CLOUD COMPUTING APPLICATIONS

Scaling Storm to 4000 nodes with Bobby Evans, Yahoo

Roy Campbell & Reza Farivar

Open Source Big Data @Yahoo

Bobby (Robert) Evans
bobby@apache.org
@bobbydata
Architect @ Yahoo

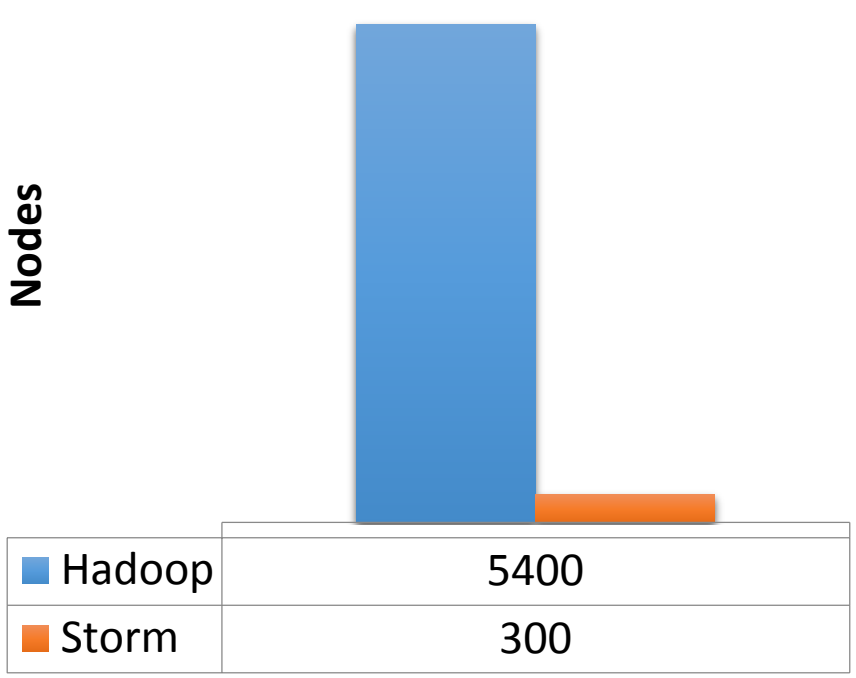
Provide a Hosted Platform for Yahoo

What We Do

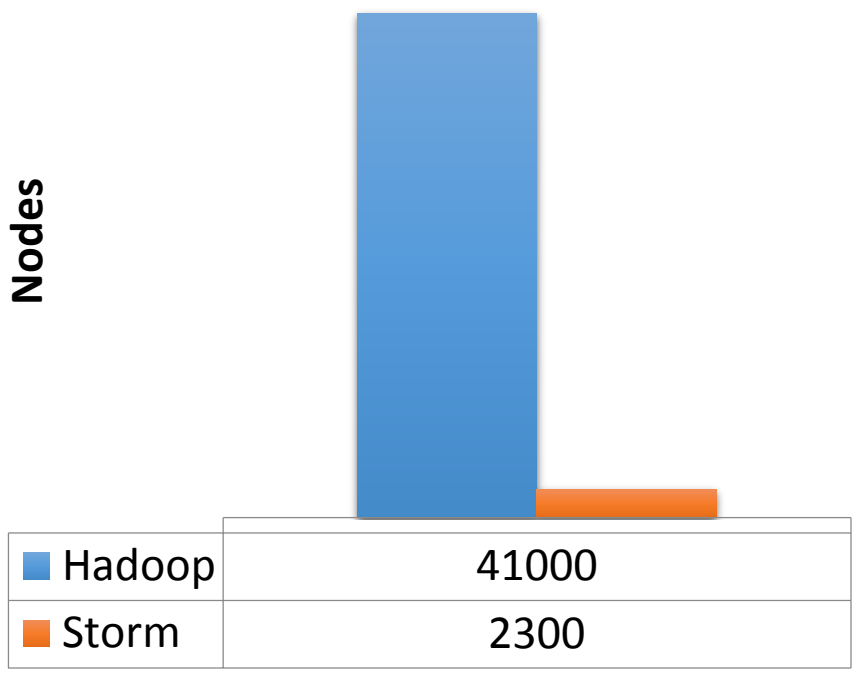
- Yahoo Scale
- Make it Secure
- Make it Easy

Yahoo Scale

Largest Cluster Size

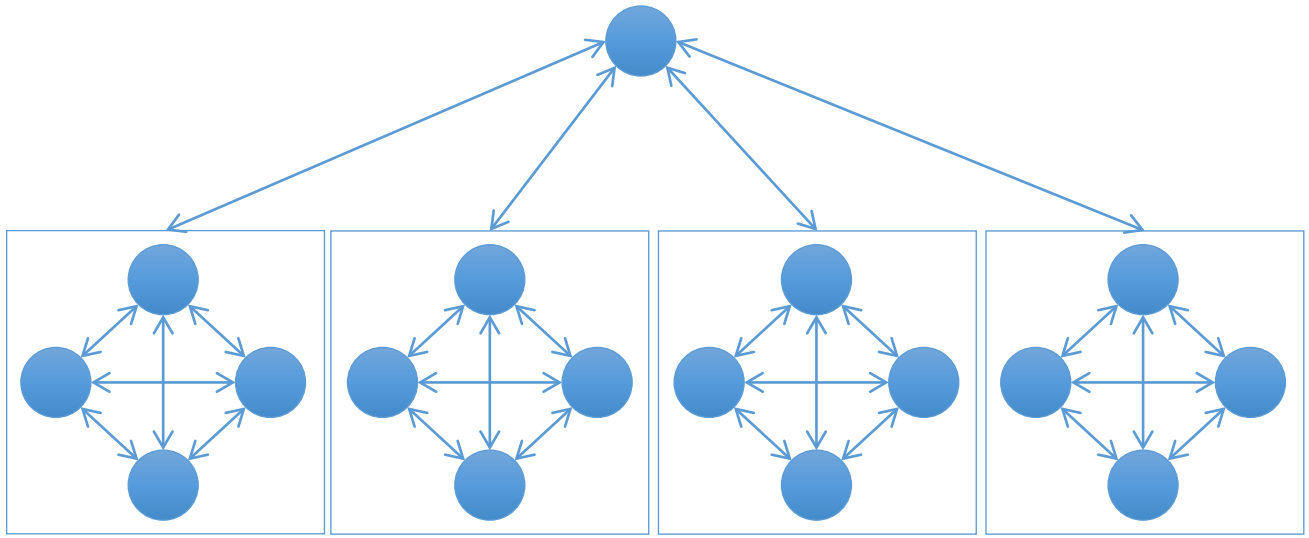


Total Nodes



Yahoo Scale (Solving Hard Problems)

Network Topology Aware Scheduling



https://en.wikipedia.org/wiki/Network_topology
https://en.wikipedia.org/wiki/Knapsack_problem

Understanding Software and Hardware

State Storage (ZooKeeper):

- Limited to disk write speed (80MB/sec typically)
- Scheduling
  - O(num_execs * resched_rate)
- Supervisor
  - O(num_supervisors * hb_rate)
- Topology Metrics (worst case)
  - O(num_execs * num_comps * num_streams * hb_rate)

On one 240-node Yahoo Storm cluster, ZK writes 16 MB/sec, about 99.2% of that is worker heartbeats

Theoretical Limit:
80 MB/sec / 16 MB/sec * 240 nodes = 1,200 nodes

Apply it to Work Around Bottlenecks

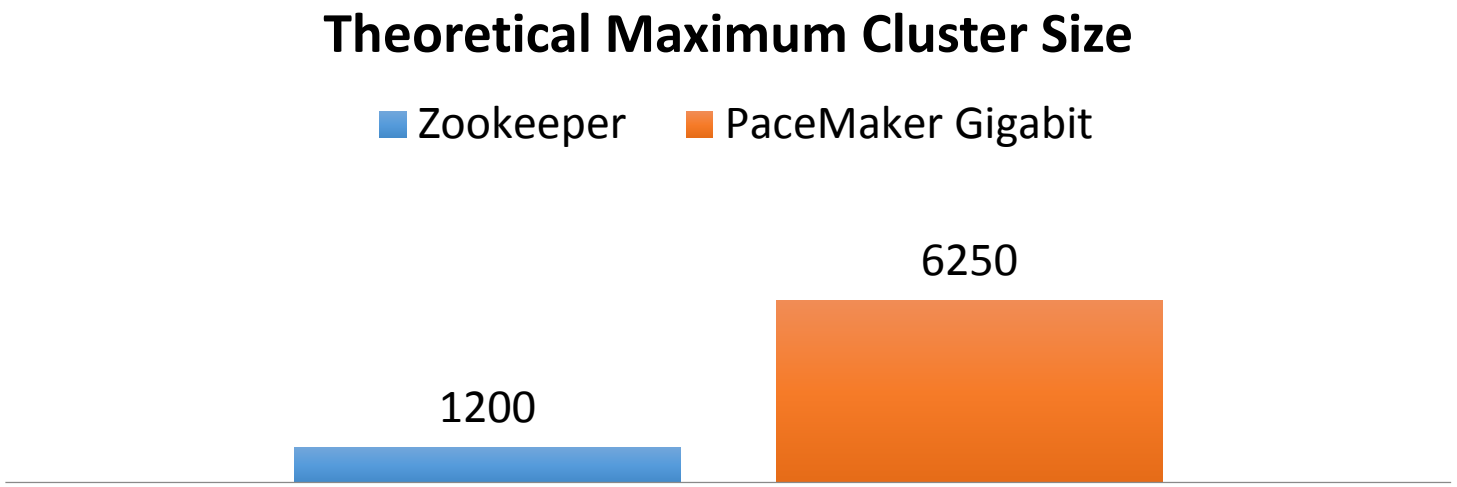
Fix: Secure In-Memory Store for Worker Heartbeats (PaceMaker)

- Removes Disk Limitation
- Writes Scale Linearly

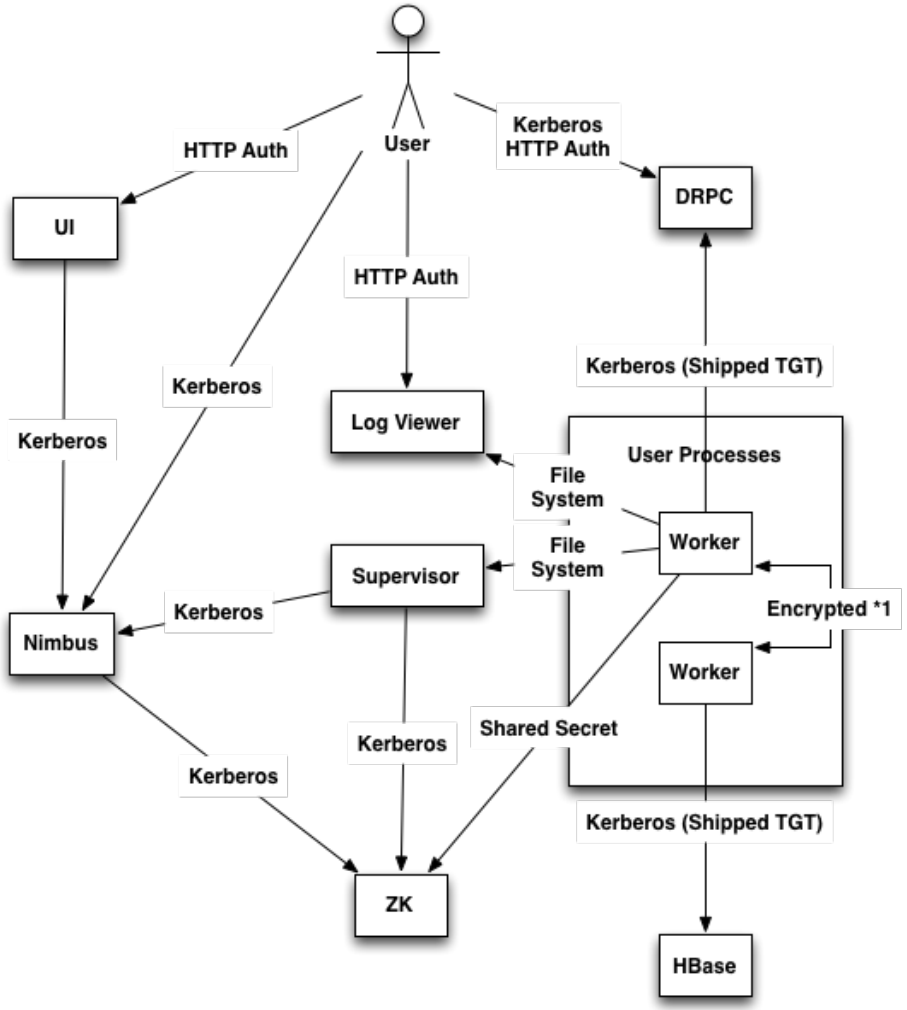
(but nimbus still needs to read it all, ideally in 10 sec or less)

240 node cluster's complete HB state is 48MB, Gigabit is about 125 MB/s

10 s / (48 MB / 125 MB/s) * 240 nodes = 6,250 nodes



Make it Secure



*1 Encrypted is not ideal, we still need to add SASL with a shared secret to netty transport

Make it Easy

- Simple API
- Easy to Debug
- Easy to Setup
- Easy to Upgrade (no downtime ideally)

Heavy lifting done by the platform