Burning Money

Current State:

we have currently 3 replicas of DB and they are at 90% utilization. Latency is stuck at 20ms. We are not able to serve the more number of people as traffic count and increases and we have limited replica.

20ms(latency) * 50(DB request/calls) = 1000ms and hence causing delay.

The Cost of Inaction:

At 20ms, we are losing 0.2% in processing delays and also 9% in decision change . Our daily business is of 10,00,000 and loss of Rs2000 per day.

The Cost of Out of Service:

Daily user visit on average 50,000 but through last month we are noticing the average users is increased by 20,000 by max traffic of 1,00,000 on last week. But due to insufficient DB and response delay we are not able to serve users causing loss of RS 2 lakh daily.

The Goal:

Sub-1ms latency to enable "Instant-Action" features such as high-frequency trading, real-time bidding, or global sync.

The Architecture

Geographic Distribution:

It's not just a clone add, it will move of data towards the Edge. More replica offer user to access the data faster, without delay and data availability will increase

Read-Heavy Scaling:

Moving 100% of read traffic from the primary to 10 replicas helps keep the primary healthy DB for writes. We will be heavily secured and fault tolerated offing user's free state of mind.

Total Throughput capacity goes up by 333%.

The Risk and Mitigation

Risk 1: Replication Lag.

With 10 nodes, it becomes more difficult to maintain all nodes as "Current."

Mitigation:

Asynchronous replication for non-critical reads and semi-sync replication for critical reads.

Risk 2: Consistency Problems.

Users may view **old** data for a few milliseconds.

Mitigation:

"Read-your-writes" consistency rules should be implemented at the application level. Using the primary DB in case of read write request simultaneously.

Risk 3: Management Complexity.

With 10 servers, there is a 10 times greater chance of a disk failure.

MITIGATION:

Each server can maintain the copy of another server and help other in case of fault.