DO YOU REALLY NEED A DISTRIBUTED SYSTEM?

W-PSDS'19

Bradley King

Field CTO Scality

LIGHTKONE

Lightweight computation for networks at the edge

SCALITY

DO YOU REALLY NEED A DISTRIBUTED SYSTEM?

SCALITY

- Scalability
- Resilience
- Redundancy
- Economy
- Overcome latency
- Parallelism

- Partitions are possible
- Complexity
- Correctness compromises
- Synchronization worries

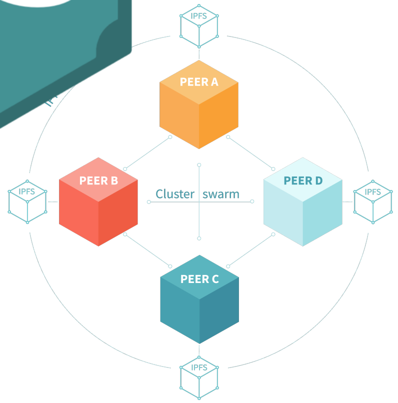
ORACLE DATABASE

Strong Serializable Consistency

Google Spanner

Azure Cosmos DB

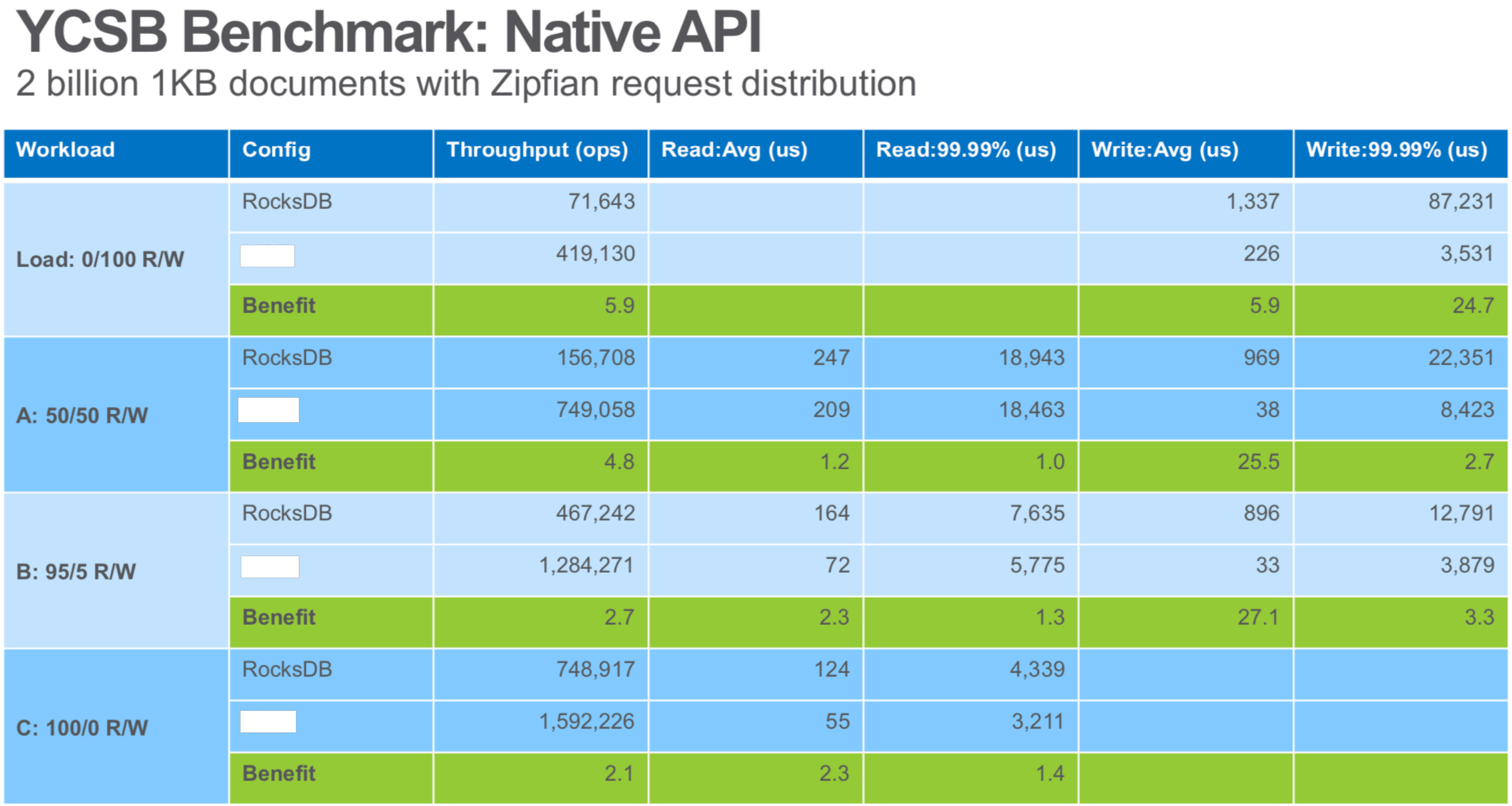
IPFS



cassandra

Eventually Consistent or hopefully consistent or?

SOME LIMITS OF VERTICAL SCALING

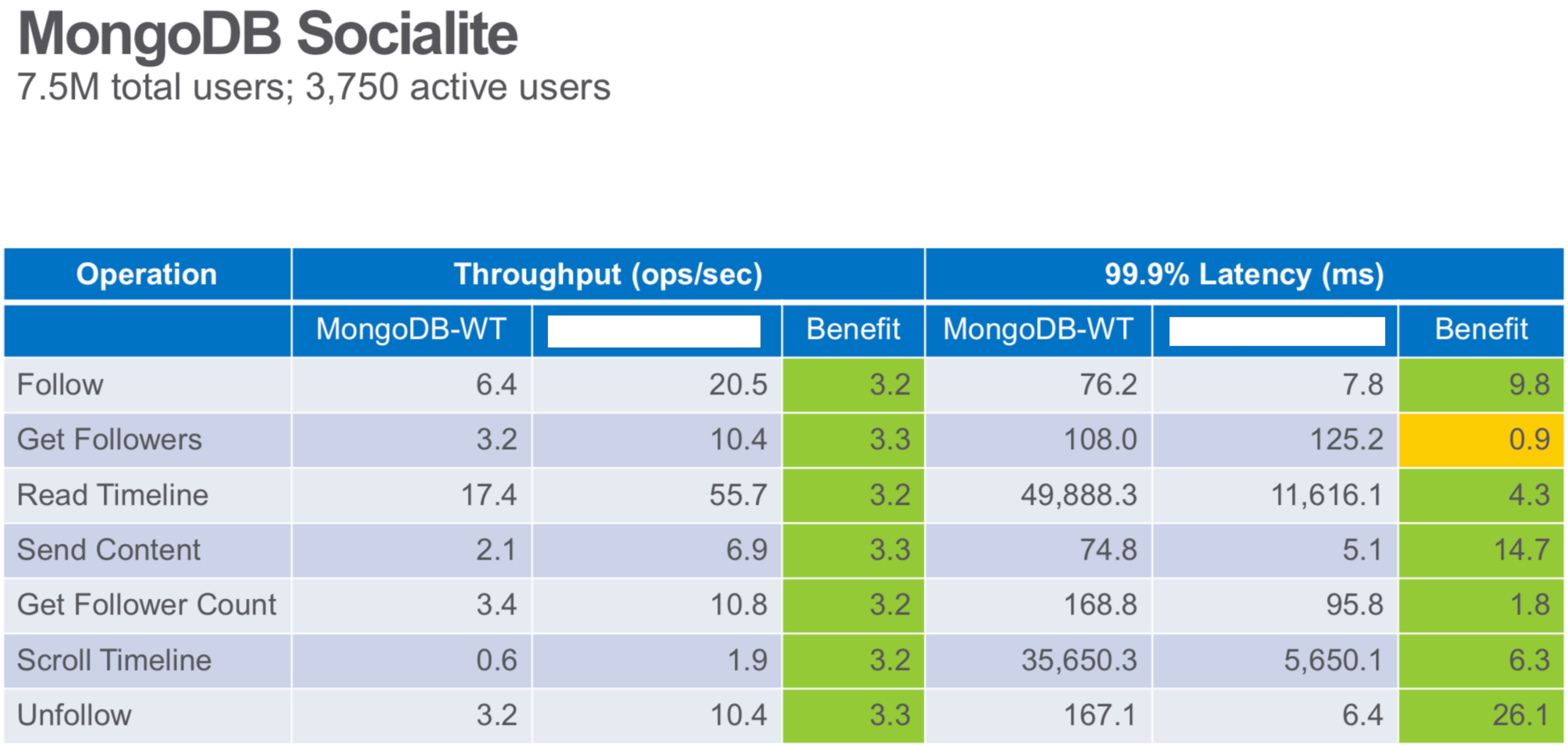


YCSB Benchmark: Native API

2 billion 1KB documents with Zipfian request distribution

Workload	Config	Throughput (ops)	Read:Avg (us)	Read:99.99% (us)	Write:Avg (us)	Write:99.99% (us)
Load: 0/100 R/W	RocksDB	71,643			1,337	87,231
	[redacted]	419,130			226	3,531
	Benefit	5.9			5.9	24.7
A: 50/50 R/W	RocksDB	156,708	247	18,943	969	22,351
	[redacted]	749,058	209	18,463	38	8,423
	Benefit	4.8	1.2	1.0	25.5	2.7
B: 95/5 R/W	RocksDB	467,242	164	7,635	896	12,791
	[redacted]	1,284,271	72	5,775	33	3,879
	Benefit	2.7	2.3	1.3	27.1	3.3
C: 100/0 R/W	RocksDB	748,917	124	4,339		
	[redacted]	1,592,226	55	3,211		
	Benefit	2.1	2.3	1.4		

SOME MORE LIMITS OF VERTICAL SCALING



MongoDB Socialite

7.5M total users; 3,750 active users

Operation	Throughput (ops/sec)			99.9% Latency (ms)		
	MongoDB-WT	[redacted]	Benefit	MongoDB-WT	[redacted]	Benefit
Follow	6.4	20.5	3.2	76.2	7.8	9.8
Get Followers	3.2	10.4	3.3	108.0	125.2	0.9
Read Timeline	17.4	55.7	3.2	49,888.3	11,616.1	4.3
Send Content	2.1	6.9	3.3	74.8	5.1	14.7
Get Follower Count	3.4	10.8	3.2	168.8	95.8	1.8
Scroll Timeline	0.6	1.9	3.2	35,650.3	5,650.1	6.3
Unfollow	3.2	10.4	3.3	167.1	6.4	26.1

LARGE EMAIL PLATFORMS

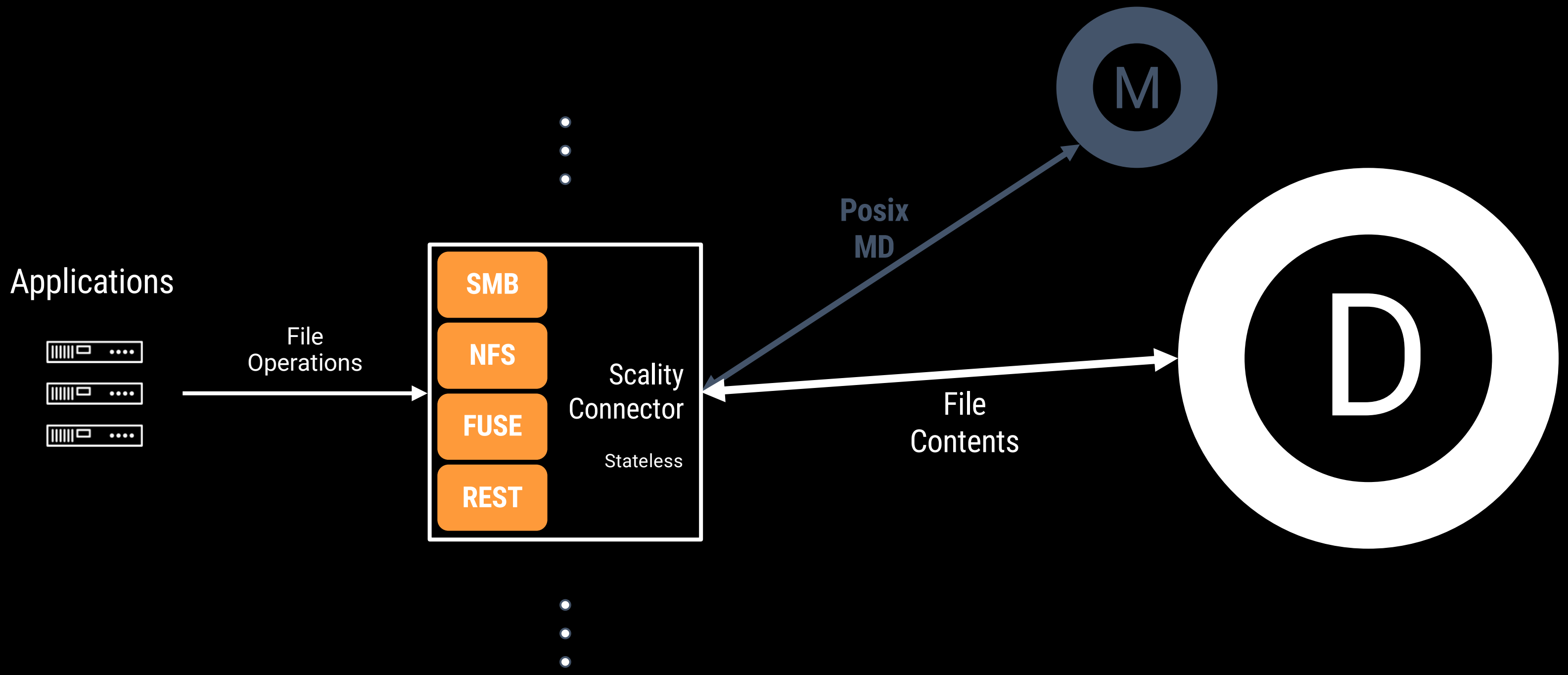
- 50~150Billion messages causes inode issues on all traditional filesystems
- Data volumes 5~15PB
- > 100K R/W IOPS + 40K deletes/sec
- 3~30 million users on business or consumer systems, outages cause support disasters, 100% uptime is an expectation
- Traditional IT Tools to handle loads include load-balancing and sharding, but state cannot be load-balanced and sharding invariably struggles as volumes or load grows
- Data immutability allows an ideal environment for fully distributed shared nothing storage

CIRCULAR BUFFER OF LOG STORAGE

- > 1PB/day storage of logs
- > 300 servers generating or accessing data
- Using largest HDDs available 14-16TB 150MB/s/drive maximum : all drives must work together
- 20GB/s continuous writes, platform has ~ 500drives > 50GB/sec bandwidth
- Evenly distributed parallelism is essential – no centralized service component
- Allows the use of ordinary hardware for a HPC like workload

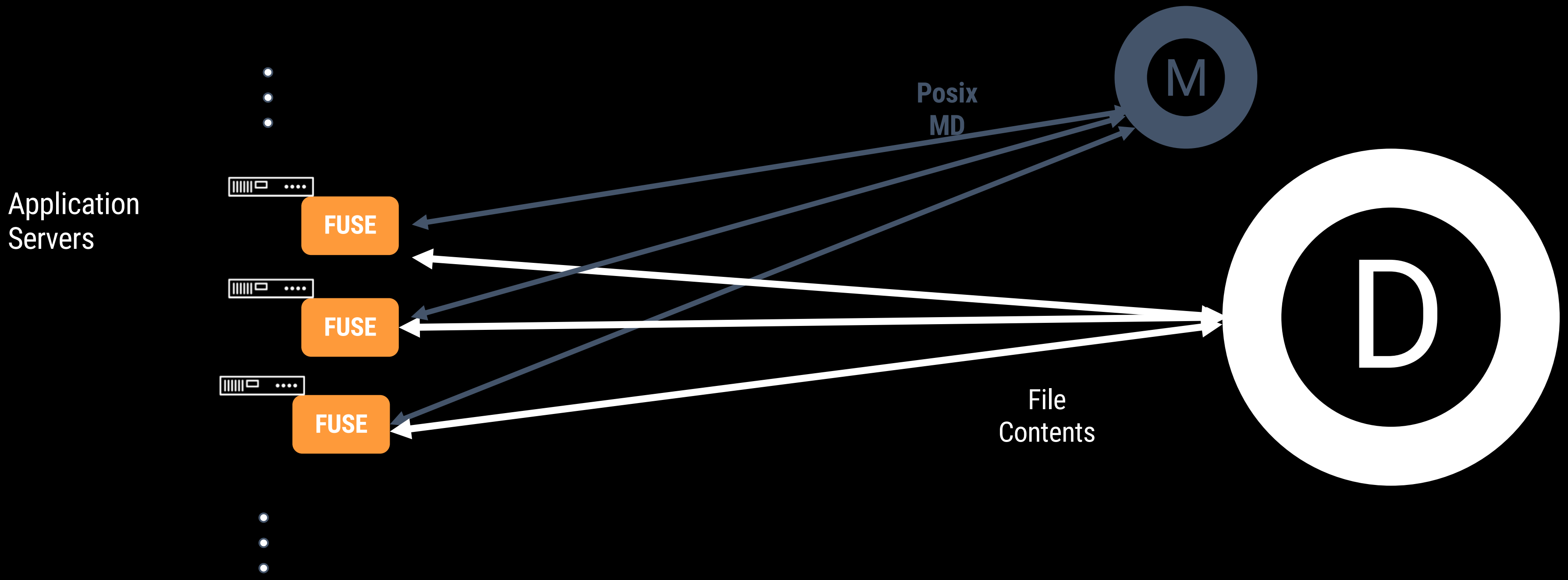
RING SCALE OUT FILE SYSTEM

SMBv3, NFSv3, Linux FUSE, and REST access
unlimited amount of volumes and files
distributed POSIX metadata · stateless connectors



SCALITY

Parallel scale-out with ~ 300 app servers



SCALITY

LIFE CRITICAL – HOSPITALS AND SOLAS SYSTEMS

- 1-2PB storage – single name-space
- Future growth is sustained and significant
- 24x7x365 availability is potentially life critical
- Historical and hi-resolution medical images are increasingly important in diagnosis and medical interventions

Hospices Civils de Lyon

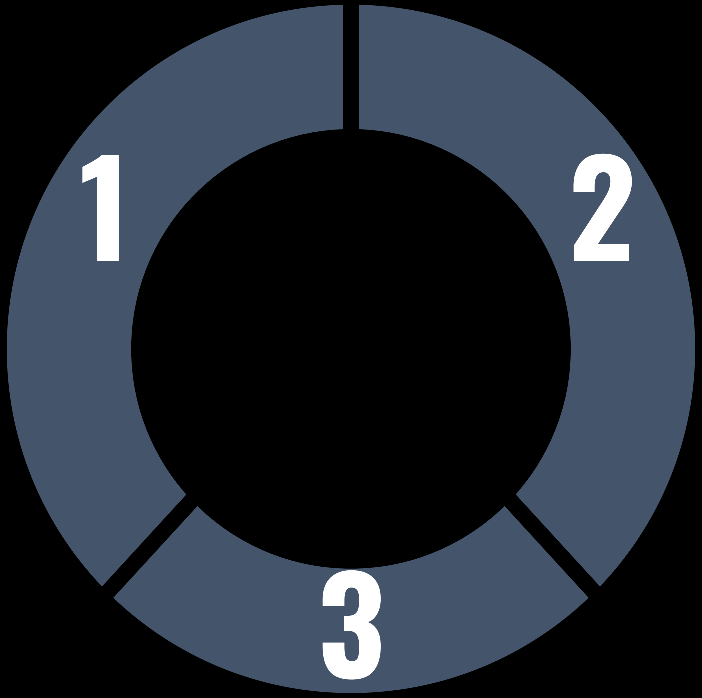
Using multi-site scale-out filesystem for HA/DR

3-SITE STRETCHED

synchronous operations across 3 sites
any single volume *belongs* to one site · any site can *host* volumes
best durability and storage efficiency combination of all multi-geo models

supports the failure of an entire site without service downtime
RPO = 0 · RTO = 0

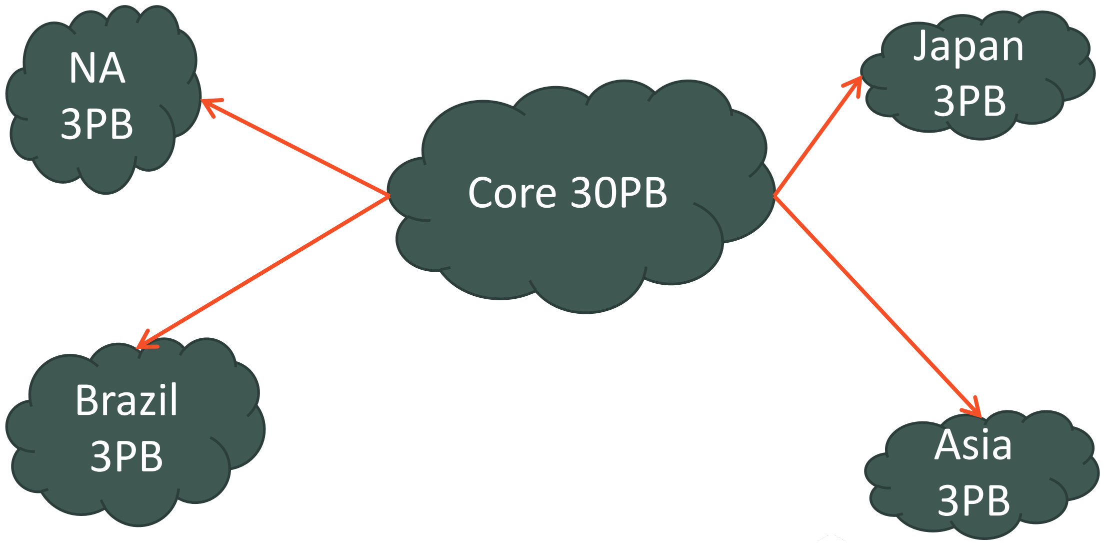
sites in the same metro area (<5ms latency)



SCALITY

INTERNET SCALE – VIDEO SITE

- > 30PB storage – single name-space
- Future growth is sustained and significant
- ~ 300 million unique visits/month
- Reliably low latency for recent data
- Acceptable latency for long tail
- 24h access



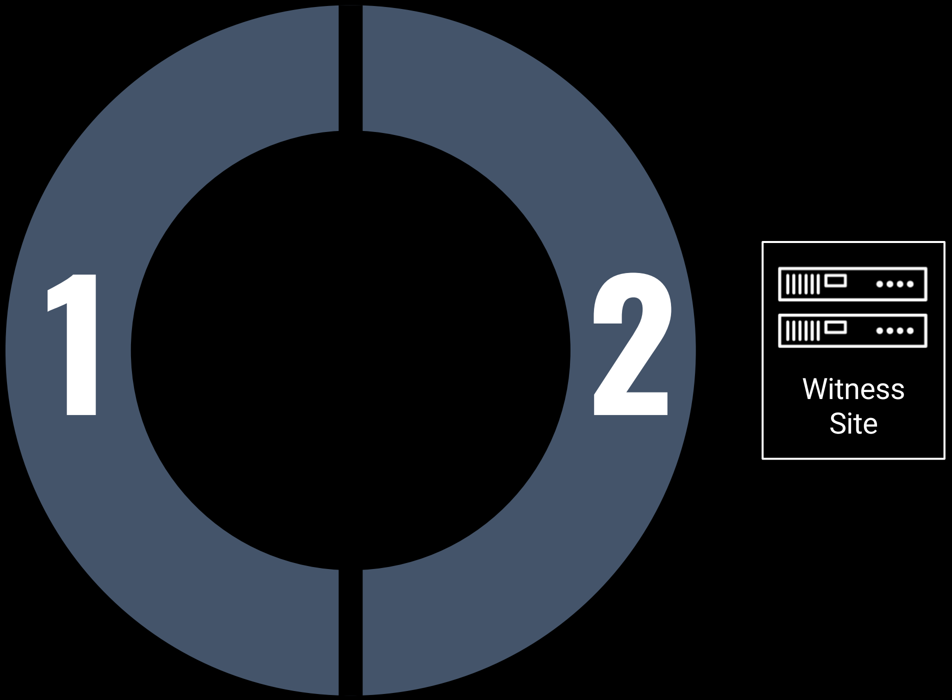
Using S3 multi-site for DR/HA

2-Site stretch immutable data with RAFT based Metadata replication

synchronous operations across 2 sites for data & across 2 sites + quorum site for metadata
active/active read/write access from everywhere
better durability & storage efficiency than 2-site asynchronously replicated

supports the failure of an entire site without service downtime
RPO = 0 · RTO = 0

sites in the same metro area (<10ms latency)



SCALITY

- If scalability demands it
- If growth is unbounded
- If availability is critical
- If the economics are better
- If data is immutable
- CRDTs for disconnected or many sites

- If 100% availability is unnecessary
- If vertical scaling is viable
- If your consistency contract requires it
- If CRDTs don’t apply