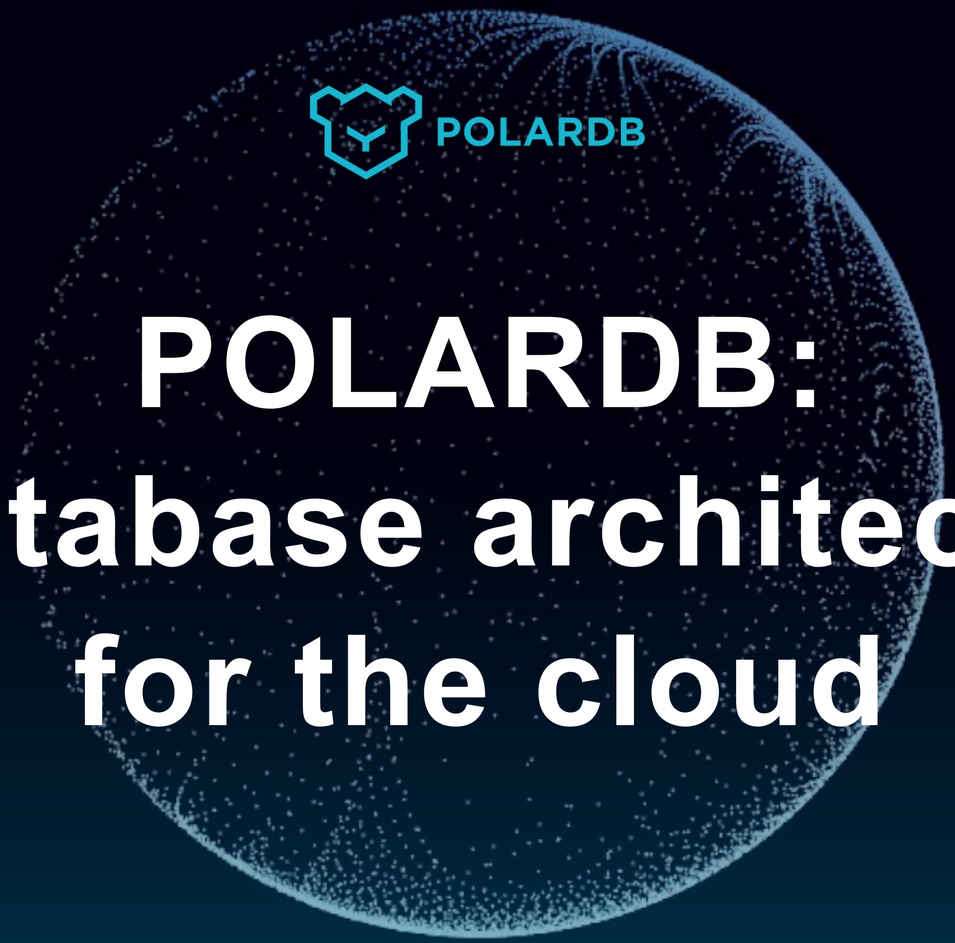
Alibaba Group
阿里巴巴集团



POLARDB: A database architecture for the cloud

ØYSTEIN GRØVLEN
Sr. Staff Engineer @ Alibaba Cloud

Bio:
Before joining Alibaba, Øystein worked for 10 years in the MySQL optimizer team at Sun/Oracle. At Sun Microsystems, he was also a contributor on the Apache Derby project and Sun's Architectural Lead on Java DB. Prior to that, he worked for 10 years on development of Clustra, a highly available DBMS.

Databases inside Alibaba Group

1 Trillion USD

100M

PB level

2018 Sales ($)

Alibaba Singles' Day(11.11)

30.8B

Cyber Monday

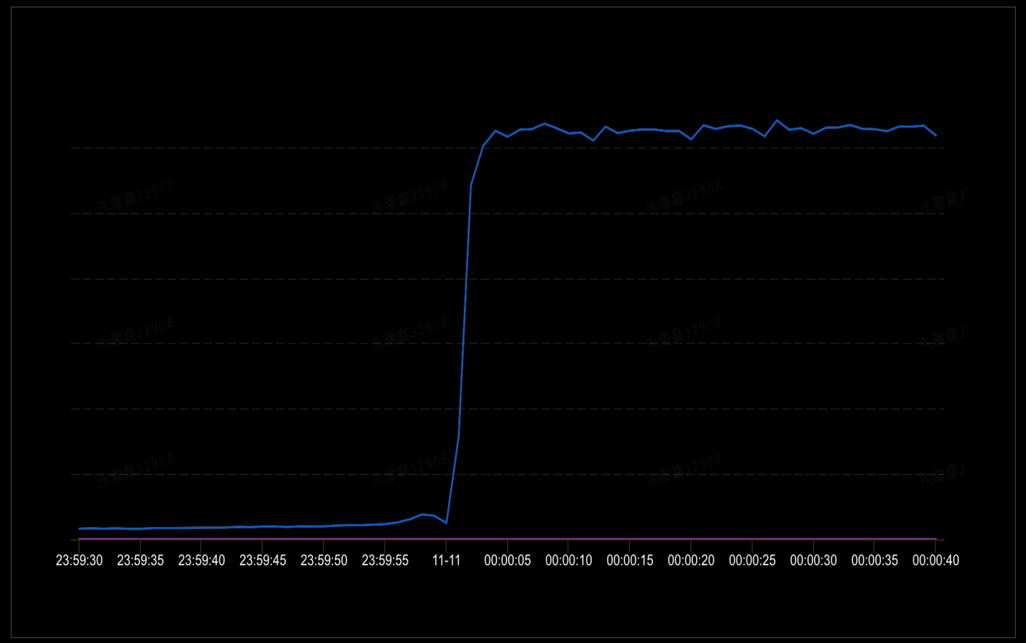
7.9B

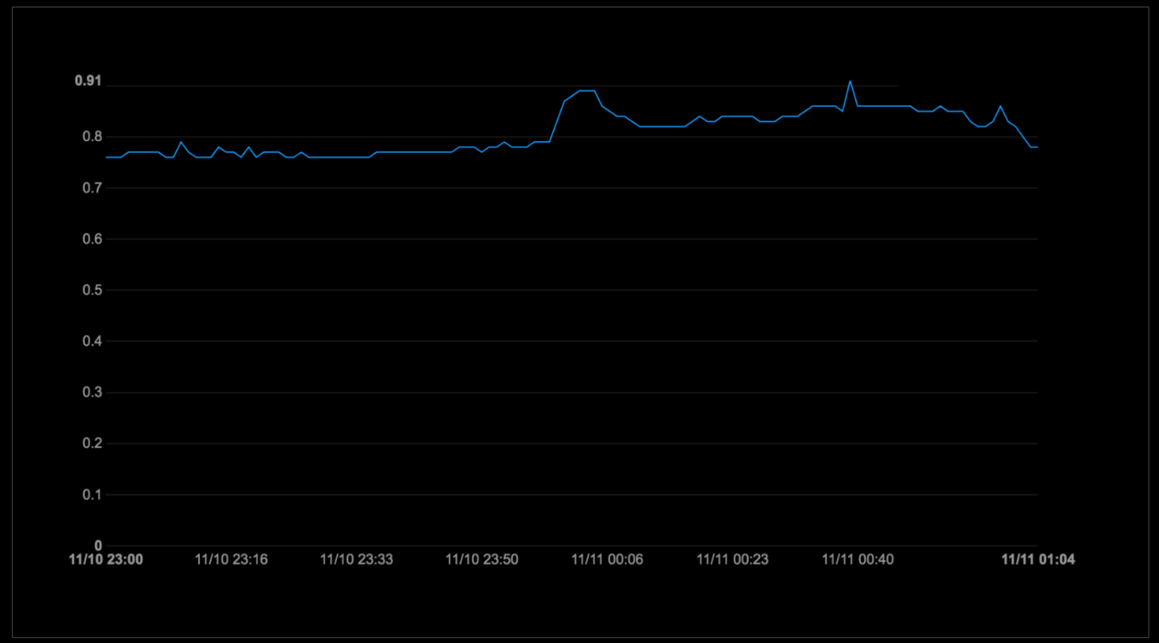
Amazon Prime Day

4.19B

*Data source: forbes, cnbc, practicalecommerce and digitalcommerce360

Database Scalability Challenge in Alibaba Single's Day





Load: ~100x

RT latency: unchanged

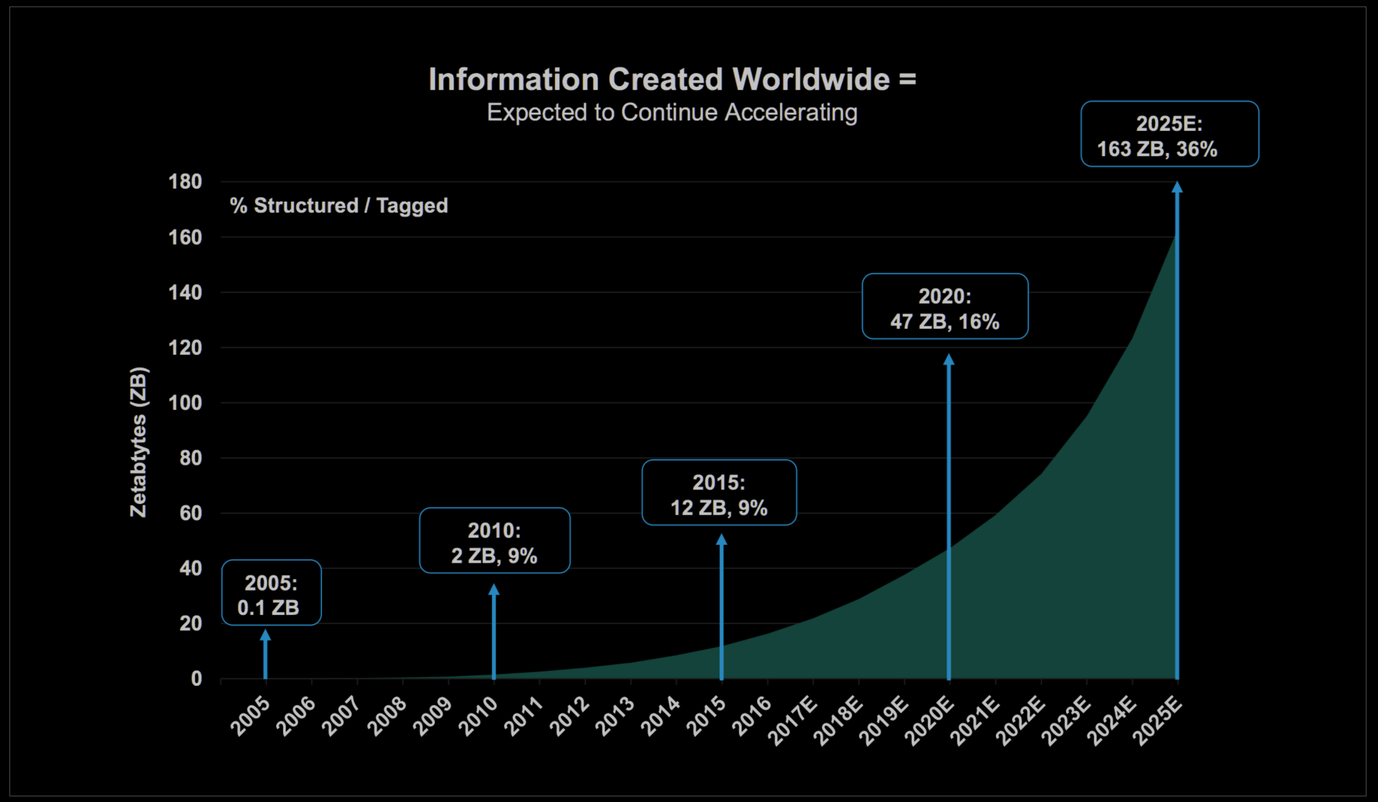
Cloud shifts fixed **CapEx** expenses to variable **OpEx** expenses

Alibaba Group 阿里巴巴集团

83% of Enterprise Workloads Will Be In The Cloud By 2020

— Forbes

Data Explosion



Generated by Human → Generated by Things

- Data in Large Scale
- Increased expense
- Hard to utilize

Cloud Native Database — Requirements

Scalable

- Auto-scaling
- Load
- Storage

Highly available

- Data Redundancy
- Automatic Failover
- Zero Downtime

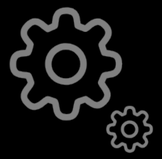
Integrate with Cloud Services

- DBaaS
- Security
- AI
- Serverless
- Monitoring

POLARDB — Cloud Native Database

CLOUD NATIVE

User Oriented



Emerging Hardware

- NVM
- RDMA
- FPGA

Serverless

- Auto Scaling
- Paid by Usage
- Zero Downtime

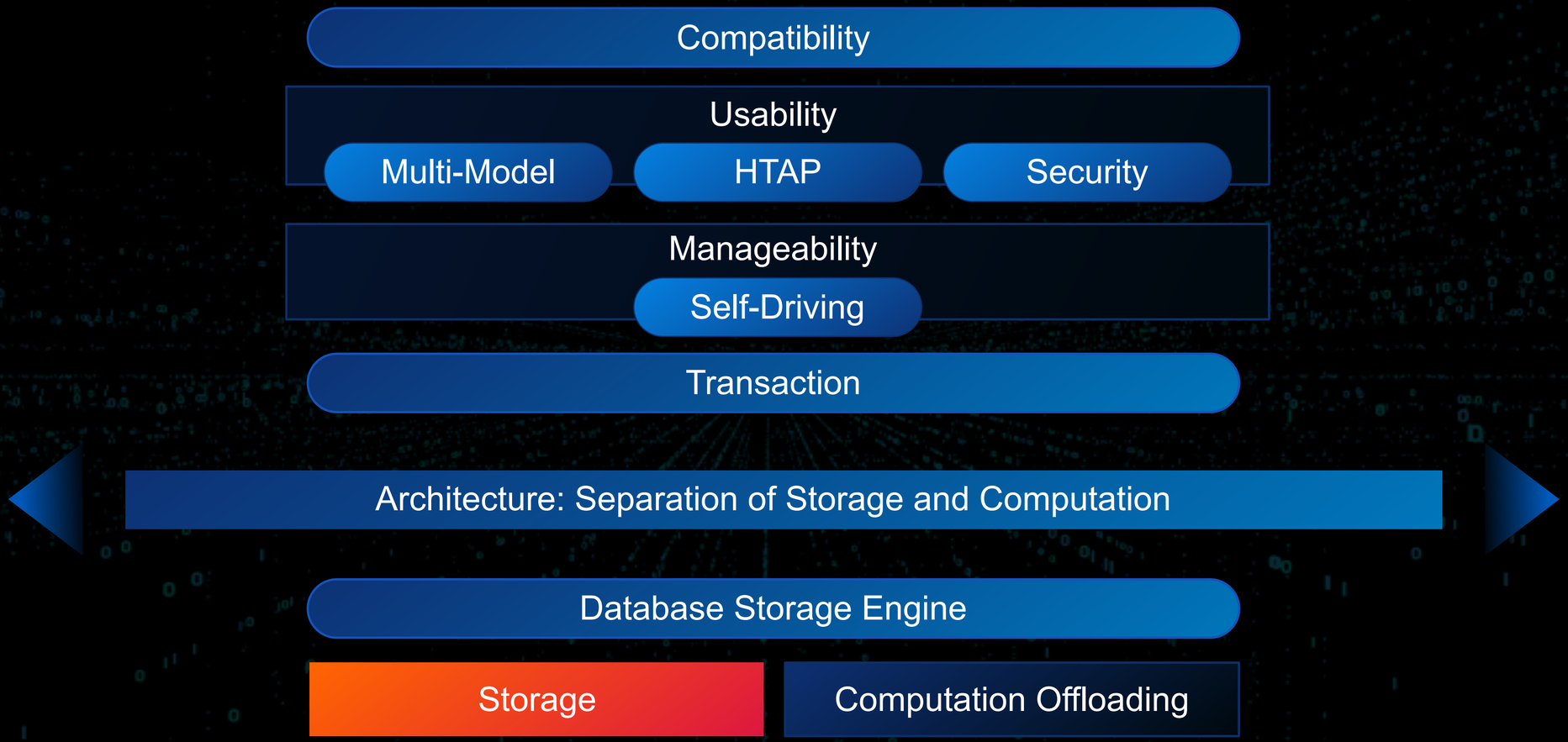
Security

- Encryption
- Audit
- Access Control

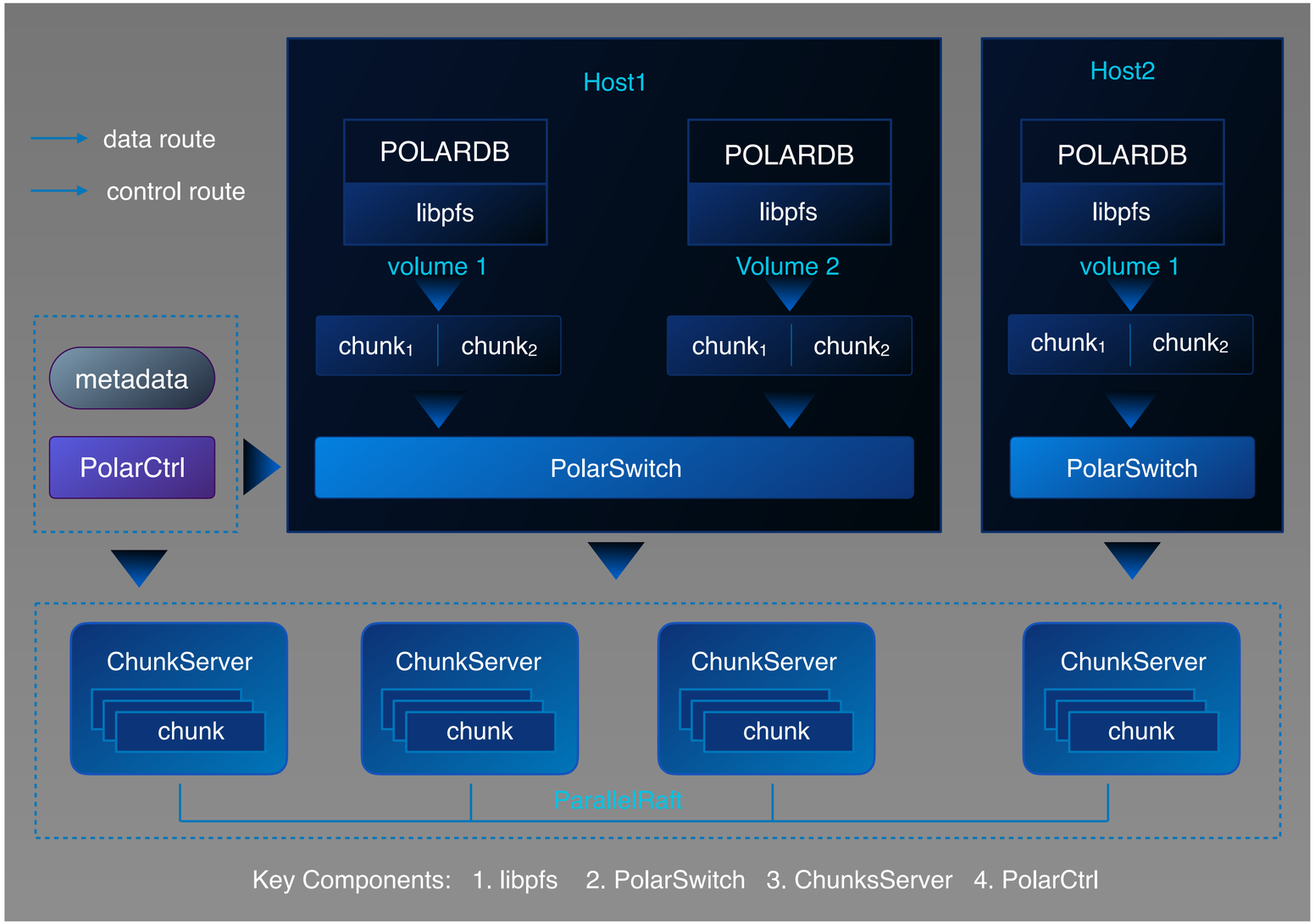
Intelligence

- Self-configuration
- Self-optimization
- Self-diagnosis
- Self-healing

Storage Revolution： PolarStore



- Design for Emerging Hardware
- Low Latency Oriented
- Active R/W – Active RO
- High Availability

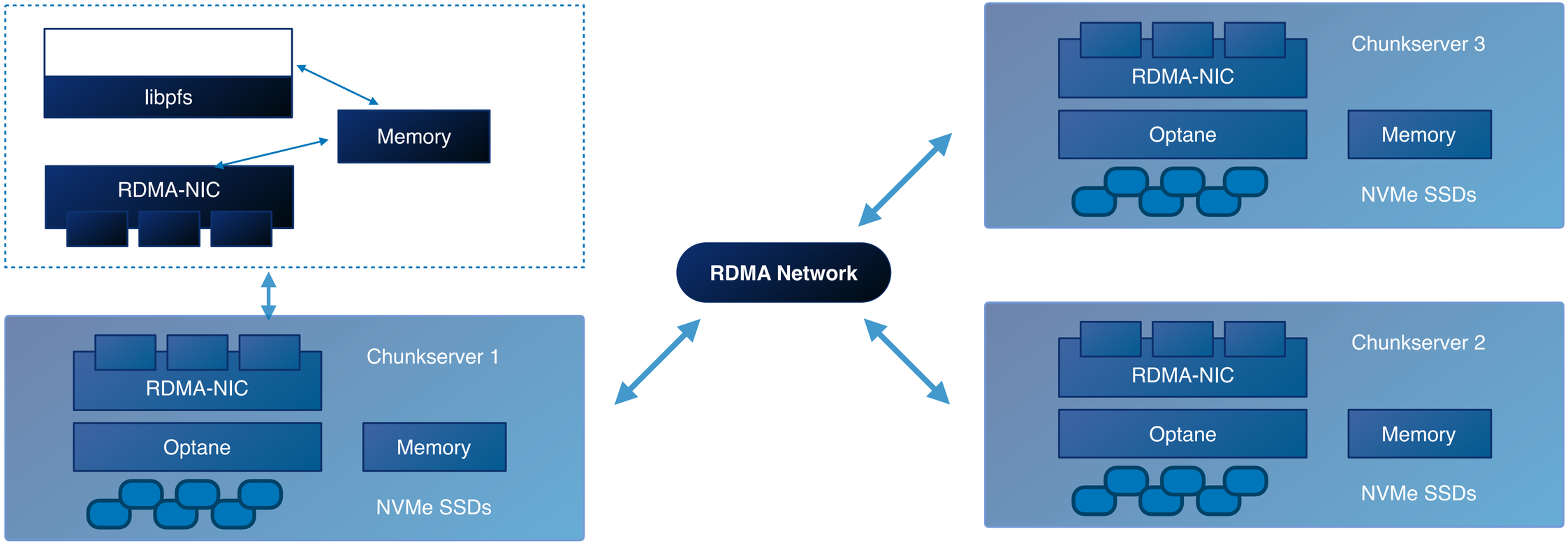


Network Over RDMA

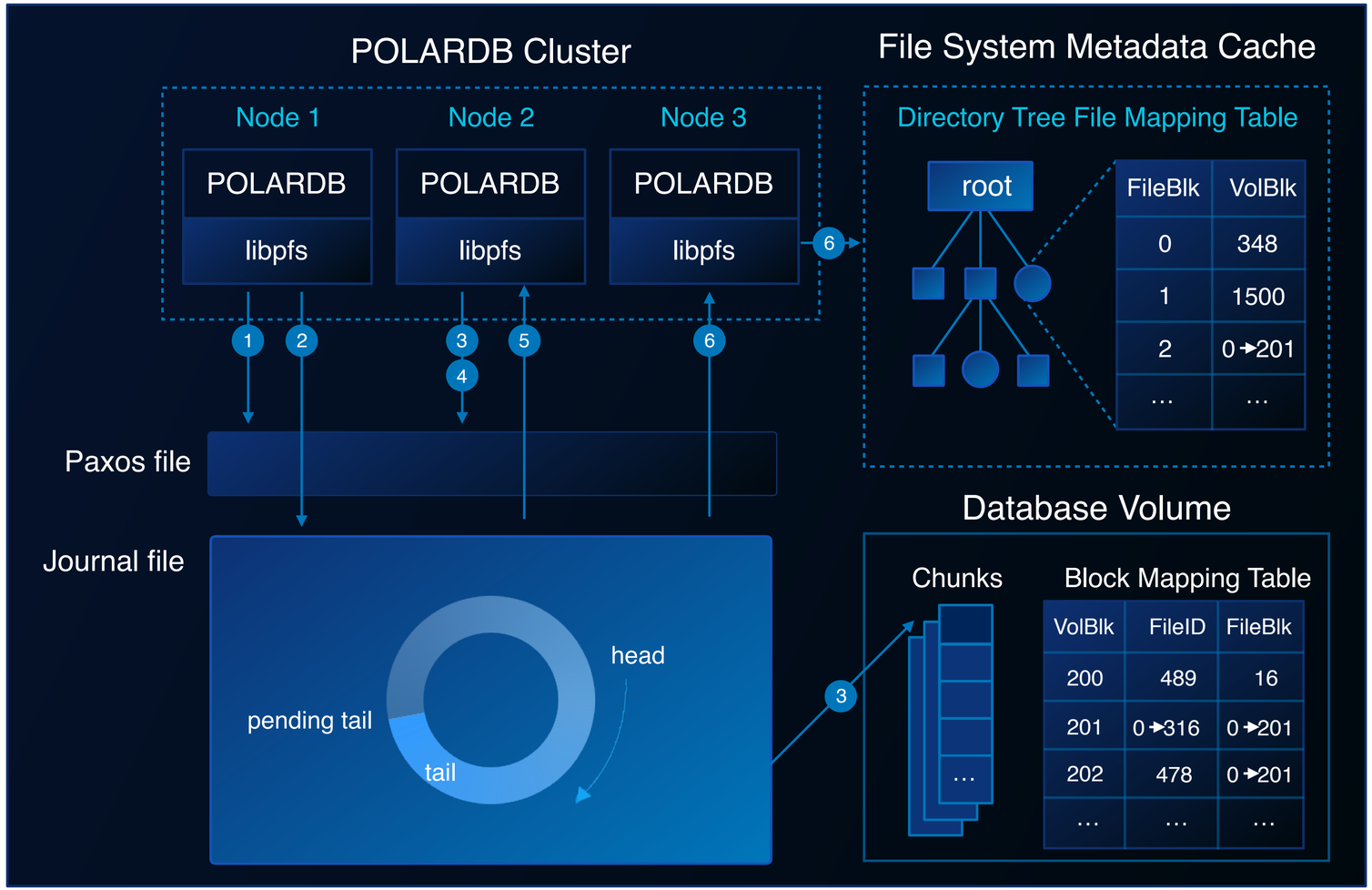
- No Context Switch
- OS-bypass & zero-copy

WAL Log in 3Dxpoint optane

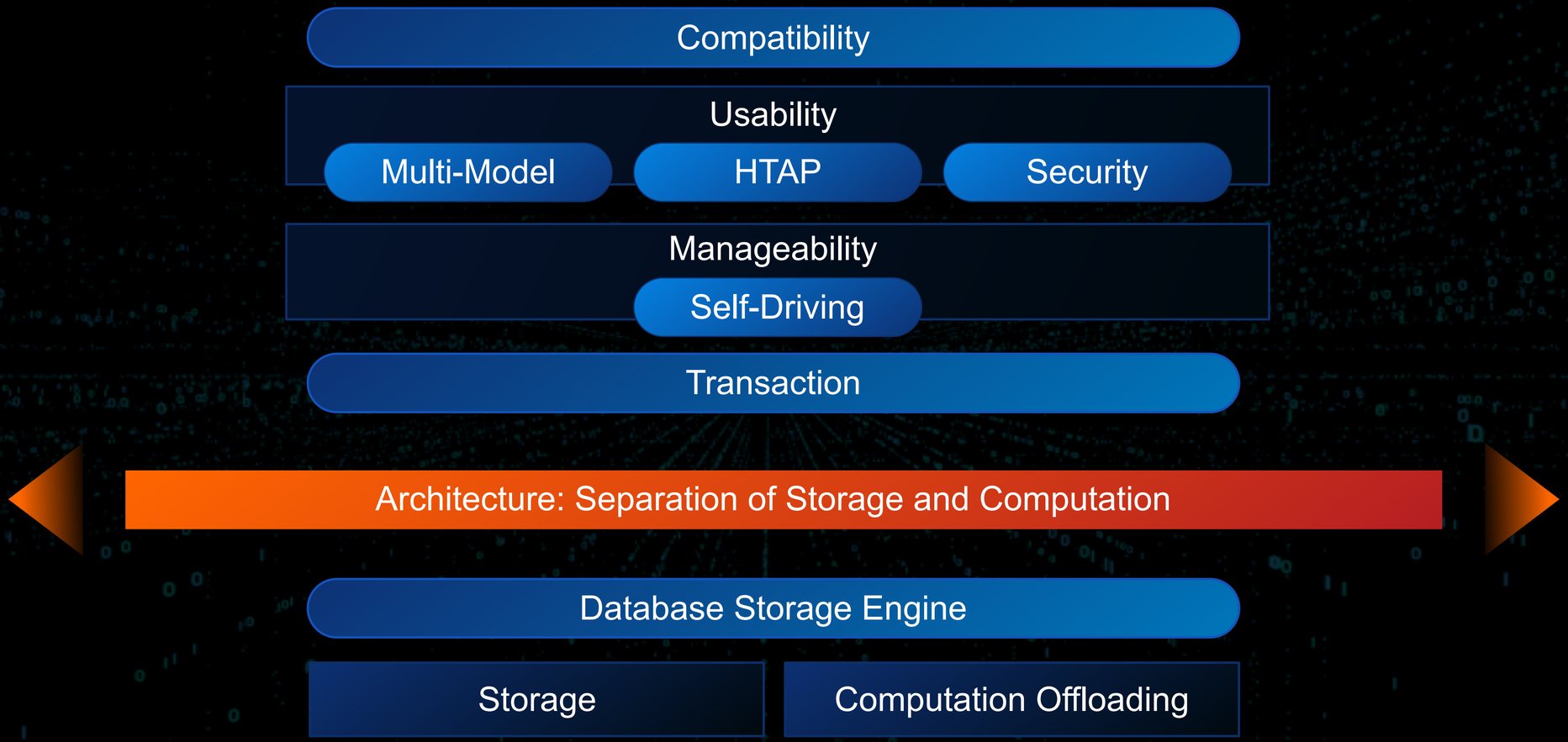
- Parallel Random I/O absorbed by Optane
- Excellent performance with less long tail latency issue
- No need of Over Provisioning



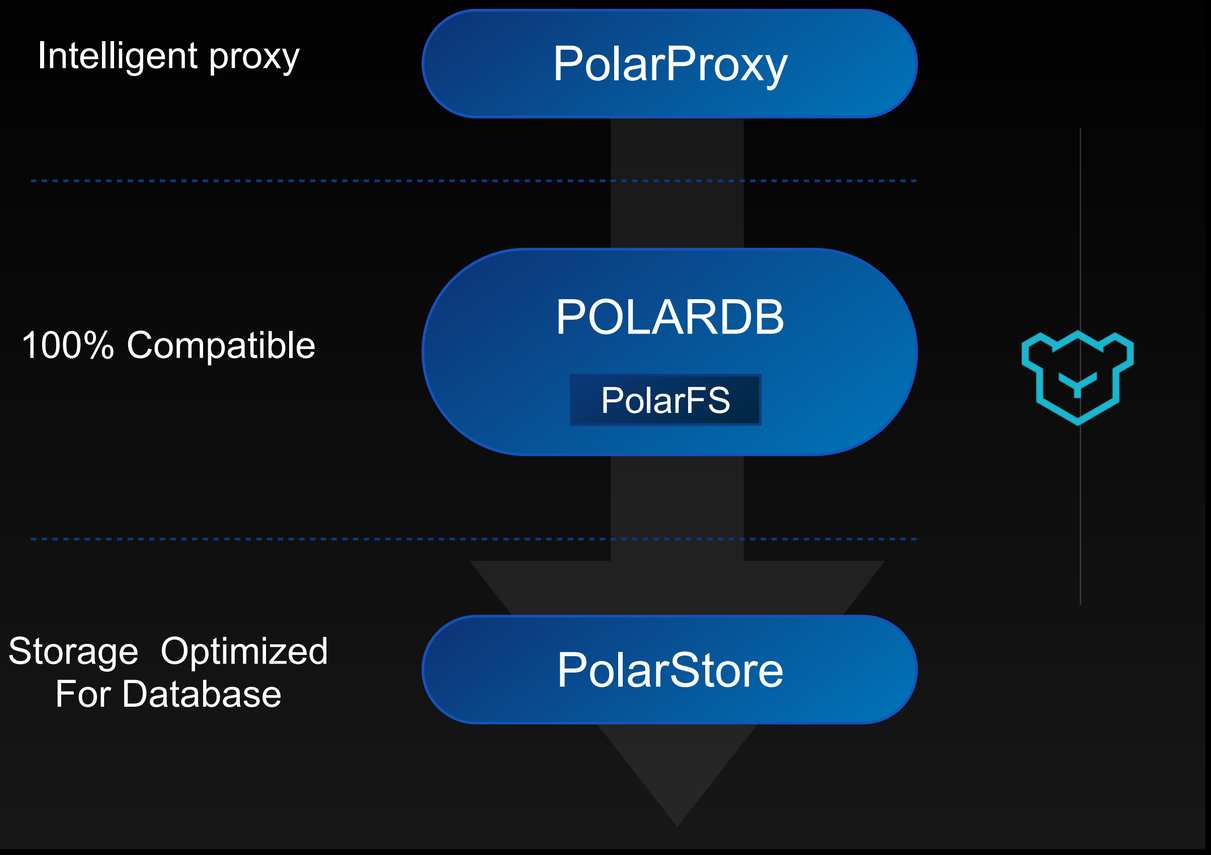
Low Latency Oriented



Database Architecture Revolution: Separation of Storage and Computation

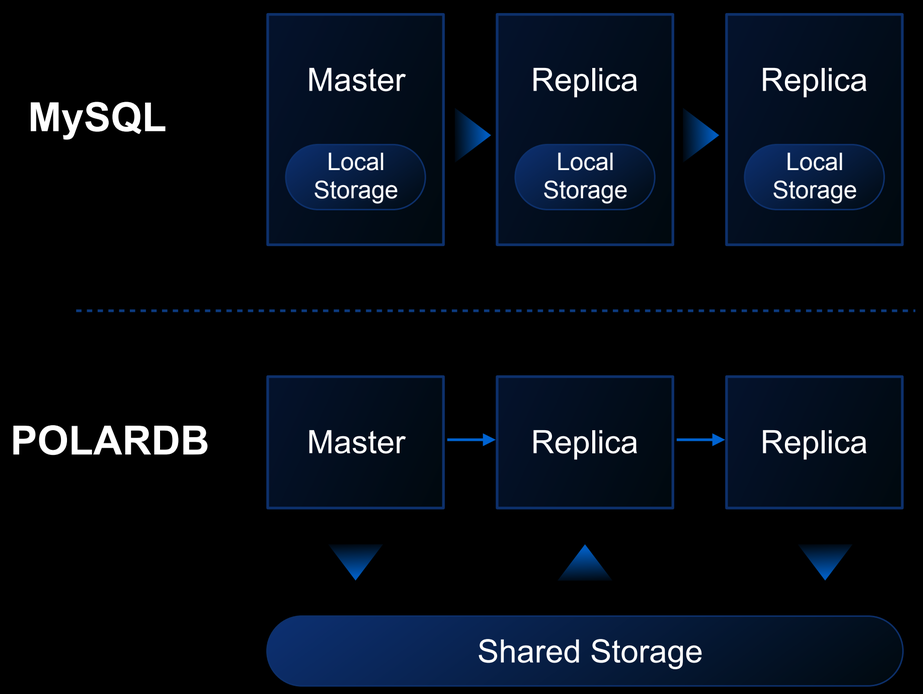


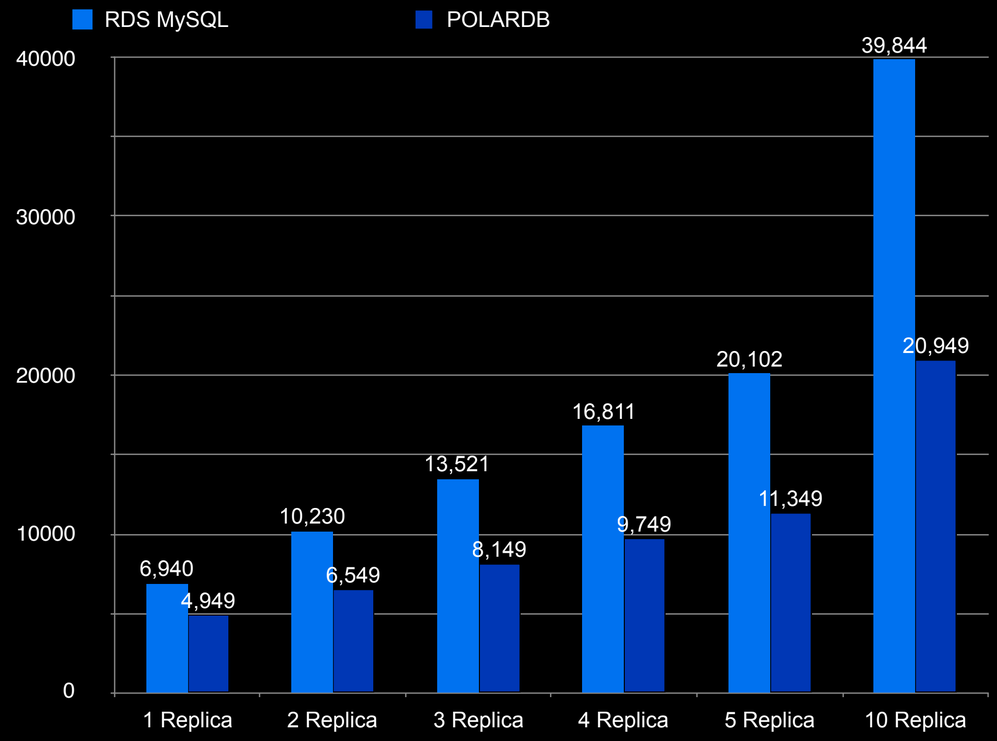
Cloud Native Architecture



- Scale compute and storage independently
- Shared storage
- Across AZ fail-over without data loss
- Optimize division of functionality between storage and compute
- Tight integration with other cloud components like metering, monitoring, control plan
- Optimize for hardware in the data centers
- Compatible with MySQL/PG etc
- Security

Dynamic Scaling





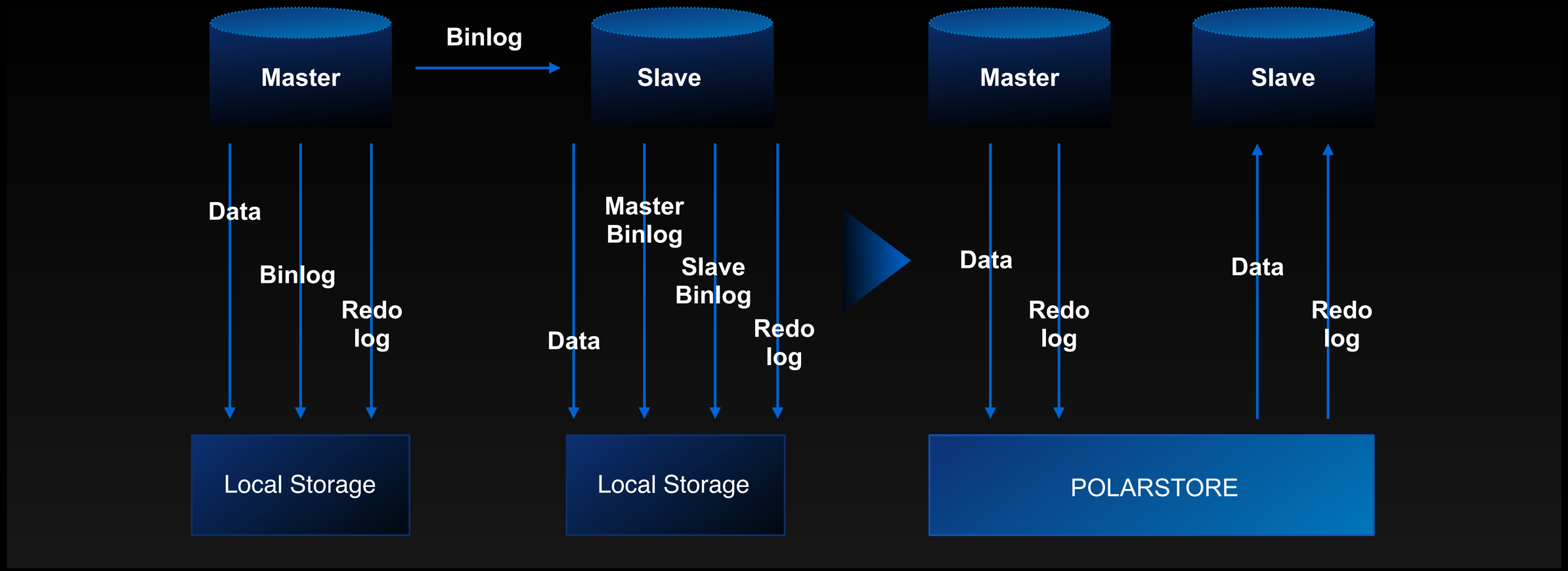
Fast Scaling

Upgrade 2vCPU to 32vCPU, only in 5 minutes
Add more Replicas, only in 5 minutes.

Lower Cost: 30%~50% OFF

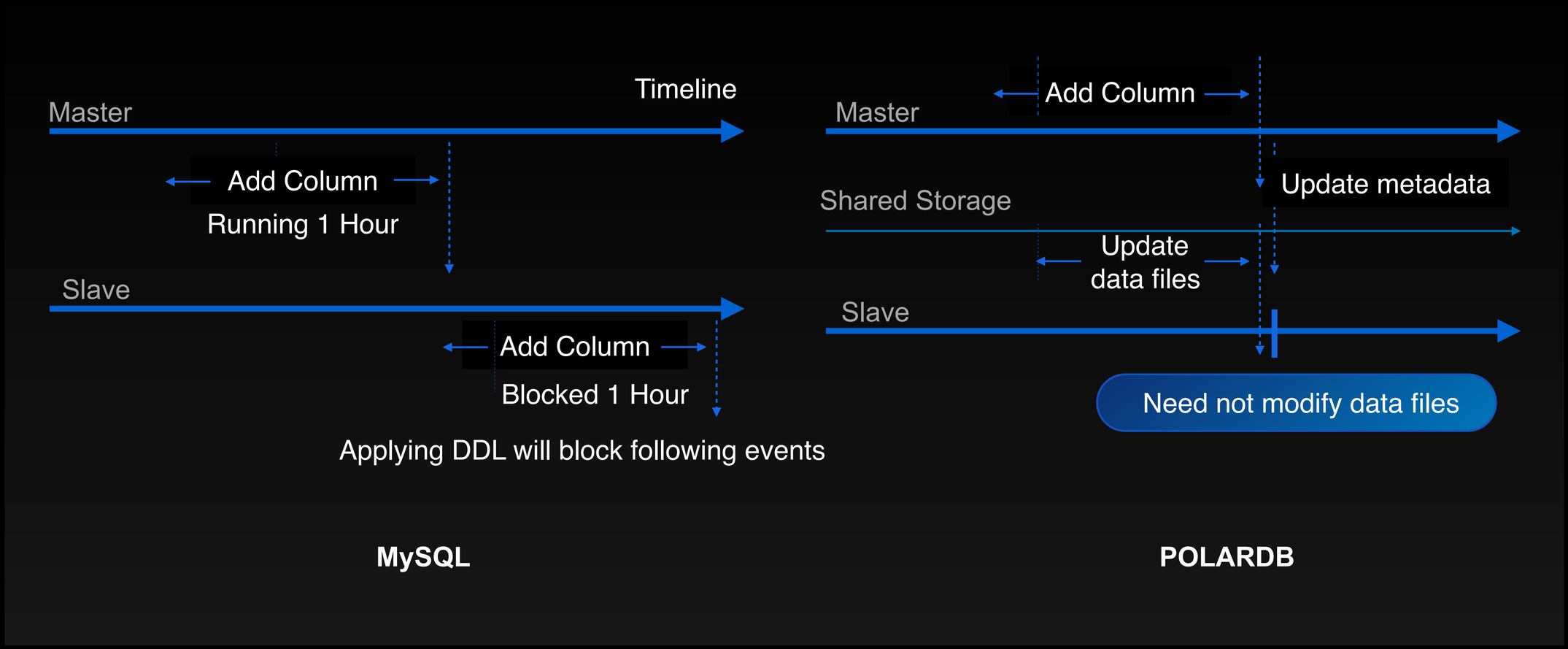
Total costs of 4vCPU 32G Memory 500G Storage with different replica numbers

Shared Nothing Logical Replication vs. Shared Storage Physical Replication



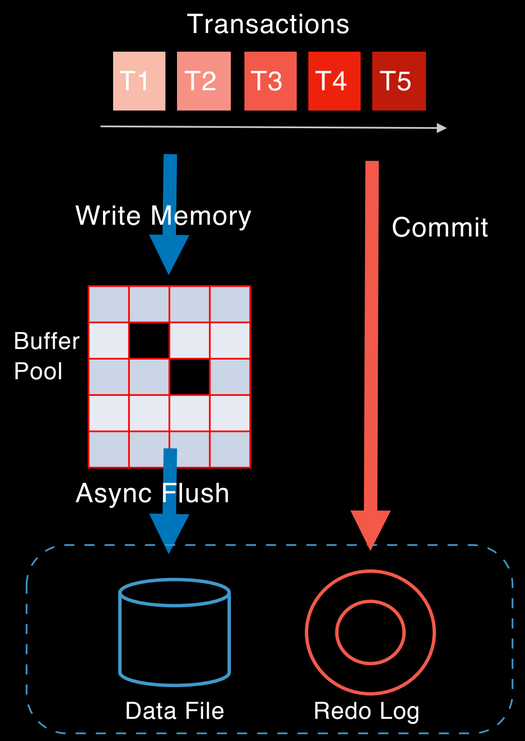
Physical Replication is much more reliable than Logical Replication

Shared Nothing Logical Replication vs. Shared Storage Physical Replication

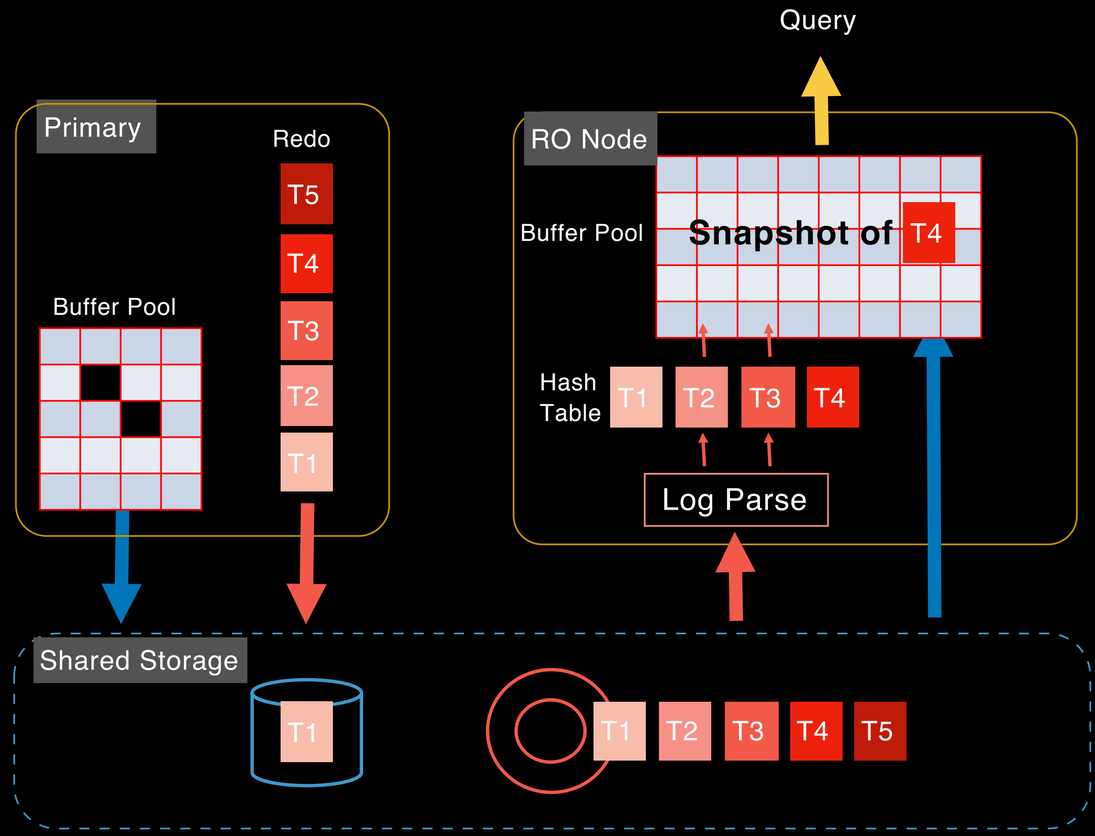


Non-blocking low-latency DDL synchronization

Physical Replication by Redo Log



DATA、LOG & MEMORY

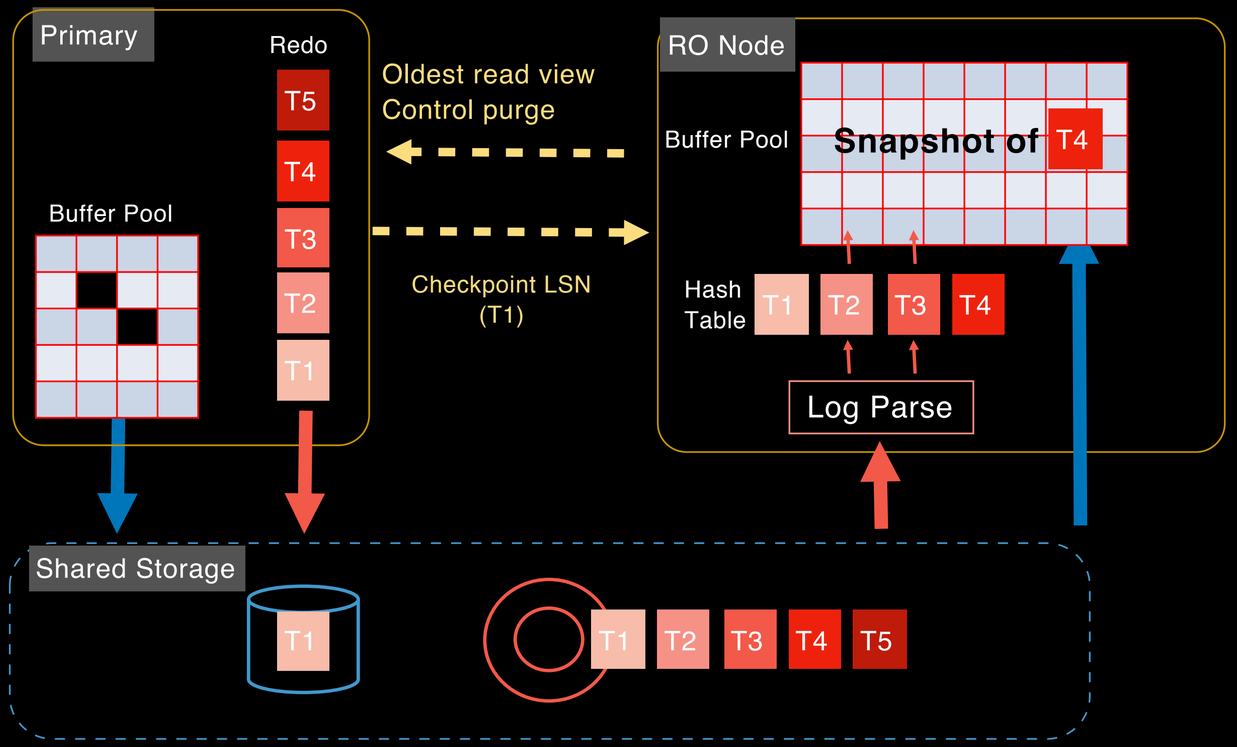


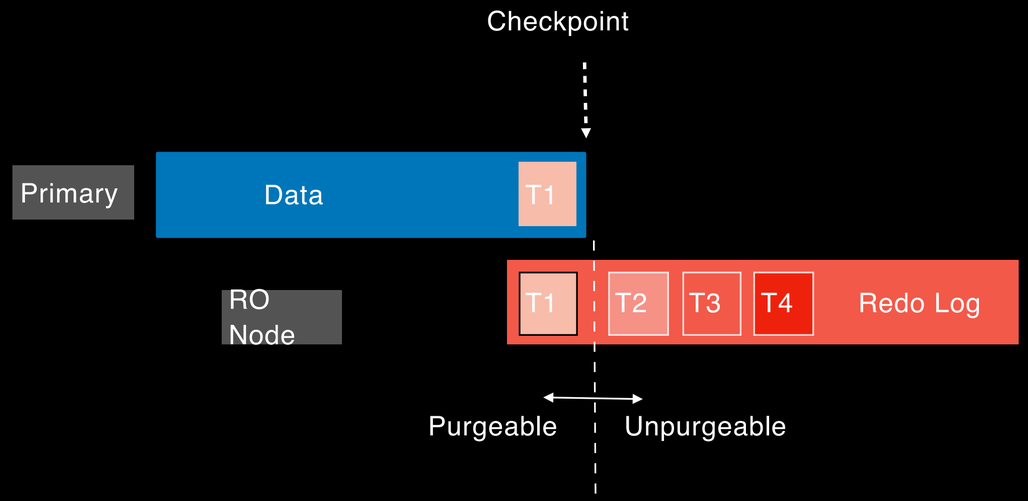
Shared Storage

Continuous Recovery

Consistent Snapshot Read

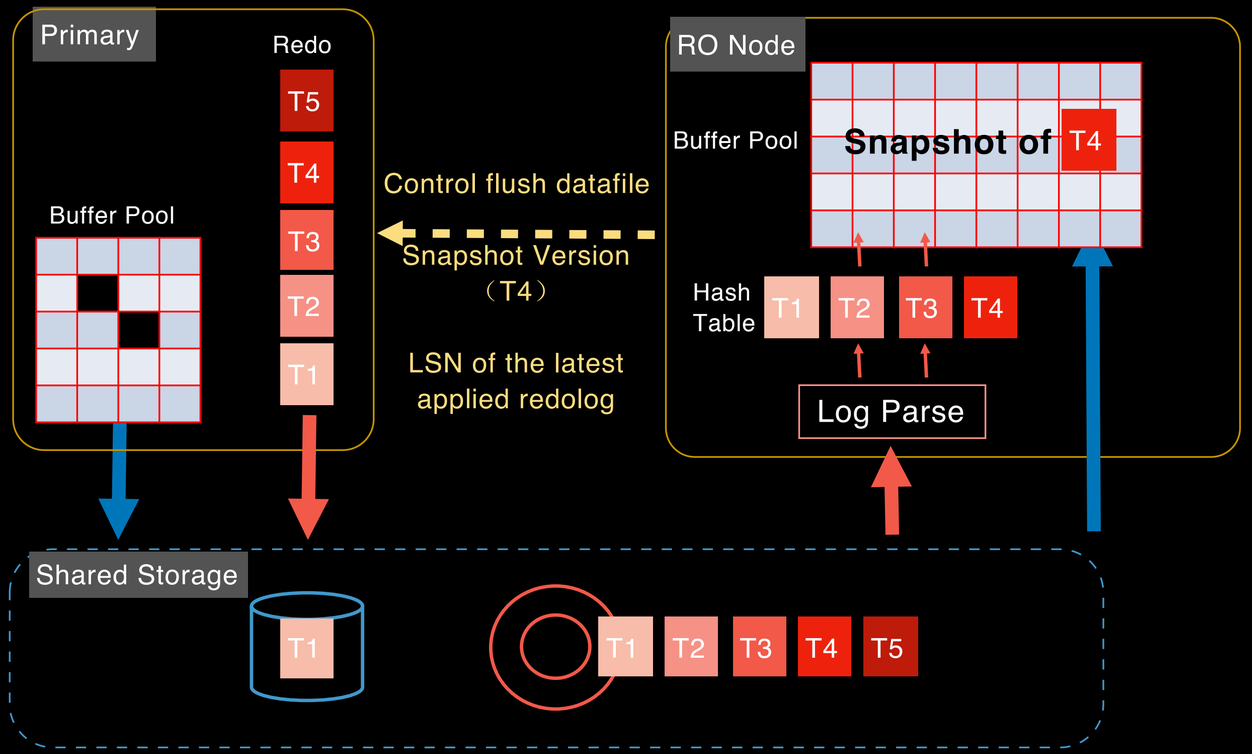
Physical Replication — Page from Past

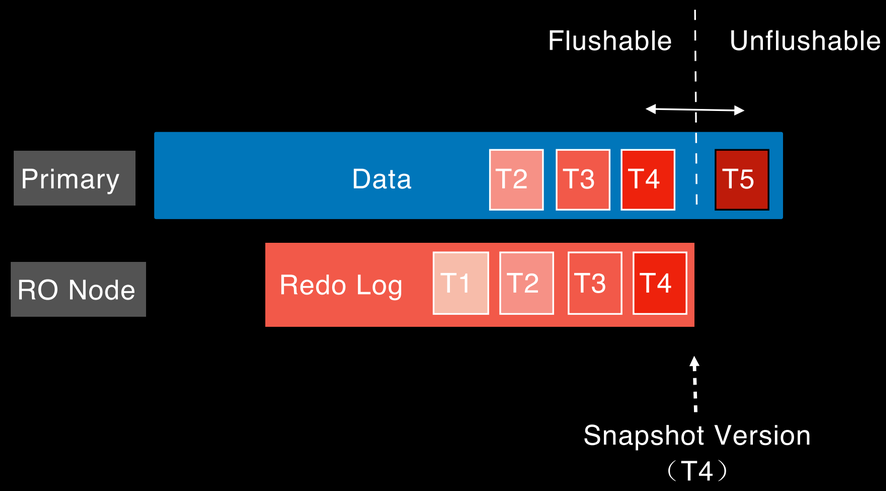




Avoid『Data Gap』

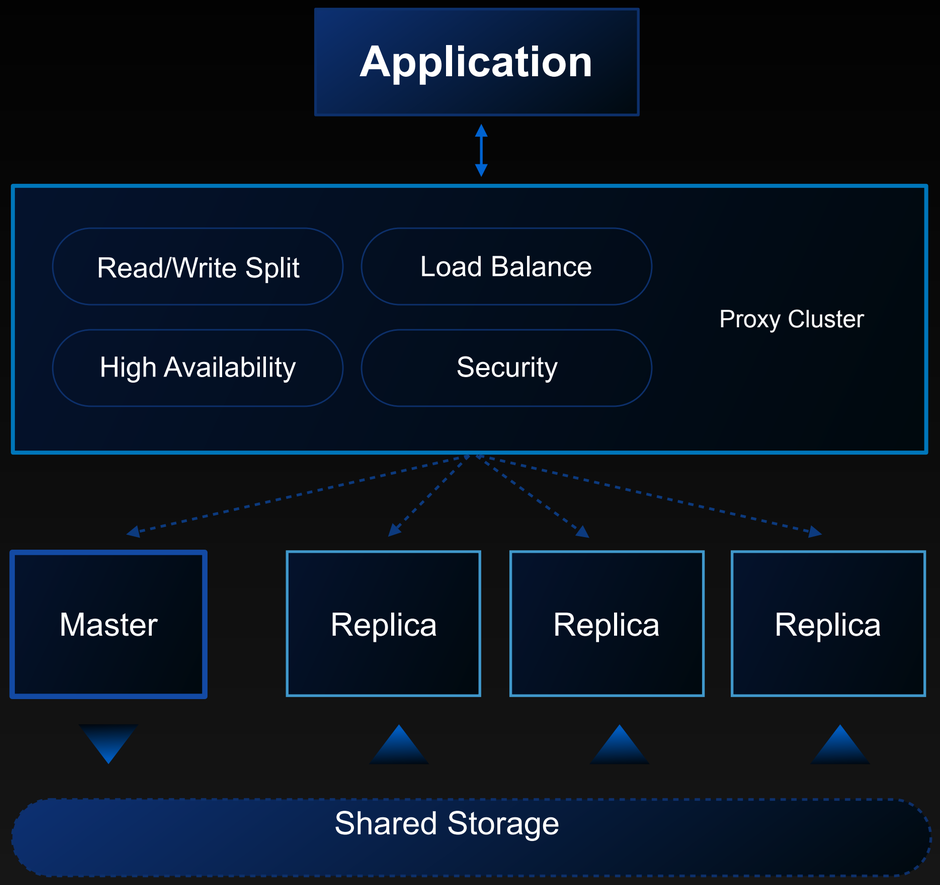
Physical Replication — Page from Future





Avoid『Data Overstep』

Single Master



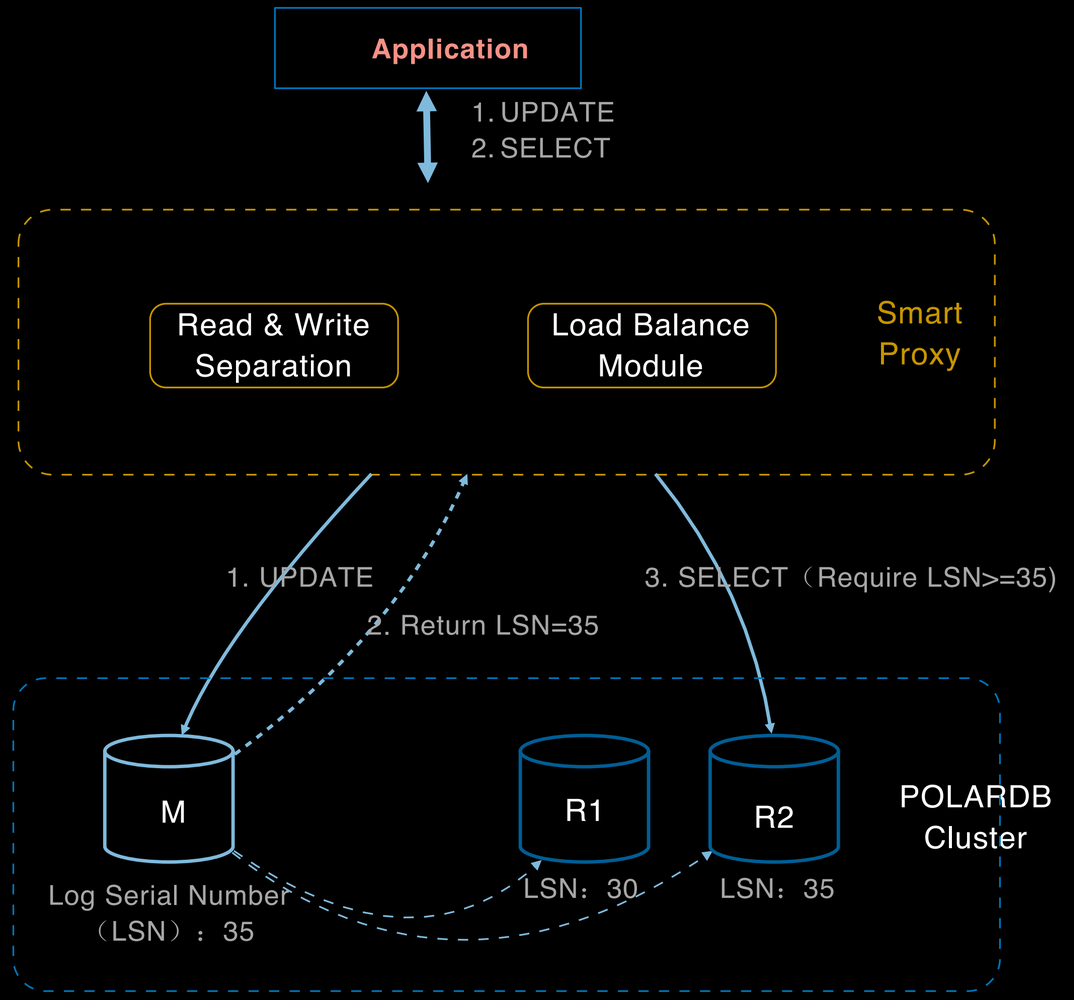
- Single Endpoint
- Transparent Failover
- Attacks Protection
- Causal Consist Read

Read and Write Separation — Session Consistent

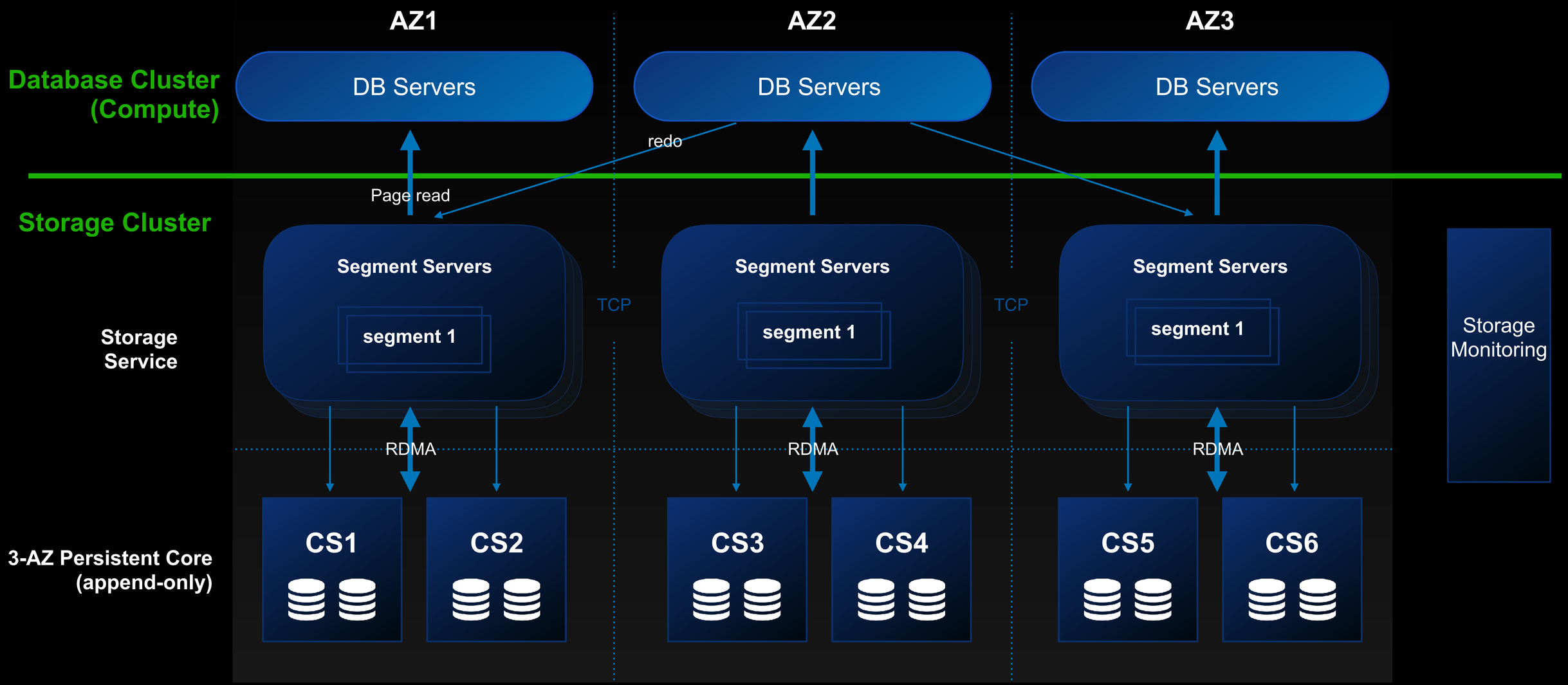
```
connection.query
{
  UPDATE user SET name='Jimmy'  WHERE id=1;
  COMMIT;
  SELECT name FROM user WHERE id=1;   // name is Jimmy
}
```

（SELECT can always get the latest data）

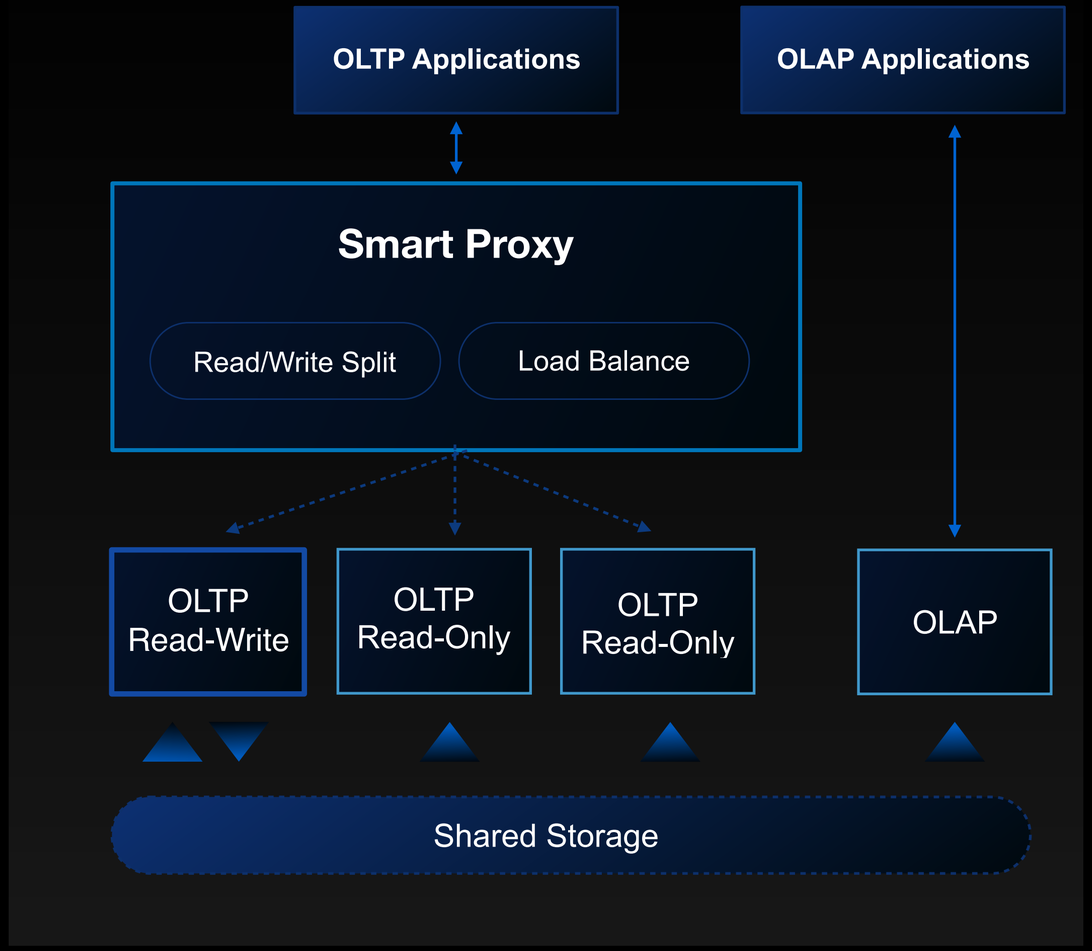
Problem『Can't read latest data』 Solved!



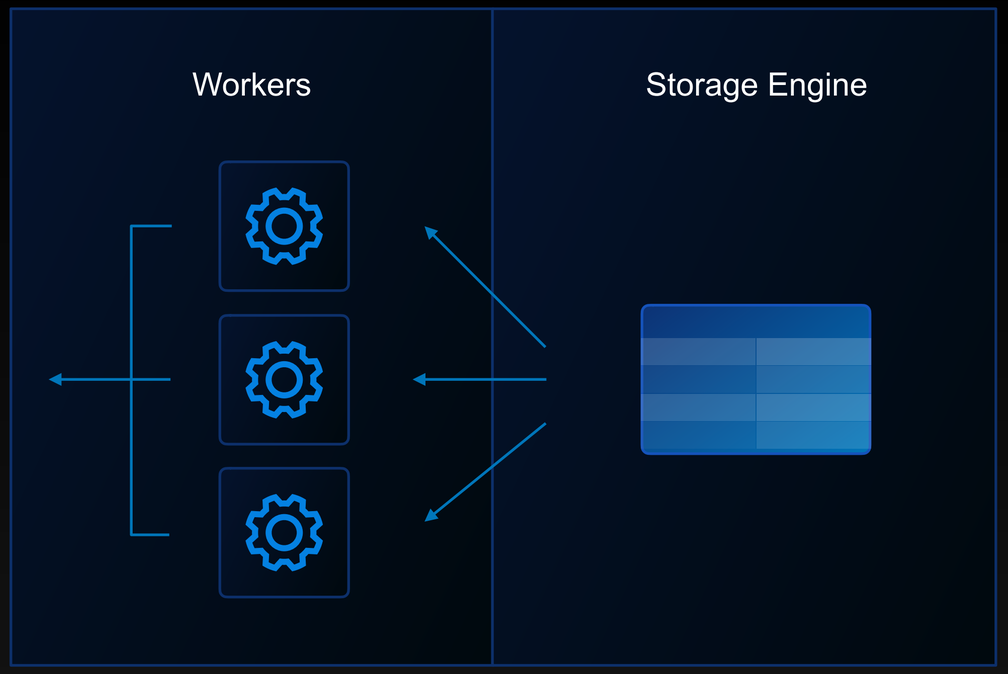
Multi-master



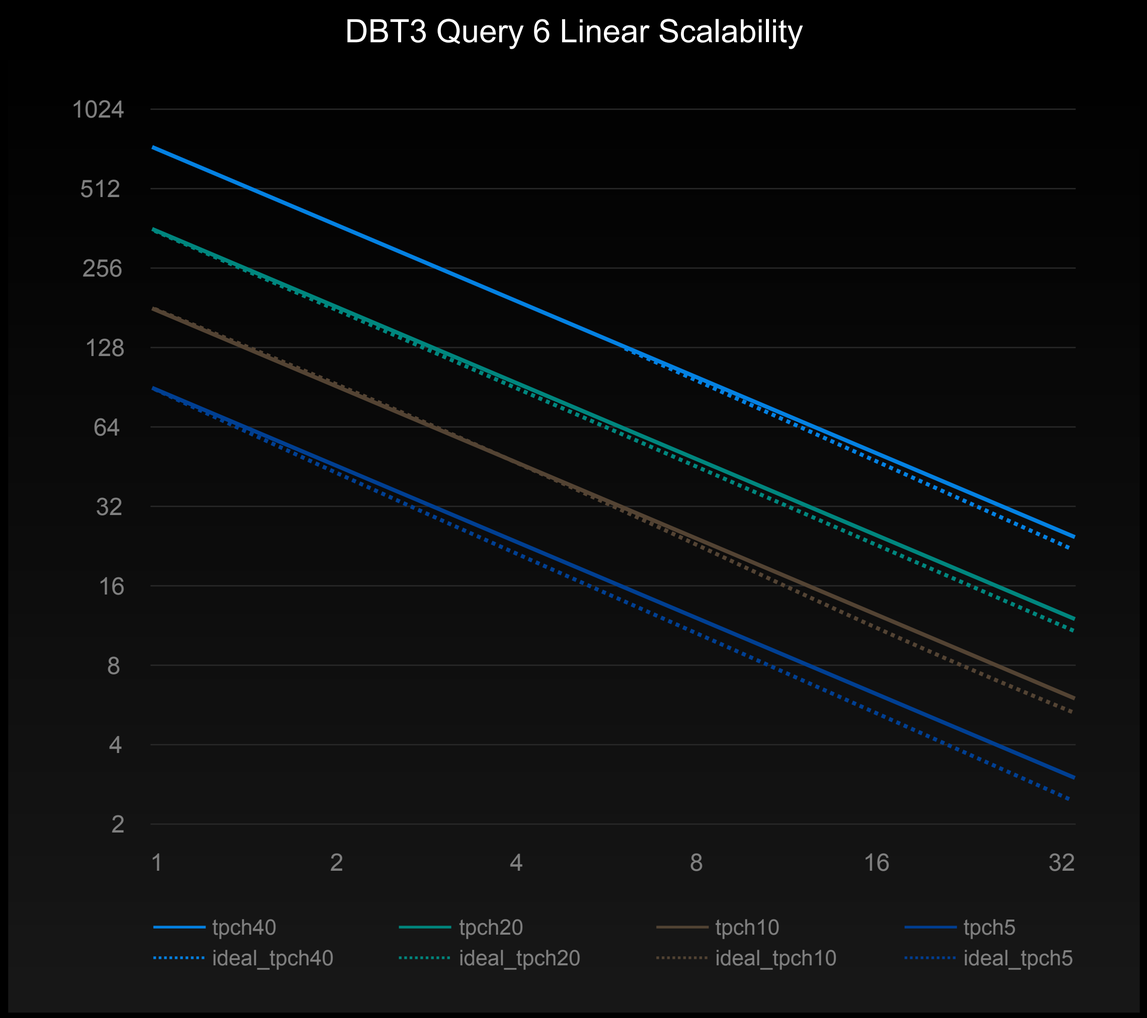
HTAP —
Hybrid Transaction and Analytical Processing



HTAP — Parallel Query
Reduce Latency of Complex Queries



One Query
Multiple Workers on Server
Parallel Scan on Storage Engine.



Alibaba Group 阿里巴巴集团

POLARDB — Database for the Cloud

- **Separation of Storage and Compute**
  - Independent scaling
  - Lower cost
- **Shared Storage**
  - High throughput
  - Low latency
  - High availability
  - Fast scaling (no data copy)
- **Physical replication**
  - Less I/O
  - Non-blocking DDL
  - Efficient parallel redo on slaves
- **Parallel Query Execution**
  - Lower latency for complex queries

Alibaba Group
阿里巴巴集团

Thank You