Millions of Tiny Databases

NSDI'20

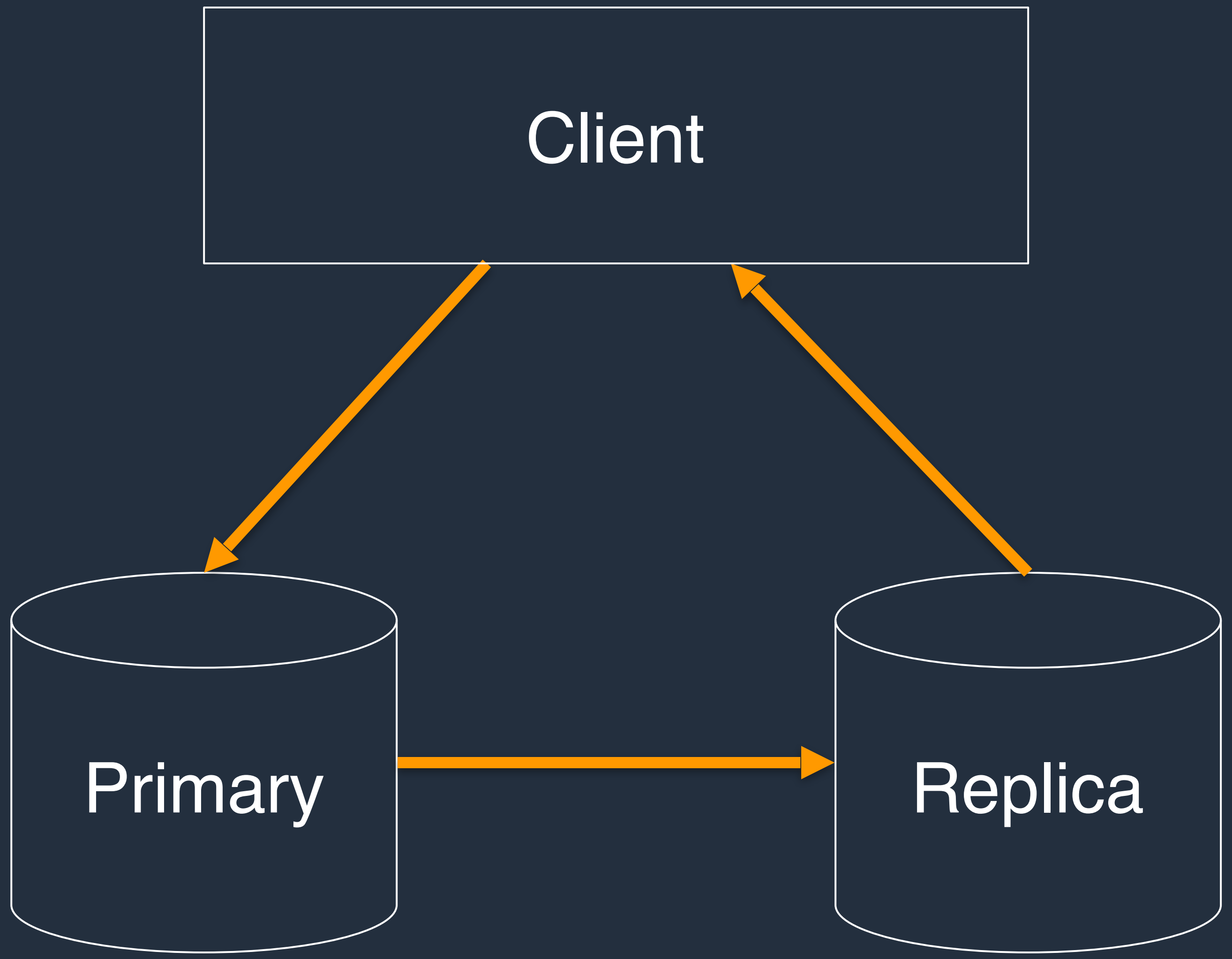
Marc Brooker, Tao Chen and Fan Ping

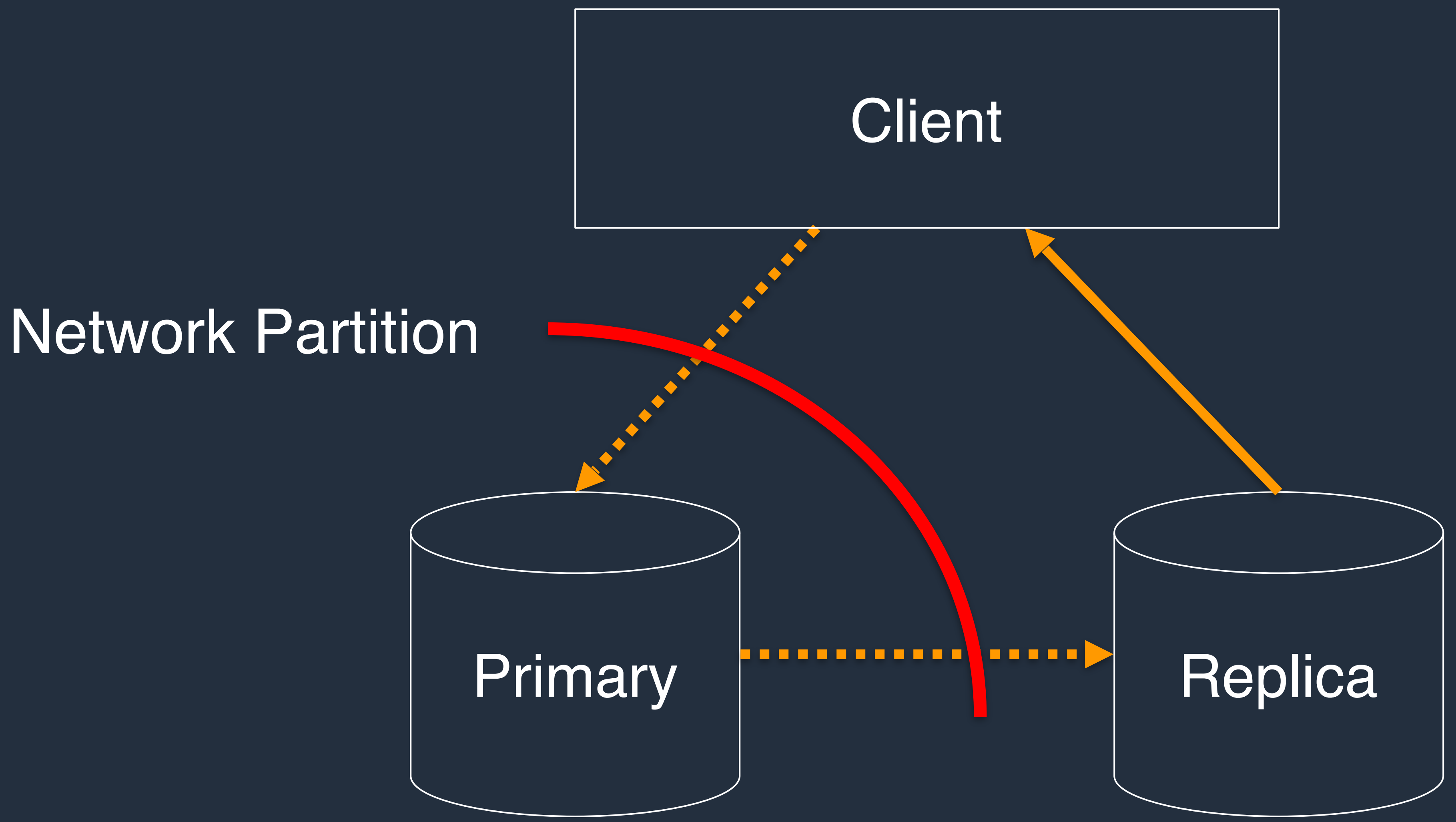
February 2020

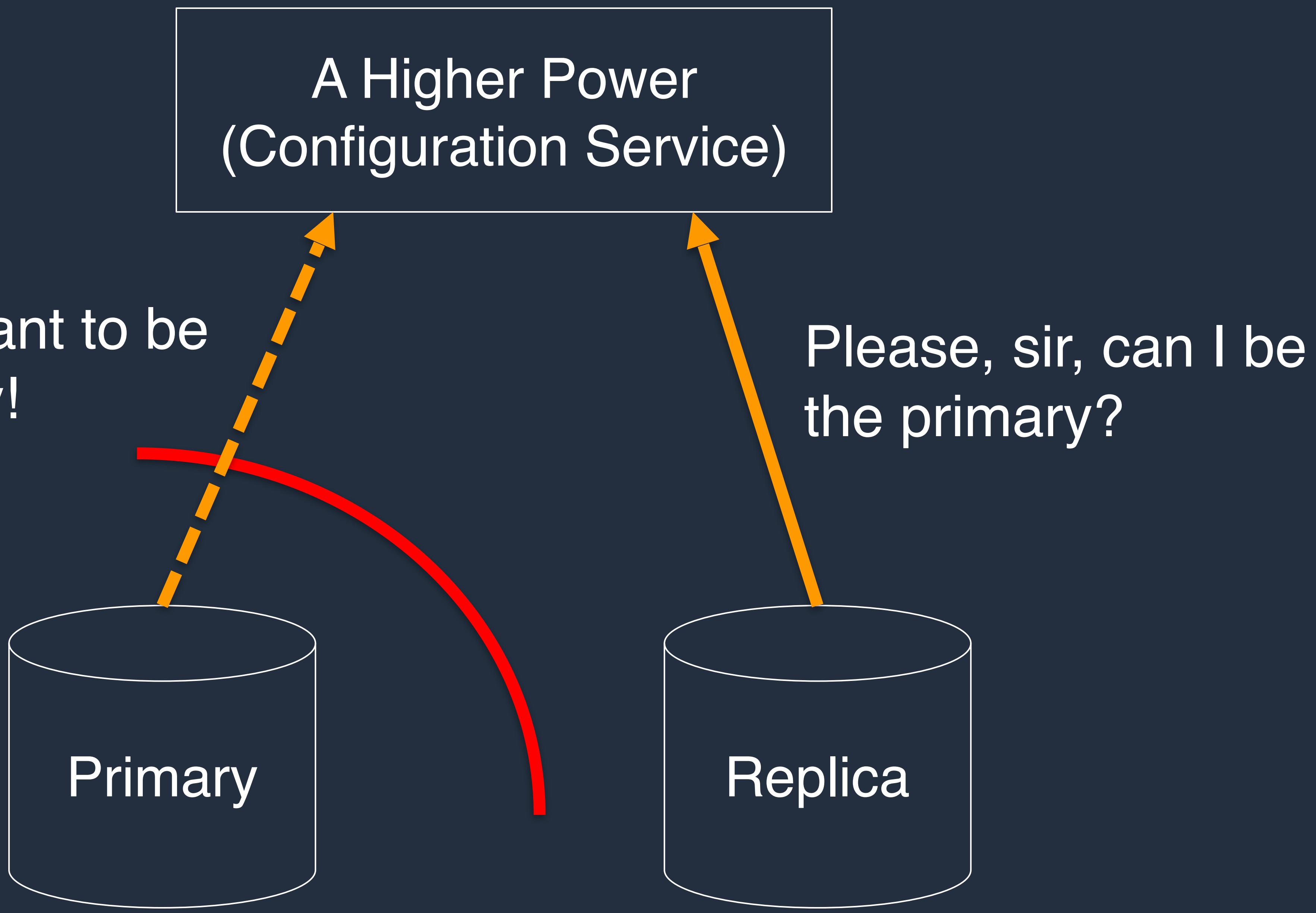
Table of contents

- Tough CAP Tradeoffs
- Availability and Blast Radius
- Physalia Architecture

Simplified Storage System




Client
Network Partition
Primary
Replica



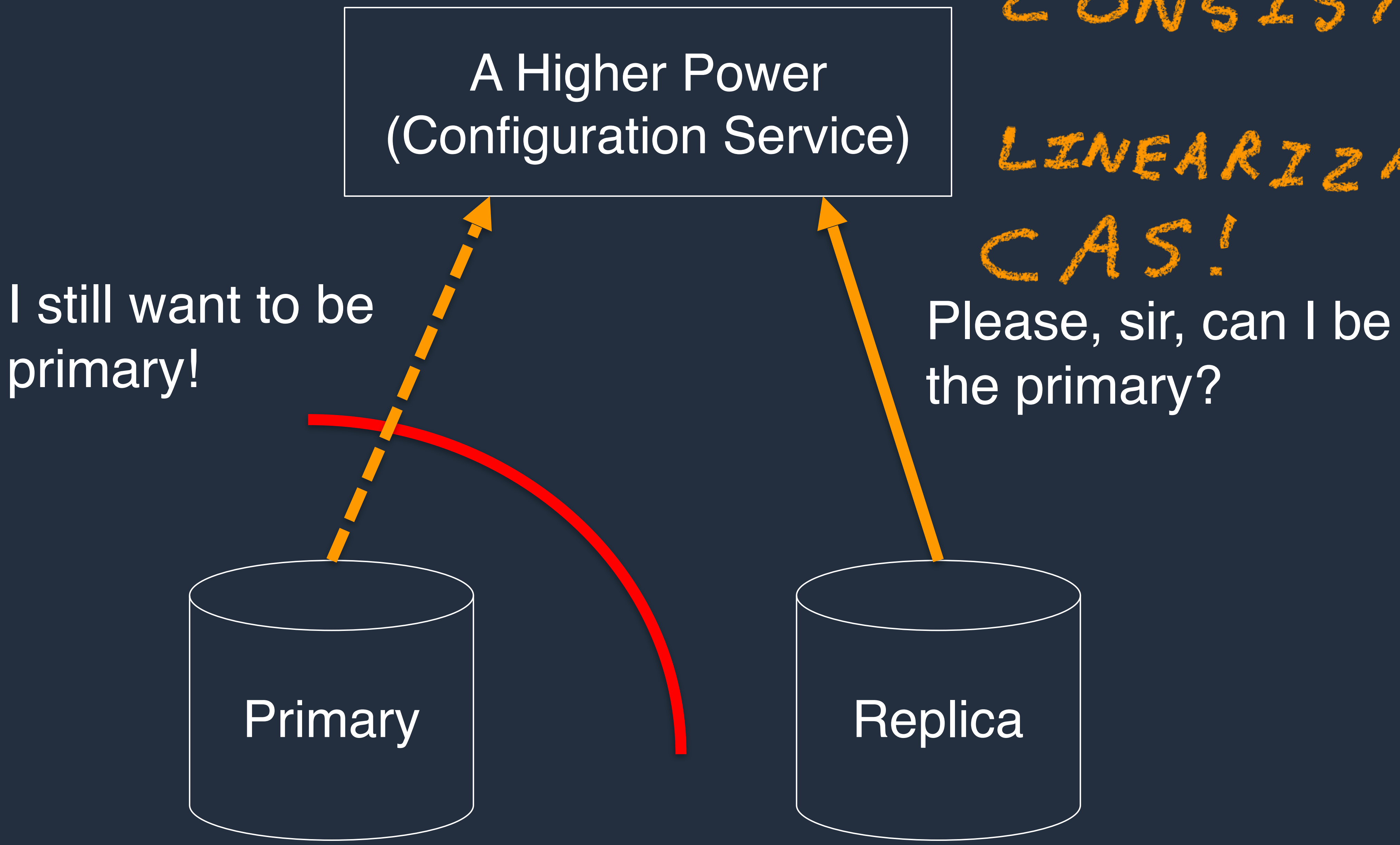
A Higher Power
(Configuration Service)

I still want to be primary!

Please, sir, can I be the primary?

Primary

Replica



A Higher Power
(Configuration Service)

I still want to be primary!

Please, sir, can I be the primary?

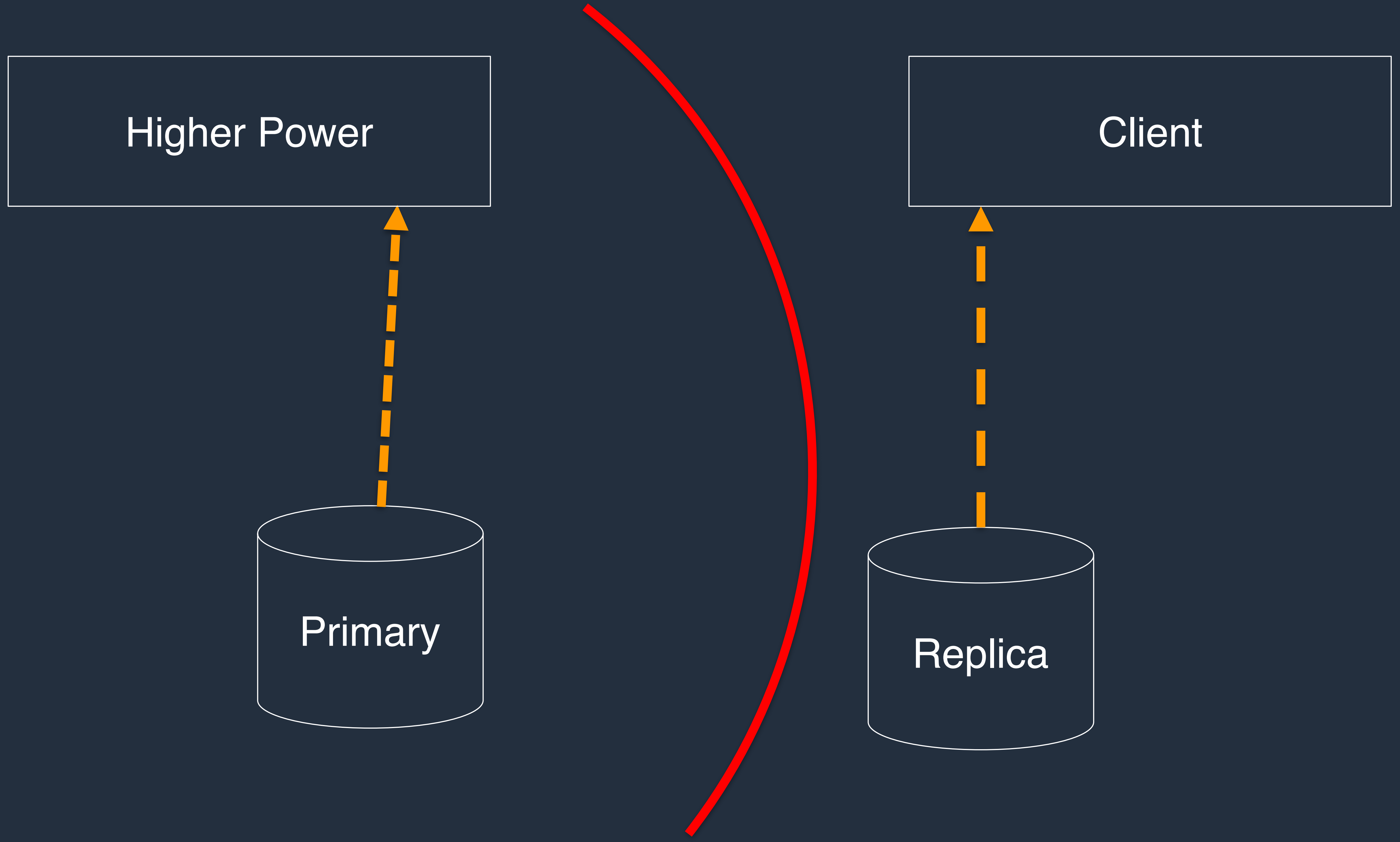
Primary

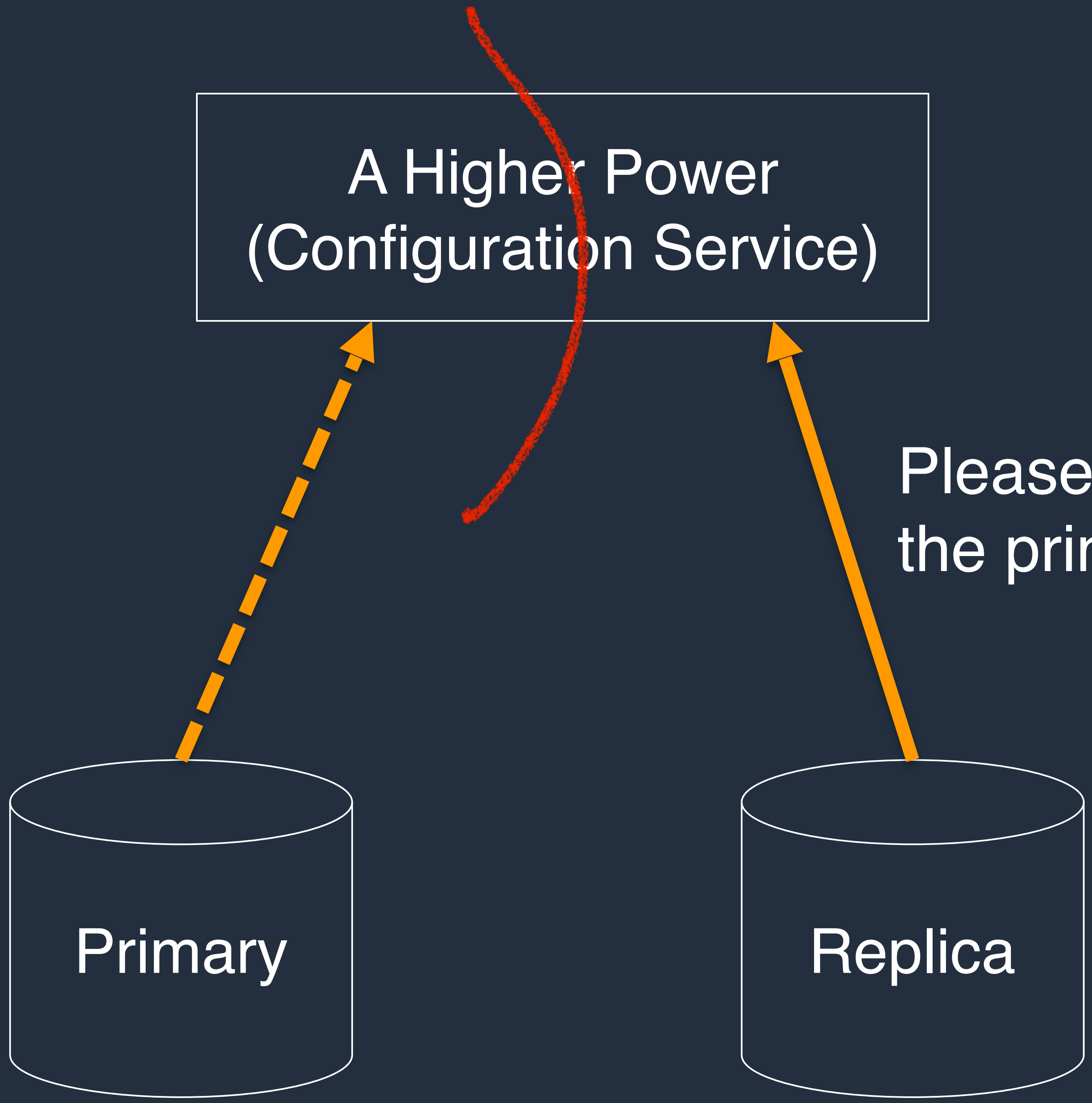
Replica

CONSISTENT!

LINEARIZABLE!

CAS!


Higher Power
Client
Primary
Replica

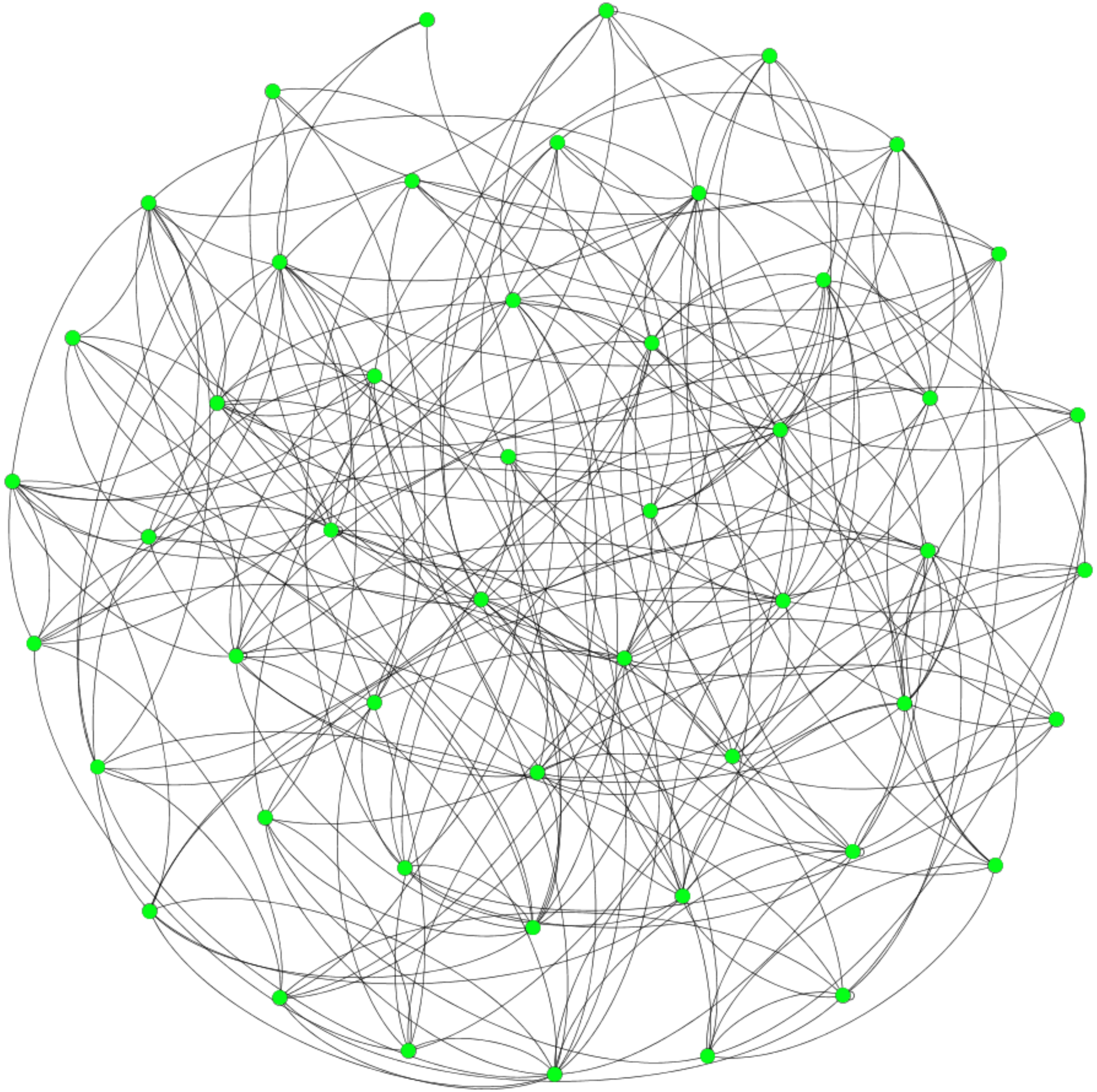


A Higher Power
(Configuration Service)

Primary

Replica

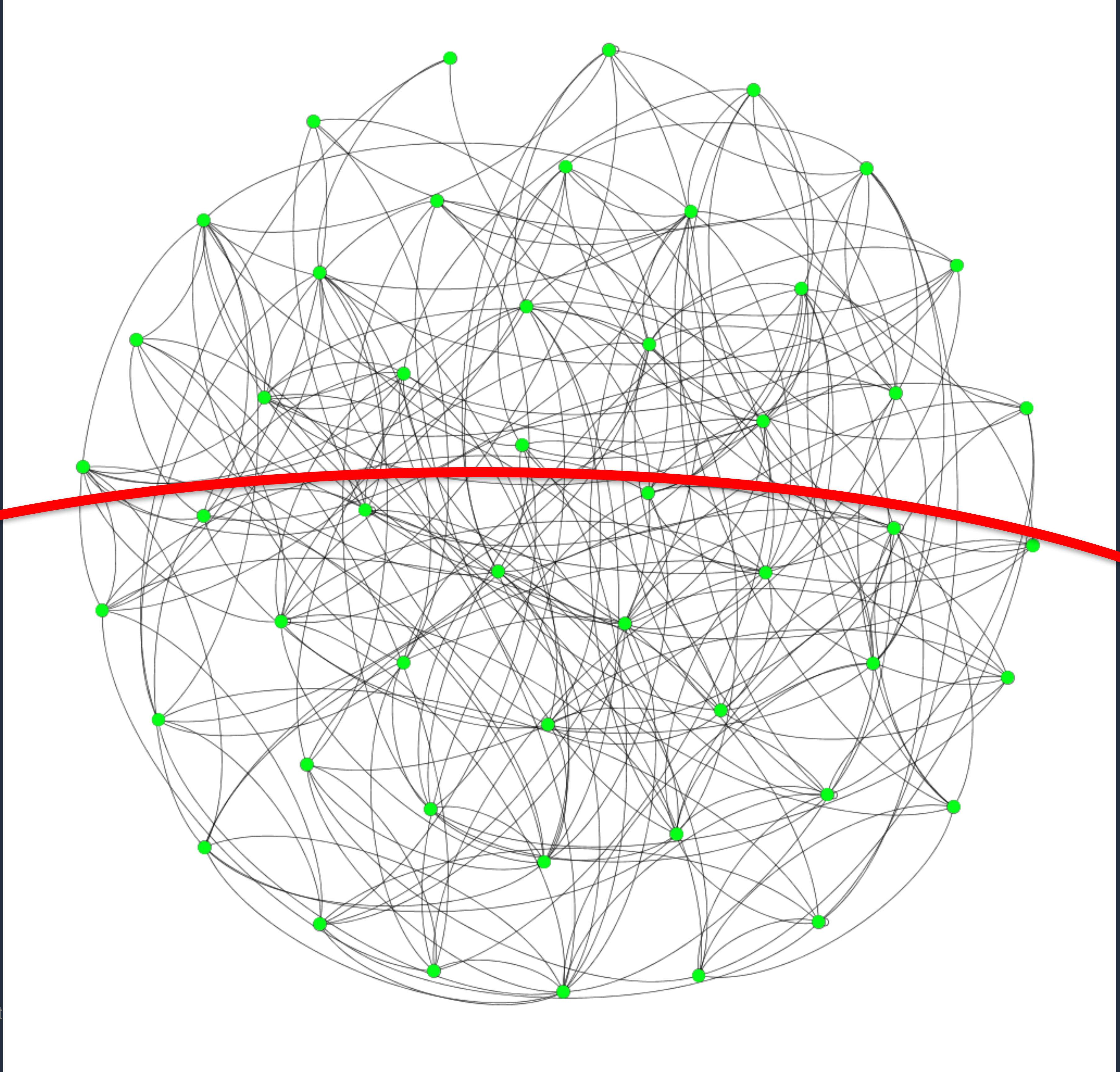
Please, sir, can I be
the primary?



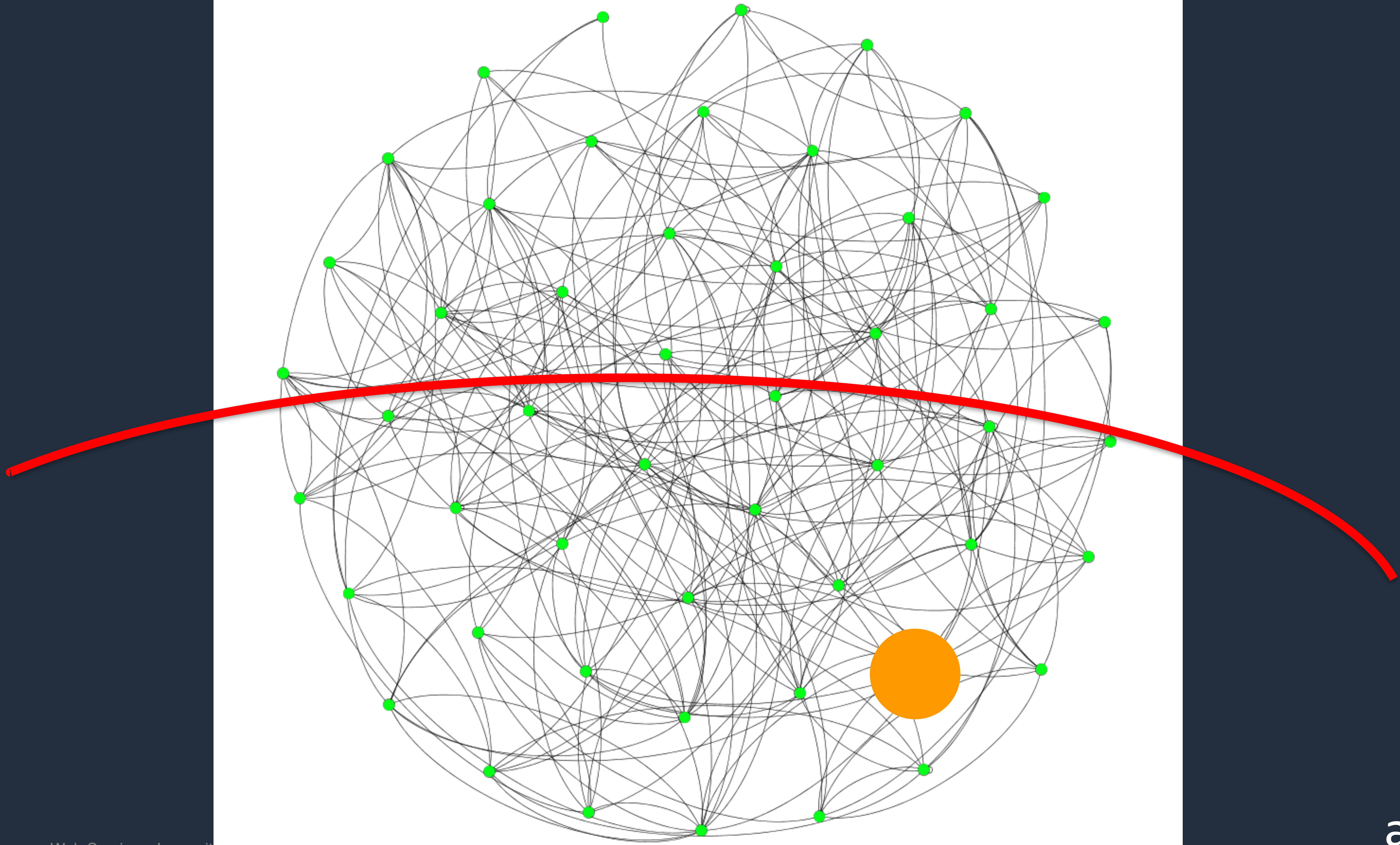
100k+

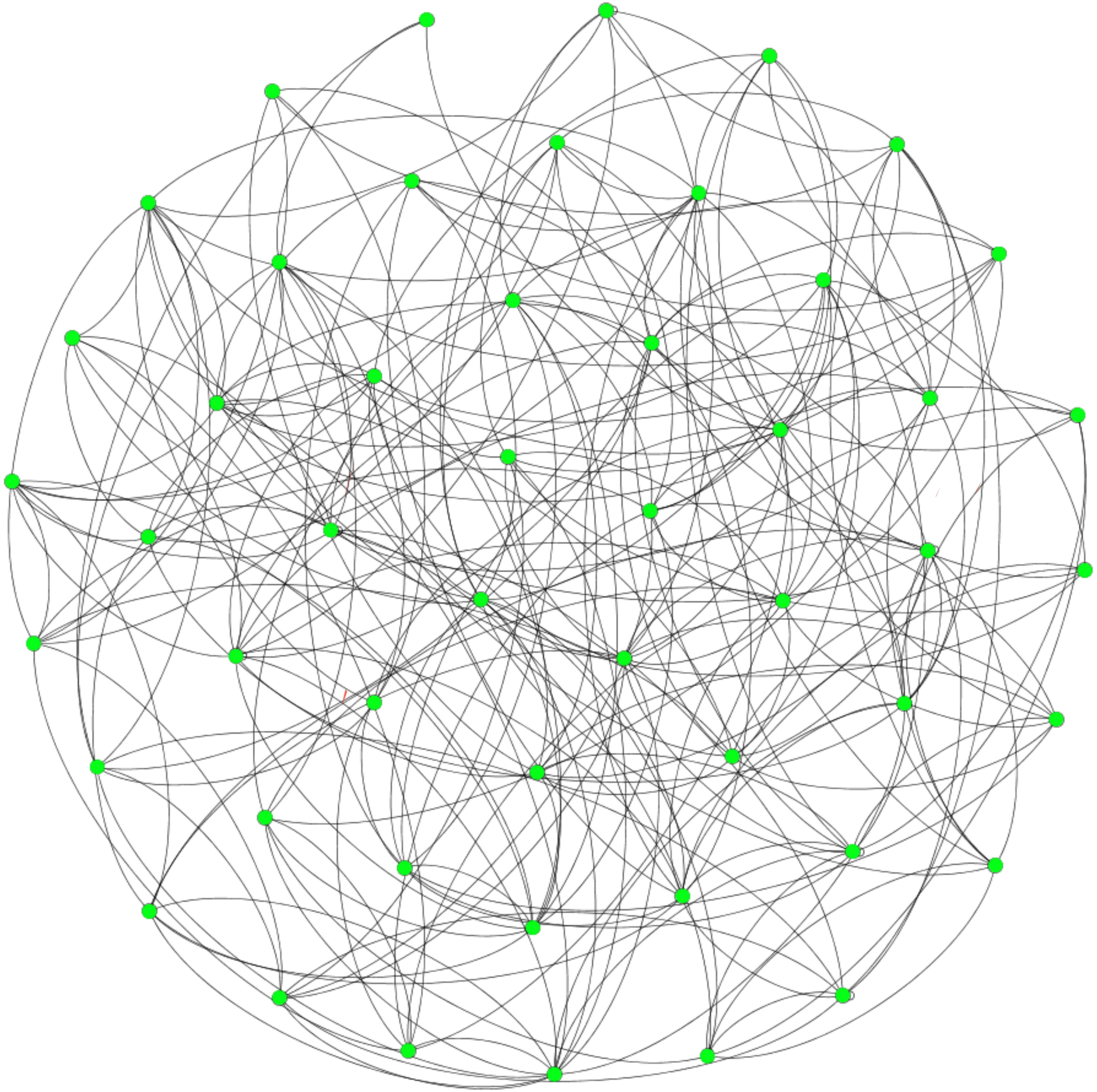
SERVERS.

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

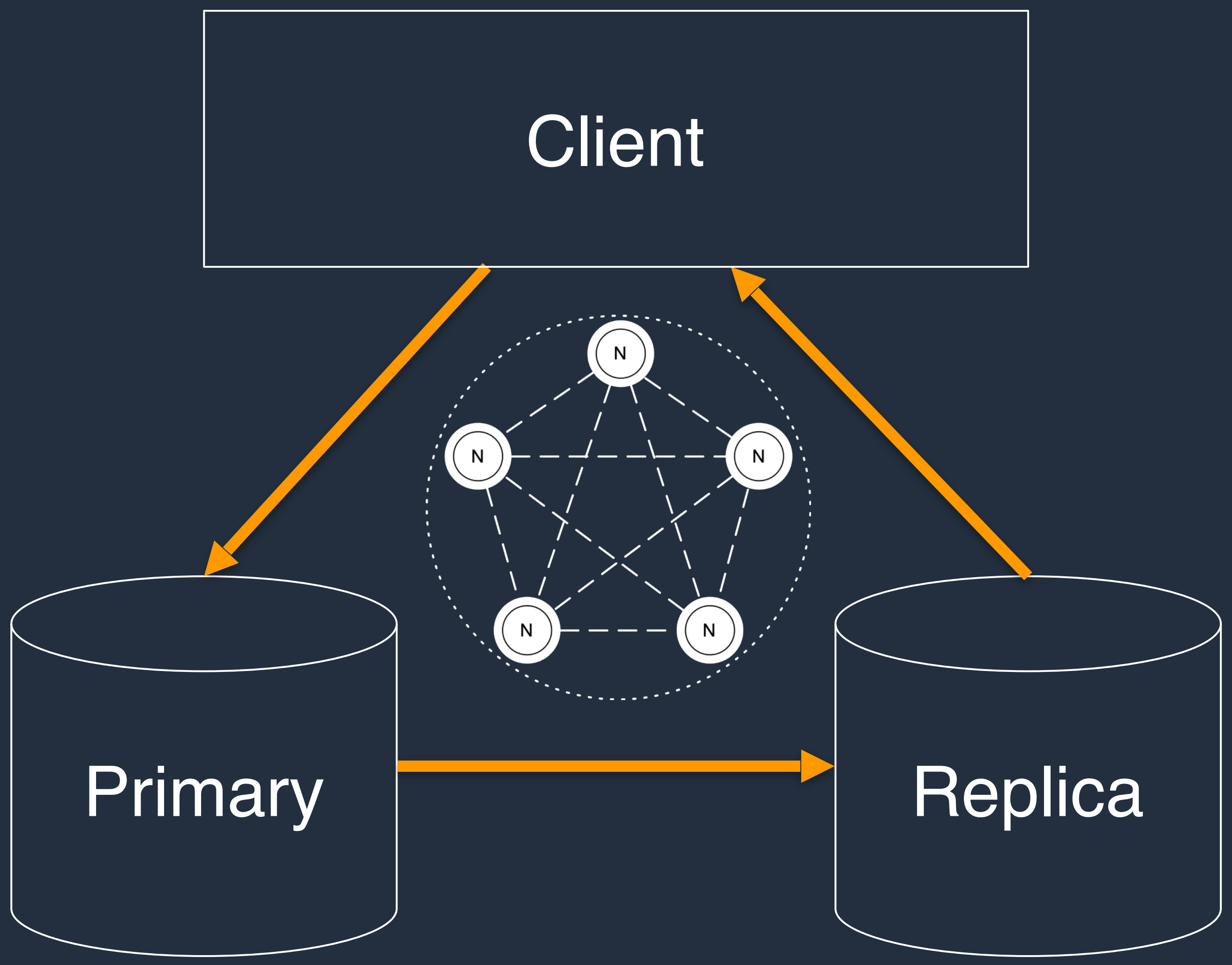
Infrequent

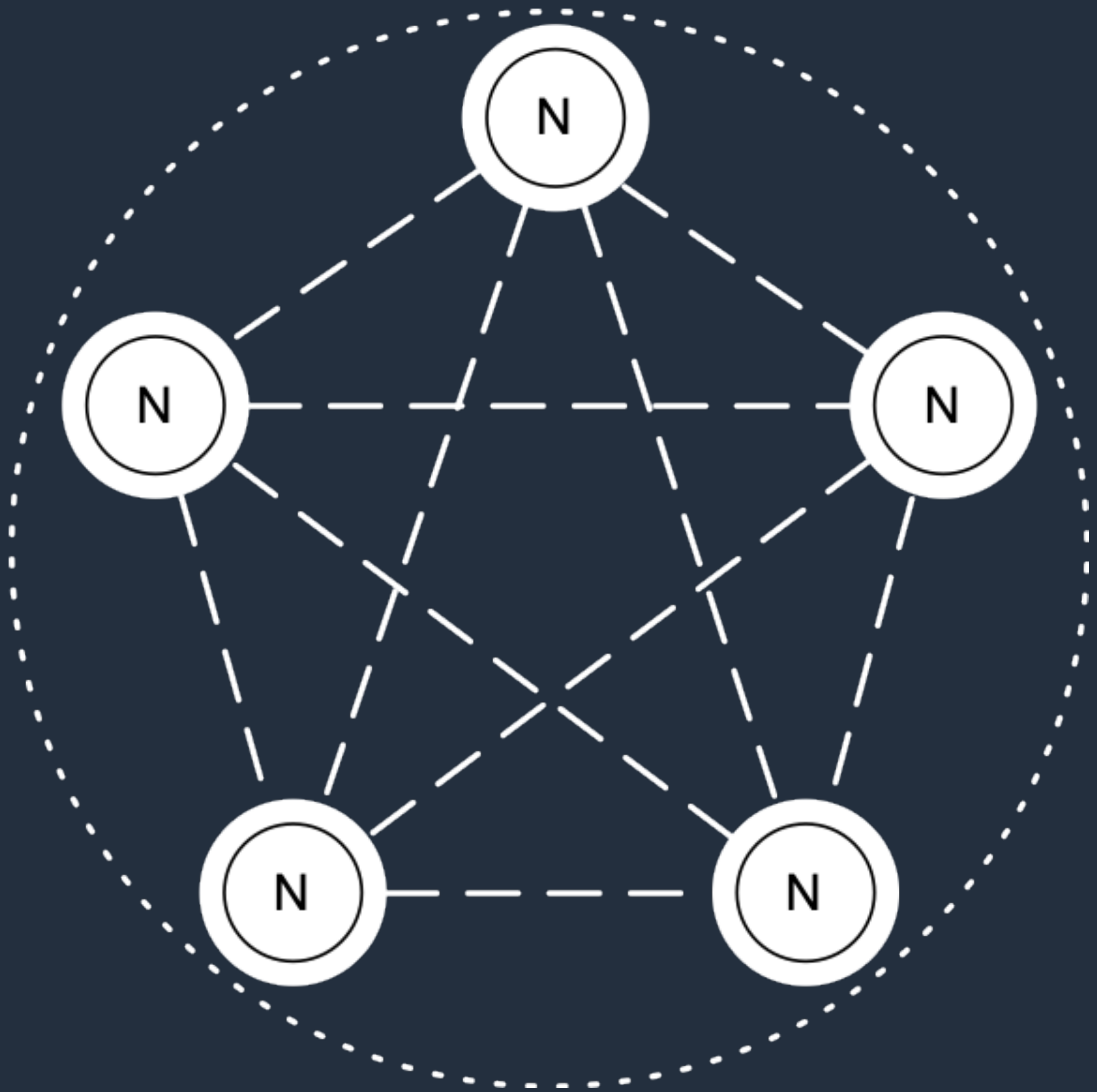
Short

Small = UNCORRELATED

“Blast Radius”

Physalia Architecture

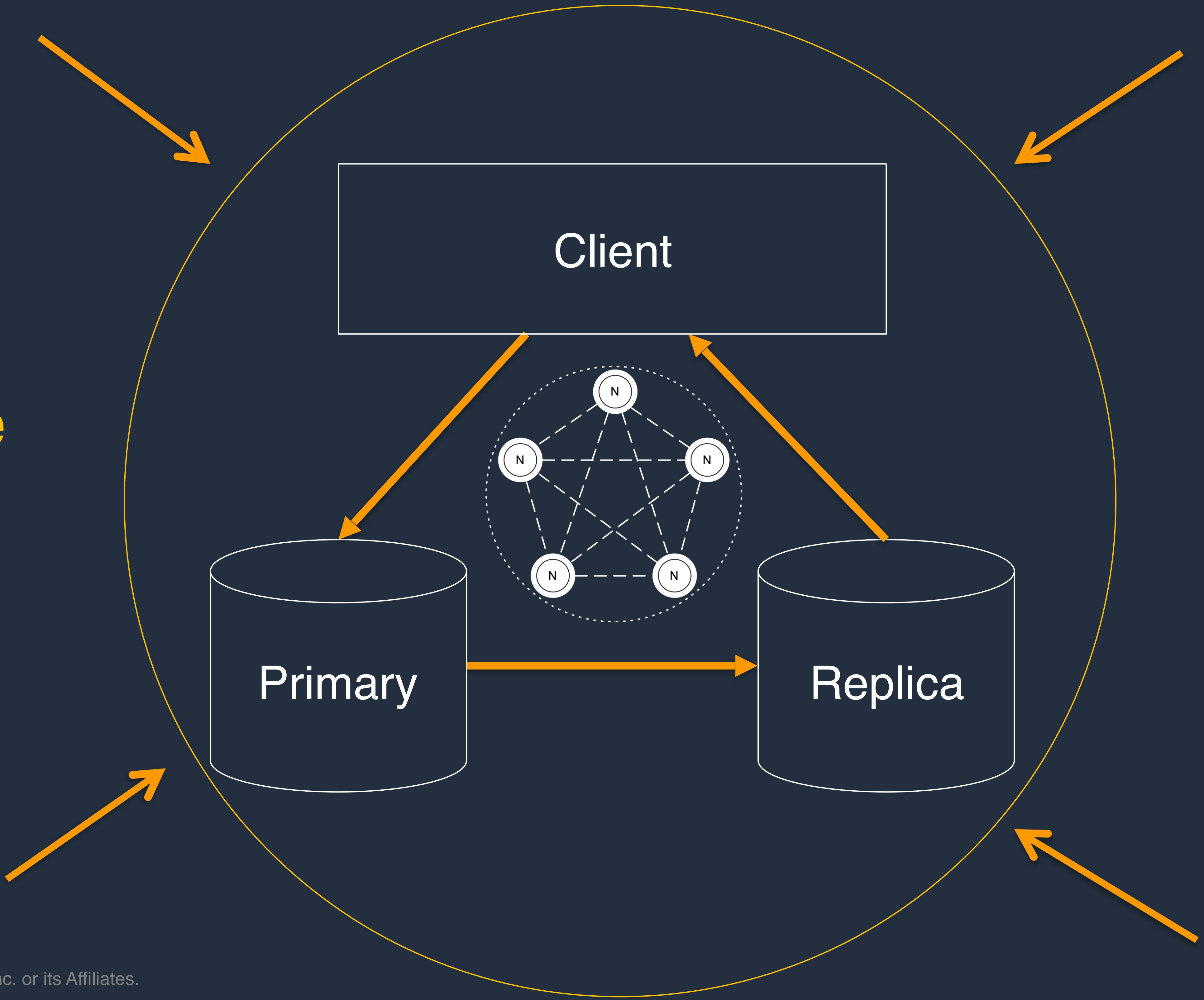

Client
N
N
N
N
N
Primary
Replica



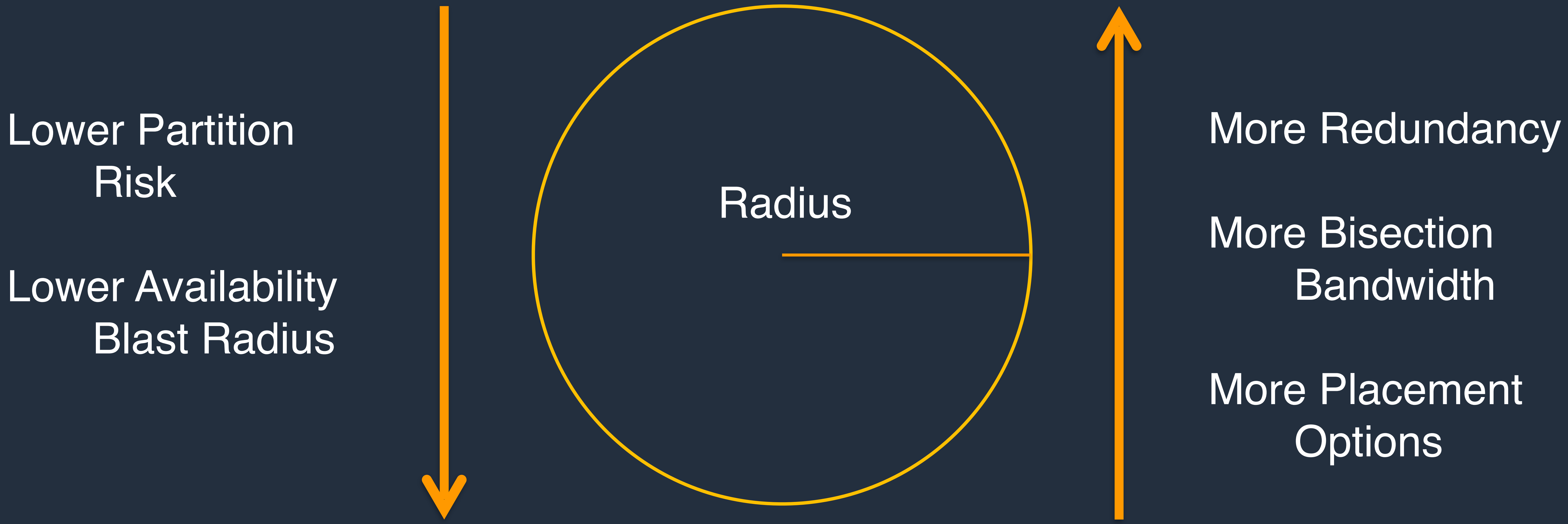
Physalia *Cell*

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

Minimize The Radius



Topology Details Matter



Lower Partition Risk

Lower Availability Blast Radius

Radius

More Redundancy

More Bisection Bandwidth

More Placement Options

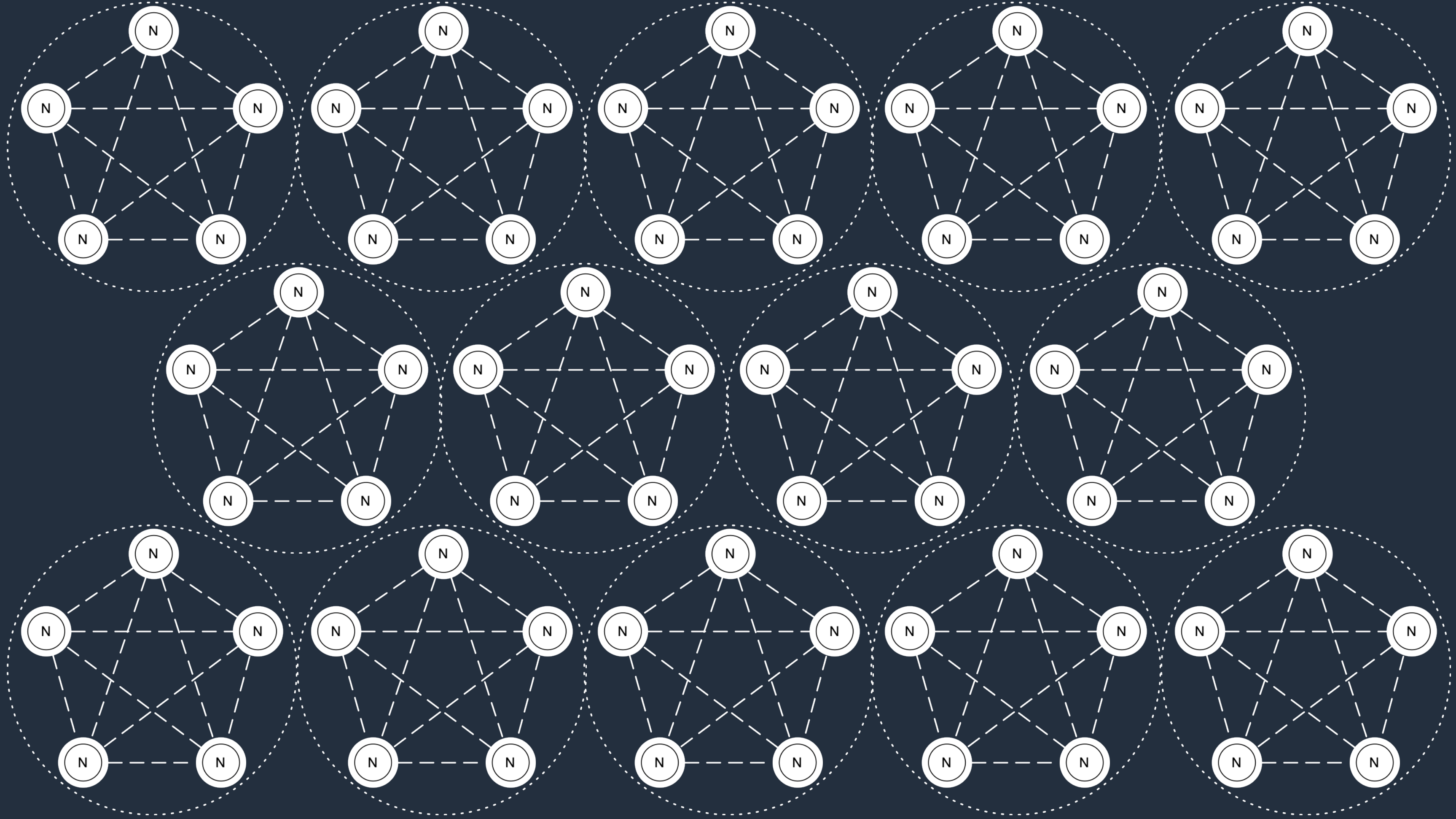
Take Advantage of Eventual Consistency When You Can!

- Discovery Cache (clients discover nodes)
- Monitoring
- ”Meta” control plane

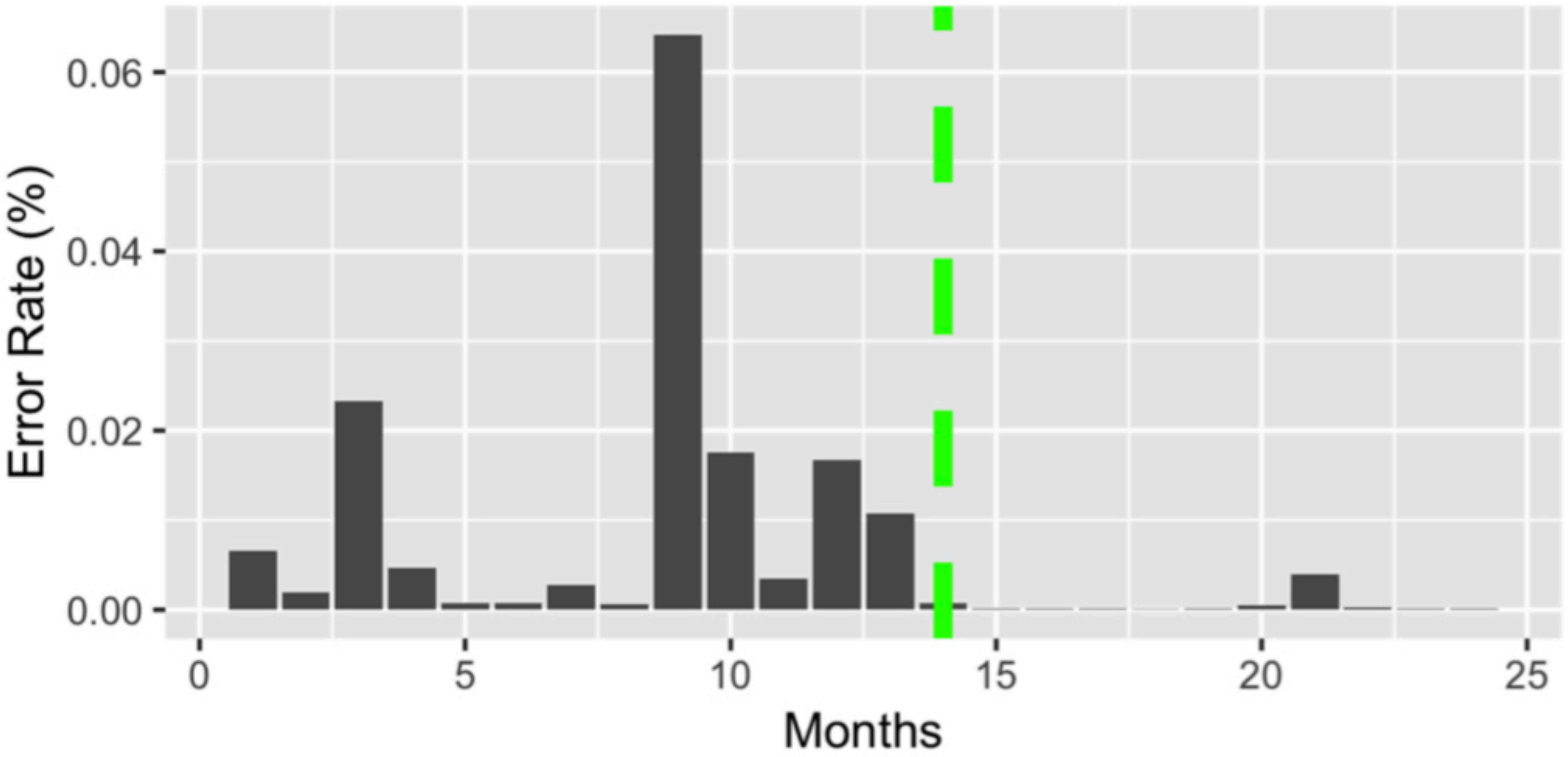
Optimize for Blast Radius

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

Build *humility* into the system.



N


Error Rate (%)
0.06
0.04
0.02
0.00
0
5
10
15
20
25
Months

Q&A

Marc Brooker

mbrooker@amazon.com

@marcjbrooker

The End