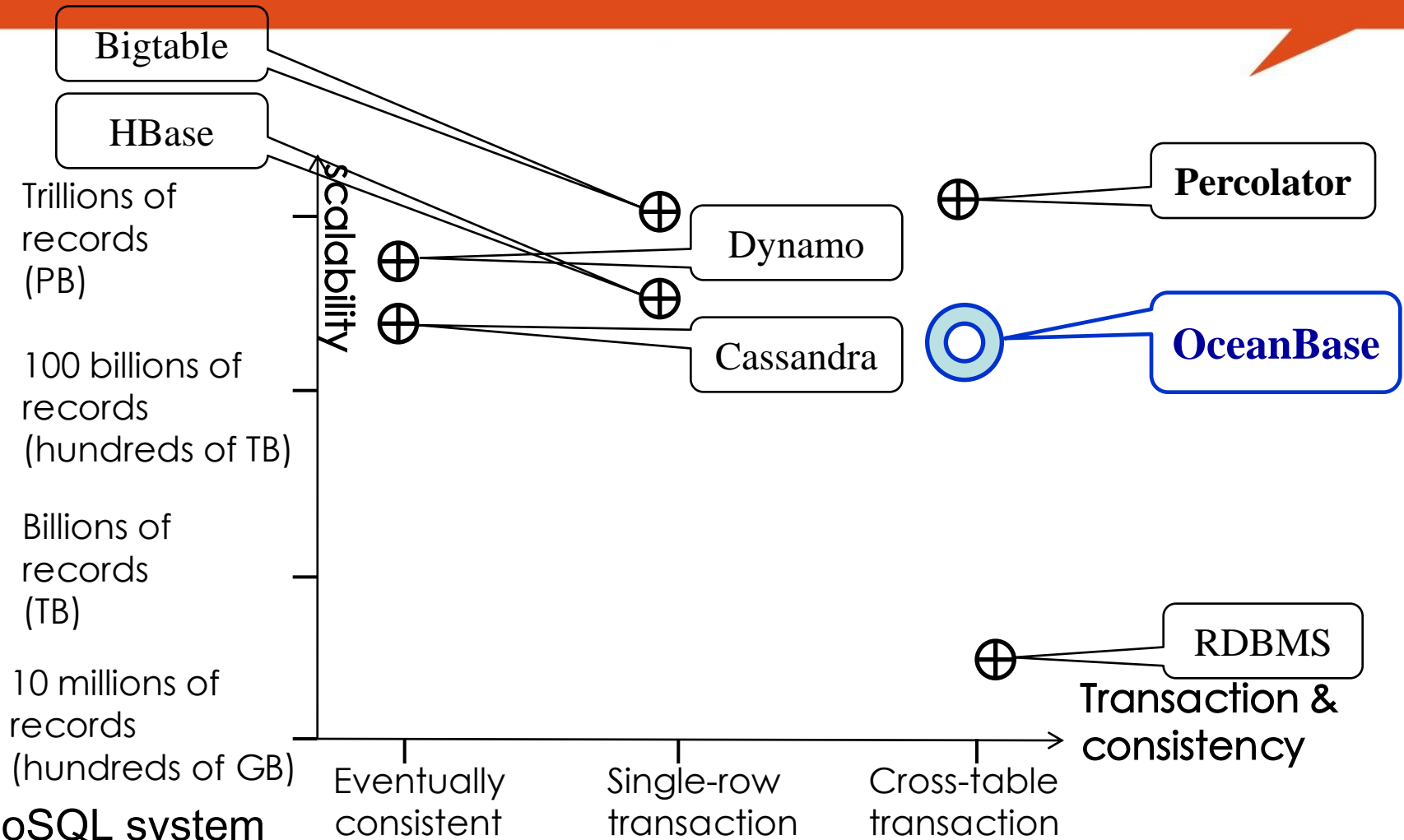


OceanBase

taobao-obdevelop@list.alibaba-
inc.com

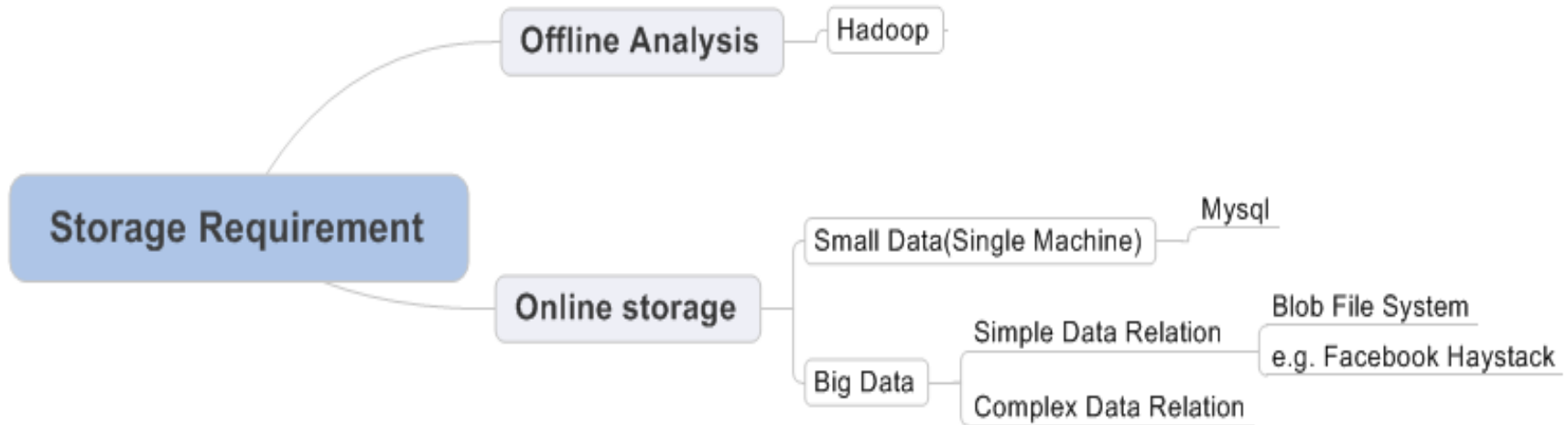
- Operating Data @Taobao, 2010
 - 370 millions registered users
 - 60 millions independent visitors per day
 - more than 2 billions page view per day
 - more than 800 millions items on line
 - more than 800 items sold per second
 - GMV: >1 billion per day
- The amount of data may increase several times, or even dozens of times in the next few years.
- RDBMS Sharding may be unpractical

Comparison of existing solutions



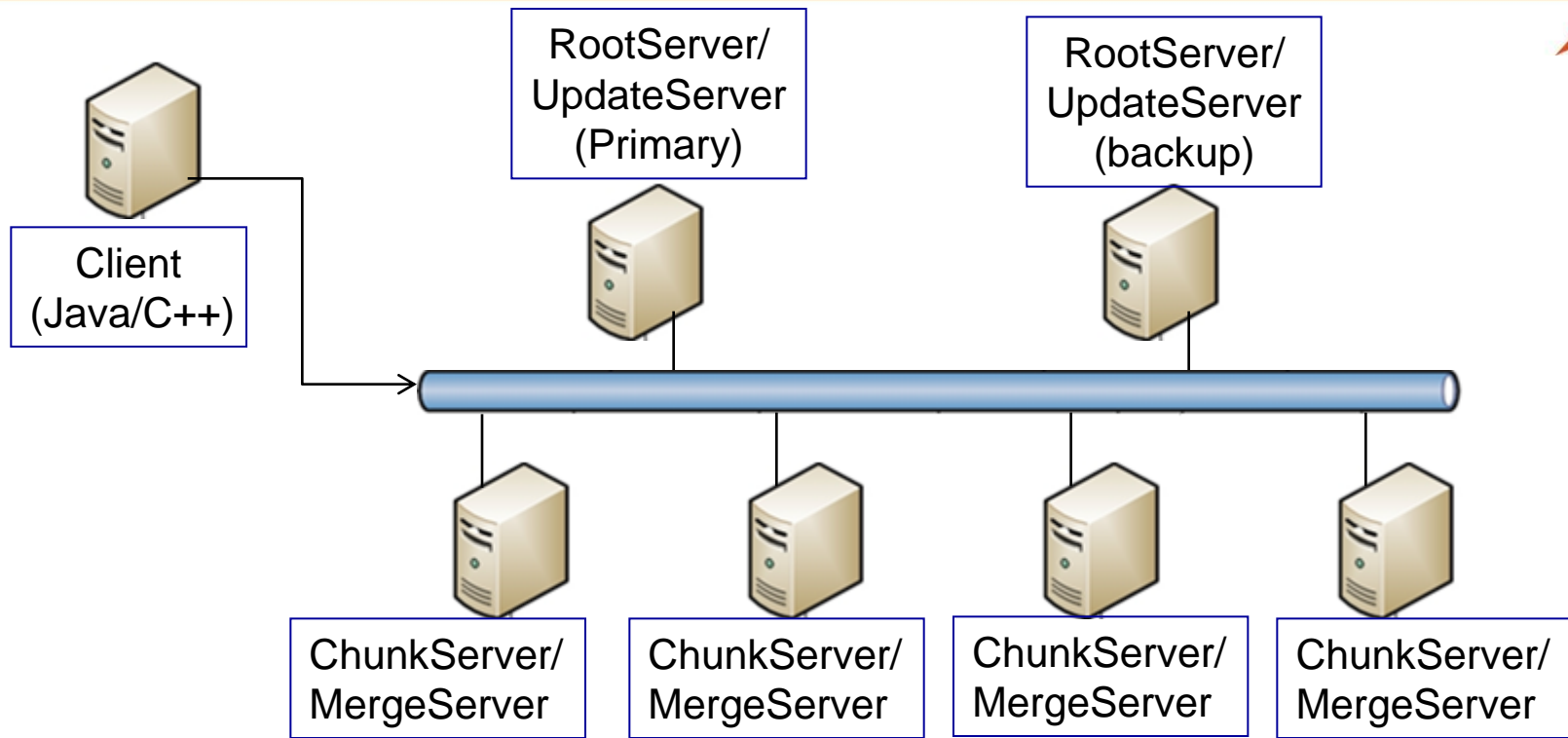
● NoSQL system

- ☑ Big data、scalability、fault-tolerant
- ☒ No cross-table transaction、weak consistency model



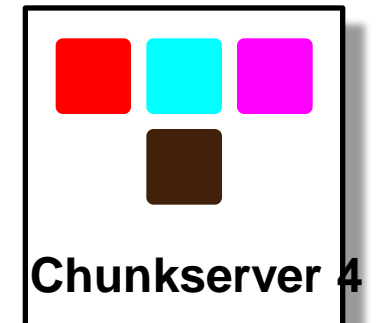
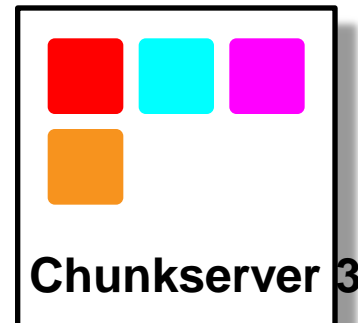
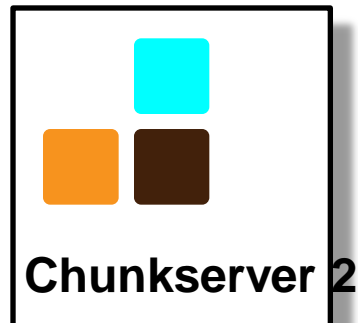
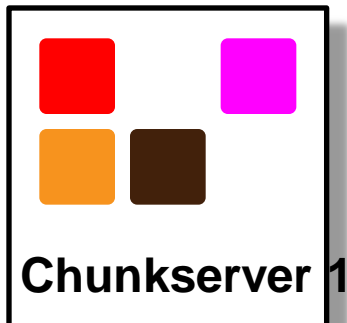
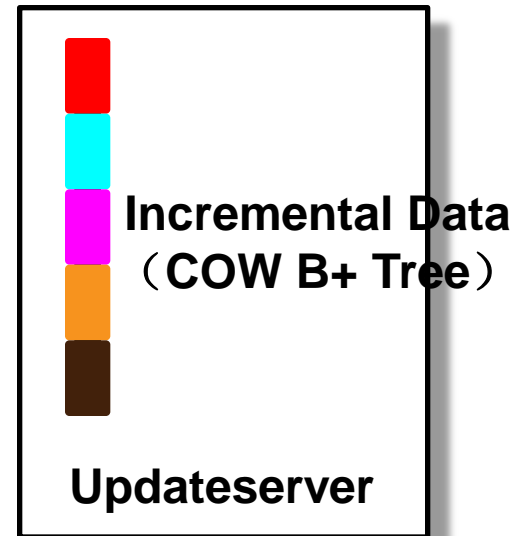
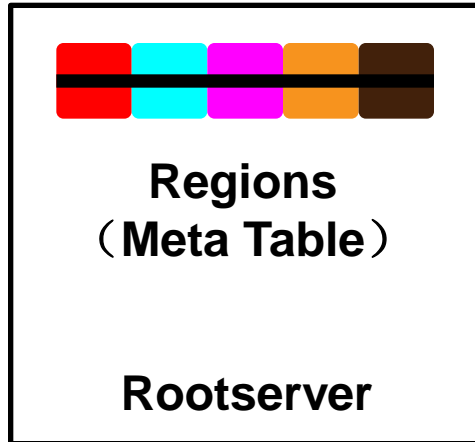
- Oceanbase: online storage system with big data and complex data relation
- Key feature of online storage: large data scale but recent write is relatively small
 - Separating incremental data with historical data
 - Historical data: large scale, distributed, SAS or SSD;
 - Incremental data: relatively small, centralized, memory or SSD;

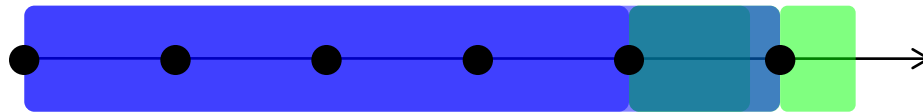
OceanBase Architecture



- RootServer(Master): Primary/backup, meta table / cluster management / schema...
- UpdateServer(Incremental data): Primary/backup, real time write (Memory+SSD)
- ChunkServer(Historical data): Distributed, static data (SAS or SSD)
- MergeServer: Distributed, Merge historical and incremental data => final result
- Daily Merge: historical data + incremental data => new historical data

OceanBase Data Distribution





**Historical data
(Chunkserver)**



**Incremental data
(Updateserver)**

- Scalability

- ChunkServer

- Automatically add and remove machine

- UpdateServer (Performance)

- Memory+SSD, multiple NICs, 10Gb NIC

- Reading from several backups

- Reliability

- ChunkServer

- Replication, default 3 copies

- UpdateServer & RootServer

- Commit log + RAID 1

- Synchronous real-time backup locally

- Almost real-time backup remotely

- Transaction

- Centralized write transaction and distributed read transaction
- Supporting cross-row and cross-table transaction

- Consistency

- Strong Consistency: Successful mutation should be applied at both primary and backup

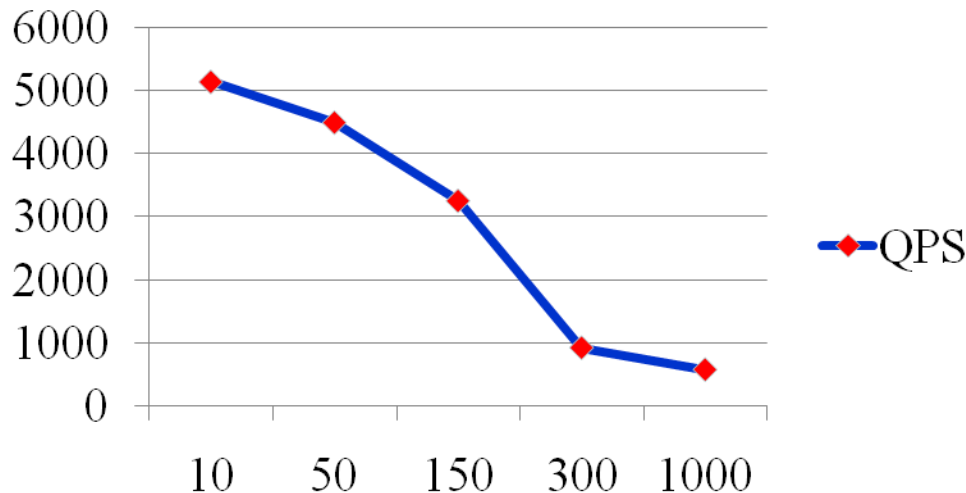
- Automatic load balance
 - Coordinated by RootServer
 - Load factor: memory and disk usage, read/write load, etc.
 - Data migration will not disturb regular read/write service
- Lock elimination
 - ChunkServer: read only, no lock needed
 - UpdateServer: Copy-on-write B+ tree, no lock for read

- Other Features

- Online schema change
- Abandoned random disk write, well suited for SSD
- Built-in data compression
- Non-stop system upgrade

● Oceanbase Performance

- 4 ChunkServer, 2 * E5520 @2.27HZ, 10 * 300GB SAS, 16GB
- 2 billions records, 3 * 160GB (compressed by LZO), 10KB block, random read, cache is closed
- Theoretical value: $4 * 180 * 10 = 7200$; OceanBase: 5100



● UpdateServer Performance

- 2 * E5520 @ 2.27HZ, 24G, 1Gb NIC

Size(byte)	20	100	1024	2048
QPS	78000	76000	70000	55000
Context Switch	26W	25W	21W	13W

➤ Optimization Point

- The first policy (Reduce malloc/free): QPS > 10W
- The second policy (Reduce context switch): QPS > 20W
- Conclusion: UPS won't be bottleneck

- Collect
 - Mysql 16*2=> Oceanbase 12+2
 - the load of machine decreased dramatically
- CTU: 2.5 billions records (2.5TB) per day
 - MongoDB => Oceanbase
 - 5 instance, 500GB per instance
- SNS feed index: Cassandra => Oceanbase
- Shop decoration system
- More than 10 billions records, several hundreds of millions of real time read and write operations
- stably running on line for more then 6 monthes

- Community
 - <http://code.taobao.org/project/view/587/>
 - <http://oceanbase.taobao.org/>
- Open source time: 2011/08/31
- Licence: GPL
- Google search: more than 90 thousands results in one month after open source
- Technical Speaking
 - Industry: Database Technology Conference China, System Architect Conference China , Taobao Open Develop Conference
 - Community: OpenSourceCamp
- Dependency: Pacemaker, Linux operating system

- Supports SQL-Like query language
- Supports distributed index
- Column-oriented storage
- Supports multi-machine parallel computing
- Supports distributed UpdateServer
- Integrates Hadoop MapReduce
- Sustained performance optimization
- Full open source
- ...