Apache Hadoop YARN: Present and Future 2017

Vinod Kumar Vavilapalli
Apache Hadoop PMC, Co-founder of YARN project
Hortonworks Inc

HORTONWORKS®
POWERING THE FUTURE OF DATA™

About.html

- Apache Hadoop PMC, ASF Member
- Yahoo! -> Hortonworks
- 10 years of (only) Hadoop
  - Finally the job-adverts asking for “10 years of Hadoop experience” have validity
- ’Rewritten’ the Hadoop processing side – Became Apache Hadoop YARN
- Running compute platform teams at Hortonworks: YARN, MapReduce, Slider, container cloud on YARN

Agenda

- Introduction
- Past
- Present & Future

Hadoop Compute Platform – Today and Tomorrow

- It's all about data!
- Layers that enable applications and higher order frameworks that interact with data
- Multi-colored YARN
  - Apps
  - Long running services
- Admins and admin tools (Ambari) for cluster management and monitoring

https://www.flickr.com/photos/happyskrappy/15699919424

Why?

- Different asks from different actors
- On isolation, capacity allocations, scheduling

Faster!

More!

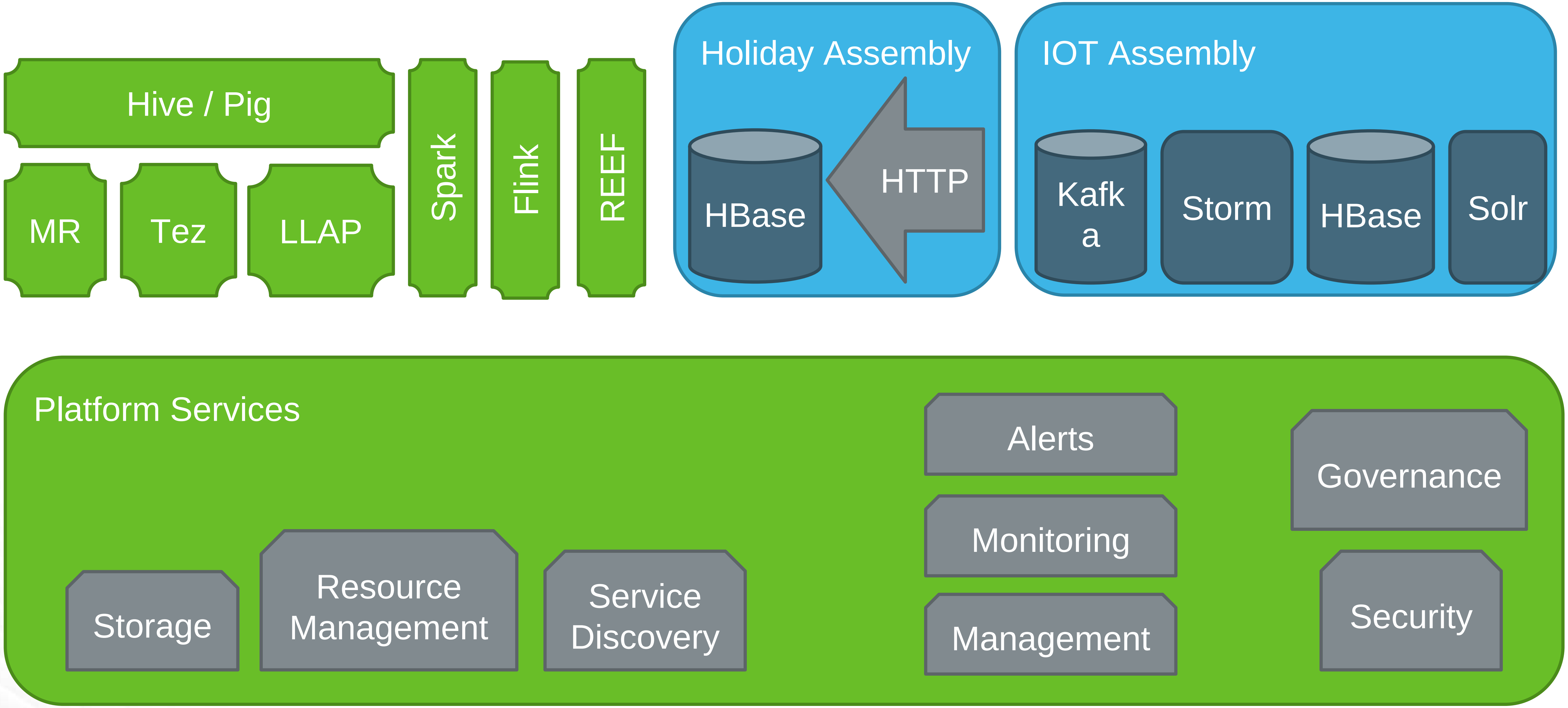
SLA!

Everything!

Right now!

Best for my cluster
Throughput
Utilization
Elasticity
Service uptime
Security
ROI

Hadoop Compute Platform – Today and Tomorrow



Past: A quick history

A brief Timeline: Pre GA

- **Sub-project of Apache Hadoop**
- **Releases tied to Hadoop releases**
- **Alphas and betas**
  - In production at several large sites for MapReduce already by that time

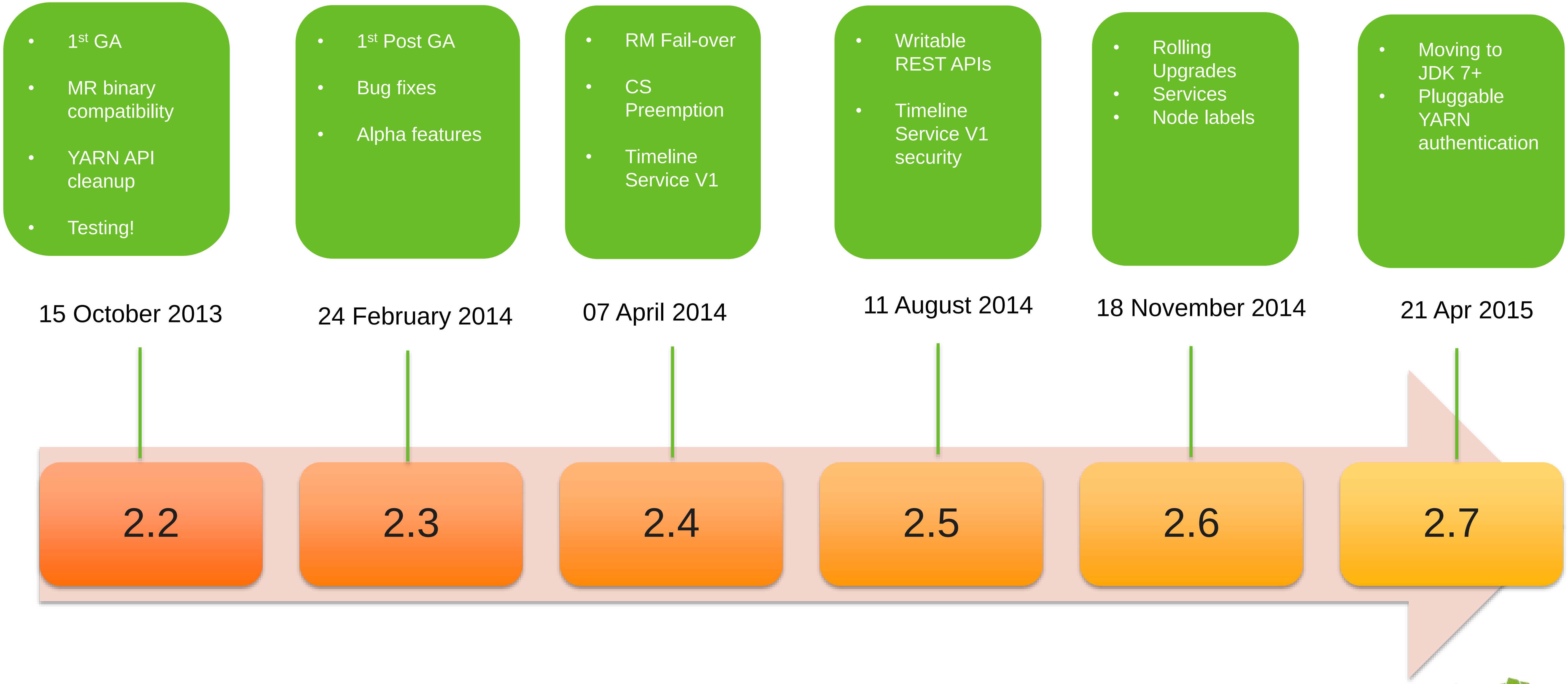
June-July 2010

August 2011

May 2012

August 2013

A brief Timeline: GA Releases 1/2



- 1st GA
- MR binary compatibility
- YARN API cleanup
- Testing!

15 October 2013

2.2

- 1st Post GA
- Bug fixes
- Alpha features

24 February 2014

2.3

- RM Fail-over
- CS Preemption
- Timeline Service V1

07 April 2014

2.4

- Writable REST APIs
- Timeline Service V1 security

11 August 2014

2.5

- Rolling Upgrades
- Services
- Node labels

18 November 2014

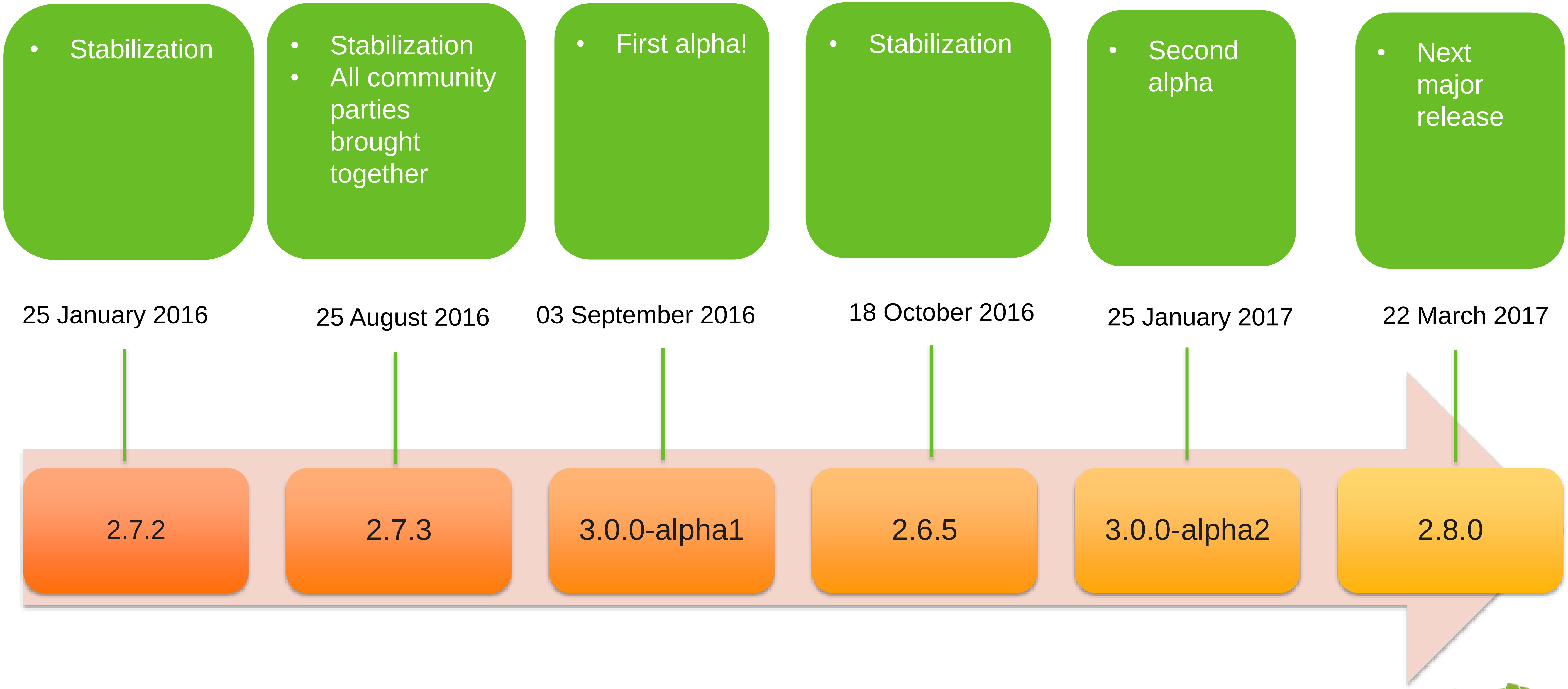
2.6

- Moving to JDK 7+
- Pluggable YARN authentication

21 Apr 2015

2.7

A brief Timeline: GA Releases 2/2



25 January 2016 – 2.7.2
- Stabilization

25 August 2016 – 2.7.3
- Stabilization
- All community parties brought together

03 September 2016 – 3.0.0-alpha1
- First alpha!

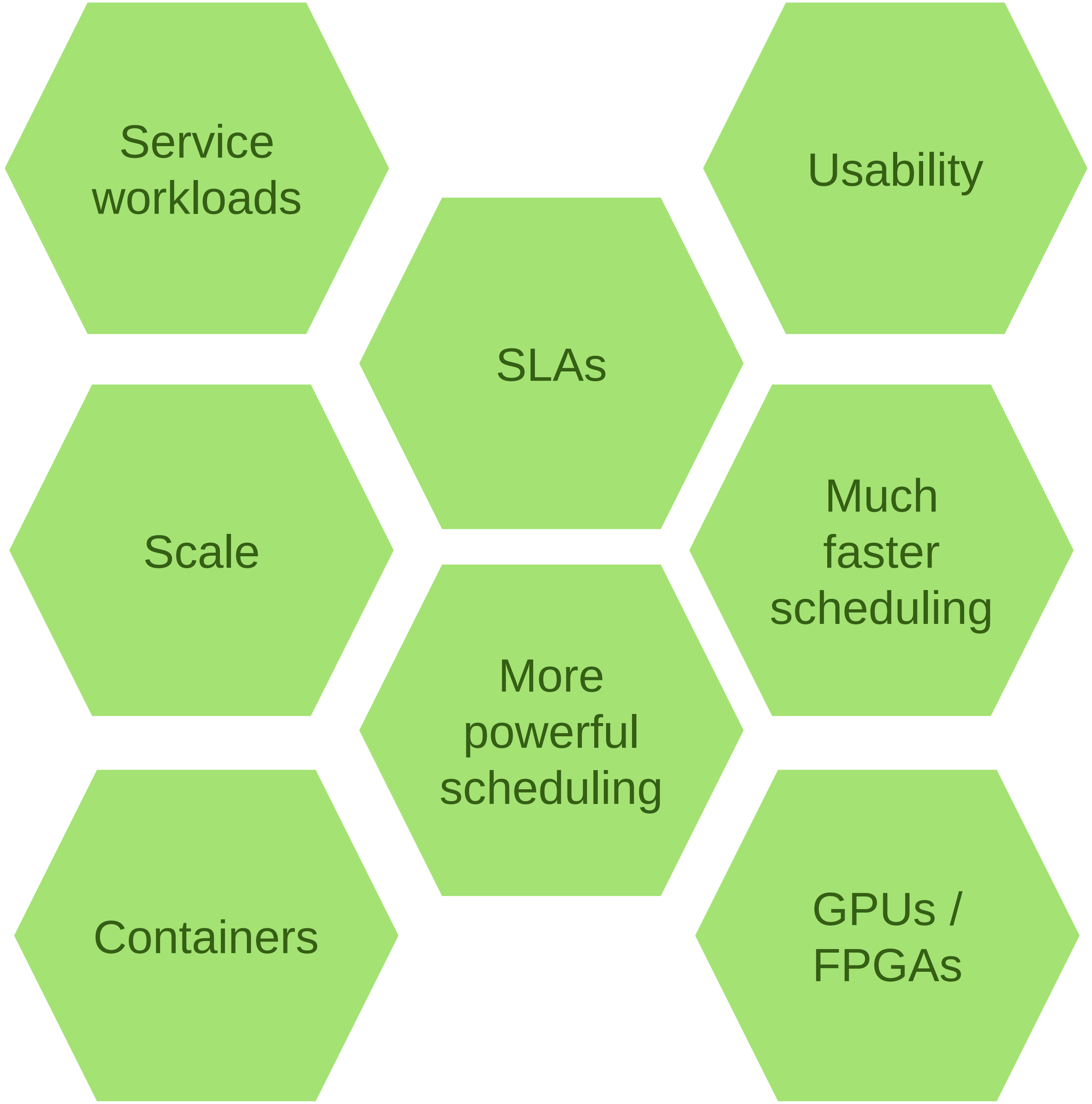
18 October 2016 – 2.6.5
- Stabilization

25 January 2017 – 3.0.0-alpha2
- Second alpha

22 March 2017 – 2.8.0
- Next major release

Hortonworks

Present & Future



Service workloads

Scale

Containers

SLAs

More powerful scheduling

Usability

Much faster scheduling

GPUs / FPGAs

Last few Hadoop releases

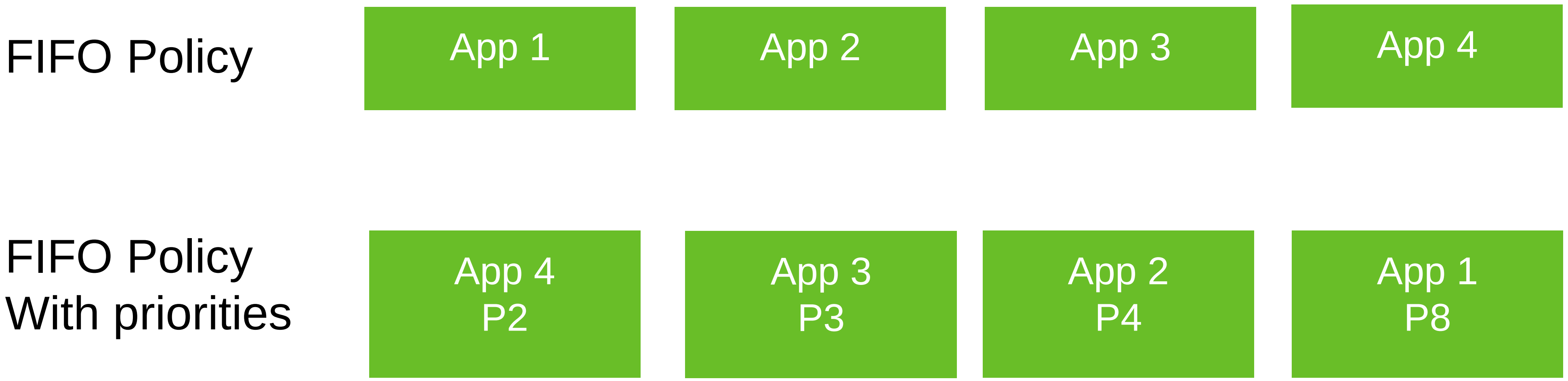
Apache Hadoop 2.8.0

Apache Hadoop 3.0.x

Apache Hadoop 2.8.0

Application priorities

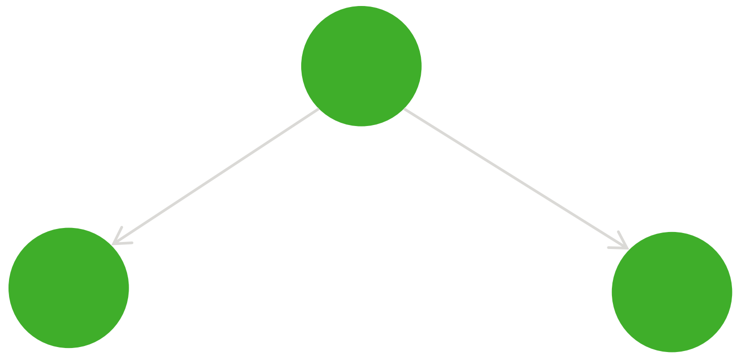
- YARN-1963
- Within a leaf-queue



FIFO Policy	App 1	App 2	App 3	App 4
FIFO Policy With priorities	App 4 P2	App 3 P3	App 2 P4	App 1 P8

Queue priorities

- Today
  - Give to the least satisfied queue first



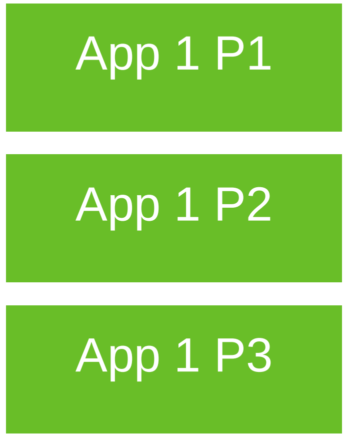
20% Configured Capacity
But 5% of used Capacity
Usage = 5/20 = 25%

80% Configured Capacity
But 8% of used Capaity
Usage 8/80 = 10%

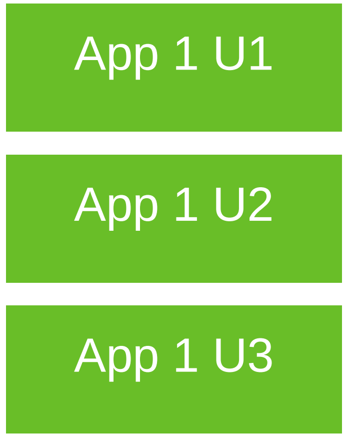
- With priorities
  - Give to the highest priority queue first

Preemption within a queue

- **Between apps of different priorities**

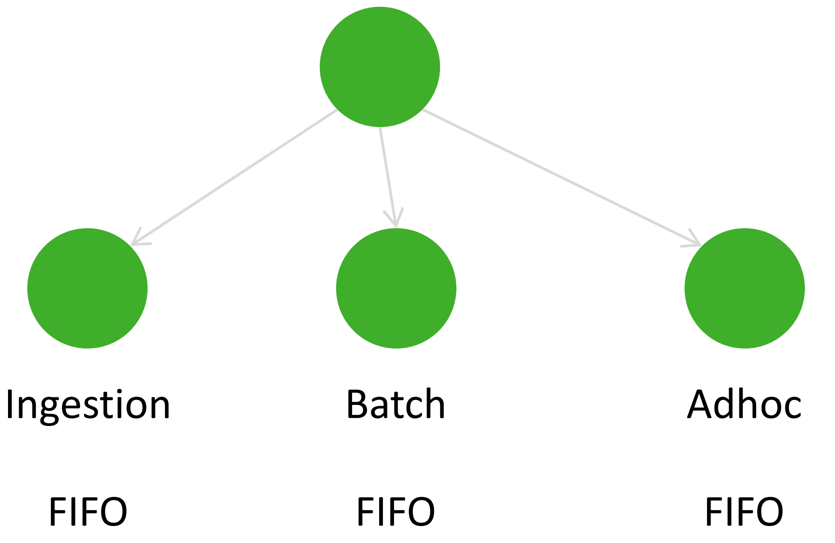


- **Between apps of different users**



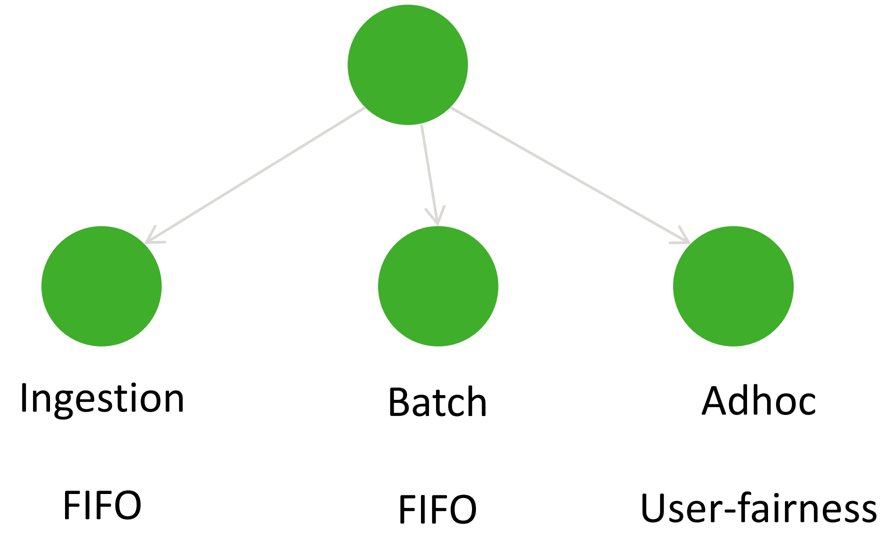
Per-queue Policy-driven scheduling

Previously



- Coarse policies
- One scheduling algorithm in the cluster
- Rigid
- Difficult to experiment

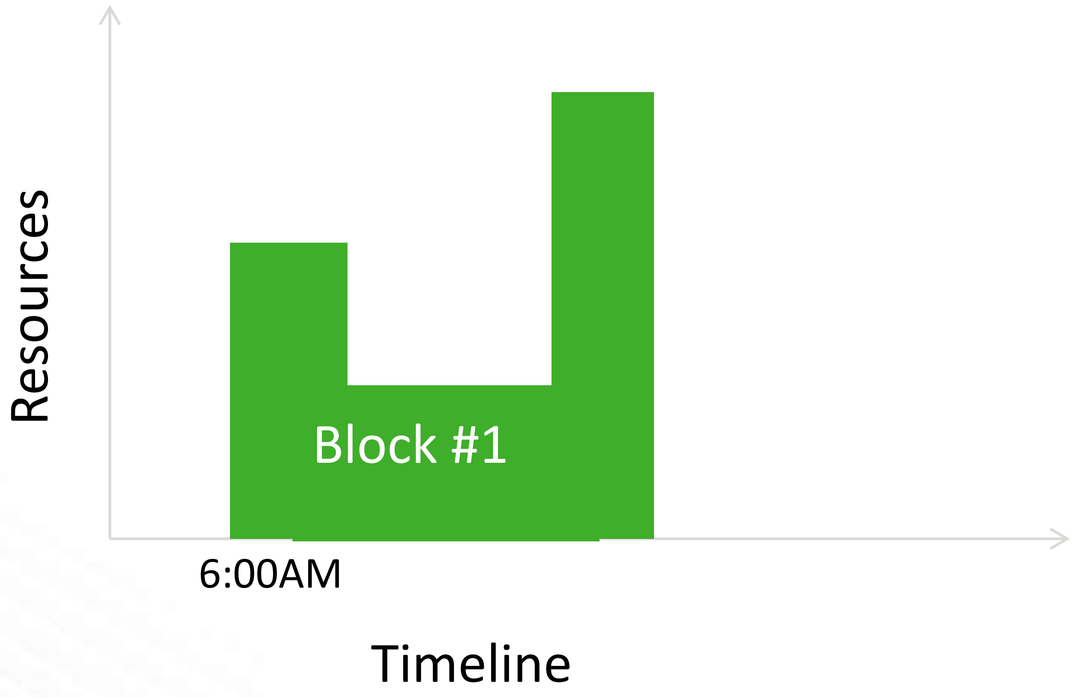
Now

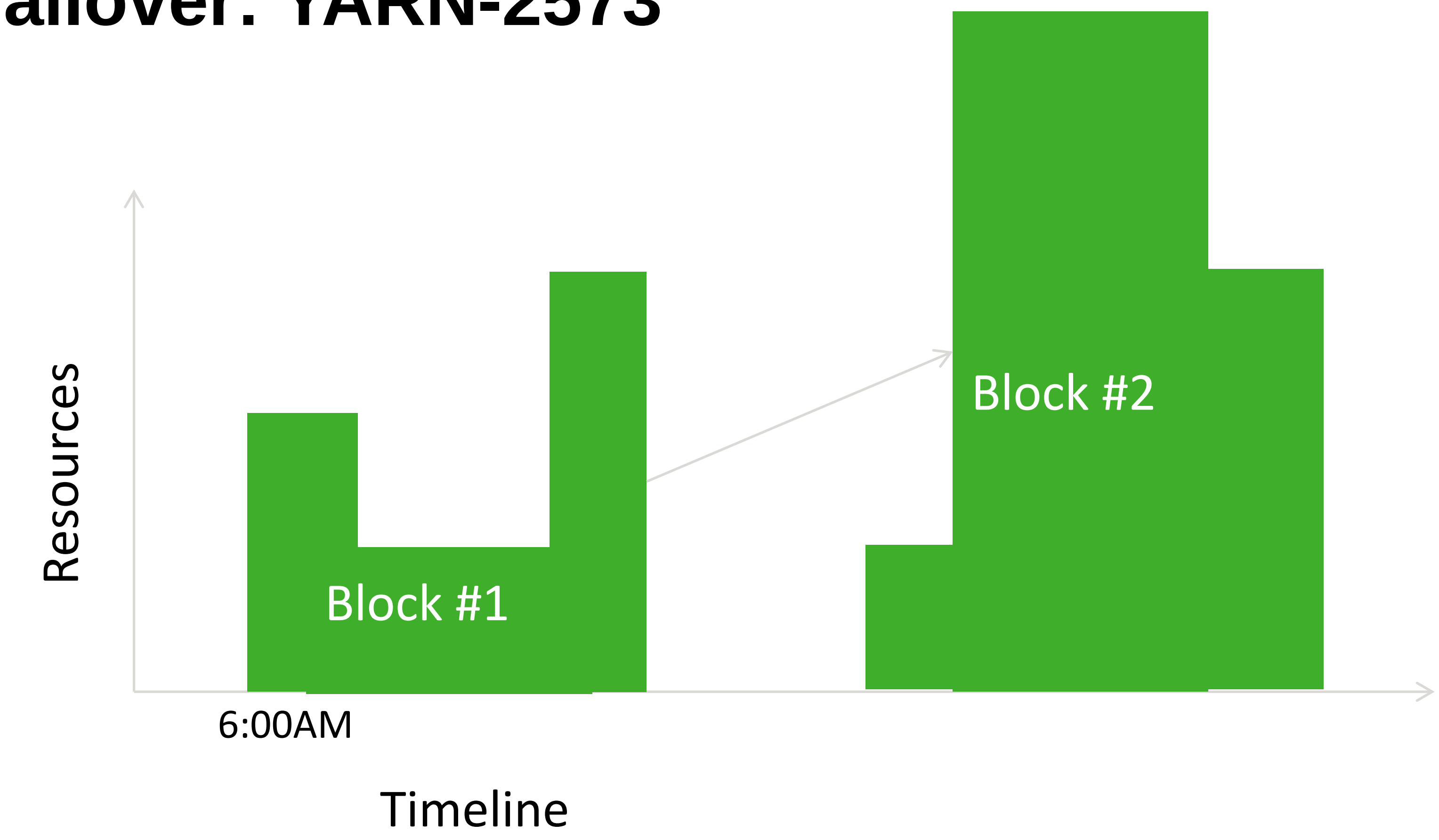


- Fine grained policies
- One scheduling algorithm per queue
- Flexible
- Very easy to experiment!

Reservations

- ***“Run my workload tomorrow at 6AM”***
- **Persistence of the plans with RM failover: YARN-2573**





-- Related Session --

Never late again! Job-Level deadline SLOs in YARN

Subru Krishnan
Microsoft

Carlo Curino
Microsoft

https://dataworkssummit.com/san-jose-2017/sessions/never-late-again-job-level-deadline-slos-in-yarn

Wednesday June 14th Room 230A

Apache Hadoop 3.x

-- Related Session --

Apache Hadoop 3.0

Junping Du
Hortonworks

Andrew Wang
Cloudera

https://dataworkssummit.com/san-jose-2017/sessions/apache-hadoop-3-0-community-update
Tuesday June 13th Room 210C

Scale!

- Only focusing on sizes of individual clusters
- Tons of sites with clusters made up of multiple thousands of nodes
  - Yahoo!, Twitter, LinkedIn, Microsoft
- Largest clusters the last couple of years
  - 6K-8K
- Roadmap: To 100K thousands and beyond
- Current progress: 40K nodes!

-- Related Session --

Lessons learned from scaling YARN to 40k machines in a multi tenancy environment

Hitesh Sharma
Microsoft

Roni Burd
Microsoft

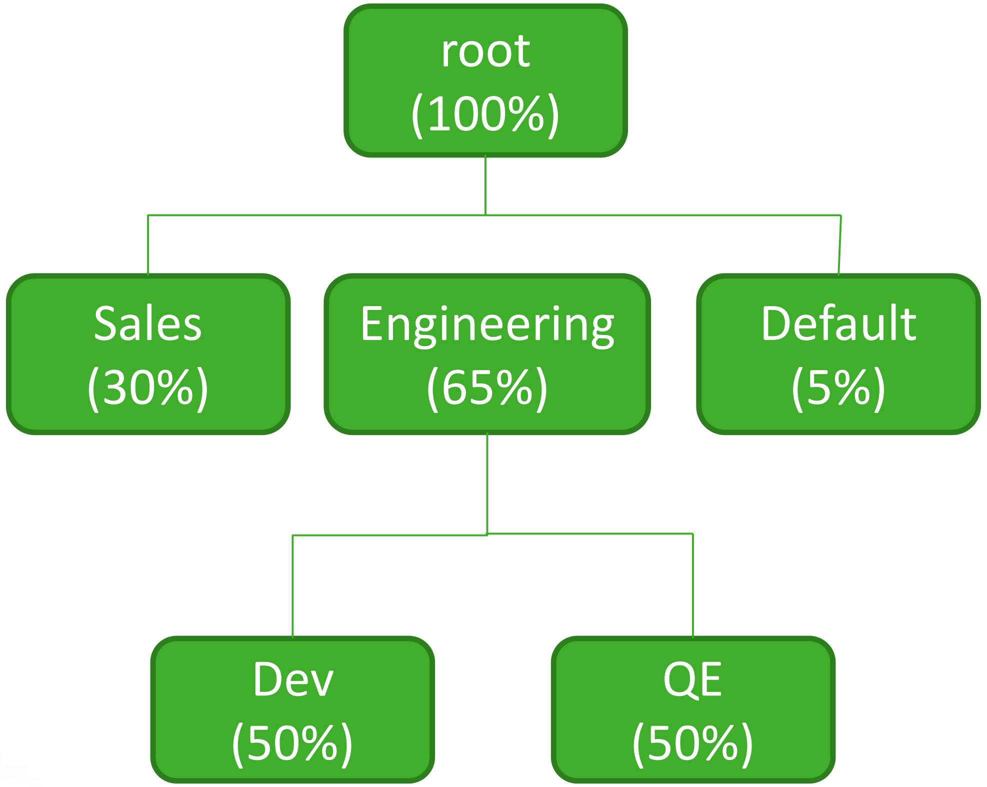
https://dataworkssummit.com/san-jose-2017/sessions/lessons-learned-from-scaling-yarn-to-40k-machines-in-a-multi-tenancy-environment

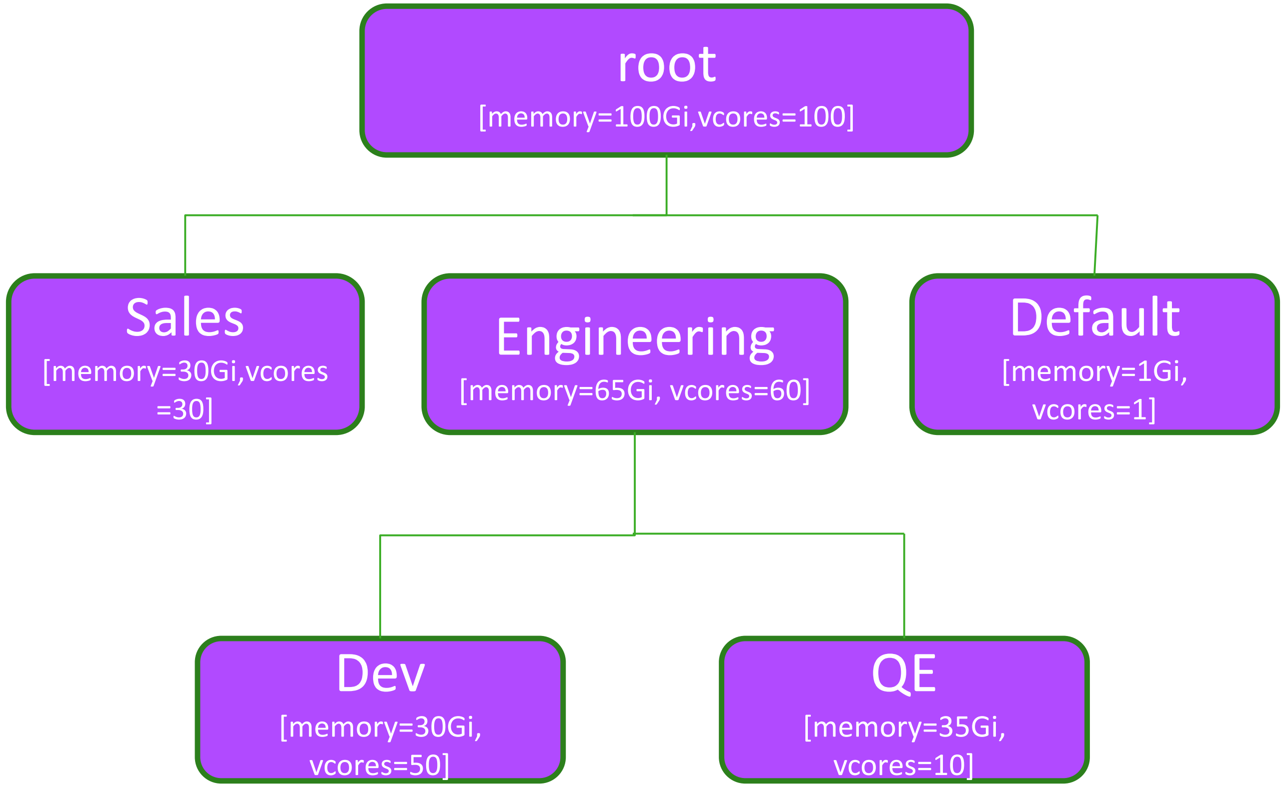
Wednesday June 14th Room 210A

Global & Fast Scheduling

- Effort led by Wangda Tan
- Problems
  - Current design of one-node-at-a-time allocation cycle can lead to suboptimal decisions.
  - Several coarse grained locks
- On trunk
  - Look at several nodes at a time
  - Fine grained locks
  - Multiple allocator threads
  - YARN scheduler can allocate 3k+ containers per second ≈ 10 mil allocations / hour!
  - 10X throughput gains with enhancement added recently
  - Opportunities for better placement

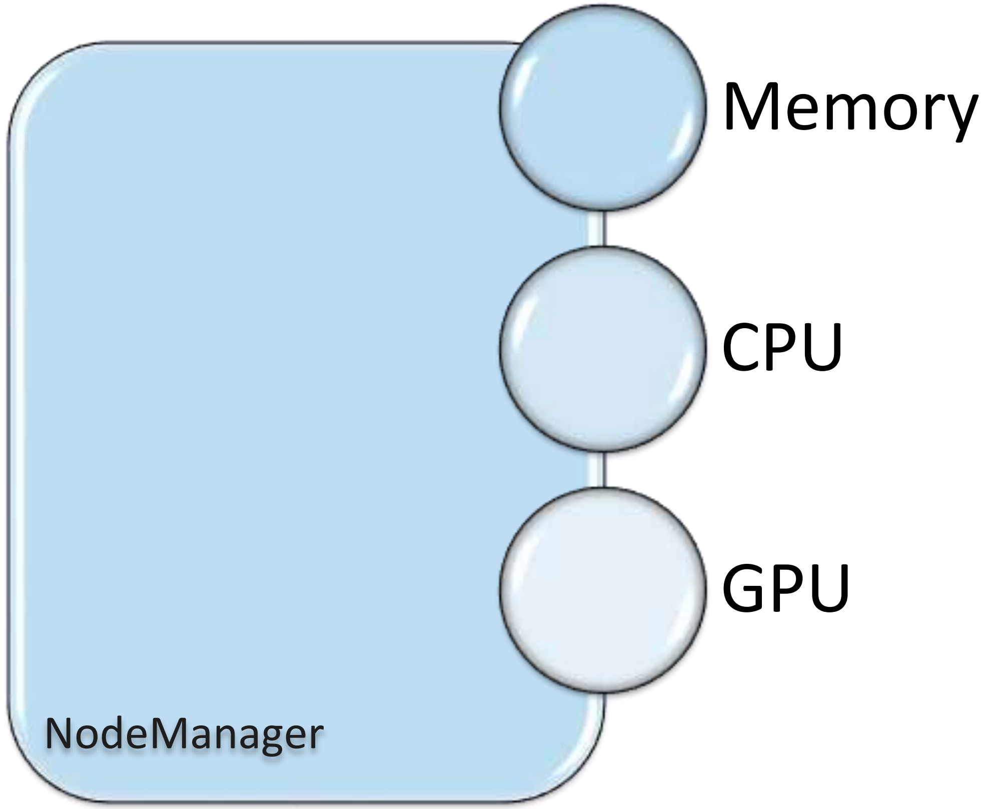
Capacities in numbers vs percentages





Resource vectors

- Till now
  - Hard coded resources
  - Memory and CPU
- Now
  - A generalized vector
  - Admins can create custom Resource Types!



GPUs on a YARN cluster!

- GPU can speed up compute-intensive applications 10x - 300x times
- Different levels of support
  - Take me to a machine where GPUs are available with Partitions / Node Labels
  - Take me to a machine where GPUs are available
    - give me a **full device** **only to me** for the **lifetime of my container**
    - give me **multiple full devices** **only to me** for the **lifetime of my container**
    - give me **full device(s)** **only to me** for a **portion of the lifetime of my container**
    - give me a **slice of device(s)** **to me** for a **full / portion of the lifetime of my container**
- More dimensions:
  - CPUs and memory and GPUs and on-GPU memory
  - Topology of multiple GPUs

-- Related Session --

Hadoop ecosystem boosts Tensorflow and machine learning technologies

Wangda Tan
Hortonworks

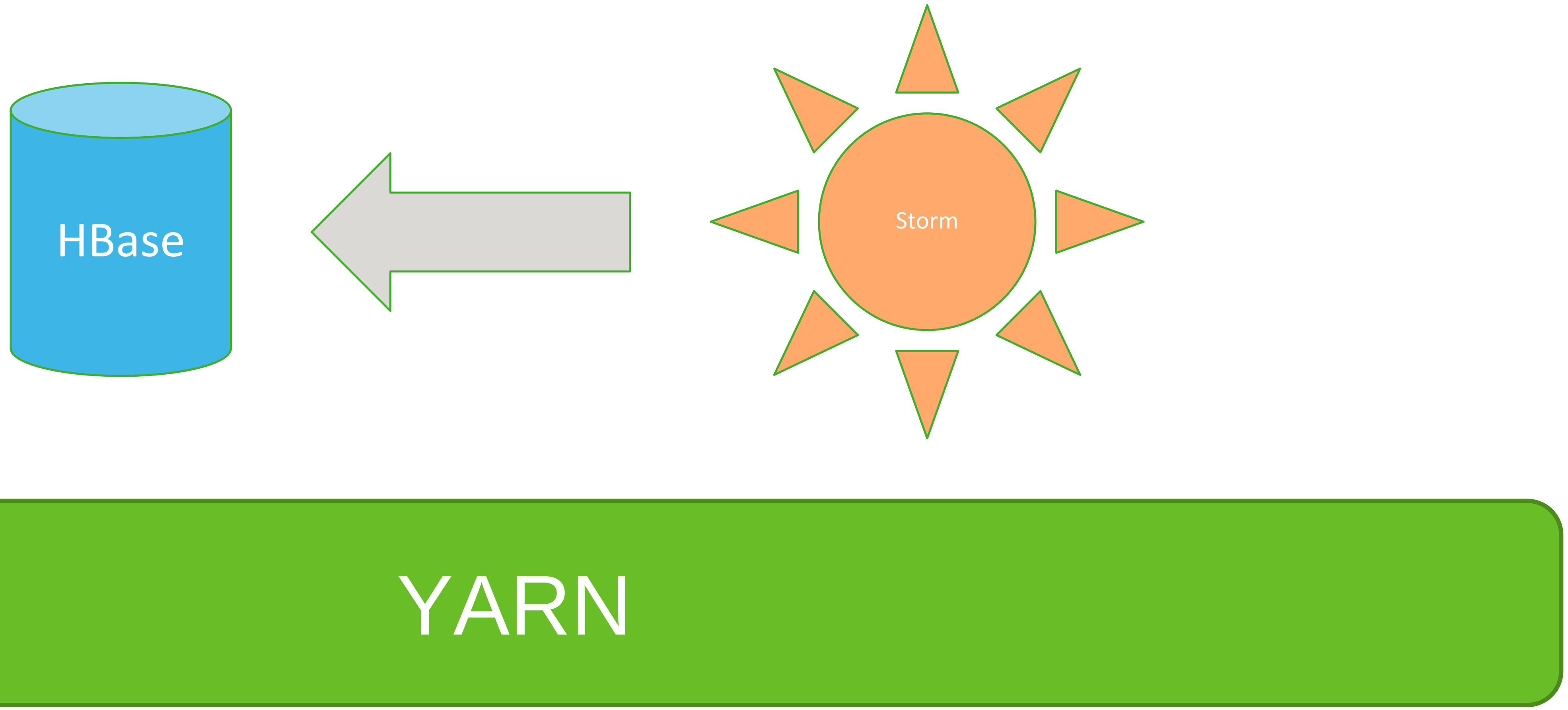
Yanbo Liang
Hortonworks

https://dataworkssummit.com/san-jose-2017/sessions/hadoop-ecosystem-boosts-tensorflow-and-machine-learning-technologies

Wednesday June 14th Ballroom B

Better placement strategies

- Affinity
- Anti-affinity



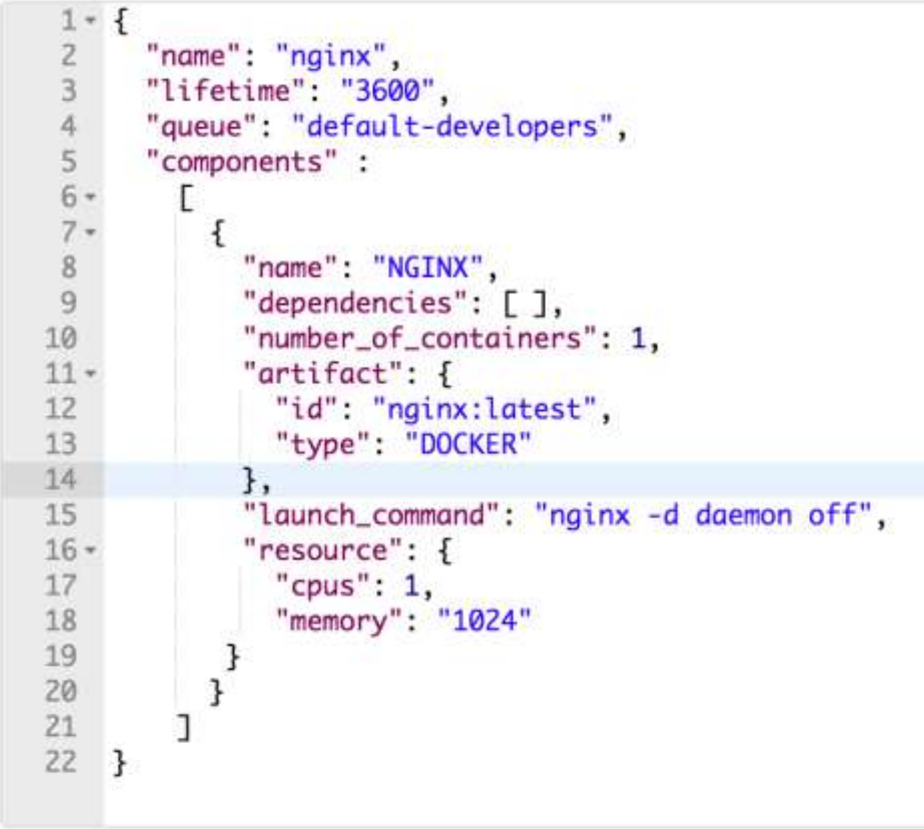
Packaging

- Containers
  - Lightweight mechanism for packaging and resource isolation
  - Popularized and made accessible by Docker
  - Can replace VMs in some cases
  - Or more accurately, VMs got used in places where they didn't need to be
- Native integration ++ in YARN
  - Support for "Container Runtimes" in LCE: YARN-3611
  - Process runtime
  - Docker runtime

Simplified APIs for service definitions

- Applications need simple APIs
- Need to be deployable "easily"

- Simple REST API layer fronting YARN
  - https://issues.apache.org/jira/browse/YARN-4793
  - [Umbrella] Simplified API layer for services and beyond
- Spawn services & Manage them



```
{
  "name": "nginx",
  "lifetime": "3600",
  "queue": "default-developers",
  "components" :
    [
      {
        "name": "NGINX",
        "dependencies": [ ],
        "number_of_containers": 1,
        "artifact": {
          "id": "nginx:latest",
          "type": "DOCKER"
        },
        "launch_command": "nginx -d daemon off",
        "resource": {
          "cpus": 1,
          "memory": "1024"
        }
      }
    ]
}
```

Services support

- Application & Services upgrades
  - ”Do an upgrade of my Spark / HBase apps with minimal impact to end-users”
  - YARN-4726
- Simplified discovery of services via DNS mechanisms: YARN-4757
  - ***regionserver30.hbase-app-3.0.vinodkv.yarn.site***

Services Framework

- Platform is only as good as the tools
- A native YARN services framework
  - https://issues.apache.org/jira/browse/YARN-4692
  - [Umbrella] Native YARN framework layer for services and beyond
- Assembly: Supporting a DAG of apps:
  - https://issues.apache.org/jira/browse/SLIDER-875

-- Related Session --

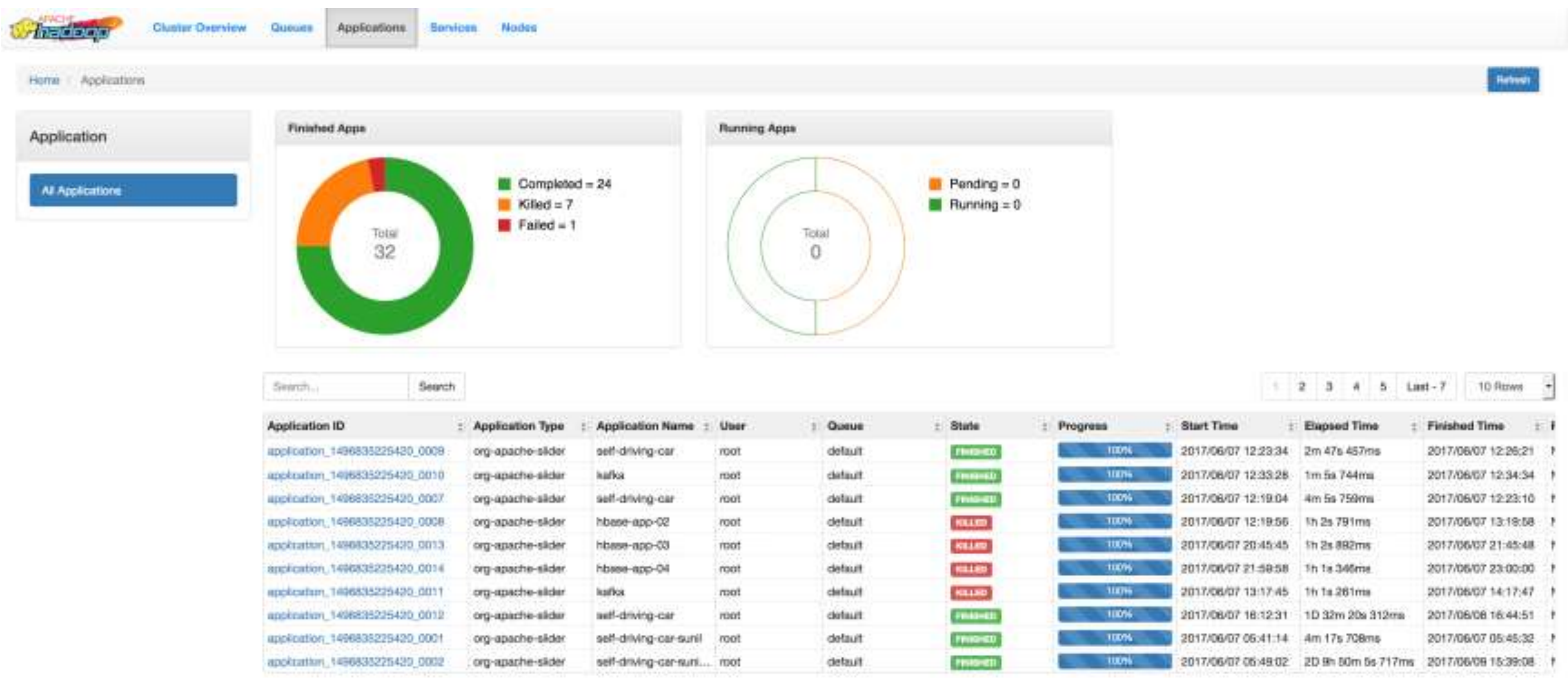
Running a container cloud on YARN

Shane Kumpf
Hortonworks

Jian He
Hortonworks

https://dataworkssummit.com/san-jose-2017/sessions/running-a-container-cloud-on-yarn
Thursday June 15th Room 210C

User experience



Application ID	Application Type	Application Name	User	Queue	State	Progress	Start Time	Elapsed Time	Finished Time
application_1496835225420_0009	org-apache-slider	self-driving-car	root	default	FINISHED	100%	2017/06/07 12:23:34	2m 47s 457ms	2017/06/07 12:26:21
application_1496835225420_0010	org-apache-slider	kafka	root	default	FINISHED	100%	2017/06/07 12:33:28	1m 5s 744ms	2017/06/07 12:34:34
application_1496835225420_0007	org-apache-slider	self-driving-car	root	default	FINISHED	100%	2017/06/07 12:19:04	4m 5s 759ms	2017/06/07 12:23:10
application_1496835225420_0008	org-apache-slider	hbase-app-02	root	default	KILLED	100%	2017/06/07 12:19:56	1h 2s 791ms	2017/06/07 13:19:58
application_1496835225420_0013	org-apache-slider	hbase-app-03	root	default	KILLED	100%	2017/06/07 20:45:45	1h 2s 892ms	2017/06/07 21:45:48
application_1496835225420_0014	org-apache-slider	hbase-app-04	root	default	KILLED	100%	2017/06/07 21:59:58	1h 1s 346ms	2017/06/07 23:00:00
application_1496835225420_0011	org-apache-slider	kafka	root	default	KILLED	100%	2017/06/07 13:17:45	1h 1s 261ms	2017/06/07 14:17:47
application_1496835225420_0012	org-apache-slider	self-driving-car	root	default	FINISHED	100%	2017/06/07 16:12:31	1D 32m 20s 312ms	2017/06/08 16:44:51
application_1496835225420_0001	org-apache-slider	self-driving-car-sunil	root	default	FINISHED	100%	2017/06/07 05:41:14	4m 17s 708ms	2017/06/07 05:45:32
application_1496835225420_0002	org-apache-slider	self-driving-car-suni...	root	default	FINISHED	100%	2017/06/07 05:49:02	2D 9h 50m 5s 717ms	2017/06/09 15:39:08

API based queue management

Decentralized

New web UI

Improved logs management

Live application logs

Timeline Service

- Application History
  - “Where did my containers run?”
  - “Why is my application slow?”
  - “Is it really slow?”
  - “Why is my application failing?”
  - “What happened with my application? Succeeded?”

- Cluster History
  - Run analytics on historical apps!
  - “User with most resource utilization”
  - “Largest application run”
  - “Why is my cluster slow?”
  - “Why is my cluster down?”
  - “What happened in my clusters?”
- Collect and use past data
  - To schedule “my application” better
  - To do better capacity planning

Timeline Service 2.0

- **Next generation**
  - Today’s solution helped us understand the space
  - Limited scalability and availability

- ***“Analyzing Hadoop Clusters is becoming a big-data problem”***
  - Don’t want to throw away the Hadoop application metadata
  - Large scale
  - Enable near real-time analysis: *“Find me the user who is hammering the FileSystem with rouge applications. Now.”*

- **Timeline data stored in HBase and accessible to queries**

Building a modern end-to-end open source Big Data reference application

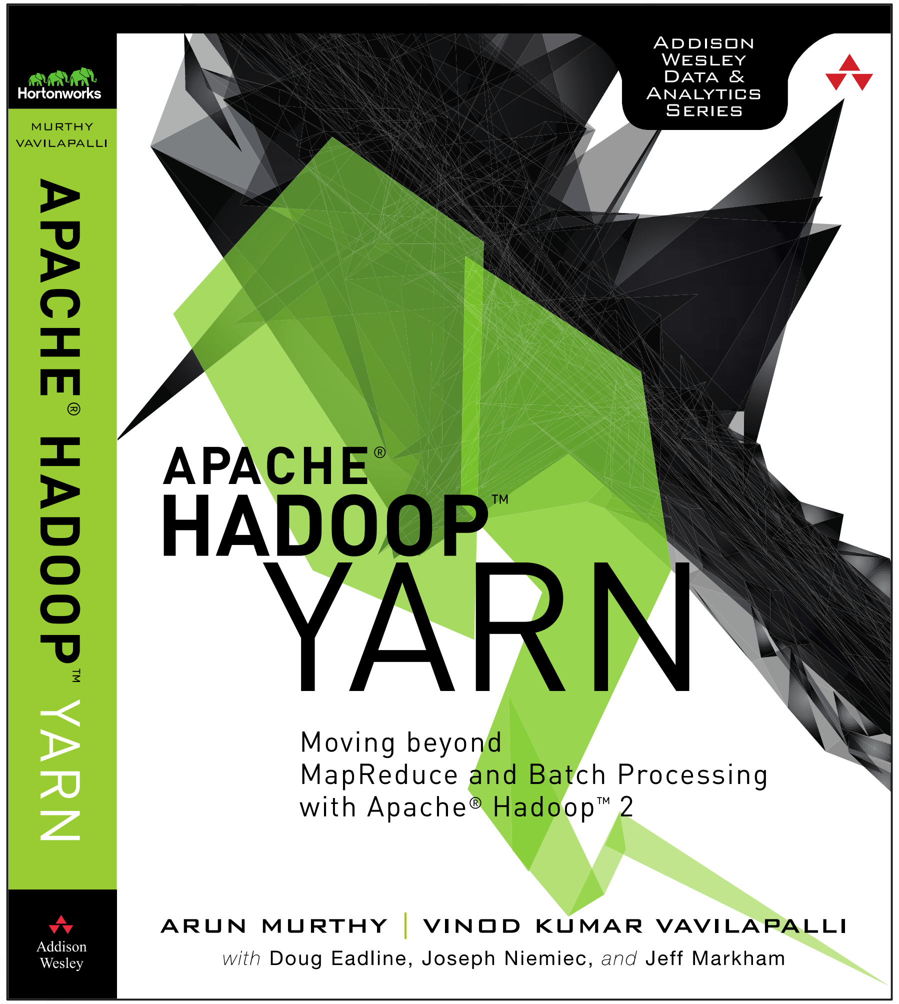
Edgar Orendain

UC Berkeley / Hortonworks

https://dataworkssummit.com/san-jose-2017/sessions/building-a-modern-end-to-end-open-source-big-data-reference-application

Wednesday June 14th Ballroom C

Thank you!



APACHE® HADOOP™ YARN

Moving beyond
MapReduce and Batch Processing
with Apache® Hadoop™ 2

ARUN MURTHY | VINOD KUMAR VAVILAPALLI

with Doug Eadline, Joseph Niemiec, *and* Jeff Markham

Hortonworks Sandbox

VERSION 2.0

Try out the latest in Hadoop Innovation

Sandbox is a personal, portable Hadoop environment that comes with a dozen interactive Hadoop tutorials. Sandbox includes many of the most exciting developments from the latest HDP distribution, packaged up in a virtual environment that you can get up and running in 15 minutes!

Learn Hadoop
Sandbox comes with a dozen hands-on tutorials that will guide you through the basics of Hadoop; tutorials built on the experience gained from training thousands of people in our Hortonworks University Training classes.

Build a Proof of Concept
The Sandbox includes the Hortonworks Data Platform in an easy to use form. You can add your own datasets, and connect it to your existing tools and applications. With this, you can prove out your use of Hadoop and plan the integration points for your first Hadoop project.

Test New Functionality
You can test new functionality with the Sandbox before you put it into production. Simply, easily and safely.