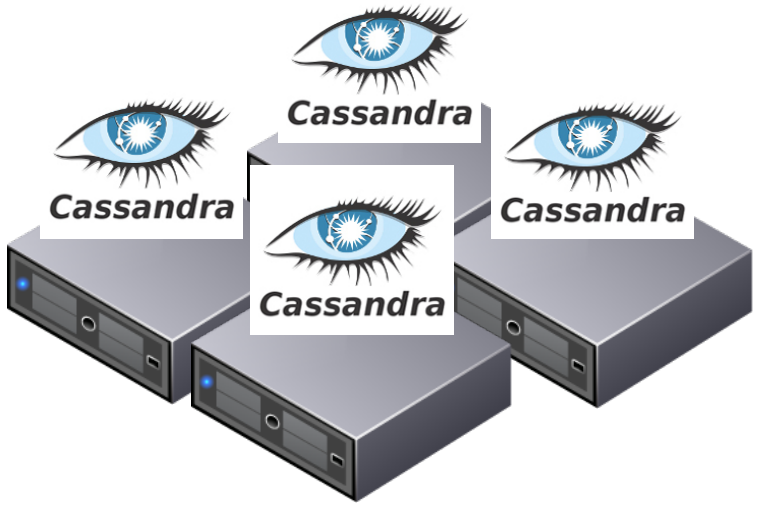
Corfu: A Cloud Scale Consistency Platform

Michael Wei
VMware Research Group
Code Mesh 2016

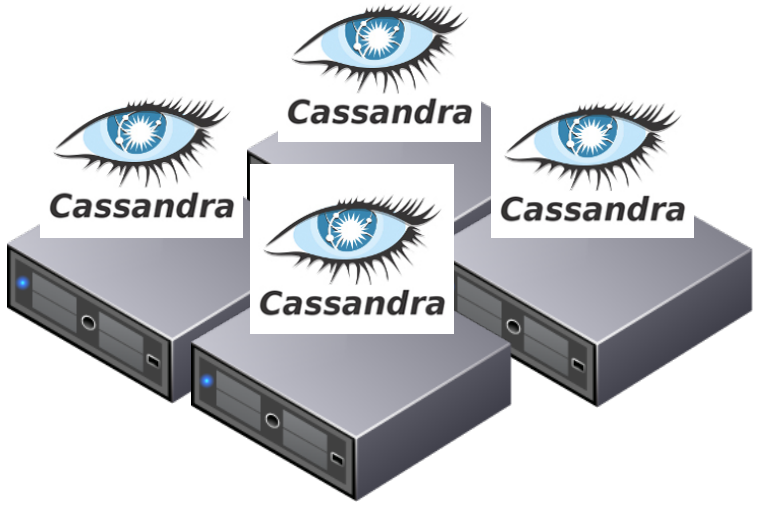
vmware

How a Typical Application Becomes Distributed

How a Typical Application Becomes Distributed



How a Typical Application Becomes Distributed



How a Typical Application Becomes Distributed

stackoverflow

Questions

Search

build a distributed system from scratch

search

31 results

relevance | newest | votes | active

1
vote

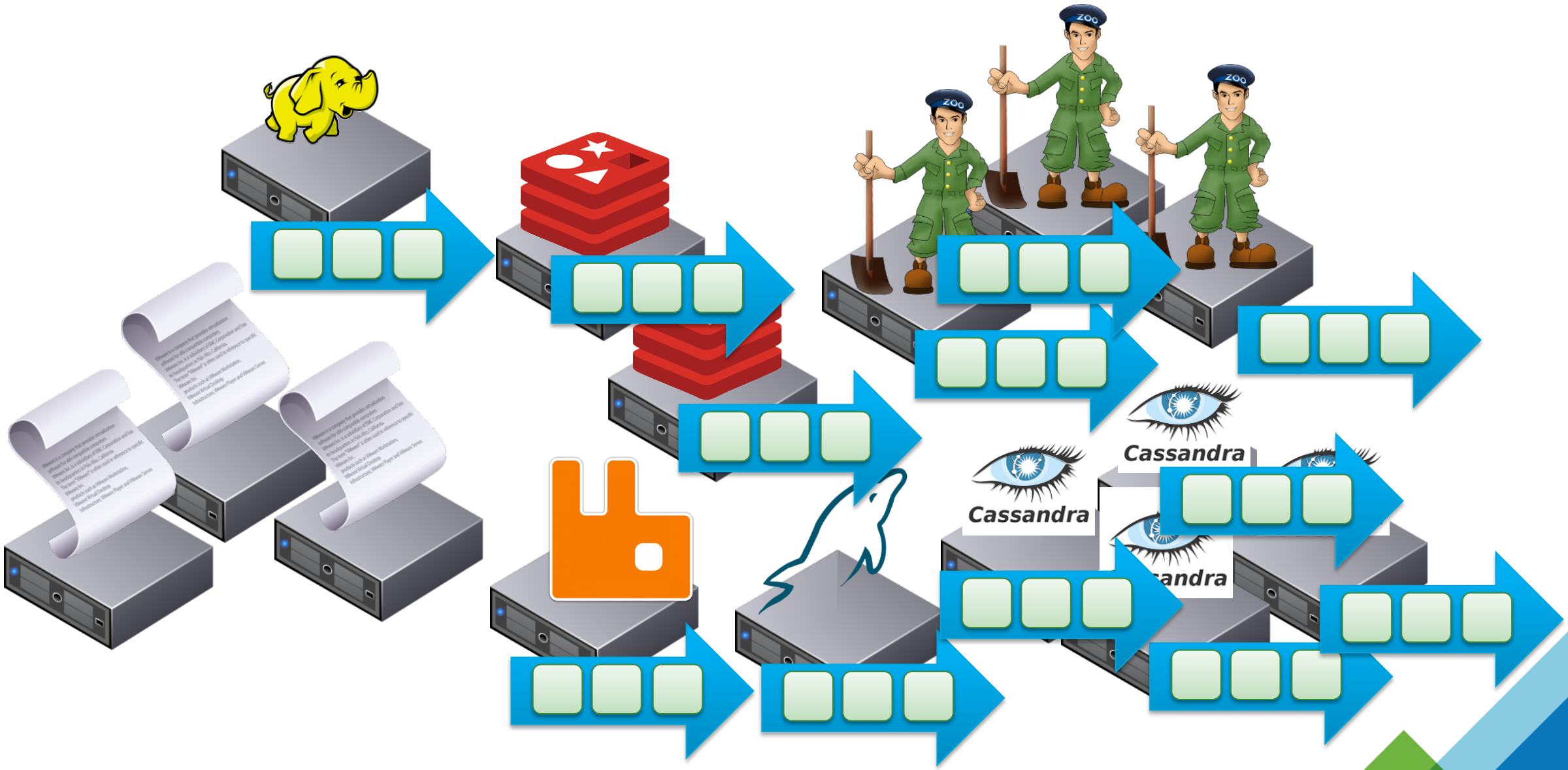
0
answers

Q: What would a clean slate design of a distributed system look like?

Why would you ever want to do such a thing?

distributed-computing distributed-system

How a Typical Application Becomes Distributed



What is Corfu?

Corfu

What is Corfu?

Corfu

Paxos

Paxos – A Family of Consensus Protocols

A = 0

A = 0;

Paxos – A Family of Consensus Protocols

A = 1

A = 0;
A = 1;

Paxos – A Family of Consensus Protocols

A = 1

A = 1

A = 1

A = 0;
A = 1;

Paxos – A Family of Consensus Protocols

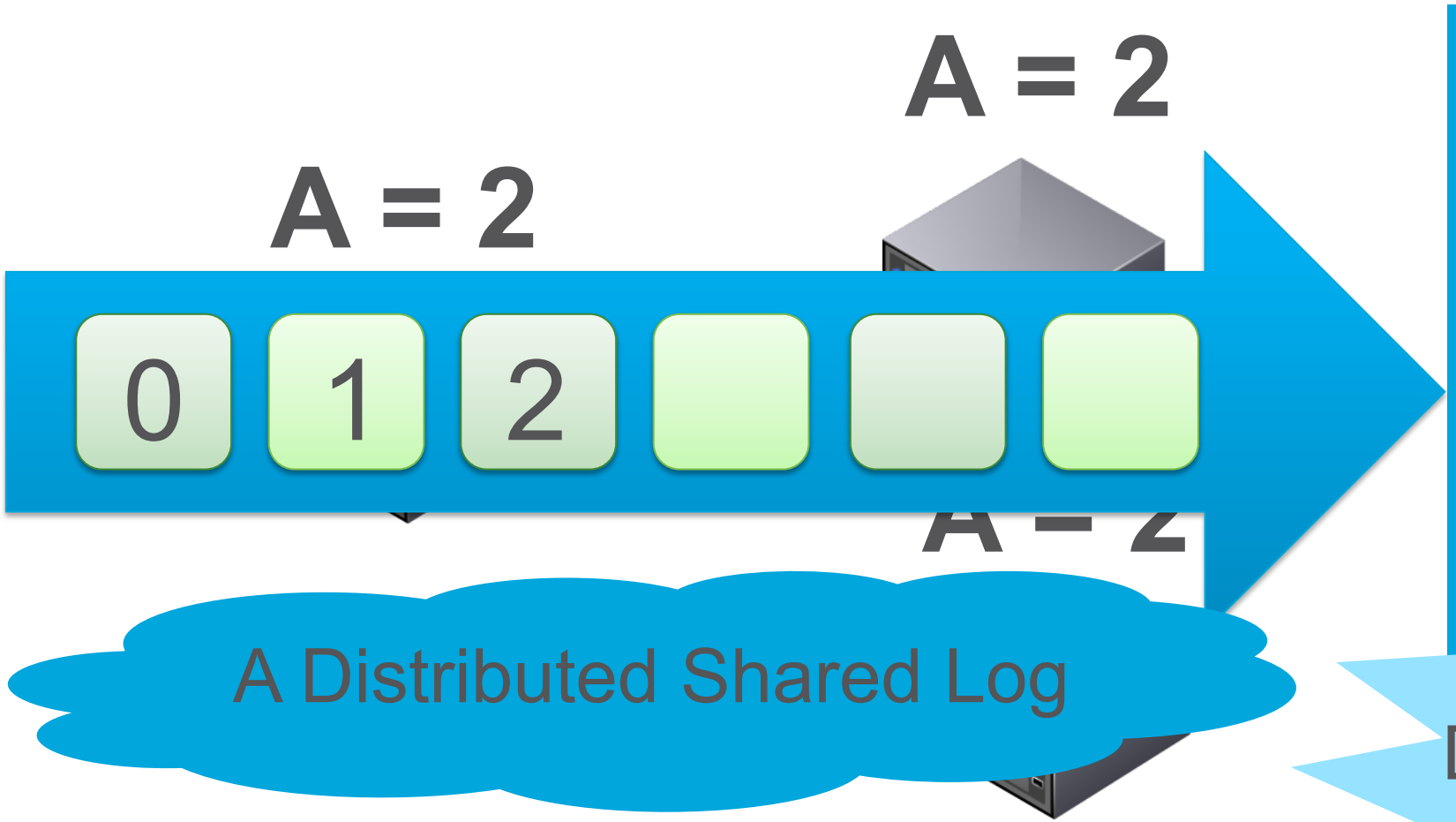
A = 2

A = 1

A = 1

A = 0;
A = 1;
A = 2;

Paxos – A Family of Consensus Protocols



A = 0;
A = 1;
A = 2;

10K ops/s
Doesn't Scale!

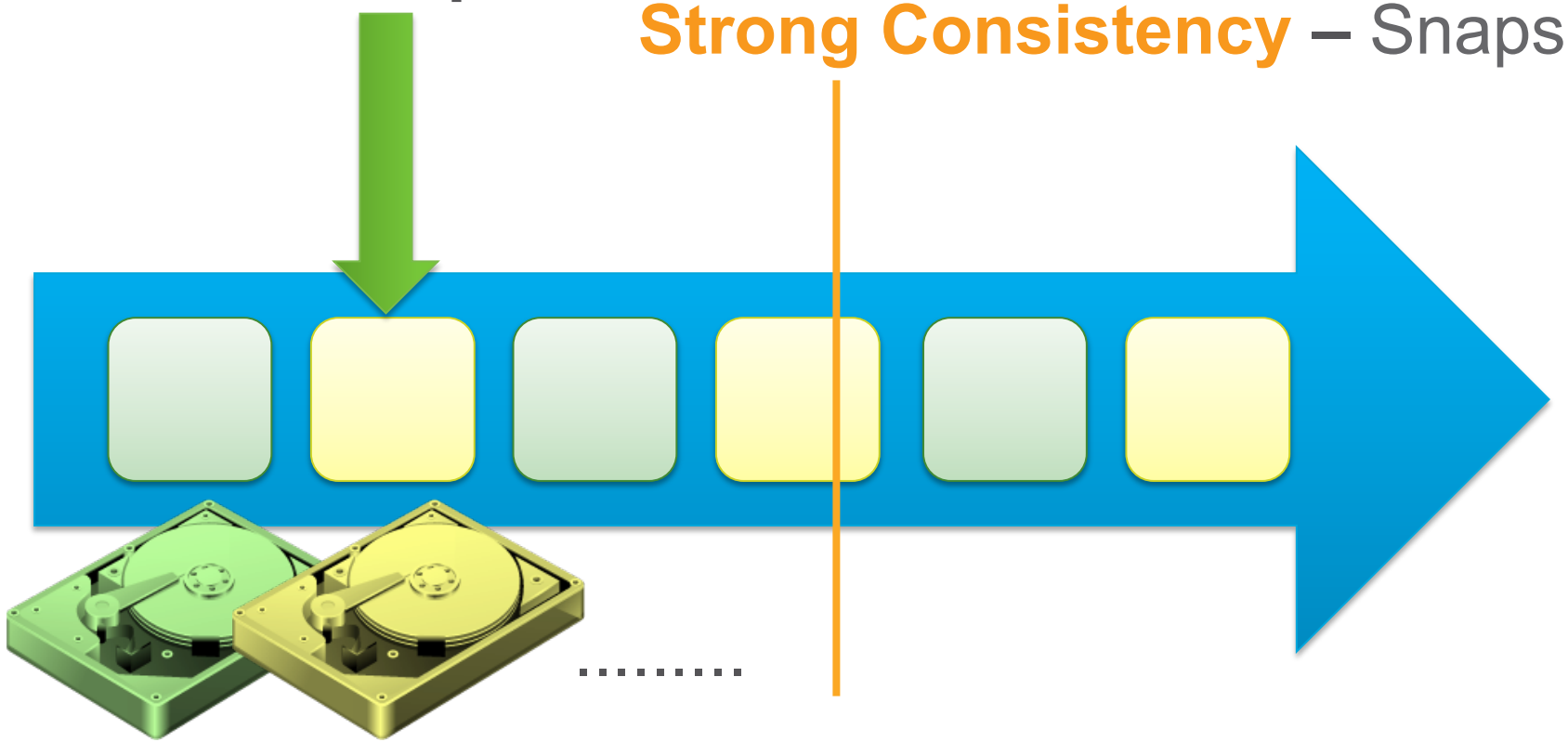
Corfu: A Distributed Shared Log

Sequencer

Layout

"Control Plane"
Traditional I/O
bottleneck

Fast Datapath

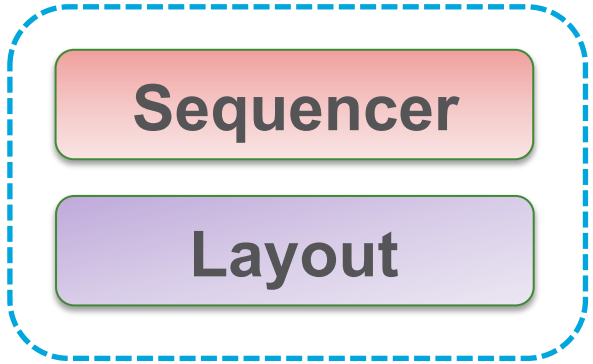


Strong Consistency – Snapshots

Write Scaling – Fast Updates

Read Scaling – Fast Random Reads

Corfu: A Distributed Shared Log



Sequencer is fast
600K op/s – not part of I/O path

Corfu: A Distributed Shared Log

Sequencer

Layout

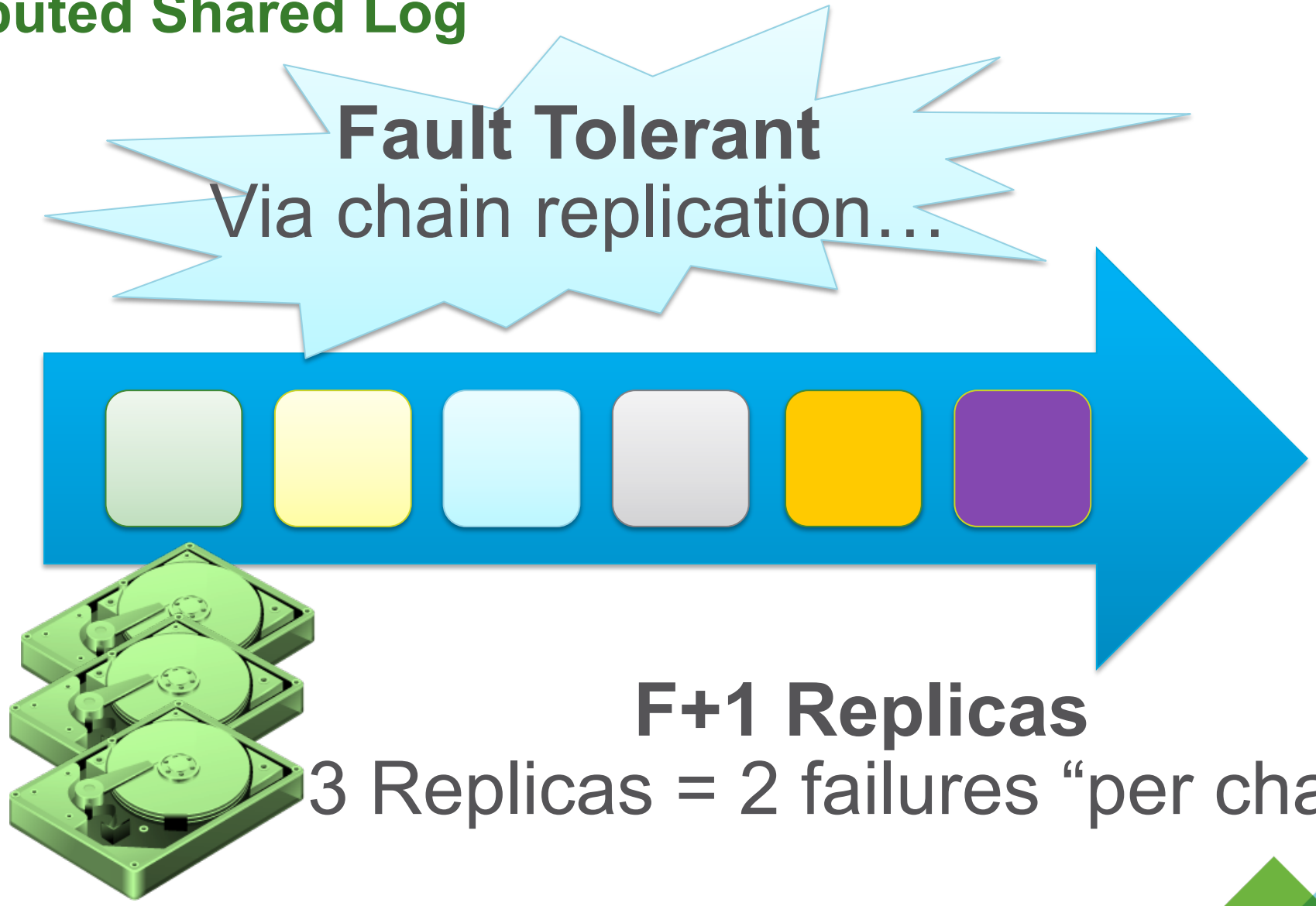
Scalable
Add more stripes…

Corfu: A Distributed Shared Log

Sequencer

Layout

Fault Tolerant
Via chain replication…



F+1 Replicas
3 Replicas = 2 failures “per chain”

Corfu: A Distributed Shared Log

Sequencer

Layout

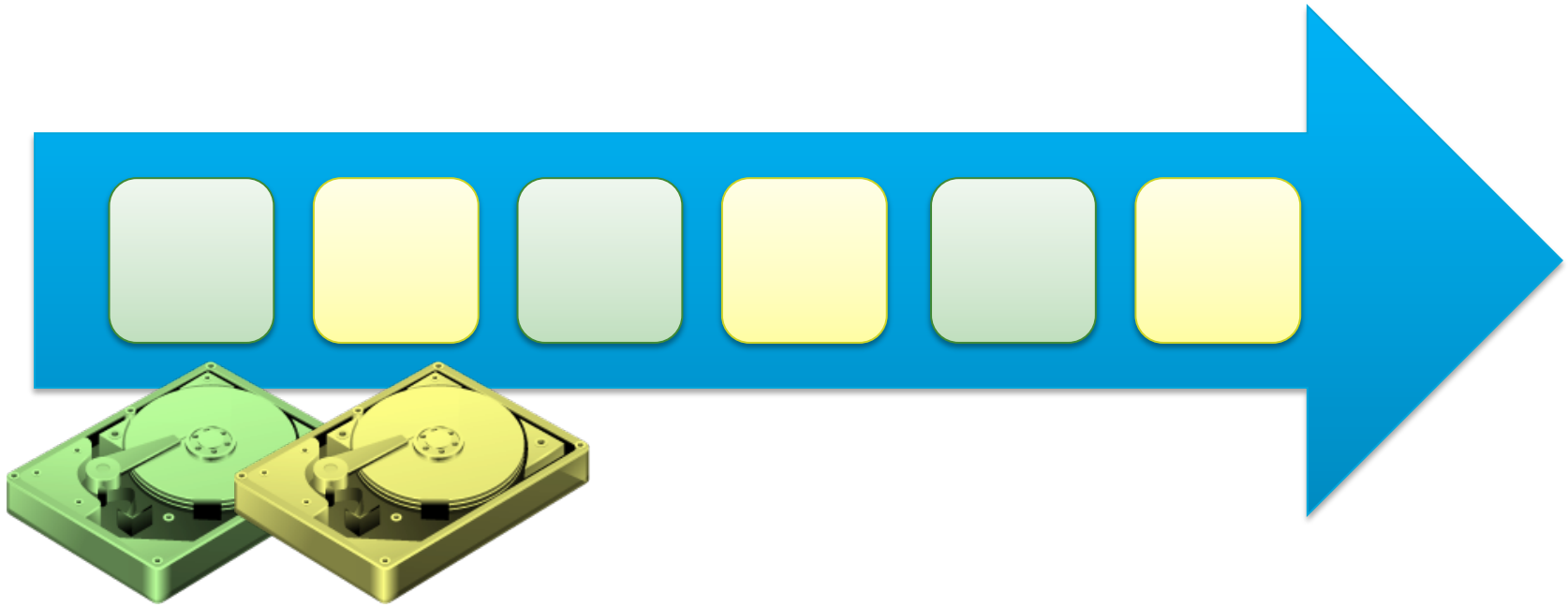
Strong Consistency

The log … is the truth!

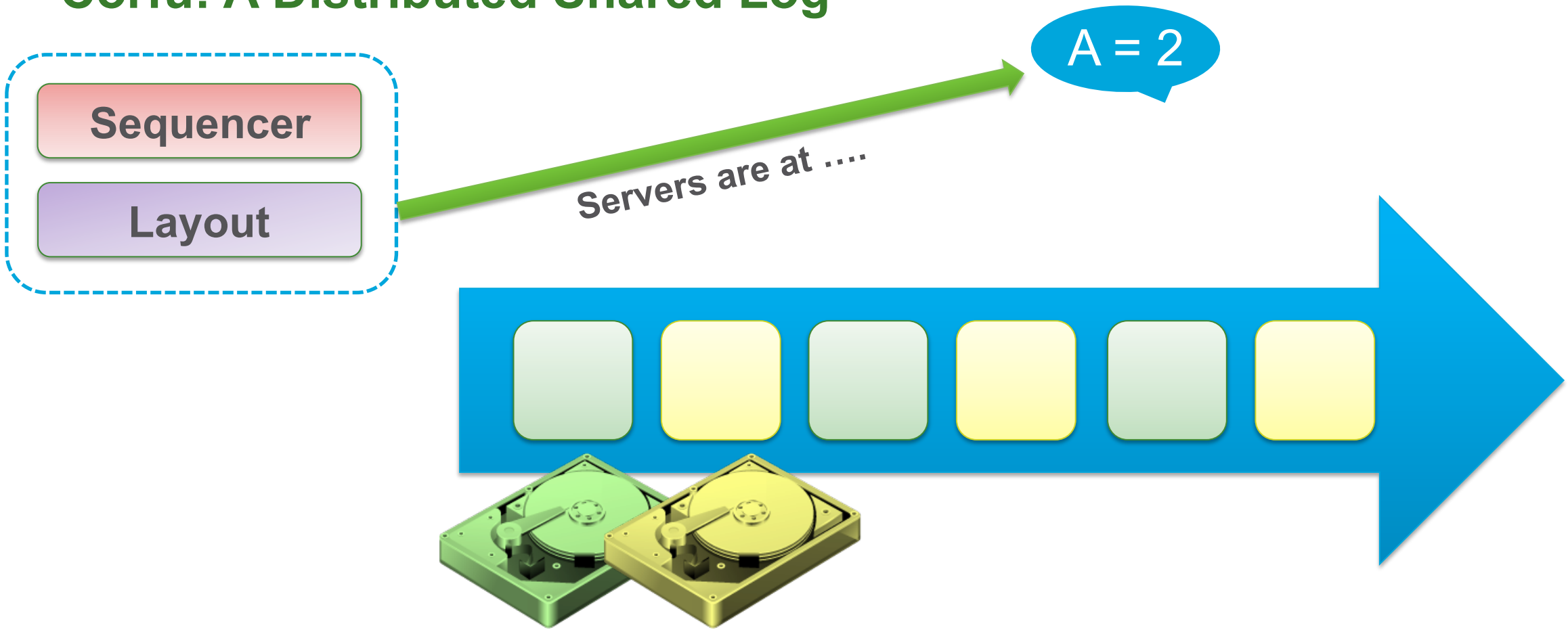
Corfu: A Distributed Shared Log

Sequencer

Layout



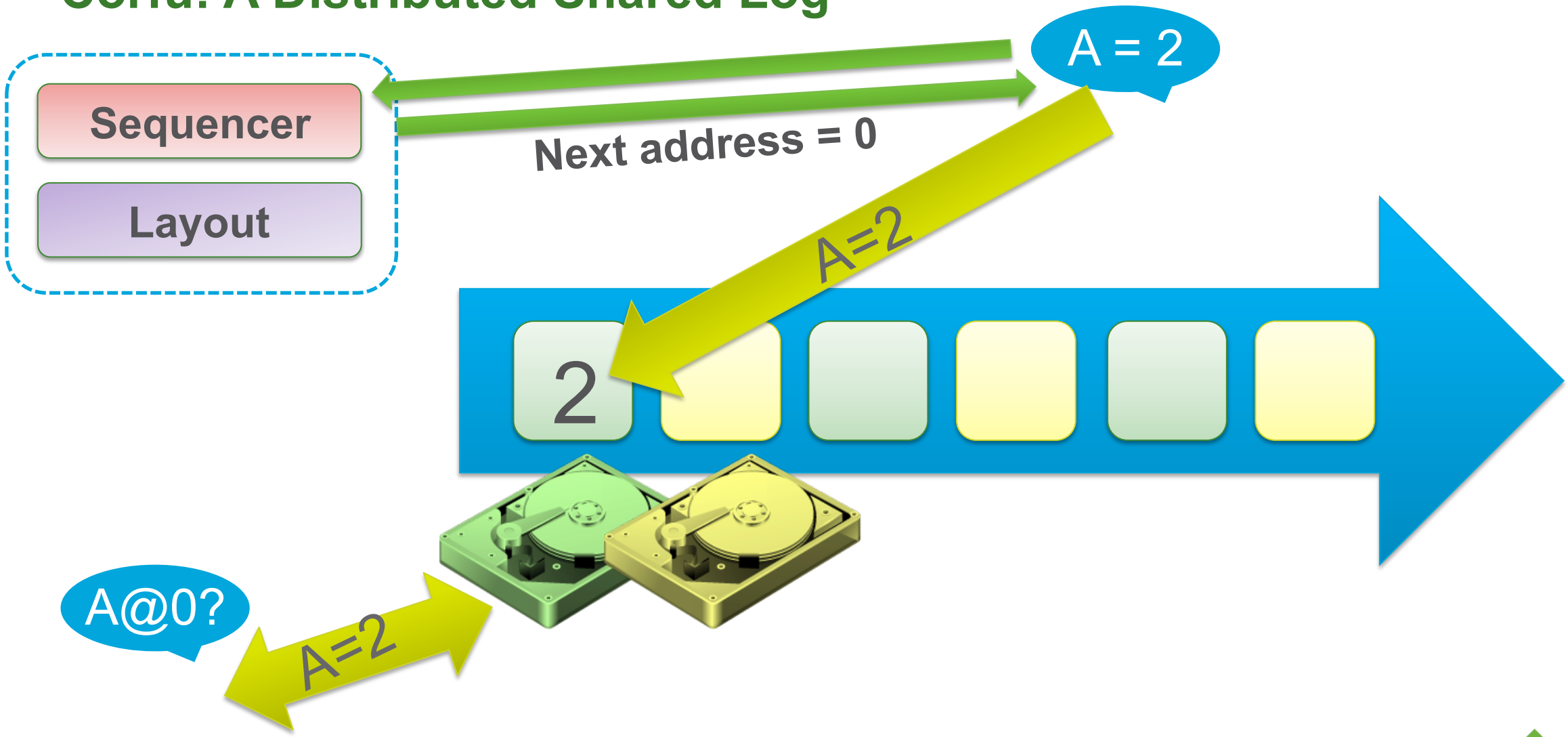
Corfu: A Distributed Shared Log



Corfu: A Distributed Shared Log

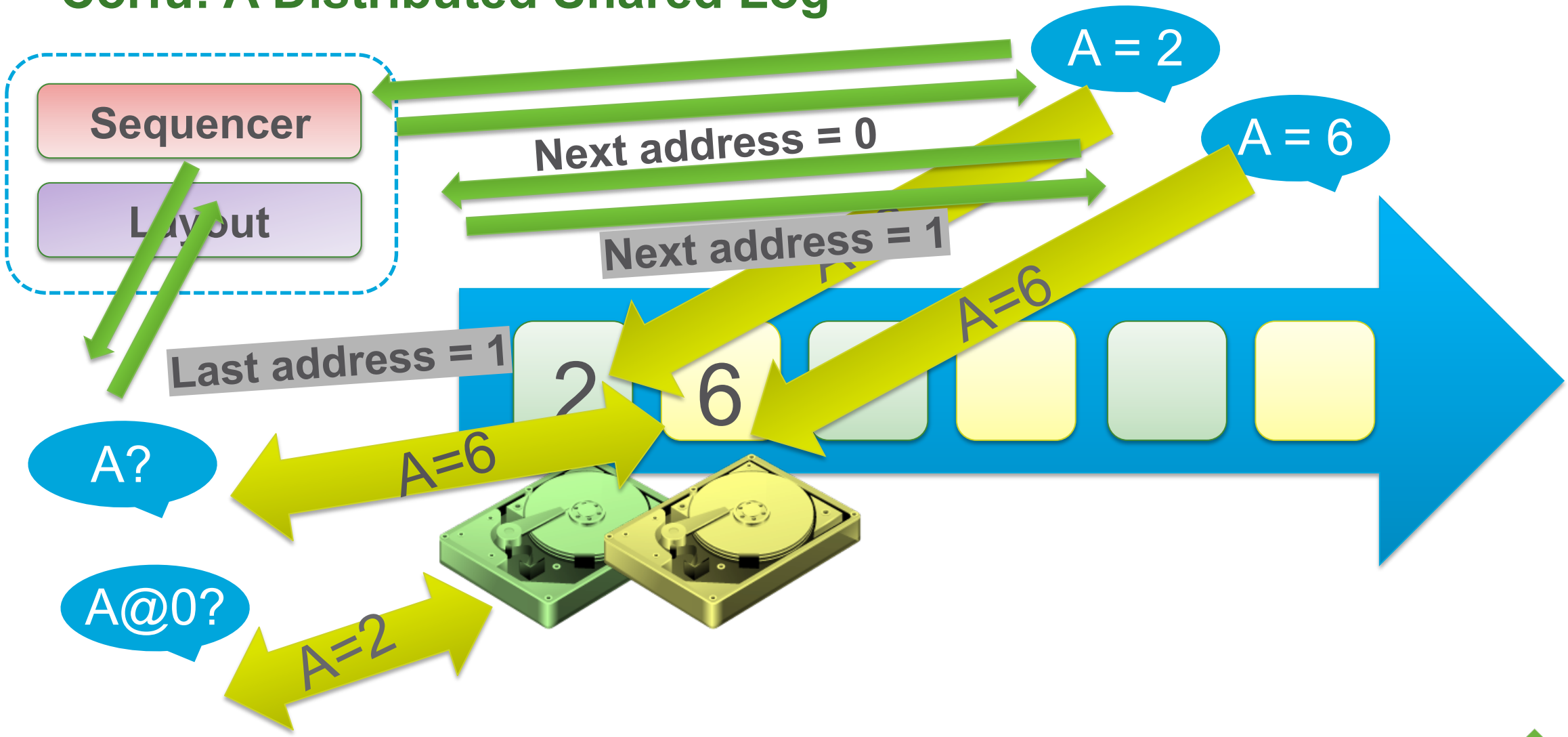
Corfu: A Distributed Shared Log

Corfu: A Distributed Shared Log



Corfu: A Distributed Shared Log

Corfu: A Distributed Shared Log

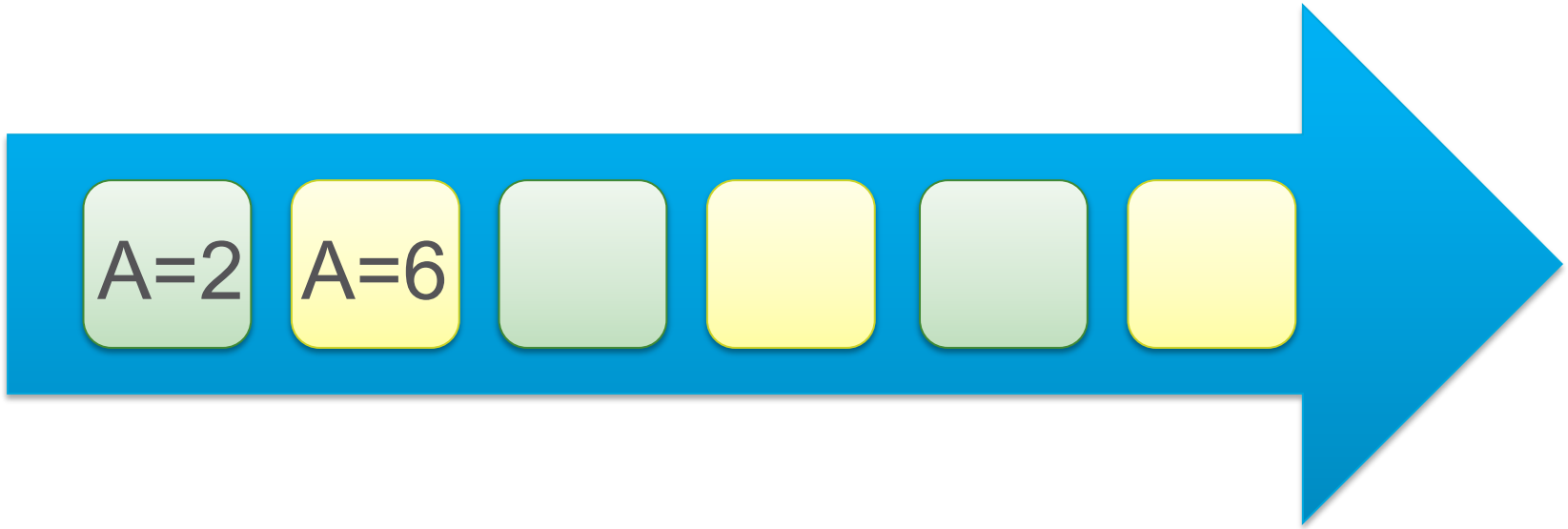


Corfu: A Distributed Shared Log

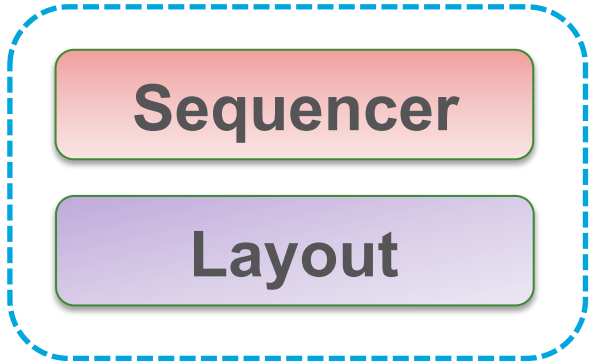
A
6

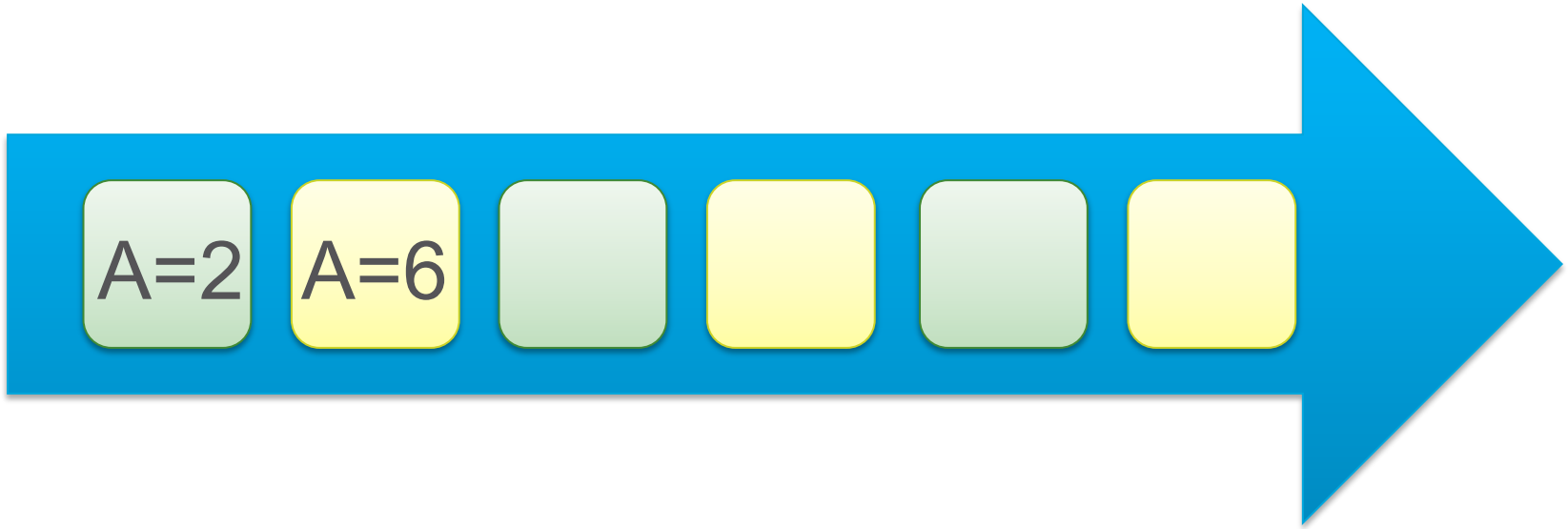
Sequencer

Layout



Corfu: A Distributed Shared Log





Corfu: A Distributed Shared Log

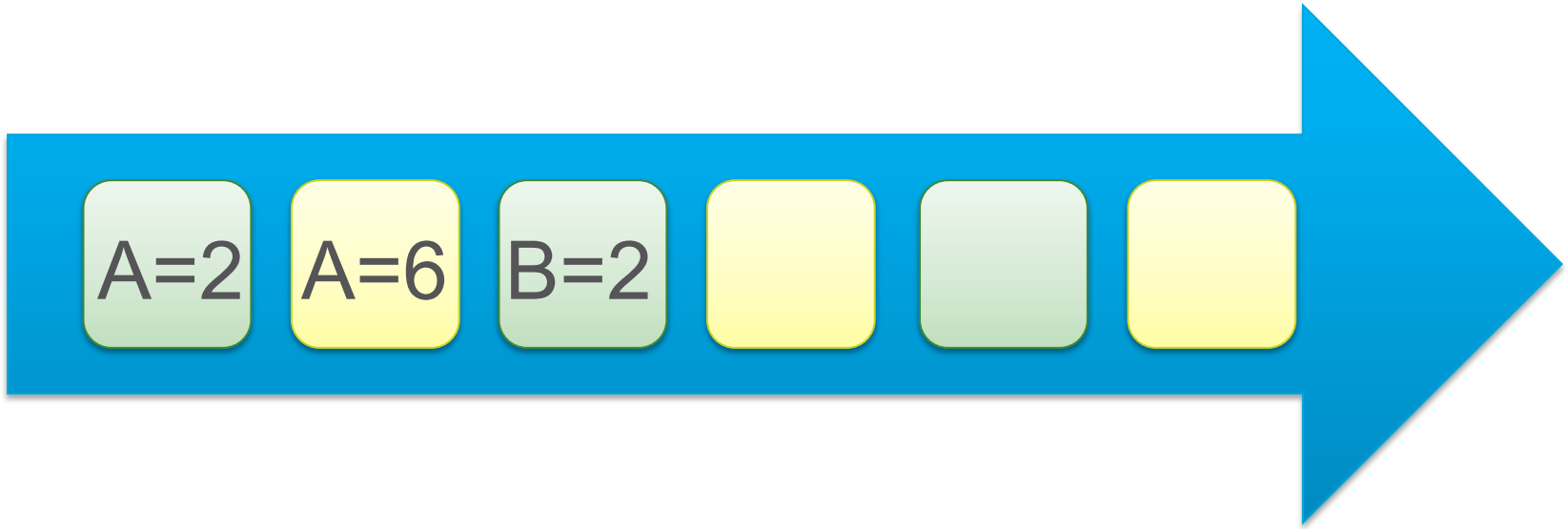
A@1 B@2

Sequencer

Layout

A
6

B
2



Corfu: A Distributed Shared Log

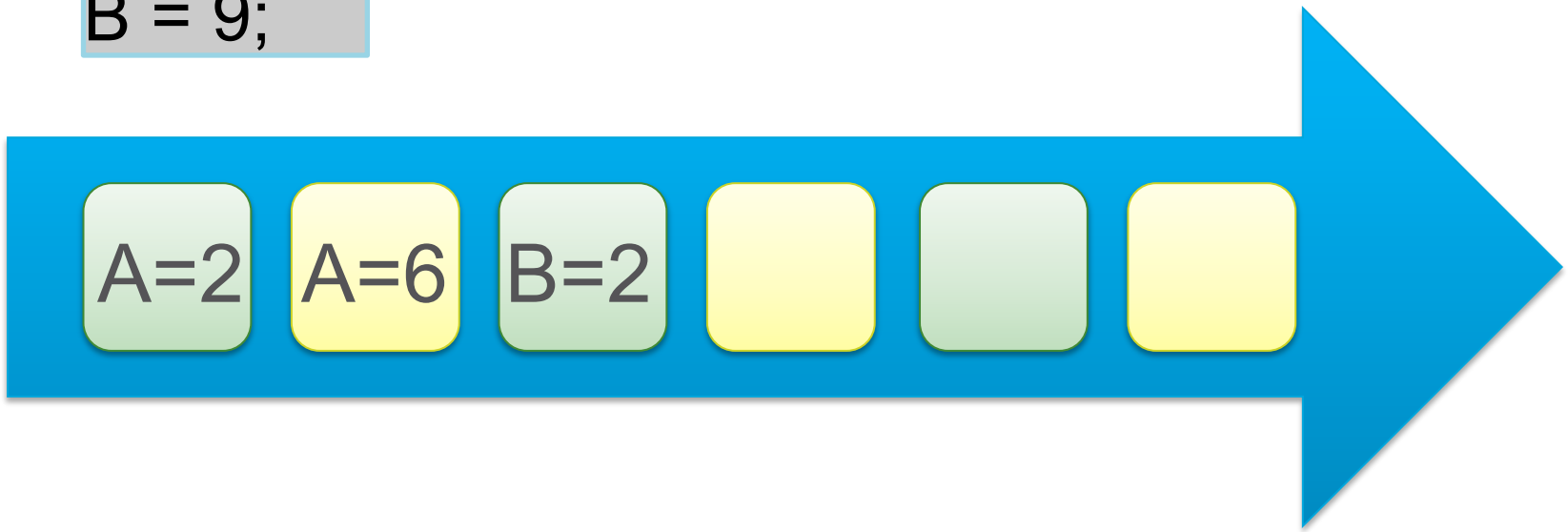
A@1 B@2

Sequencer

Layout

Transaction

A = 1;
B = 9;



A
6

B
2

Corfu: A Distributed Shared Log

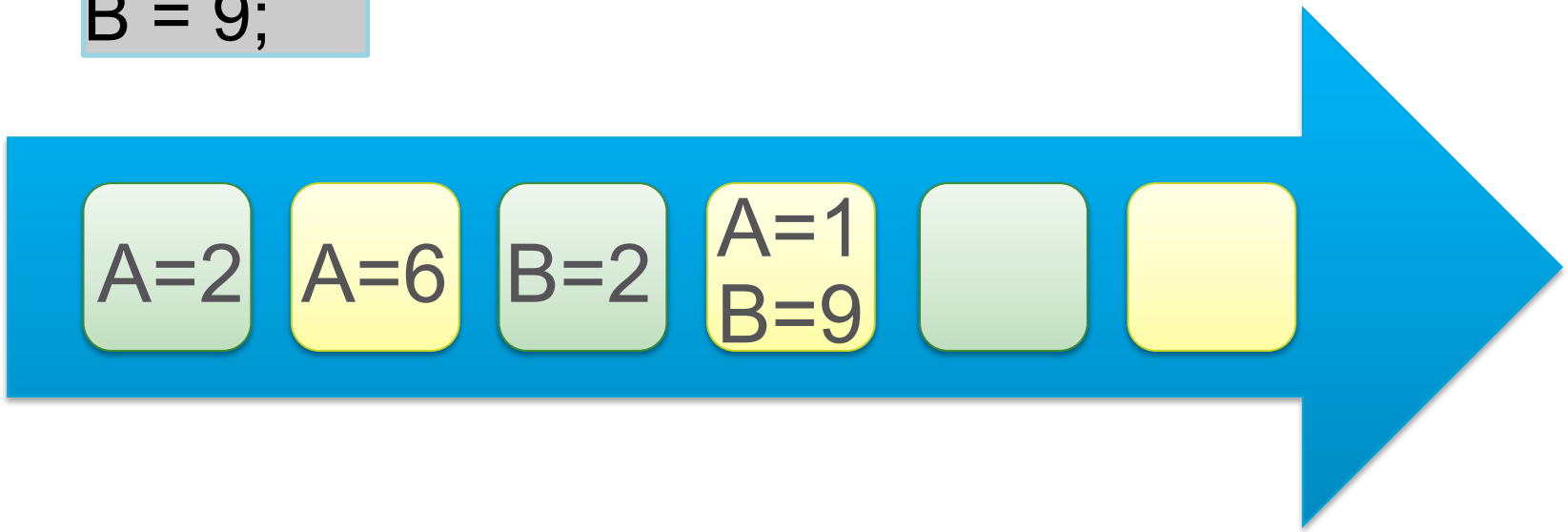
A@3 B@3

Sequencer

Layout

Transaction

A = 1;
B = 9;



A
1

B
9

Corfu: A Distributed Shared Log

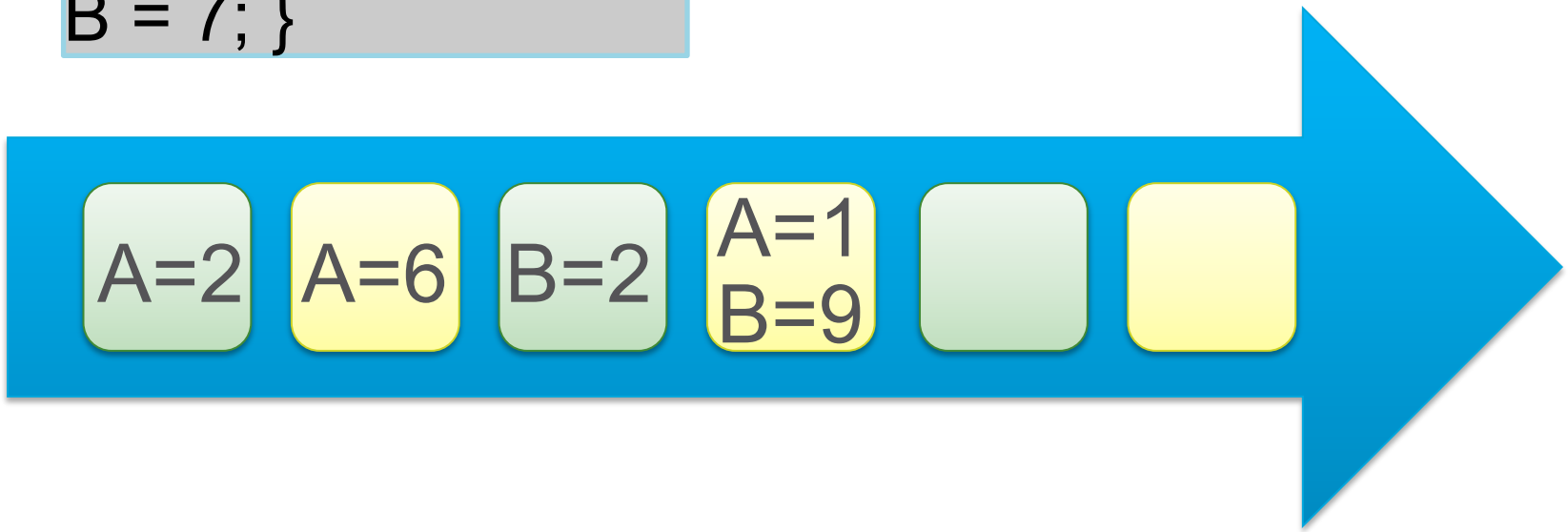
A@3 B@3

Sequencer

Layout

Transaction

```
if (A == 1) {
B = 7; }
```



A
1

B
9

Corfu: A Distributed Shared Log

A@3 B@3

Sequencer

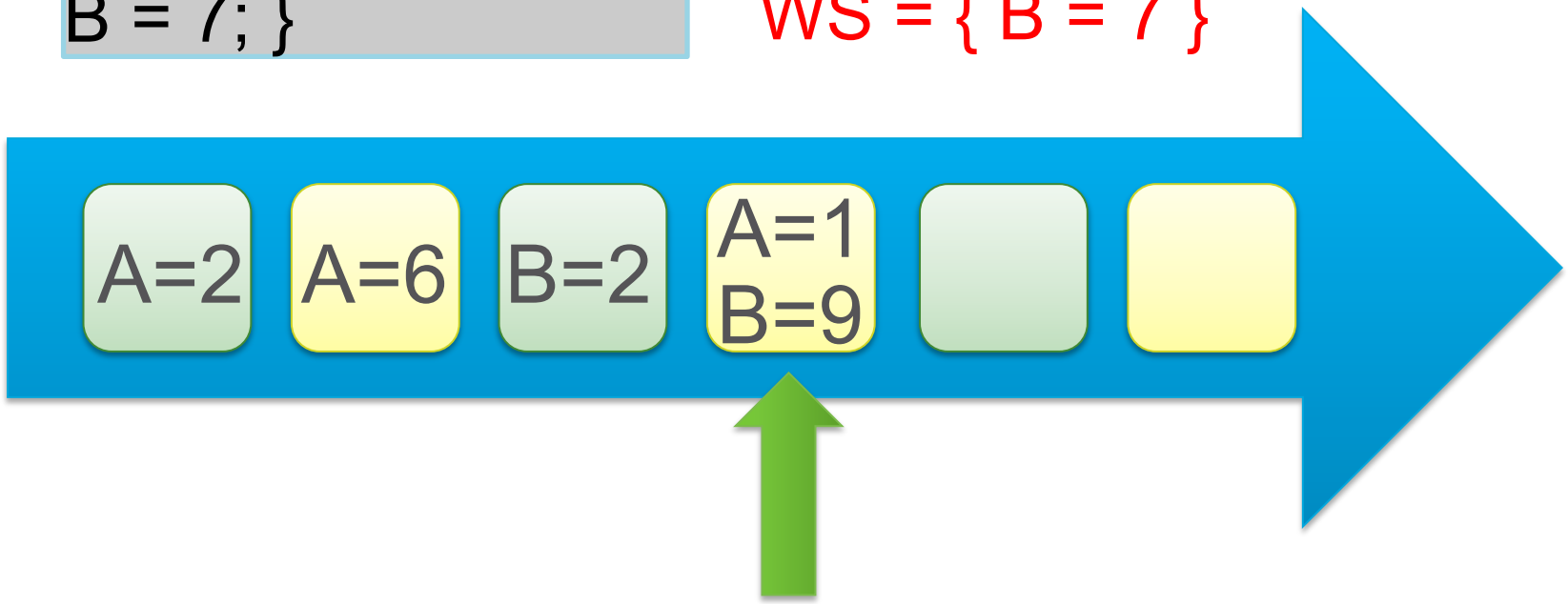
Layout

Transaction

```
if (A == 1) {
B = 7; }
```

RS = A @ 3
WS = { B = 7 }

A	B
1	9



Address = 3

Corfu: A Distributed Shared Log

A@3 B@4

Sequencer

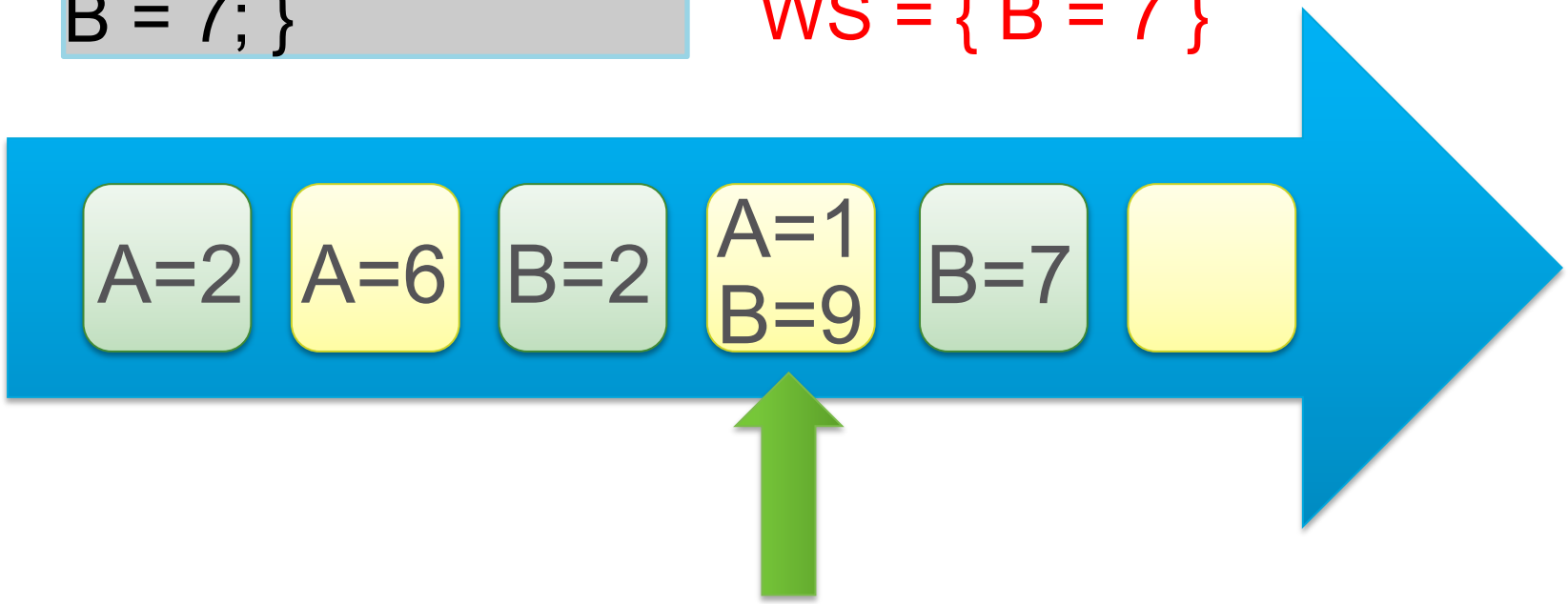
Layout

Transaction

```
if (A == 1) {
B = 7; }
```

RS = A @ 3
WS = { B = 7 }

A	B
1	7



Address = 3

Corfu: A Distributed Shared Log

A@4 B@3

Sequencer

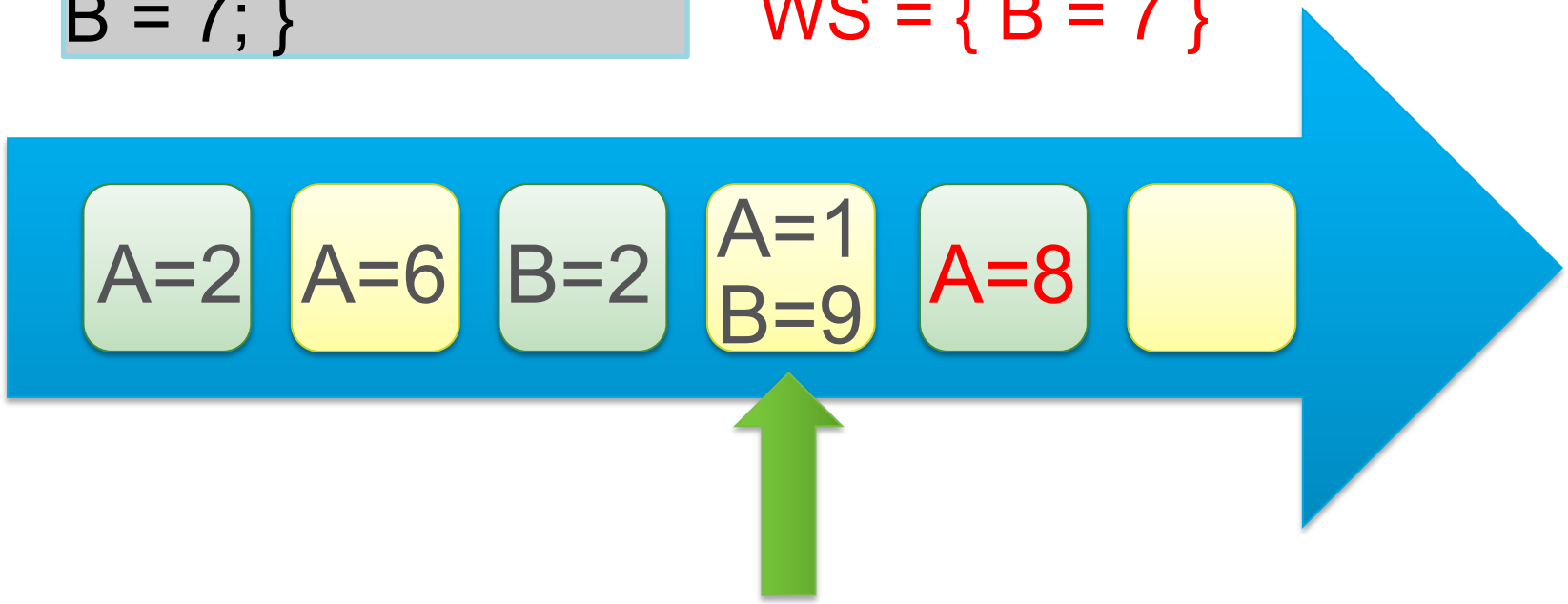
Layout

Transaction

```
if (A == 1) {
B = 7; }
```

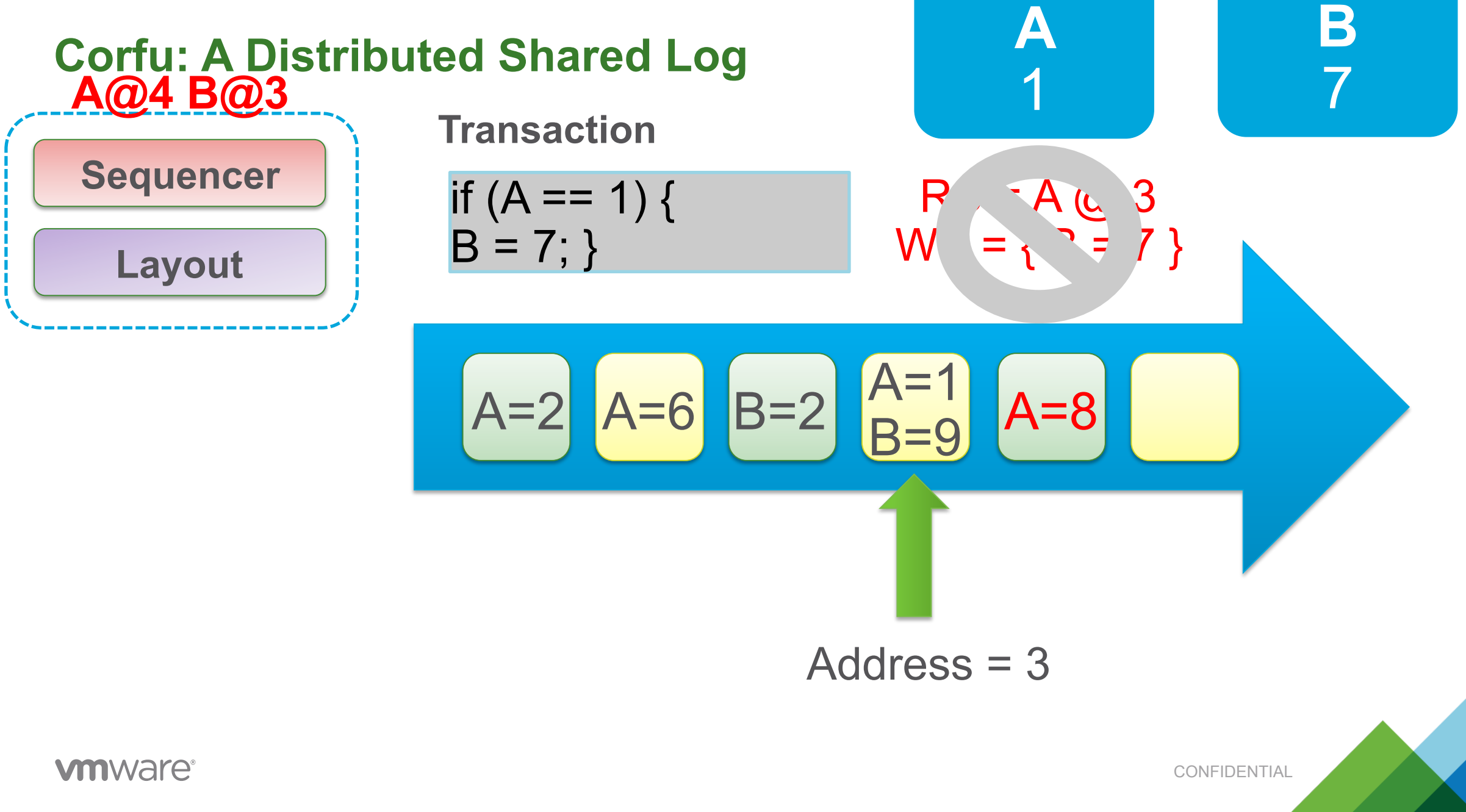
RS = A @ 3
WS = { B = 7 }

A	B
1	7



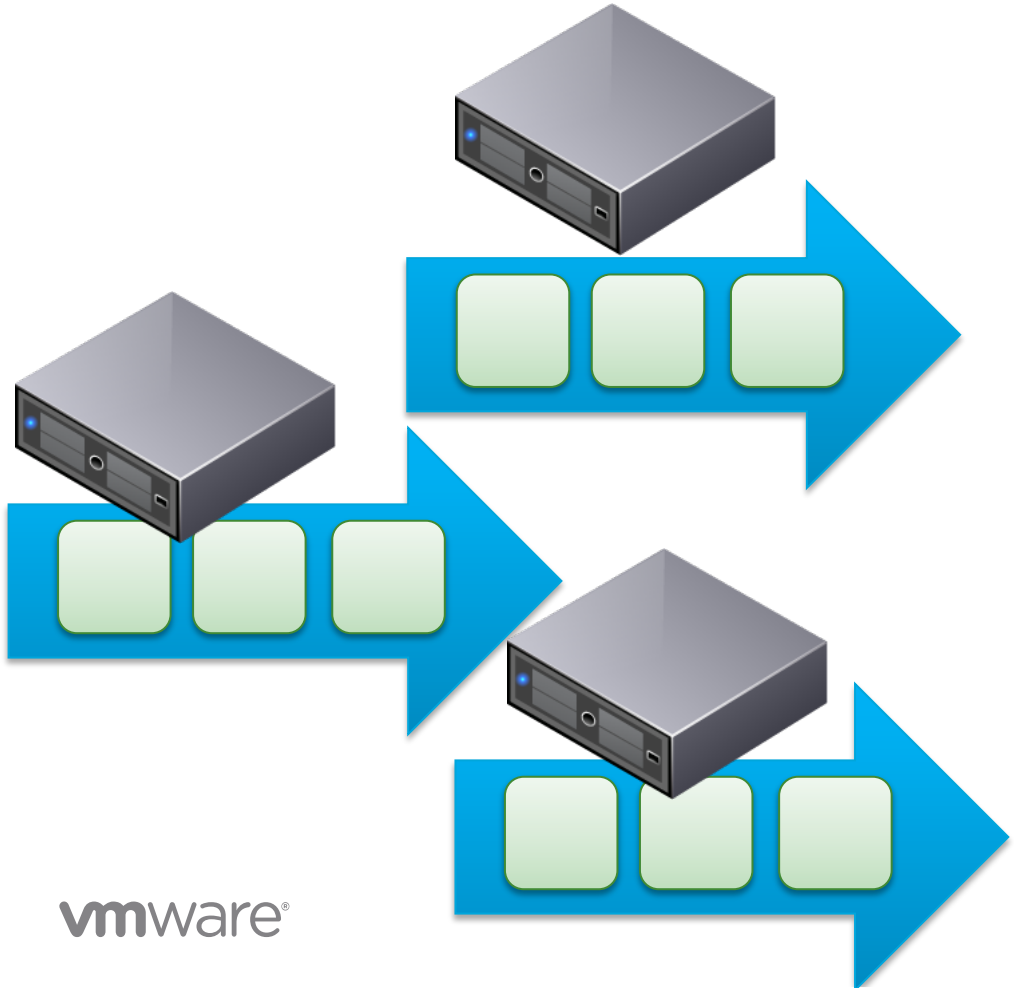
Address = 3

Corfu: A Distributed Shared Log

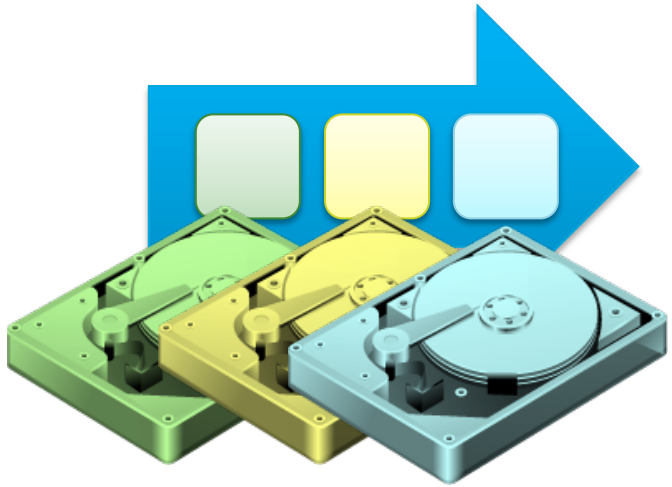


Corfu vs other “Distributed Logs”

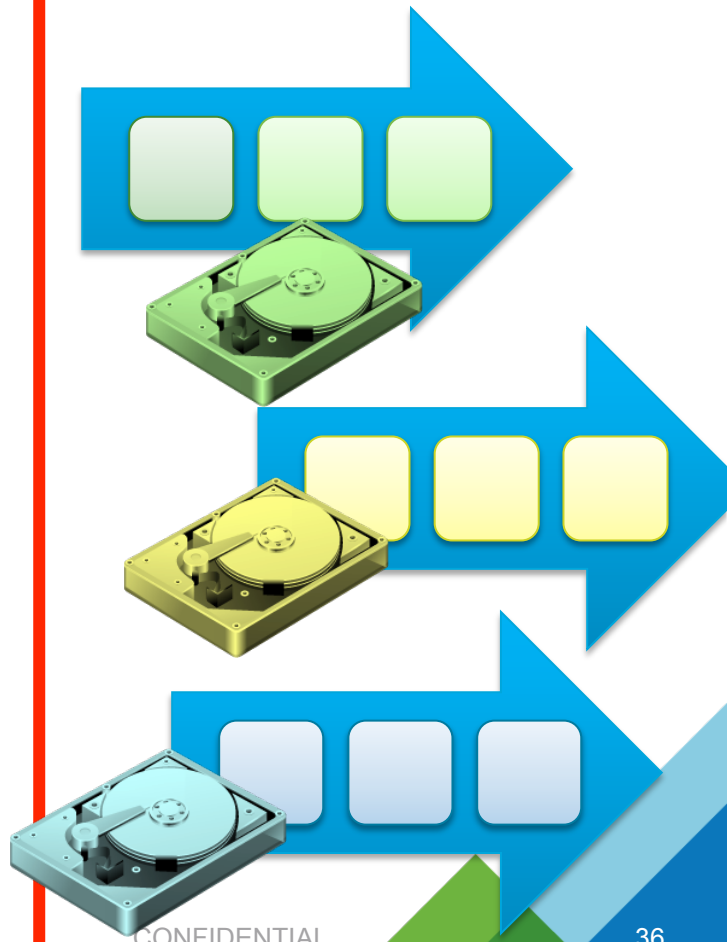
Paxos



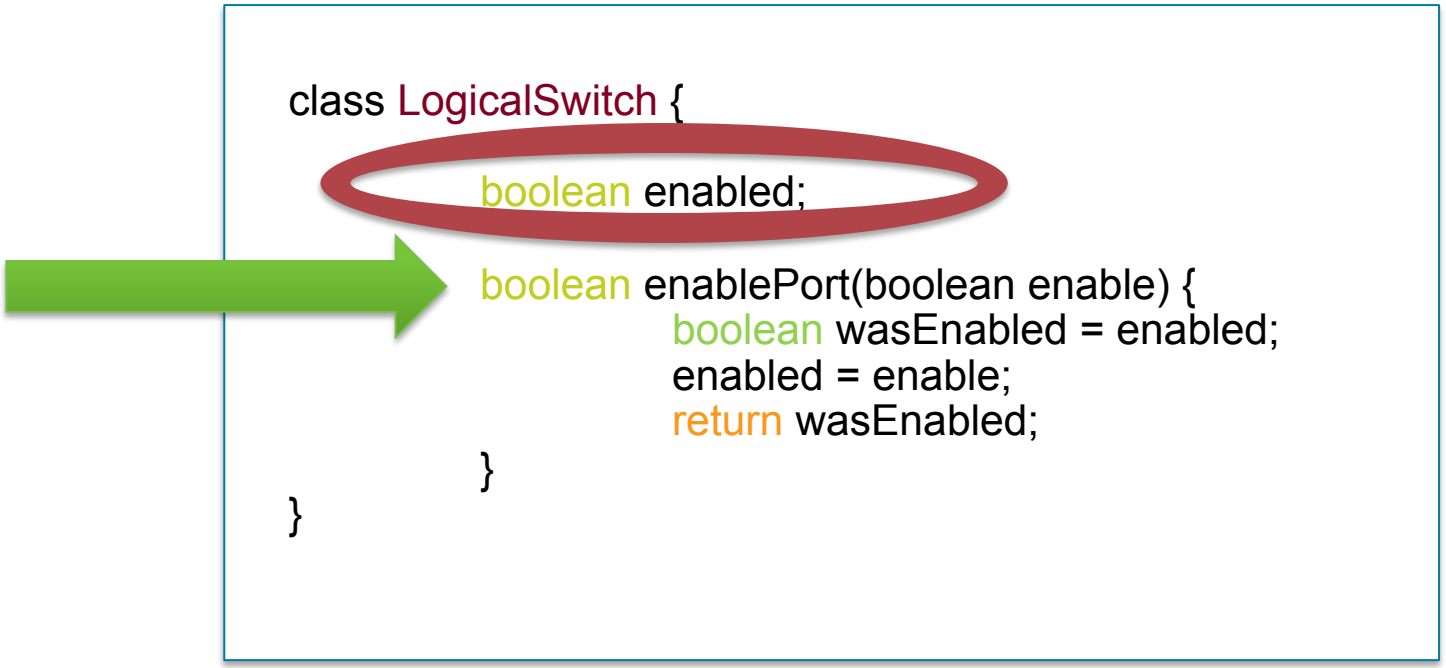
Corfu



Kafka



New Programming Model



New Programm

```
class LogicalSwitch {

        boolean enabled;

        boolean enablePort(boolean enable)
                boolean wasEnabled = enabled;
                enabled = enable;
                return wasEnabled;
        }
```

```
CorfuRuntime cr =
        new CorfuRuntime("10.0.0.1:9000");

LogicalSwitch mySwitch =
        cr.build()
                .setStreamName("switch-a")
                .setType(LogicalSwitch.class)
                .open();

mySwitch.enablePort(true);
```

New Programm

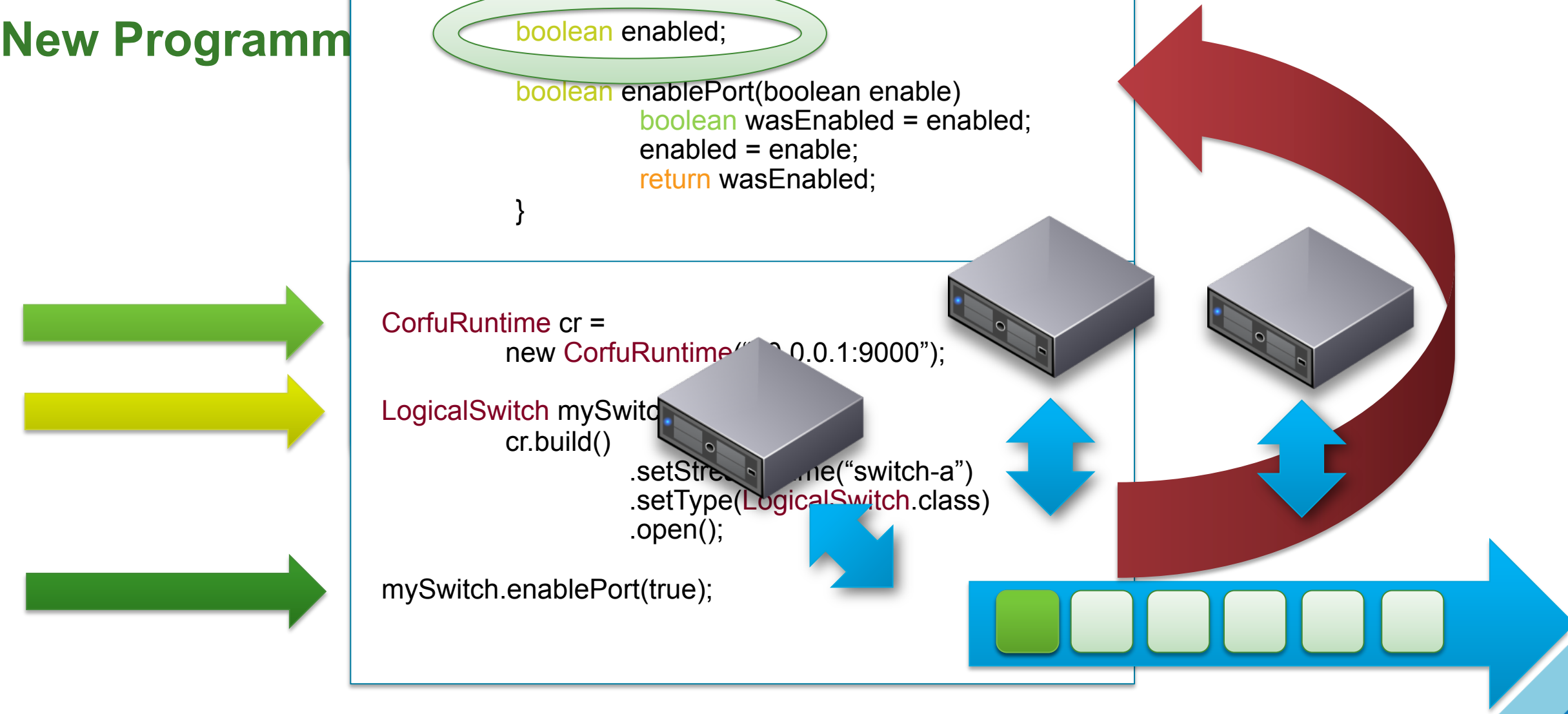
```
class LogicalSwitch {
        boolean enabled;

        boolean enablePort(boolean enable)
                boolean wasEnabled = enabled;
                enabled = enable;
                return wasEnabled;
        }
```

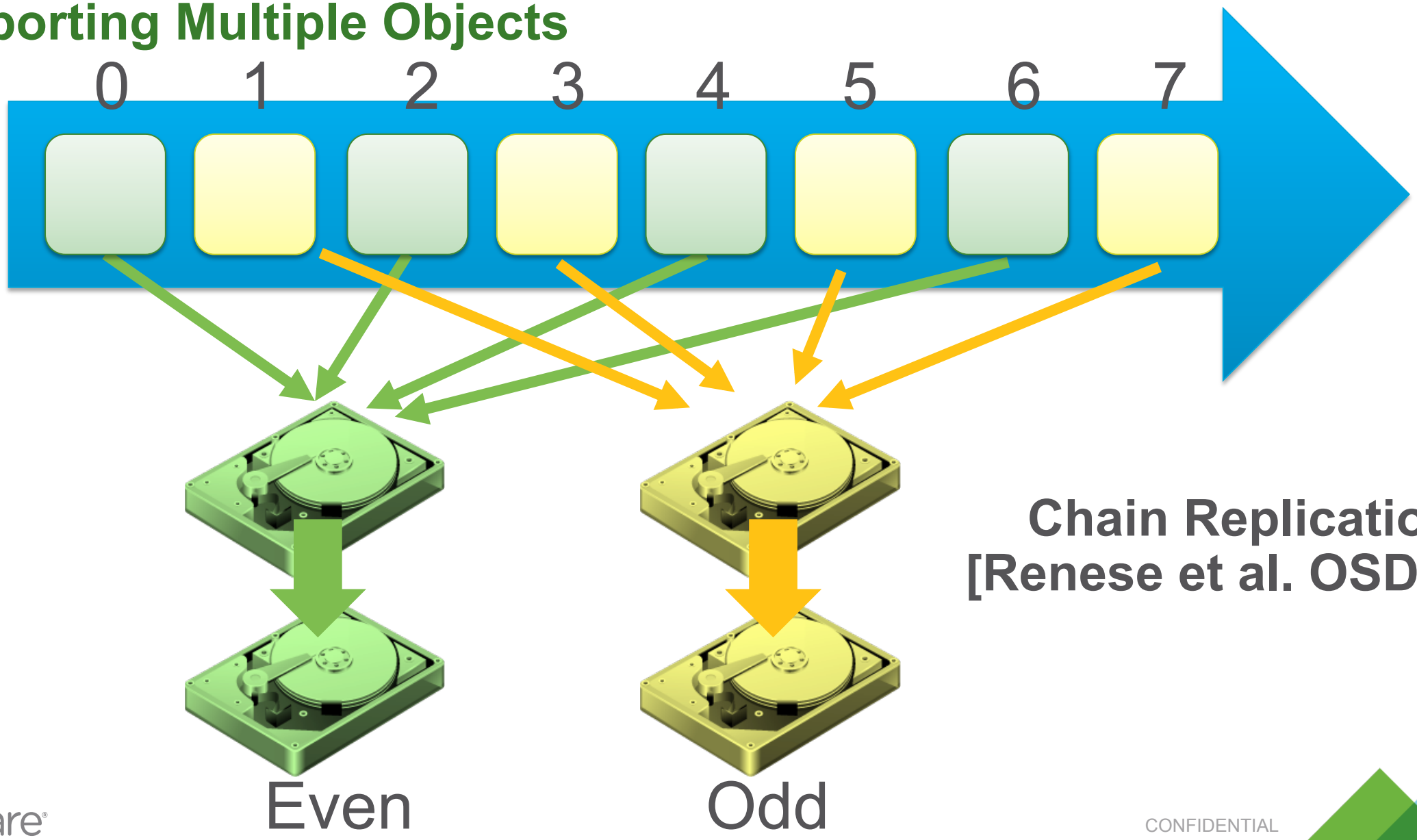
```
CorfuRuntime cr =
        new CorfuRuntime("...0.0.1:9000");

LogicalSwitch mySwitch
        cr.build()
                .setStre...me("switch-a")
                .setType(LogicalSwitch.class)
                .open();

mySwitch.enablePort(true);
```

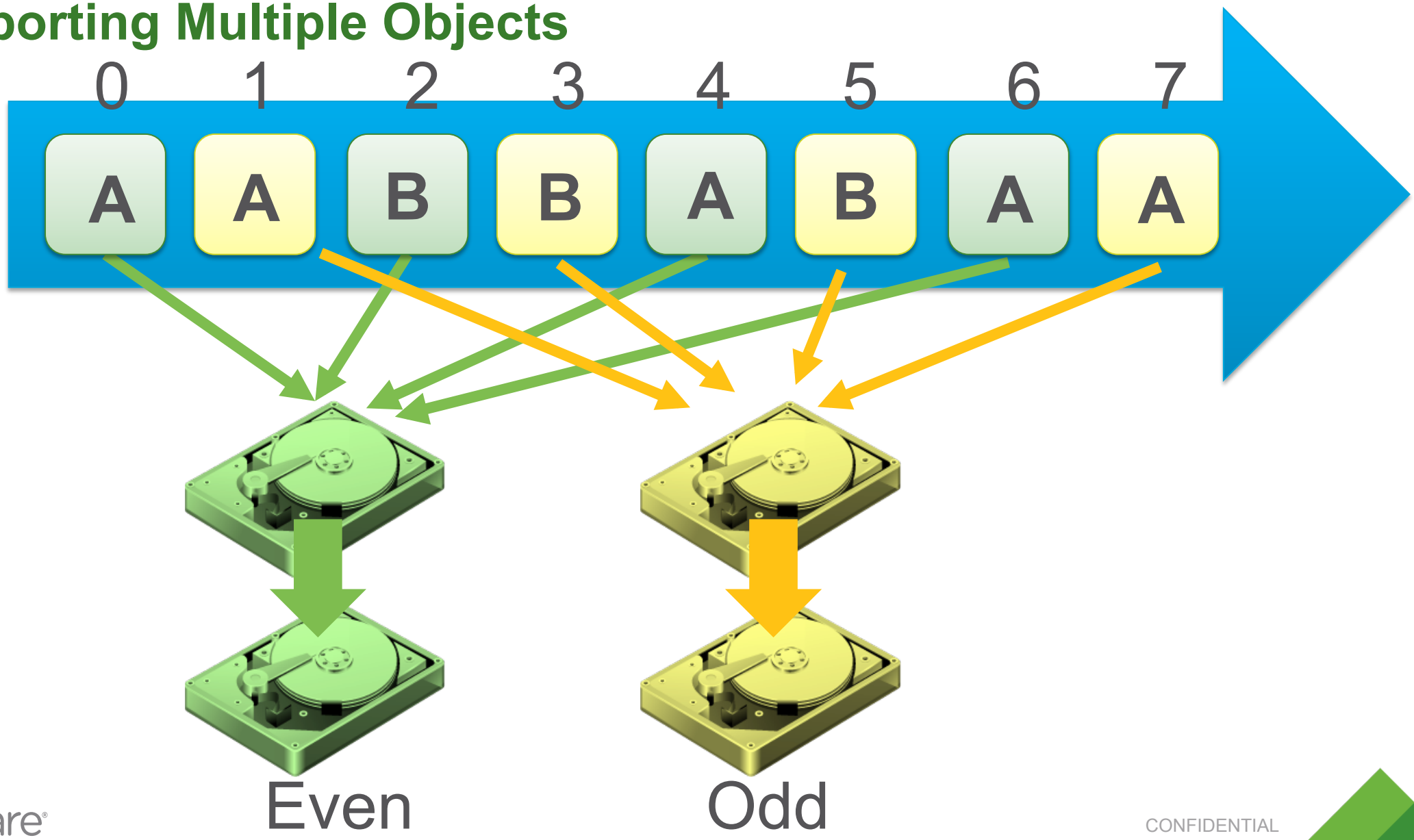


Supporting Multiple Objects

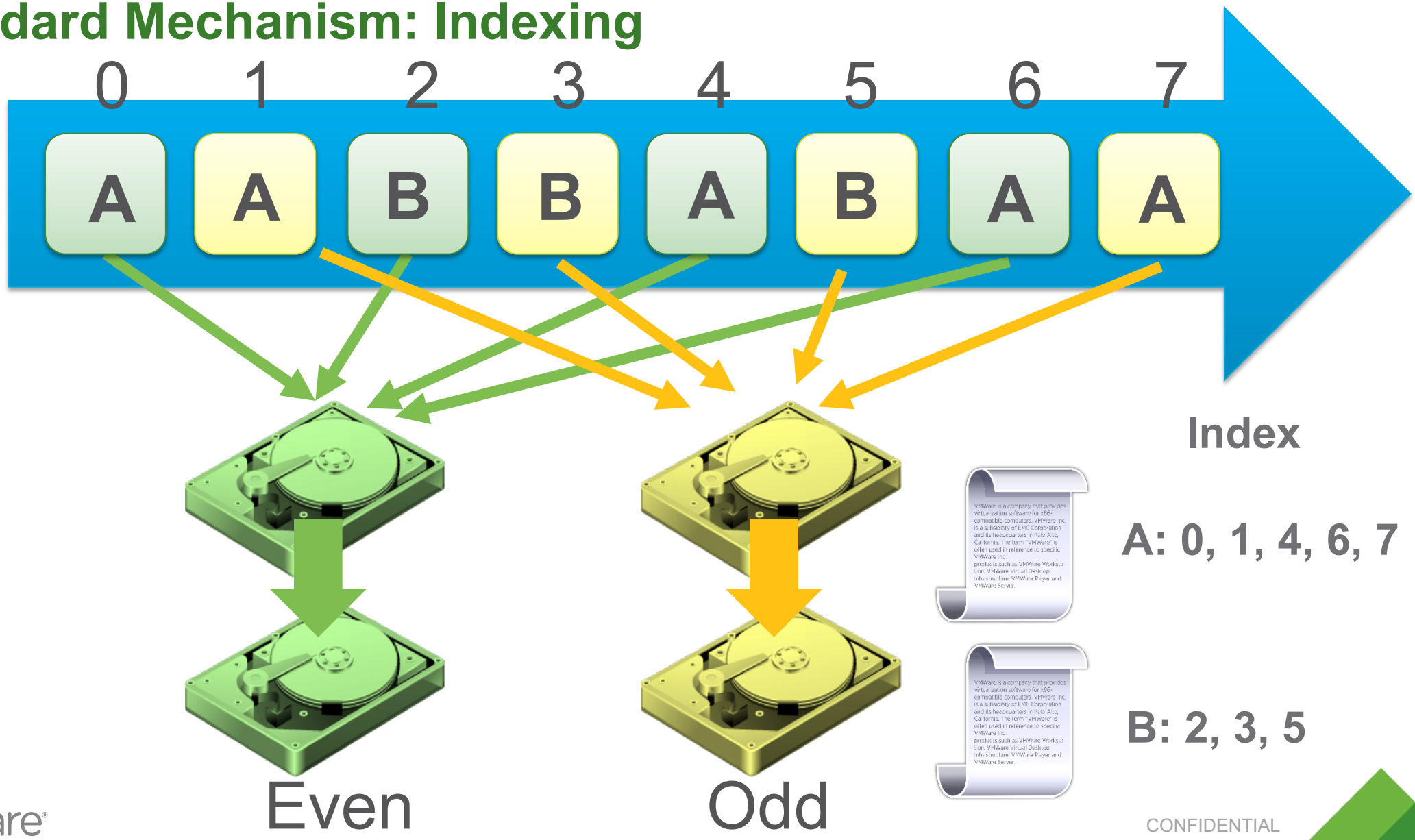


Chain Replication
[Renese et al. OSDI‘04]

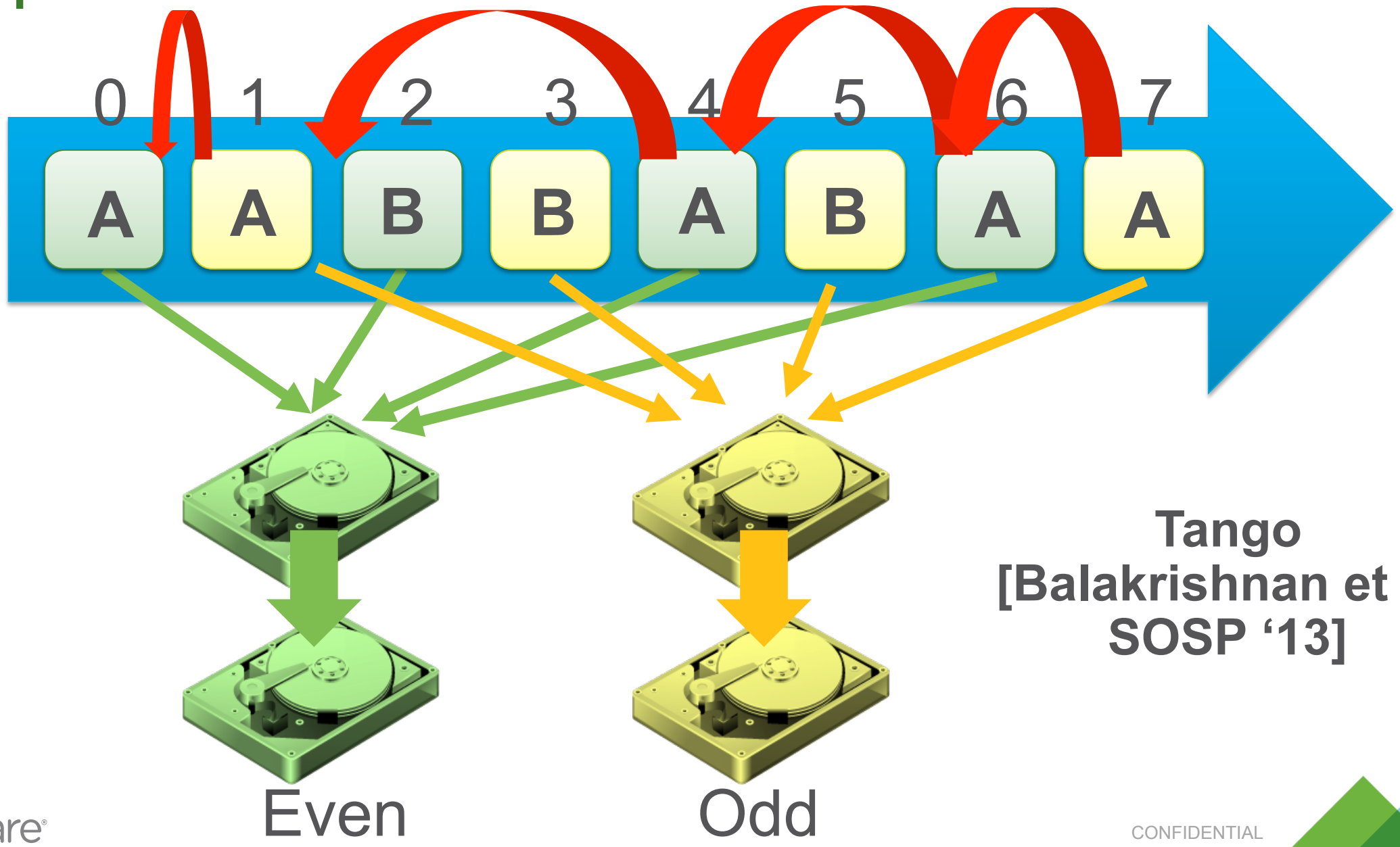
Supporting Multiple Objects



Standard Mechanism: Indexing

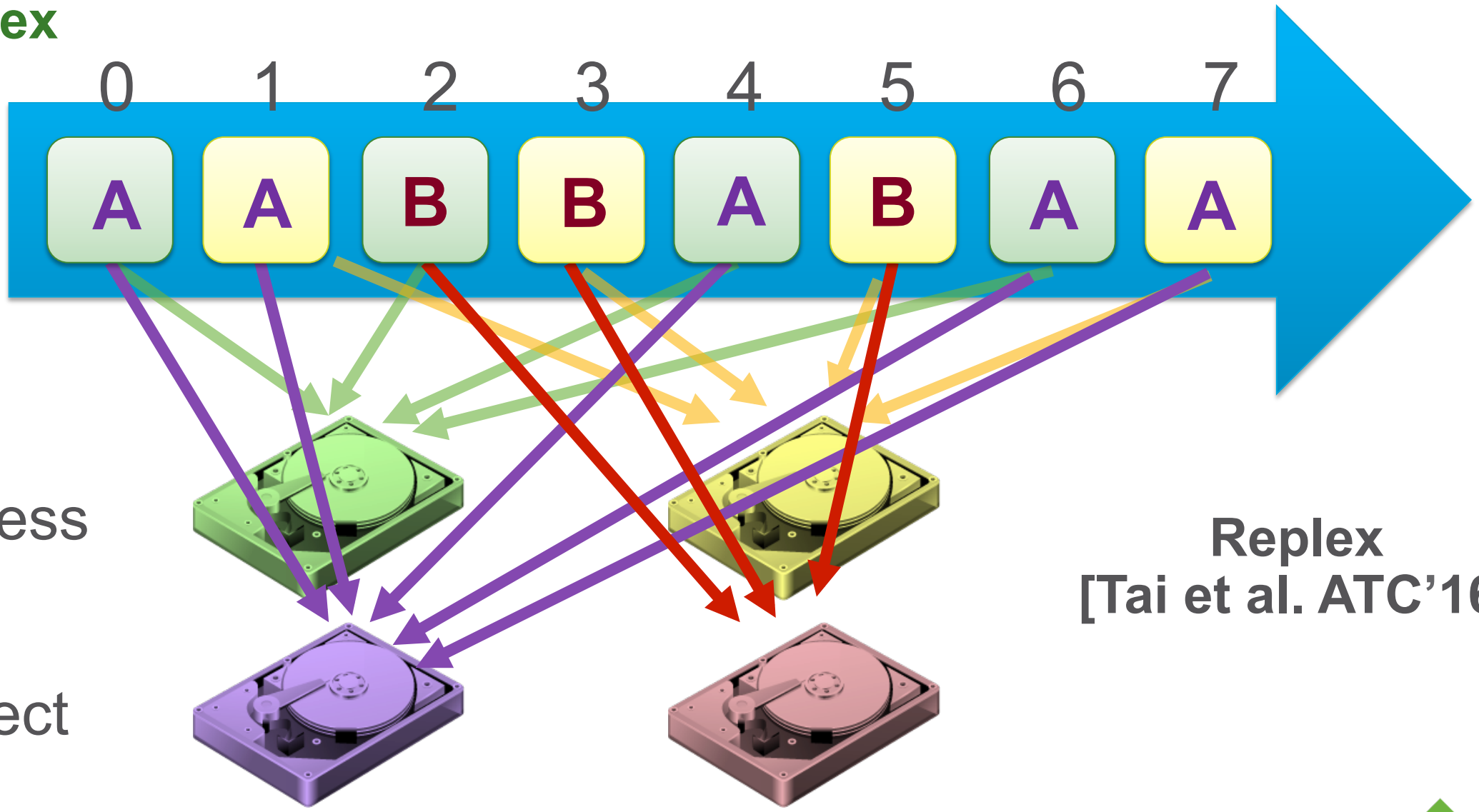


Backpointers



Tango [Balakrishnan et al. SOSP '13]

Replex

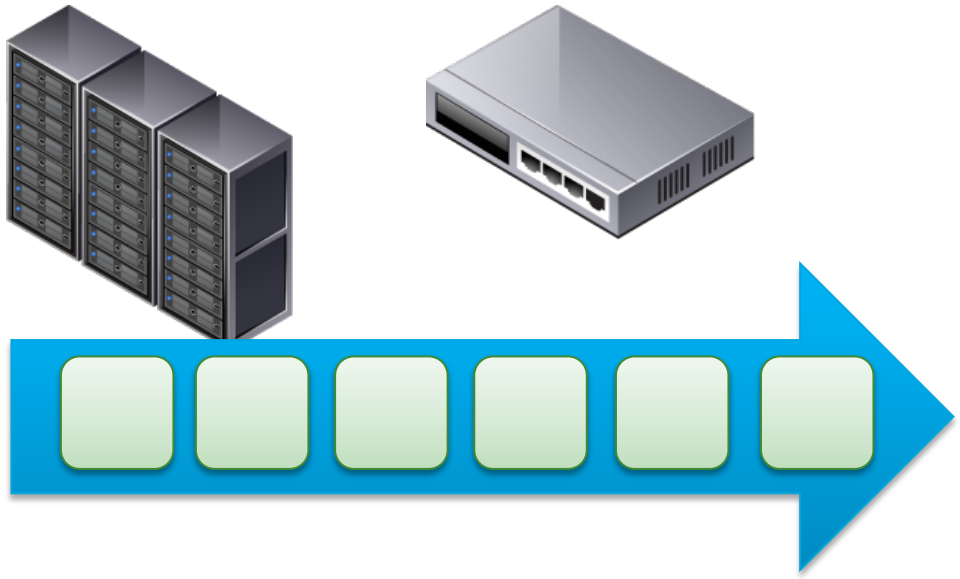


Replex
[Tai et al. ATC'16]

CMPLAT: Cluster Management Platform for NSX

Left to Right:
Scott Fritchie
Michael Wei(PoC),
Chris Rossbach,
Dahlia Malkhi (PoC),
Udi Wieder

Jim Stabile, Ittai Abraham,
Sandeep Uttamchandani
Medhavi Dawan,
Maithem Munshed



- Re-engineer NSX cluster management to use Corfu “under the covers”
- Avoid the need for every NSX installation to include Zookeeper, etc.
- Corfu evolved significantly as a result of this joint effort

CMPLAT: Cluster Management Platform for NSX

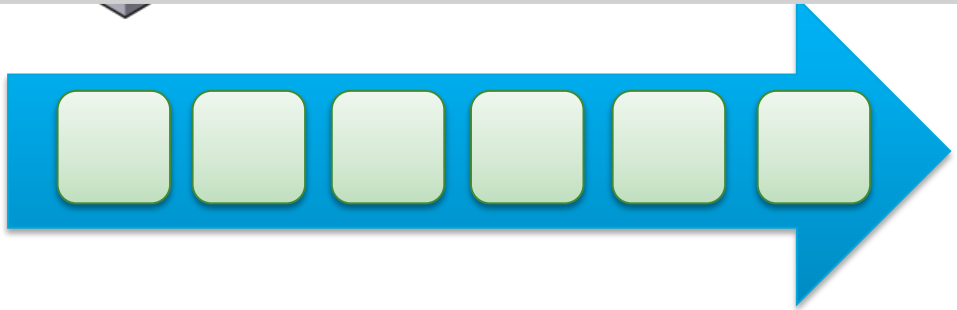
Left to Right:
Scott Fritchie
Michael Wei(PoC),
Chris Rossbach,
Dahlia Malkhi (PoC),
Udi Wieder

Jim Stabile, Ittai Abraham,
Sandeep Uttamchandani
Medhavi Dawan,
Maithem Munshed

API Requests

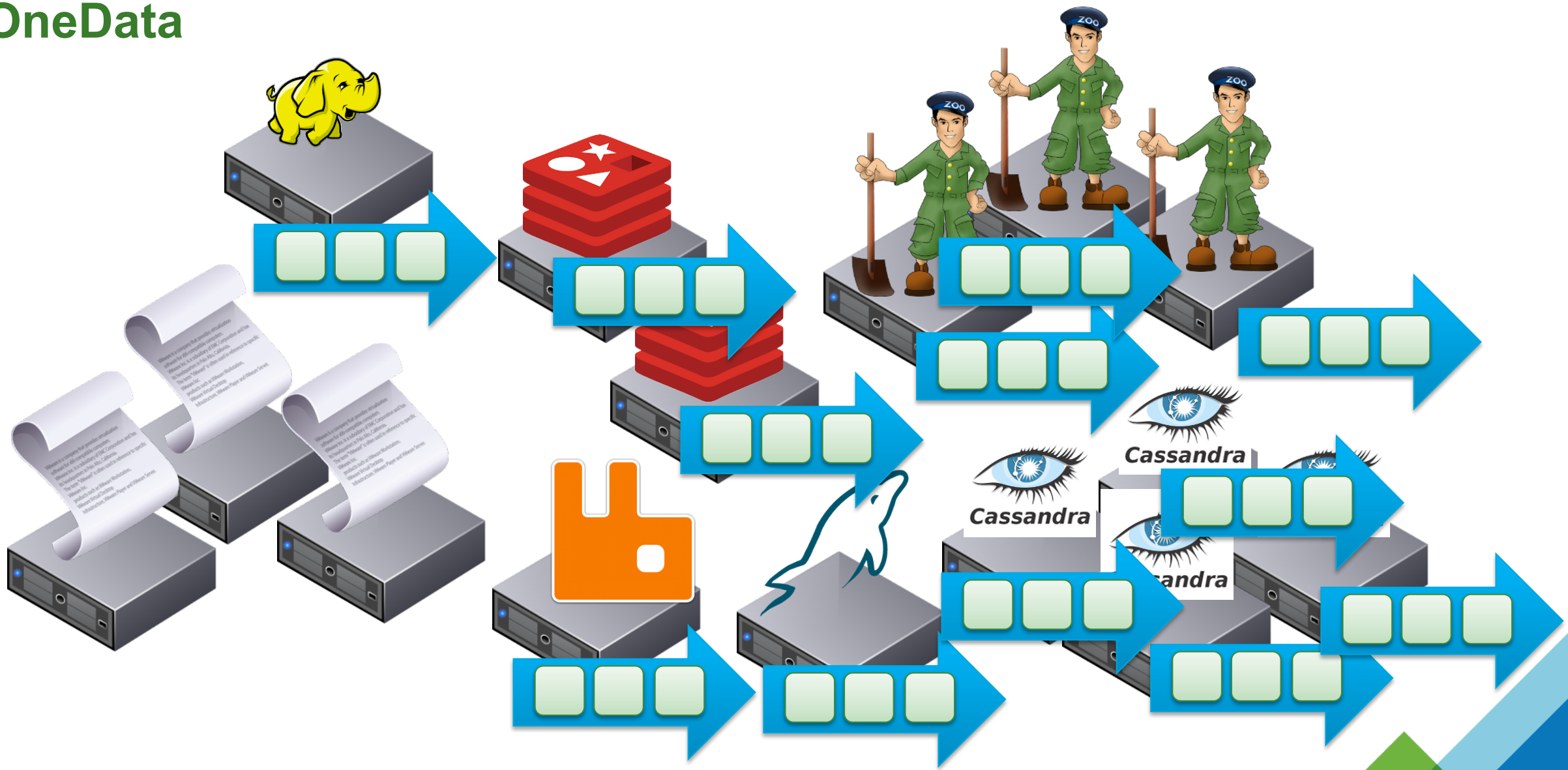
NSX Scale-Out Management Plane

NSX Java Object Model



Corfu Distributed Log

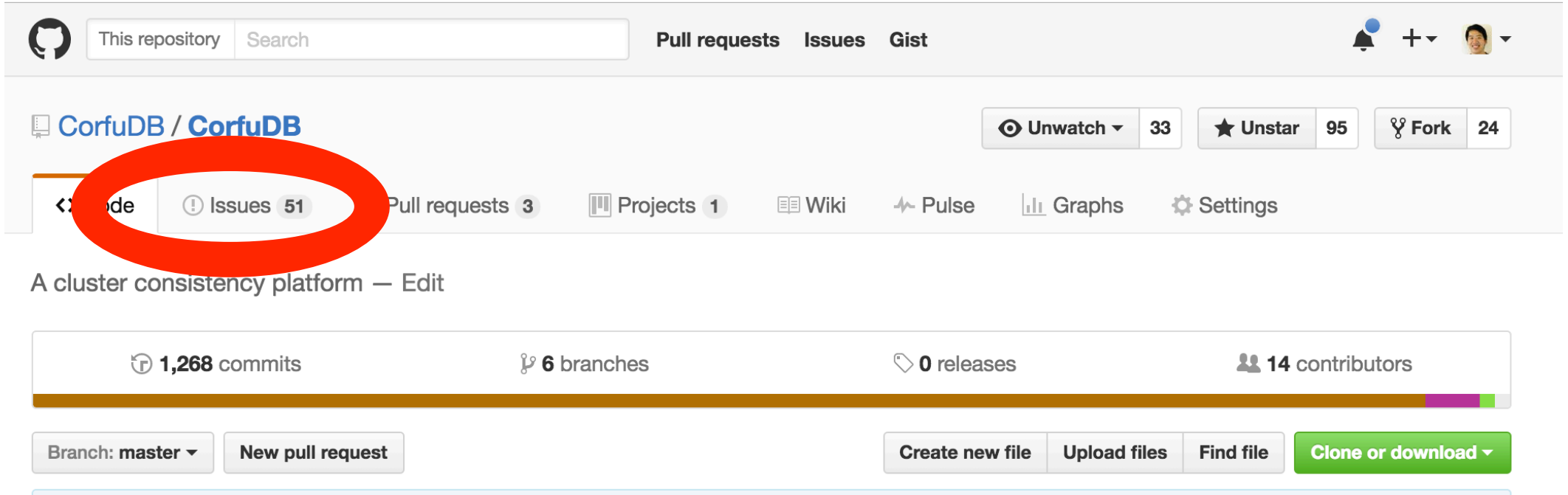
OneData



Open Source Efforts

Yes, we're on Github…

github.com/CorfuDB/CorfuDB



No, this is not a source code dump…