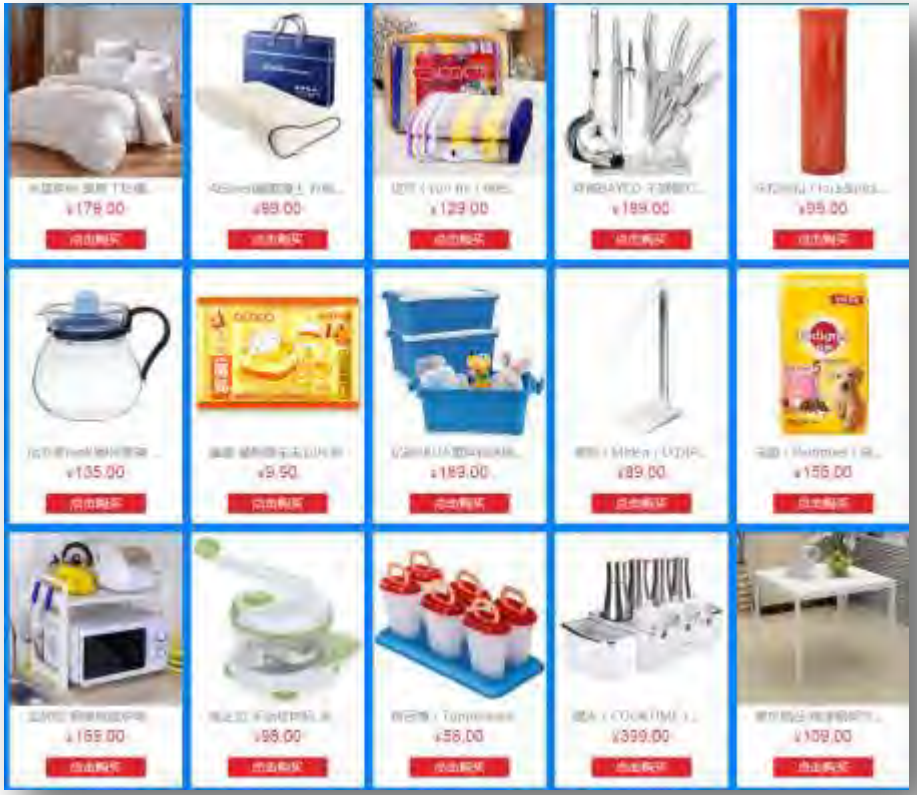
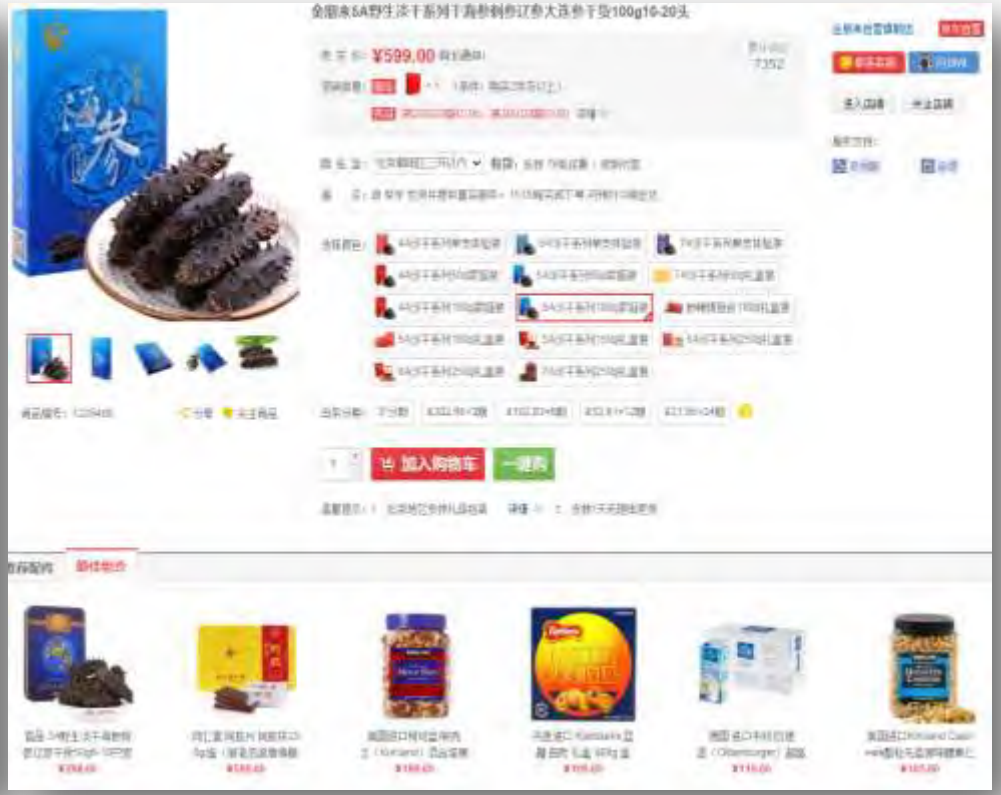
大规模内存数据库JIMDB：从2014到2016

刘海锋@京东

JD.COM 京东

E-Commerce Needs Fast Data





Memory is the new disk.

-- Jim Gray

JD.COM 京东

The Jingdong In-Memory Database

以内存为中心的数据存储

过去两年持续建设

支撑京东大多数动态内容

JD.COM 京东

演进历程

史前时代

- 分布式平台化
- 底层技术研发
- 快速规模增长
- 全自动化维护
- 正在做的事情

JD.COM 京东

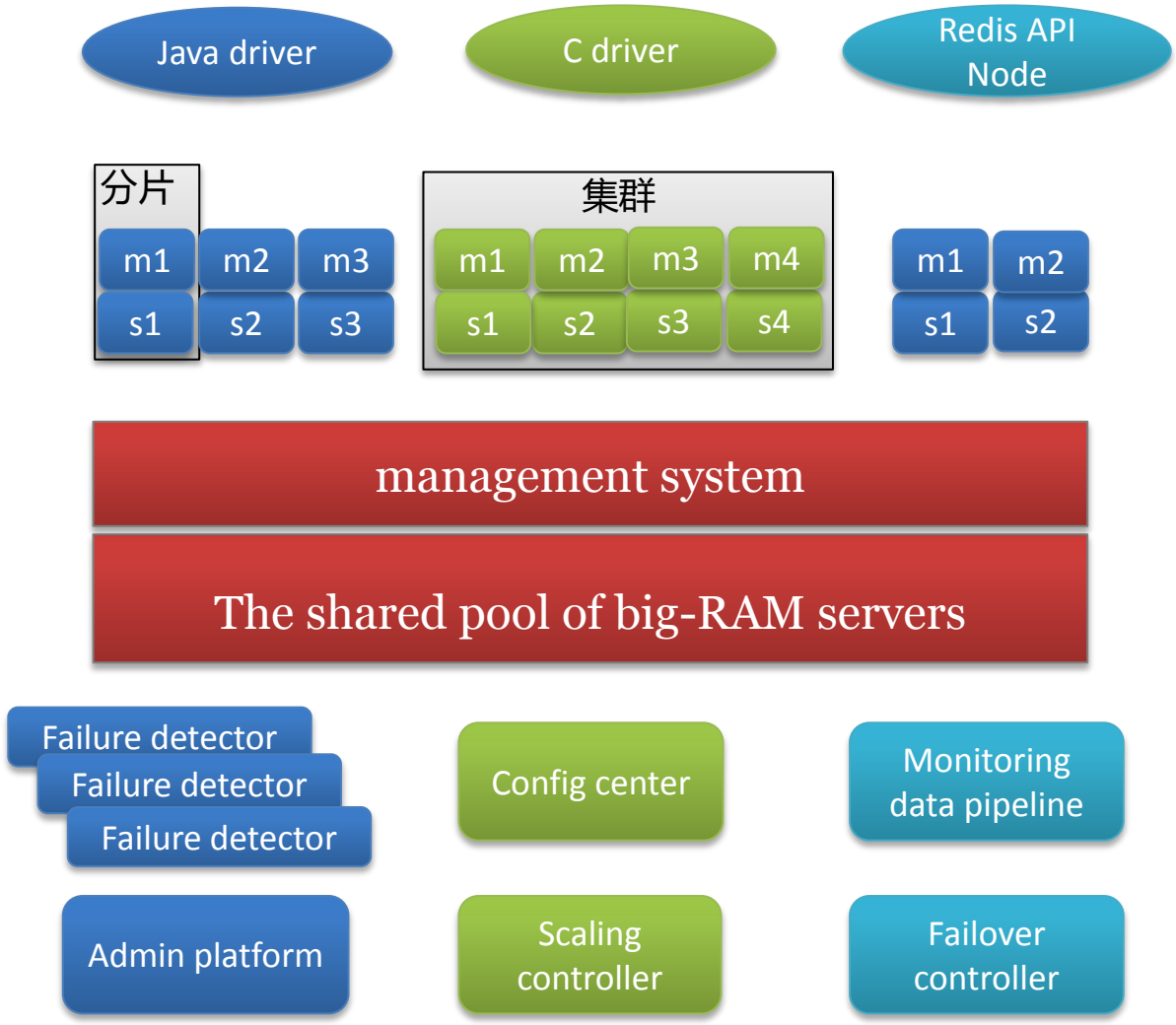
Before JIMDB

百余redis实例

一套监控系统

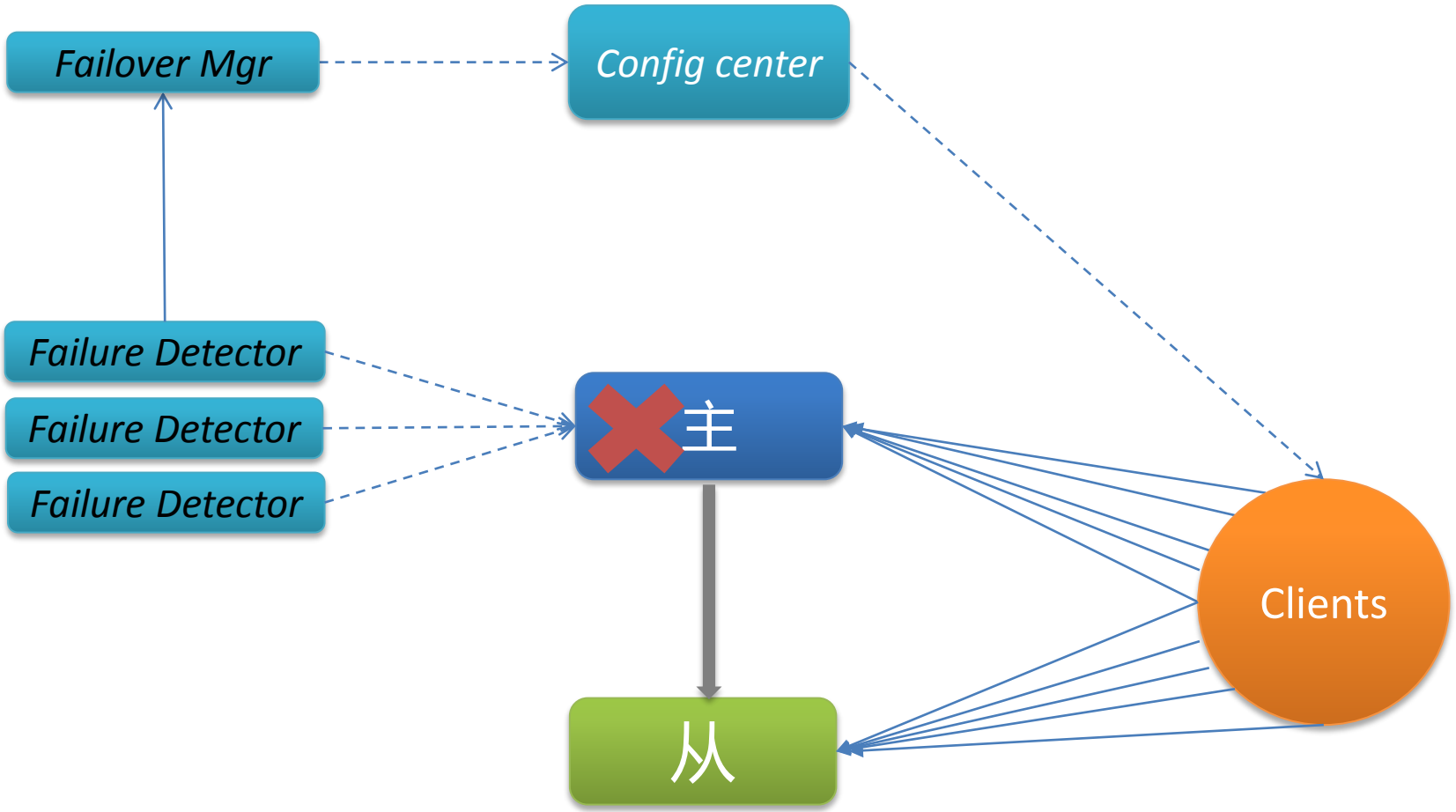
JD.COM 京东

Build a distributed system



JD.COM 京东

Auto Failover

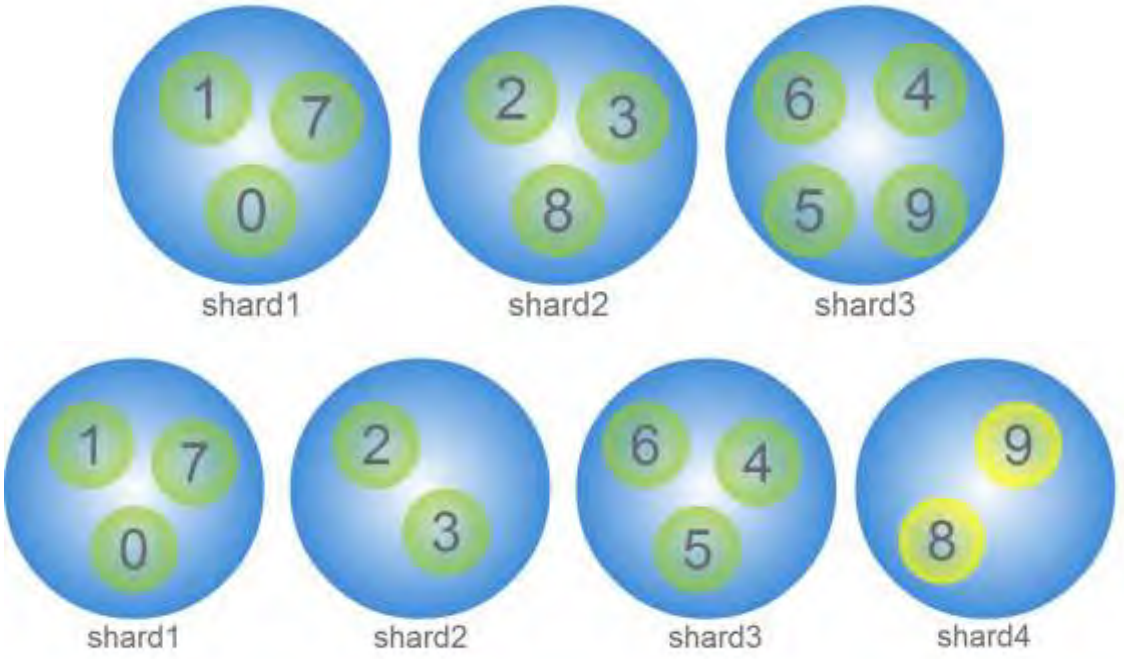


JD.COM 京东

Online Dynamic Re-Sharding

❑Partial replication

- Cluster ➔ bucket ➔ shard



JD.COM 京东

底层技术研发

存储引擎

- Dict
- LSM with RAM-SSD hybrid
- B+Tree

复制协议

- async, sync
- filtered, partial replication
- State Machine Replication

分片策略

- Hash
- Range

JD.COM 京东

分而治之

根据业务场景交付不同集群

纯缓存

不复制或异步复制

哈希分片

LRU淘汰

可靠存储

同步或SMR复制

范围或哈希分片

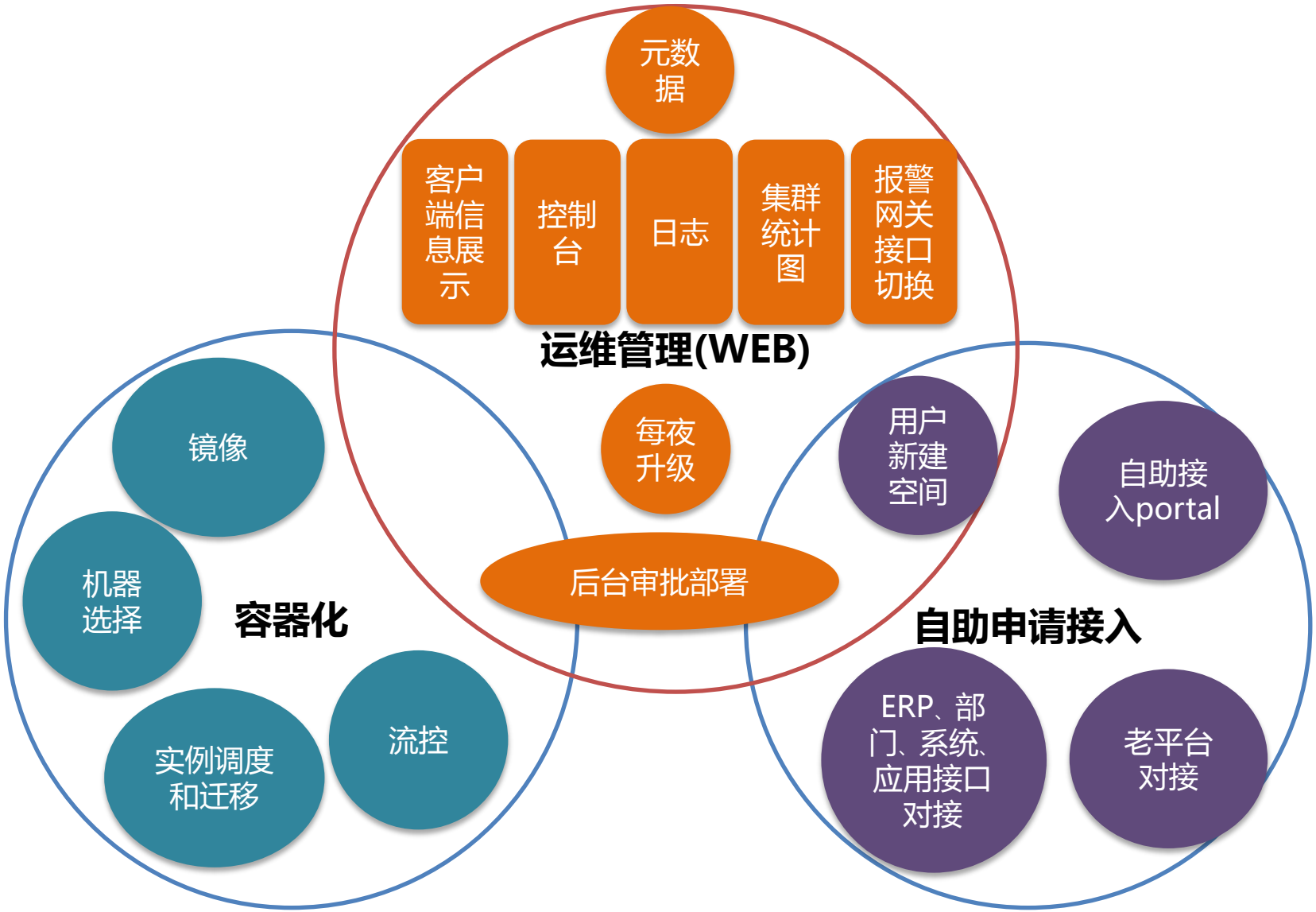
快照备份

JD.COM 京东

完善的监控体系

JD.COM 京东

基于容器的自动化运维



JD.COM 京东

目前规模

数千台大内存机器，多个数据中心

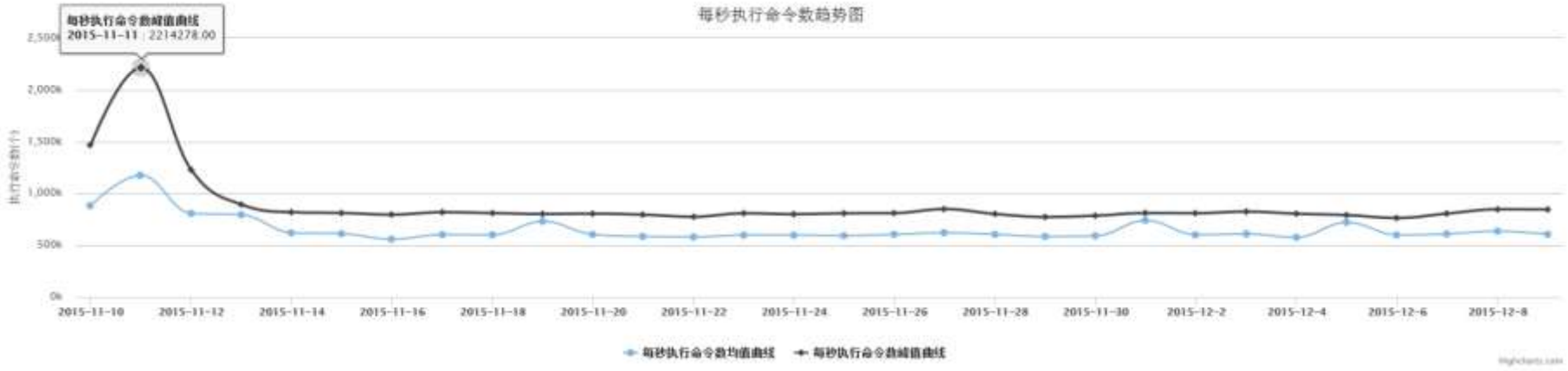
- 256GB RAM, 10Gb NIC

1000+线上集群

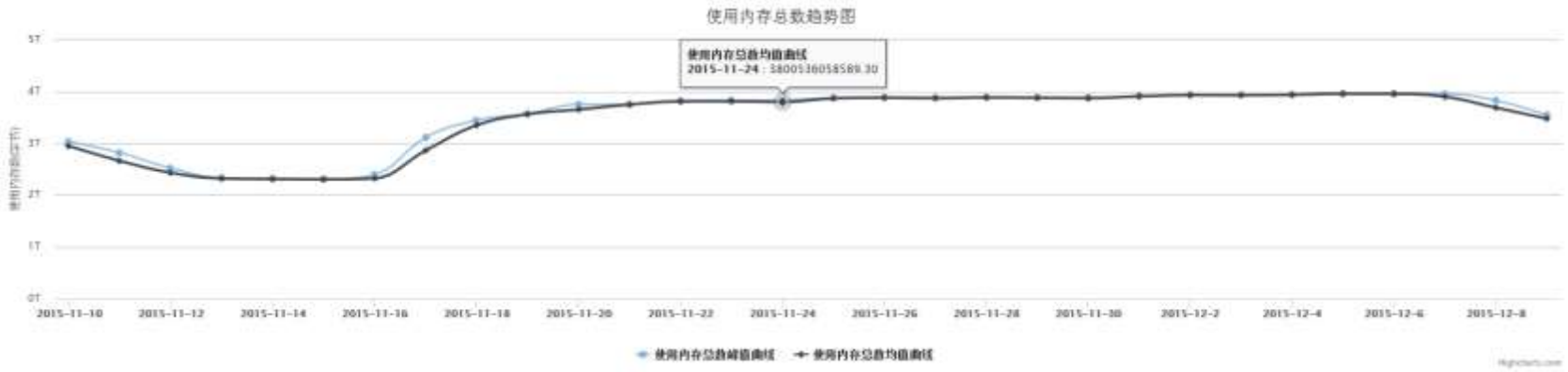
数万个Docker实例

JD.COM 京东

极佳的性能

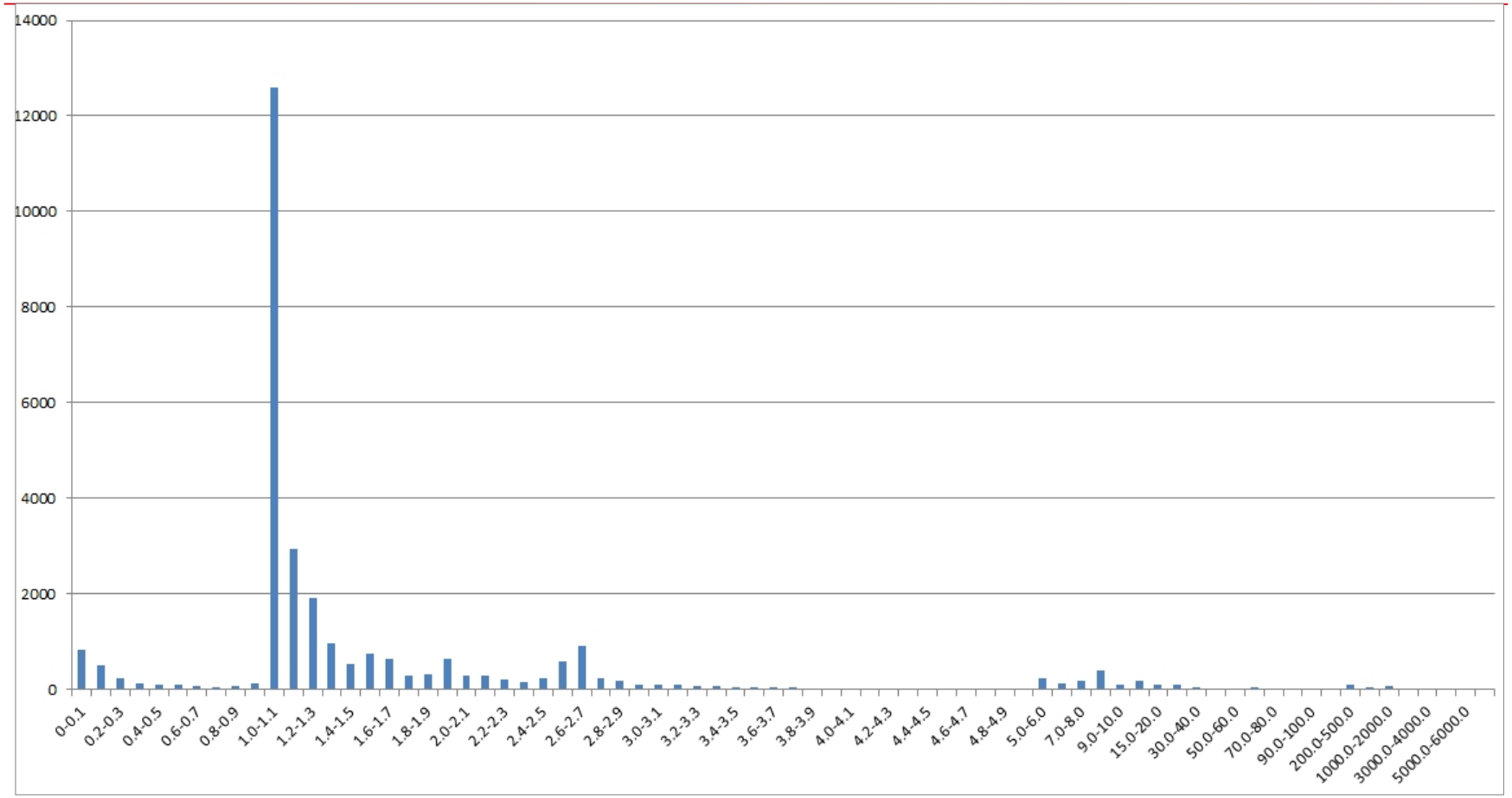


线上某集群，双11当日峰值TPS > 200W，99%延迟低于2ms



JD.COM 京东

Memory Fragment Statistics



jemalloc是目前最棒的分配器，再行开发意义不大。

JD.COM 京东

正在做的事情 – 更强的性能

- 定制网络协议栈
  - 用户态直接驱动网卡
  - 提升小包处理能力
- 更大的内存、更快的网络

正在做的事情 – 增强功能

From NoSQL to NewSQL

- A scale-out, flexibly replicated, in-memory data structure store
- Multiple SQL query processors running on it

JD.COM 京东

正在做的事情 – 复用资源

- 从专属集群到复用空闲内存
  - 弹性计算云机器
  - 非结构化存储机器
- 聚合整个IDC的RAM资源

JD.COM 京东

总结

一件事，
分步做，持续做。

内存，
是存储的未来。