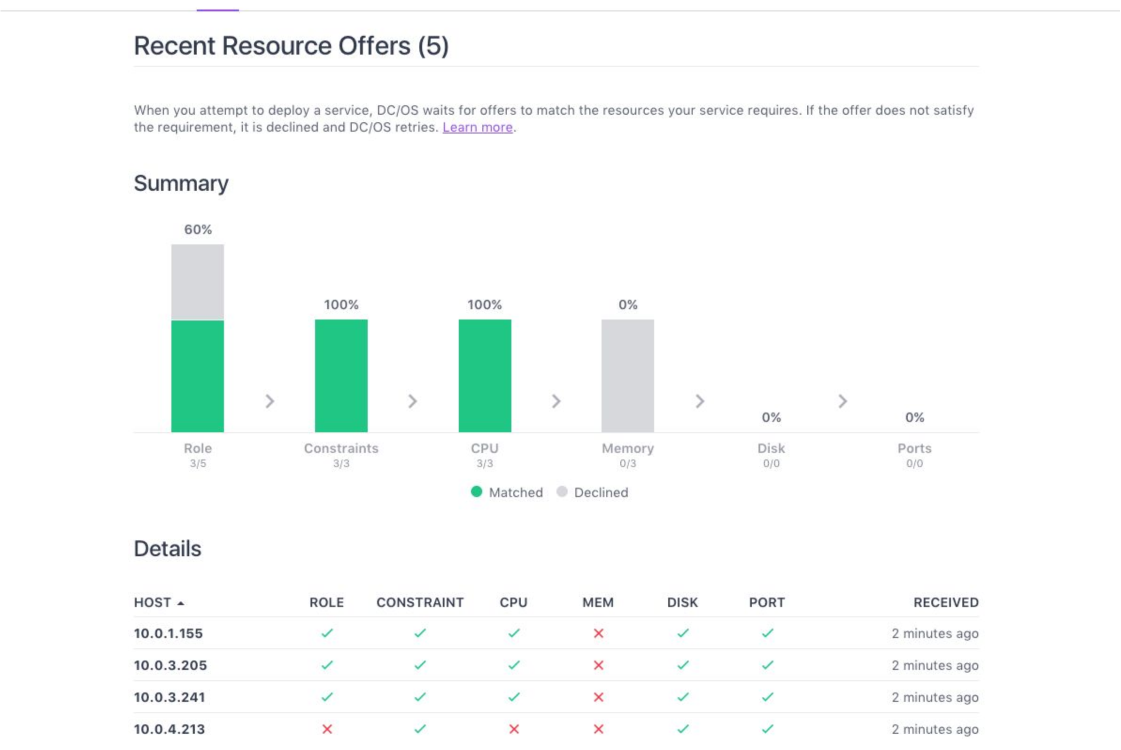
© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9

DC/OS: OPERATIONS

DEPLOYMENT FAILURE DEBUGGING

- Understand why your application is not deploying
- Understand which nodes in the cluster can accommodate the role, constraints, cpu, mem, disk and port requirements for your app



Recent Resource Offers (5)

When you attempt to deploy a service, DC/OS waits for offers to match the resources your service requires. If the offer does not satisfy the requirement, it is declined and DC/OS retries. Learn more.

Summary

Details

HOST ▲	ROLE	CONSTRAINT	CPU	MEM	DISK	PORT	RECEIVED
10.0.1.155	✓	✓	✓	✗	✓	✓	2 minutes ago
10.0.3.205	✓	✓	✓	✗	✓	✓	2 minutes ago
10.0.3.241	✓	✓	✓	✗	✓	✓	2 minutes ago
10.0.4.213	✗	✓	✗	✗	✓	✓	2 minutes ago

© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS: OPERATIONS

UPGRADES AND CONFIG UPDATES

- Generate new config for cluster nodes

```
    $ dcos_generate_config.sh --generate-node-upgrade-script
<installed_cluster_version>
```

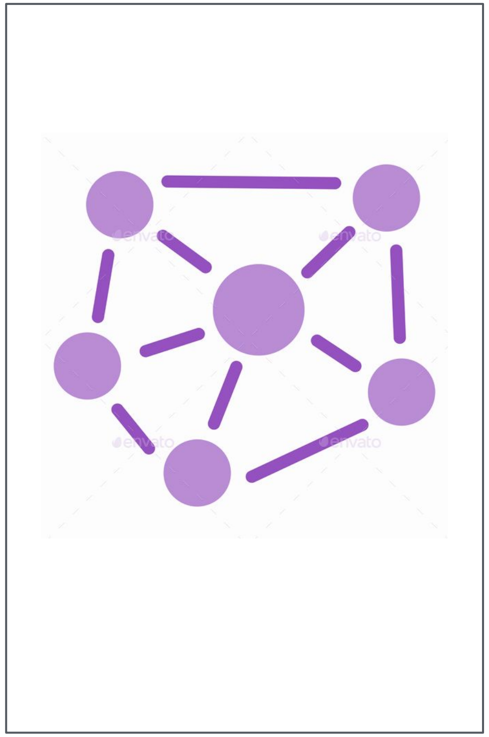
- Single command upgrade script for individual nodes

```
    $ curl -O <Node upgrade script URL>
    $ sudo bash ./dcos_node_upgrade.sh
```

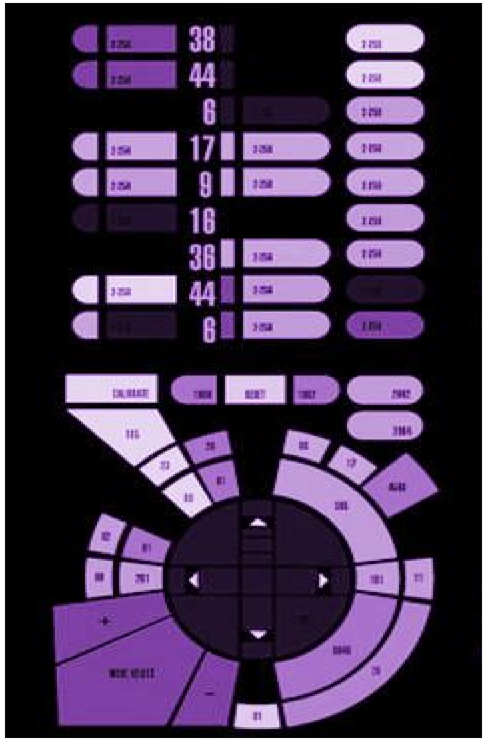
© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9 - Workloads

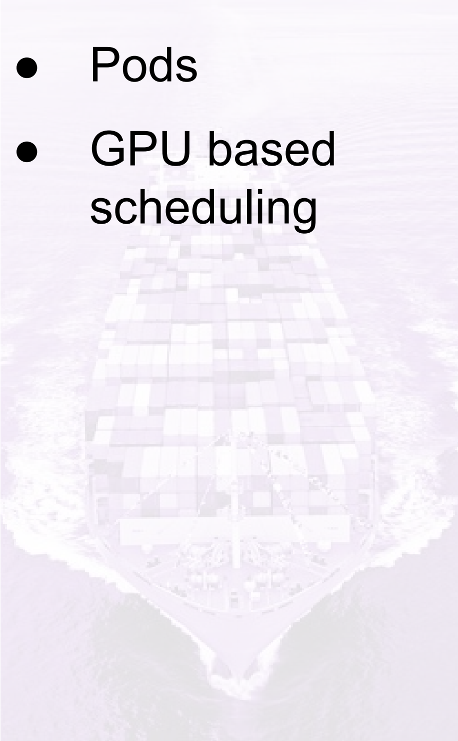
DATA SERVICES ECOSYSTEM



OPERATIONS



WORKLOADS



- Pods
- GPU based scheduling

© 2017 Mesosphere, Inc. All Rights Reserved.

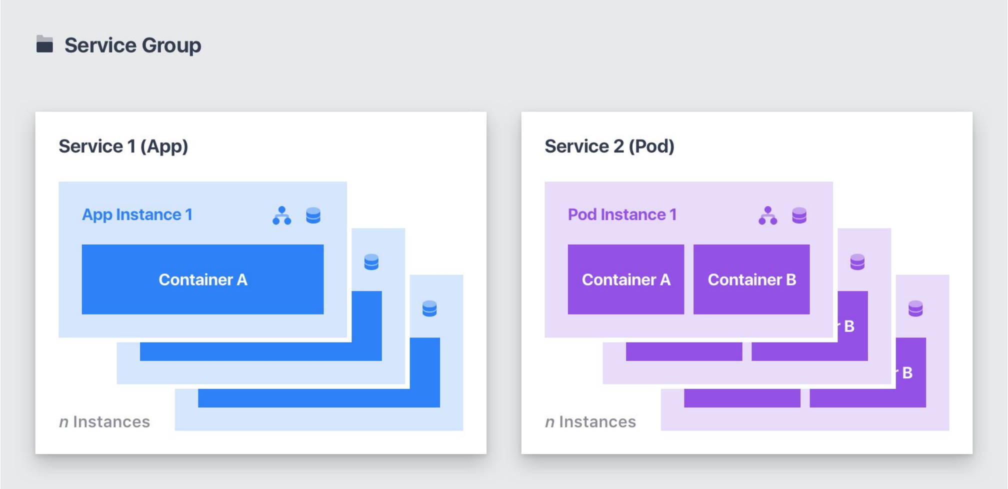
DC/OS: WORKLOADS

PODS

- Schedule, deploy and scale multiple containers on the same host(s) while sharing IP address and storage volumes

- All containers in a pod instance run as if they are running on a single host in pre-container world

- Useful for migrating legacy applications or building advanced micro services (side car containers)



© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9

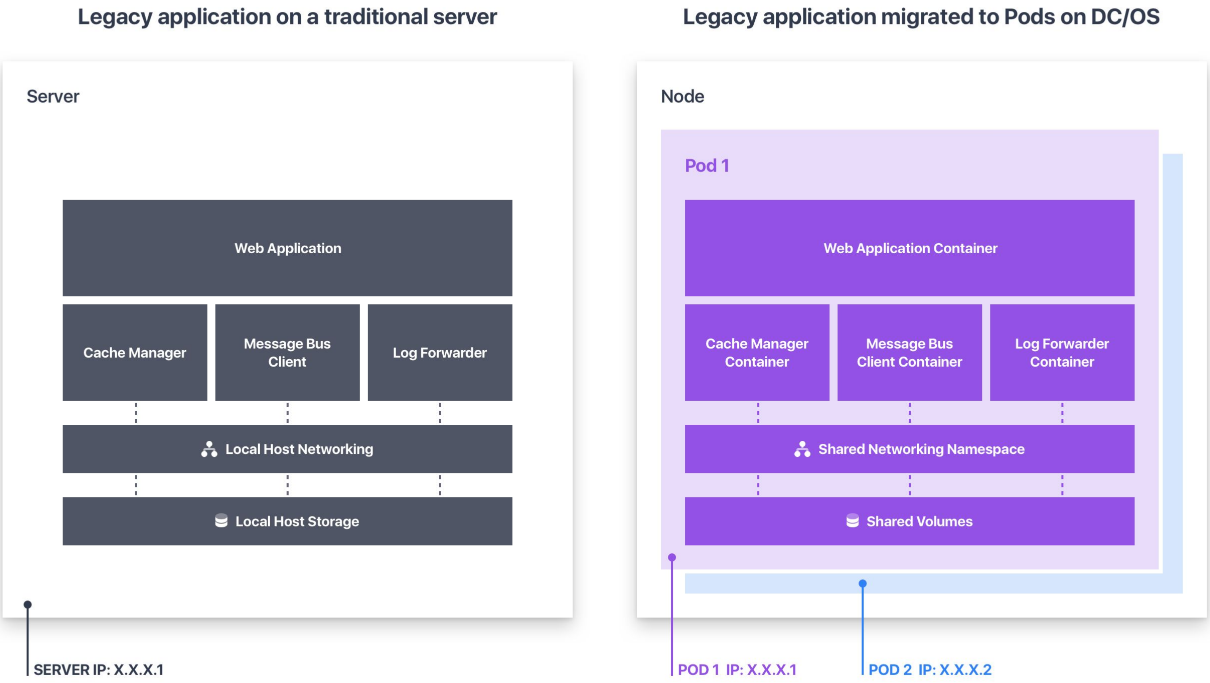
DC/OS: WORKLOADS

PODS: MIGRATING LEGACY APPS TO CONTAINERS

- Traditional monolithic apps on VMs usually have support services such as log shipper, message queuing clients

- Many support services assume col-location on same host, and local-host access to networking and storage

- Pods simplify moving legacy monolithic apps to containers, reducing risk and accelerating migrations



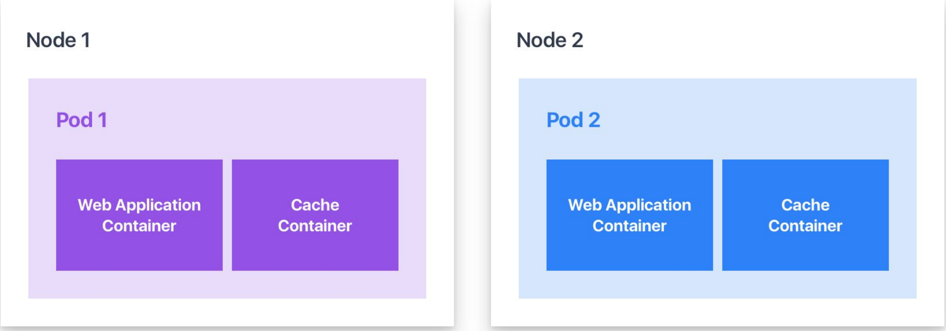
© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9

DC/OS: WORKLOADS

PODS: SUPPORT SERVICES (SIDE-CAR CONTAINERS)

- Advanced Micro Services patterns require colocating containers together
- Support services include for example:
  - Logging or monitoring agents,
  - Backup tooling & Proxies
  - Data change watchers & Event publishers
- Pods simplify the building and maintenance of complex such microservices

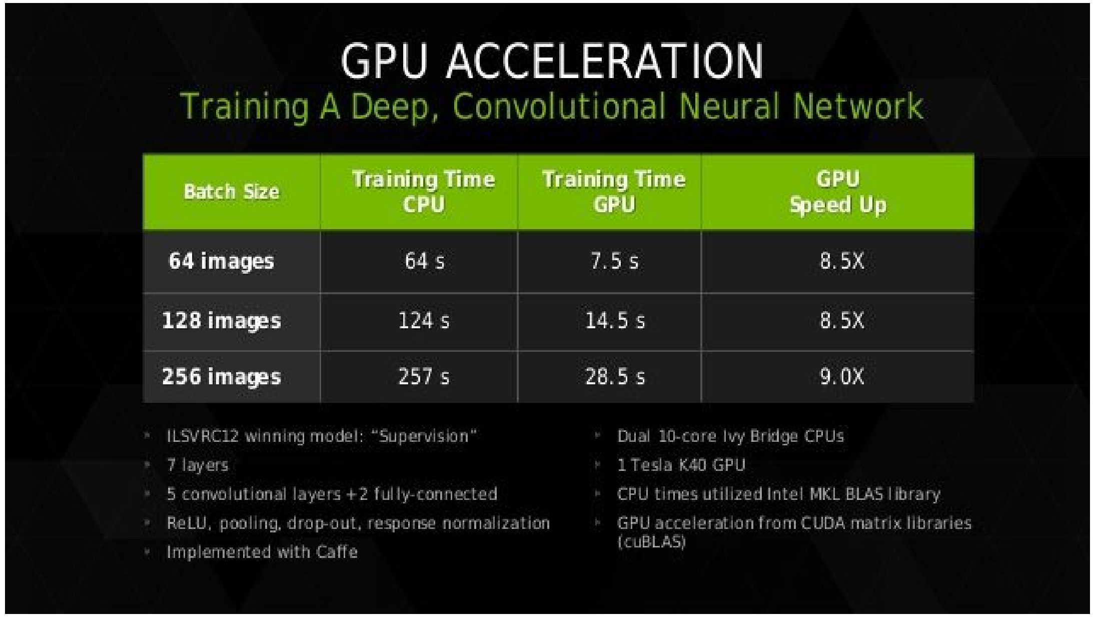


© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS: WORKLOADS

GPU: WHY GPU?

- GPUs are needed for many machine learning and deep learning applications
- GPUs are essential for real-time or near real time machine learning models
- GPUs deliver from 10X to 100X performance for some applications, resulting lower $$$/IOPS and more productivity to data science teams
- GPU applications include real time fraud detection, genome sequencing, cohort analysis and many others



GPU ACCELERATION

Training A Deep, Convolutional Neural Network

Batch Size	Training Time CPU	Training Time GPU	GPU Speed Up
64 images	64 s	7.5 s	8.5X
128 images	124 s	14.5 s	8.5X
256 images	257 s	28.5 s	9.0X

- ILSVRC12 winning model: "Supervision"
- 7 layers
- 5 convolutional layers +2 fully-connected
- ReLU, pooling, drop-out, response normalization
- Implemented with Caffe

- Dual 10-core Ivy Bridge CPUs
- 1 Tesla K40 GPU
- CPU times utilized Intel MKL BLAS library
- GPU acceleration from CUDA matrix libraries (cuBLAS)

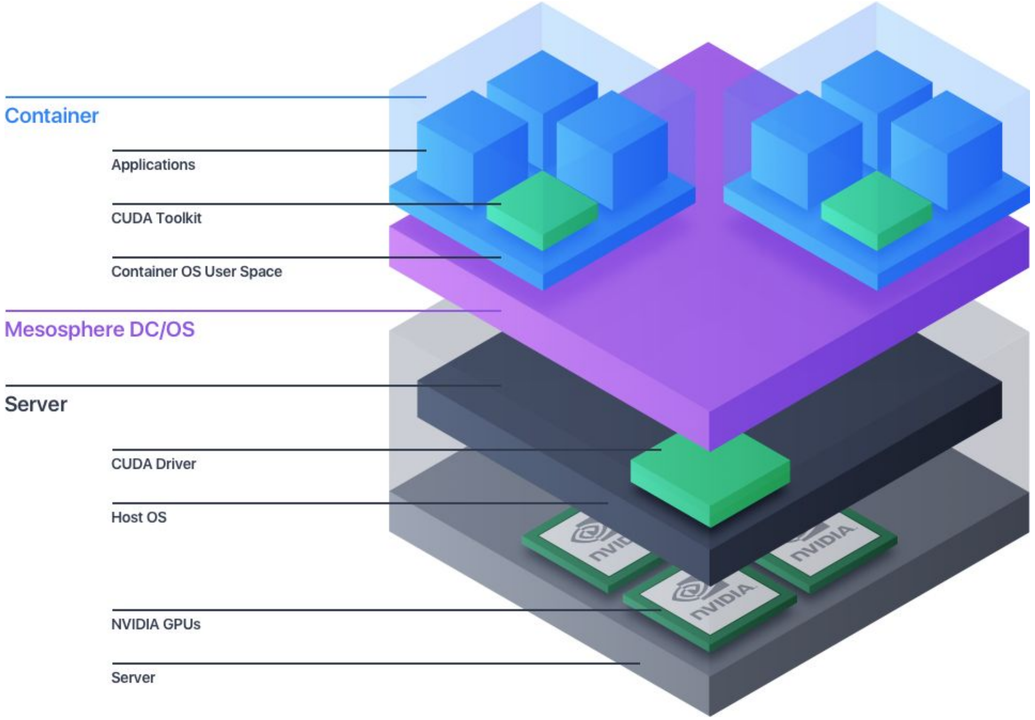
© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS 1.9

DC/OS: WORKLOADS

GPU BASED SCHEDULING

- Test Locally with Nvidia-Docker, deploy to production with DC/OS
- Isolate GPU instances and schedule workloads just like CPU and memory, guaranteeing performance
- Efficiently Share GPU resources across data science team
- Simplify migrating machine learning models across from dev to production, and across clouds



© 2017 Mesosphere, Inc. All Rights Reserved.

DC/OS

OTHER IMPROVEMENTS

- Mesos 1.2
- Marathon 1.4
- Docker 1.12 and 1.13 (17.03-ce) support
- Centos 7.3 and CoreOS 1235.12.0 support
- Performance improvements across all networking features.
- CNI support for 3rd party CNI plugins.
- 100s of additional bugfixes and tests

© 2017 Mesosphere, Inc. All Rights Reserved.

MESOSPHERE