CS 739 Demo

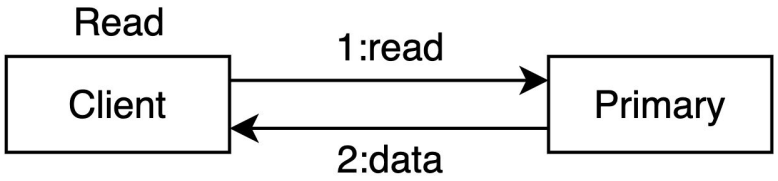
FireStore - Replicated Block Storage

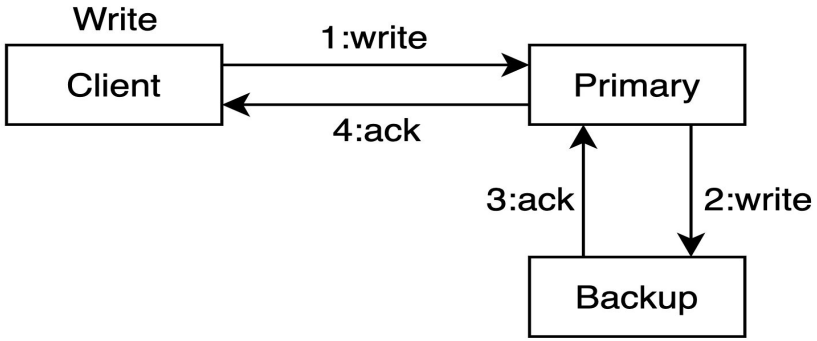
Akshat Sinha
Himanshu Sagar
Kaushik Kota

FireStore - Details

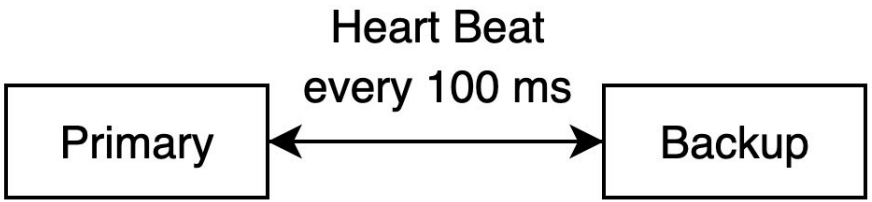
- Replication Strategy: Primary - Backup
- Client connects to Primary only
  - If primary is down, connects to Backup
- Data is stored as single log file
- Read and write to Primary

Write & Read Protocol



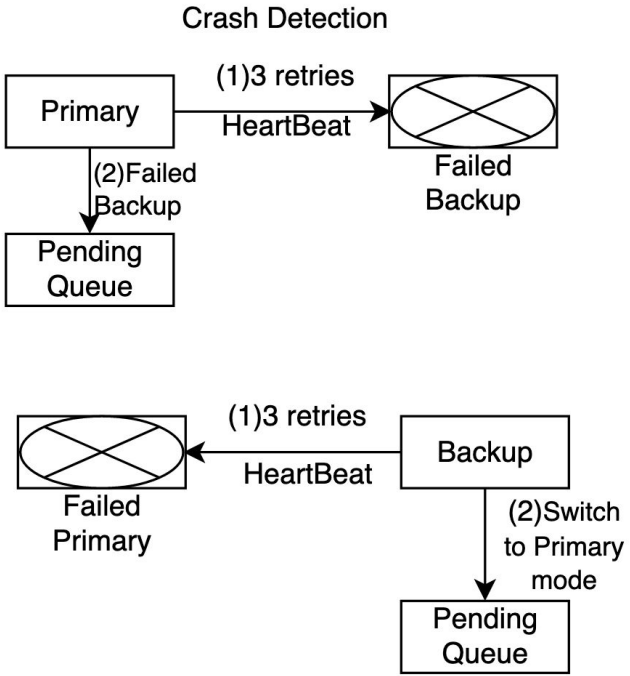


Primary and Backup HeartBeat



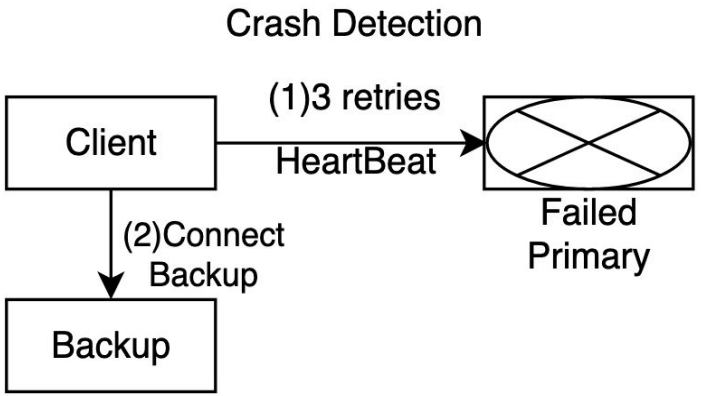
Crash Detection at Primary/Backup

- Retries 3 times
- Pending queue gets created
- All new write requests are queued
- Condition variable is used to wait



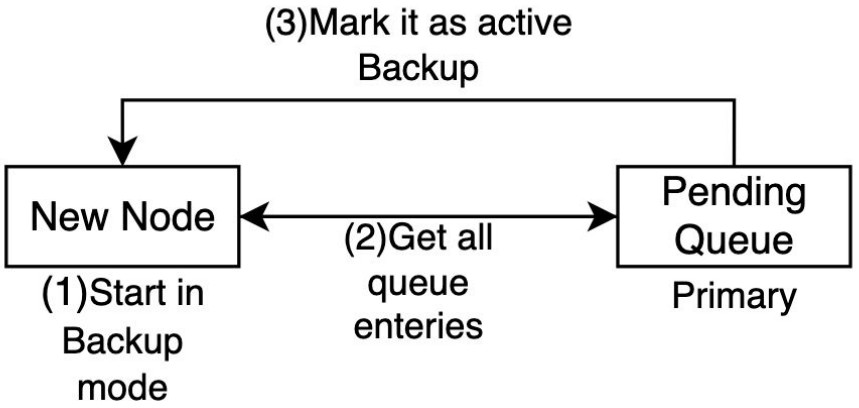
Crash Detection at Client

- Backup failure: No impact
- Primary Failure: Connect to Backup



Crash Recovery at Primary/Backup

- New node is always Backup
- Signal condition variable
- New writes are queued
- Once queue is empty
  - Mark it active

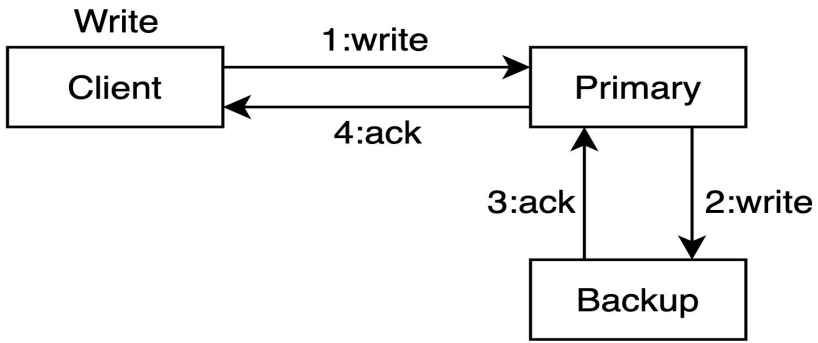


Hardware Specifications

- 3 CloudLab Nodes
- Hardware Type: c220g5
  - CPU: Two Intel Xeon Silver 4114 10-core CPUs at 2.20 GHz
  - RAM: 192GB ECC DDR4-2666 Memory
  - Disk: One 1 TB 7200 RPM 6G SAS HDs + One Intel DC S3500 480 GB 6G SATA SSD
  - NIC: Dual-port Intel X520-DA2 10Gb NIC (PCIe v3.0, 8 lanes) + Onboard Intel i350 1Gb

Crash Points

```
enum S_POINTS
{
    S_NO_CRASH = 0,
    PRIMARY_AFTER_WRITE_REQ_RECV = 1,
    PRIMARY_AFTER_WRITE = 2,
    BACKUP_AFTER_WRITE_REQ_RECV = 3,
    BACKUP_AFTER_WRITE = 4,
    PRIMARY_AFTER_ACK_FROM_B = 5,
    PRIMARY_AFTER_ACK_TO_C = 6,
};
```



Demo - Primary Crash and Recovery

Correctness - Test

Ground truth: All writes to local log file

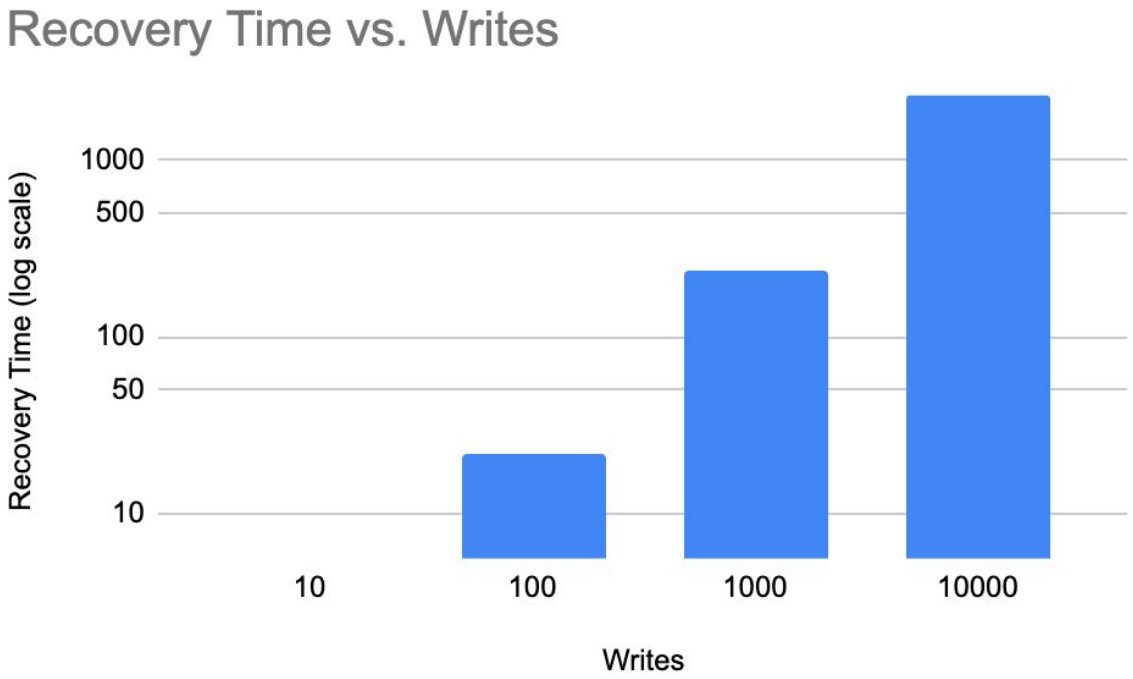
Workload: Single client

- Random read/write
- Sequential read/write
- Repeated with and without failures

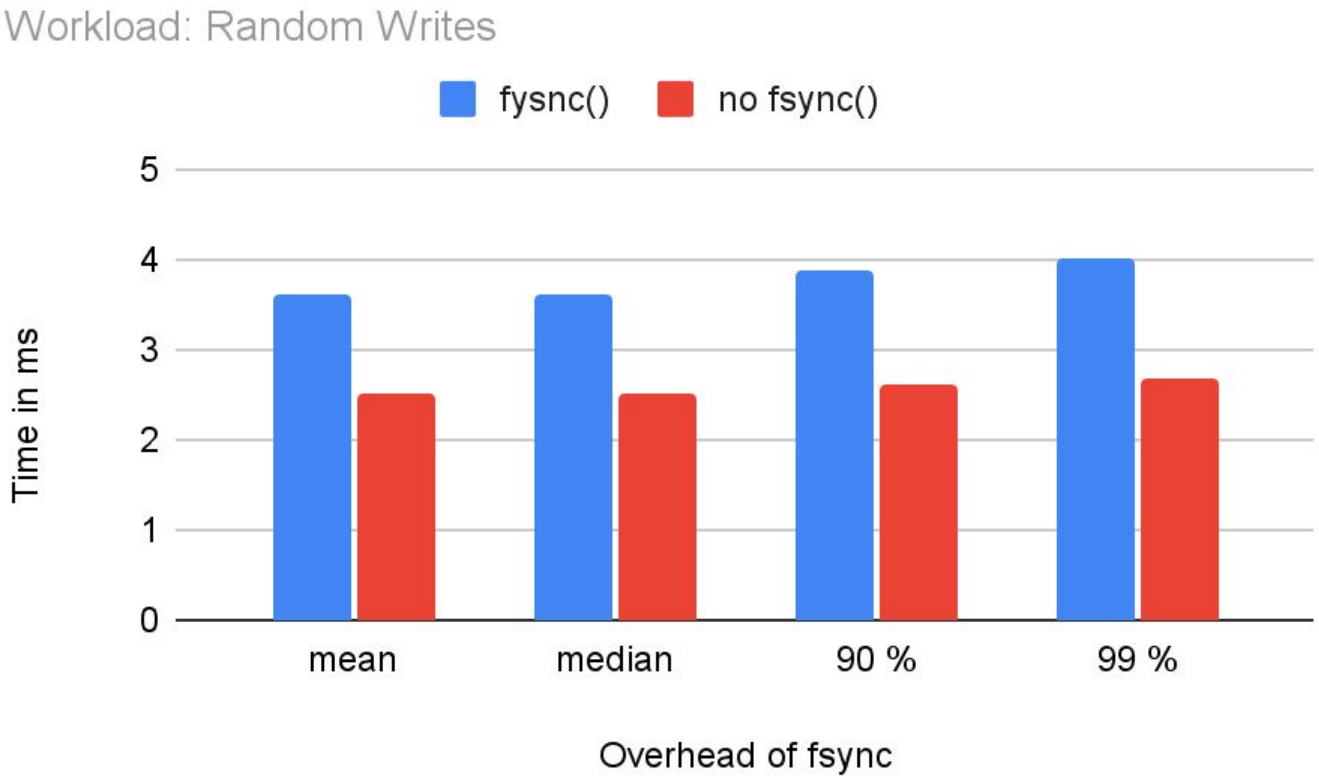
Correctness Verification

- Checksum is checked at every read
- Need to explore for concurrent clients

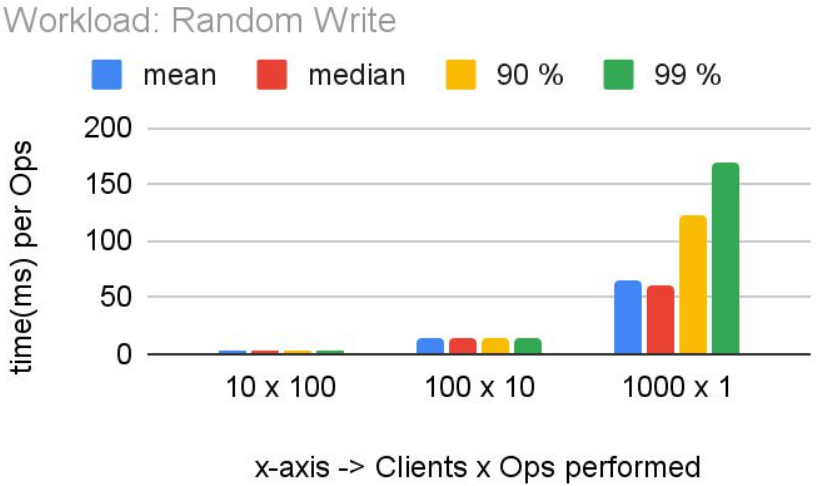
Recovery Time of Backup with varying workloads

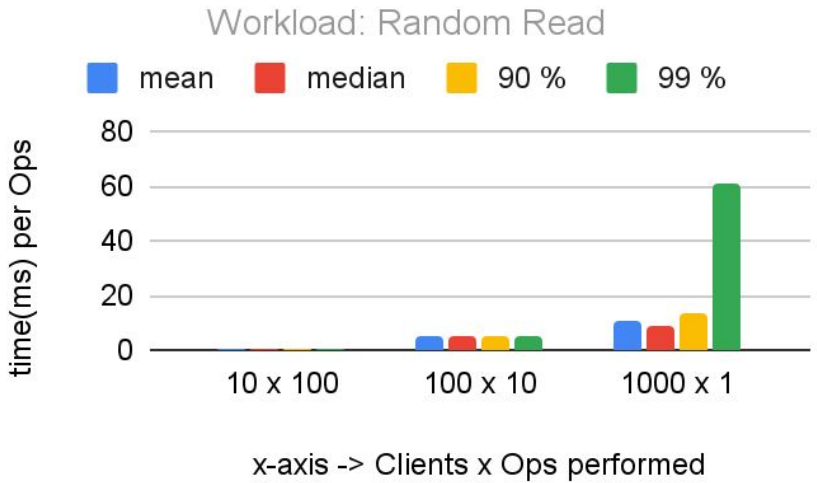


Latency measurement of Write with and without fsync()



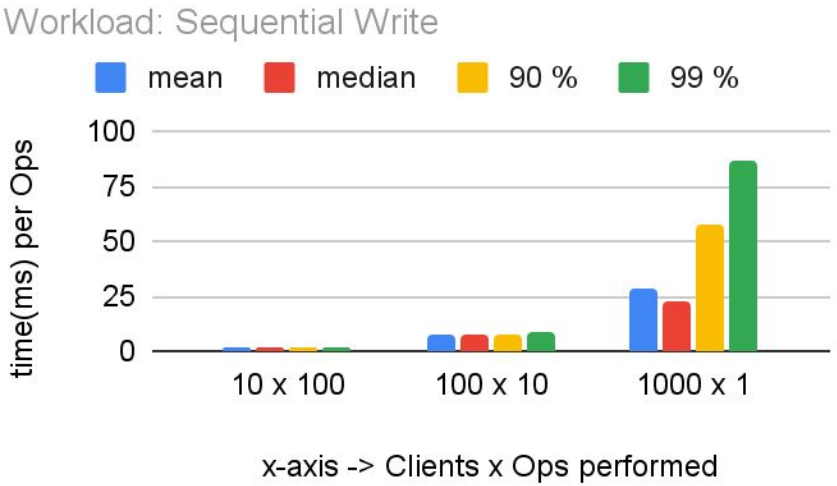
Read/Write Latency measurements with varying clients and operations





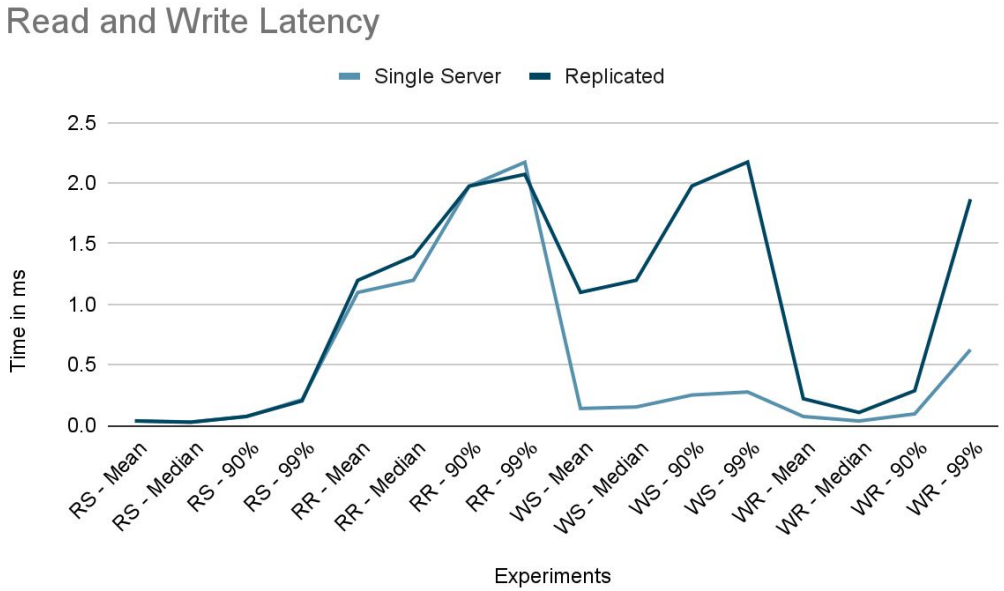
10 x 100 implies 10 clients concurrently doing 100 operations

Read/Write Latency measurements with varying clients and operations



Results - Read/Write

Reads and Writes(Sequential vs Random) single client (x1000 ops)



Legend:

RS - Read Sequential

RR - Read Random

WS - Write Sequential

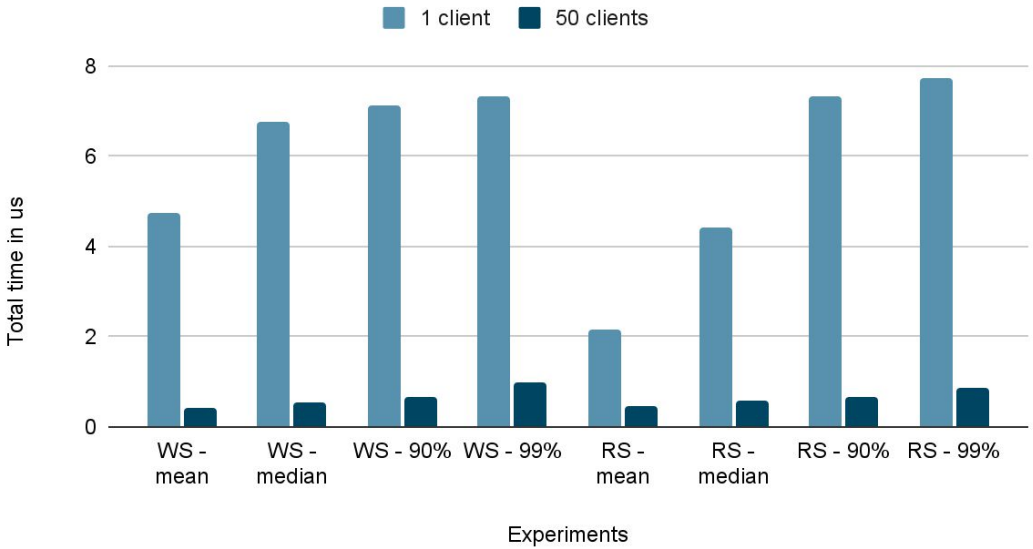
WR - Write Random

Results - Read/Write

Reads and Writes(Sequential vs Random) multiple client

50 clients (4 ops) vs 1 client (200 ops)

Latency Measurements



Legend:

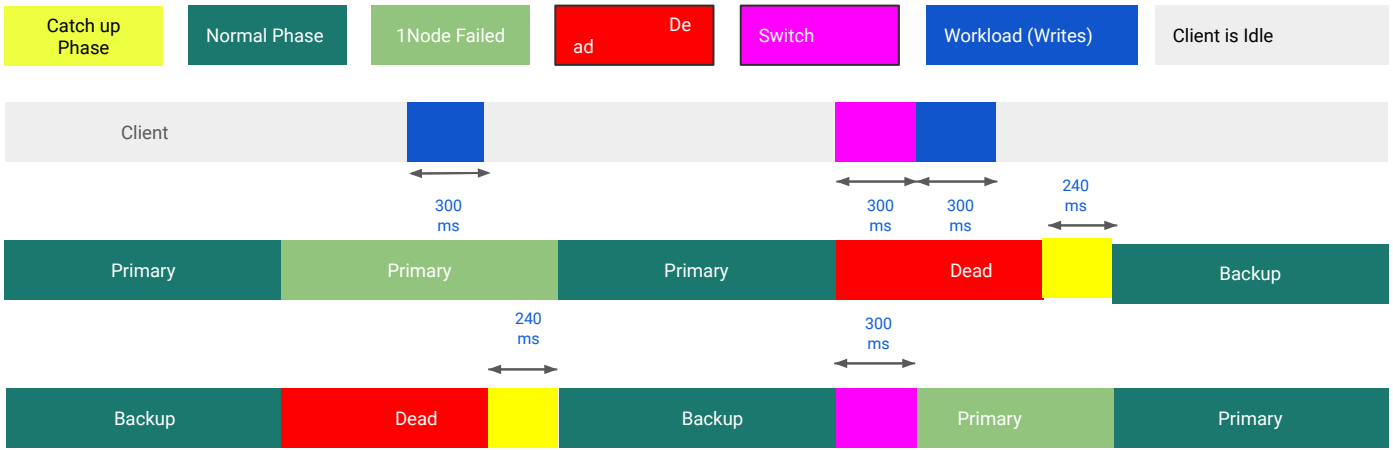
RS - Read Sequential

RR - Read Random

WS - Write Sequential

WR - Write Random

Crash Detection & Recovery - Timeline of actual events



Questions ?