Alibaba Cloud

# Cloud Native Database Systems at Alibaba: Opportunities and Challenges


FeiFei Li

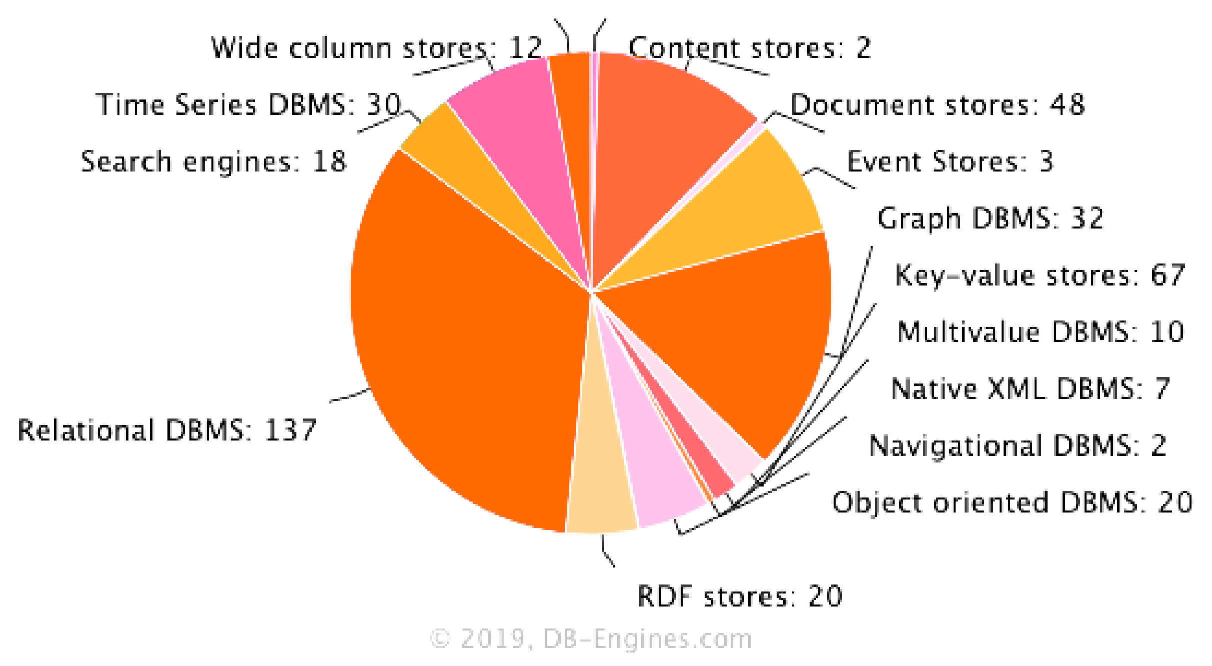
VP of Alibaba Group
President of the Database Products Business Units, Alibaba Cloud Intelligence
Director of Database and Storage Lab, DAMO Academy

# DB-Engines Database Trend (May 2019)

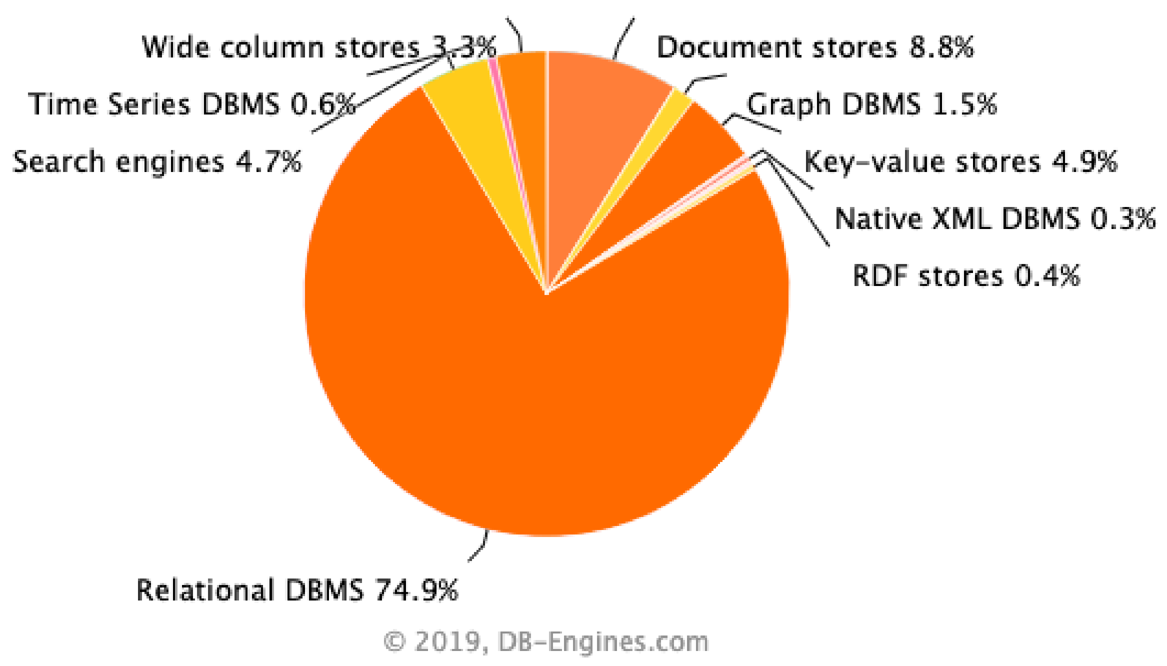
Alibaba Cloud | Worldwide Cloud Services Partner

## Relational DBMS (137) & NoSQL (210)

### Number of systems per category, May 2019

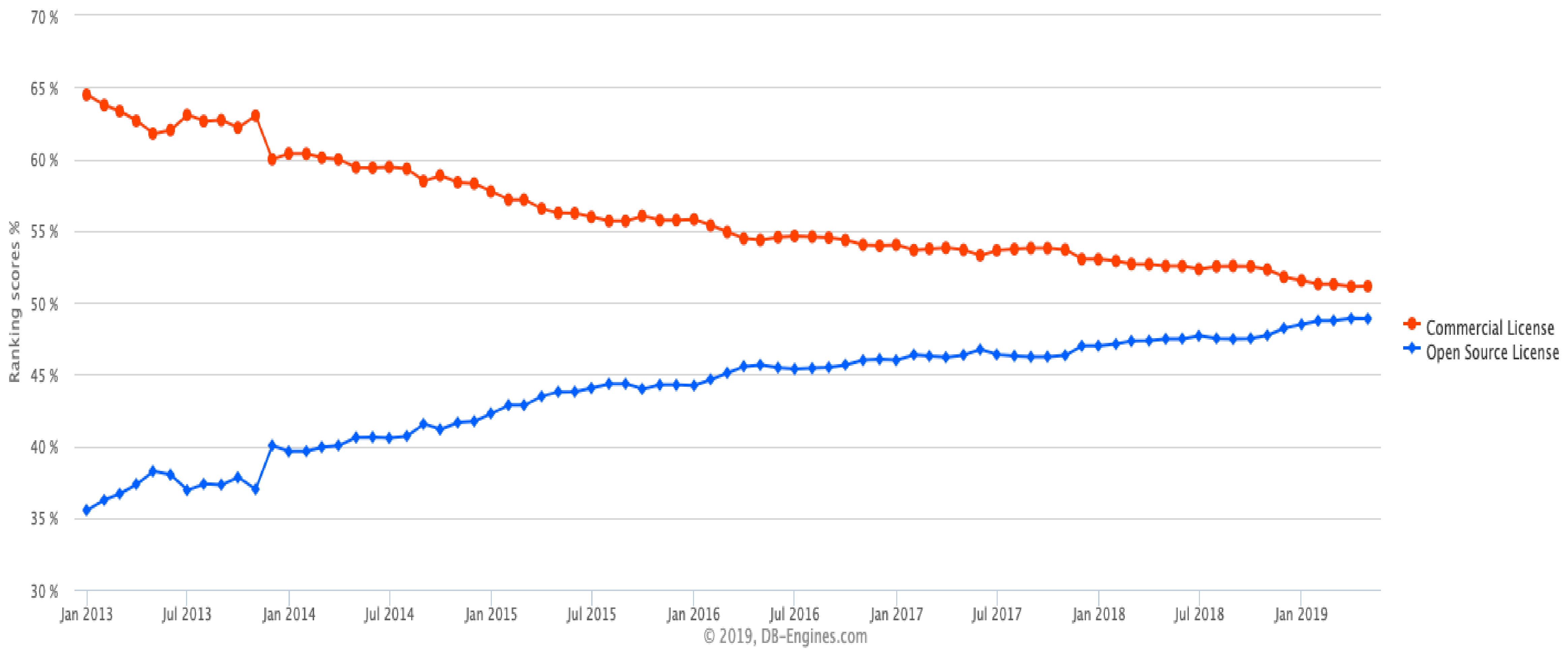



### Ranking scores per category in percent, May 2019




## Commercial License (178) vs Open Source (169)

### Popularity trend




**The top 5 commercial systems, May 2019**

| Rank | System | Score | Overall Rank |
|---|---|---|---|
| 1. | Oracle | 1286 | 1. |
| 2. | Microsoft SQL Server | 1072 | 3. |
| 3. | IBM Db2 | 174 | 6. |
| 4. | Microsoft Access | 144 | 9. |
| 5. | Splunk | 85 | 13. |

**The top 5 open source systems, May 2019**

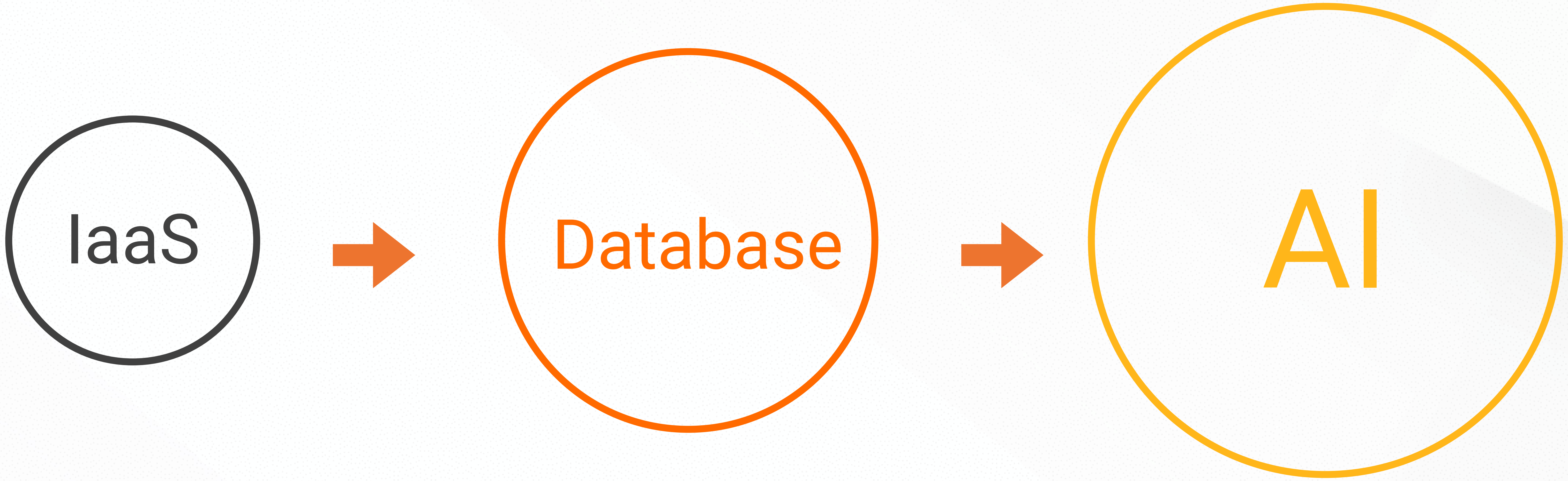
| Rank | System | Score | Overall Rank |
|---|---|---|---|
| 1. | MySQL | 1219 | 2. |
| 2. | PostgreSQL | 479 | 4. |
| 3. | MongoDB | 408 | 5. |
| 4. | Elasticsearch | 149 | 7. |
| 5. | Redis | 148 | 8. |

# Database: A key component of the cloud

Alibaba Cloud | Worldwide Cloud Services Partner

## “The real battle will be in databases”

Source:
- “How Amazon Web Services aims to win cloud computing’s next big battle” SiliconANGLE
- “AWS to Oracle: Now it's our turn and we got next”ZDNet




**Produce, storage, and processing of data**

Oracle、Google、Amazon、Apple、Microsoft、IBM、Facebook、SAP、Alibaba、Huawei、Tencent、Baidu、etc.

# Outline



- **Background**
- POLARDB
- AnalyticDB
- Self-Driving Database Platform
- Conclusion

# Alibaba Database Usage — Alibaba Group

Alibaba Cloud | Worldwide Cloud Services Partner

## Digital Media & Entertainment

YOUKU, UC, 土豆, 阿里体育, 阿里巴巴影业集团, 阿里音乐, 大麦, 阿里游戏, UC News, 微博

## Core E-Commerce

阿里巴巴 1688.com, 淘宝 Taobao, TMALL天猫, 天猫国际 TMALL GLOBAL, 聚划算 juhuasuan.com, 闲鱼, Alibaba.com, 天猫超市, AliExpress, LAZADA, 农村淘宝 cun.taobao.com, 阿里拍卖

## Local Services

高德地图 amap.com, 飞猪 Fliggy, 盒马, 淘票票, 口碑 Koubei, 饿了么

| | |
|---|---|
| 蚂蚁金服 ANT FINANCIAL | Financial Service |
| CAI NIAO 菜鸟 | Logistics |
| alimama | Marketing |
| Alibaba Cloud | Cloud Computing |

**870 Million**

**Active Users of Alipay Around the Globe**

Serving 15 million SMEs worldwide

**100 Million**

**Packages Processed by Cai Niao Logistic Network Each Day**

Covering 15 hundred cities and counties in mainland China

**$768 Billion**

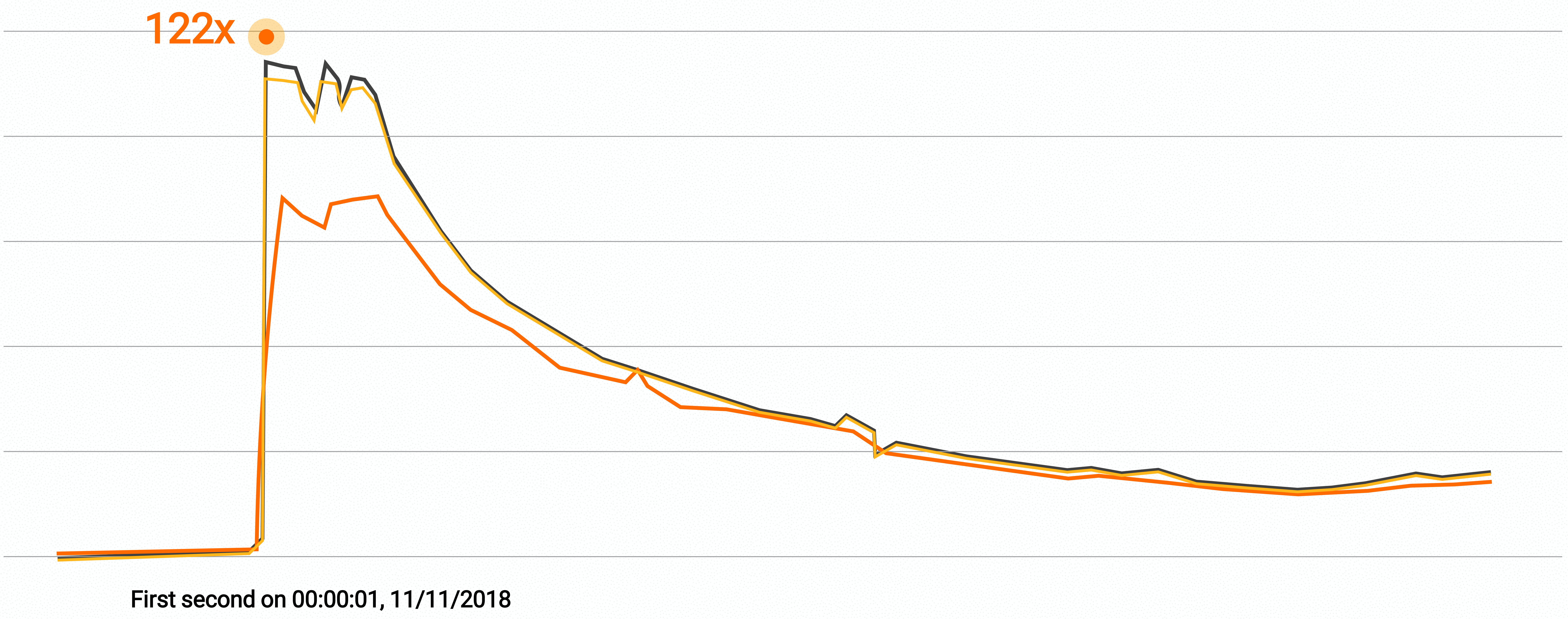
**GMV of FY2018**

552 million monthly active users

# Singles Day (11/11) 2018

Alibaba Cloud | Worldwide Cloud Services Partner

Database workloads (~500,000 sales transactions per second at peak, which leads to a few million TPS)




Hundreds of thousands of DB instances, >tens of PB of data

# Alibaba Database Usage — Public Cloud

Finance

Retailer

Manufacturer

Media and Entertainment

International Clients

# DB Revenue Report by Gartner

Alibaba Cloud | Worldwide Cloud Services Partner

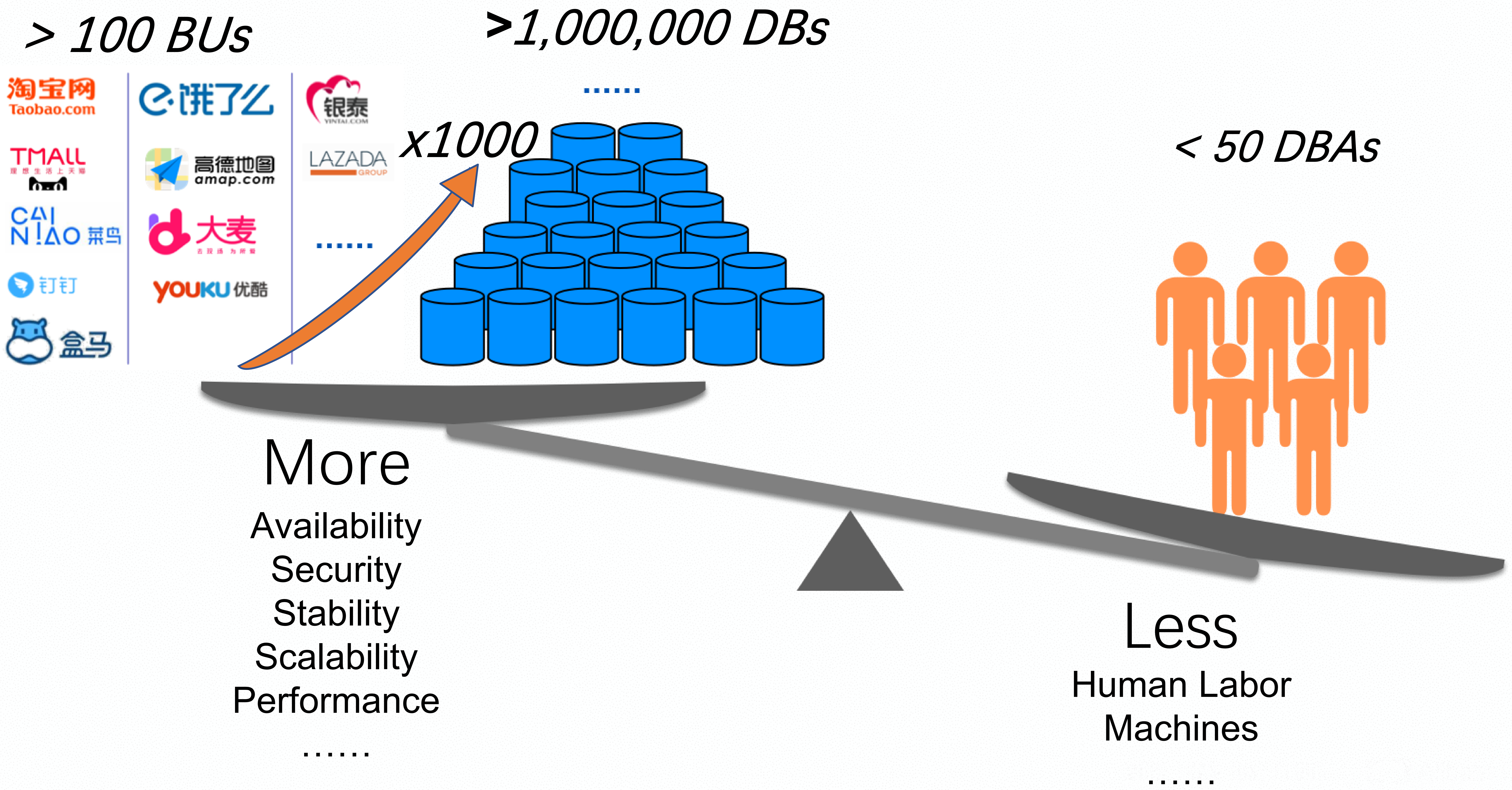
Table 1. DBMS Cloud Services Revenue (2016-2018)

| | Revenue | | | Revenue Growth | |
|---|---|---|---|---|---|
| | 2016 | 2017 | 2018 | 2017 | 2018 |
| Amazon | 1,700.64 | 3,615.90 | 6,319.11 | 112.62% | 74.76% |
| Microsoft | 53.38 | 918.27 | 2,149.40 | 1620.11% | 134.07% |
| Alibaba | 96.93 | 213.44 | 460.55 | 120.21% | 115.77% |
| Oracle | 100.16 | 224.76 | 373.12 | 124.40% | 66.01% |
| Google | 101.47 | 164.36 | 285.49 | 61.98% | 73.70% |
| Tencent | 21.87 | 110.85 | 247.30 | 406.96% | 123.09% |
| Huawei | 77.42 | 70.99 | 137.87 | –8.32% | 94.22% |
| IBM | 57.28 | 73.35 | 120.22 | 28.06% | 63.90% |
| Cloudera | 23.85 | 45.35 | 79.21 | 90.15% | 74.66% |
| MongoDB | 9.12 | 8.77 | 65.70 | –3.90% | 649.33% |
| Other | 127.07 | 171.69 | 250.74 | | |
| **Grand Total** | **2,369.19** | **5,617.73** | **10,488.70** | **137.12%** | **86.71%** |
| *% of Total DBMS* | *6.87%* | *14.43%* | *22.75%* | | |
| Data sourced from "Market Share: Enterprise Platform as a Service, Worldwide, 2018" | | | | | |

Source: Gartner (June 2019)

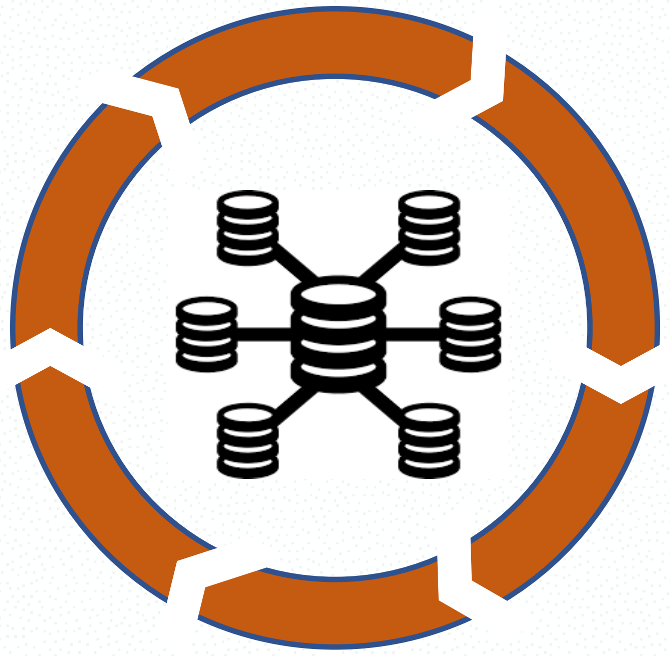
| | Revenue | | | Revenue Growth | |
|---|---|---|---|---|---|
| | 2016 | 2017 | 2018 | 2017 | 2018 |
| Alibaba | 96.93 | 213.44 | 460.55 | 120.21% | 115.77% |

# Background - Do More with Less



# Challenges

Alibaba Cloud
Worldwide Cloud Services Partner



GLOBAL SHOPPING FESTIVAL 2018

## Management at Scale

Scheduling
Protection
Runtime Management
Optimization
Backup/Restore
Security
......

## Hotspot

Scalability
Elasticity
Stability
Cost
......

## Workloads Diversity

SLA-driven
Workload-aware
Agility
>100 BUs
>20,000 Developers
......

# Journey to Cloud Native Database Systems

Alibaba Cloud | Worldwide Cloud Services Partner

**Human Labor**

DBAs

Commercial Database

Oracle

~2010

**DBPaaS**

Large-scale DB Lifecycle Management Infra

Open Source Database

AliSQL/MySQL

2010~2017

**SDDP**

Self-Driving Database Platform

Cloud-Native Database

POLARDB (VLDB 2018, SIGMOD 2019)

AnalyticDB (VLDB 2019)

2017~now

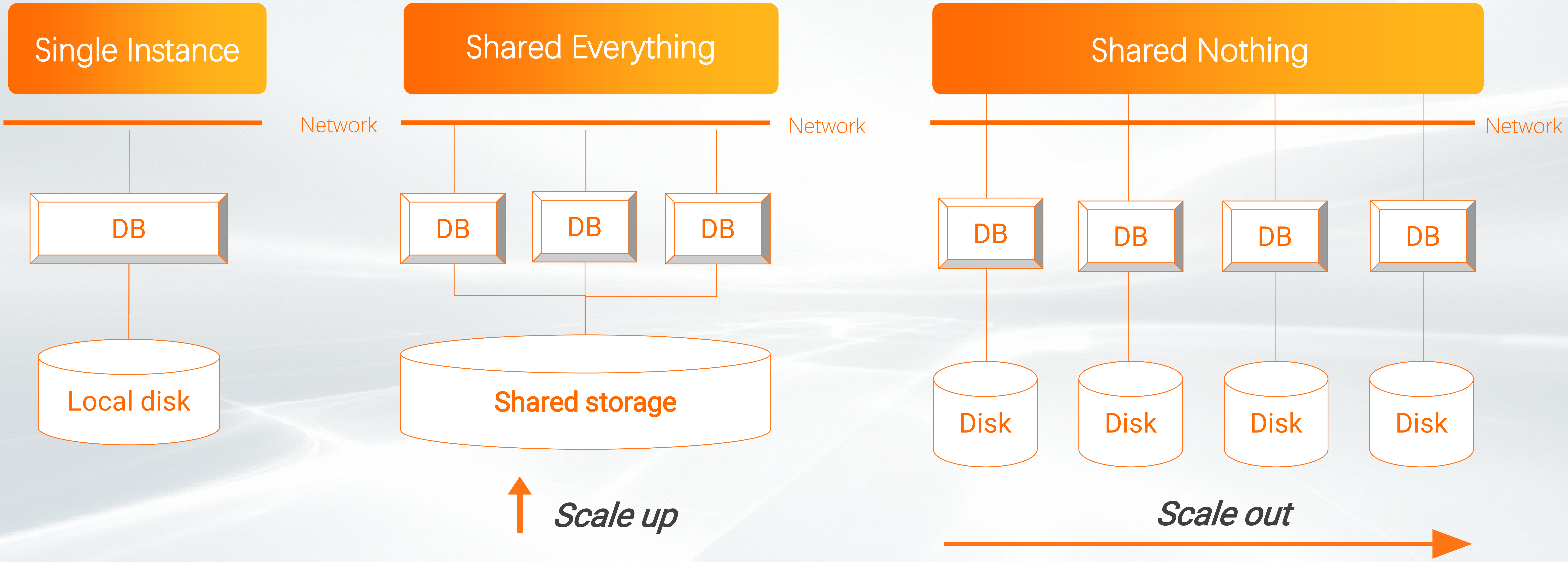
# Outline

Worldwide Cloud Services Partner

- Background
- POLARDB
- AnalyticDB
- Self-Driving Database Platform
- Conclusion

# Three Different Architectures






e.g. MySQL, PG

e.g. POLARDB

e.g. PolarDB-X, AnalyticDB

# Background of POLARDB



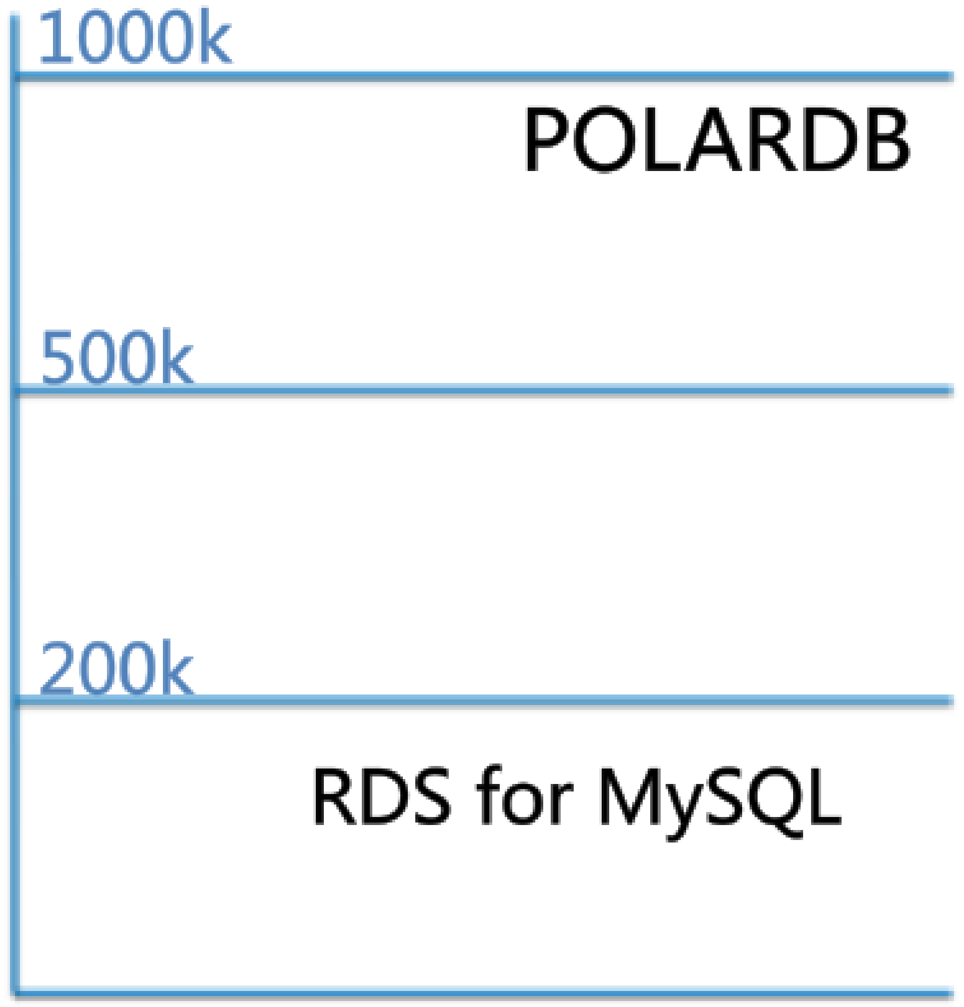
## - Design of POLARDB – A Cloud Native Database

- Elasticity
- Low Cost
- Performance
- Continuous solution

# POLARDB Key Features

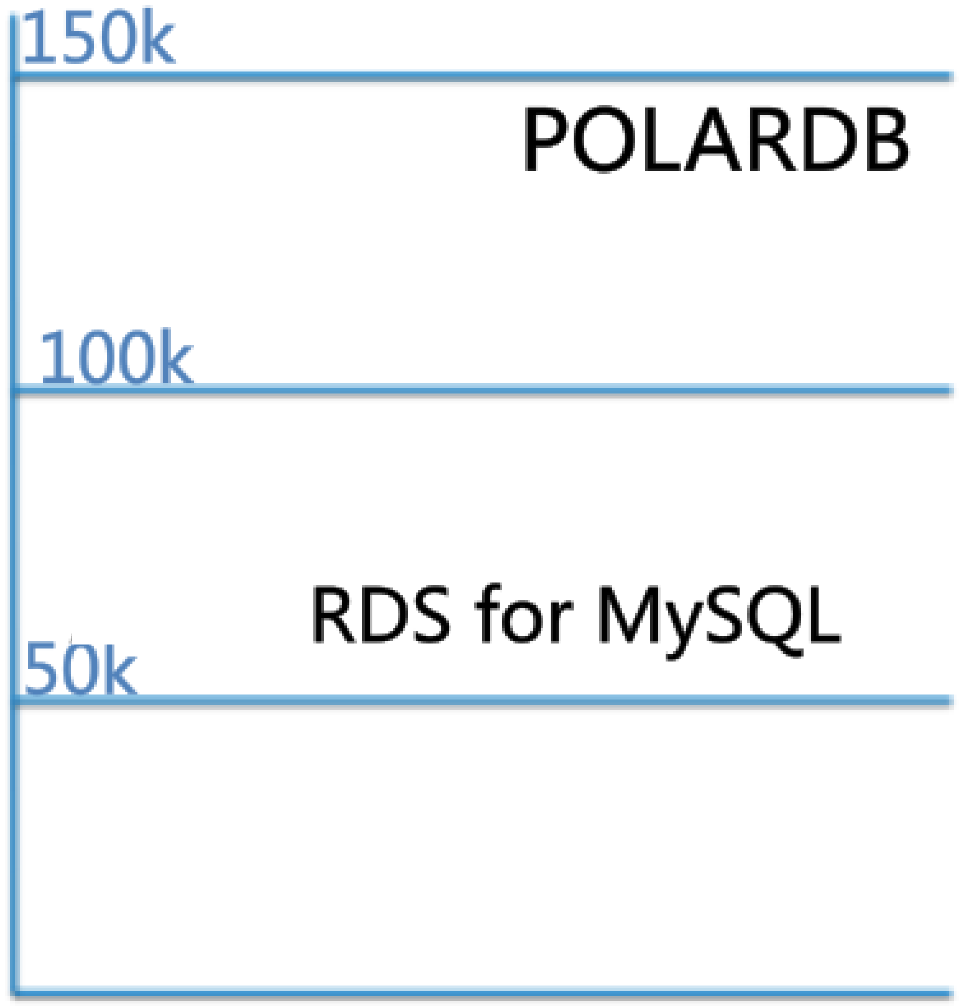
Alibaba Cloud | Worldwide Cloud Services Partner

## 1 M QPS



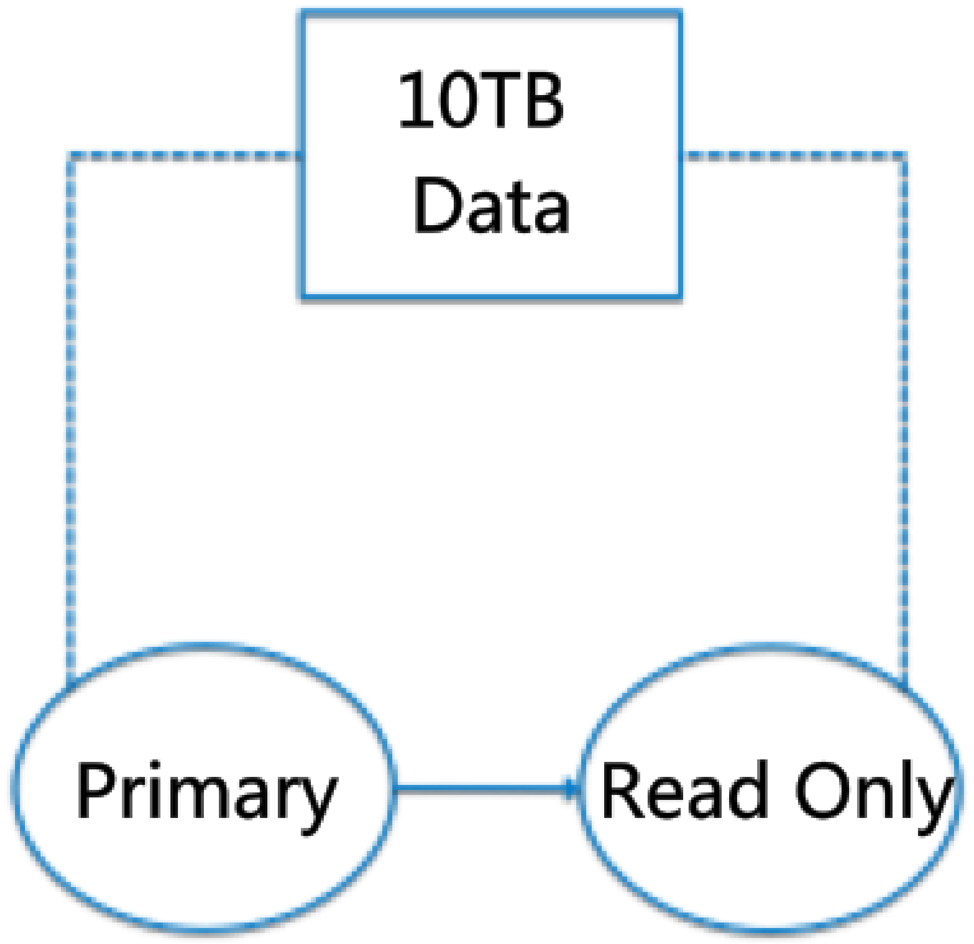

Read Performance

## 130K TPS



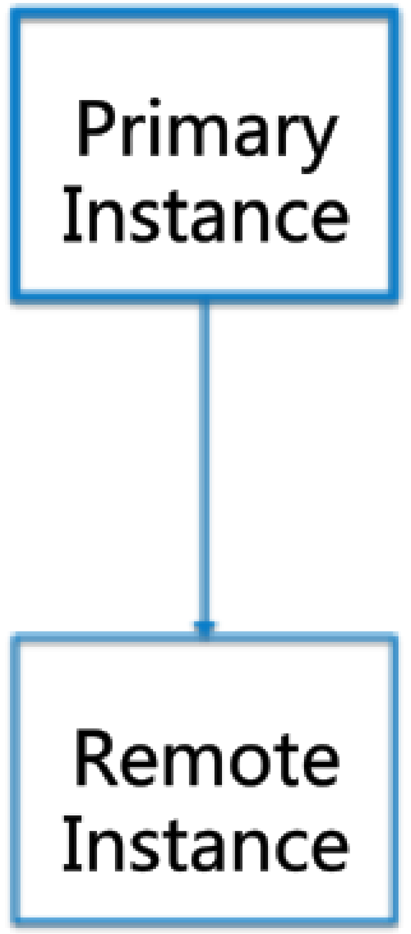

Write Performance

## 2min




RO Copies

## 3min




Remote Backup

**- 100TB Capacity**

**- 100% MySQL/PG Compatible**

# POLARDB Architecture -- Overview



## - Low Latency Oriented

- Pure Userspace I/O Stack
- Zero Copy, RDMA

## - Design for Emerging Hardware

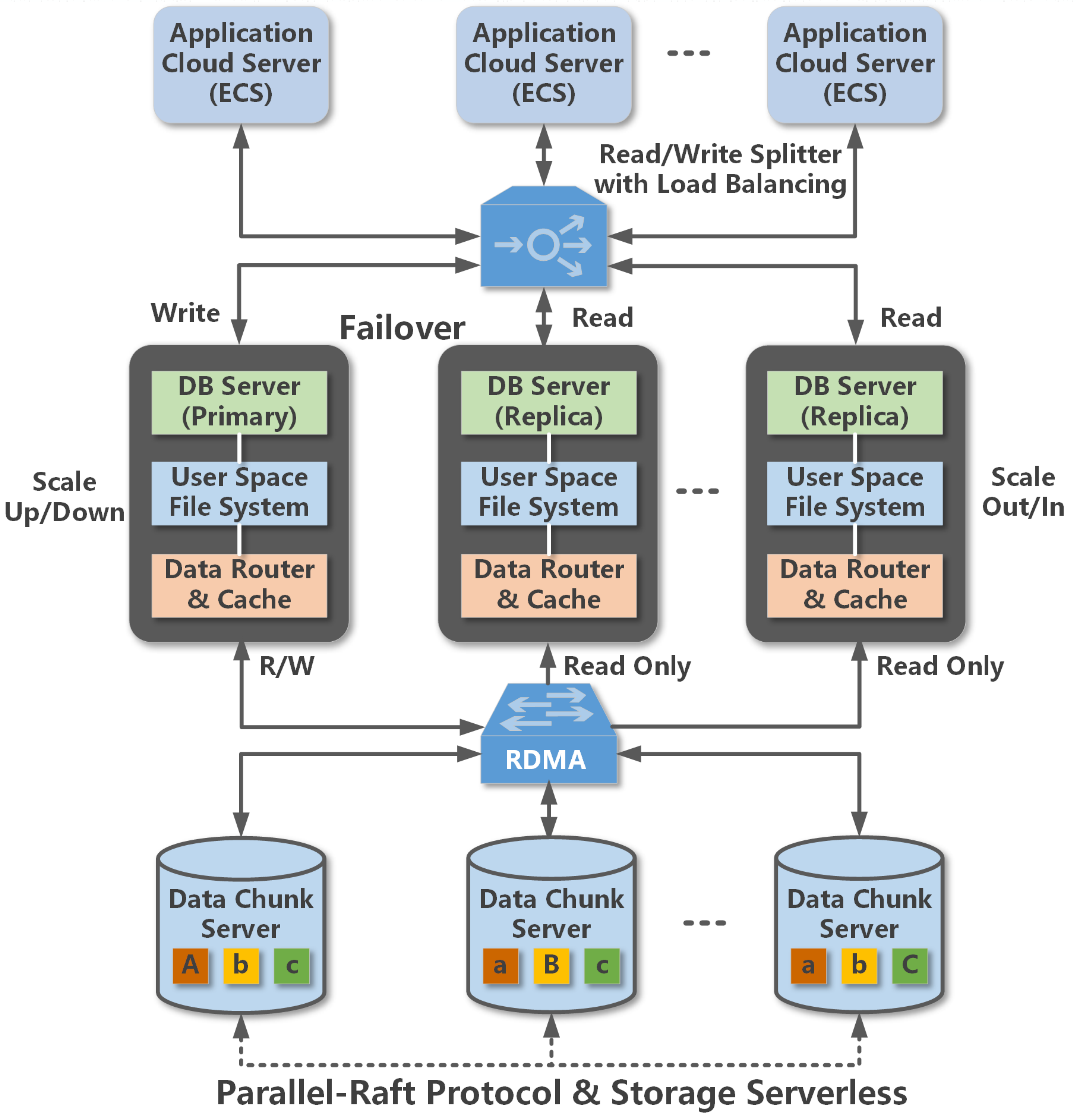
- RDMA NIC
- NVMe SSD, Optane

## - Active R/W – Active RO

- One Write / Multiple Read

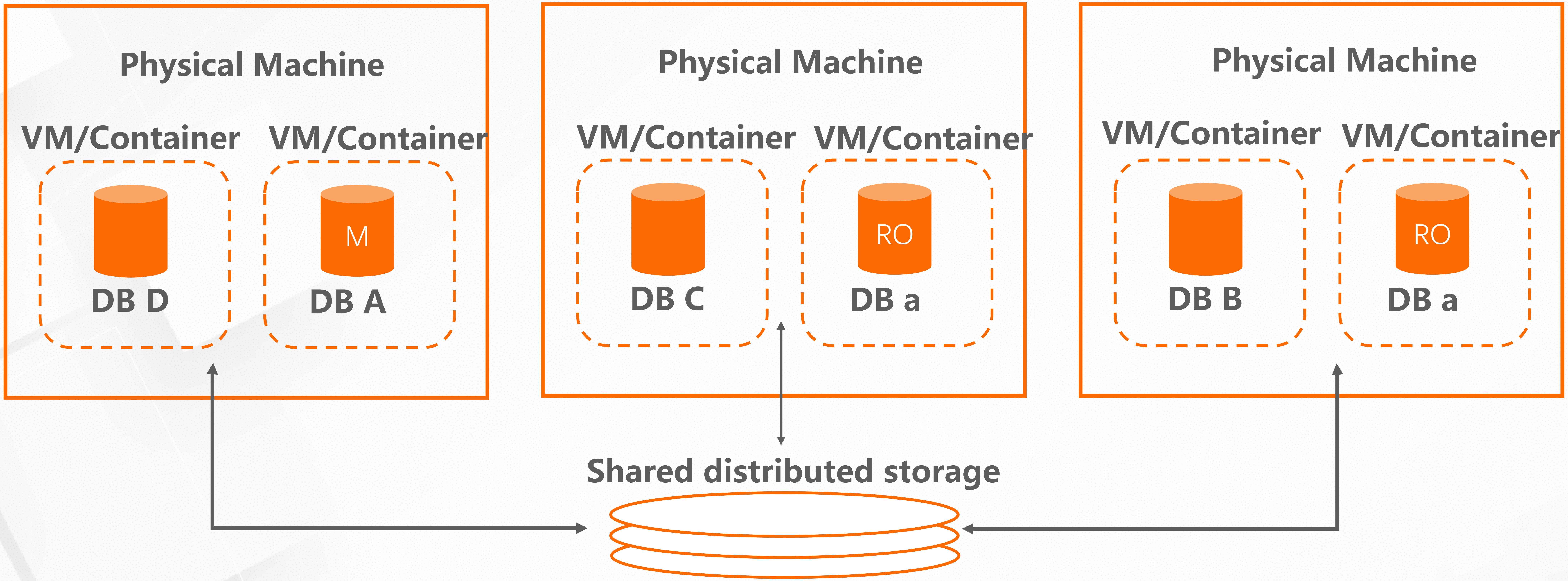
## - High Avalability

- Three Replicas
- ParallelRaft



# POLARDB Architecture

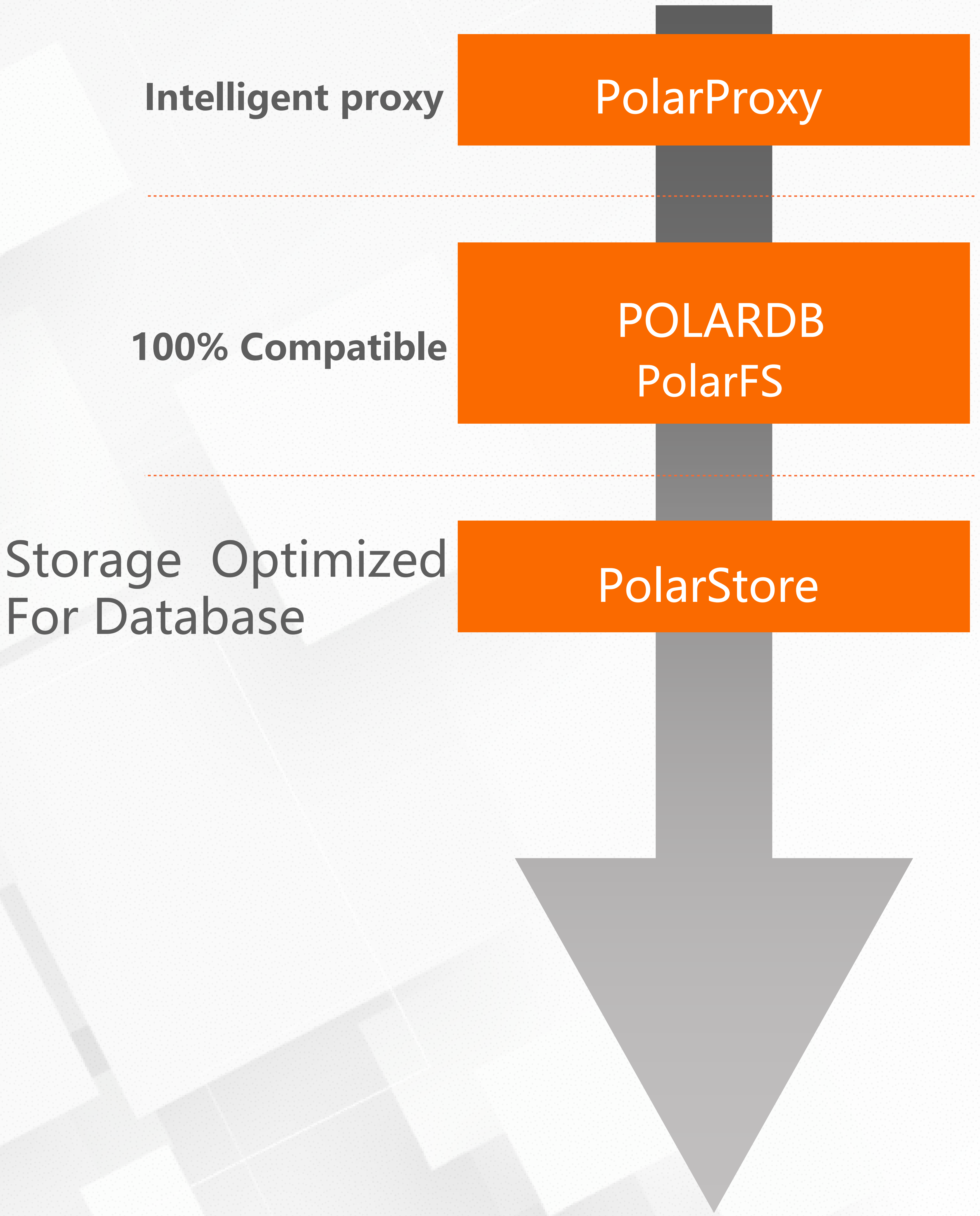


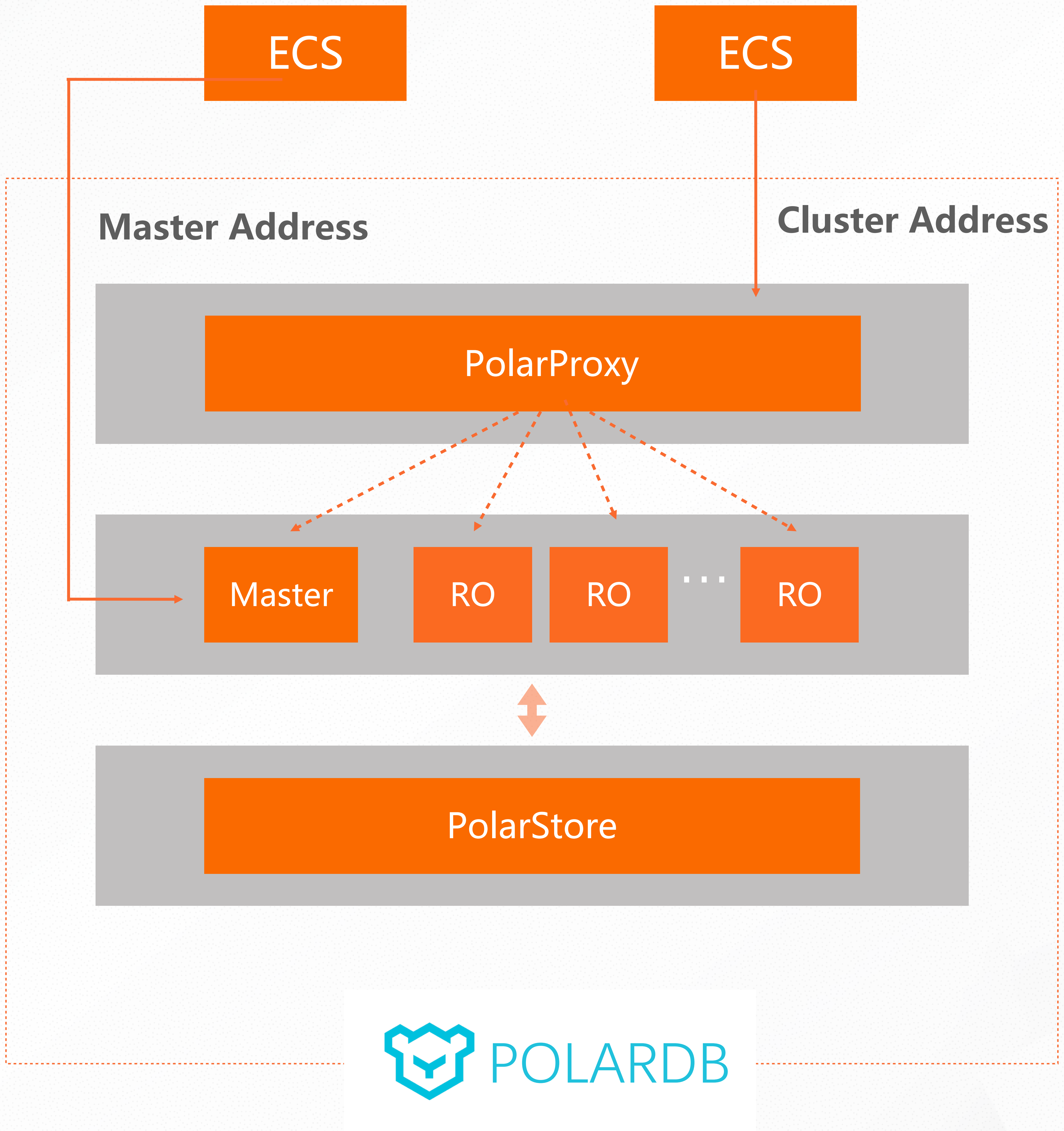



**Active-Active + Decouple of Reads and Writes + Serverless Storage**

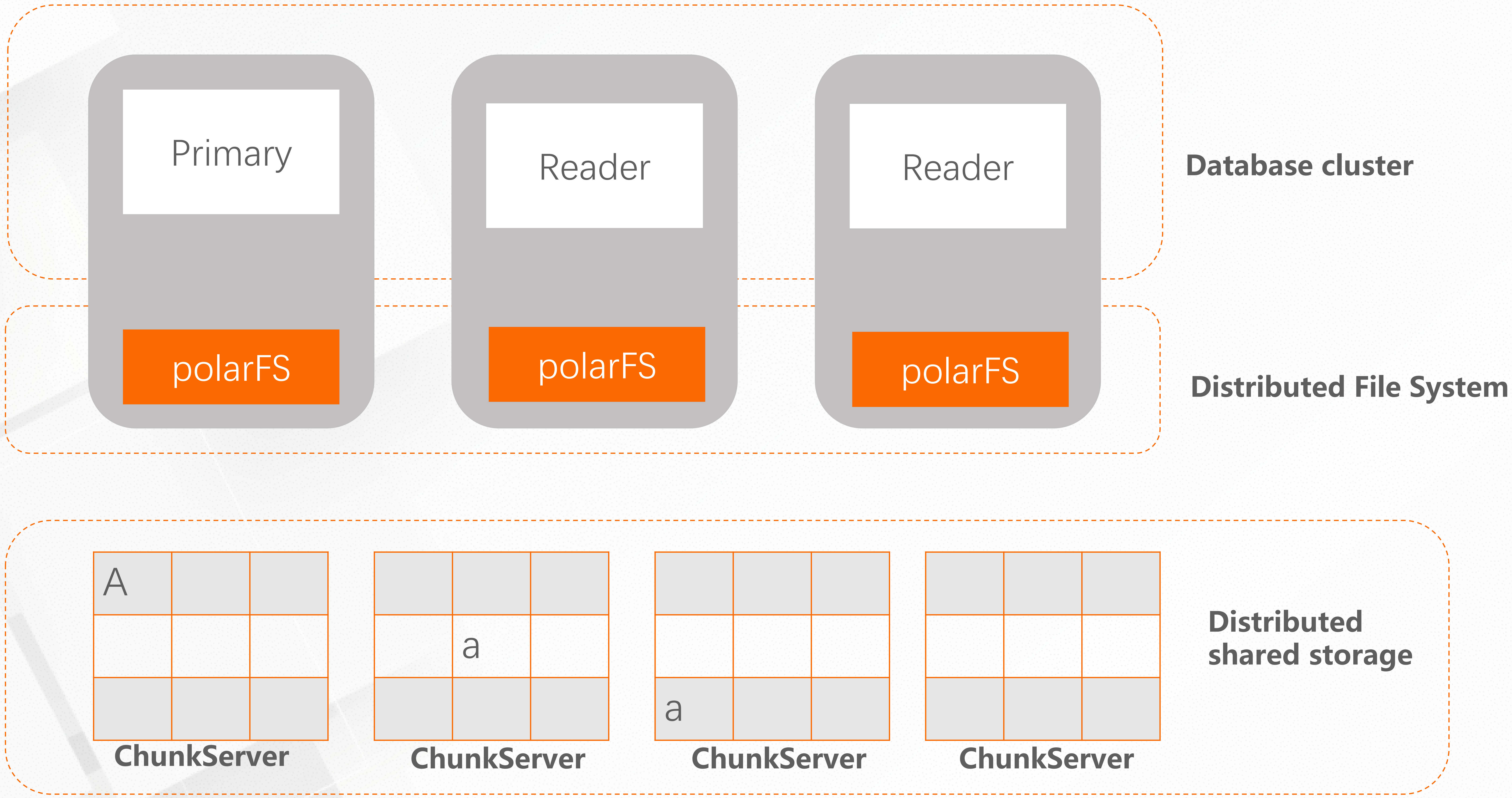
# POLARDB Architecture





# POLARDB Architecture




RDMA&SPDK

ParallelRaft

Lockless storage

**Excellent Elasticity**

2core vCPU upgrades to 32core in < 5mins
2 nodes scales out to 4 nodes < 5mins

**Cost reduction**

Serverless on-demand billing

**100T, max-capacity**

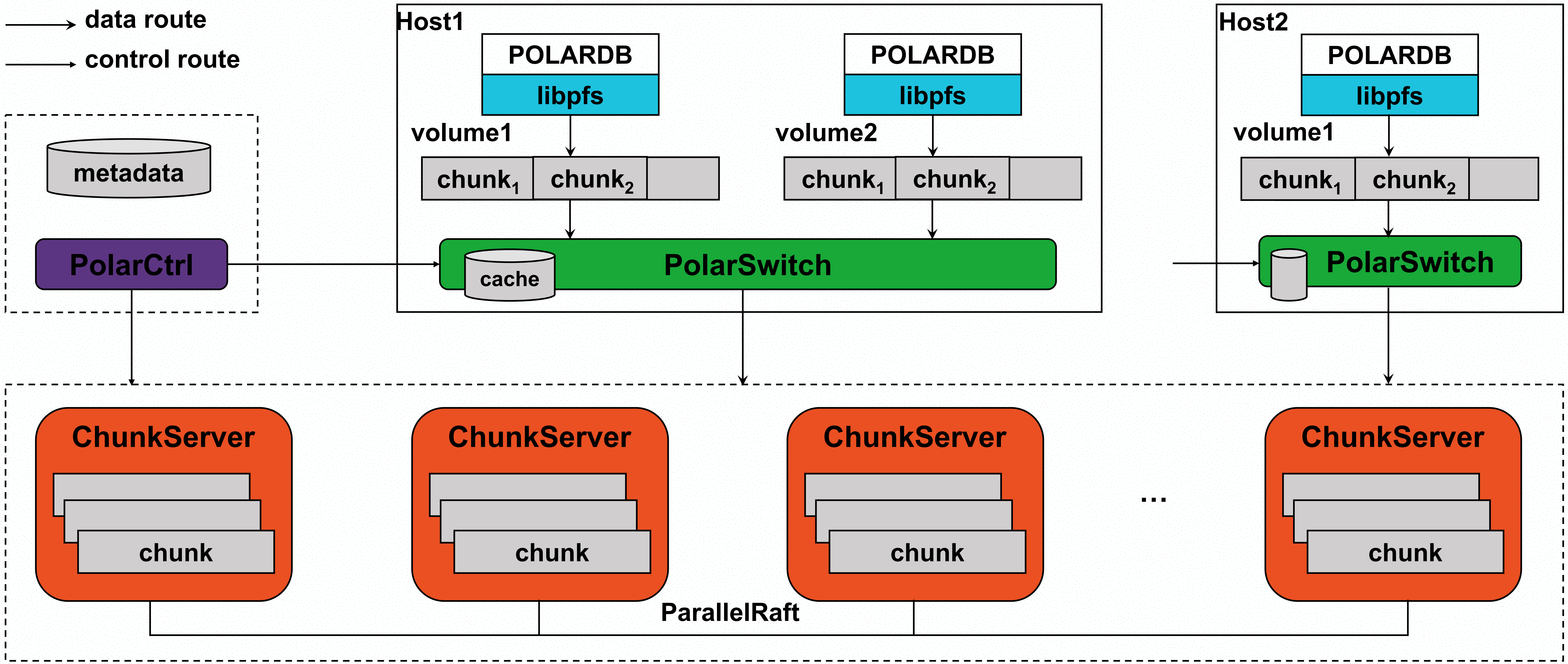
Elasticity: scale-out

**Lockless Backup**

Snapshot backup without locking

# Architecture of PolarFS (VLDB 2018) -- Inside






- Key Components: 1. libpfs 2. PolarSwitch 3. ChunksServer 4. PolarCtrl

# Architecture of PolarFS -- Components

## - 1. libpfs

### - Library in Userspace

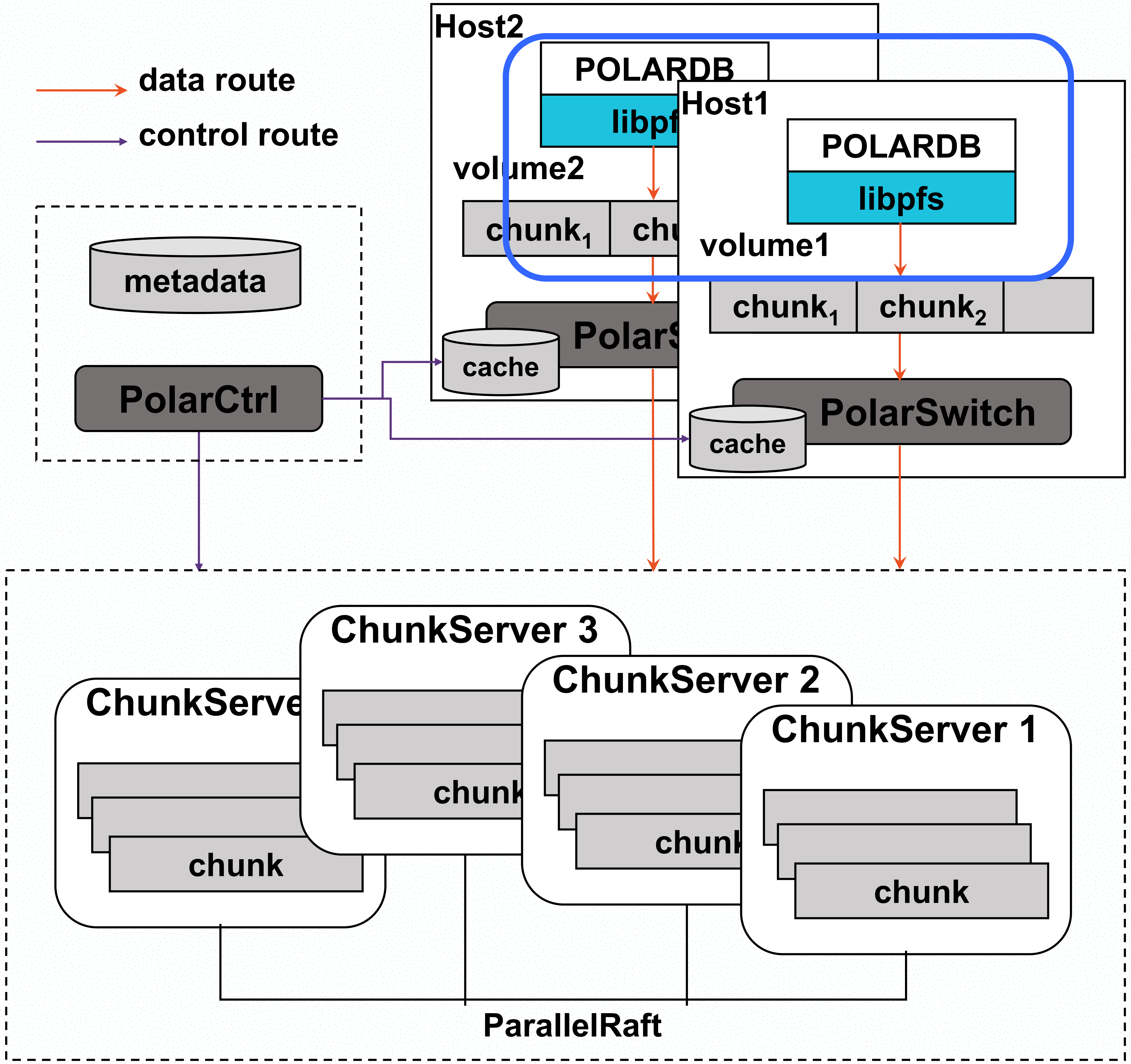
- Enable I/O without trapping into Kernel

### - POSIX-like File System API

- Easy to port different kernels to run on PolarFS

```
int     pfs_mount(const char *volname, int host_id)
int     pfs_umount(const char *volname)
int     pfs_mount_growfs(const char *volname)

int     pfs_creat(const char *volpath, mode_t mode)
int     pfs_open(const char *volpath, int flags, mode_t mode)
int     pfs_close(int fd)
ssize_t pfs_read(int fd, void *buf, size_t len)
ssize_t pfs_write(int fd, const void *buf, size_t len)
off_t   pfs_lseek(int fd, off_t offset, int whence)
---
```



# Architecture of PolarFS -- Components



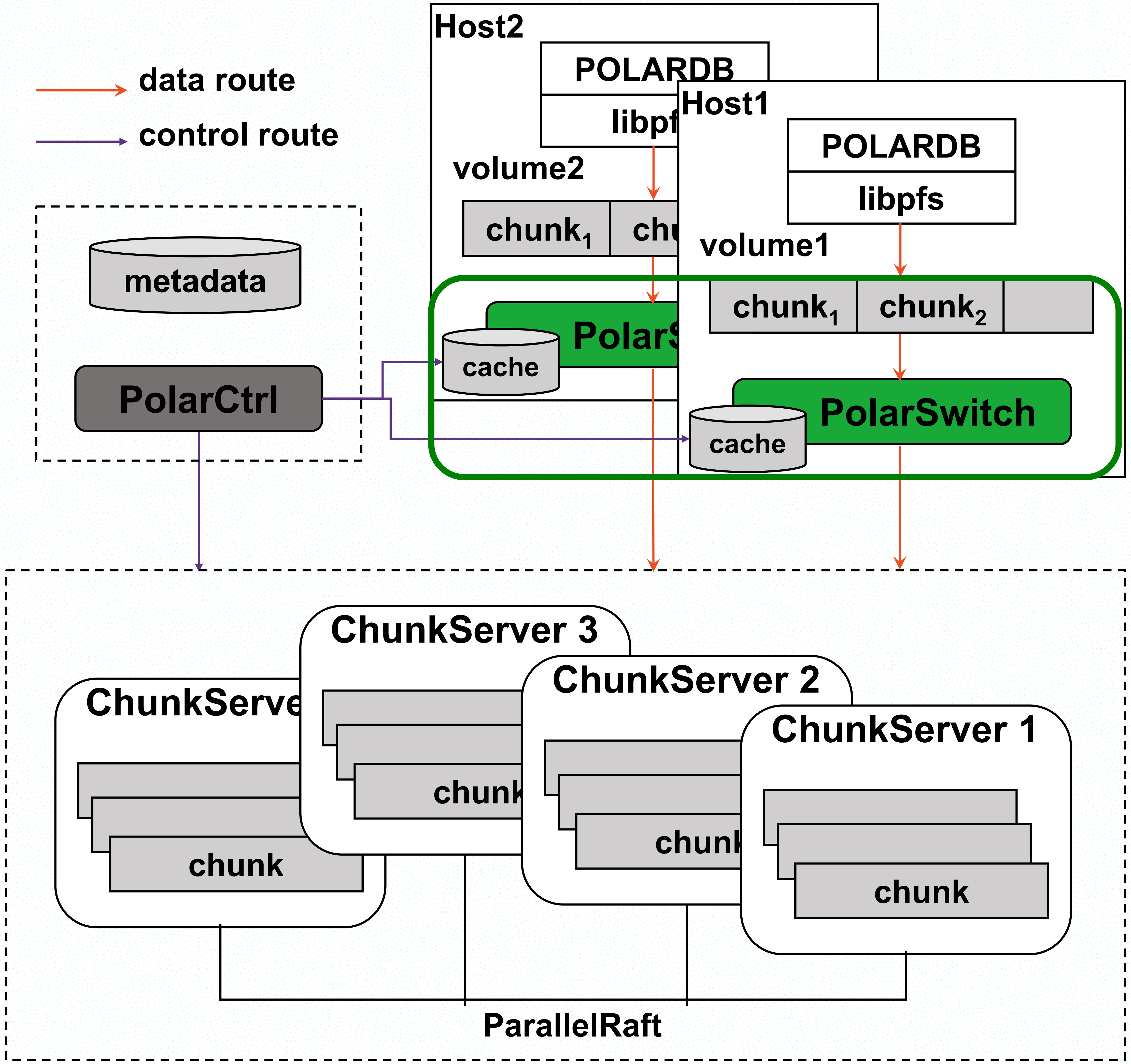
## - 2. PolarSwitch

### - Proxy Daemon
- Run on Compute Nodes

### - Remap I/O Requests
- Cache Mapping Metadata

### - Send Remapped I/O Requests
- To one or more Storage Node(s)

### - Write Data to Primary Replica



# Architecture of PolarFS -- Components

Alibaba Cloud | Worldwide Cloud Services Partner

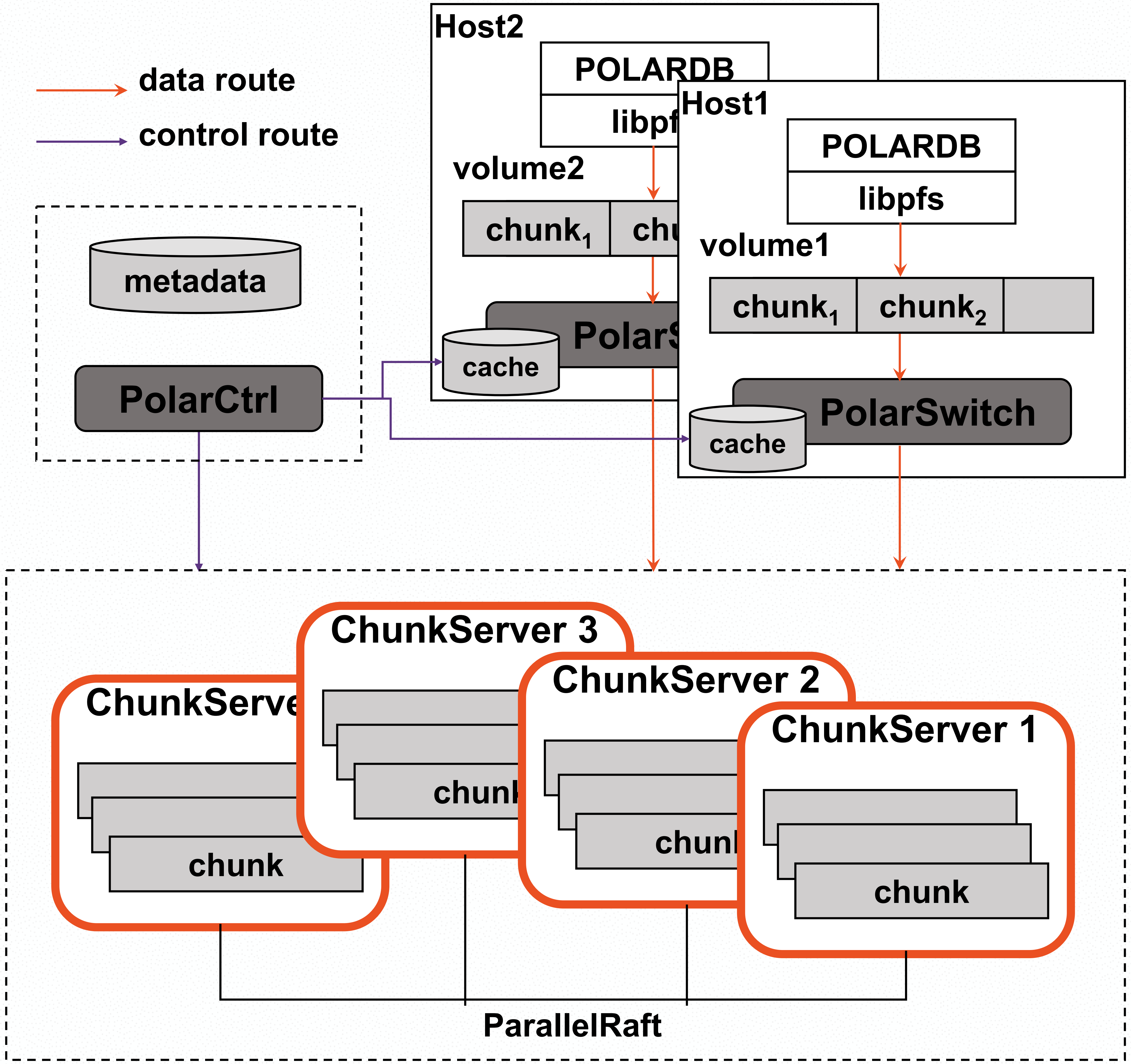
## - 3. ChunkServer

### - Run on Storage Nodes

### - Each ChunkServer
- Manage One NVMe SSD Device
- Bound to One CPU Core

### - Inner Chunk allocation
- Use WAL for A and D

### - Chunk Replication
- Parallel Raft Consensus Protocol



# Architecture of PolarFS -- Components

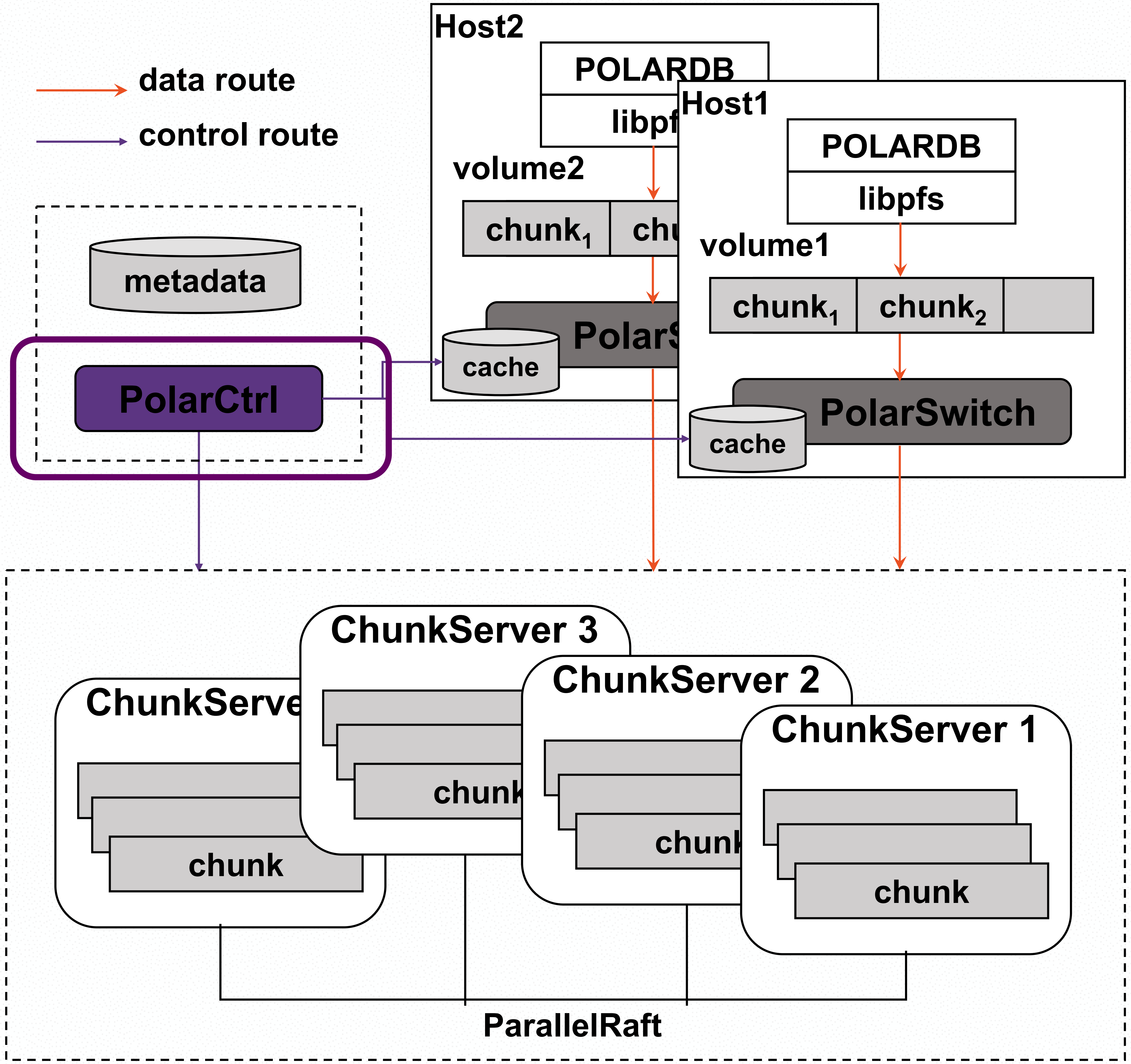
## - 4. PolarCtrl

**- Control Plane of PolarFS Cluster**

- ChunkServer Membership
- Chunk Arrangement and Load Balancing
- Metadata Center(push to PolarSwitch)
- Monitor System Health
  - I/O Performance for Each Volume (Latency, IOPS, etc.)
  - Scheduling Data CRC Check Periodically

**- Not on critical I/O Path**

- Most I/Os Only Invovle PolarSwitch and ChunkServer



# Put Them Together

Alibaba Cloud | Worldwide Cloud Services Partner

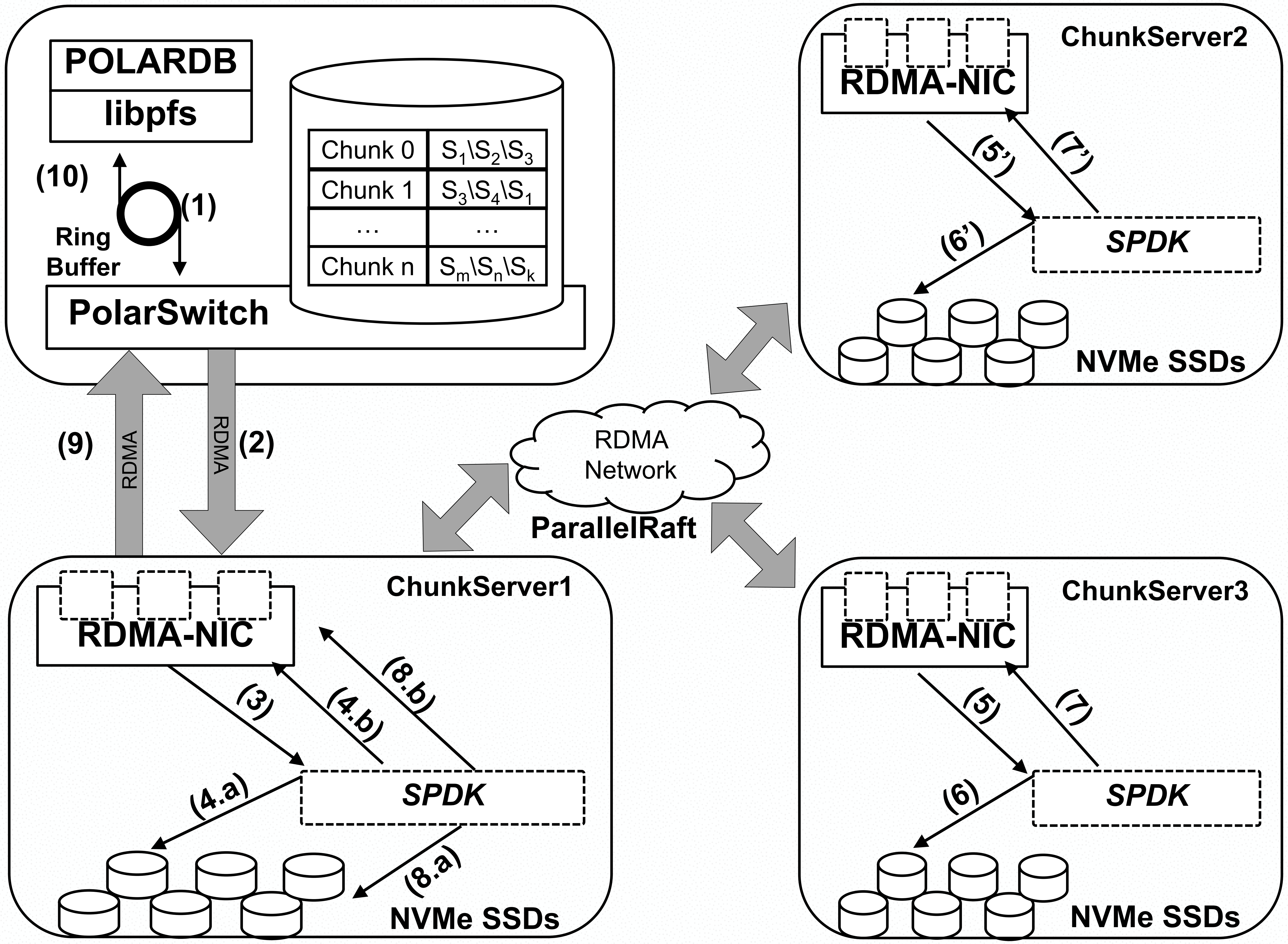
## - A Write I/O Flow

### - All in Userspace

1. A request by Database
2. Libpfs make IO requests to PolarSwitch
3. PolarSwitch RDMA to ChunkServer
4. ChunkServer write SSD by SPDK
5. Main replica RDMA to Follower Node

## - Ultra-low Latency

- No Syscall
- No Context Switch
- No Useless Data Copy



# Failure Resilient of PolarFS -- ParallelRaft



## - ParallelRaft Motivation

**- DB tolerate Out-of-Order I/O completion**

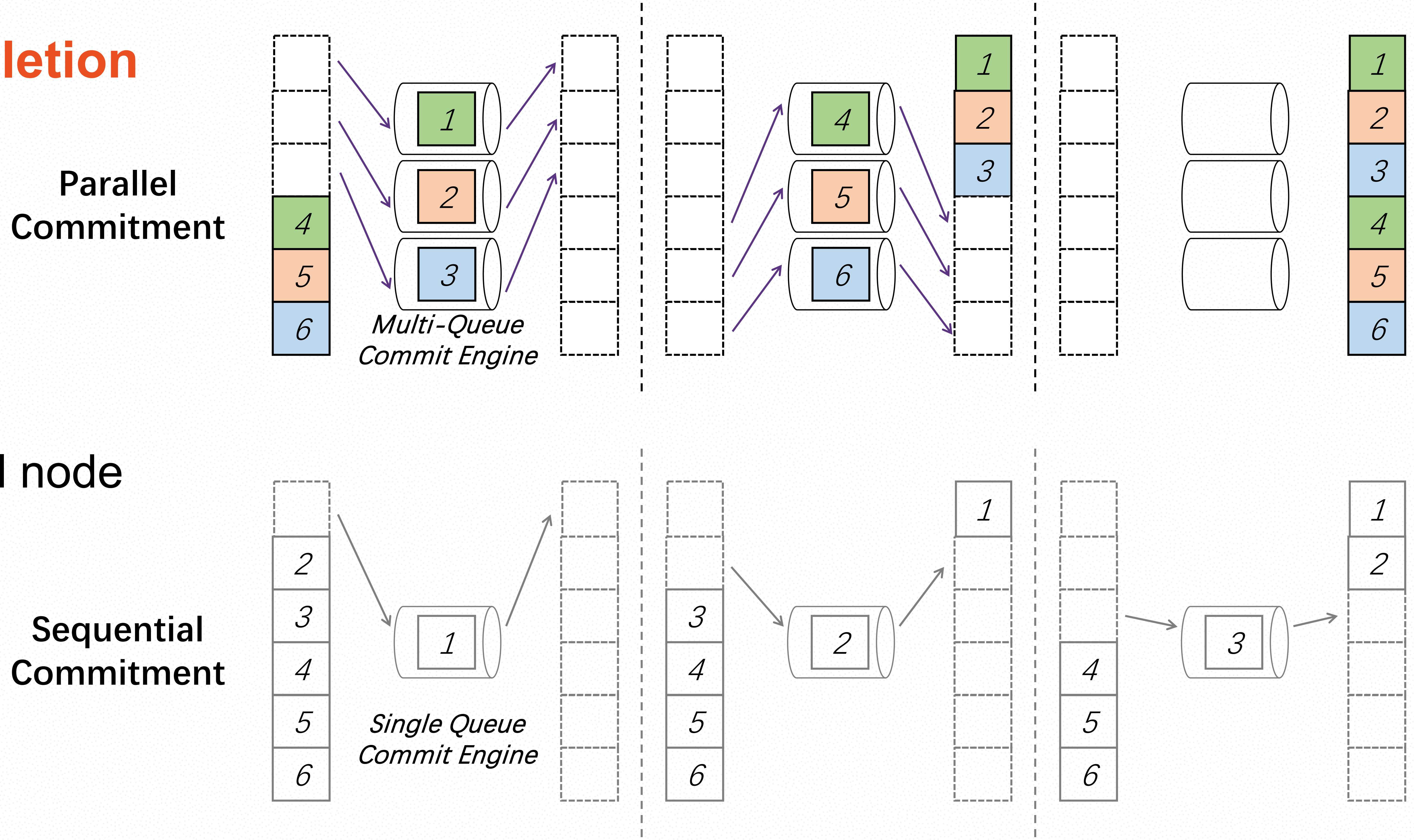
- Normal Storage Semantics

**- Loose Raft Sequential Restriction**

- Commit and Apply out-of-orderly

**- Keep Logical Serialization**

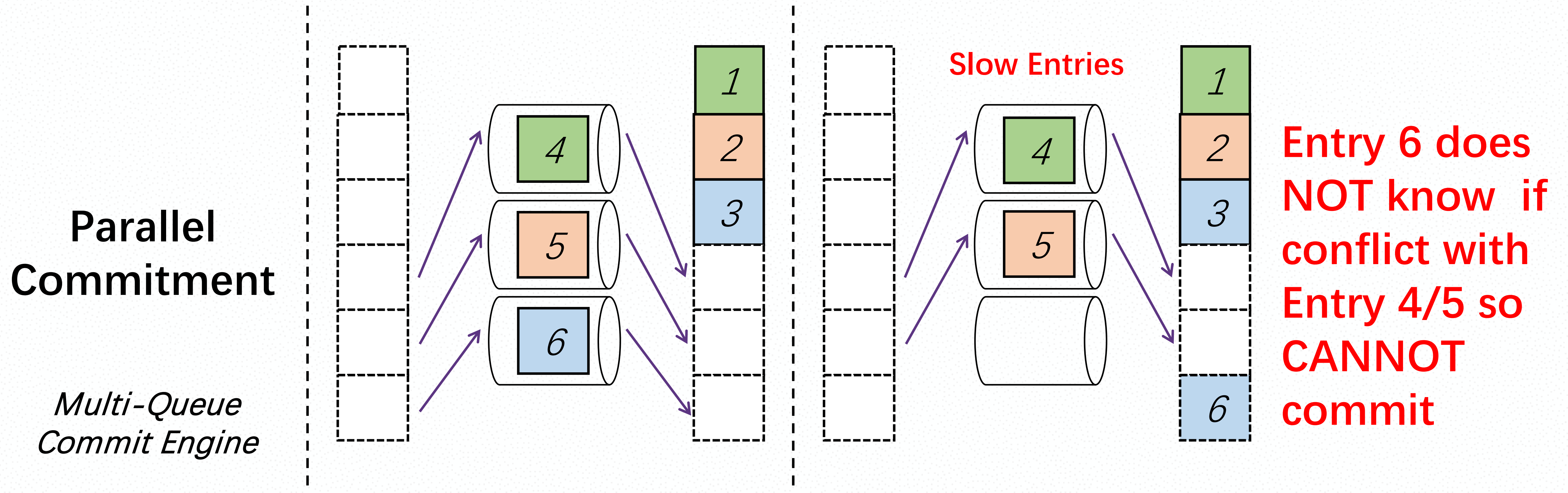
- Overlapped modification consistent on all node

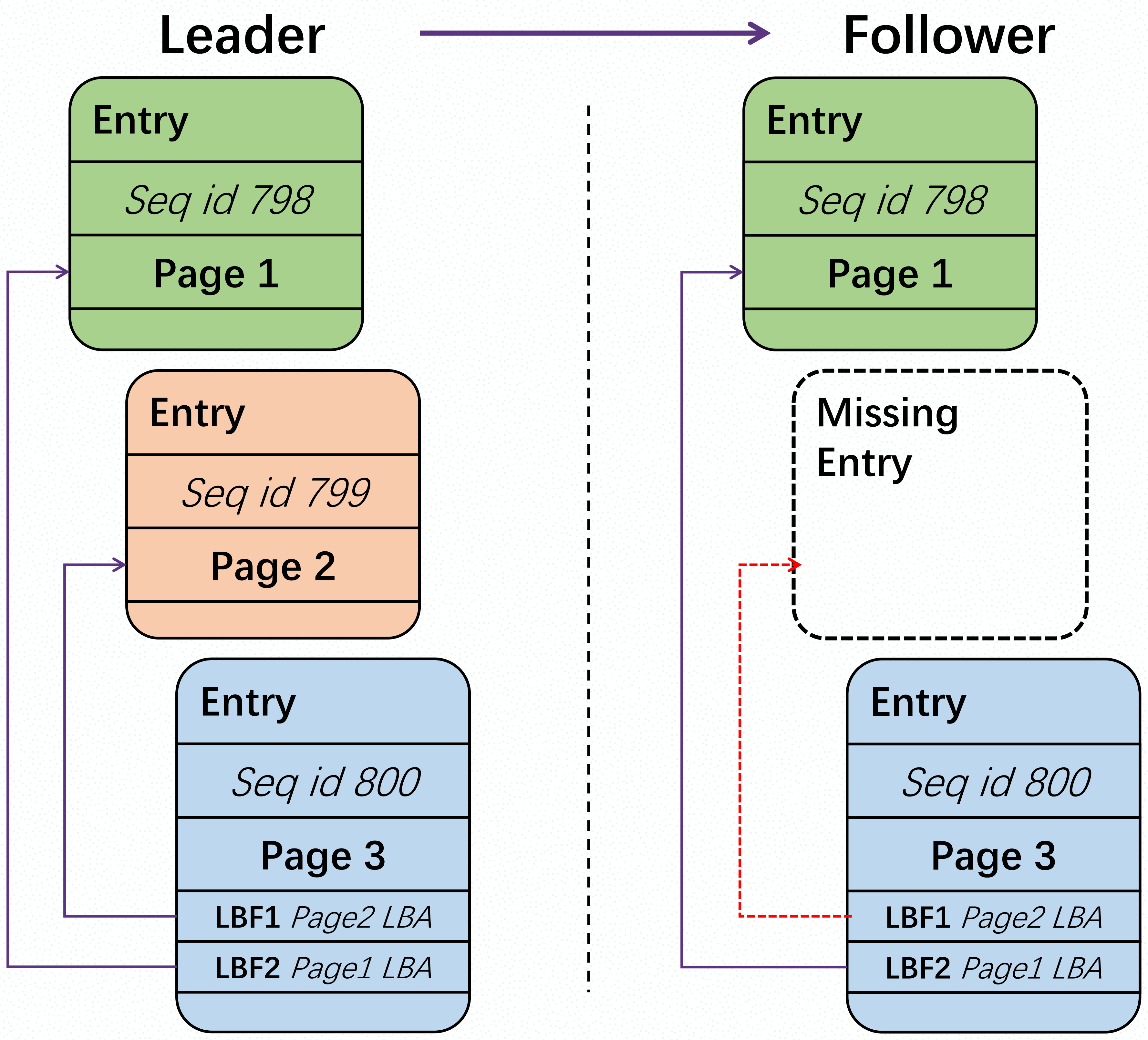


# Failure Resilient of PolarFS -- ParallelRaft

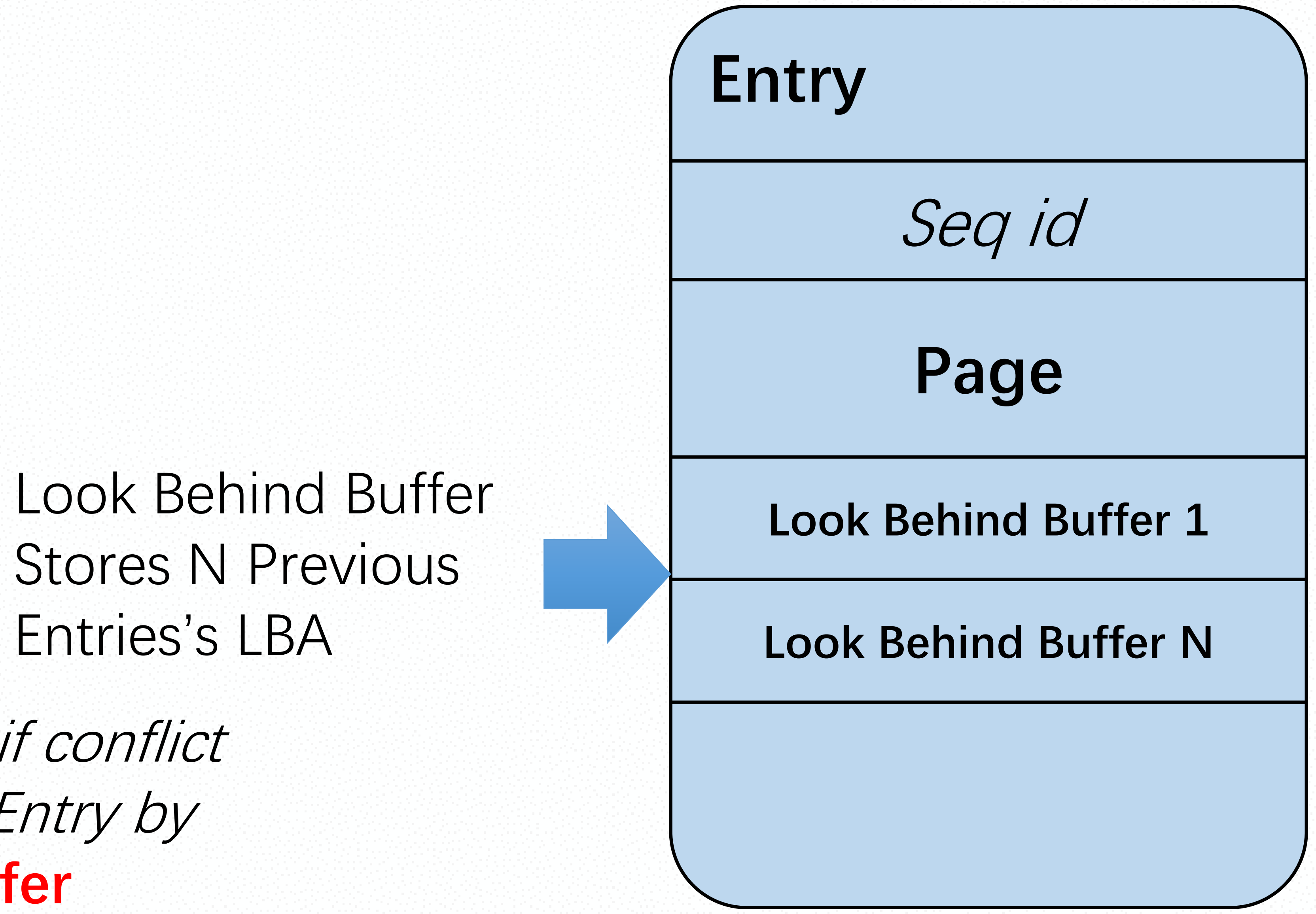


## - Key Problem
## - Missing Entry Conflict or Not??
## - Solution: Look Behind Buffer




*Now Entry Know if conflict with the Missing Entry by*
**Look Behind Buffer**
**(LBA: logical block address)**

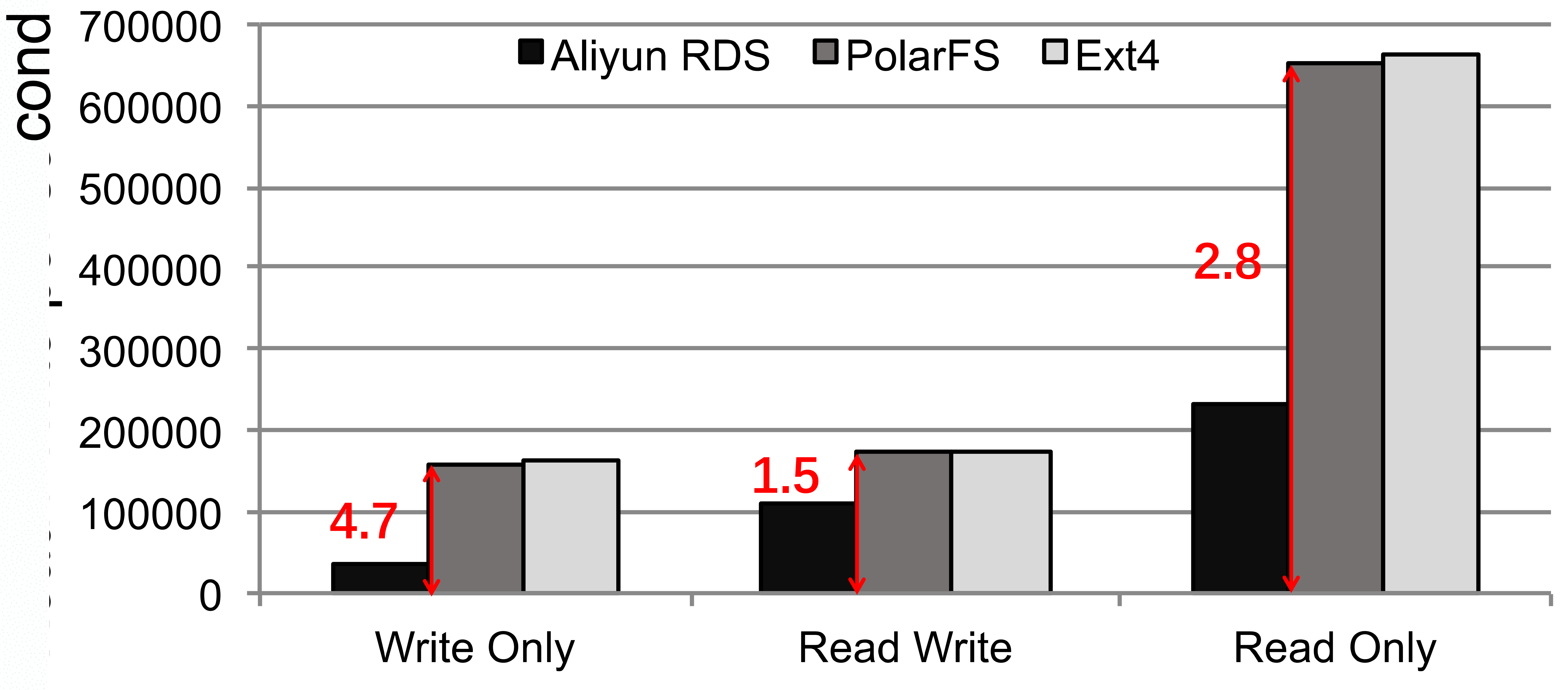


# Evaluation -- Database on PolarFS



## - Three Systems

1.Alibaba MySQL cloud service RDS 2.POLARDB on PolarFS 3.POLARDB on local Ext4



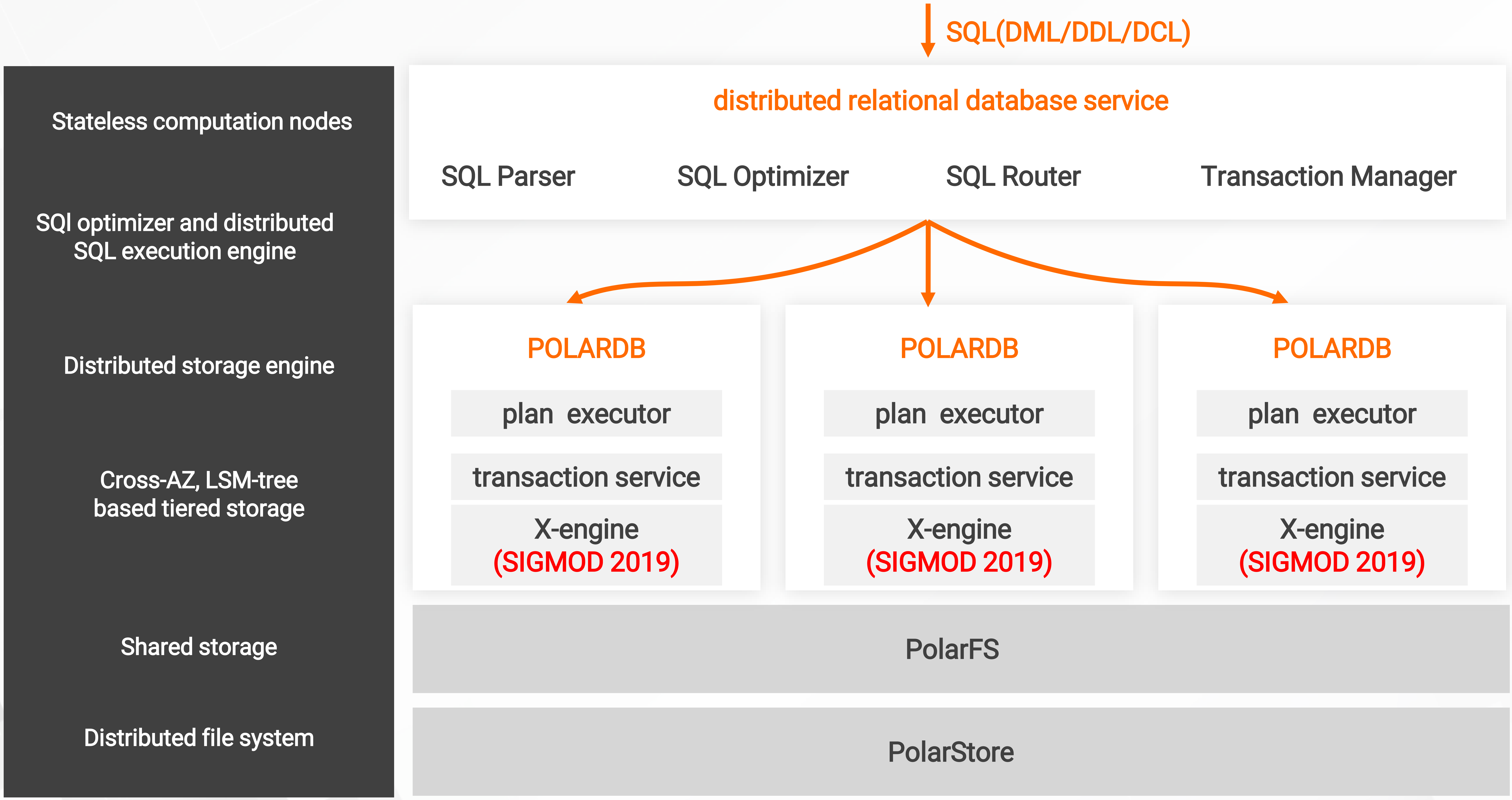

**POLARDB on PolarFS**
**Comparable to Local SSD**
**With 3 Replicas over Nodes**

## - OLTP workloads

- Read-only, Write-only (update : delete : insert = 2:1:1), R/W-mixed (read : write = 7:2)
- Test data size : 500GB, Tables : 250, Records of each table : 8,500,000

# POLARDB-X: shared-nothing layer for scaling out

Worldwide Cloud Services Partner



# Outline

Alibaba Cloud | Worldwide Cloud Services Partner

Background

POLARDB

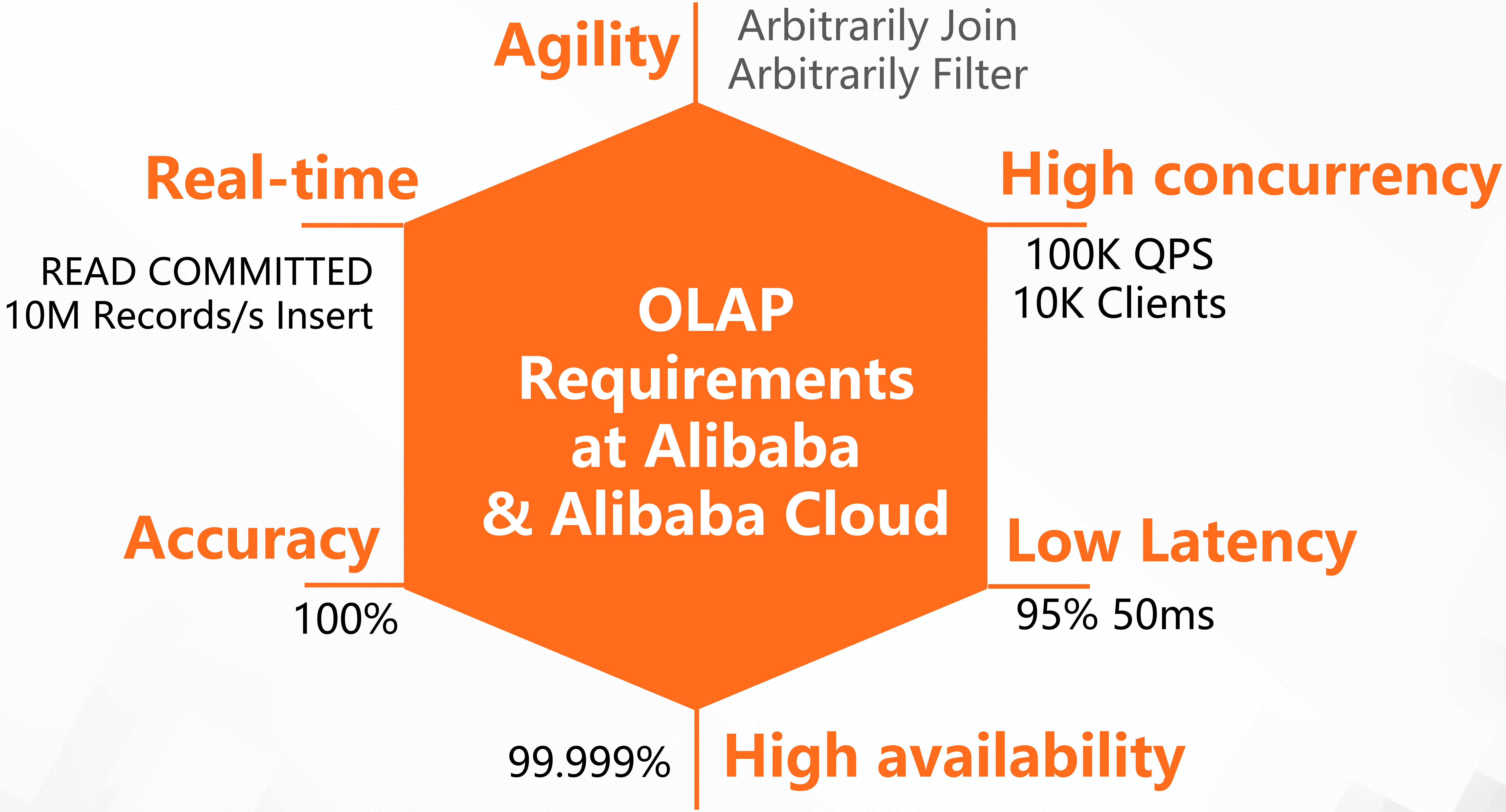
AnalyticDB

Self-Driving Database Platform

Conclusion

# Background

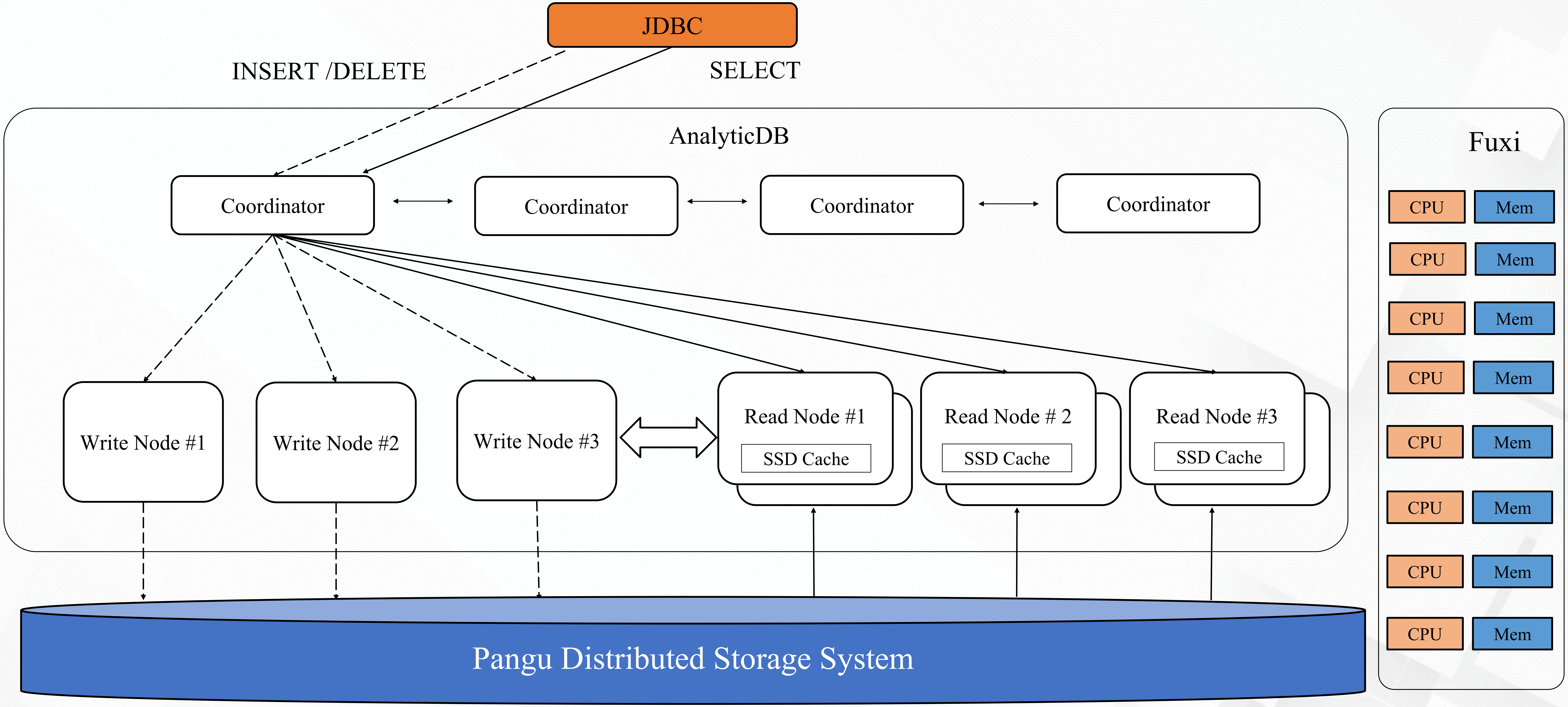





Brand
data Bank
powered by Alibaba
EMS

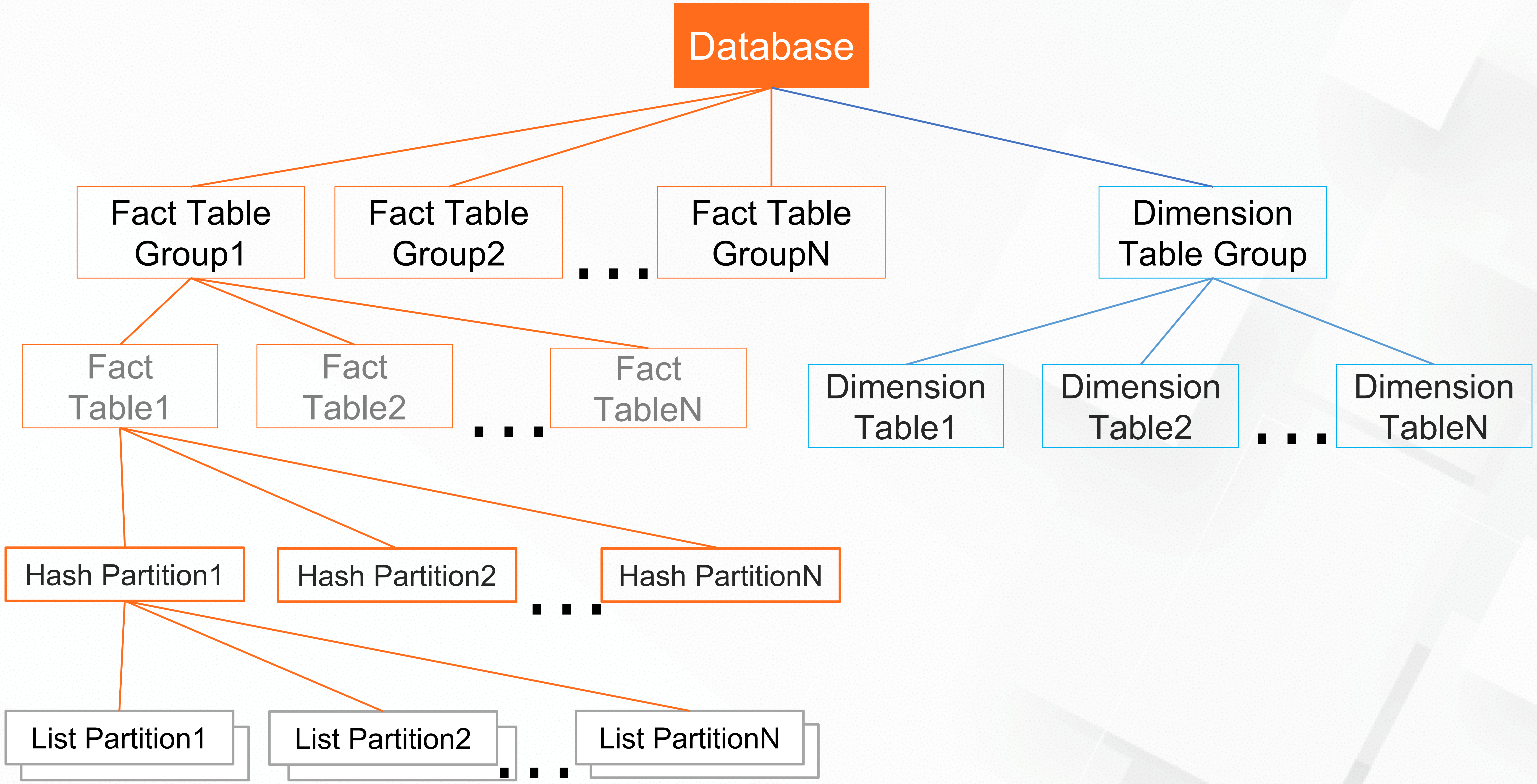
# AnalyticDB Architecture (VLDB 2019)





# Table Partitions

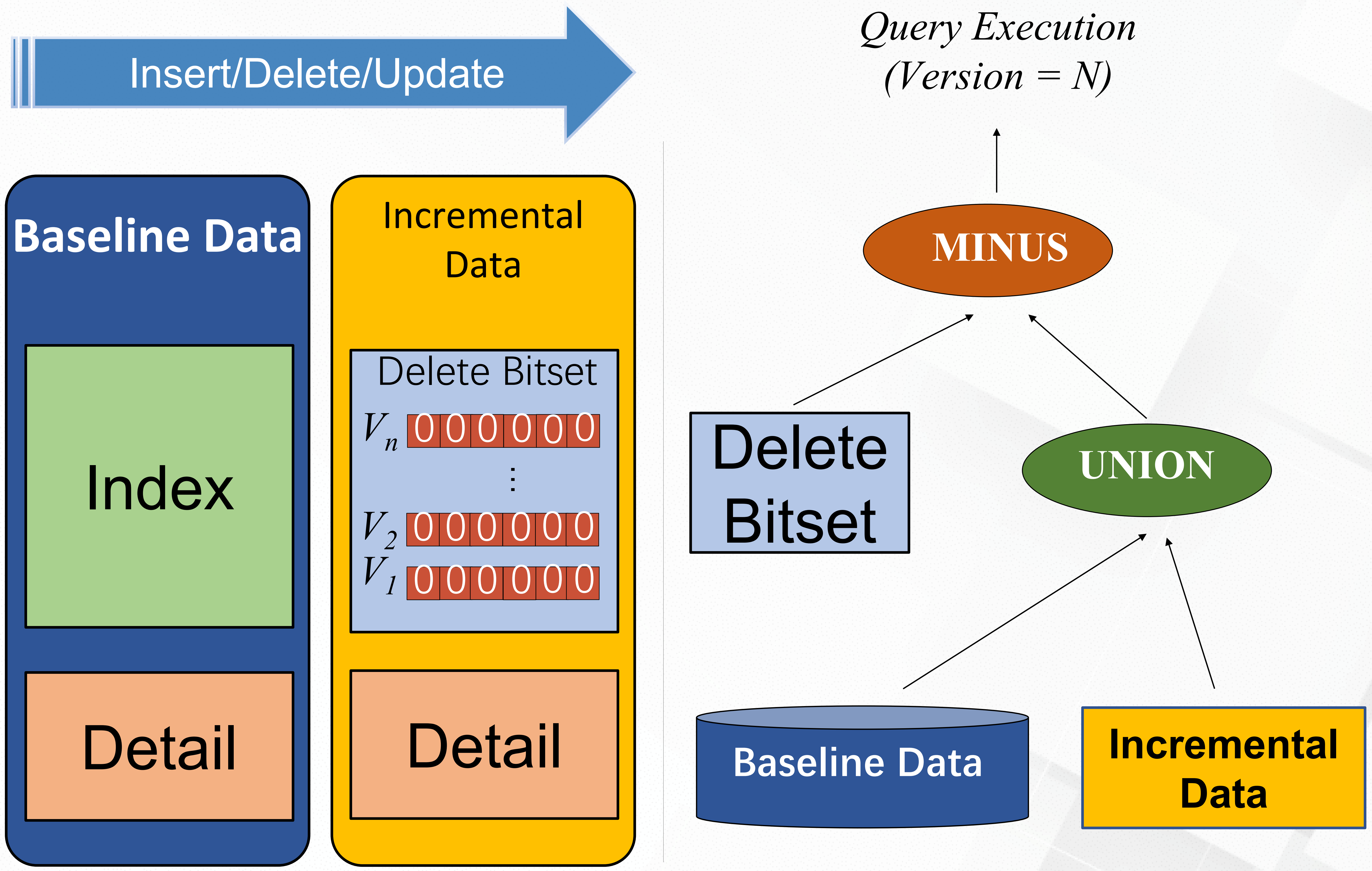
- Hash + List two level partitions for fact table
- Dimension table
- Partition pruning on both Hash and List partitions



# Lambda Storage



- Query is executed on both baseline data and incremental data
- Multi versioned delete bitsets are used to support snapshot read and delete/update.
- Baseline and incremental data are merged in background.

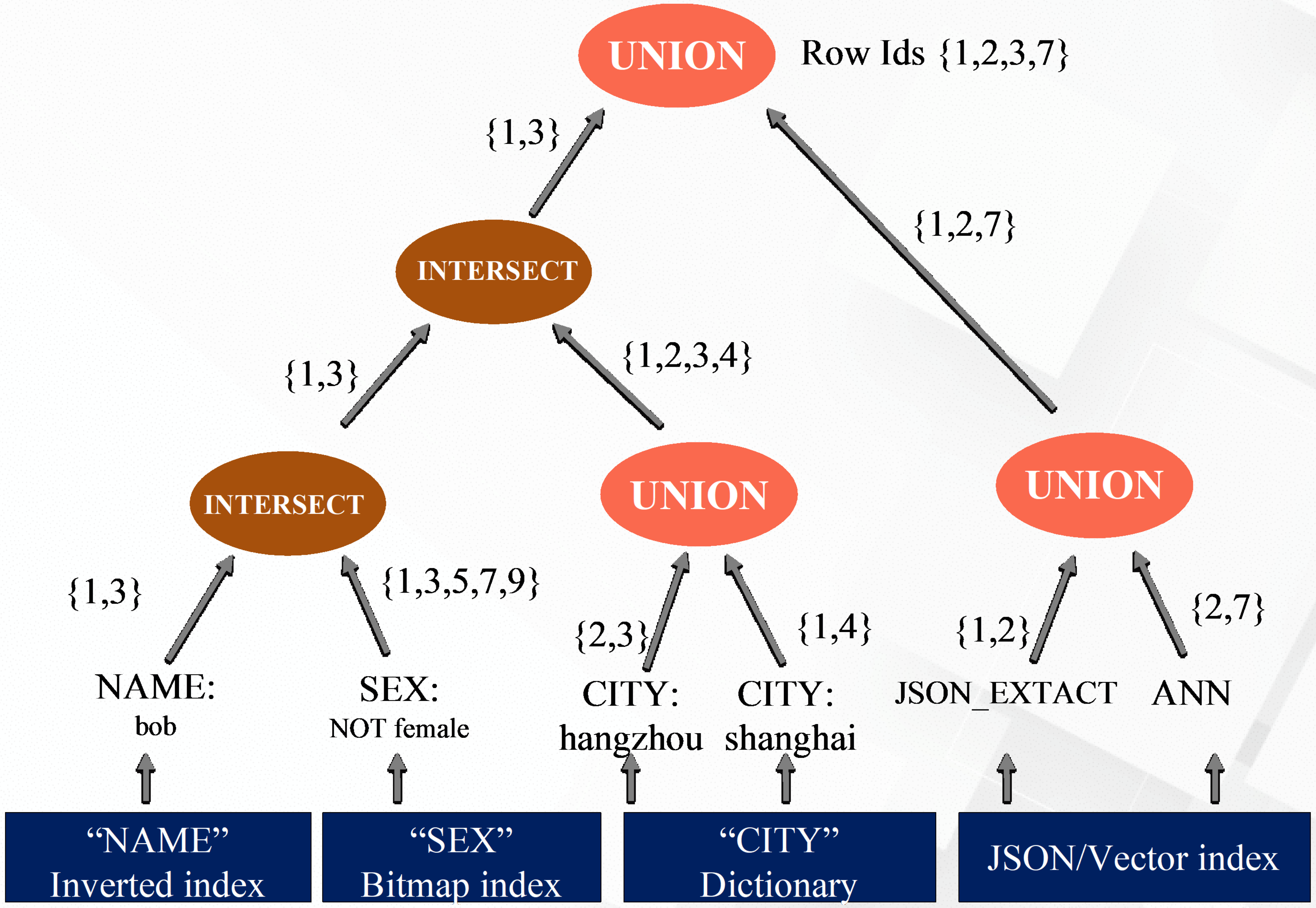


# Index for computing

Alibaba Cloud | Worldwide Cloud Services Partner

- Indexes on all columns
- High performance for ad-hoc queries
- K-ways merge for row-id sets.
- Runtime indexes selection
- Support complex-typed data
- Full text and JSON index

SELECT ... From t WHERE
( NAME='Bob' AND (CITY = 'Hangzhou' OR CITY = 'Shanghai') AND
( SEX != Female ) ) OR ( JSON_EXTACT(ATTR,'time') > 0
OR ANN(VEC, [1,1,1,1], 2) )

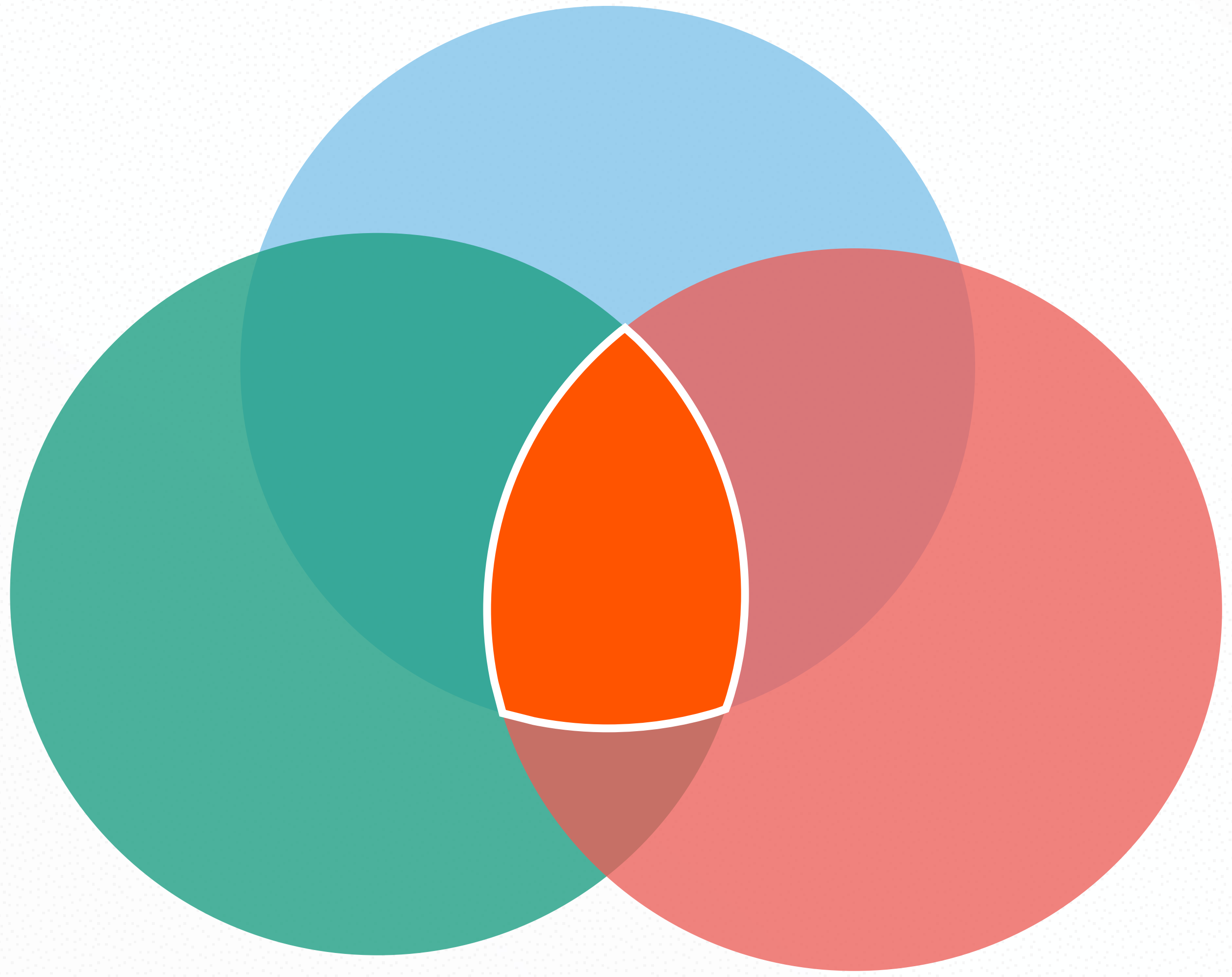


# Hybrid row-column store

Alibaba Cloud | Worldwide Cloud Services Partner

## Multi-dimensional analysis

- ✓ Any column Join
- ✓ Complex long computing tasks, ETL

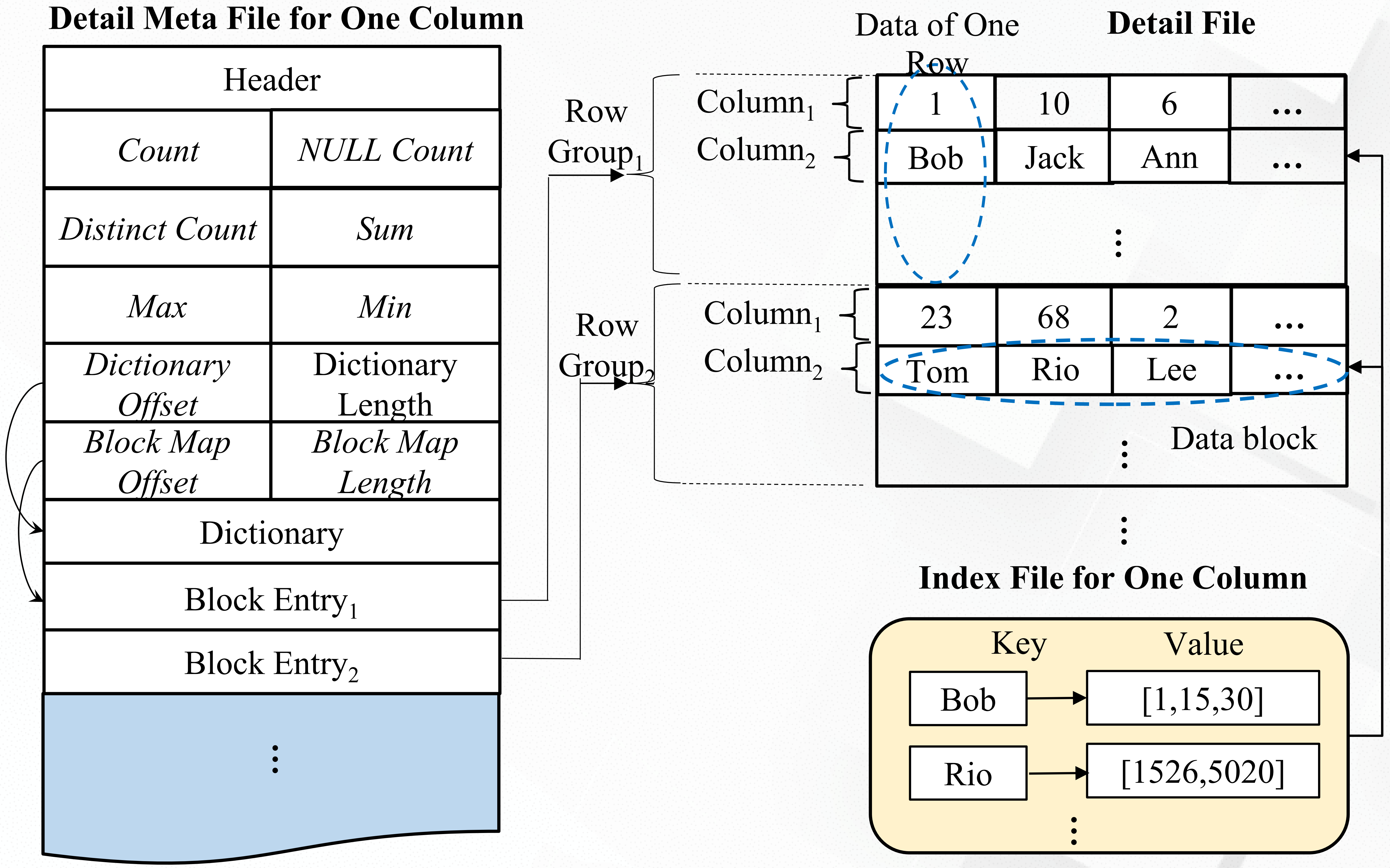


## Detail Query

- ✓ 1000+ columns extra wide table
- ✓ Semi-structured, large fields (JSON/ARRAY, etc.)

## Real-time RW

- ✓ Live updates
- ✓ 6 million TPS
- ✓ 10000+ QPS



# Real Workload Evaluation

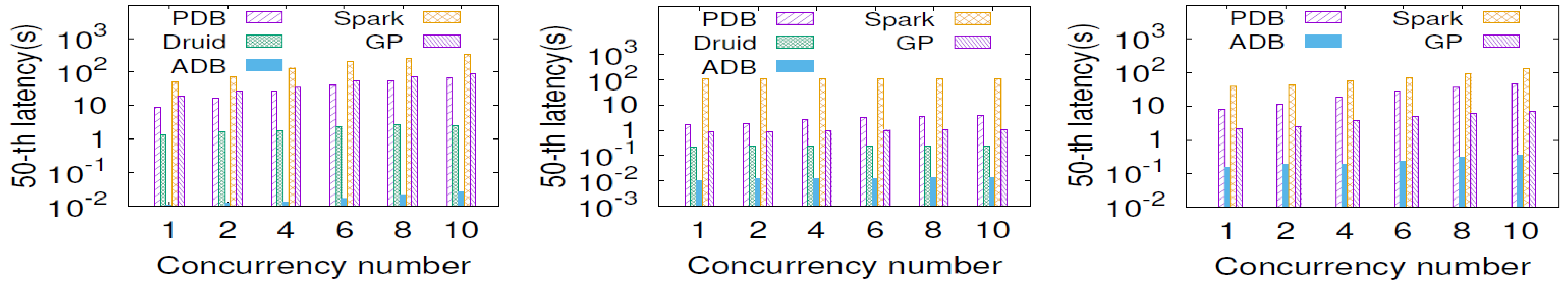
Alibaba Cloud | Worldwide Cloud Services Partner

| Query Type | Query |
| --- | --- |
| **Full Scan (Q1)** | SELECT * FROM orders ORDER BY o trade time LIMIT 10 |
| **Point Lookup (Q2)** | SELECT * FROM orders WHERE o trade time BETWEEN '2018-11-13 15:15:21' AND '2018-11-13 16:15:21' AND o trade prize BETWEEN 50 AND 60 AND o seller id=9999 LIMIT 1000 |
| **Multi-table Join (Q3)** | SELECT o seller id, SUM(o trade prize) AS c FROM orders JOIN user ON orders.o user id = user.u id WHERE u age=10 AND o trade time BETWEEN '2018-11-13 15:15:21' AND '2018-11-13 16:15:21' GROUP BY o seller id ORDER BY c DESC LIMIT 10; |

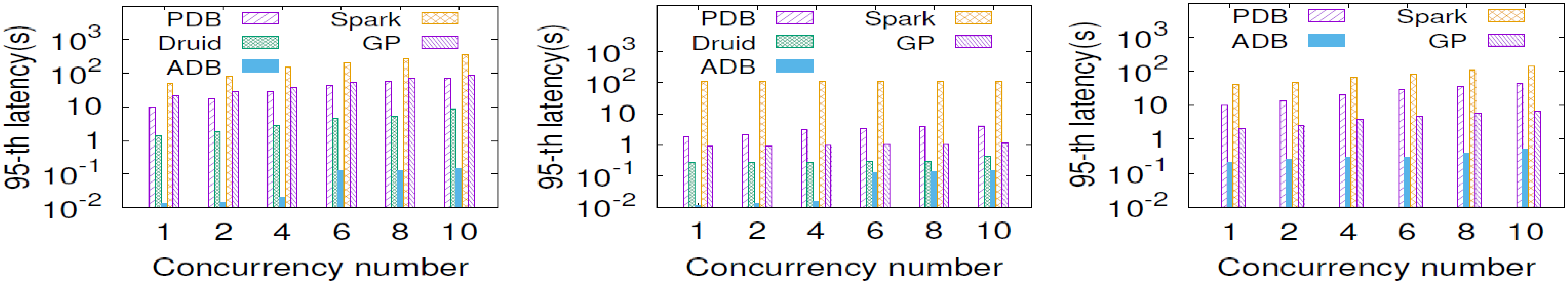
- Eight Physical Machines
- Intel Xeon Platinum 8163 CPU, @ 2.5 GHz
- 300GB main memory and 3TB SSD
- 10Gbps Ethernet network
    - 4 coordinators
    - 4 write nodes, and 32 read nodes
- Workloads
    - 1TB and 10 TB

# Comparison






(a) Q1 (b) Q2 (c) Q3

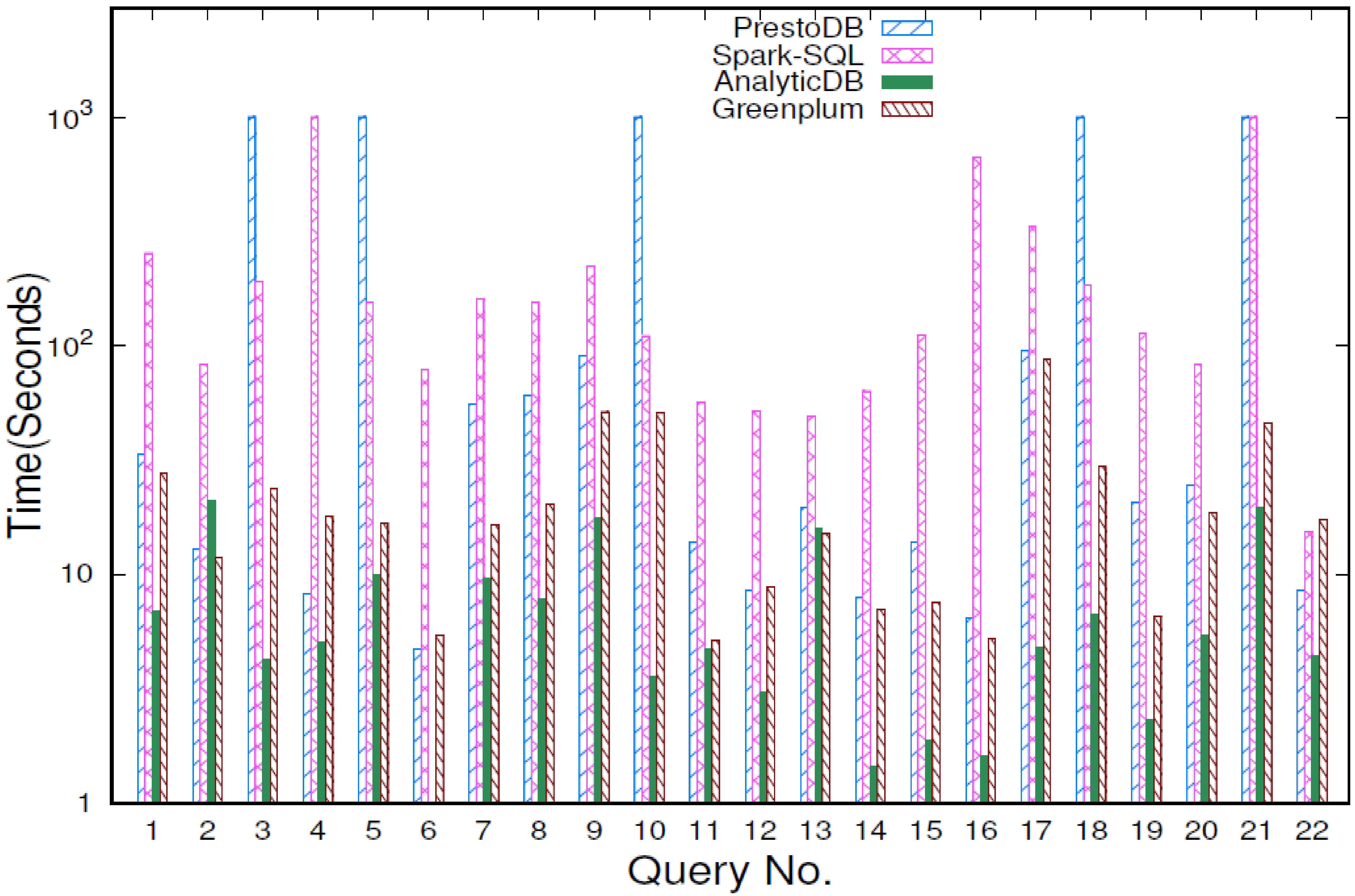



(a) Q1 (b) Q2 (c) Q3

# TPC-H Evaluation



- Pipeline-process
- All-column index
- Hybrid row-column storage
- Runtime cost-based index path selection
- K-ways merging and composite predicates pushdown
- Vectorized execution engine and optimized codeGen



# Outline



- Background
- POLARDB
- AnalyticDB
- **Self-Driving Database Platform**
- Conclusion

# Components of SDDP

Alibaba Cloud | Worldwide Cloud Services Partner

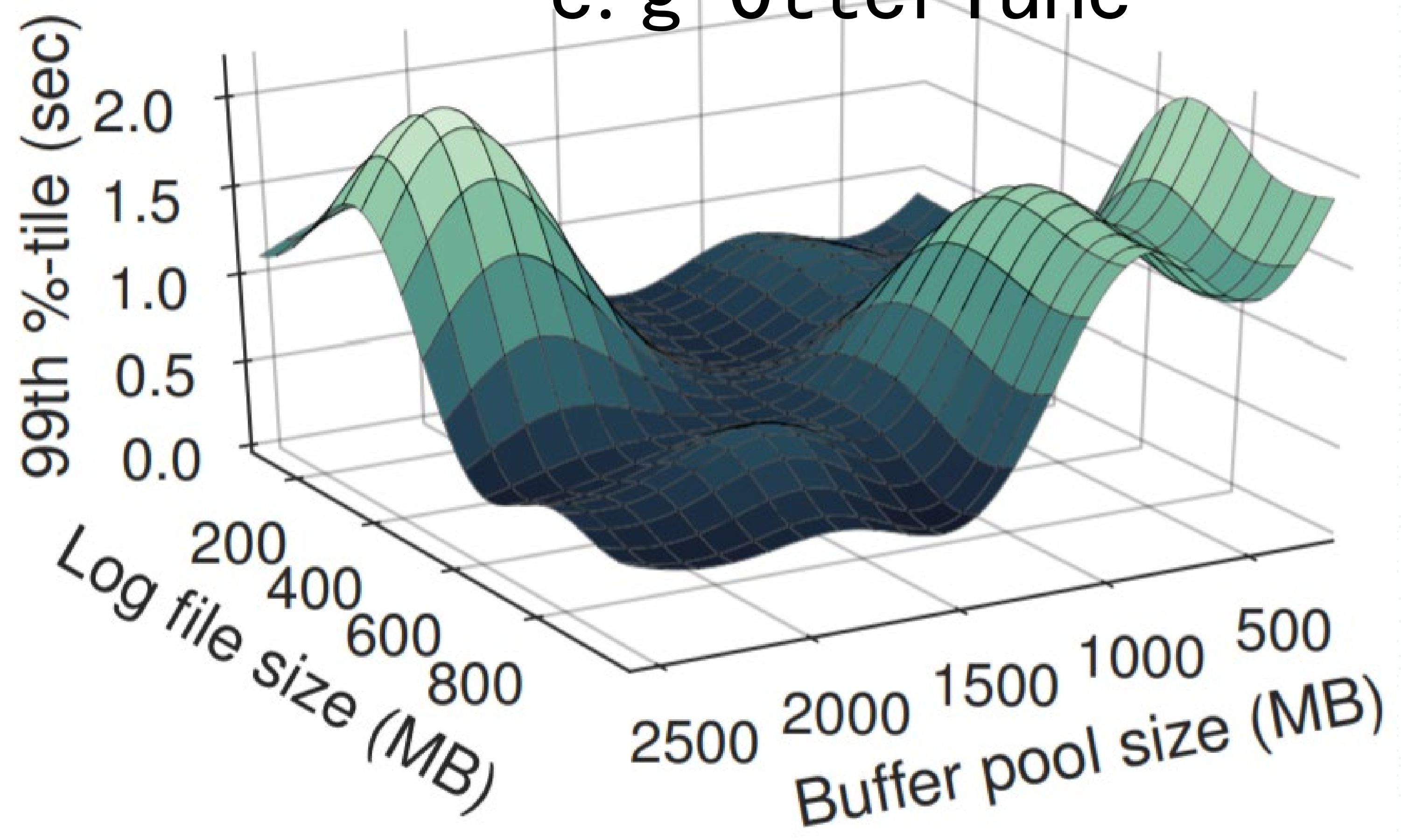

Alibaba Self-Driving Database Platform (SDDP) provides cloud databases with automatic operation and maintenance

# Key features of SDDP

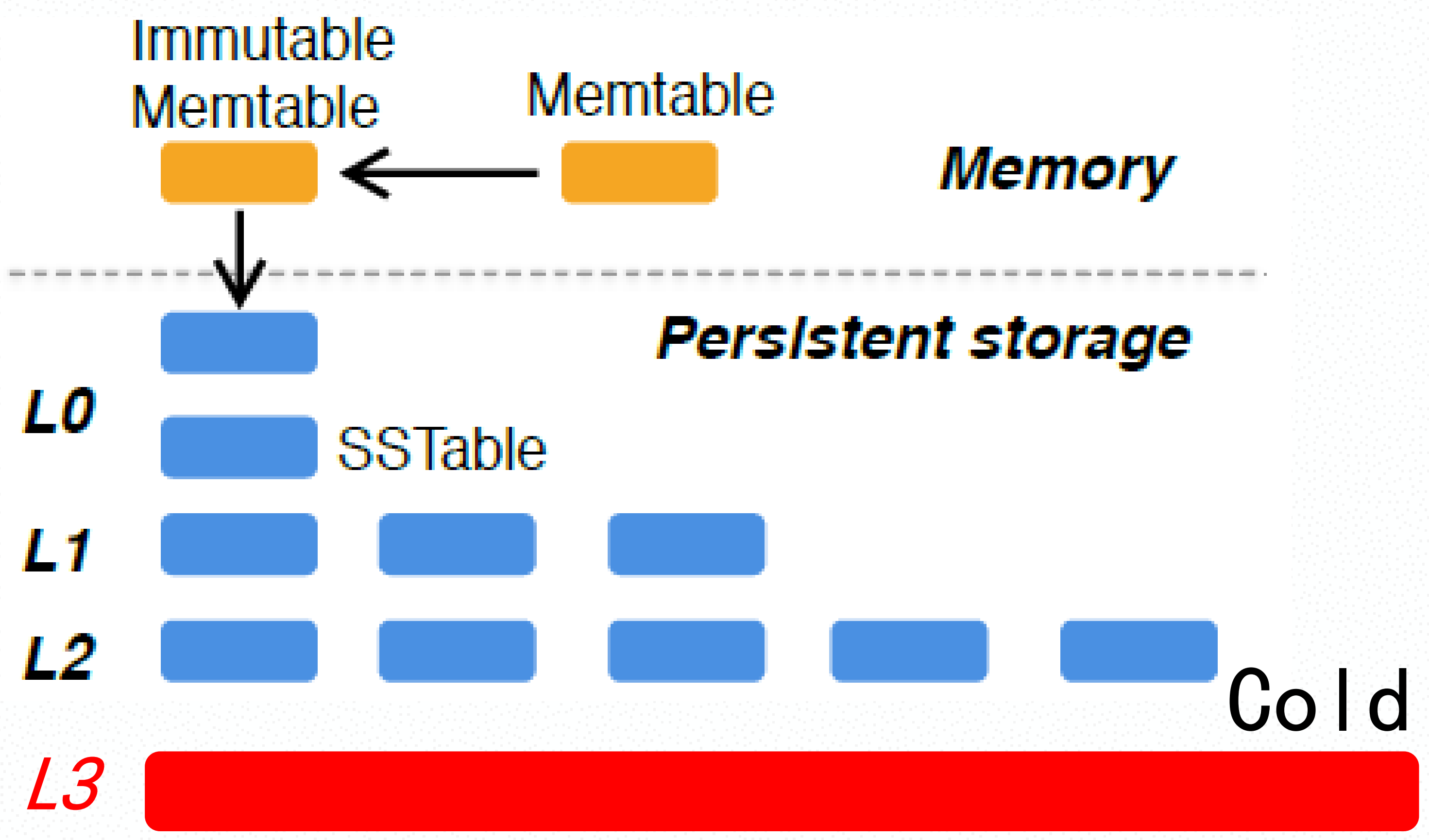


## Knobs tuning

e. g OtterTune




## Hot/cold separation




## ClouDBench

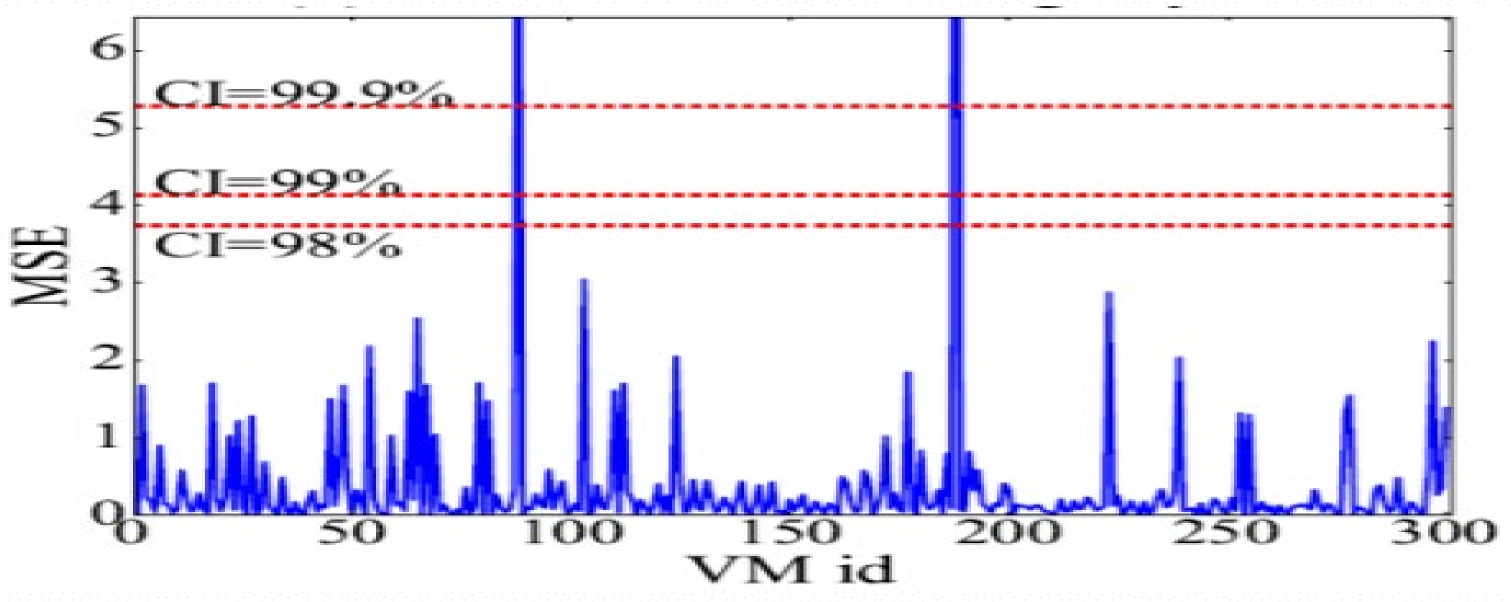
Workload Generator

Workload Replay

Select ··· where A = 1;
Select ... where B = 2;
Select ··· where A = 2;
Select ··· where A = 2;
Select ··· where B = 3;
Select ··· where A = 1;
Insert ...;
Insert ···;
Update ···;
Update ···;
···

## Anomaly detection

Slow SQL → Throttling




## NL2SQL

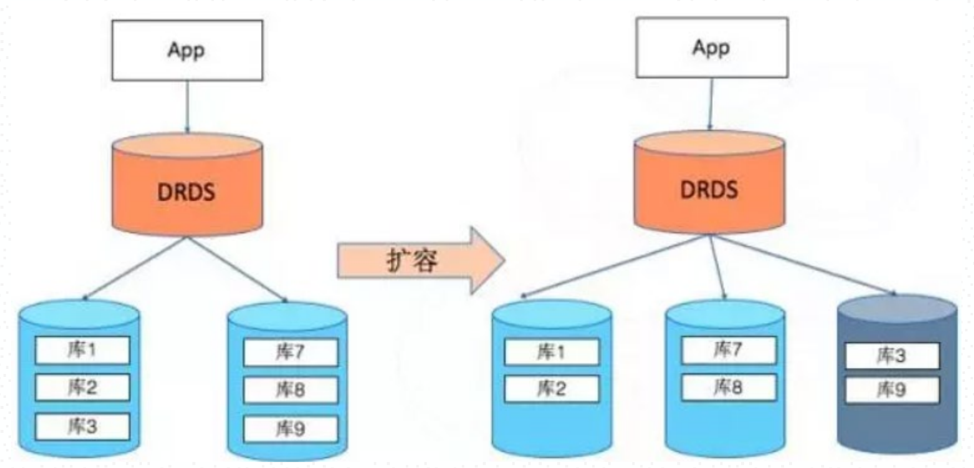
Human Language

SQL

DBMS

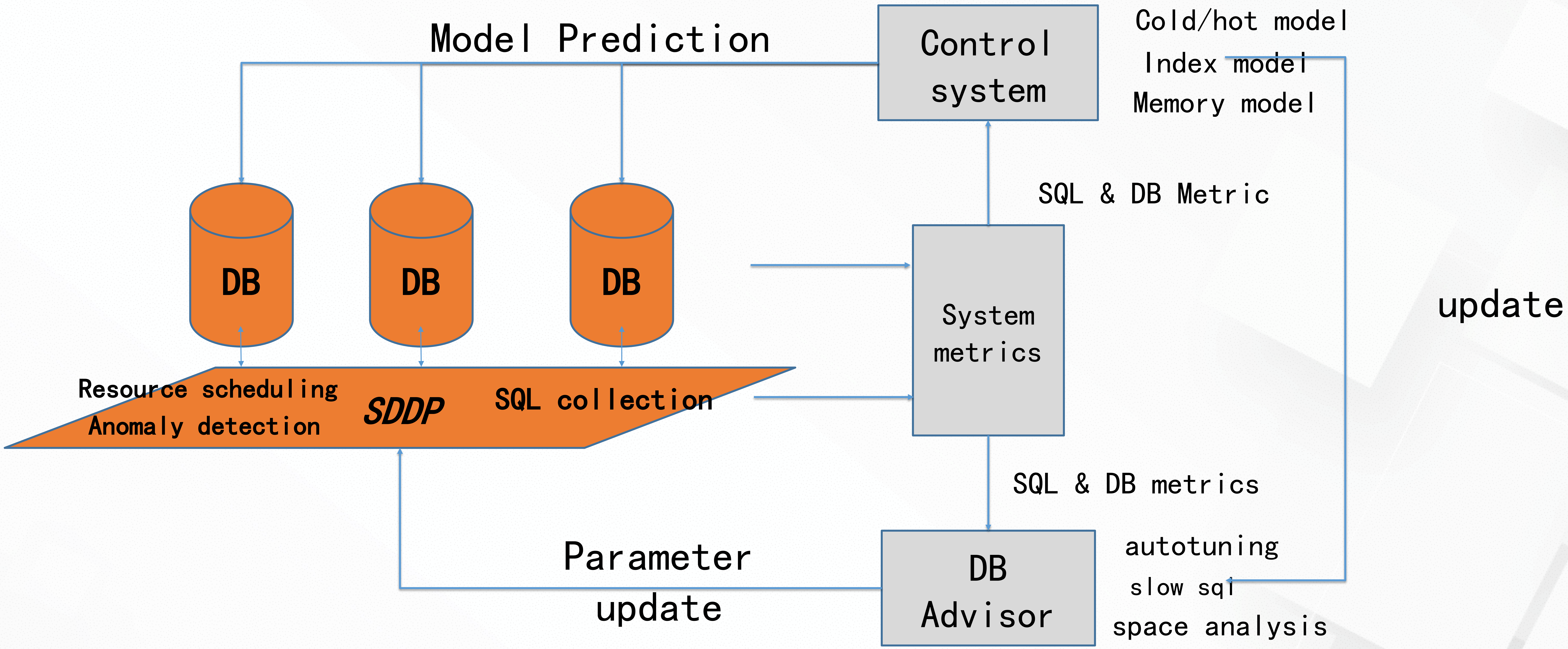

## DB Design recommendation

Auto Index/sharding



# SDDP: Self-Driving Database Platform





# iBTune - individualized Buffer Tuning: Motivation



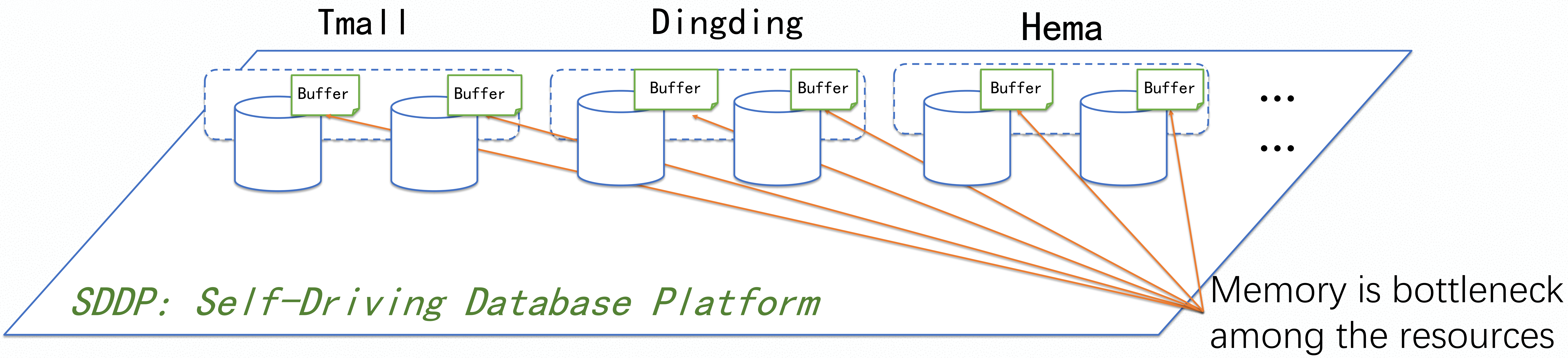



Table 1: Usage of different memory pools

| Memory Pool | buffer pool | insert buffer | log buffer | join buffer | key buffer | read buffer | sort buffer |
|---|---|---|---|---|---|---|---|
| Avg. Size | 29609.98M | 8.00M | 200.00M | 0.13M | 8.00M | 0.13M | 1.25M |
| Percent | 99.27% | 0.03% | 0.67% | 0.00% | 0.03% | 0.00% | 0.00% |

The memory uses at Alibaba product environment

Buffer pool is the largest memory consumer

# iBTune - Motivation

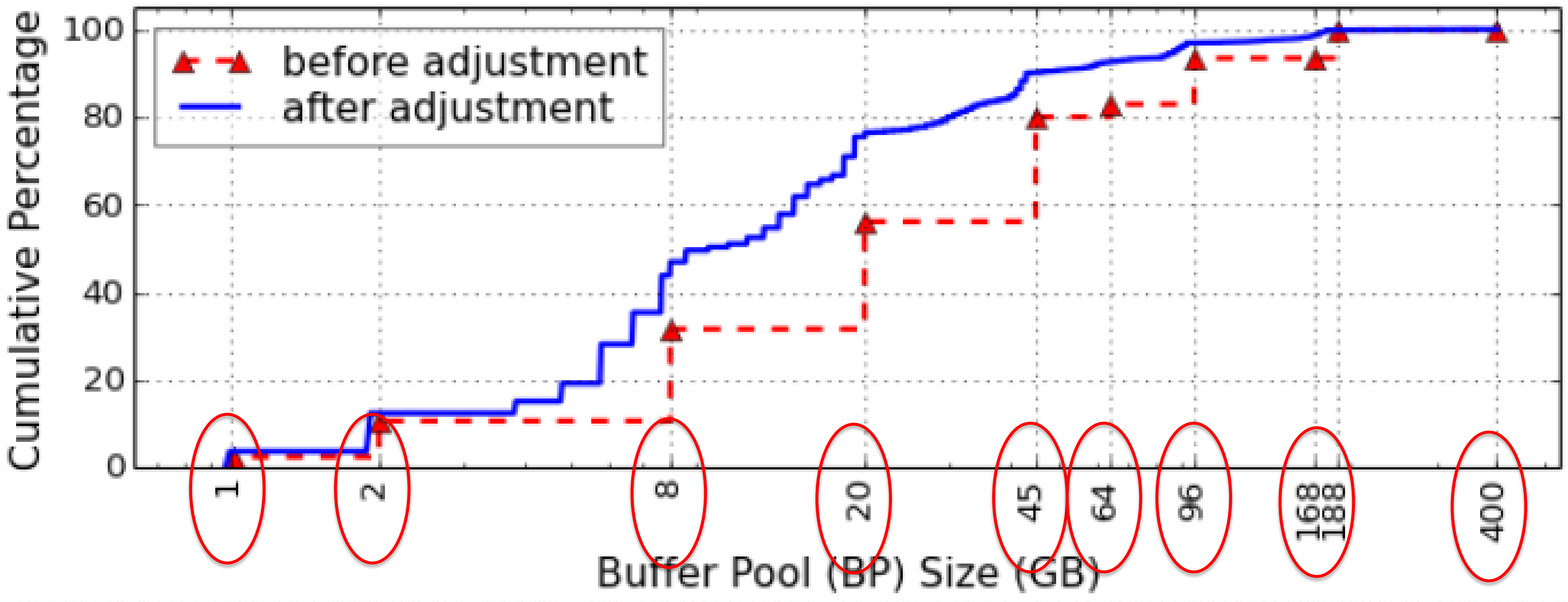
Reduce memory (buffer pool) while guaranteeing SLA (response time).

- DBA manually uses a small number of BP sizes (10 configurations in our case).

- Each instance's BP size might be different as the query workload is different.

- Manual tuning is not scalable for large cloud databases since each instance has different BP size.

CDF of individual BP sizes
before and after the iBTune applies




**iBTune: Individualized Buffer Tuning for Largescale Cloud Databases**

# iBTune – High level idea

Alibaba Cloud | Worldwide Cloud Services Partner

Practical function $\frac{\log(mr_{target}) - \log(mr_{cur})}{\log(bp_{bptarget}) - \log(bp_{cur})} \approx -\alpha_i.$

**tolerate miss ratio (t_miss_ratio)**

DB instance

F1(t_miss_ratio) = New BP size

✓ Apply new BP size

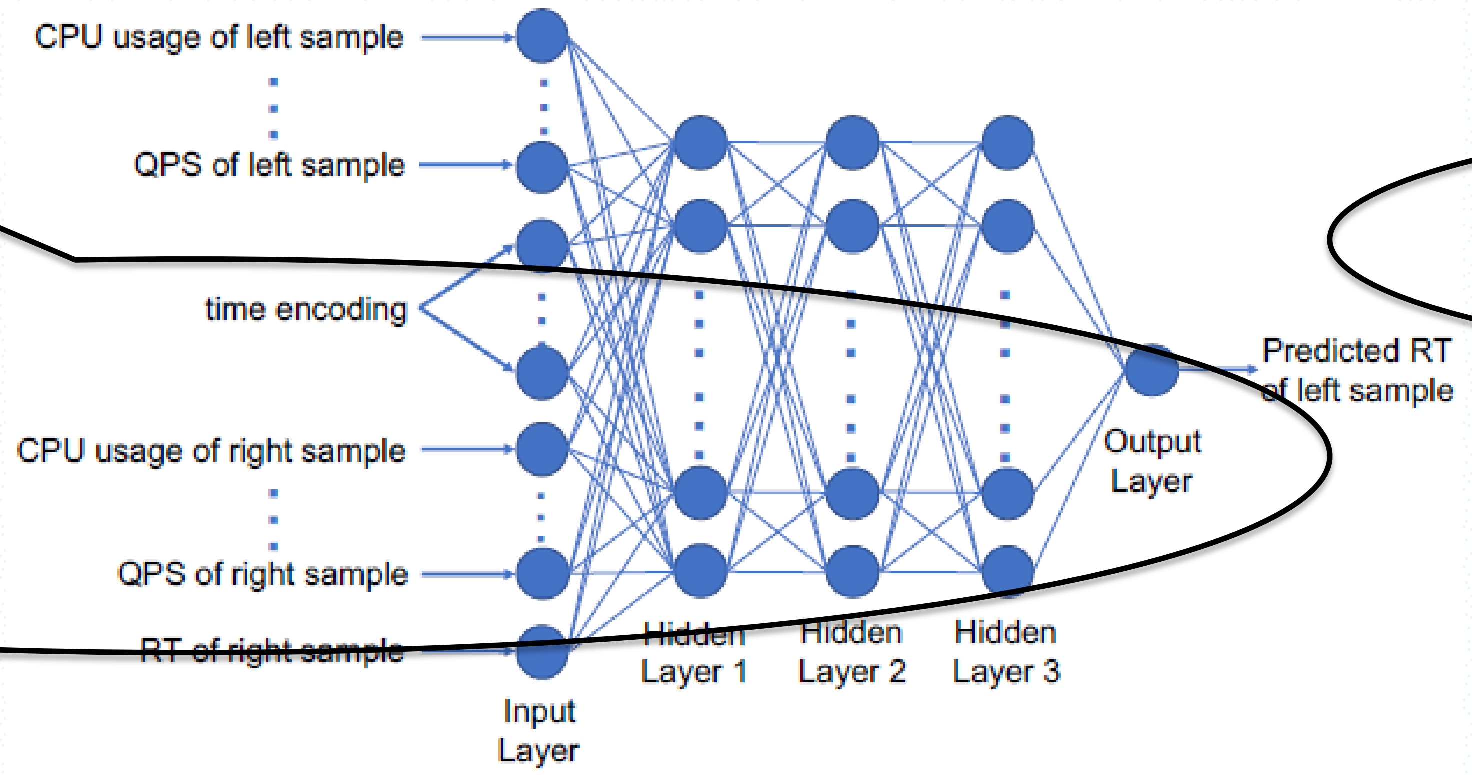
DB instance

F2(t_miss__ratio) = Response time

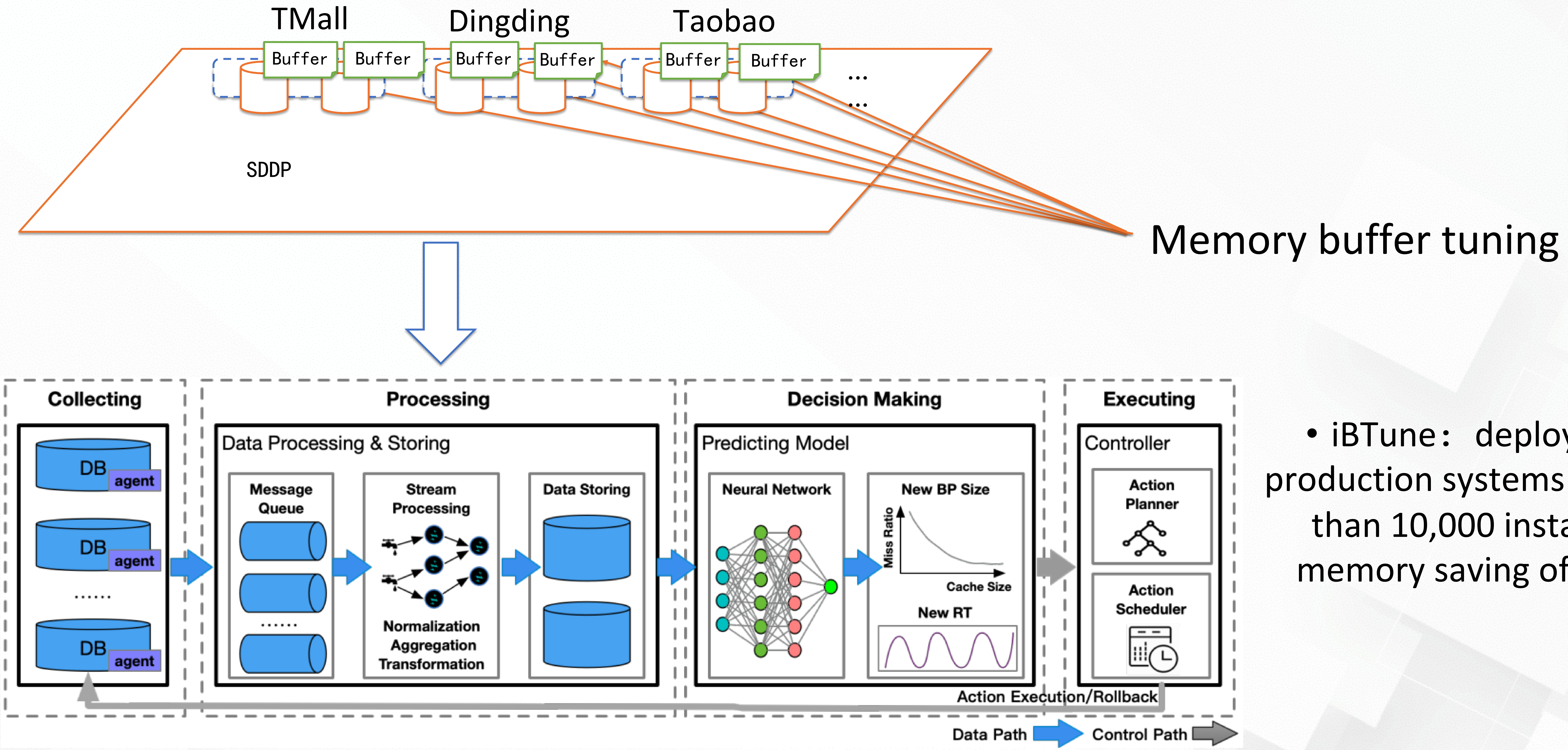
< Safe Response time (SLA)

Guarantee SLA

Pairwise DNN



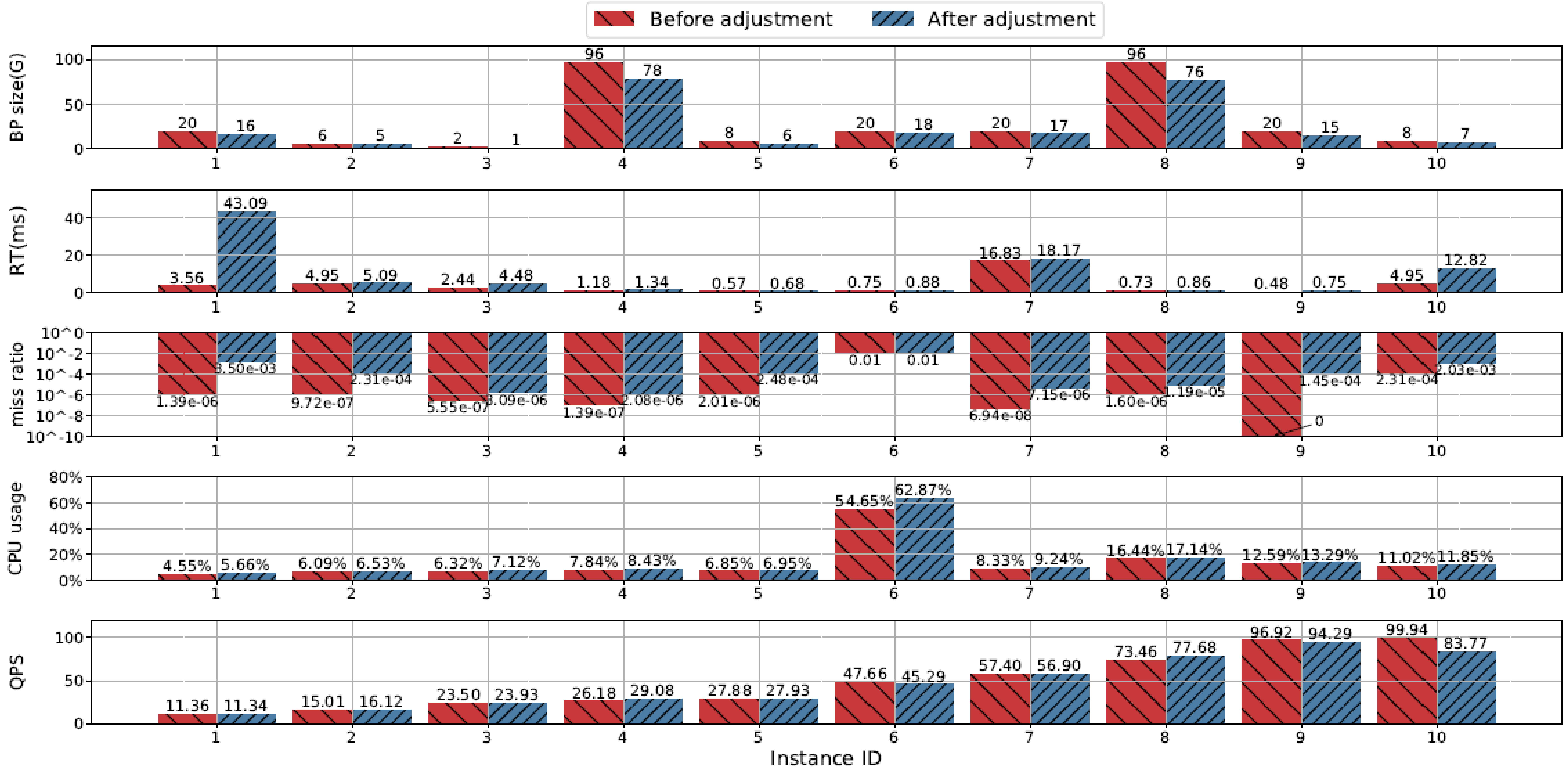
# iBTune: individualized Buffer Tuning (VLDB 2019)




- iBTune： deployed on production systems with more than 10,000 instances, memory saving of >20TB

# Sample Results

Alibaba Cloud | Worldwide Cloud Services Partner




10 representative instances. The memory saving ranges from 50% to 10%, which strongly supports that a single number does not fit all. Instance 1 has a large increase in RT after the adjustment. We find that there is one query that consumes 99.97% of the total response time. The lookup value in WHERE condition changes for this query.

# Outline

Alibaba Cloud | Worldwide Cloud Services Partner

- Background
- POLARDB
- AnalyticDB
- Self-Driving Database Platform
- **Conclusion**

# Recent publications



- [SIGMOD'19] X-Engine: An Optimized Storage Engine for Large-scale E-Commerce Transaction Processing. http://sigmod2019.org/sigmod_industry_list
- [SIGMOD'18] TcpRT: Instrument and Diagnostic Analysis System for Service Quality of Cloud Databases at Massive Scale in Real-time. https://sigmod2018.org/sigmod_industrial_list.shtml
- [VLDB'18] PolarFS: an ultra-low latency and failure resilient distributed file system for shared storage cloud database. www.vldb.org/pvldb/vol11/p1849-cao.pdf
- [VLDB'14] Realization of the Low Cost and High Performance MySQL Cloud Database. www.vldb.org/pvldb/vol7/p1742-alibaba.pdf
- [VLDB'19] AnalyticDB: Real-time OLAP Database System at Alibaba Cloud. https://www.vldb.org/2019/?papers-industrial
- [VLDB'19] iBTune: Individualized Buffer Tuning for Largescale Cloud Databases https://www.vldb.org/2019/?papers-industrial

# Conclusion and Future Work



- **Cloud-native: more elasticity, high availability, and excellent scalability**
- HTAP capability
- Multi-model
- New hardware support (software-hardware co-design)
- Security
- Self-driving databases

Thanks

Alibaba Cloud

# Database Requirement is Changing

## - Several Giant Business → A Wide Range of Enterprises

## - In China, 70% of uprising enterprises are hampered by such Data Challenges.

| | |
|---|---|
| **High Cost** | A million $ license? Professional engineers? |
| **Weak Ability** | Data Backup? Mining? Trouble Shooting? |
| **Data Explosive** | Hard to store, analyze, utilize. |

## - Everything is Online, so does Database

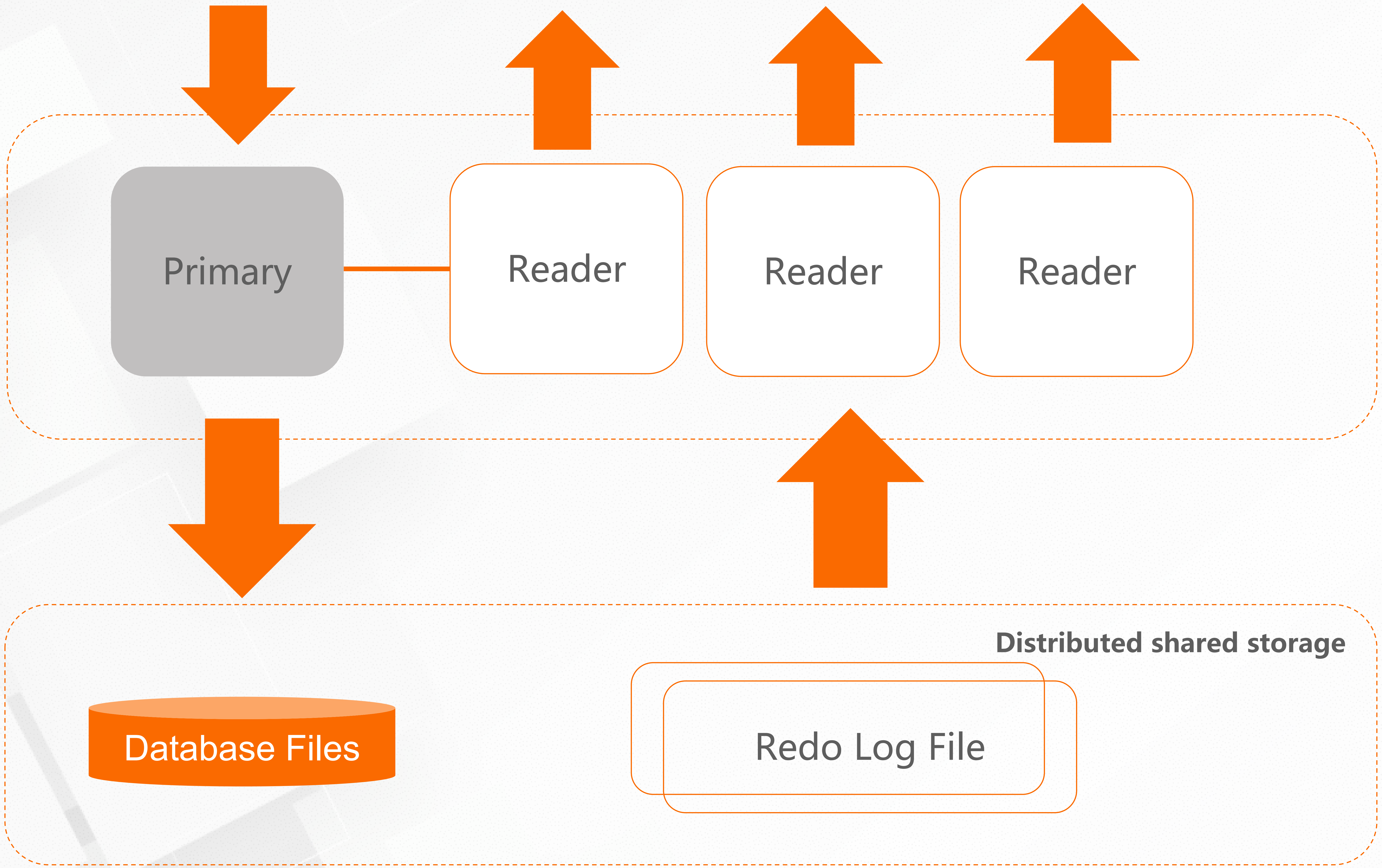
| Data | Computing | Application | Consumer |
|---|---|---|---|
| Generate | Training | Working | Global |
| Gather | Biding | Social | 7 X 24 |
| Extract | Anti-fraud | Entertainment | Digitalized |
| … | … | … | … |

# Cloud Database Features

- **More flexible**: marketing activity, hotspot → **Elastic**
- **More economical**: small profits, globalization → **Low Cost**
- **More efficient**: sensitive to customer loss → **Performance**
- **More agile**: time & opportunity is money → **Continuous Solution**

# POLARDB Architecture






**Excellent Elasticity**

**2core vCPU upgrades to 32core in < 5mins**
**2 nodes scales out to 4 nodes < 5mins**

**Cost reduction**

**Serverless on-demand billing**

# Failure Resilient of PolarFS -- ParallelRaft

## - Why develop a variant of Raft

**- At First Choose Raft**

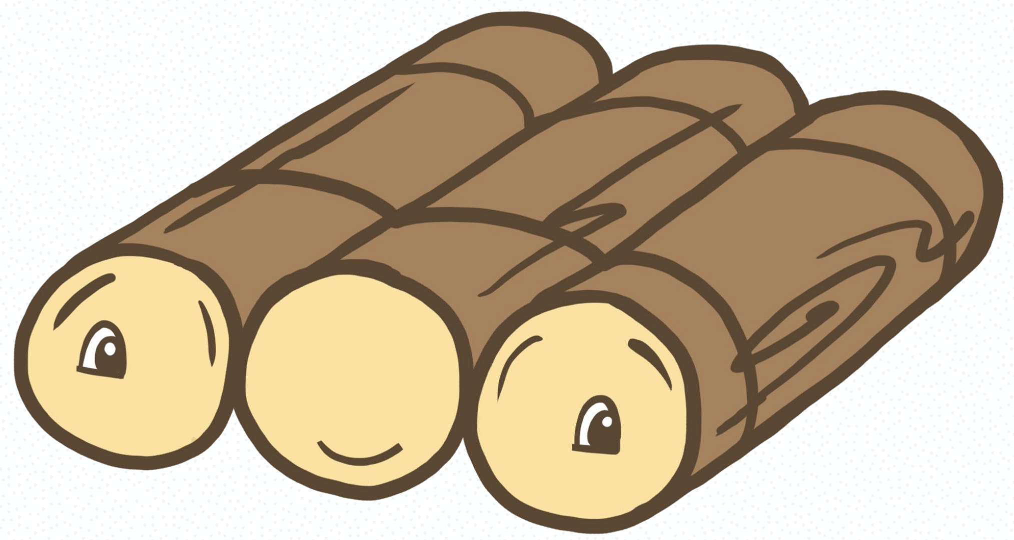
- For its implementation simplicity

**- Found not suitable for PolarFS**

- Test shows scale problem under our highly concurrent environment

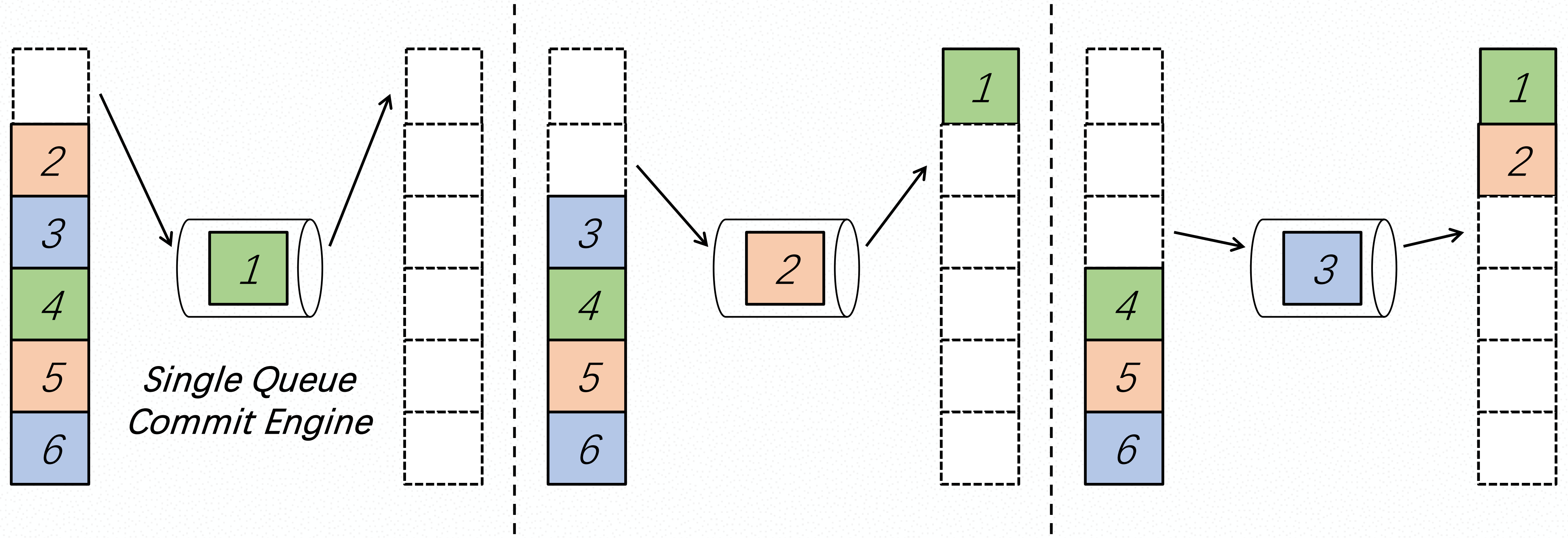
**- Sequential Commitment Limitation**

- May get stuck when some entry is slow



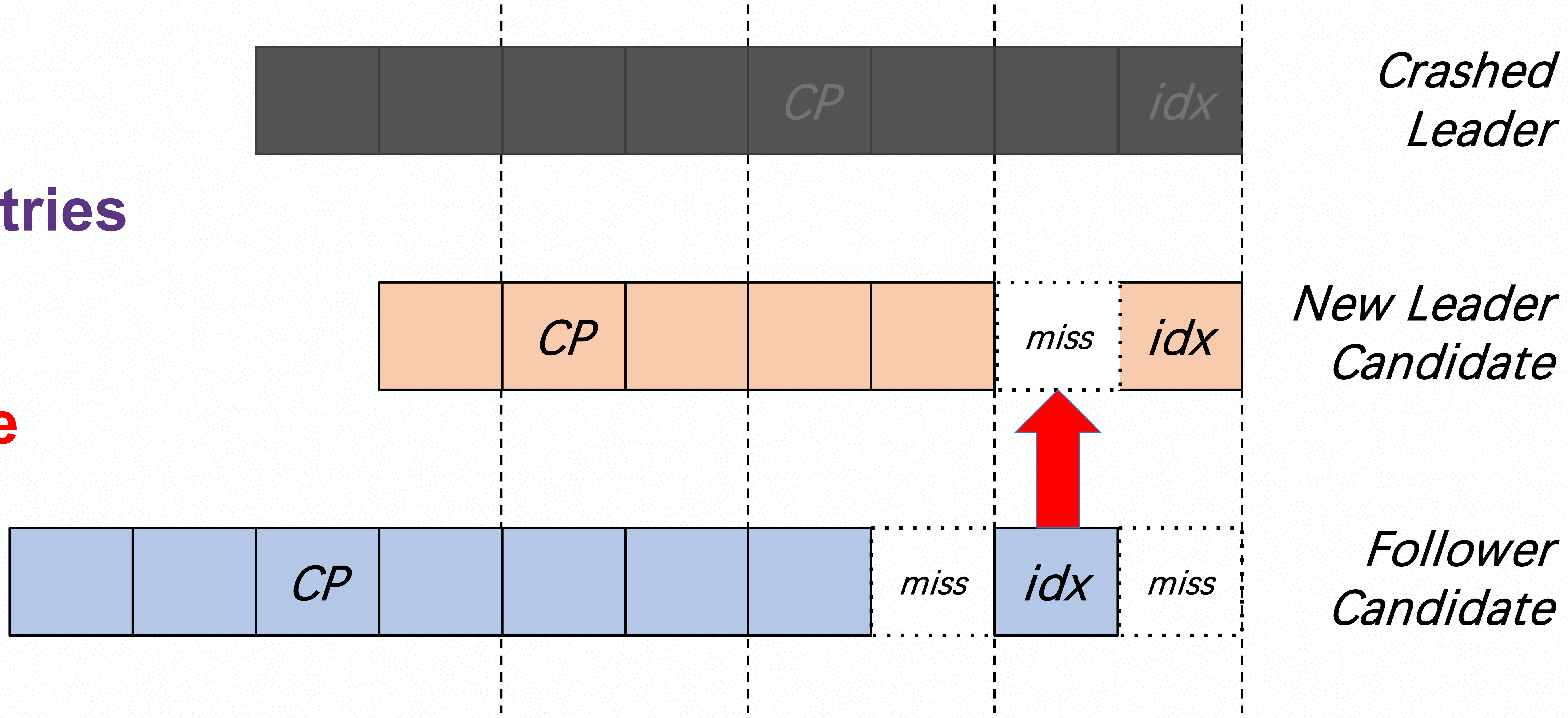
From: https://raft.github.io/

**Sequential Commitment**



# Failure Resilient of PolarFS -- ParallelRaft

## - Selecting a New Leader




**- Candidate might lack committed entries**

- Holes permitted in the log

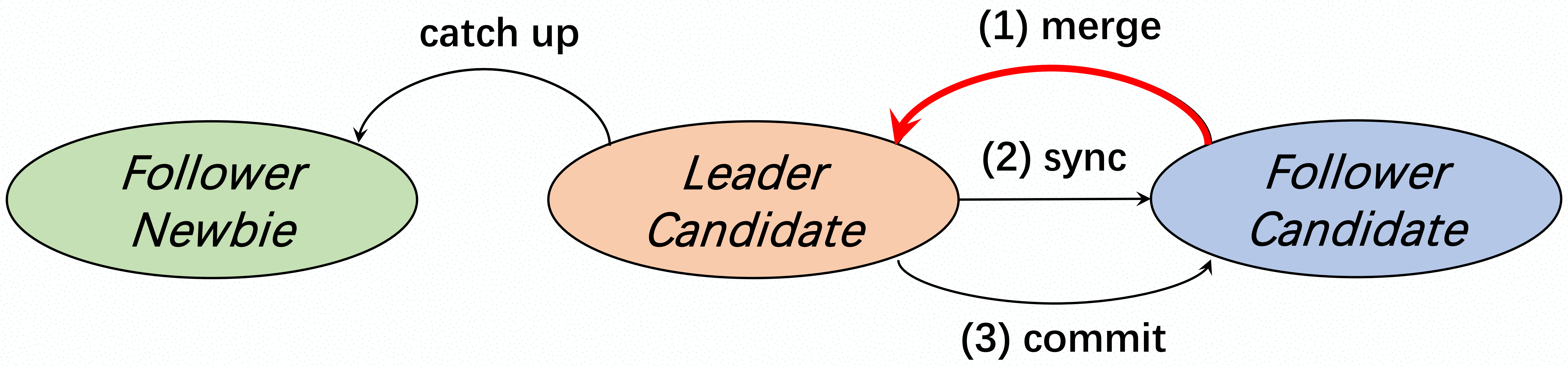
**- Introduce an additional merge stage**

- Guarantee Leader Completion Property

**- CP is CheckPoint**

- Logs old than CP are deletable

Other Details please refer the Paper



# Highlight

**- Storage can play an important role in Database Enhancement**

- QPS, TPS, Capacity, RO instance extension, Fast Backup

**- PolarFS delivers extreme performance with high reliability for Cloud Database**

- Adopting Optane, NVMe SSD, RDMA, OS-bypass and zero-copy techniques, ParallelRaft
- Allowing 3-replica write latency comparable to single replica on the local SSD

**- Sharing Storage Architecture, POSIX-like Userspace interfaces**

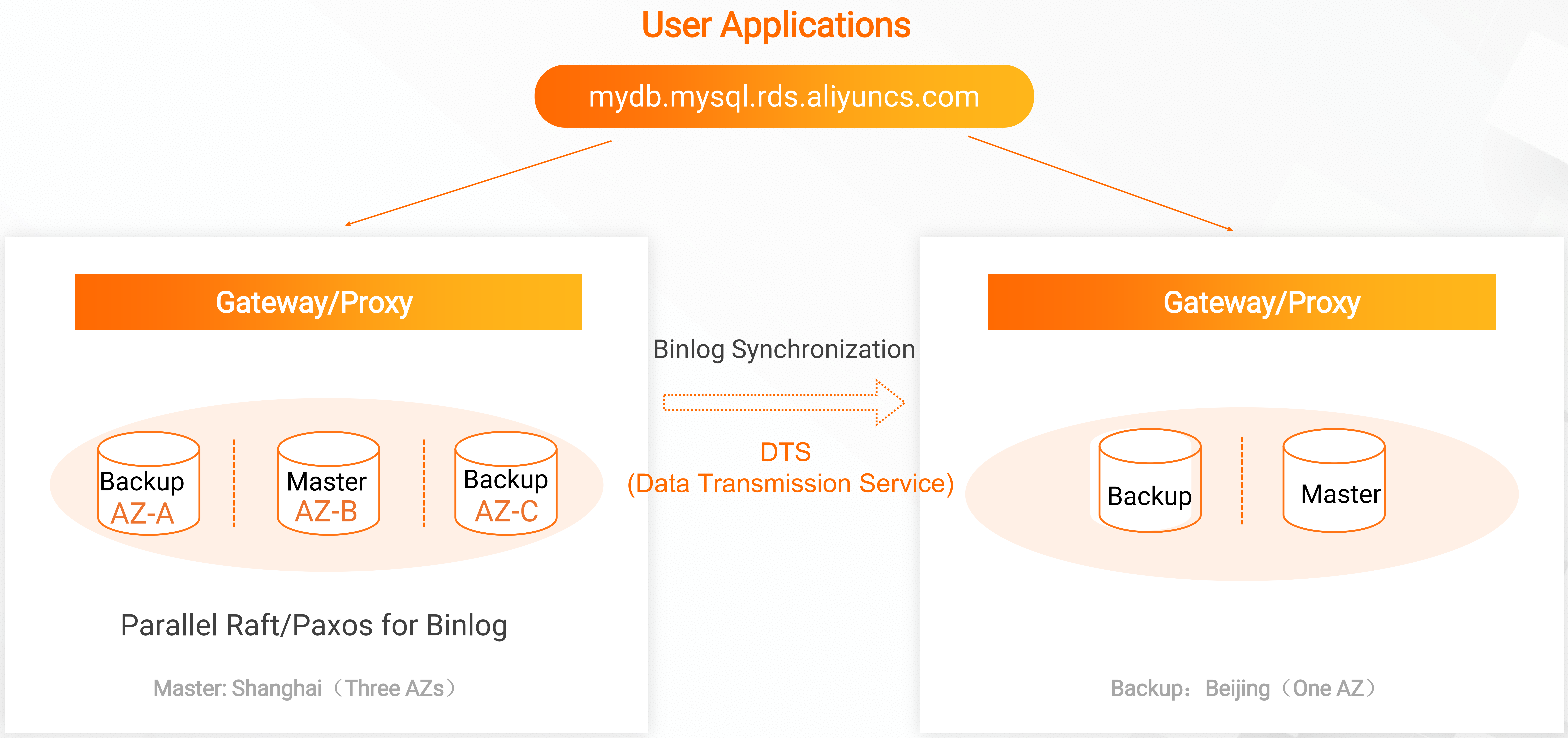
- Easily port DB: Minor modifications to get whole database database improvement

**- Future: New hardware and More DB optimizations**

- NVM, FPGA, etc.
- Parallel Query, SQL Offloading

# High Availability: Cross AZ-Cross Region

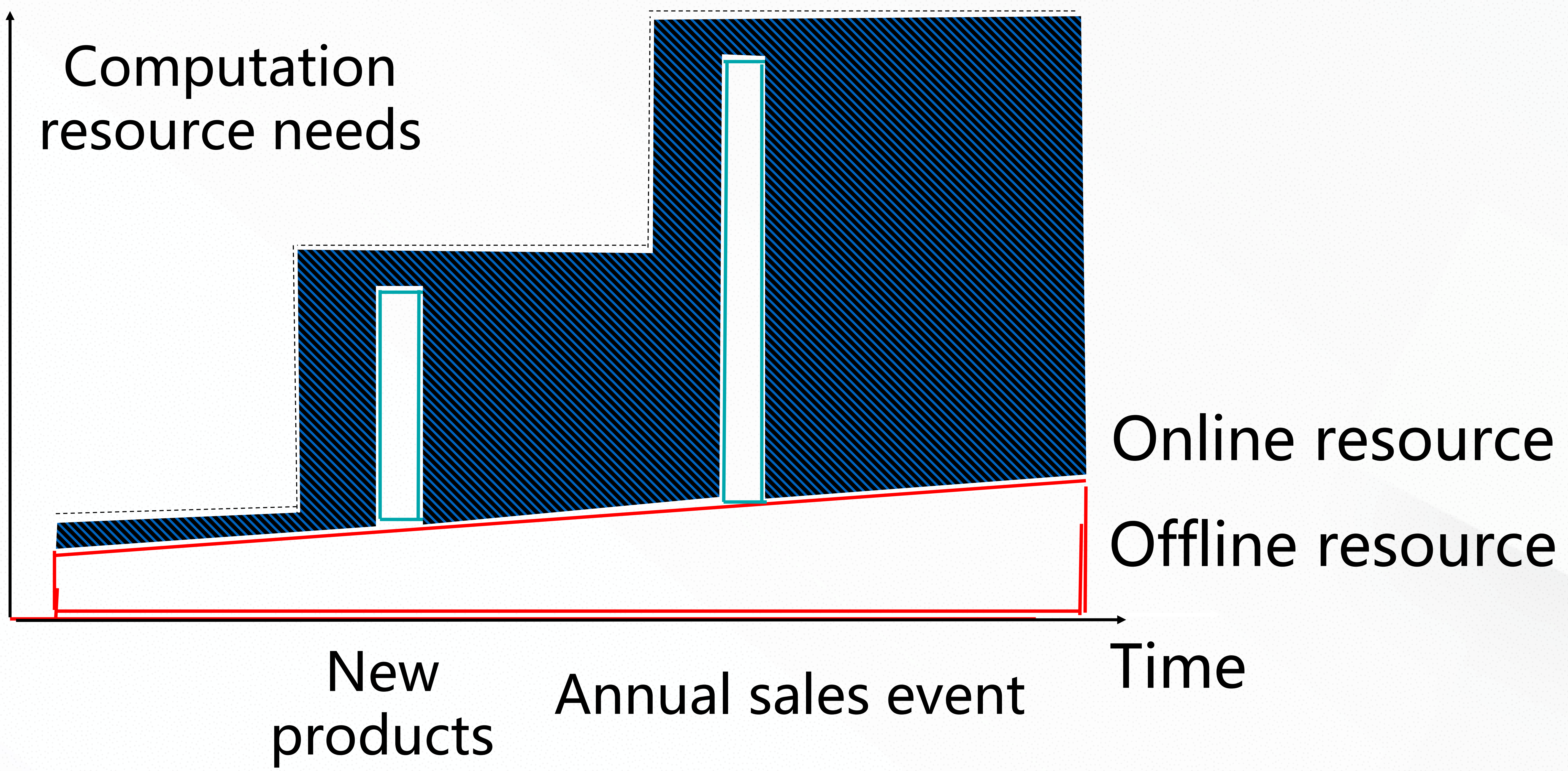




# Why Cloud Native?



**A typical use-case: on-demand usage and elastic billing**



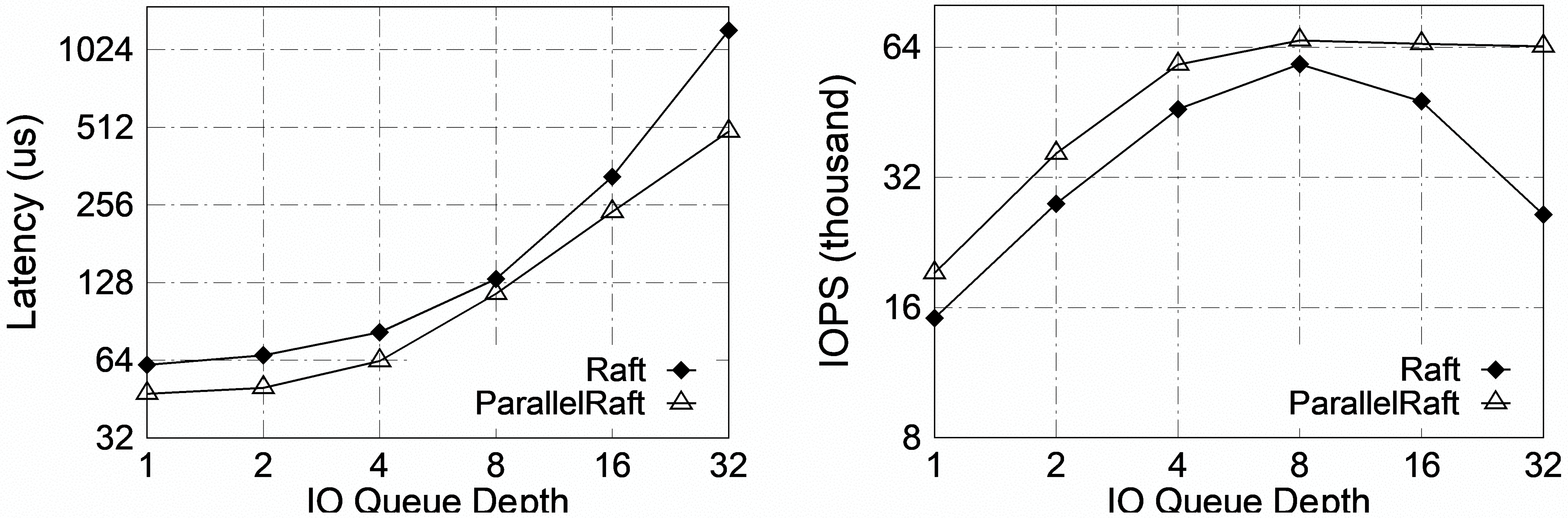

Yearly subscription---50% off

On demand usage and elastic billing – minute-level

# ParallelRaft vs. Raft

## - FIO Test in Our Concurrent Environment

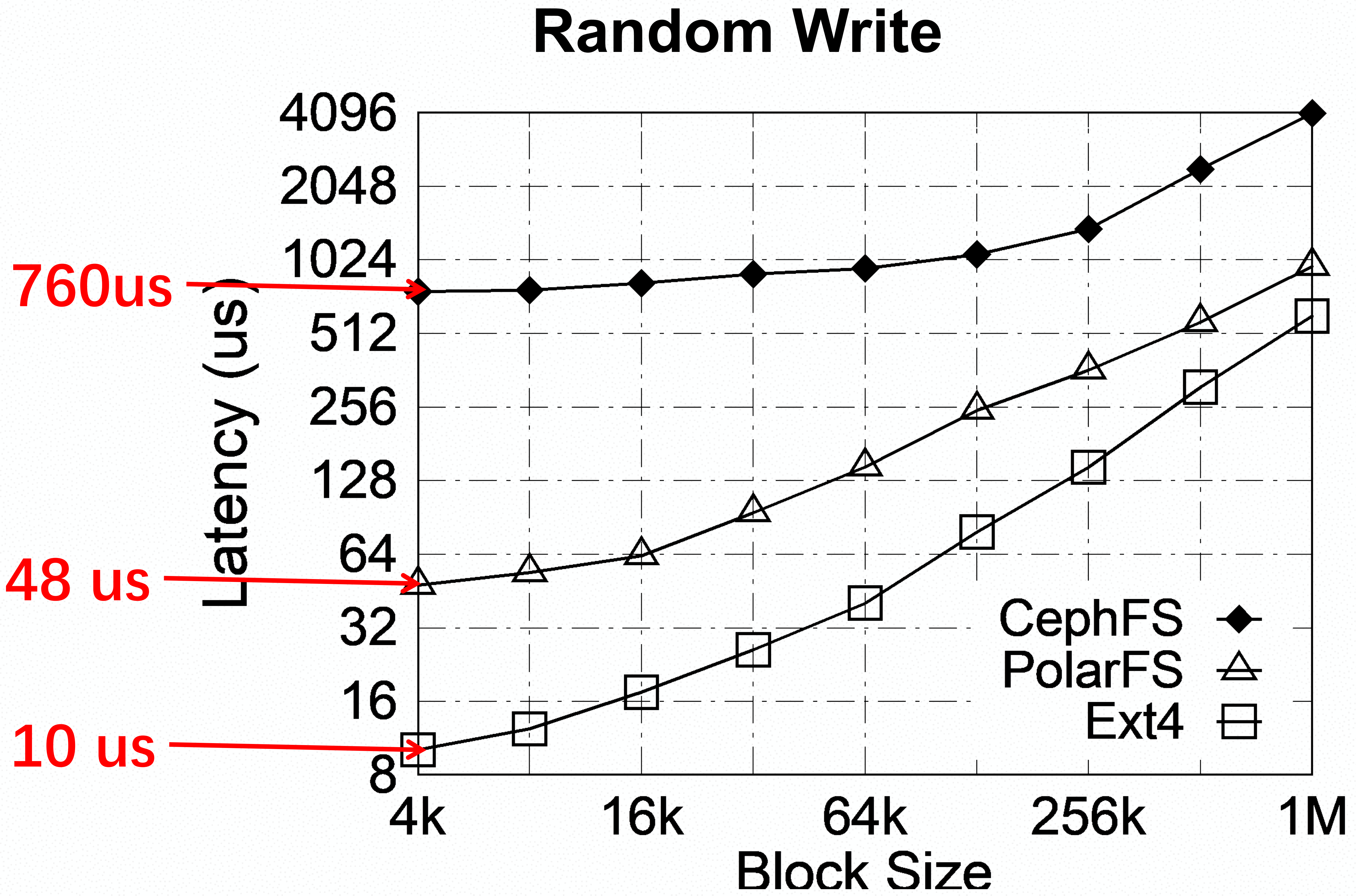
**- Performance Gap appear under large I/O depth**

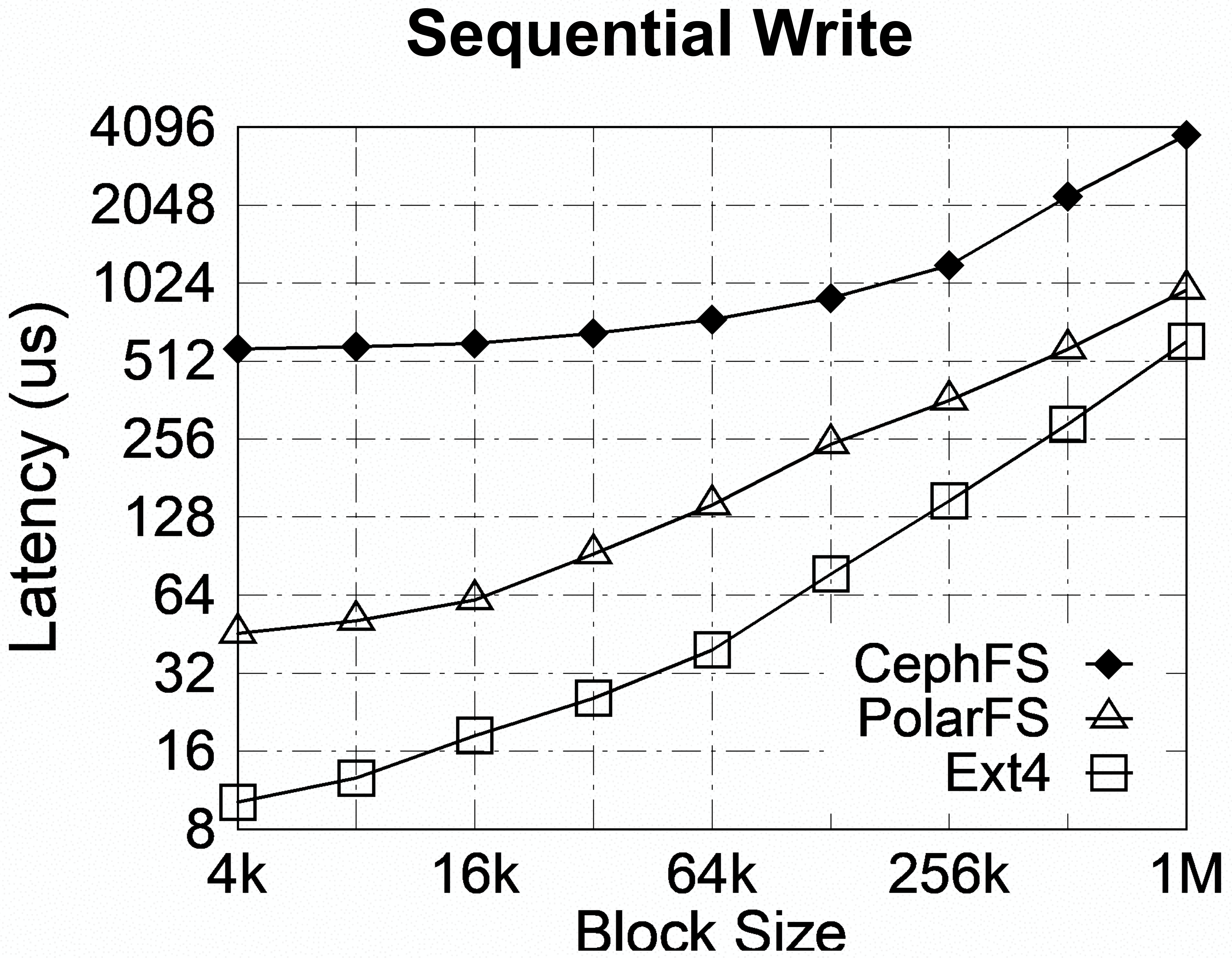



**The IOPS of Raft drops after I/O depth 8, ParallelRaft keep a steady high.**
**ParallelRaft helps PolarFS to get high performance under heavy workloads.**

# Evaluation – PolarFS Latency

- PolarFS and CephFS both have 3 replicas; Local Ext4 one replica



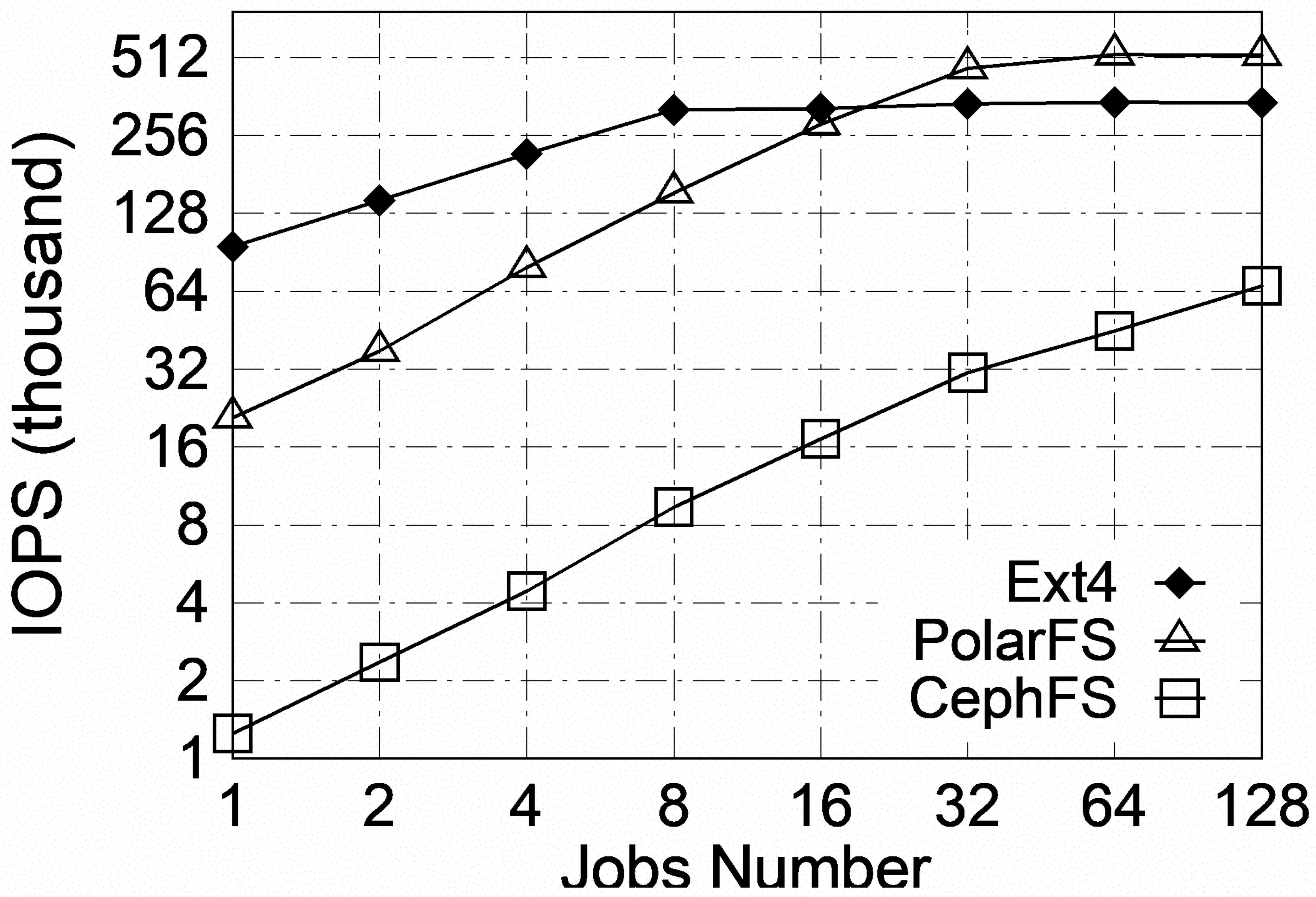




**The Average Write Latency Compared to Local Ext4**

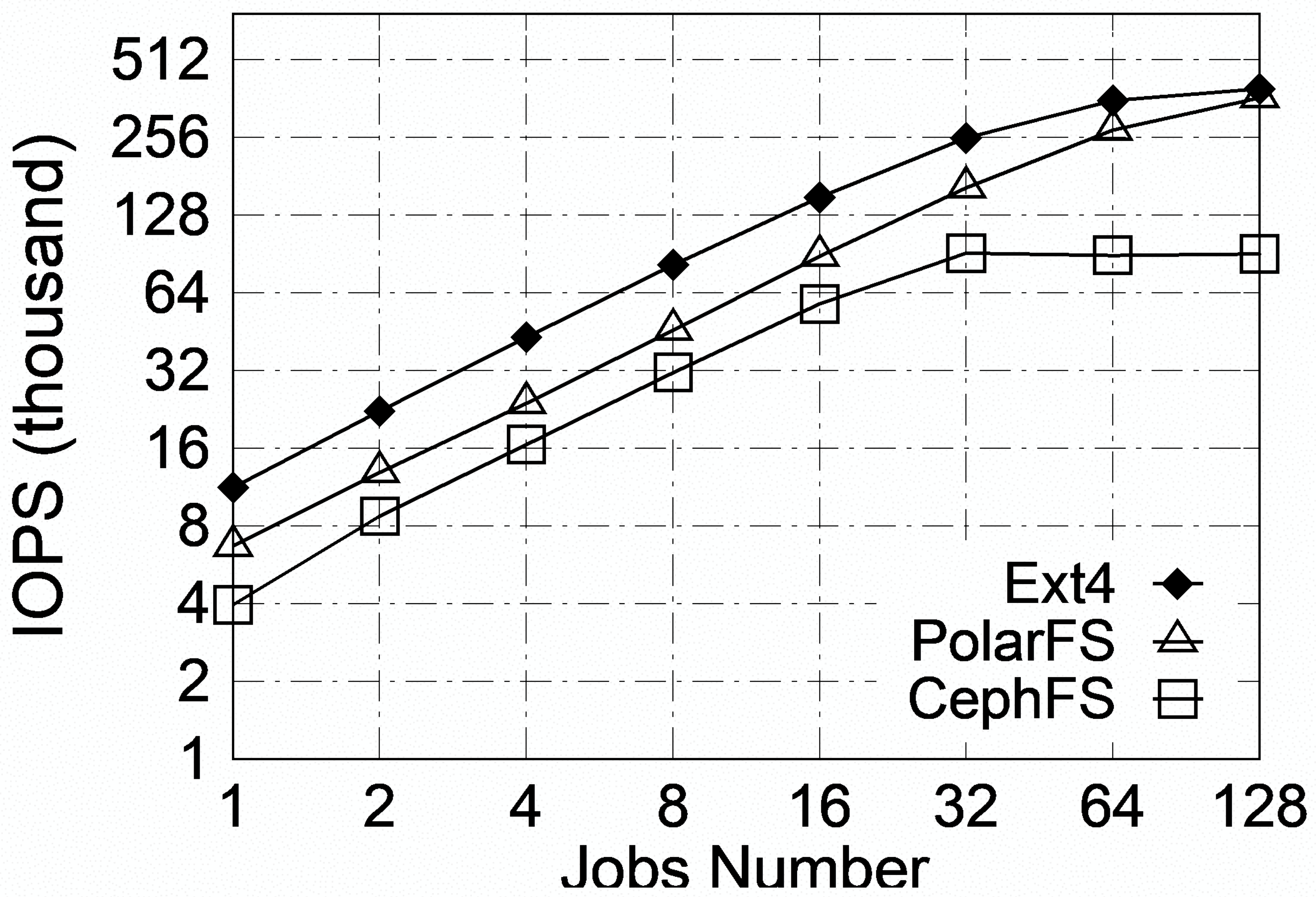
| | | |
|---|---|---|
| **Random:** | **PolarFS 1.6 to 4.7 times slower** | **CephFS 6.5 to 75 times slower** |
| **Sequential:** | **PolarFS 1.6 to 4.8 times slower** | **CephFS 6.3 to 56 times slower** |

# Evaluation – PolarFS Throughput

4KB Random Write




4KB Random Read

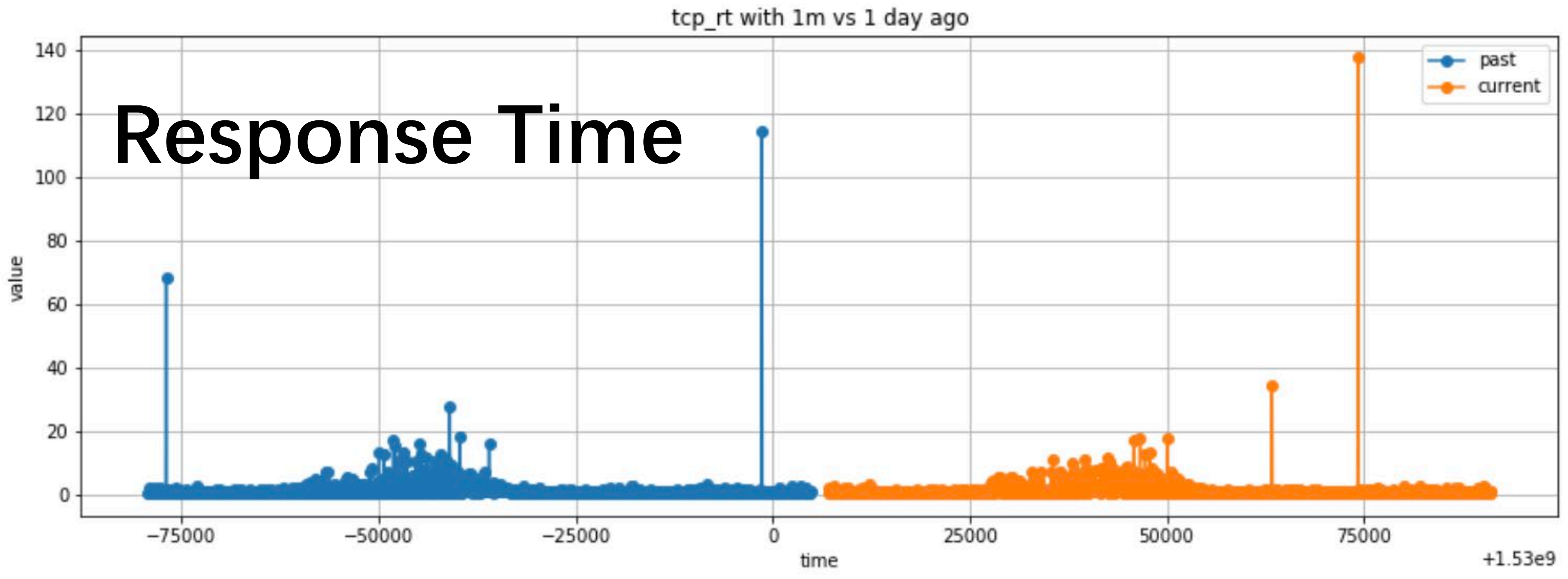


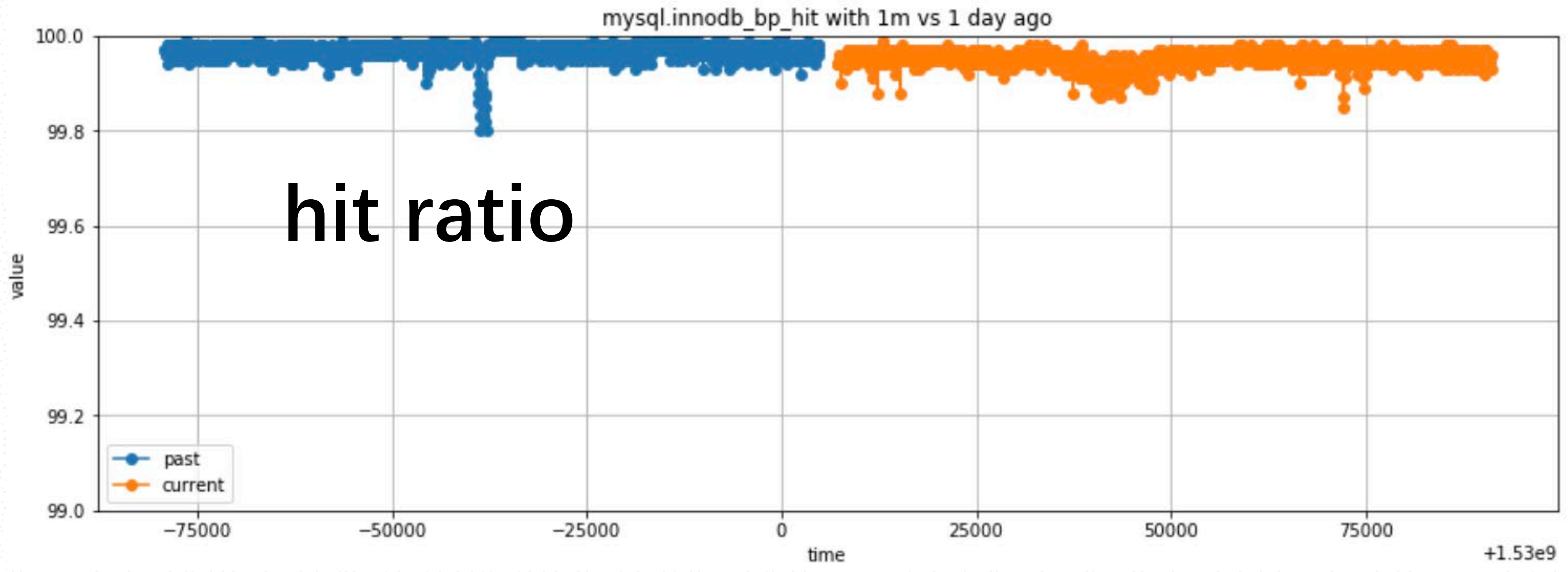

**Write / Read throughput of Ext4 and PolarFS are 4/7.7, 4.4/5.1 higher than CephFS**

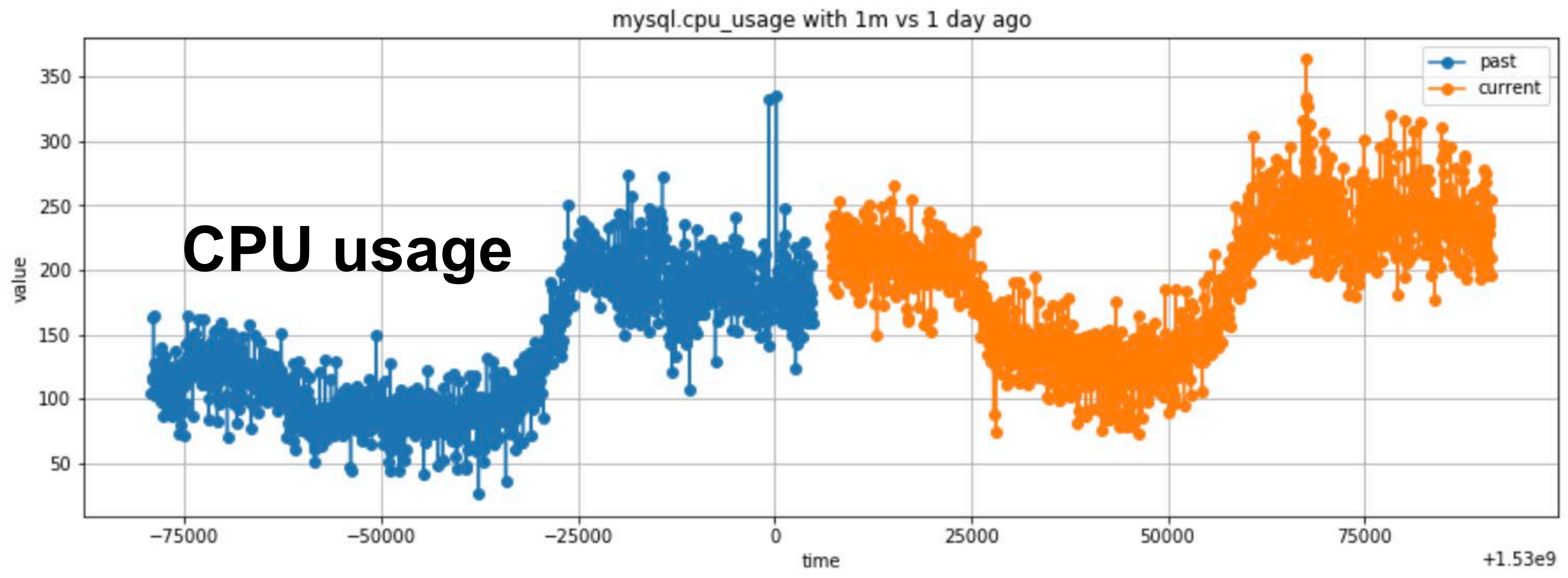
# iBTune - Preliminary Attempt

Alibaba Cloud | Worldwide Cloud Services Partner

## Buffer pool (BP) size is correlated to miss ratio




- Challenge: Heuristic method (such as shrinking 10% each time) does not work, since we have to try many times, which makes the system unstable and is unacceptable for mission-critical applications.

**Intuition：**

- Calculate BP based on hit ratio (miss ratio) to avoid restarting system multiple times
- Confirm whether the BP size meets the requirement of SLA

# System halting avoidance

Alibaba Cloud | Worldwide Cloud Services Partner

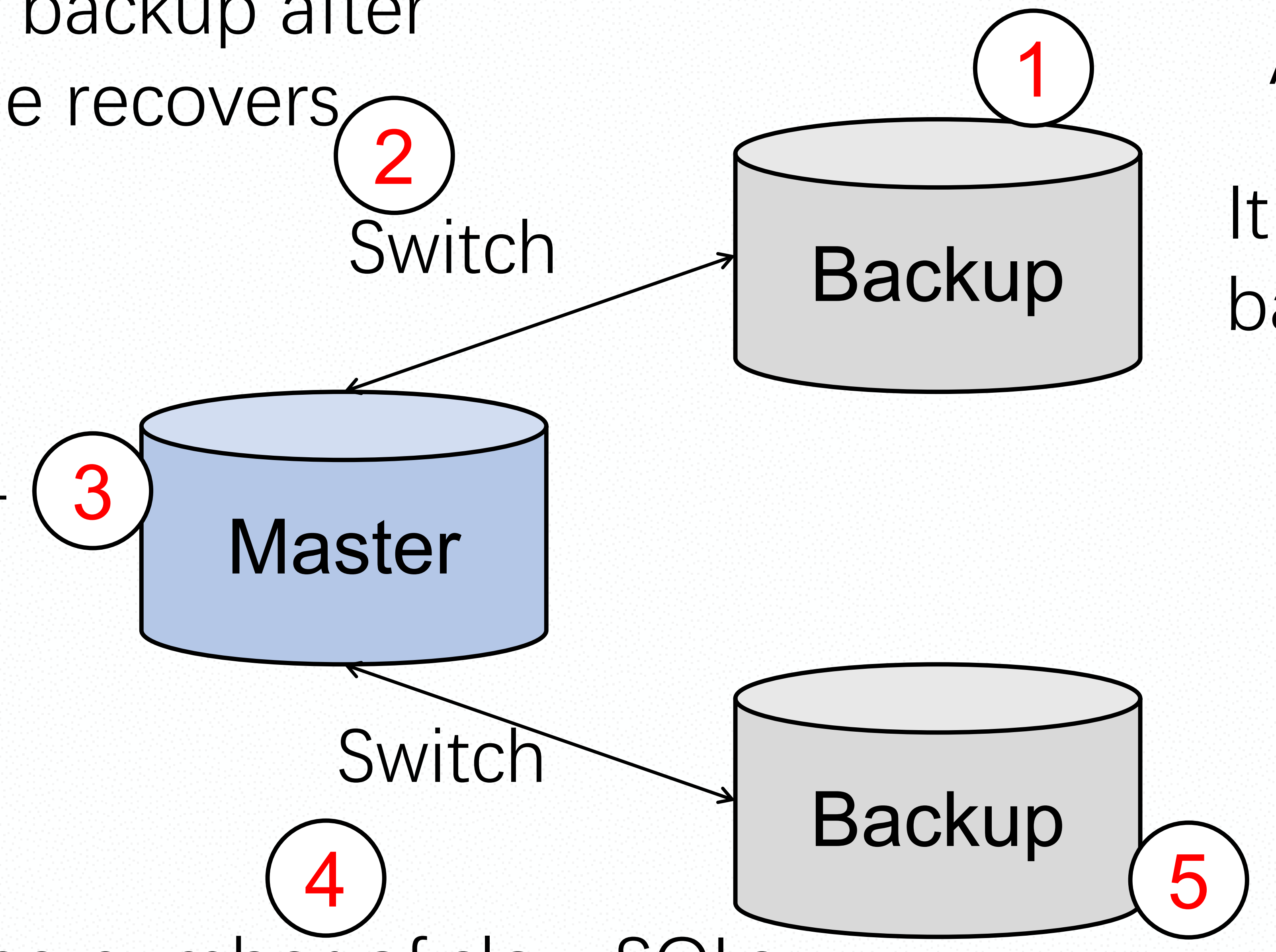
**Based on X-Paxos: high availability protocol via binlog implementation at Alibaba**

Switch master and backup after the backup node recovers

Adjust backup node’s BP size

It leads to 10~20 sec system halting to backup node, without affecting system service.

Monitor the new master for 24 hours




If the new master works fine during the following 24 hours, backup nodes’ BP will be adjusted.

If the new master is abnormal, i.e., the number of slow SQLs increases, rollback will be trigged.

Rollback means switching between master and backup nodes: since the backup node’s BP is still the old one, rollback restores the original status.

# Evolvement of database systems



**Structured Data**

RDBMS
**[ SQL+OLTP ]**

**Structured Data**

Data warehouse
Data Cube
**[ ETL+OLAP ]**

**Heterogeneous Data**

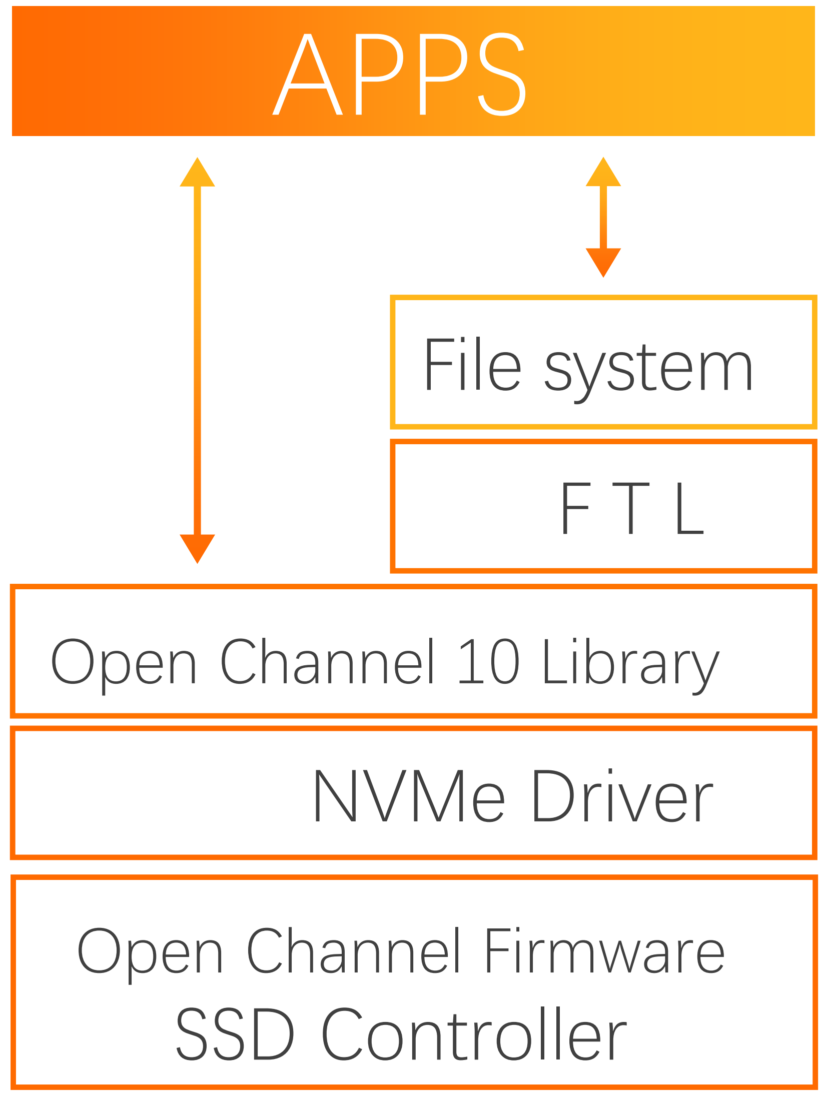
Structured Data
Graph
Time Series
Vector
Spatial Data
Text

RDBMS
NoSQL/NewSQL DB

**[ Multi-Model + HTAP ]**

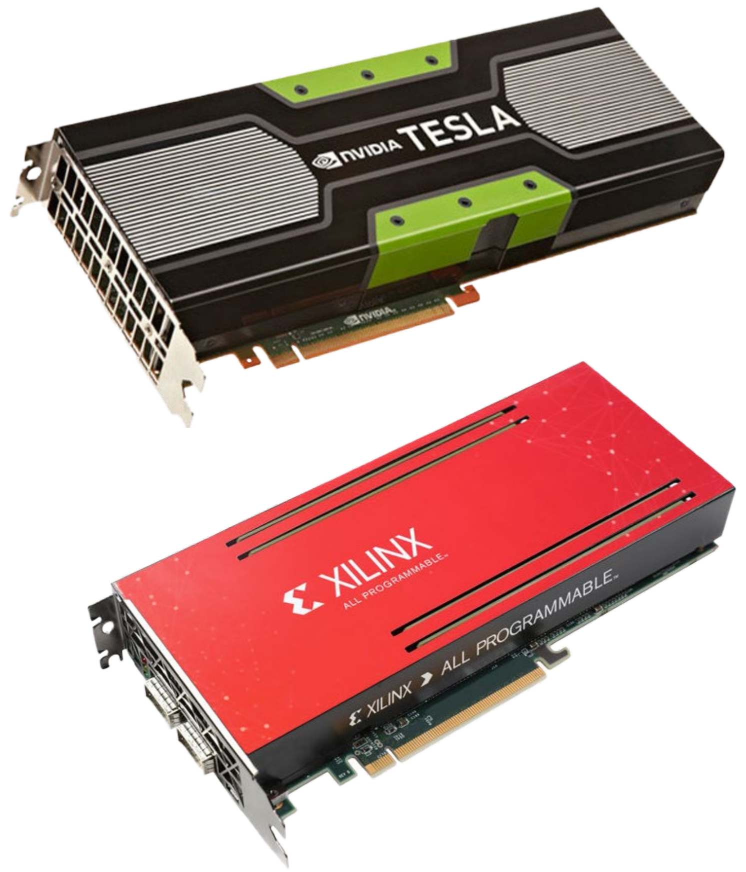
# Hardware-software co-design


RDMA




Open-Channel SSD

NVM 3D XPoint




GPU/FPGA

# Multi-Model Database System

Alibaba Cloud | Worldwide Cloud Services Partner

**Query Interface**

SQL | Put/Get | SPARQL | Doc QL | GQL | ....

## DB Engine

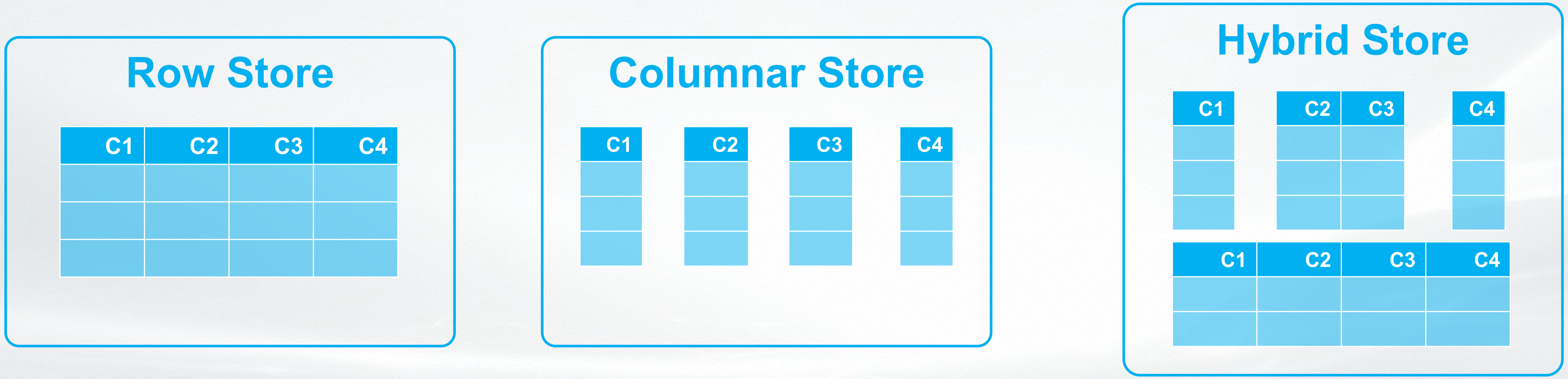
**Data format**

Structured data | Graph | Vector | Doc

Spatial-Temporal | Non-structured data | Time series

# HTAP: Hybrid Transaction and Analytical Processing

Alibaba Cloud | Worldwide Cloud Services Partner




**OLTP**

✓ Real time update, point query

✗ Compression, analytics

**OLAP**

✓ Compression、read only、Complex queries and scan

✗ Updates

**Hybrid: HTAP**

# Data Security