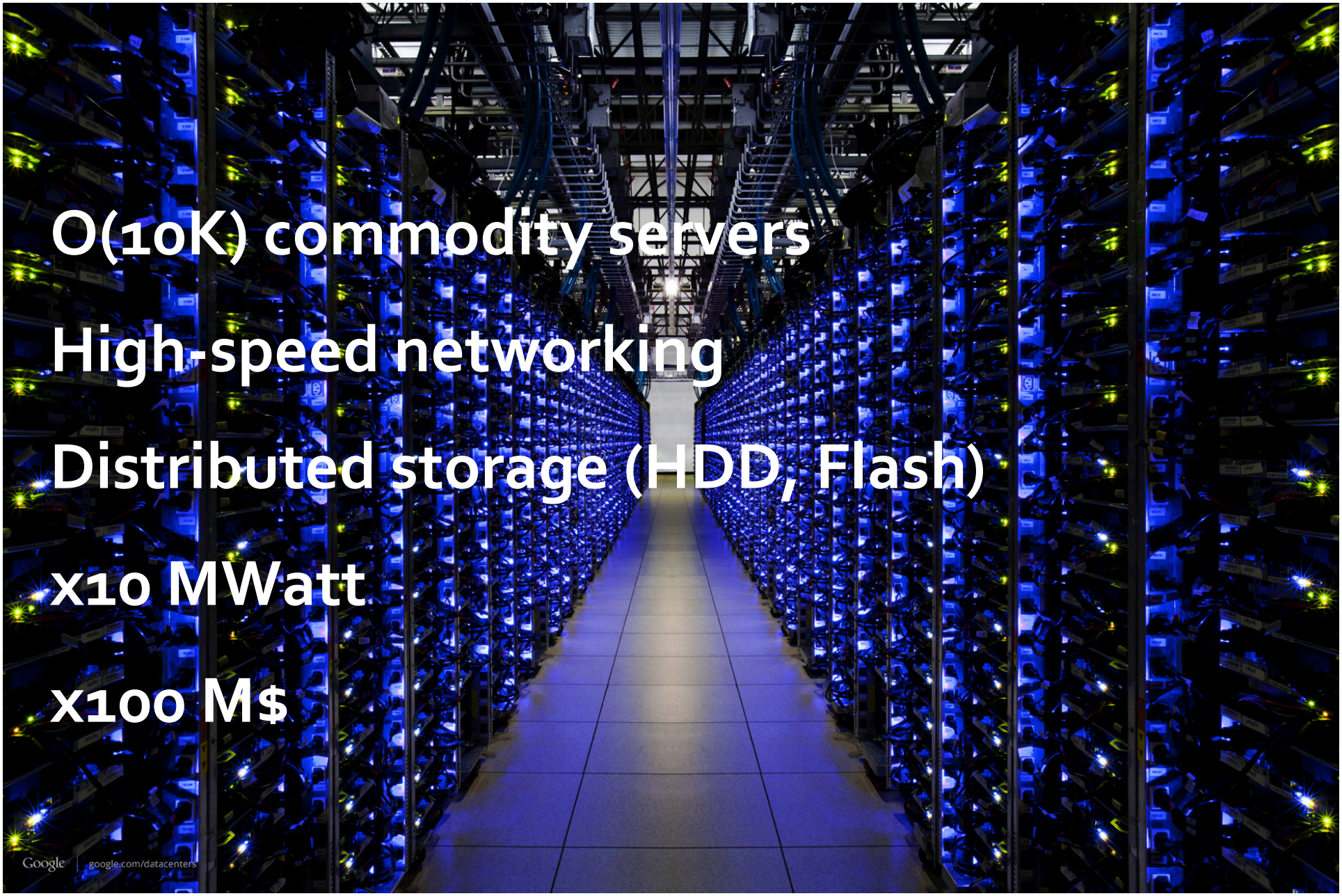
Re:platforming the Datacenter with Apache Mesos

Christos Kozyrakis


MESOSPHERE

Why your ASF project should run on Mesos


O(10K) commodity servers
High-speed networking
Distributed storage (HDD, Flash)
x10 MWatt
x100 M$
Google | google.com/datacenters

developers

automation

performance

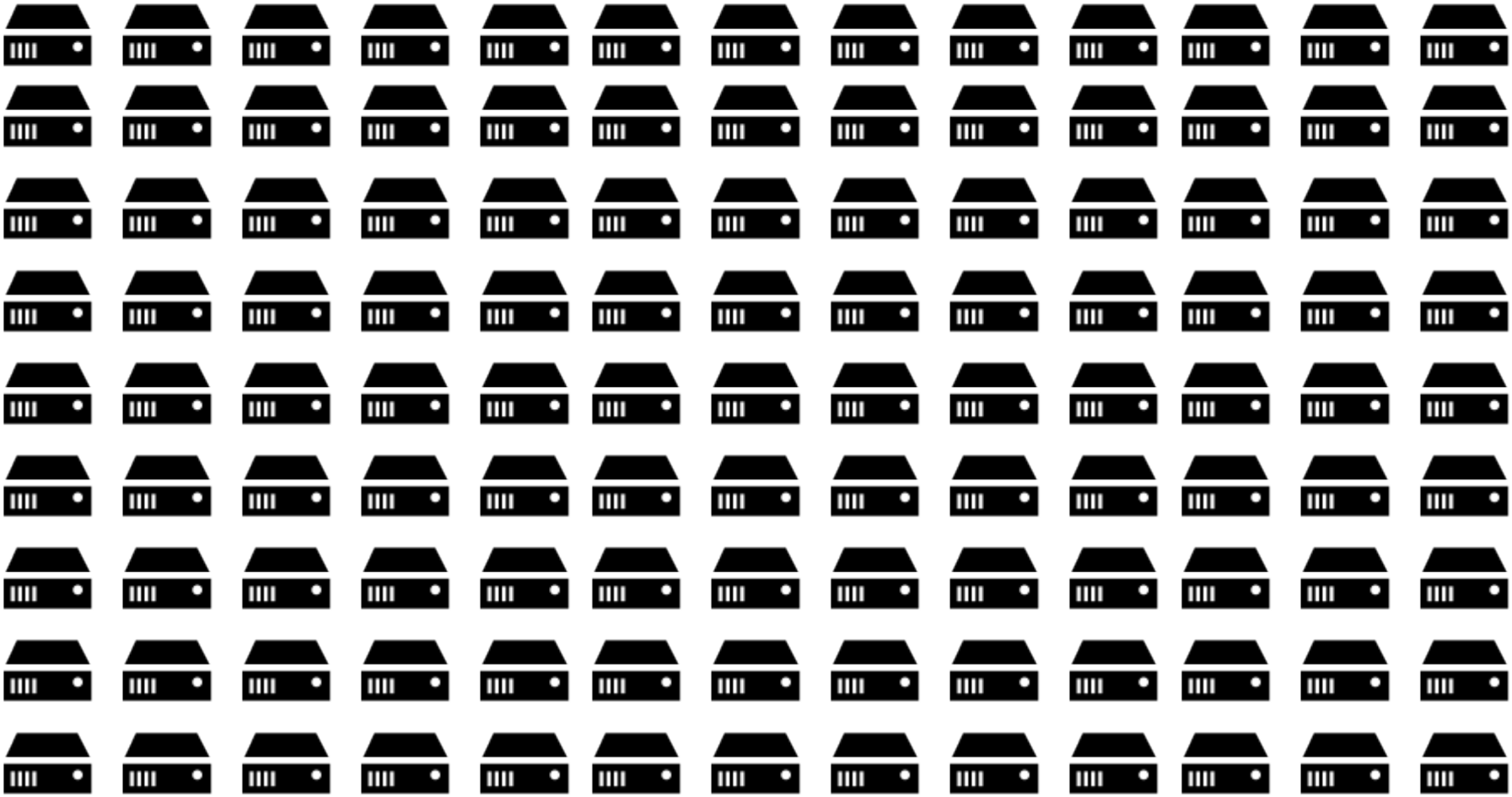
ops

automation

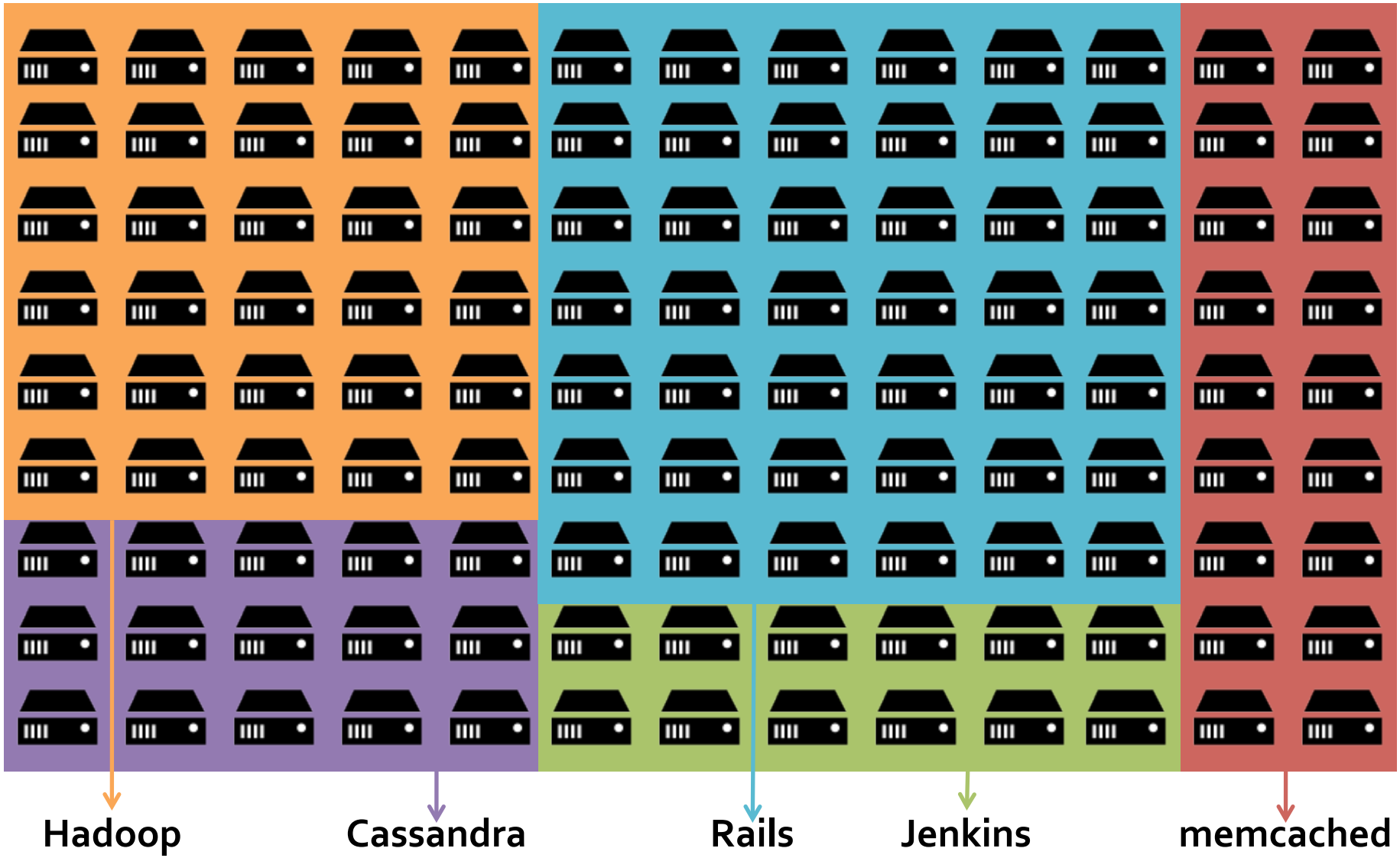
efficiency

① Datacenter past

Static Partitioning

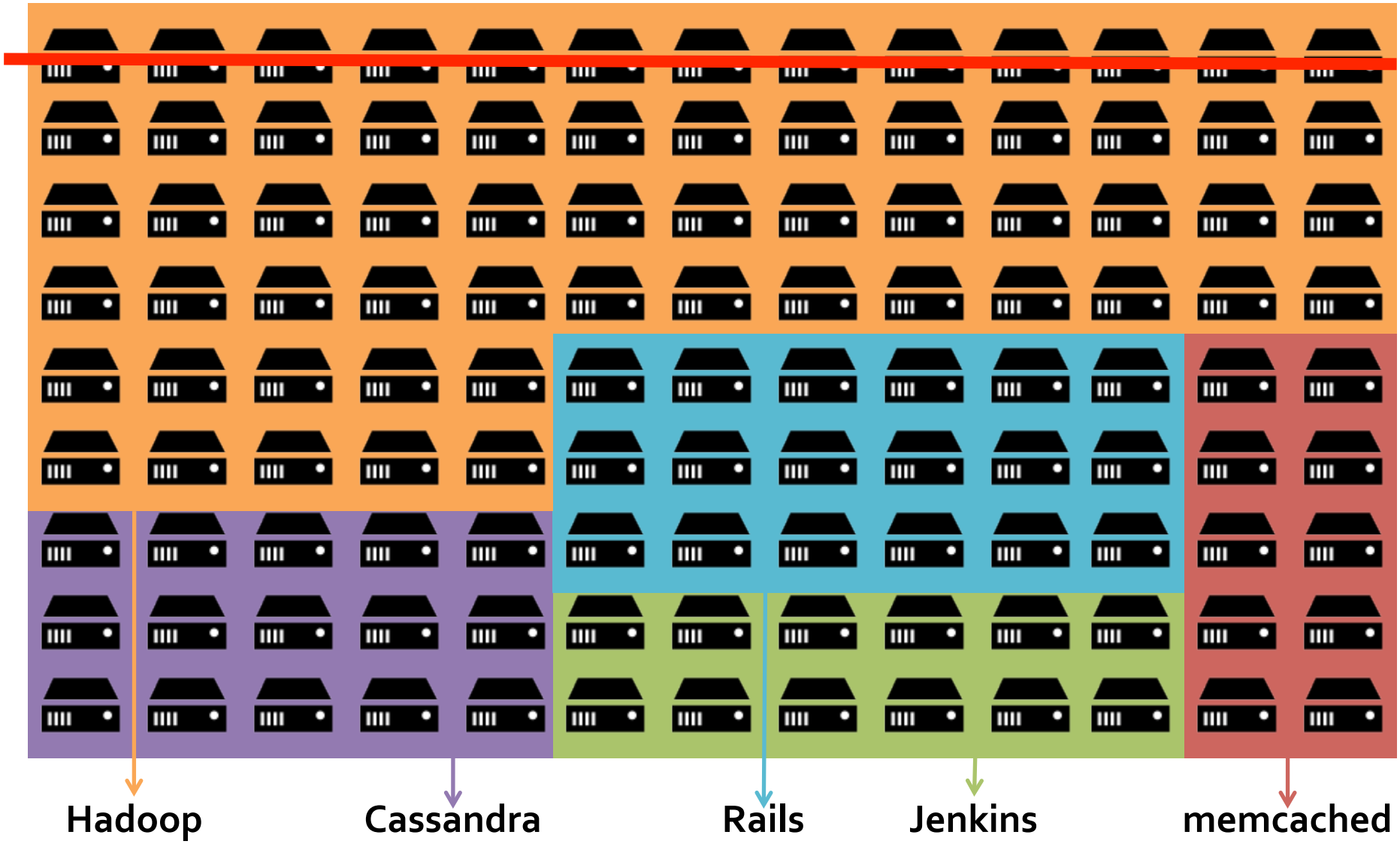


Static Partitioning

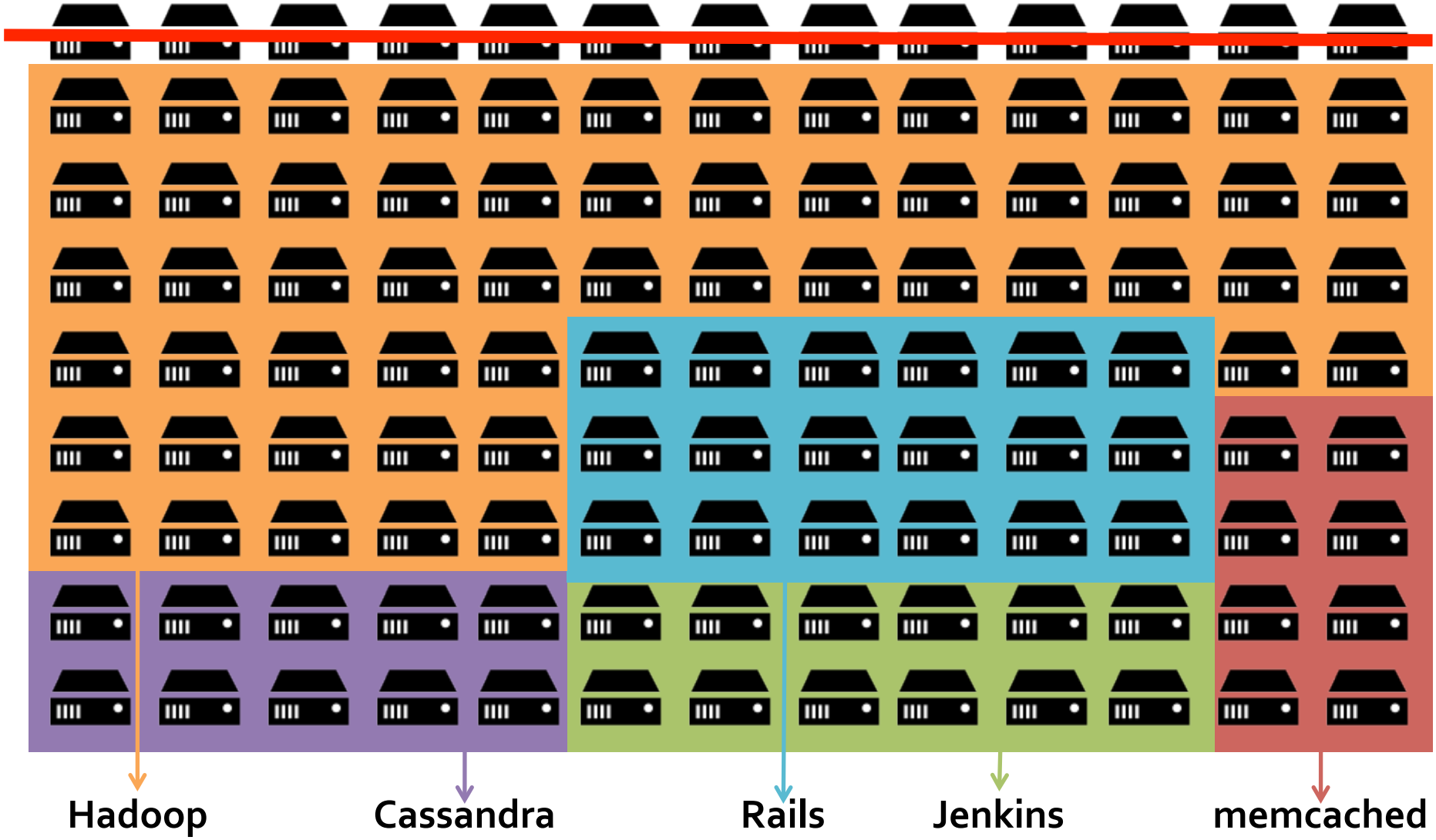


Static Partitioning

Static Partitioning



Static Partitioning



Static Partitioning

developers

☒ automation

☑ performance

ops

☒ automation

☒ efficiency

② Datacenter present

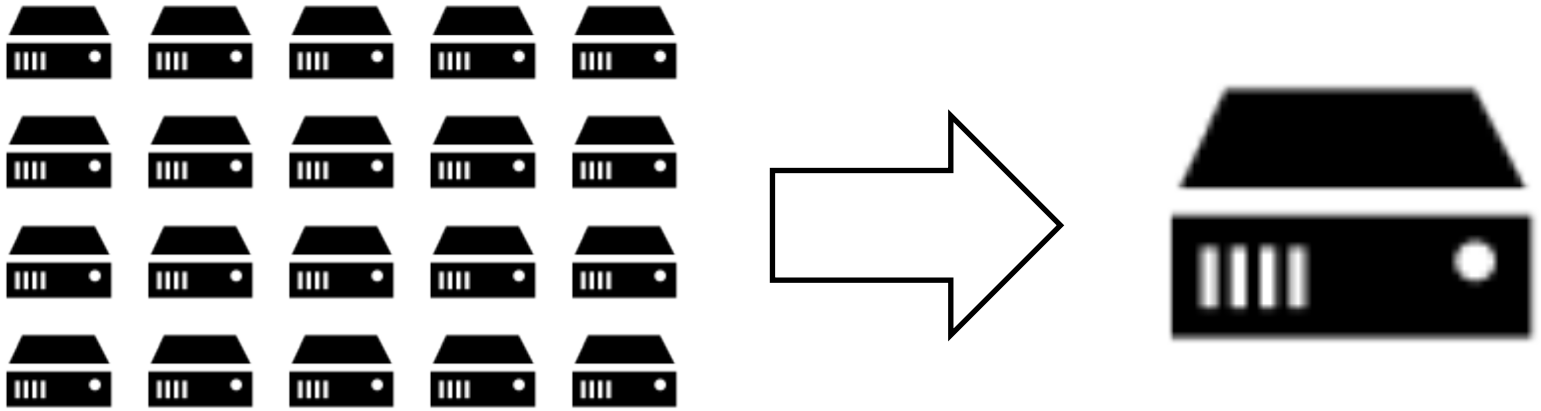
Apache Mesos

The datacenter OS kernel

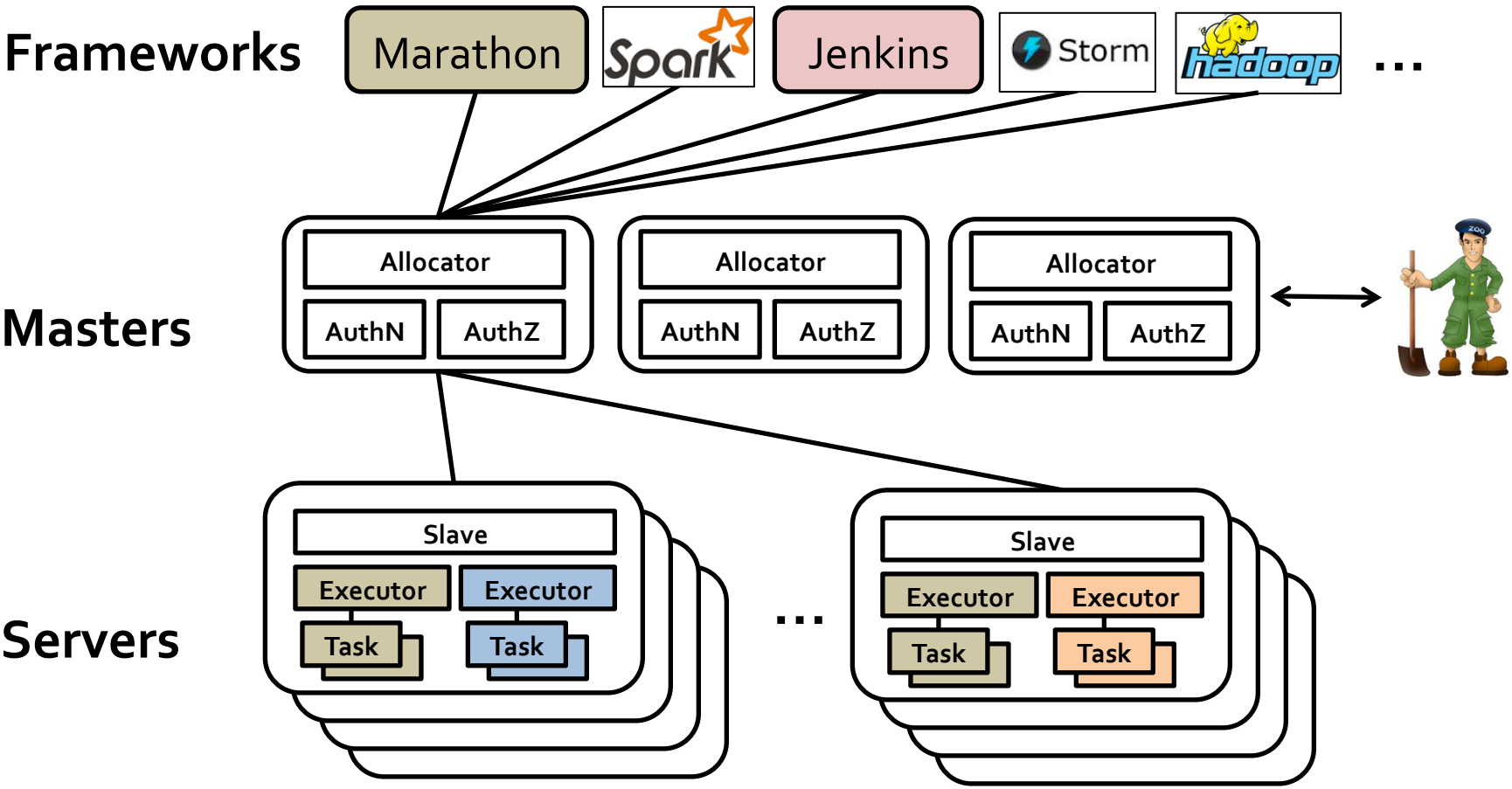
Aggregates all resources into a single shared pool

Dynamically allocates resources to distributed apps

Container management at scale (cgroups, docker, …)

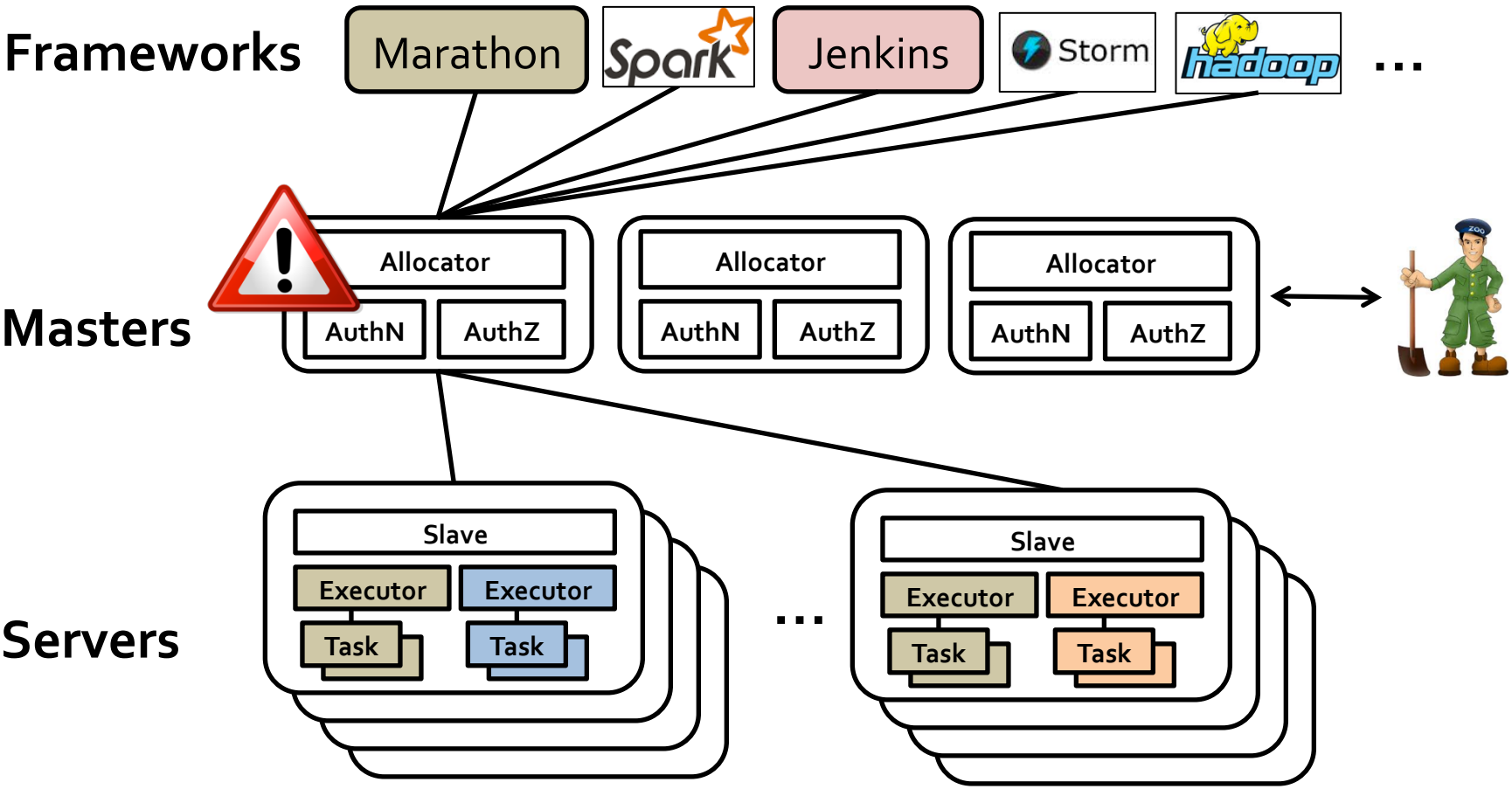


Mesos Architecture



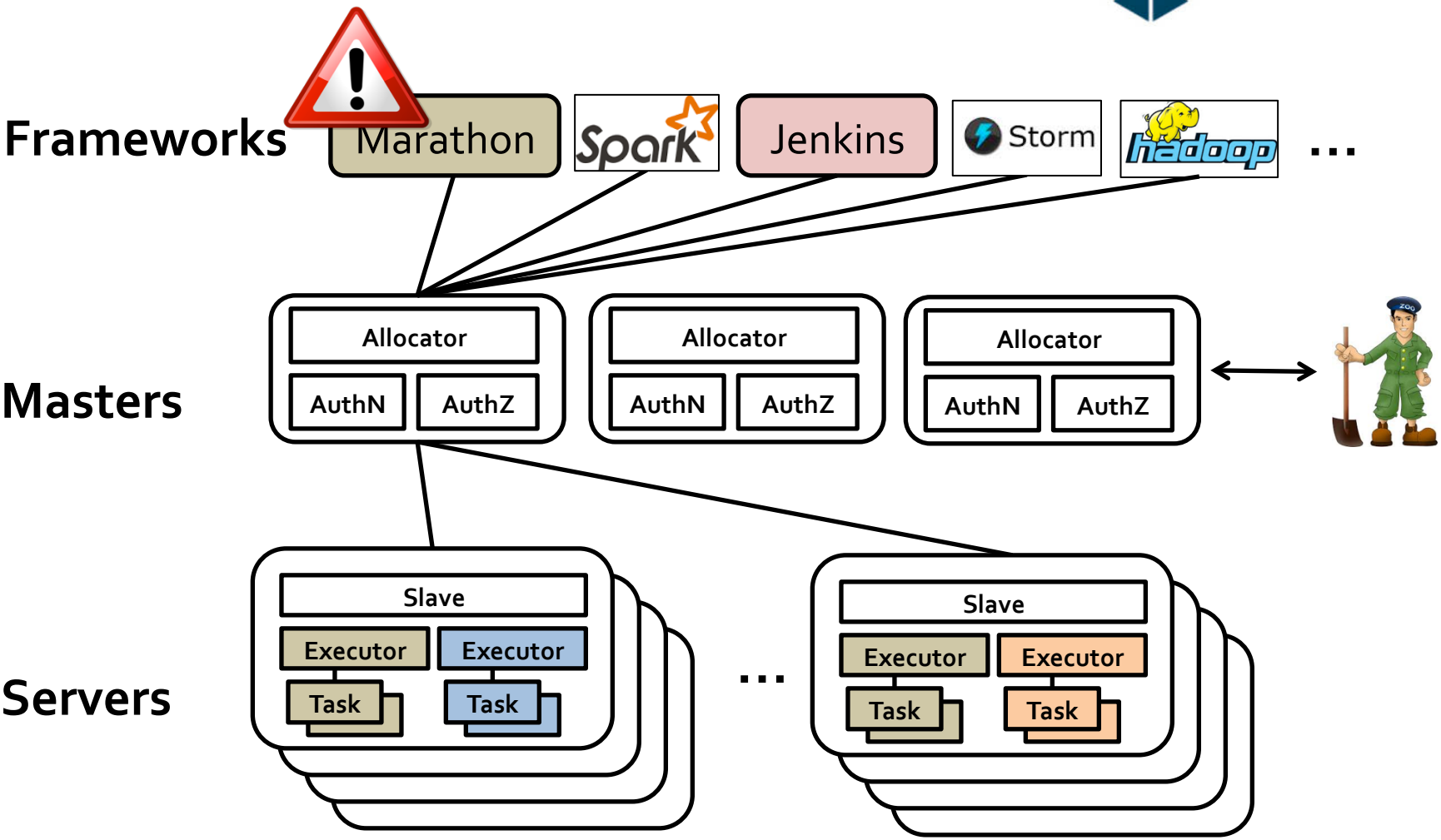
Scales to 10s of thousands of servers

Mesos Fault Tolerance



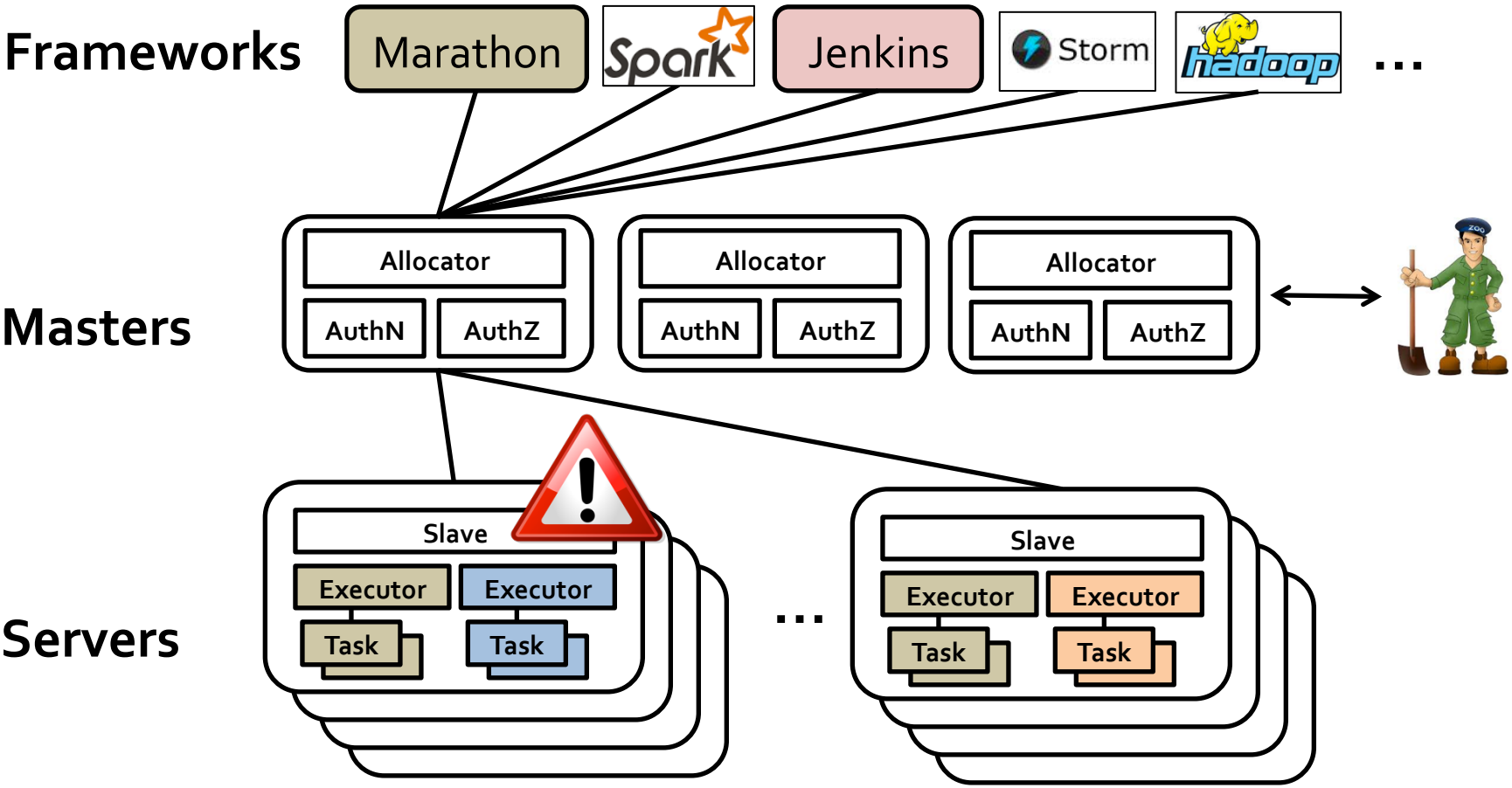
Tasks survive failures of the master

Mesos Fault Tolerance



Tasks survive failures of the framework

Mesos Fault Tolerance



Tasks survive failures of the slave process

Scheduling in Mesos

No single scheduler fits all needs

Long-running services need

- scale up/down, fault tolerance

Analytics services need

- fast task launching

2-level Scheduling

Mesos master (single API)

- Resource allocation (offers)
- Task health checks
- Task isolation

Mesos frameworks (domain-specific APIs)

- Scale up/down
- Fault tolerance
- Task grouping
- Task dependencies
- Queuing & priorities

2-level Scheduling Benefits

Multiple APIs to Mesos through frameworks

- Marathon, Aurora, Singularity
- Storm, Spark, Hadoop, Chronos
- Cassandra, Elasticsearch

Multiple task scheduling approaches

- Spark fine-grain Vs Spark coarse-grain

Simple, stable, and scalable Mesos master

Dynamic Resource Allocation

Resource allocation based on framework roles

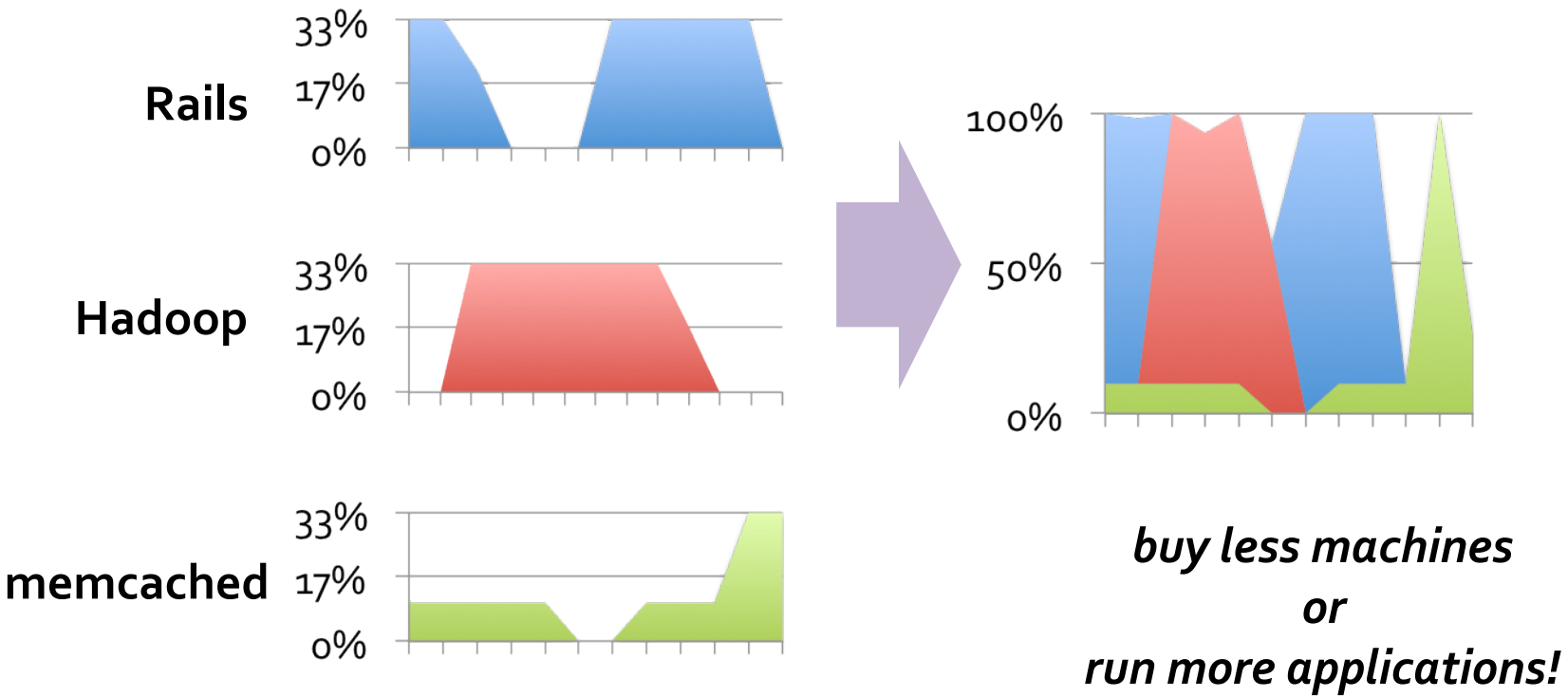
Dominant resource fairness (DRF)

- Weighted fair share calculated based on dominant resource
- Frameworks do no worse than having a weight-sized cluster

Resource reservations

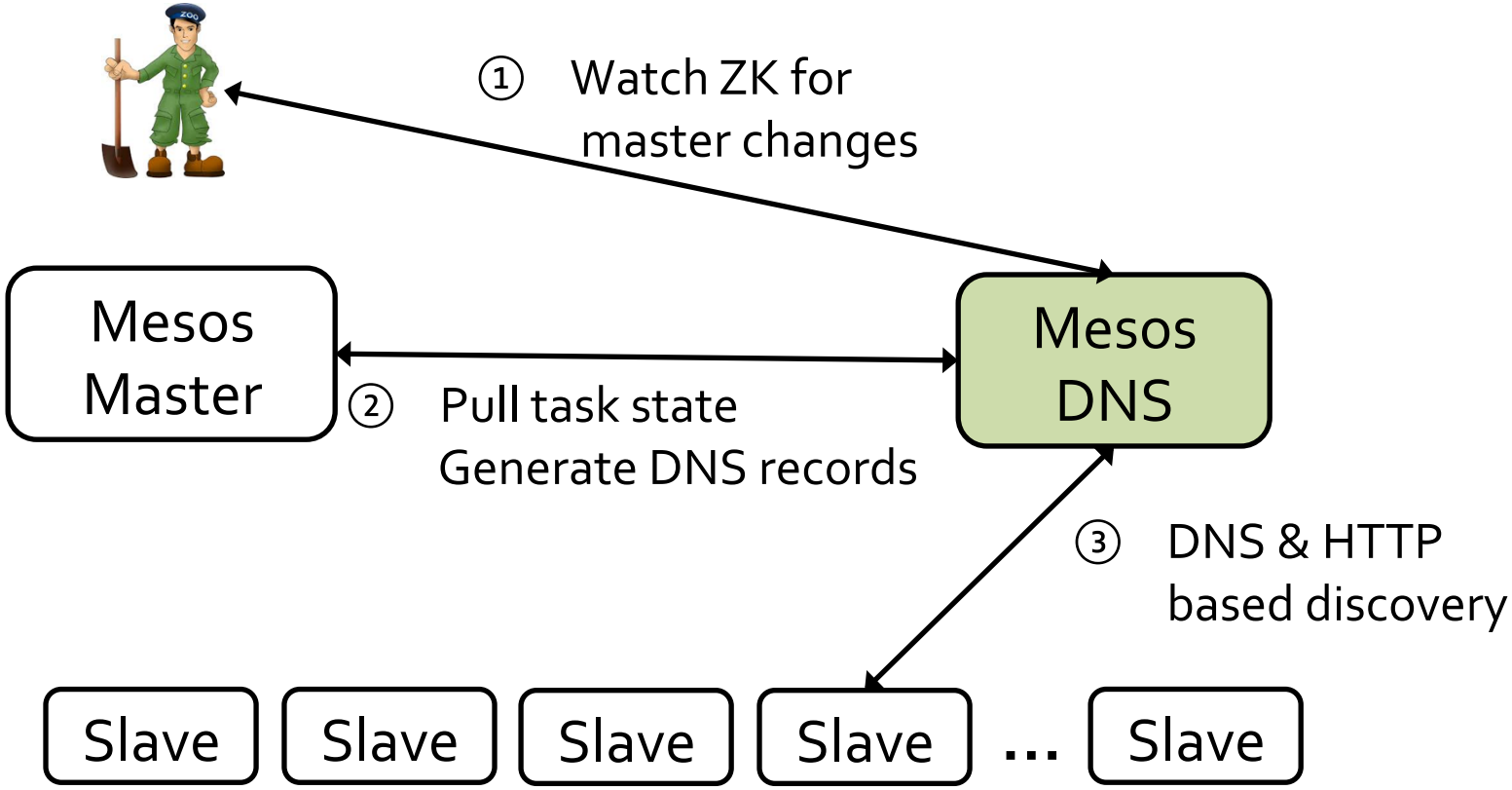
- Resources can be allocated to specific frameworks if needed

Dynamic Resource Allocation



buy less machines
or
run more applications!

Service Discovery

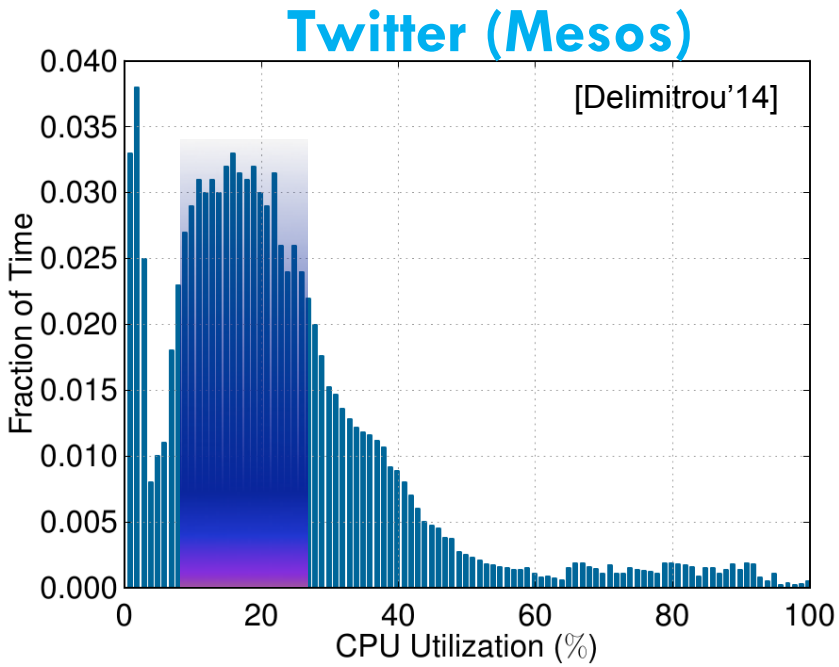


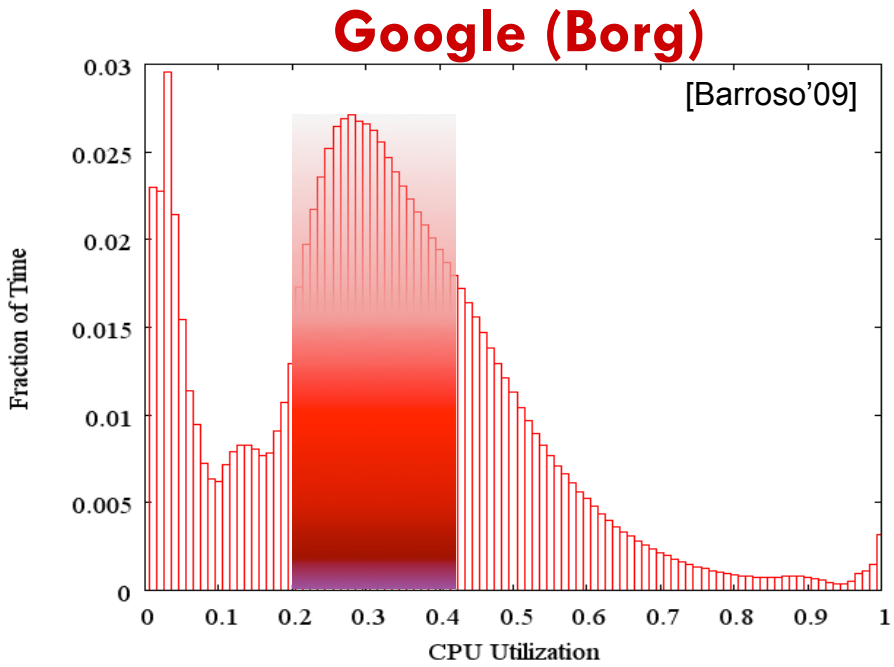
nginx.marathon.mesos → 10.13.17.95

_nginx._tcp.marathon.mesos →10.13.17.95:8181

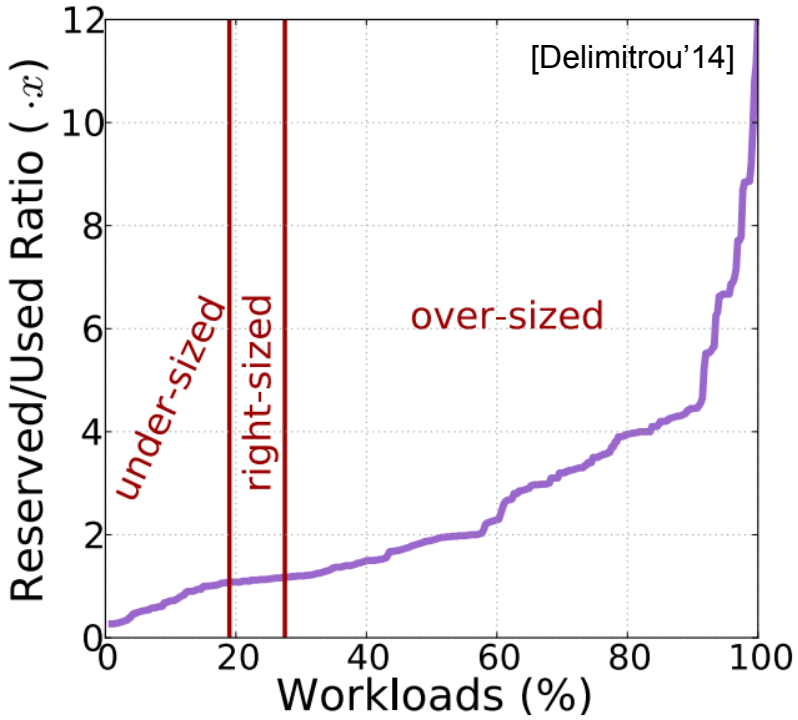
③ Datacenter future

Utilization Reality





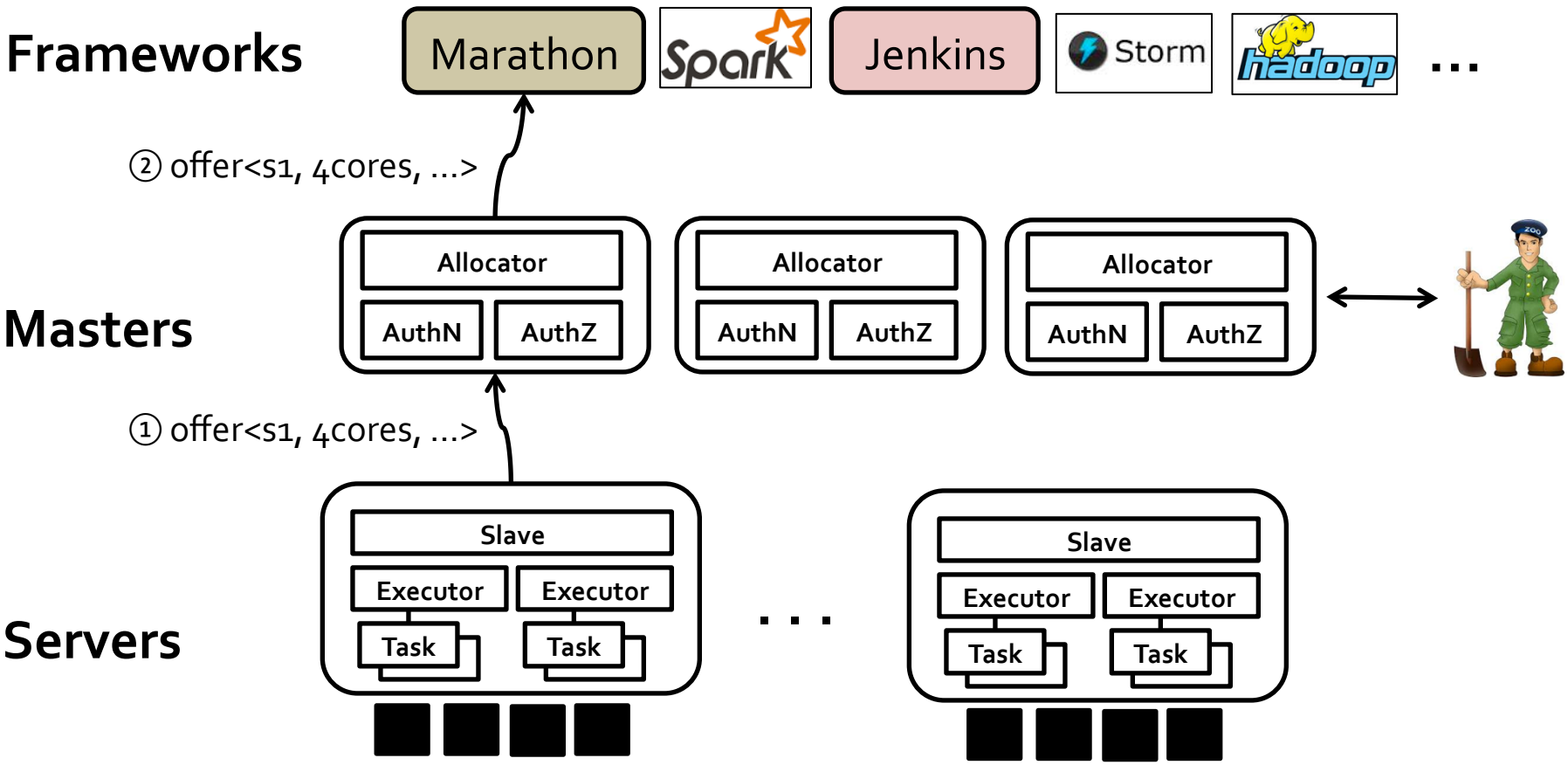
The Curse of Overprovisioning



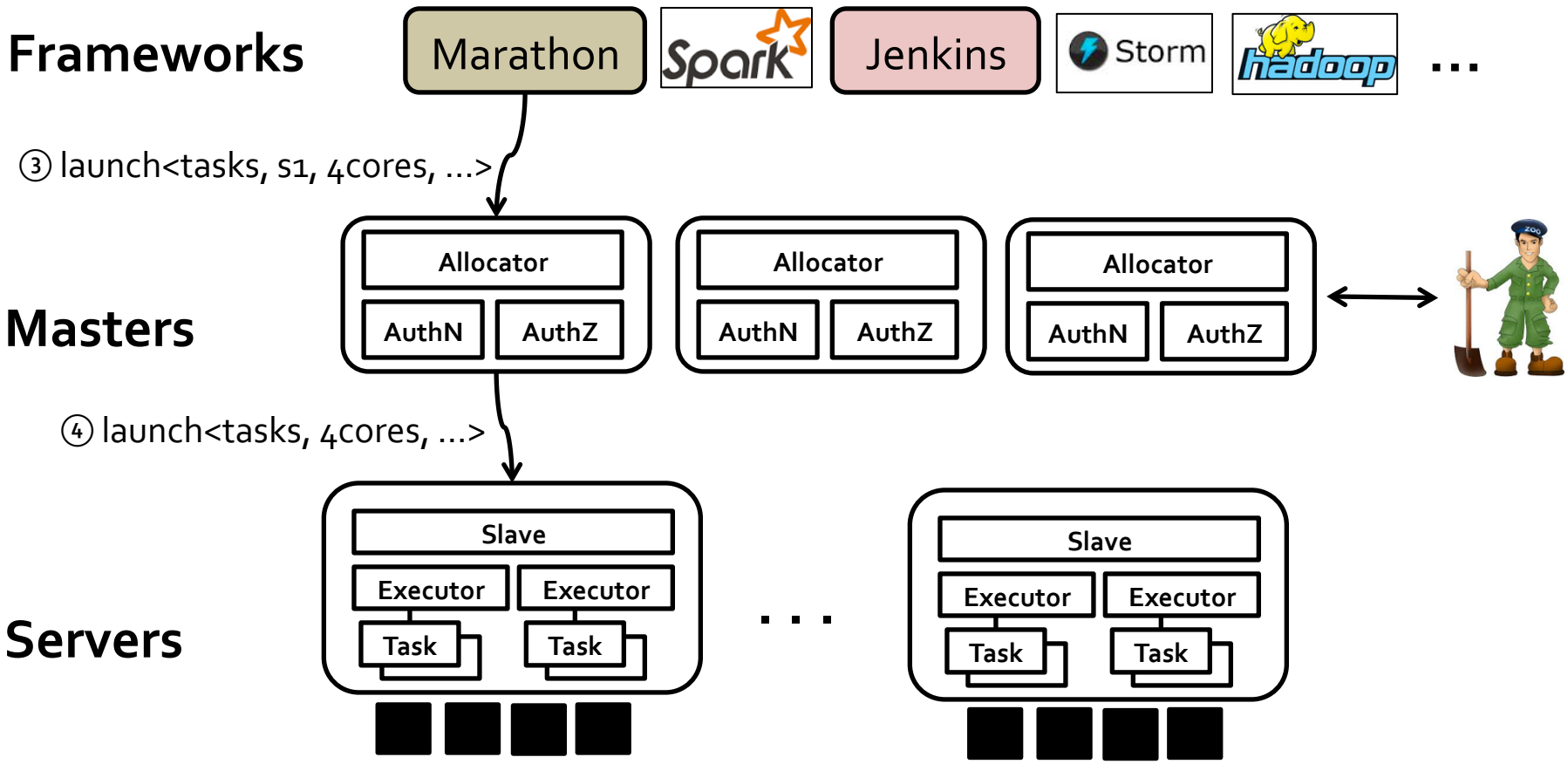
Bloated reservations to deal with

diurnal load patterns, load spikes, software & platform changes

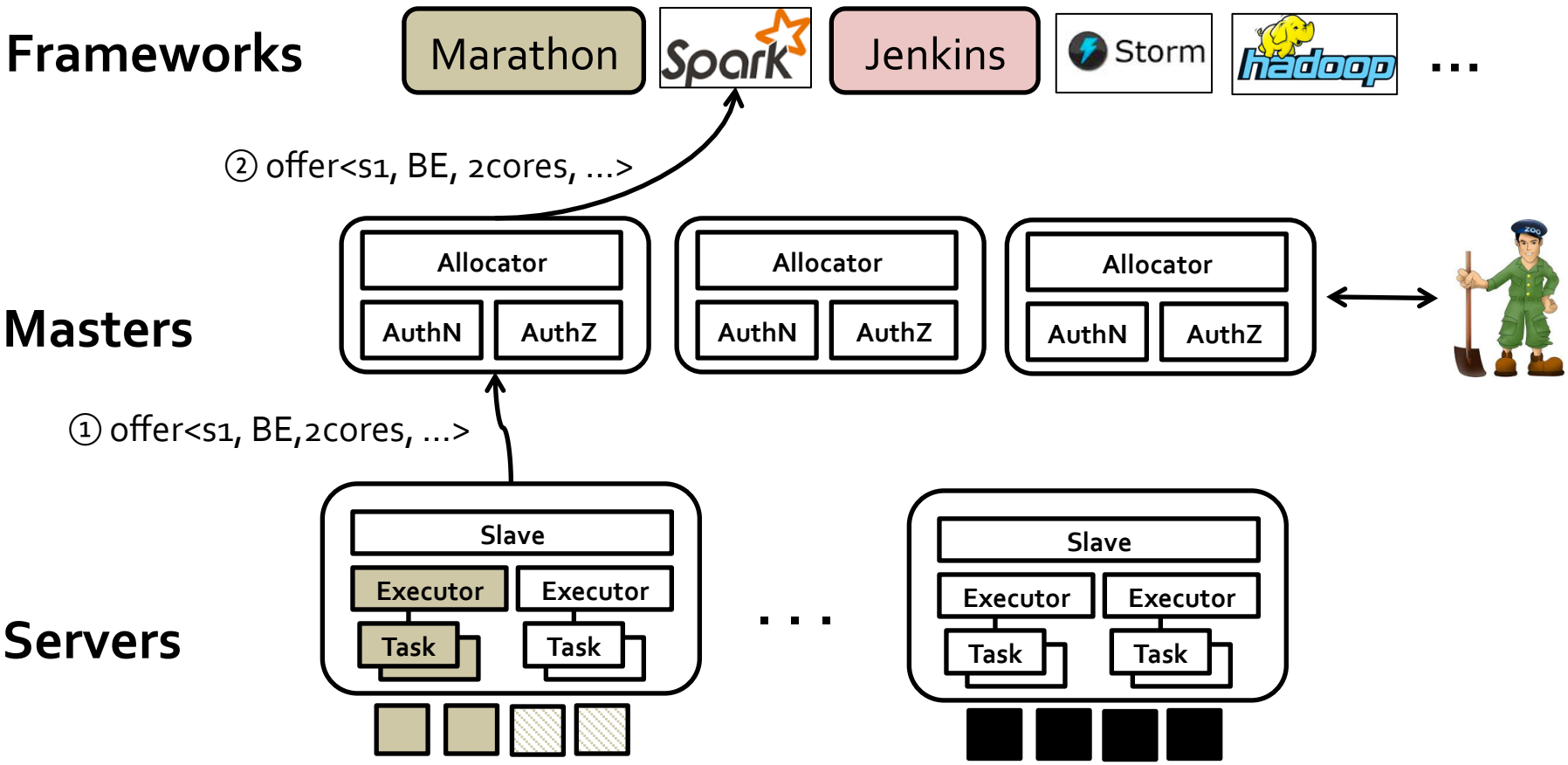
Oversubscription



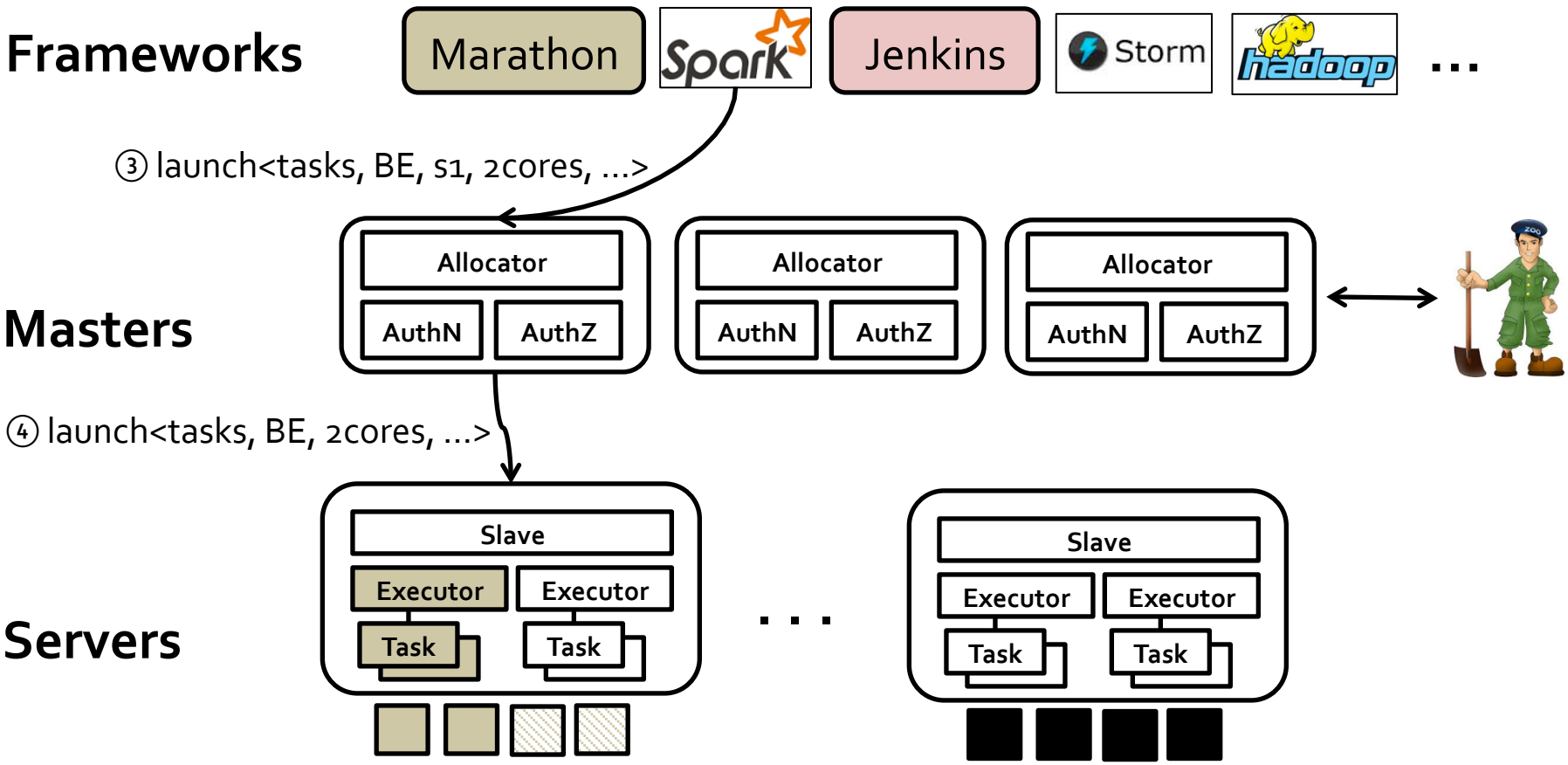
Oversubscription



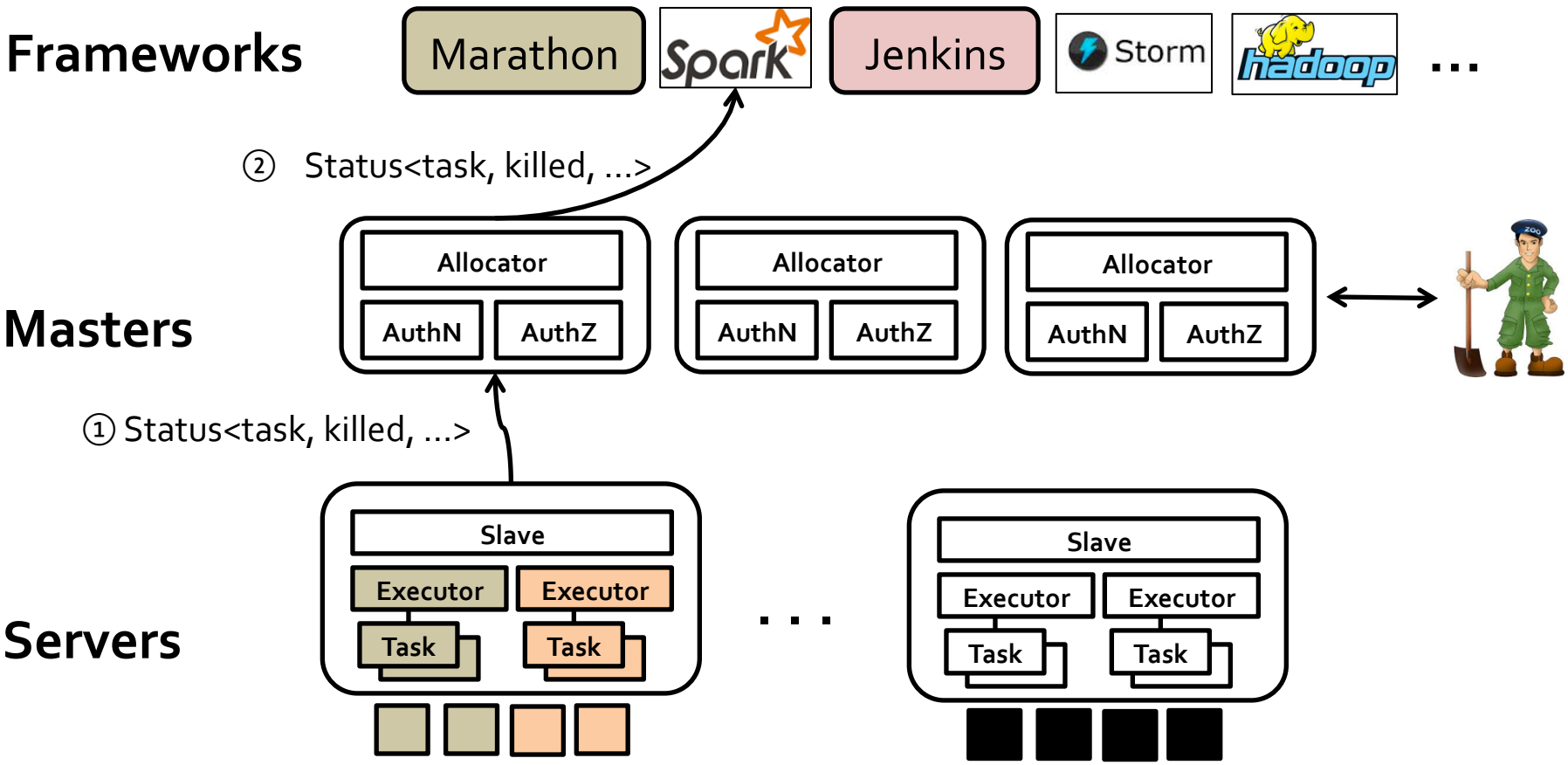
Oversubscription



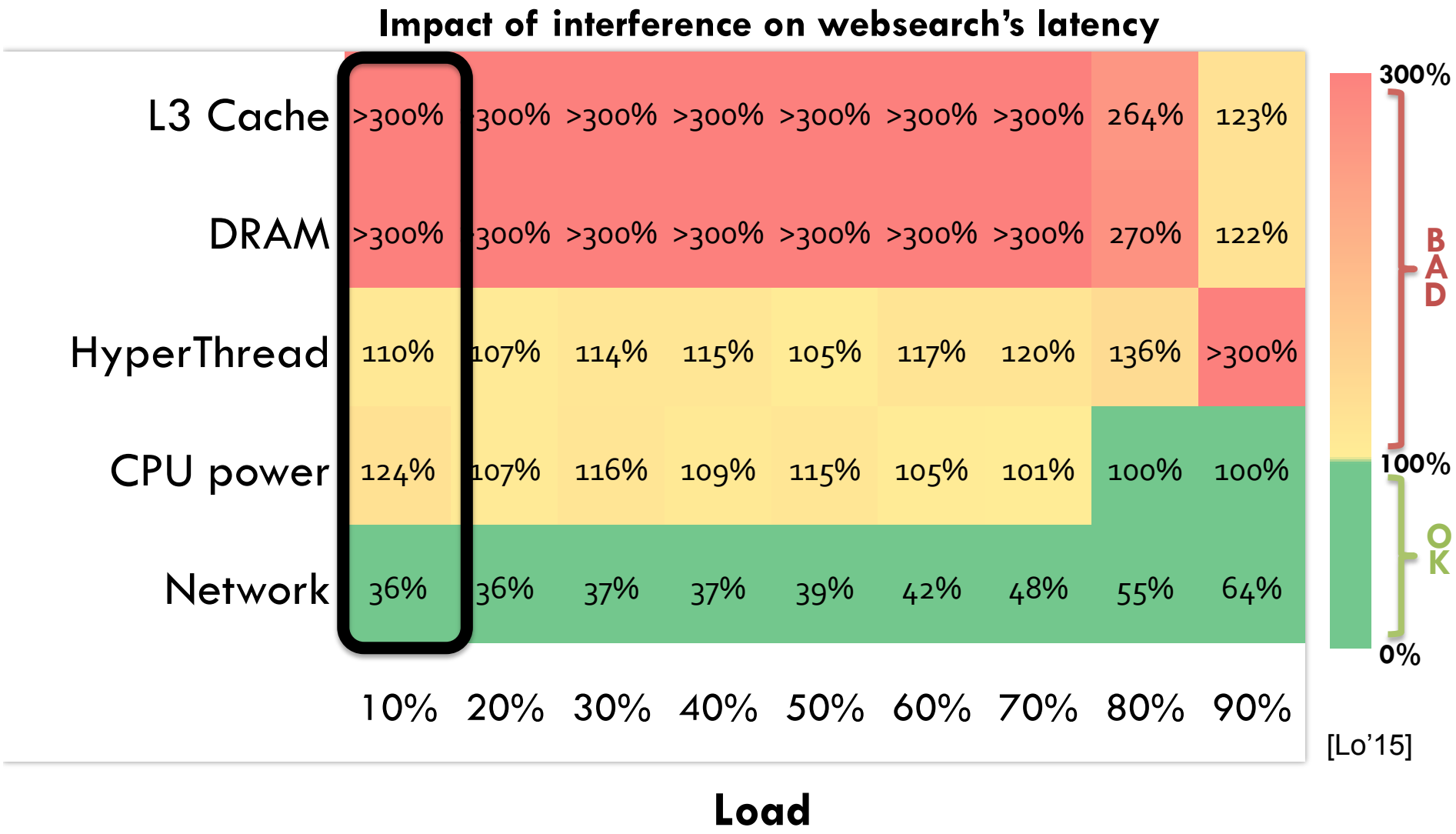
Oversubscription



Oversubscription



Interference → Performance Loss



Isolators

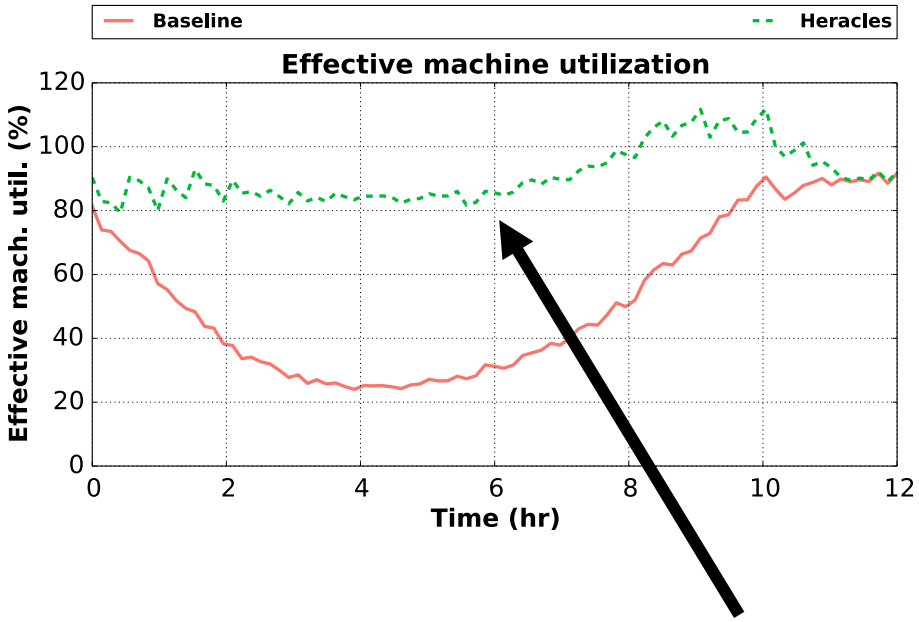
Mesos slave invokes isolators

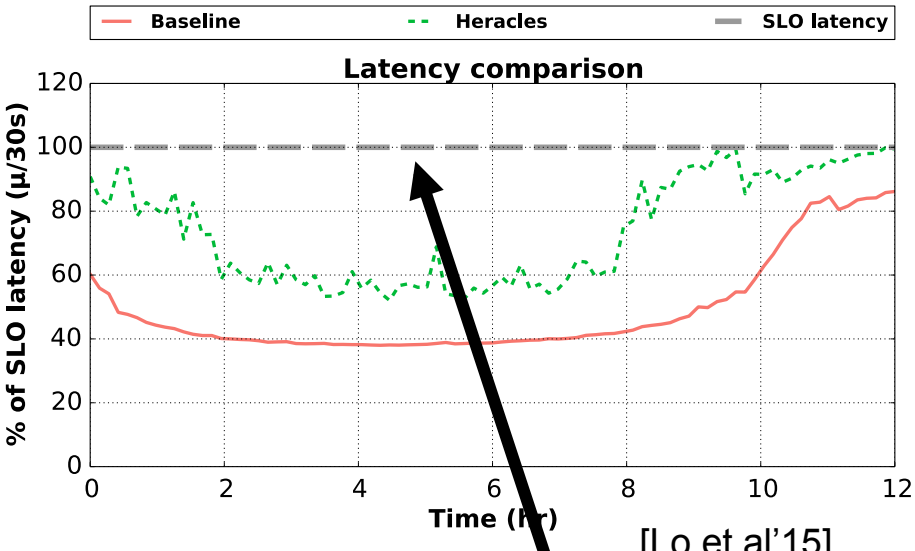
- Modules that monitor & isolate resources for executors

Isolator modules

- CPU (cgroups cpushares, cpusets)
- Memory (cgroups)
- Disk
- Network
- Cache
- Power
- …

Isolators → Performance QoS



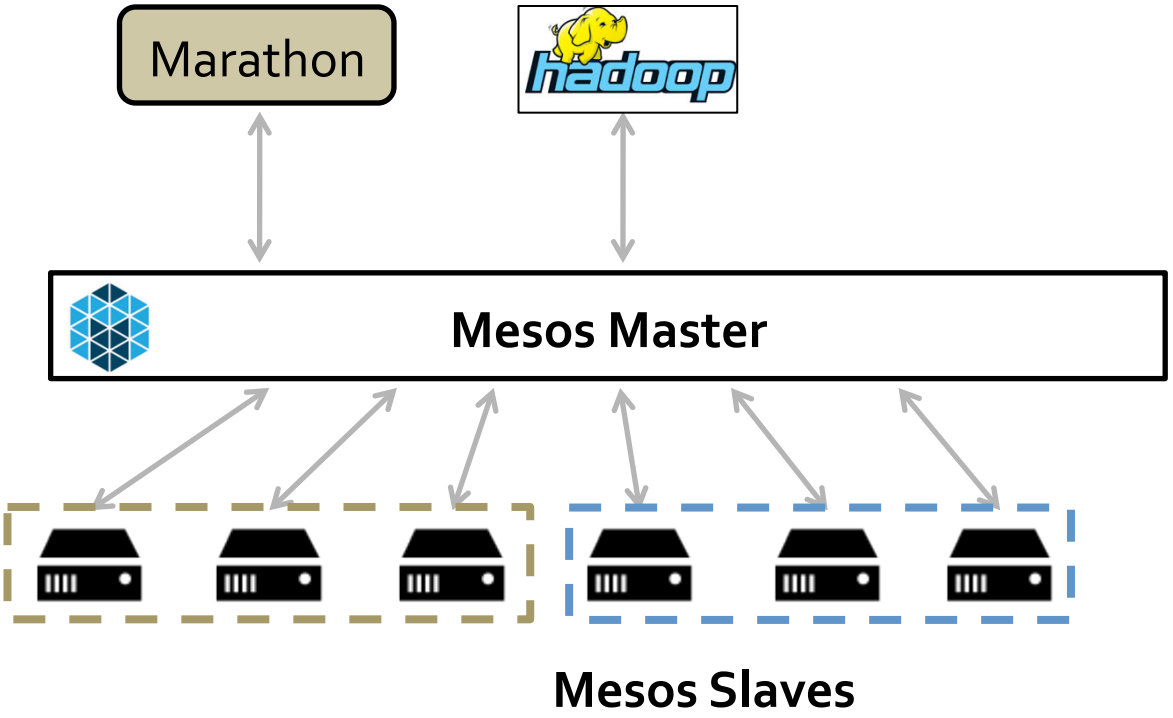


[Lo et al'15]

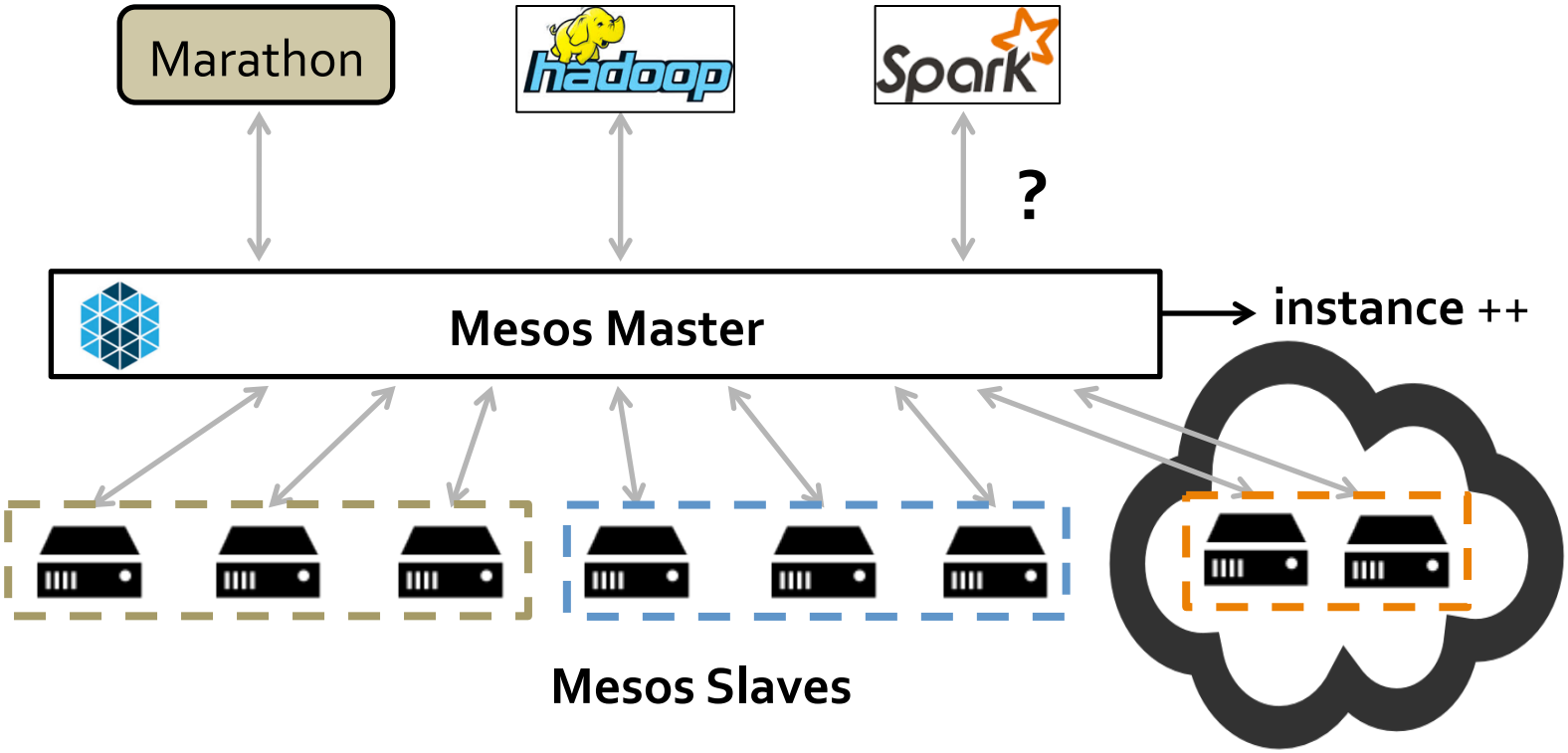
+ best-effort task
>90% HW utilization

No latency SLO problems

Hybrid Datacenters



Hybrid Datacenters



Scheduling based on workload type, data locality, pricing,…

Future Directions

Oversubscription

Container & application right sizing

Hybrid datacenters

Power aware scheduling

Locality aware scheduling

Mesos Datacenter

developers

- ☑ automation
- ☑ performance

ops

- ☑ automation
- ☑ efficiency

Want to Learn More?

Mo 3pm – Cracking the Container Scale Problem with Apache Mesos

Tue 4.20pm – The Emergence of the Datacenter Developer

Wed 4.15pm – Mesos + Yarn = Myriad

Questions?

References

- http://mesos.apache.org/
- http://www.mesosphere.com/
- https://github.com/mesosphere/mesos-dns
- https://www.cs.berkeley.edu/~alig/papers/mesos.pdf
- https://www.cs.berkeley.edu/~alig/papers/drf.pdf
- http://www.morganclaypool.com/doi/abs/10.2200/S00516ED2V01Y201306CAC024
- http://web.stanford.edu/~cdel/2014.asplos.quasar.pdf
- http://web.stanford.edu/~davidlo/resources/2014.heracles.isca.pdf