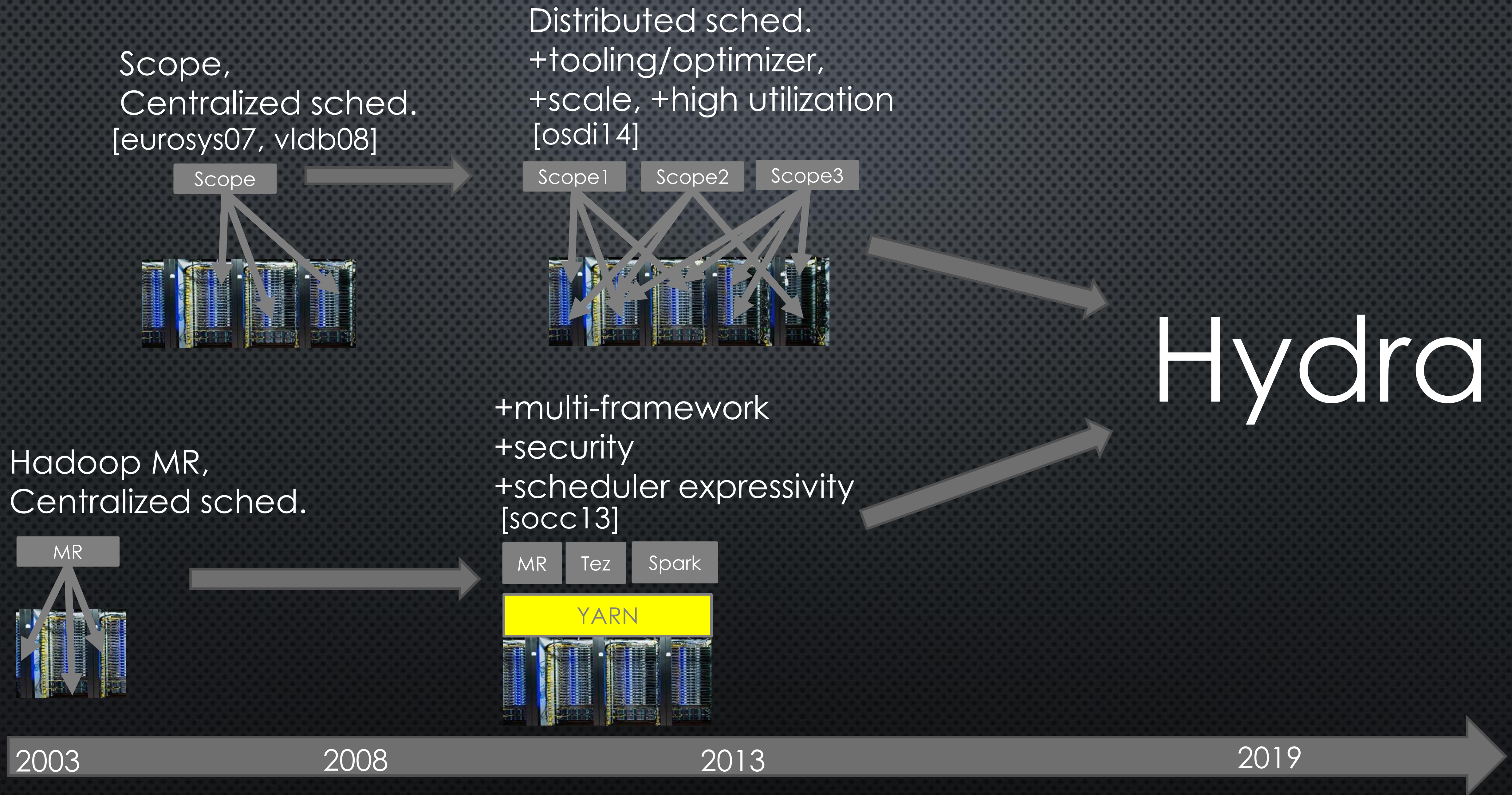
HYDRA

A FEDERATED RESOURCE MANAGER FOR DATA-CENTER SCALE ANALYTICS

Carlo Curino, Subru Krishnan, Konstantinos Karanasos, Sriram Rao, Giovanni M. Fumarola, Botong Huang, Kishore Chaliparambil, Arun Suresh, Young Chen, Solom Heddaya, Roni Burd, Sarvesh Sakalanaga, Chris Douglas, Bill Ramsey, and Raghu Ramakrishnan

Microsoft

BIGDATA SCHEDULING: A JOURNEY…



Scope,
Centralized sched.
[eurosys07, vldb08]

Distributed sched.
+tooling/optimizer,
+scale, +high utilization
[osdi14]

Hadoop MR,
Centralized sched.

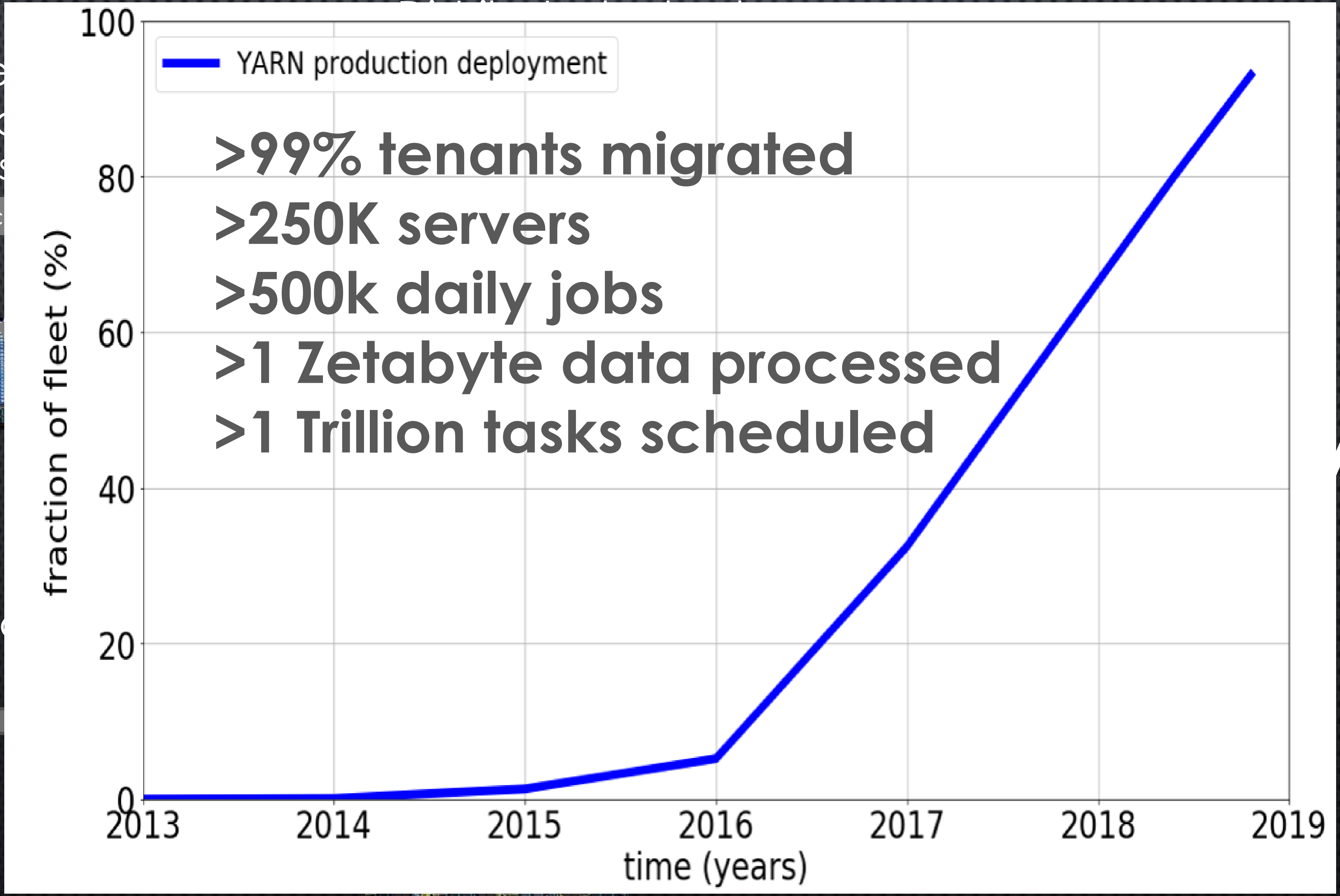
+multi-framework
+security
+scheduler expressivity
[socc13]

Hydra

BIGDATA SCHEDULING: A JOURNEY…

Scope
Centr
[eurosys

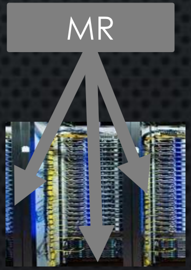
Sc



>99% tenants migrated
>250K servers
>500k daily jobs
>1 Zetabyte data processed
>1 Trillion tasks scheduled

ydra

Hadoop MR,
Centralized s



2003 2008 2013 2019

HYDRA CHALLENGES

1. Support multiple application frameworks **[socc13]**
2. Simplify writing new app frameworks **[vldb14,sigmod15,tocs17]**
3. Achieve good ROI, i.e., high CPU utilization **[atc15, eurosys16]**
4. Scale to large clusters, many jobs, large jobs **[nsdi19]**

OSS + production!

THE SCALE/UTILIZATION CHALLENGE…

Cluster(s):
> 50K nodes

Job(s):
>2M tasks, >5PB input

Scheduler:
>70K QPS

Utilization:
~60% avg CPU util

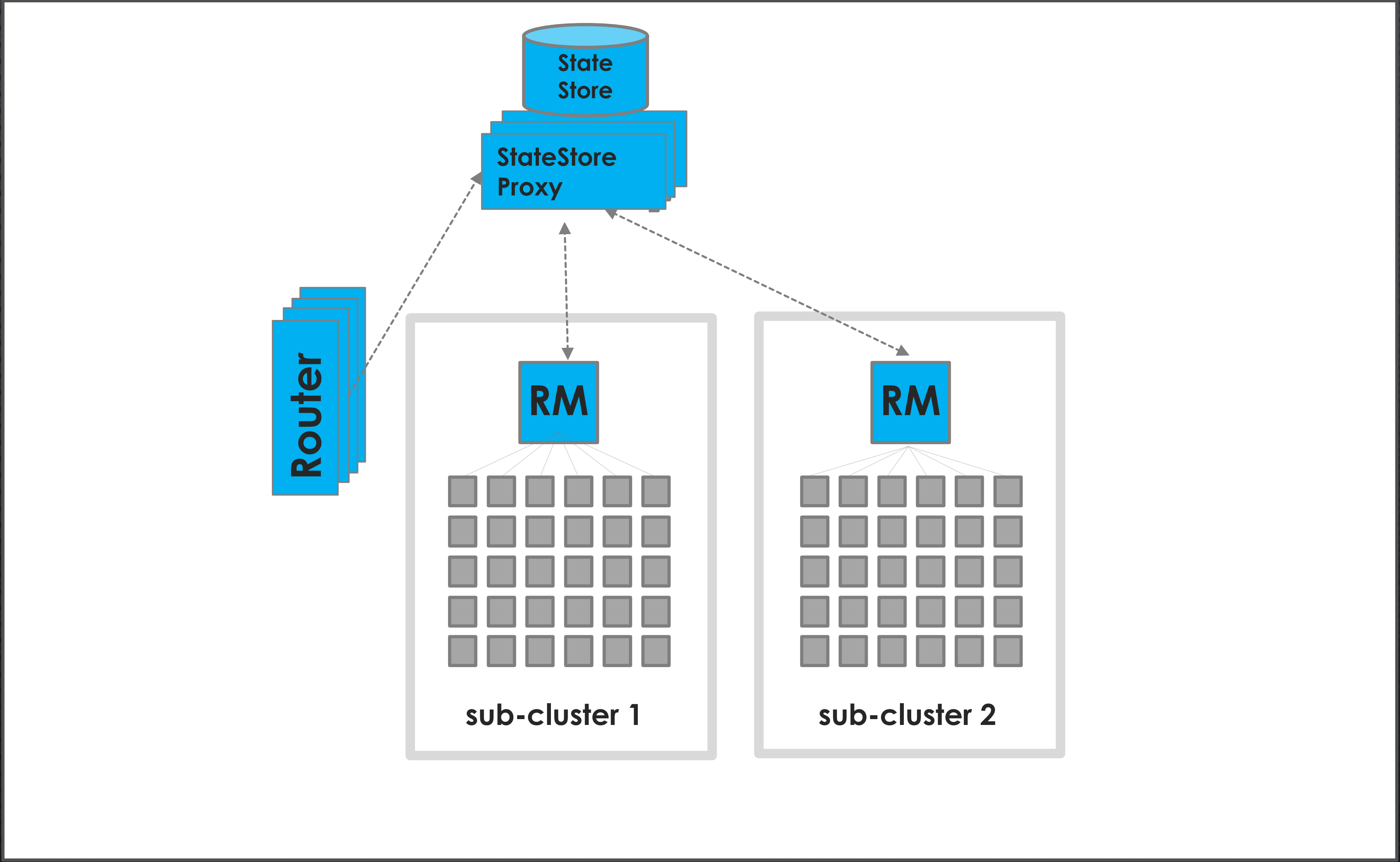
Tasks:
10 sec 50th %ile

NEED SCALE, UTILIZATION?
GO DISTRIBUTED!
NEED SCHEDULING CONTROL AND MULTI-FRAMEWORK
GO CENTRALIZED!

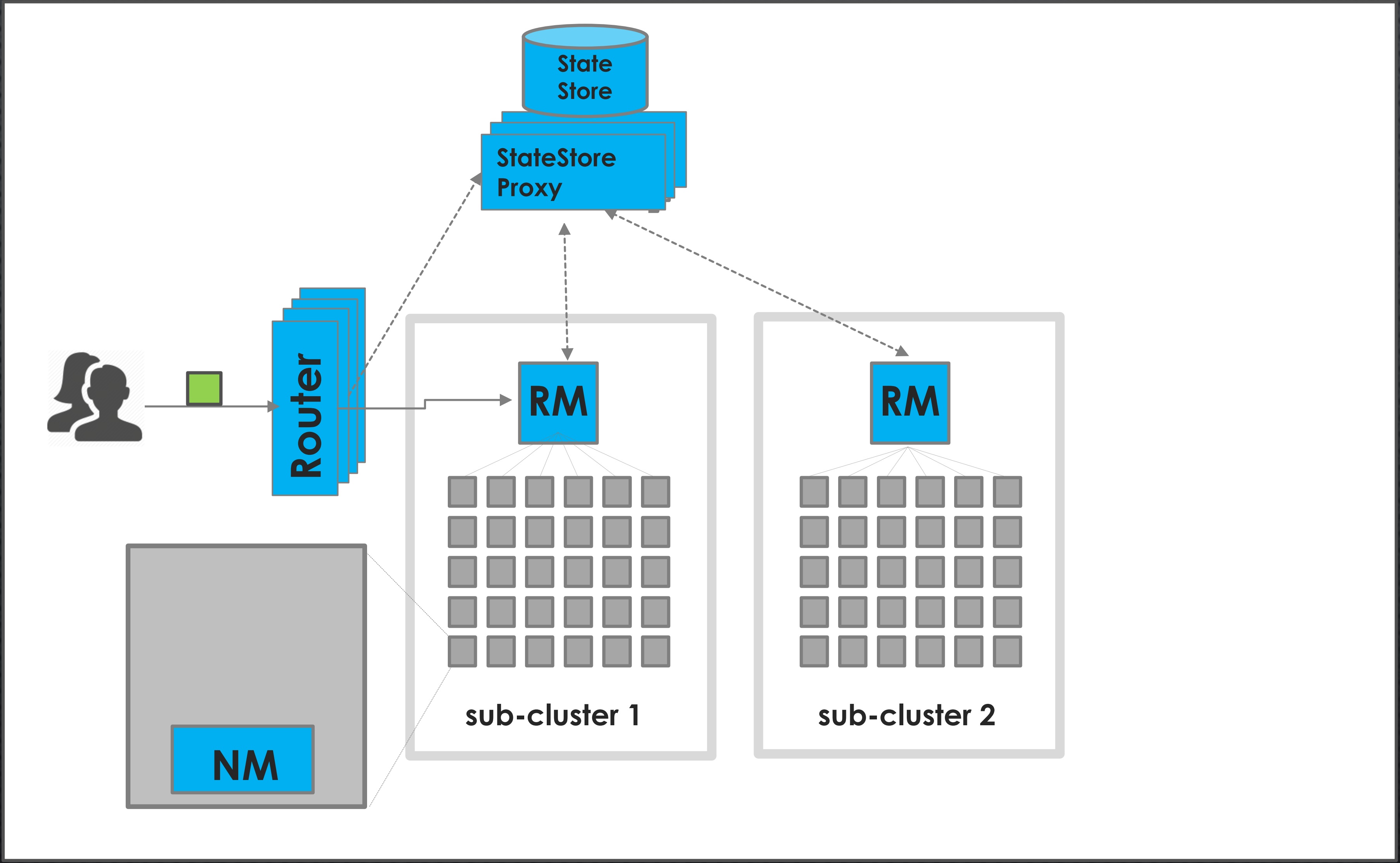
WANT IT ALL?

YOU GOTTA FEDERATE!

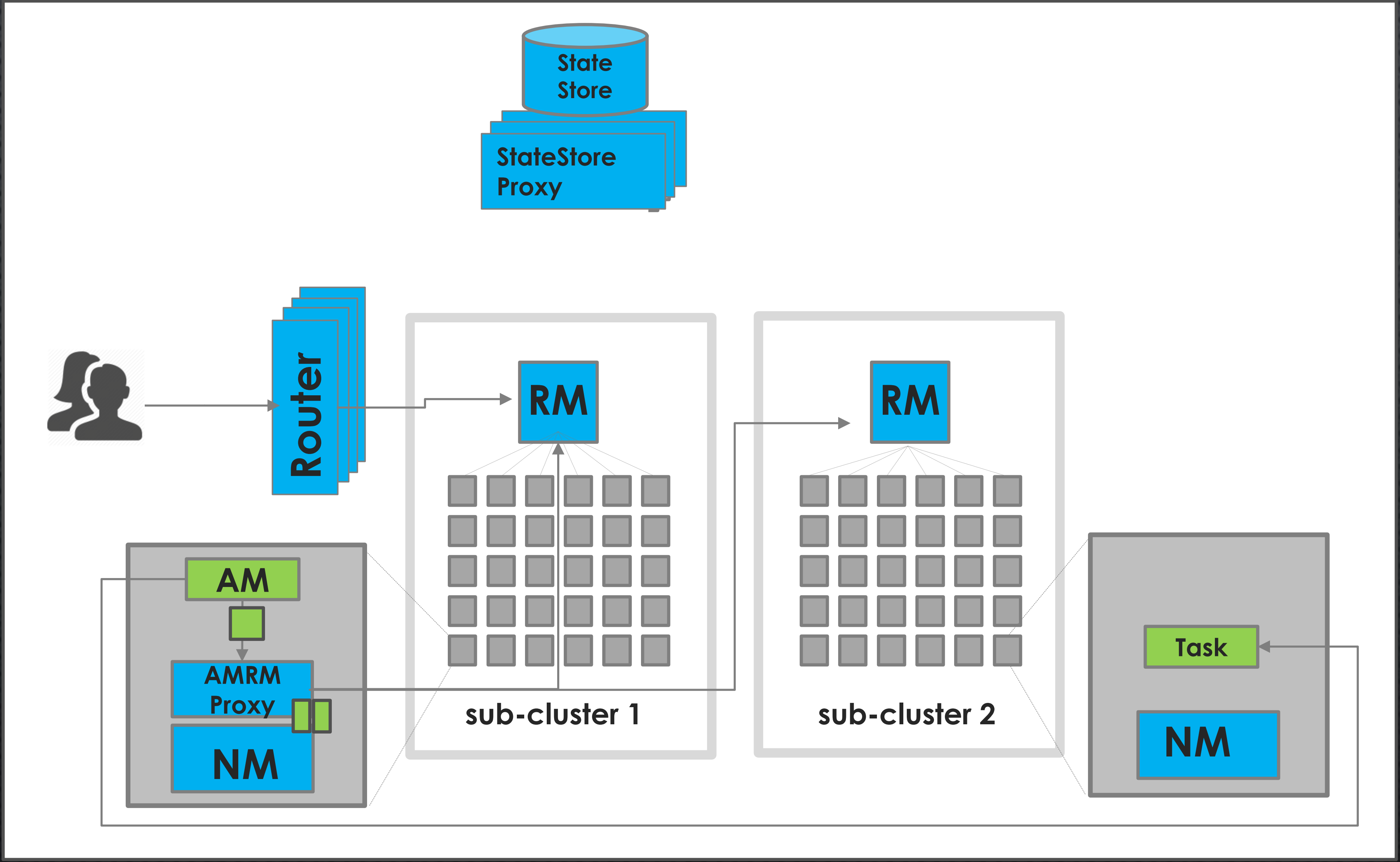
HYDRA ARCHITECTURE



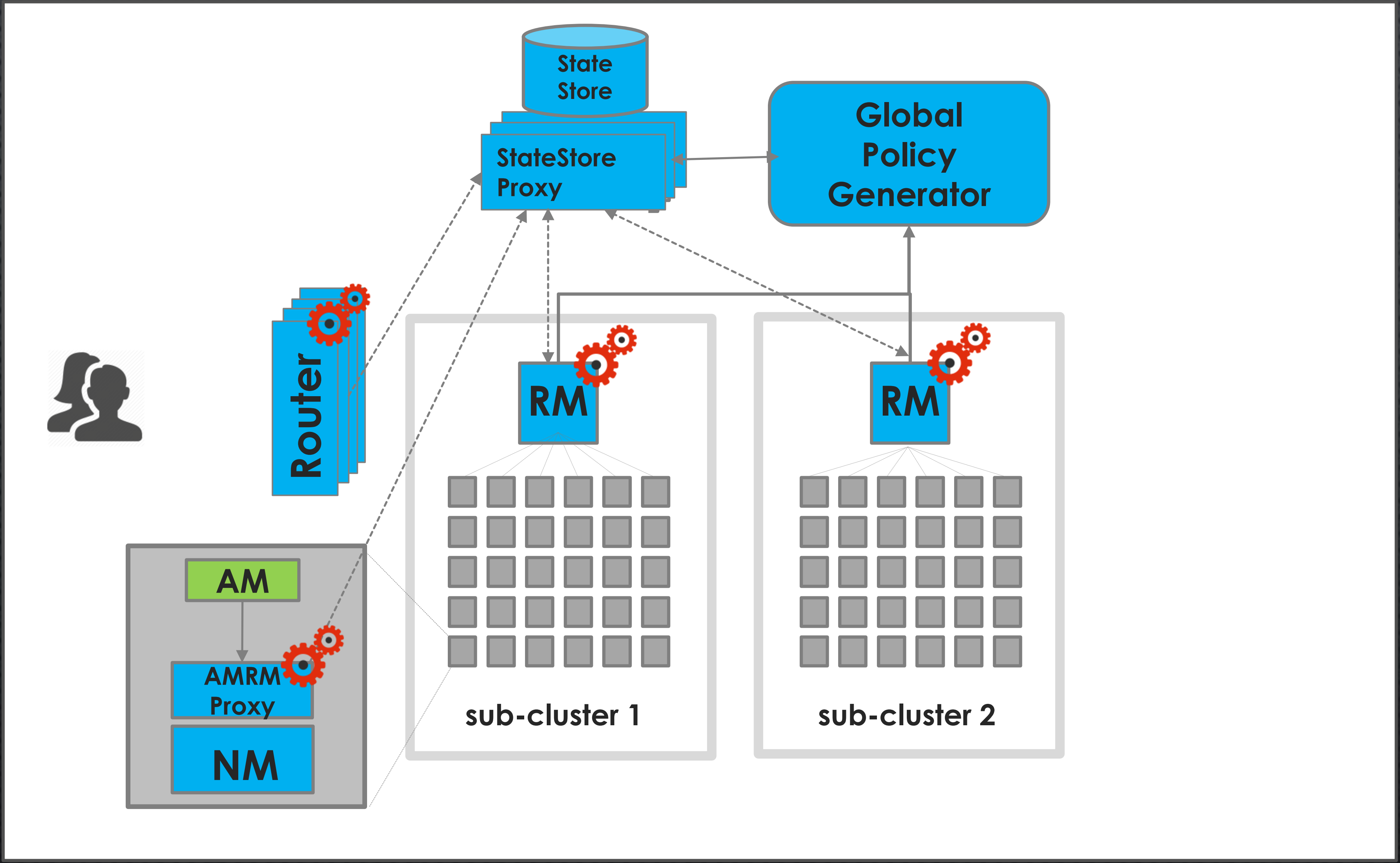
HYDRA ARCHITECTURE



HYDRA ARCHITECTURE



HYDRA ARCHITECTURE



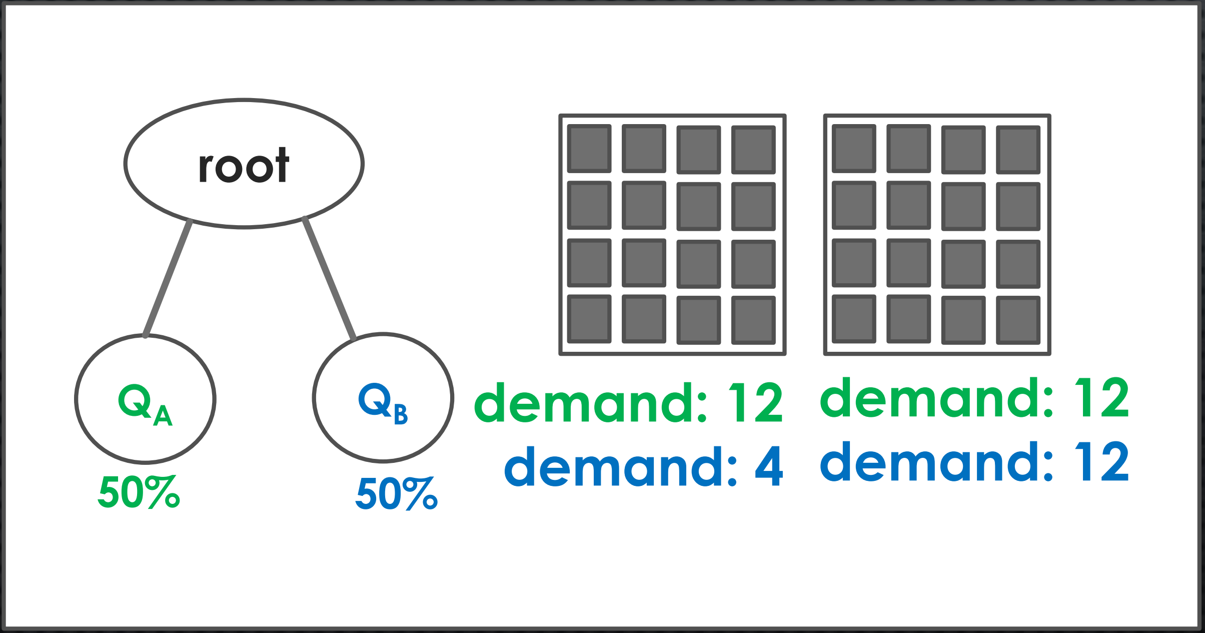
POLICIES

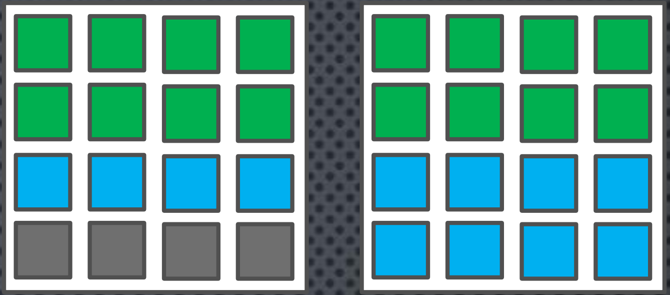
SCHEDULING DESIDERATA

- Global goals:
  - High Utilization
  - Scheduling invariants *(e.g., fairness)*
  - Locality *(e.g., machine preferences)*

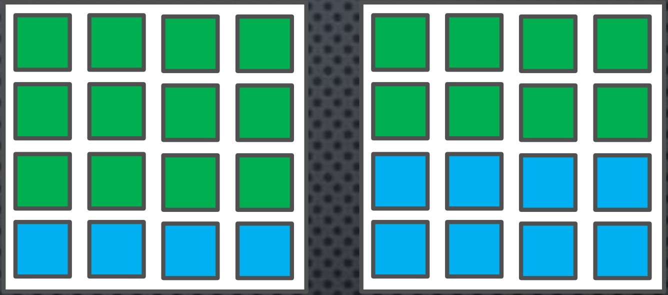
POLICIES

- AMRMProxy routing of requests
  - Enforce locality?
- Per-cluster RM scheduling decisions
  - Enforce quotas?

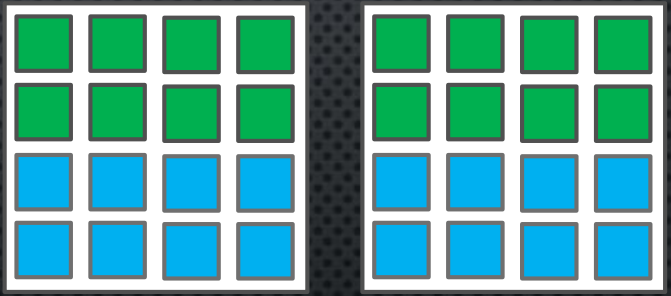




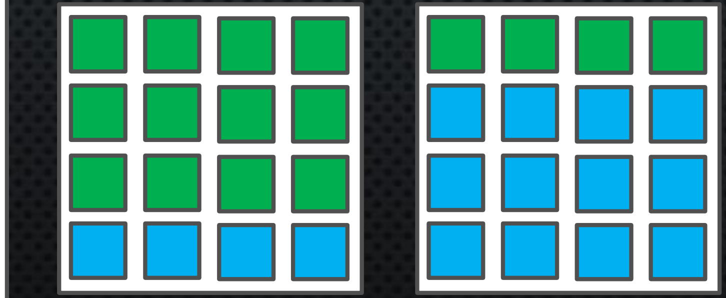
✓ Locality, Fairness
✗ Utilization



✓ Utilization, locality
✗ Fairness



✓ Utilization, Fairness
✗ Locality



✓ Utilization,
✓ Fairness,
✓ Locality.

KEY IDEA

DECOUPLE:

- Share determination
  - How many resources should a queue get?
- Placement
  - On which machine should each task run?

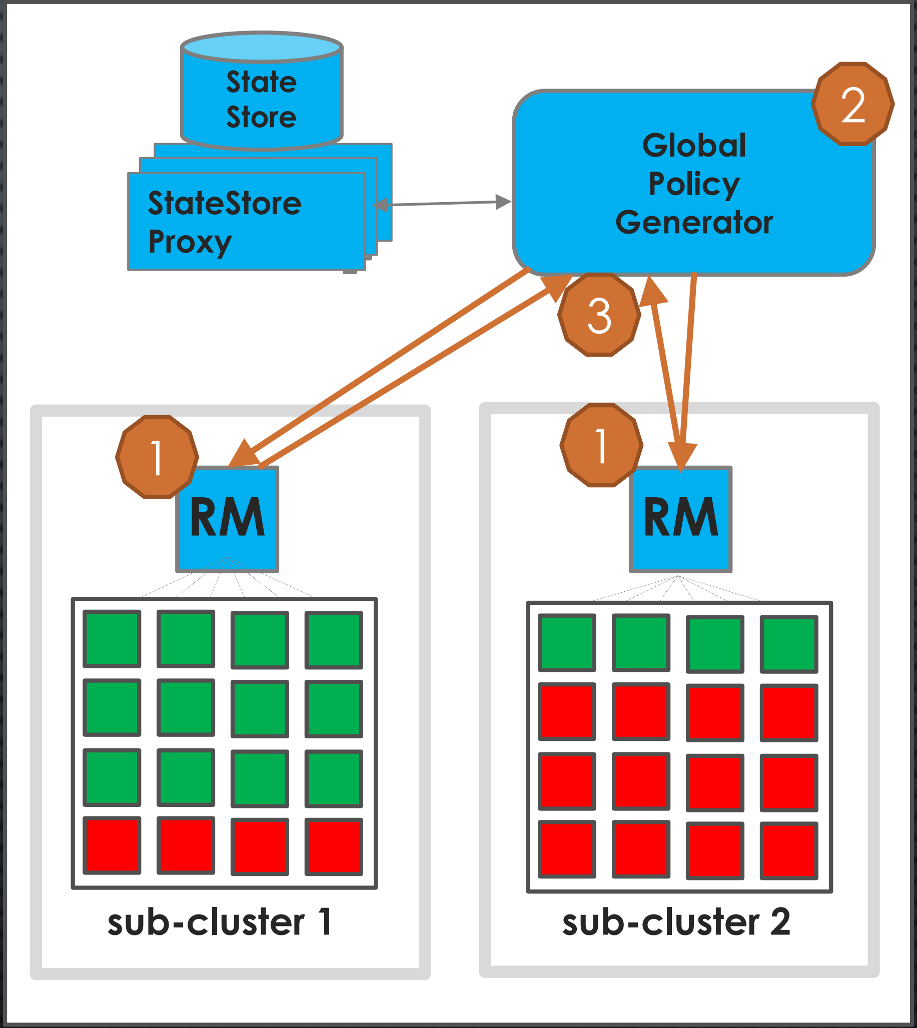
PROPOSED SOLUTION*

1 Periodically gather queue information at GPG

2 Determine resources for each queue at each sub-cluster centrally

- Logically reassign all resources, accounting for demand skew (and already assigned resources)

3 Propagate capacity decisions to each sub-cluster's RM, which perform local task allocation



* More advanced than what in prod. (details in paper)

HANDLING GPG DOWNTIME

- If GPG is down, we would fallback to local decisions
  - Problematic if they “diverge” too much from global one

- Leverage LP-based “tuning” of local queue allocation
  - Historical demand as a predictor of future demand

PRODUCTION EXPERIENCE

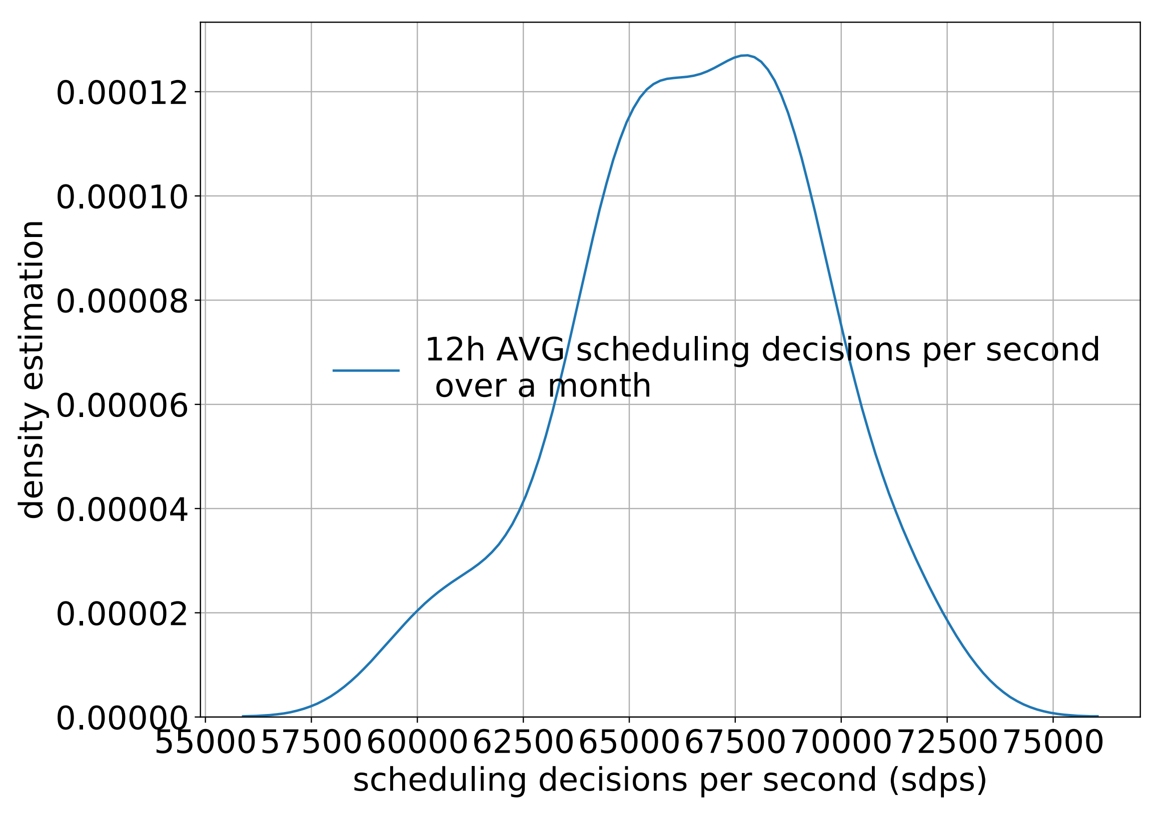
MATCHING A DISTRIBUTED SCHEDULER VIA FEDERATION

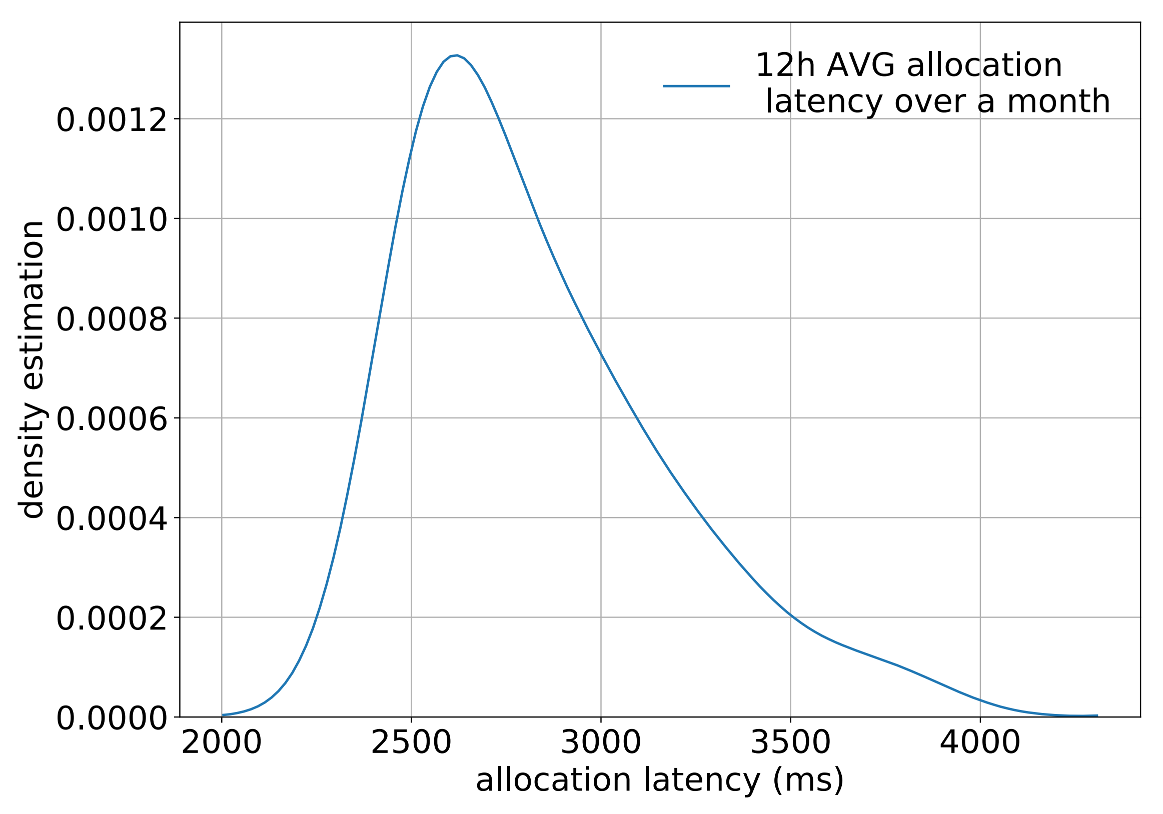
WORKLOAD

- Microsoft-wide production workload
  - 5 clusters with over >250k total servers
  - >500k daily jobs
  - ~2B daily tasks
  - Highly skewed in job and task size/duration

SCALE

HIGH SCHEDULING RATE AT LOW ALLOCATION LATENCY!

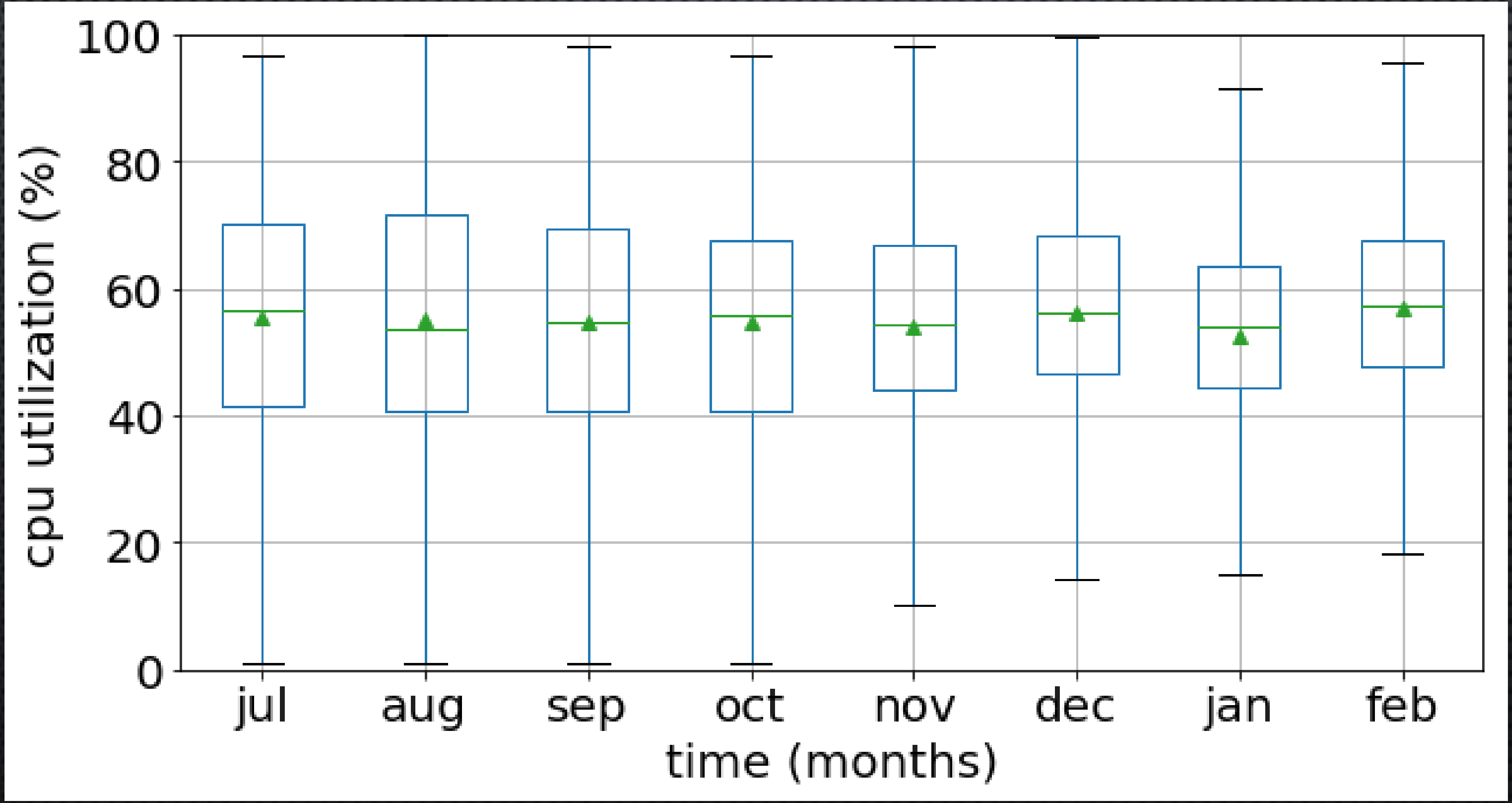




*not in the paper, updated as of jan/feb 2019

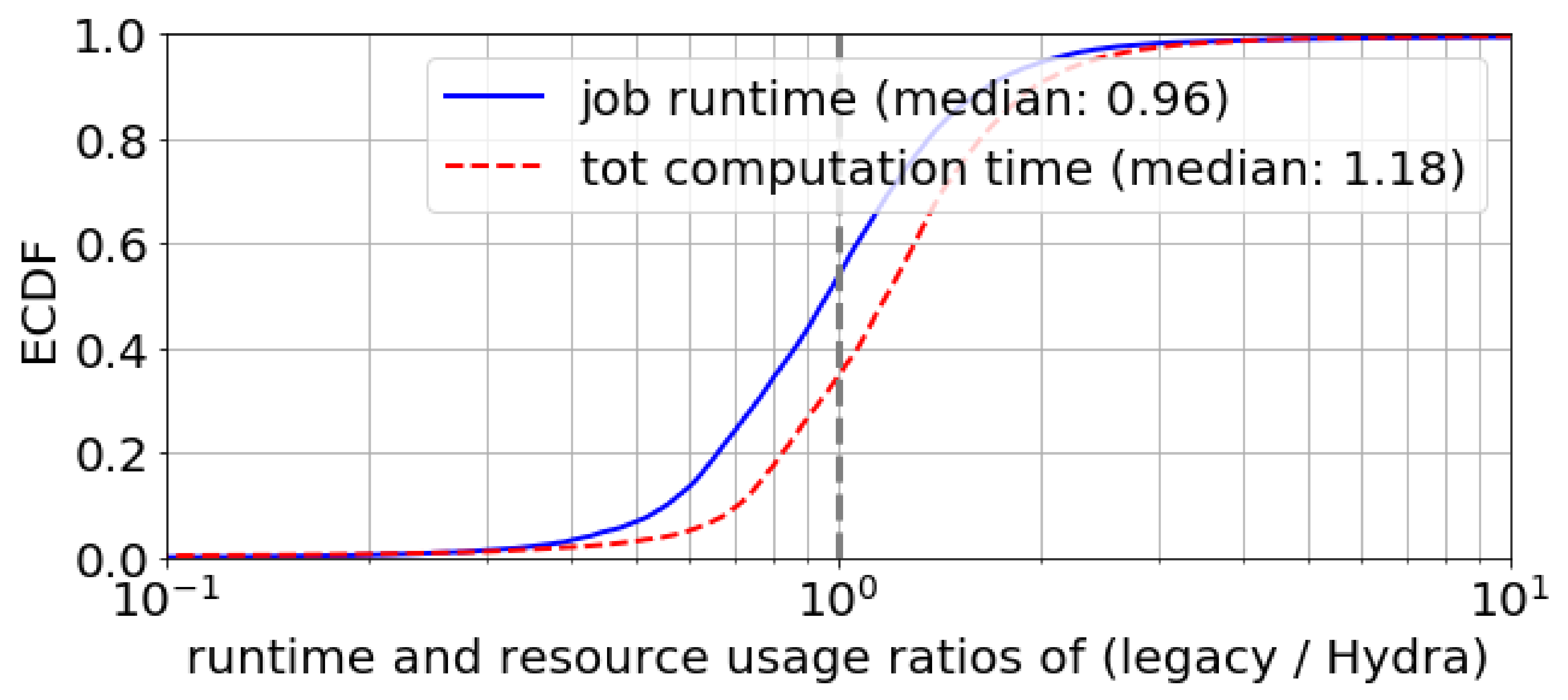
UTILIZATION

FEDERATED DESIGN IMPROVES LOAD BALANCING, WHILE RETAINING UTILIZATION!



PERFORMANCE

JOBS PERFORM JUST AS WELL (AND TASKS ARE AS EFFICIENT)!



QUALITATIVE EXPERIENCE

- In-place migration: we changed and engine mid-flight

- Happy customers can now play with OSS tech + MS stack!
- Federated design improved ***operability***:
  - Experiment at sub-cluster granularity
  - OSS innovation is easier to leverage
- Policy-driven design:
  - Allows us to dynamically adapt and experiment

CONCLUSION

- Hydra’s federated architecture got us:
  - Multi-framework / Scale / Utilization / Operability
- Exciting journey from 0 to:
  - >250k nodes
  - >200K LoC open-source code
  - 11 published papers

If you want to be part of our next project, we are hiring!