Millions of Tiny Databases

NSDI’20

Marc Brooker, Tao Chen and Fan Ping
February 2020
Table of contents

• Tough CAP Tradeoffs
• Availability and Blast Radius
• Physalia Architecture
Simplified Storage System

Client

Primary

Replica
Network Partition
I still want to be primary!

Please, sir, can I be the primary?
A Higher Power (Configuration Service)

I still want to be primary!

Please, sir, can I be the primary?

CONSISTENT!
LINEARIZABLE!
CAS!
A Higher Power (Configuration Service)

Please, sir, can I be the primary?
100k+ SERVERS.
Here?

Or

Here?
Partitions are not clean!
Availability and Blast Radius
Availability is typically improved with Redundancy
Availability
is typically improved with
Redundancy*

* Unless failures are correlated
Infrequent
Short
Small
Infrequent
Short
Small = Uncorrelated
“Blast Radius”
Physalia Architecture
Physalia *Cell*

- Replicated state machine
- Configuration for one volume, or small set of volumes.
- K/V store API
- Strict serializable transactions
Minimize The Radius

Client

Primary

Replica
Topology Details Matter

Lower Partition Risk
Lower Availability Blast Radius

More Redundancy
More Bisection Bandwidth
More Placement Options
Take Advantage of Eventual Consistency When You Can!

- Discovery Cache (clients discover nodes)
- Monitoring
- "Meta" control plane
Optimize for Blast Radius

• Minimize impact of partitions (and CAP tradeoffs),
• overload,
• software bugs &
• operational issues.

Build *humility* into the system.
Q&A
Marc Brooker
mbrooker@amazon.com
@marcjbrooker
The End