YARN Federation (YARN-2915)

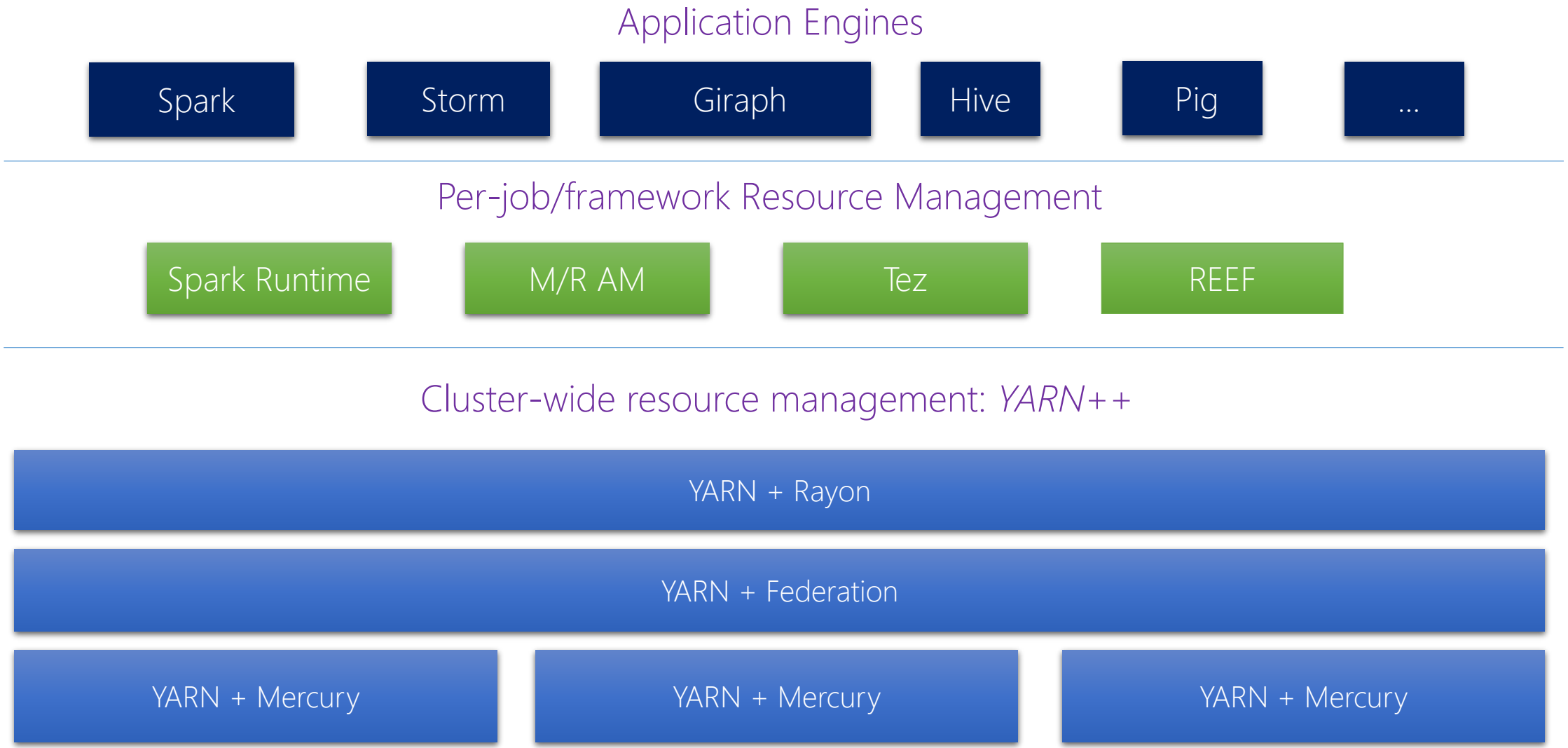
Microsoft

Motivation

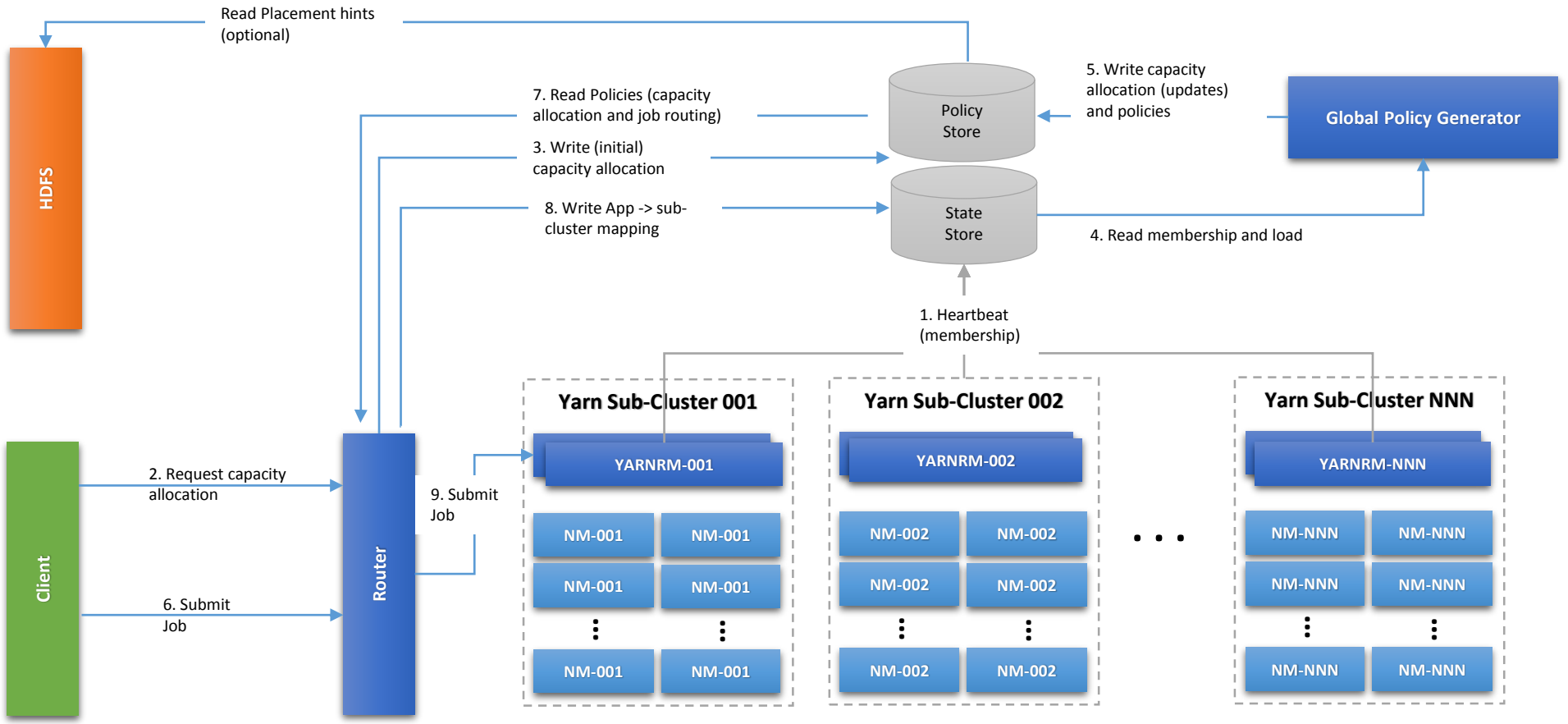
- Scale out
- Cluster pooling

YARN scale out constrained by RM scale up

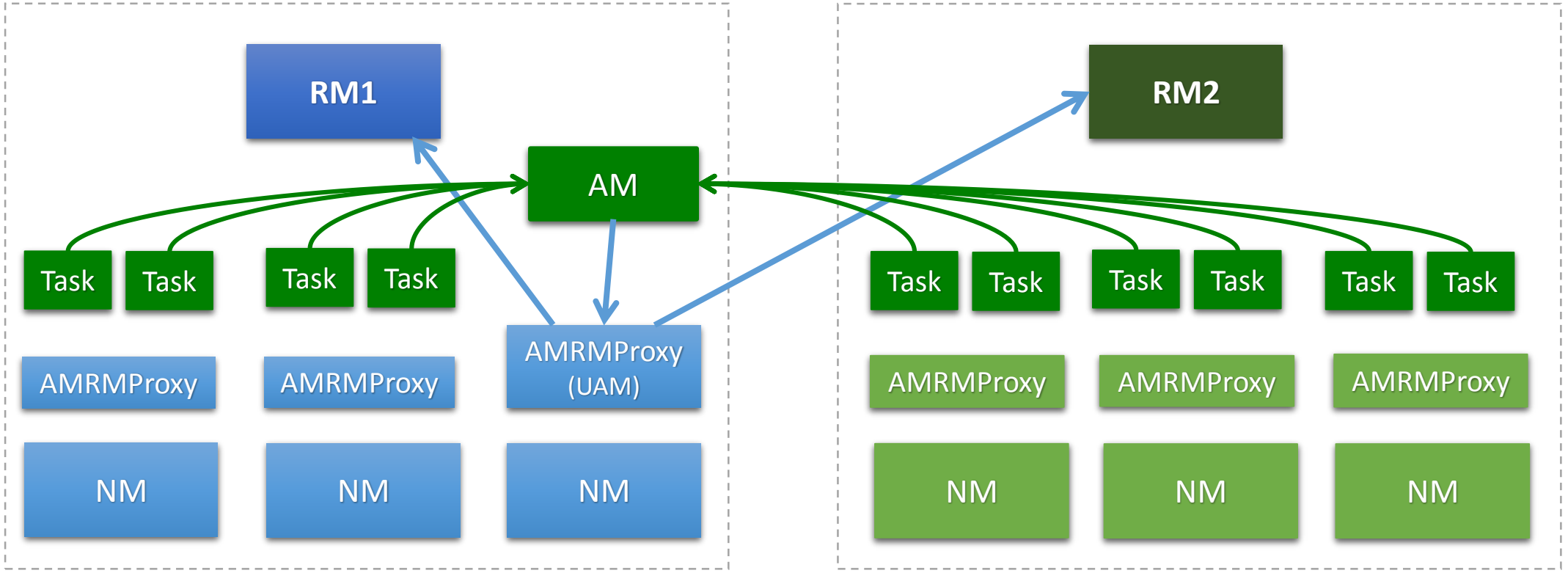
Big Picture


Application Engines
Spark
Storm
Giraph
Hive
Pig
...
Per-job/framework Resource Management
Spark Runtime
M/R AM
Tez
REEF
Cluster-wide resource management: *YARN++*
YARN + Rayon
YARN + Federation
YARN + Mercury
YARN + Mercury
YARN + Mercury

Federation Architecture



Example execution: M/R across 2 sub-clusters



YARN Federation Prototype (based on 2.6 YARN)

- RM heartbeats into StateStore
- Router send load to sub-cluster (random pick on sub-cluster enabled for user)
- AMRMProxy forwards ResourceRequests to all allowed sub-clusters

Policy Space (open problem)

Scalability

- If most apps reside within one cell → linear scale-out
- Policies to concentrate load where possible (trade offs)

Global scheduler invariants

- Fragmented enforcement (at each sub-cluster)
- Relax invariants? Eventual enforcement?