Being Afraid - How Paranoia at Dropbox Protects your Data

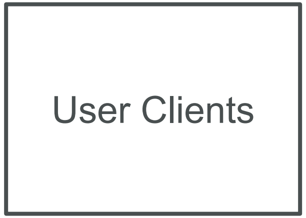
David Mah

Dropbox

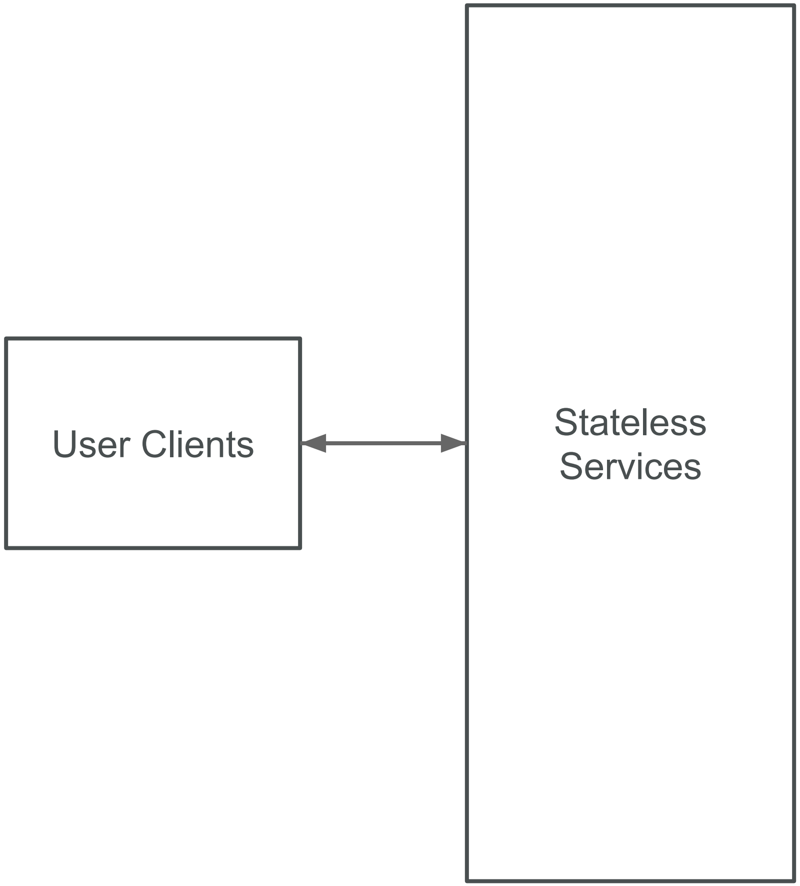
Trust

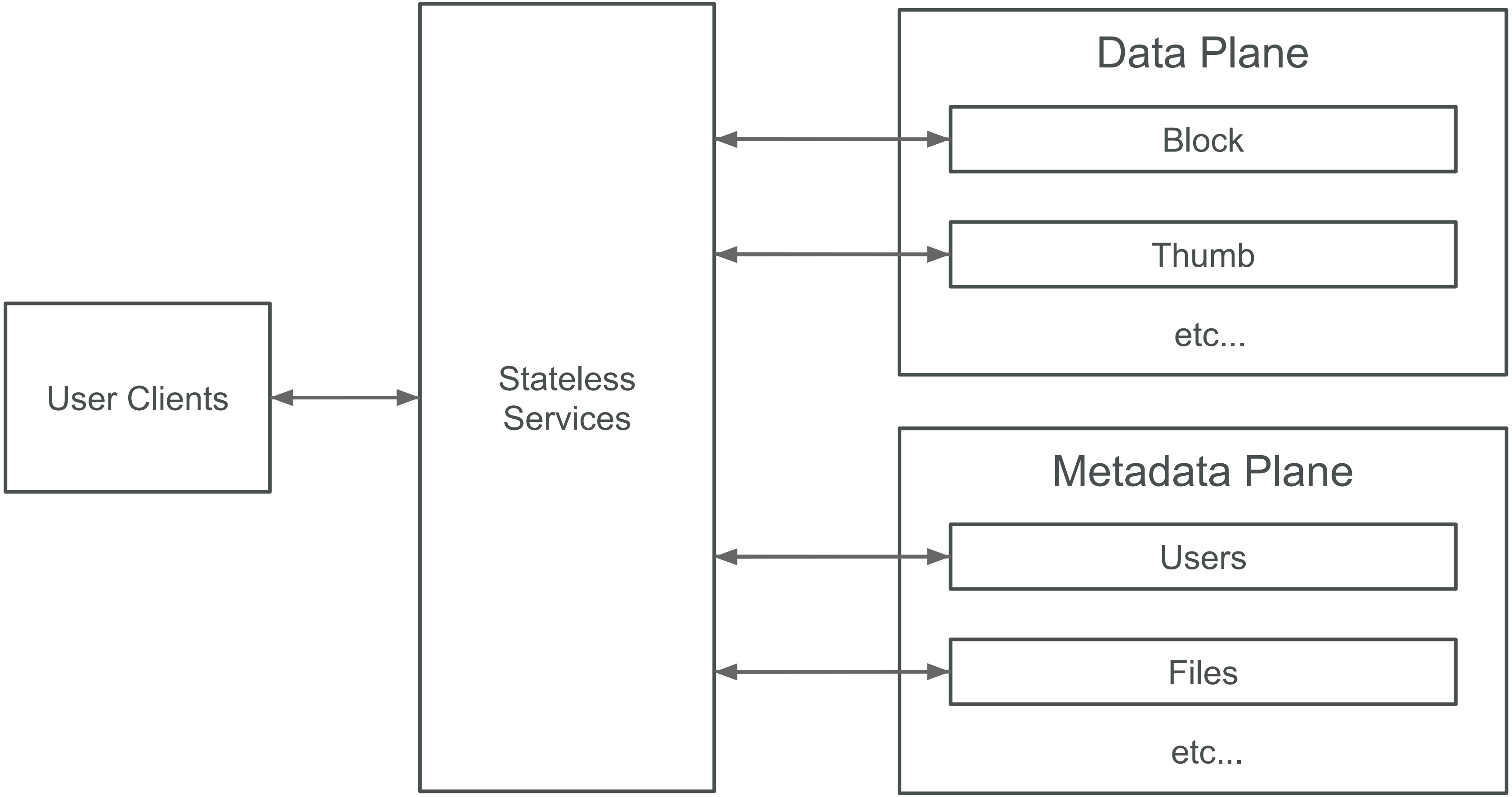
Trust!!

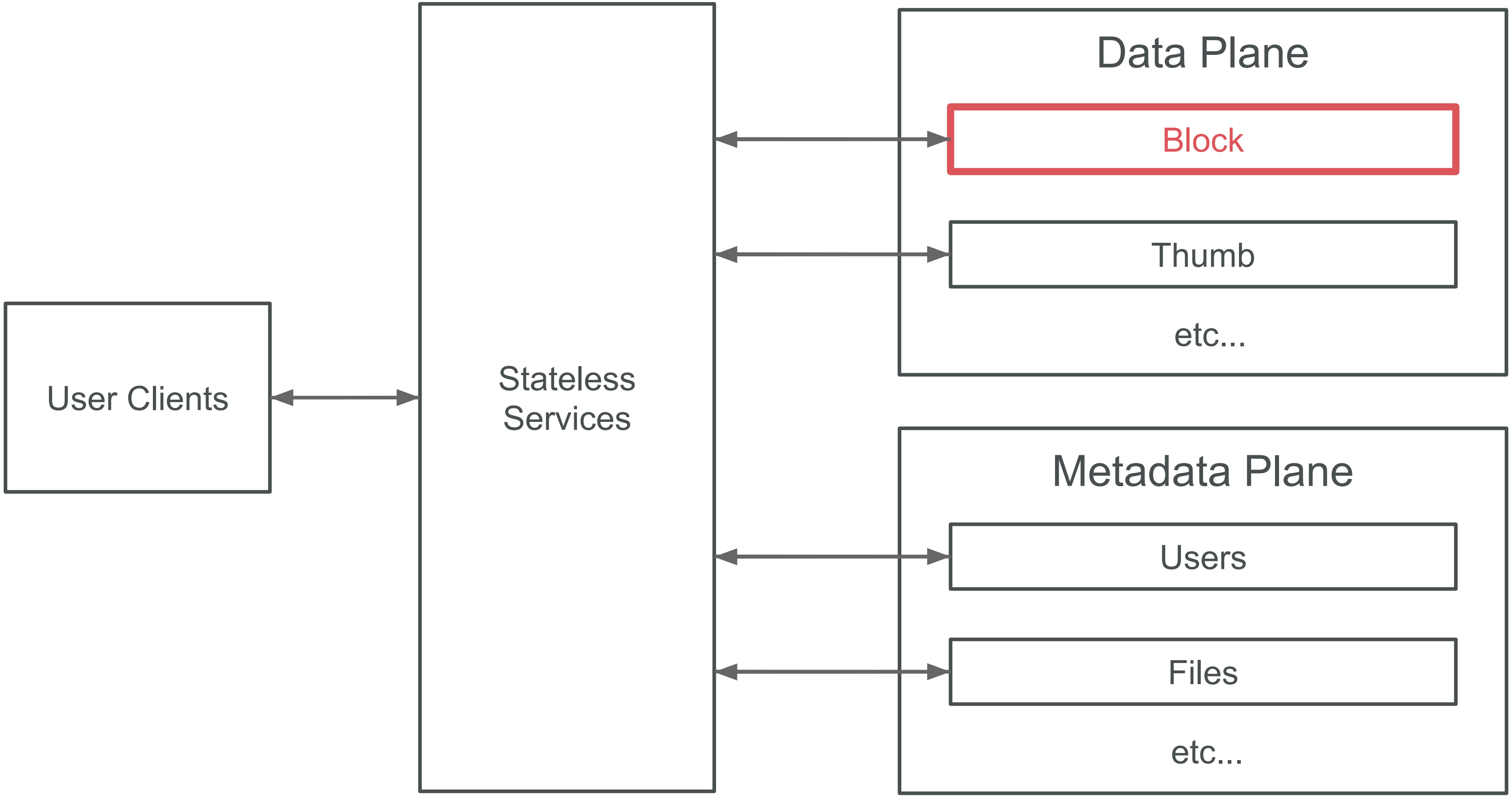
Me?



User Clients


User Clients
Stateless Services


User Clients
Stateless Services
Data Plane
Block
Thumb
etc...
Metadata Plane
Users
Files
etc...


User Clients
Stateless Services
Data Plane
Block
Thumb
etc...
Metadata Plane
Users
Files
etc...

Blockstore

Durability

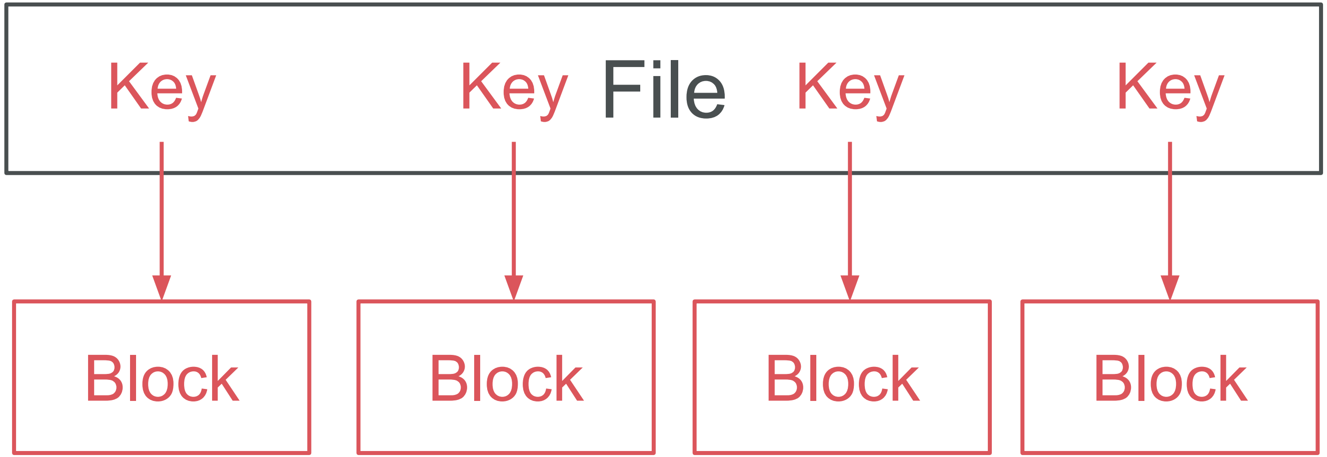
Architecture

Architecture
Fears and Defenses

Architecture

Fears and Defenses

File


Key
Key
File
Key
Key
Block
Block
Block
Block

Key

Unique ID for a block.

Block

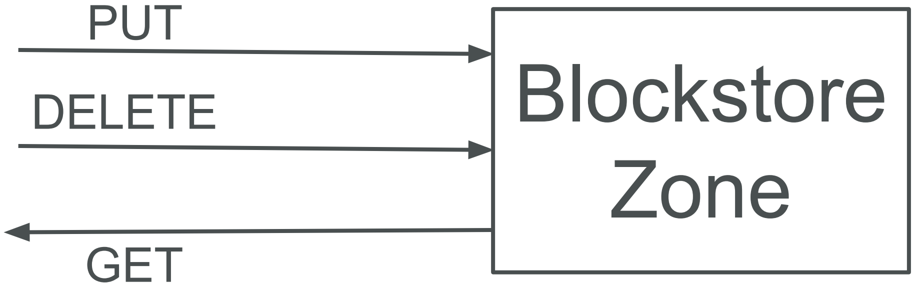
Blob of data on MB scale

Key

Block

API:

```
put(key, block)
get(key)
delete(key)
```

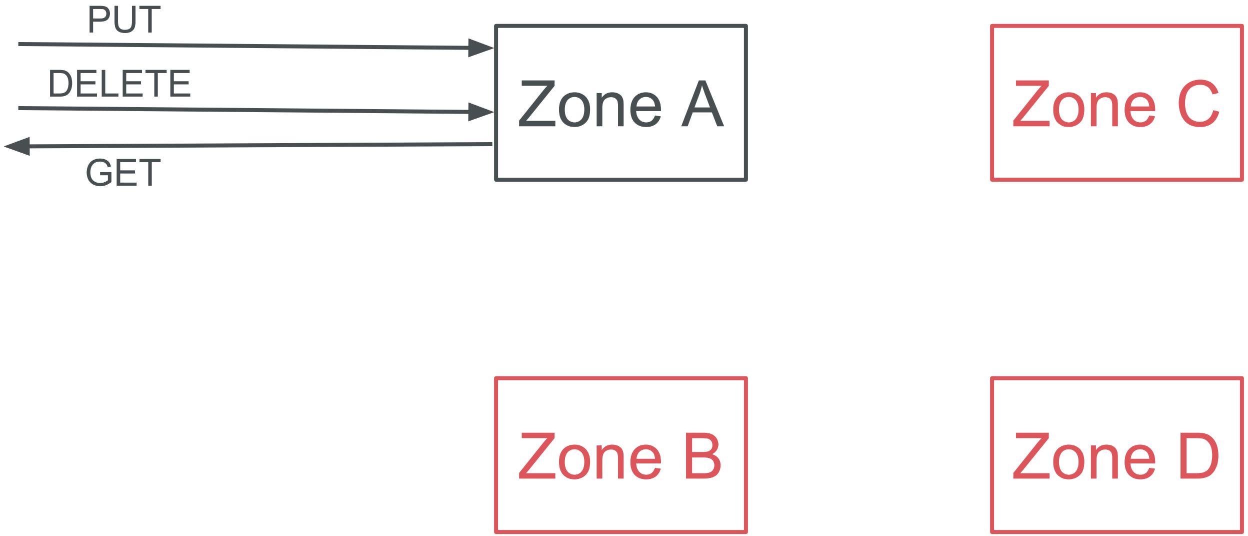


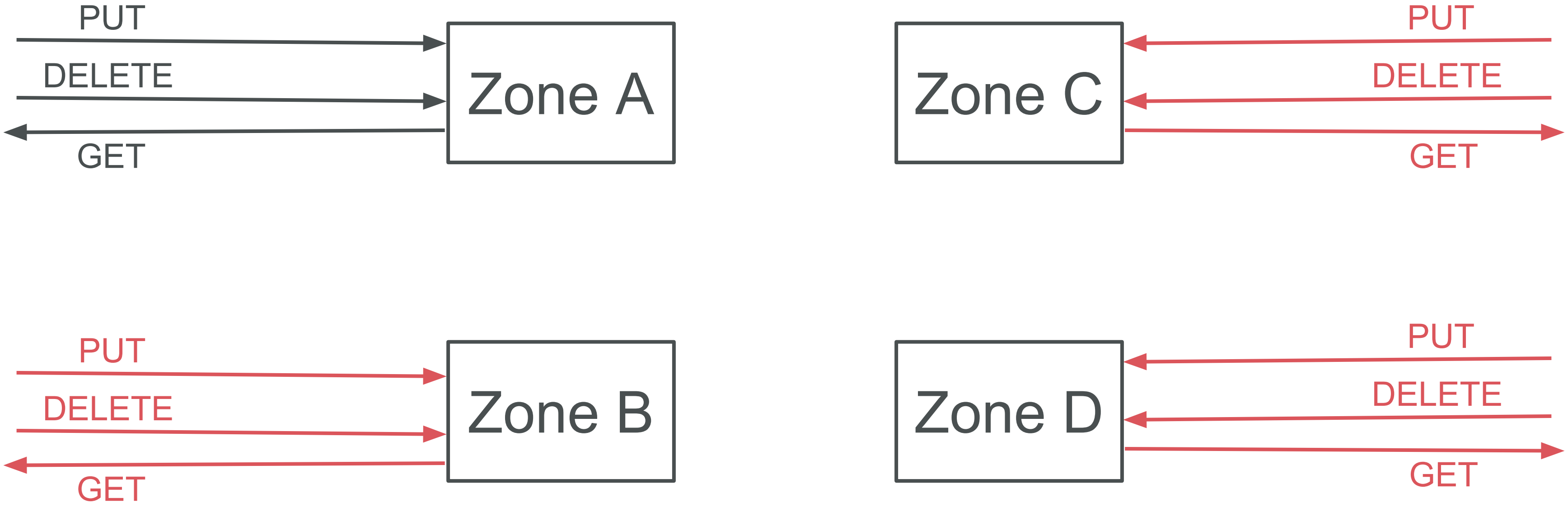
PUT

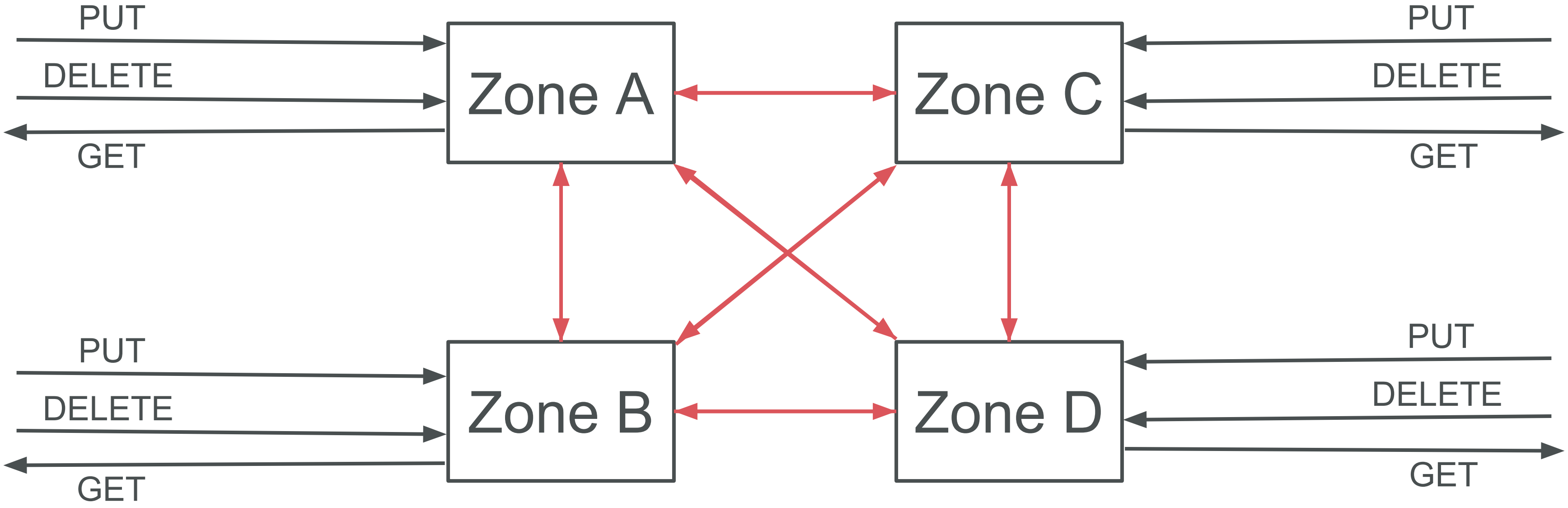
DELETE

GET

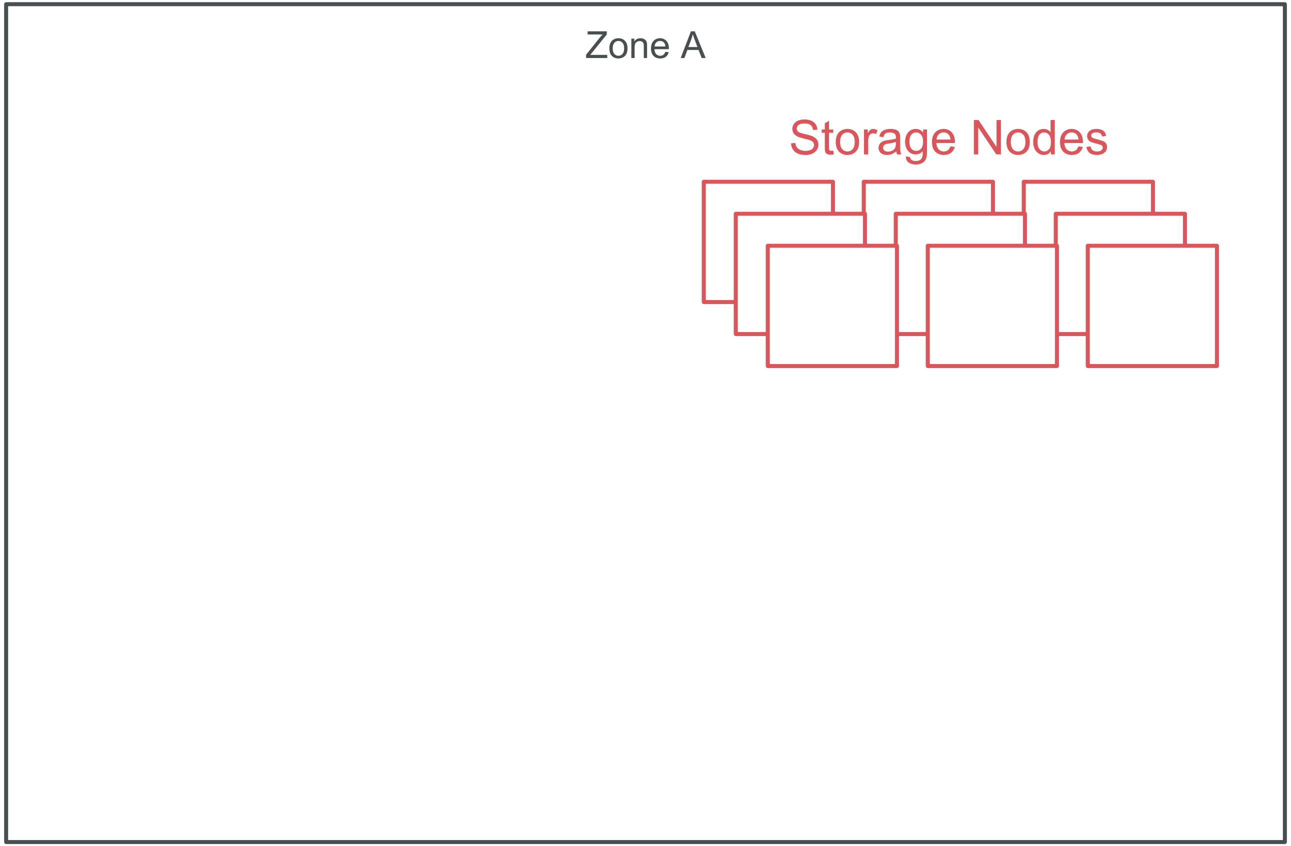
Blockstore Zone


PUT
DELETE
GET
Zone A
Zone C
Zone B
Zone D


PUT
DELETE
GET
Zone A
Zone C
PUT
DELETE
GET
PUT
DELETE
GET
Zone B
Zone D
PUT
DELETE
GET

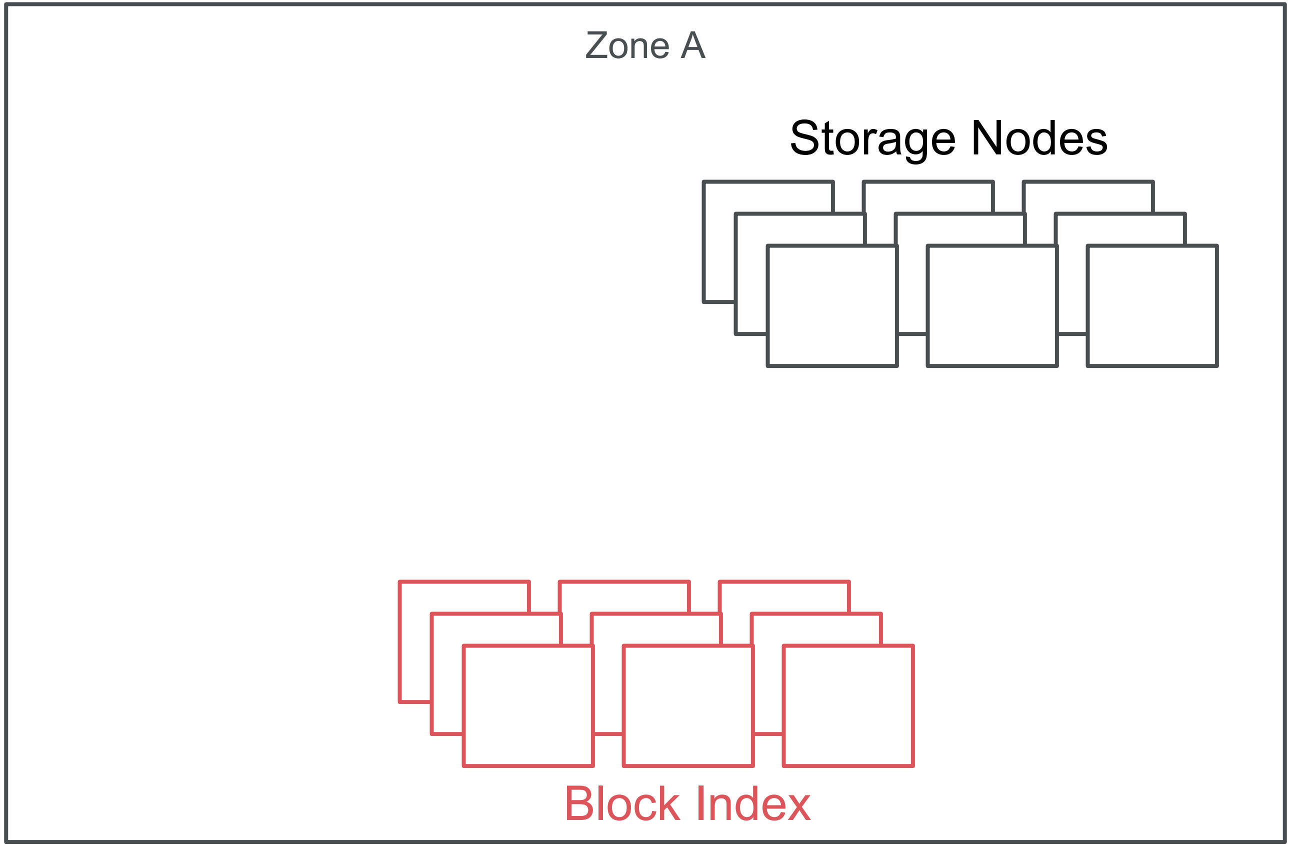

PUT
DELETE
GET
Zone A
PUT
DELETE
GET
Zone C
PUT
DELETE
GET
Zone B
PUT
DELETE
GET
Zone D

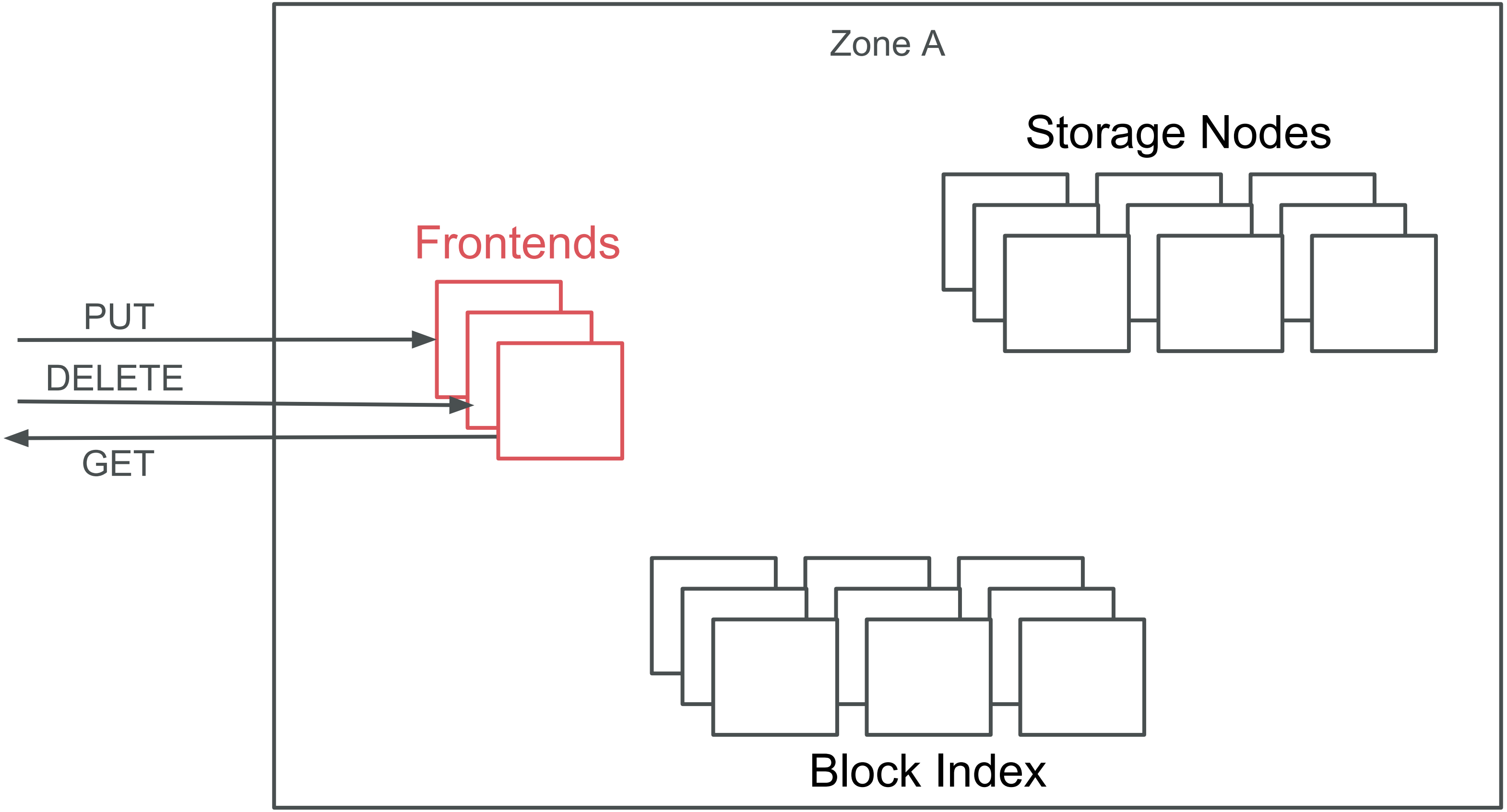
Zone A

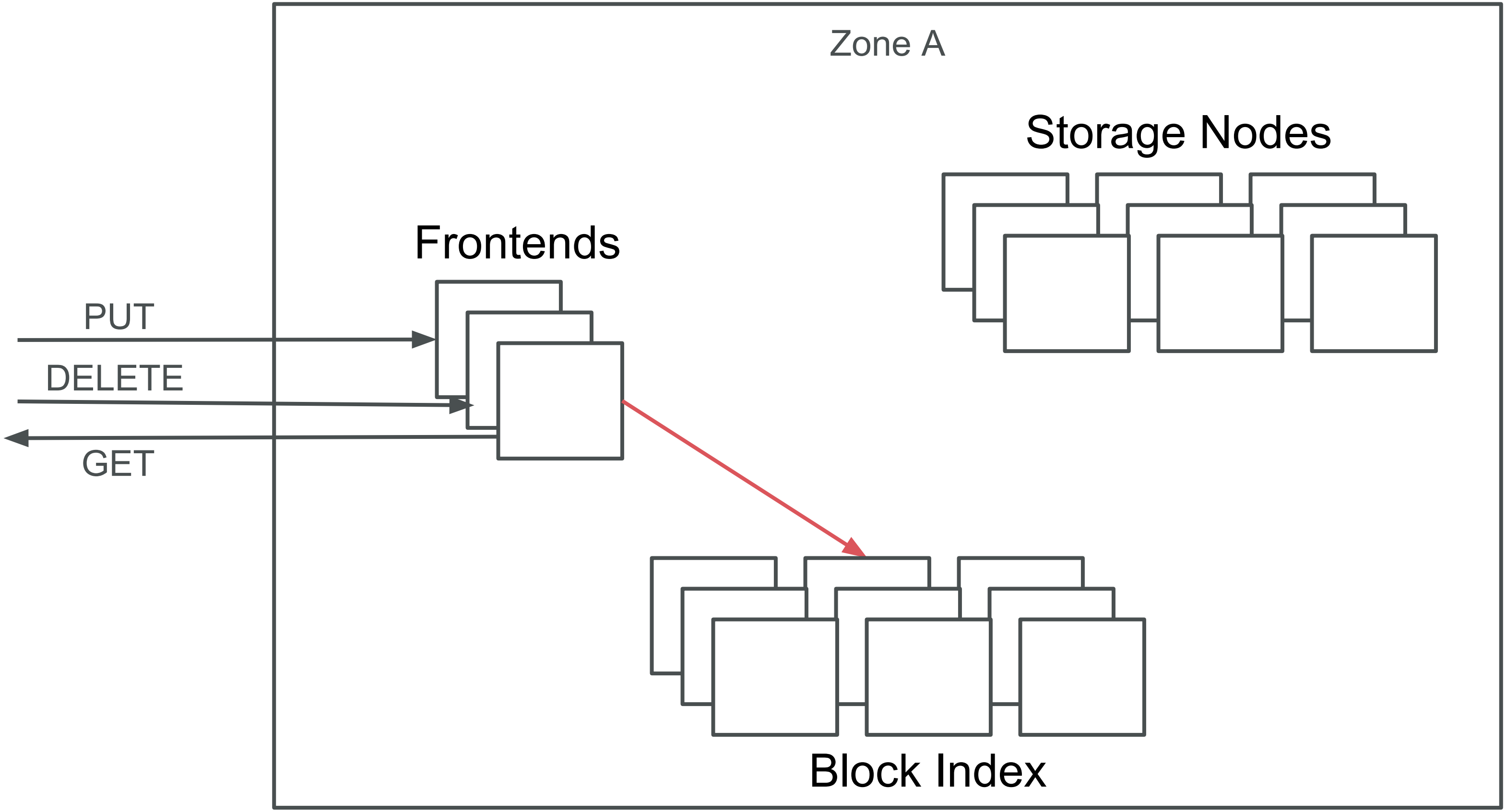


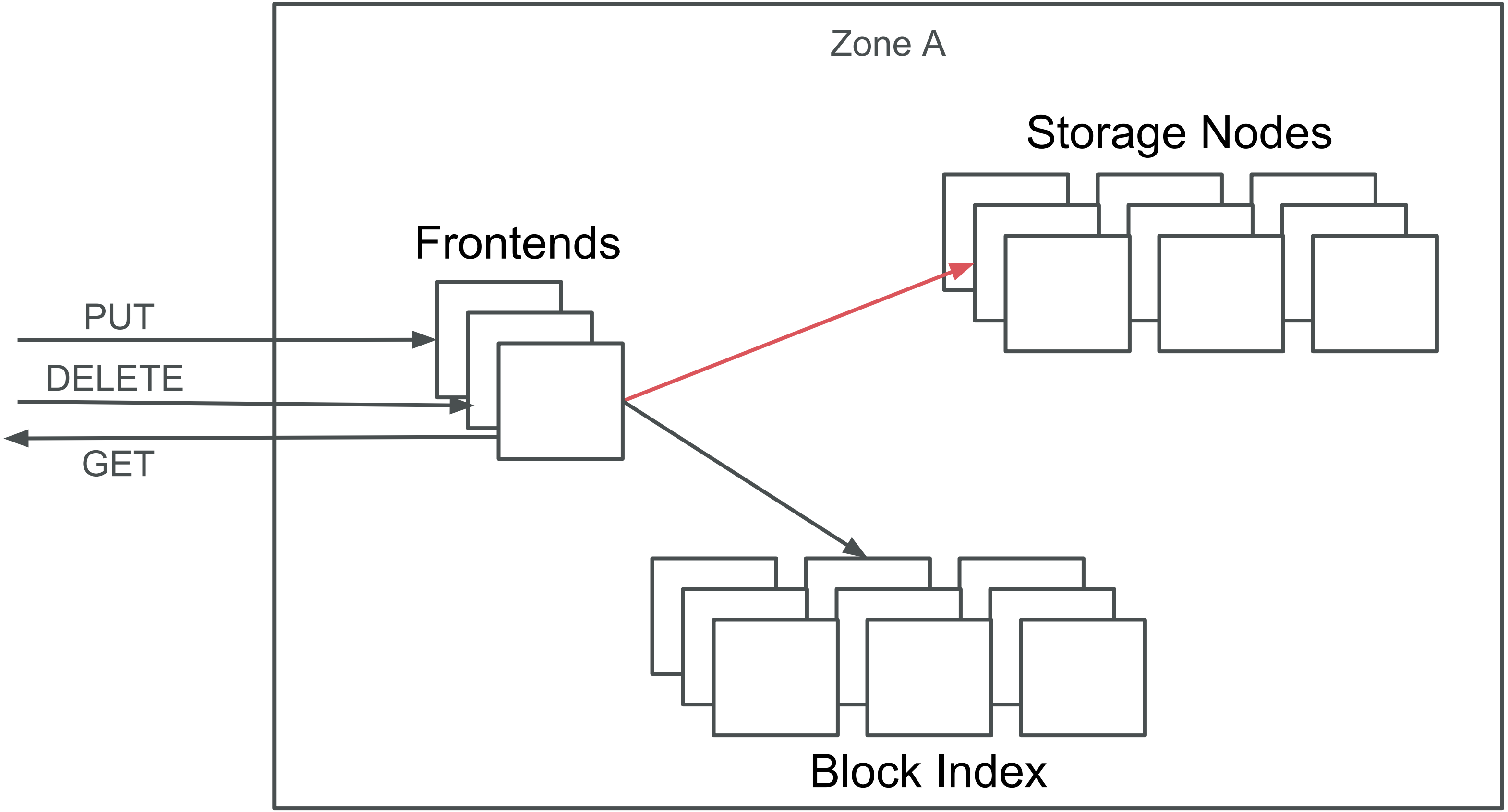
Zone A

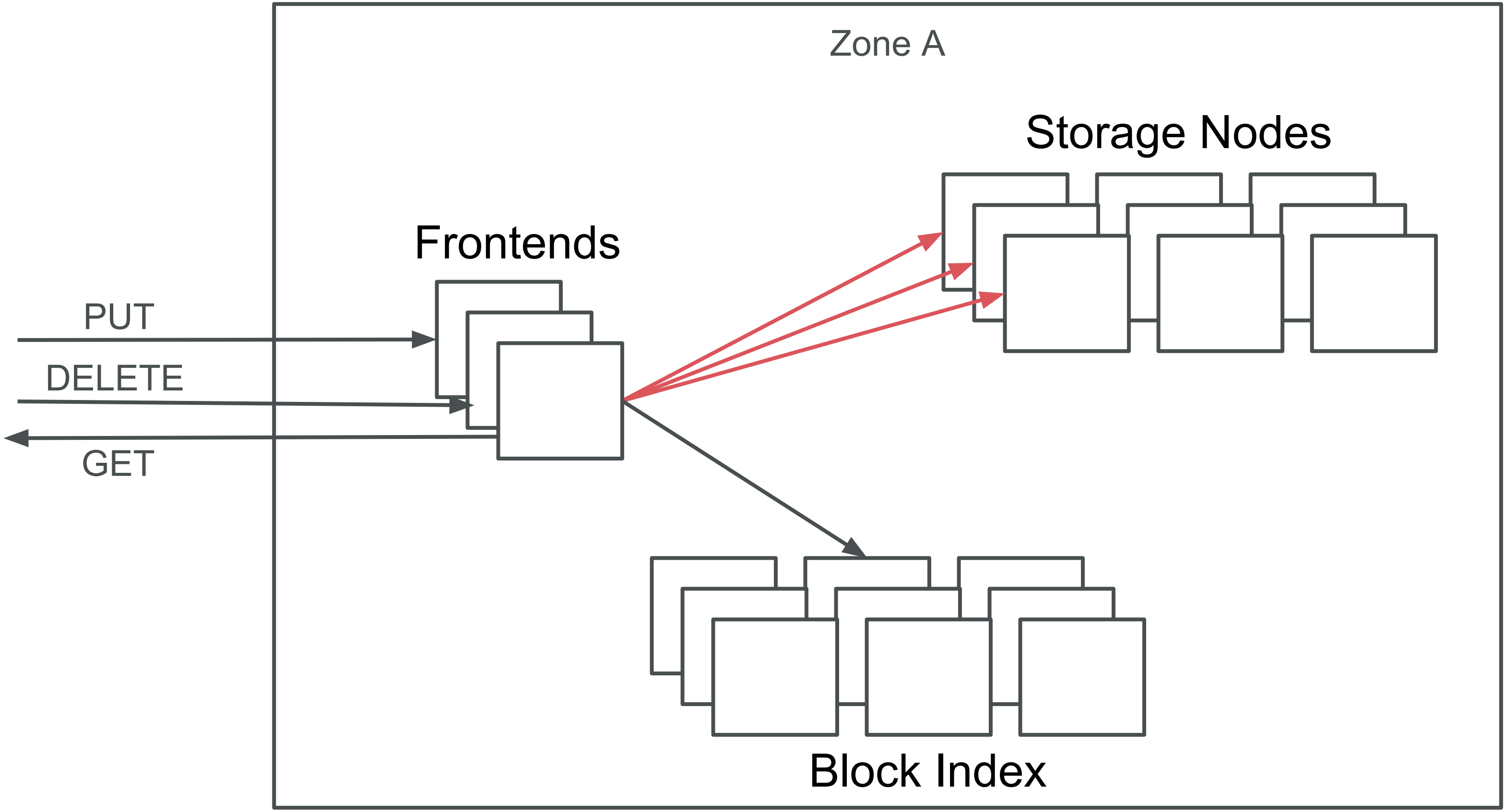
Storage Nodes

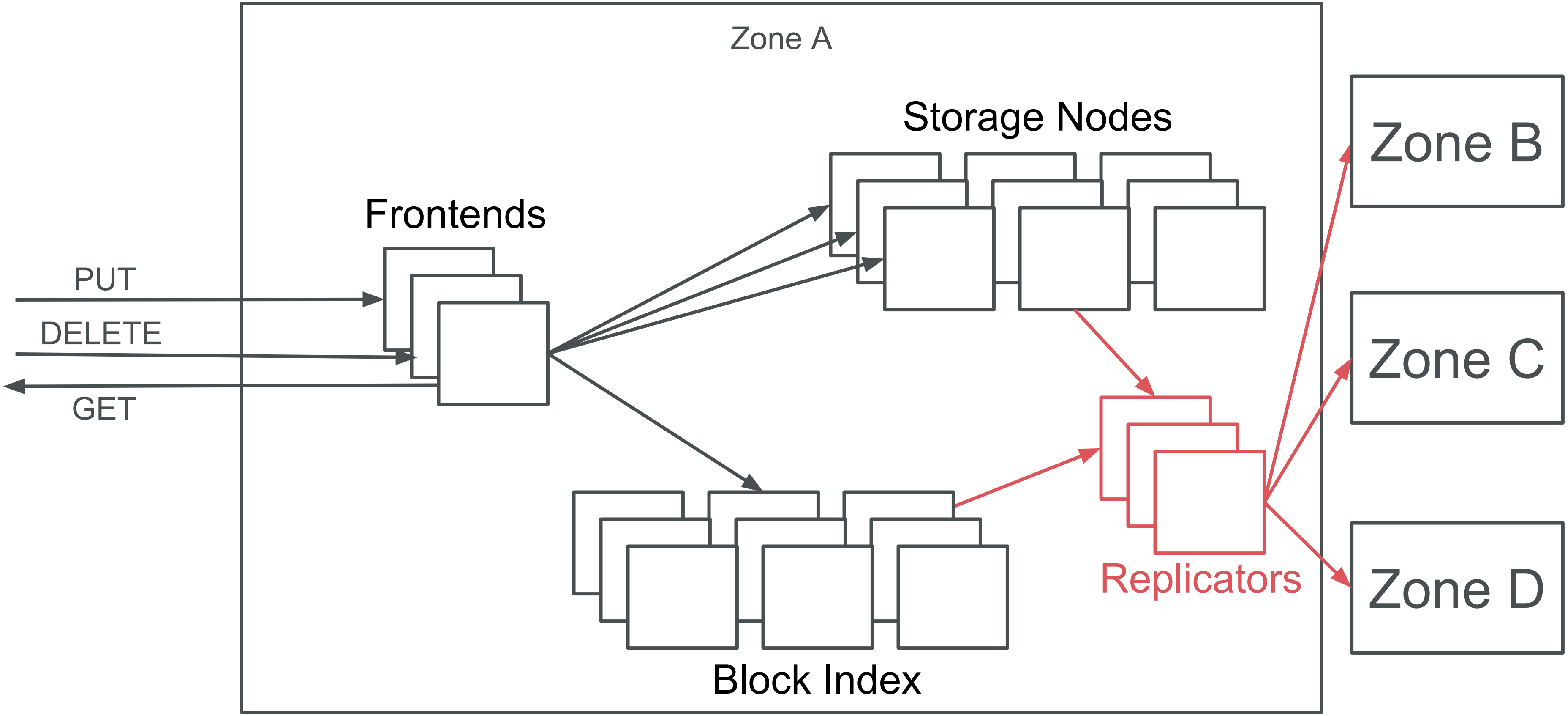

Zone A
Storage Nodes
Block Index


Zone A
Storage Nodes
Frontends
PUT
DELETE
GET
Block Index


Zone A
Storage Nodes
Frontends
PUT
DELETE
GET
Block Index


Zone A
Storage Nodes
Frontends
PUT
DELETE
GET
Block Index


Zone A
Storage Nodes
Frontends
PUT
DELETE
GET
Block Index


Zone A
Storage Nodes
Frontends
PUT
DELETE
GET
Replicators
Block Index
Zone B
Zone C
Zone D

Architecture

Fears and Defenses

Fears

Moving things are scary

Moving things create accidents

What accidents are the scariest?

chance x severity = danger

Alien attack?

not in lifetime x death → minor concern

probably

Operator reboots wrong server?

once a week? x node down→ valid concern

Automation reimages wrong disks?

once a year? x data loss→ HUGE concern

Biggest Fears

Biggest Fears

Software

Biggest Fears

Software

Hardware

Biggest Fears

Software

Hardware

Humans

Biggest Fears

Software

Hardware

Humans

Tooling/Automation

Combating fears?

Verify!

Protect!

Biggest Fears

Software

Hardware

Humans

Tooling/Automation

Biggest Fears

Software
Hardware
Humans
Tooling/Automation
Protections

Biggest Fears

Software

Hardware

Humans

Tooling/Automation

Protections

Fear of Software

Corruption inducing bugs...

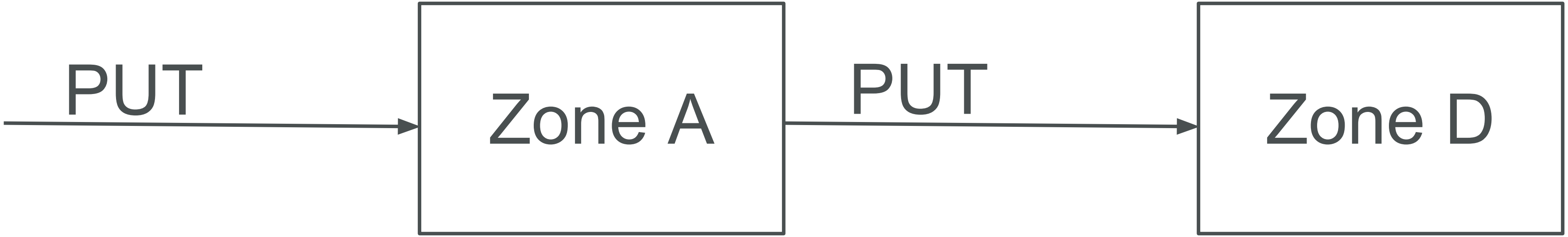
Software bugs? Verify!

Testing is cool..

Bugs/Crashes?

Those are normal case

Story Time!


PUT
Zone A
PUT
Zone D

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS
1113	TTTTTTTTT

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS
1113	TTTTTTTTT
1114	UUUUUUUU

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR

Replication Queue (MySQL)

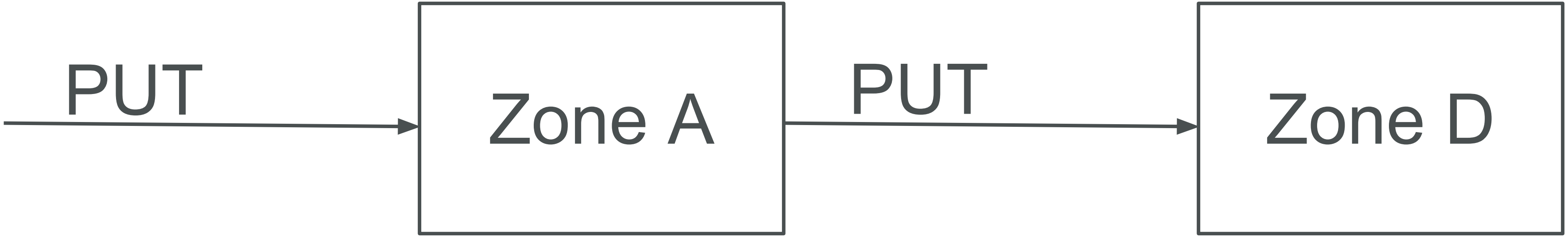
Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS
1114	UUUUUUUU

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS
1113	TTTTTTTTT
1114	UUUUUUUU

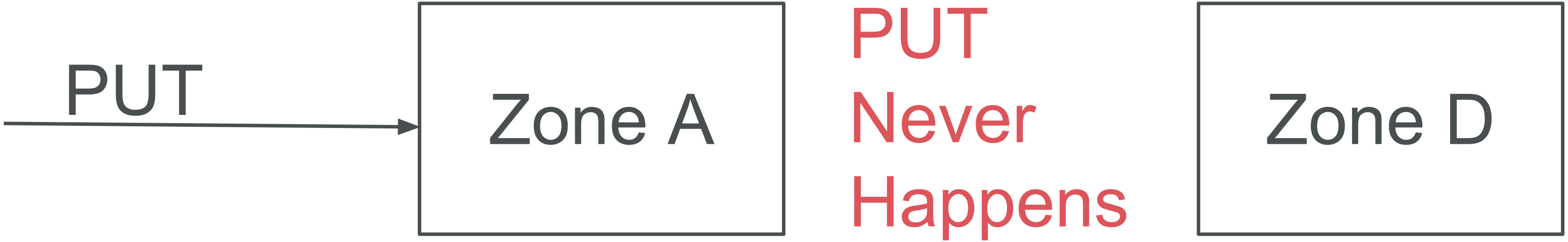


PUT

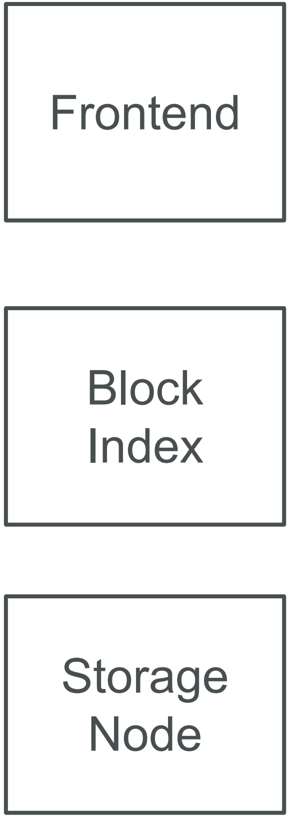
Zone A

PUT

Zone D


PUT
Zone A
PUT
Never
Happens
Zone D

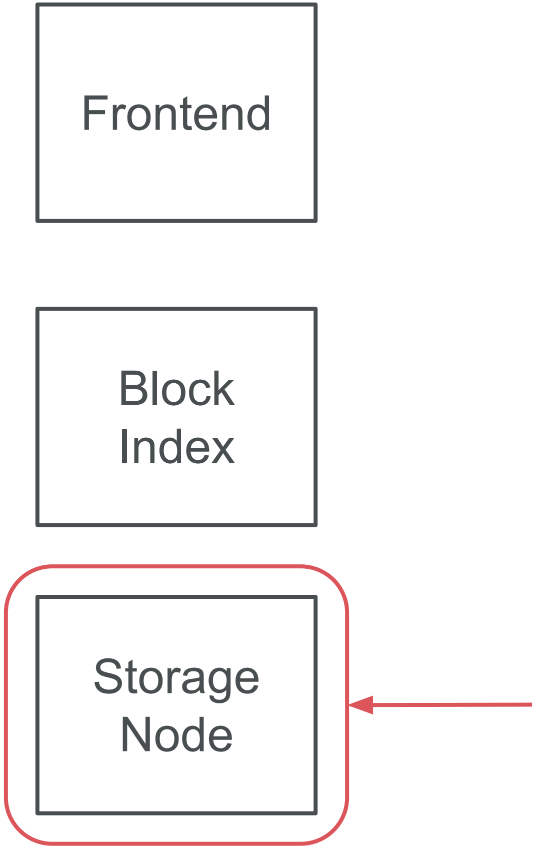
Series of Checkers



Frontend

Block Index

Storage Node



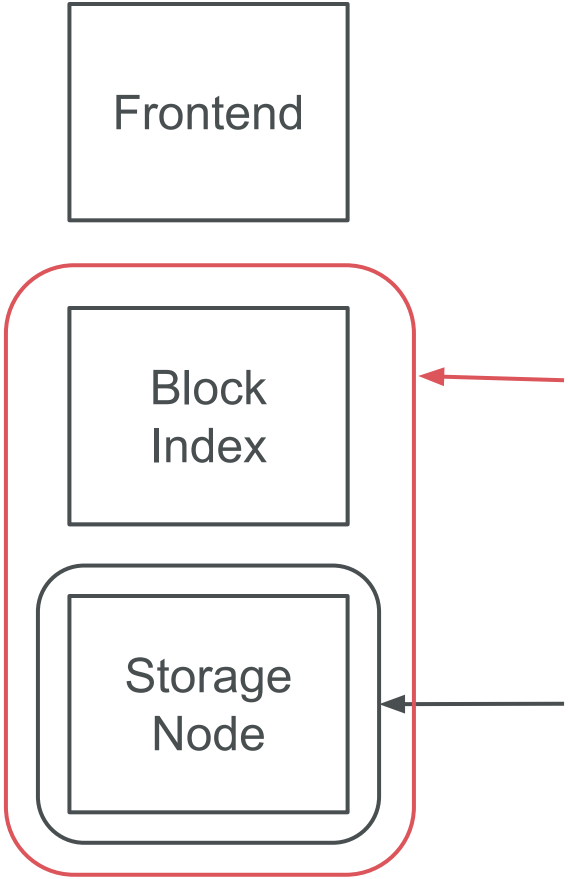
Frontend

Block Index

Storage Node

Scrubber

Can we actually get this off disk?

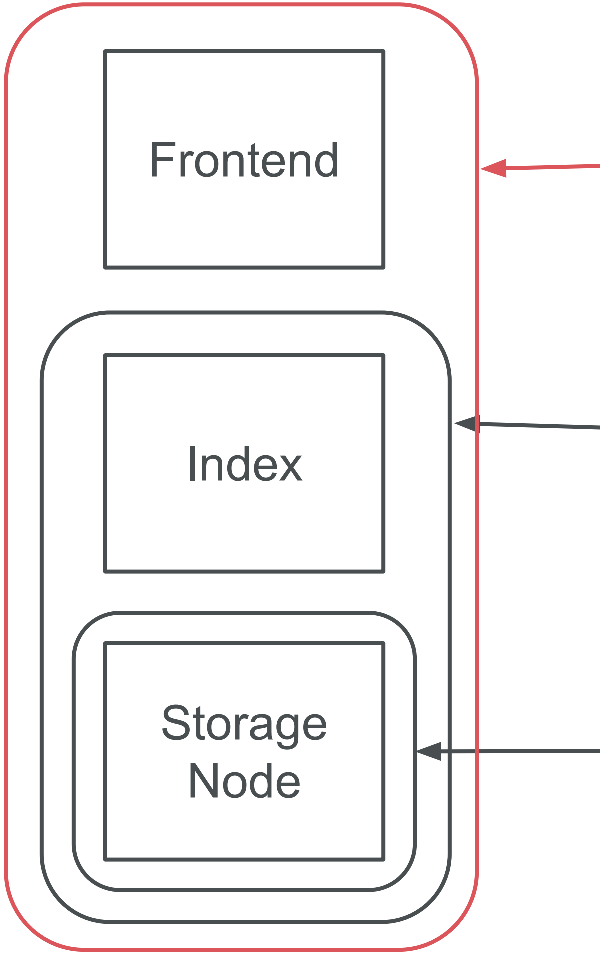


Scanner

Do we have what we think we have?

Scrubber

Can we actually get this off disk?



Watcher

Can we serve the data?

Scanner

Do we have what we think we have?

Scrubber

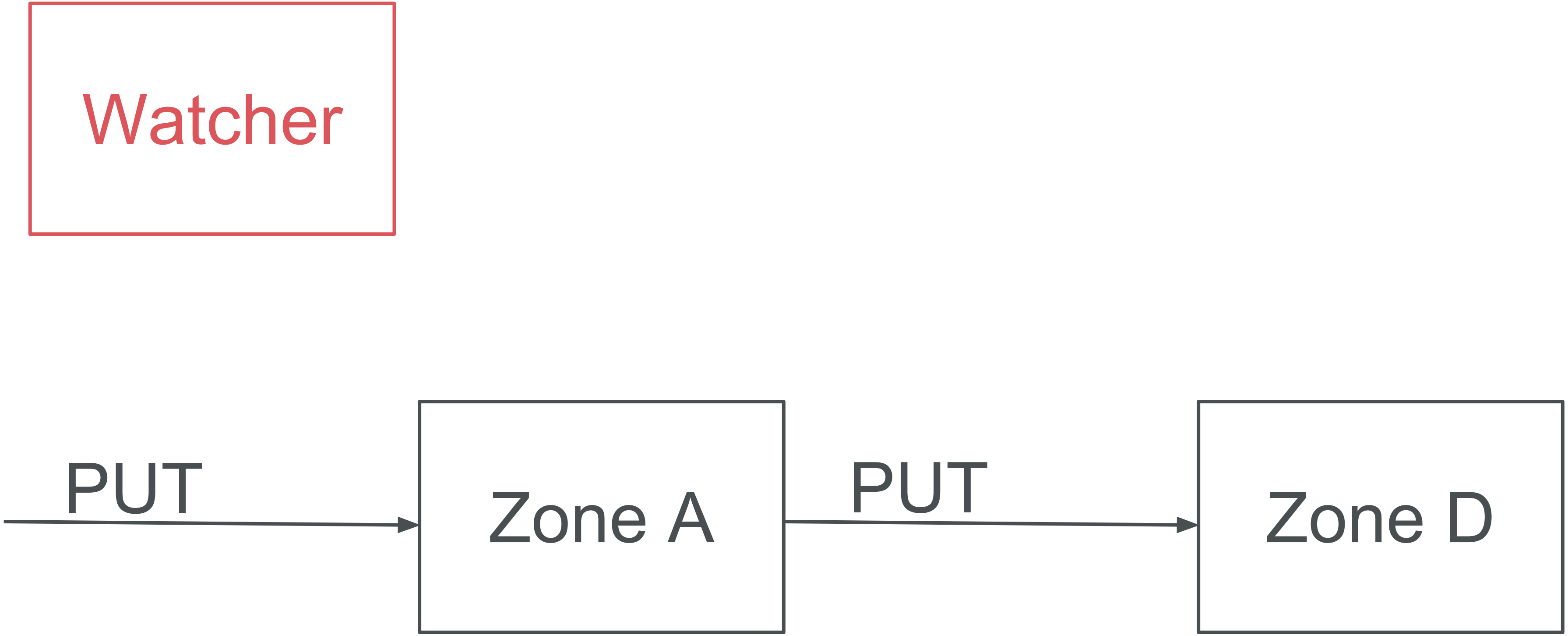
Can we actually get this off disk?

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS
1114	UUUUUUUU

Replication Queue (MySQL)

Index	Key to replicate
1111	RRRRRRR
1112	SSSSSSSS
1113	TTTTTTTTT
1114	UUUUUUUU



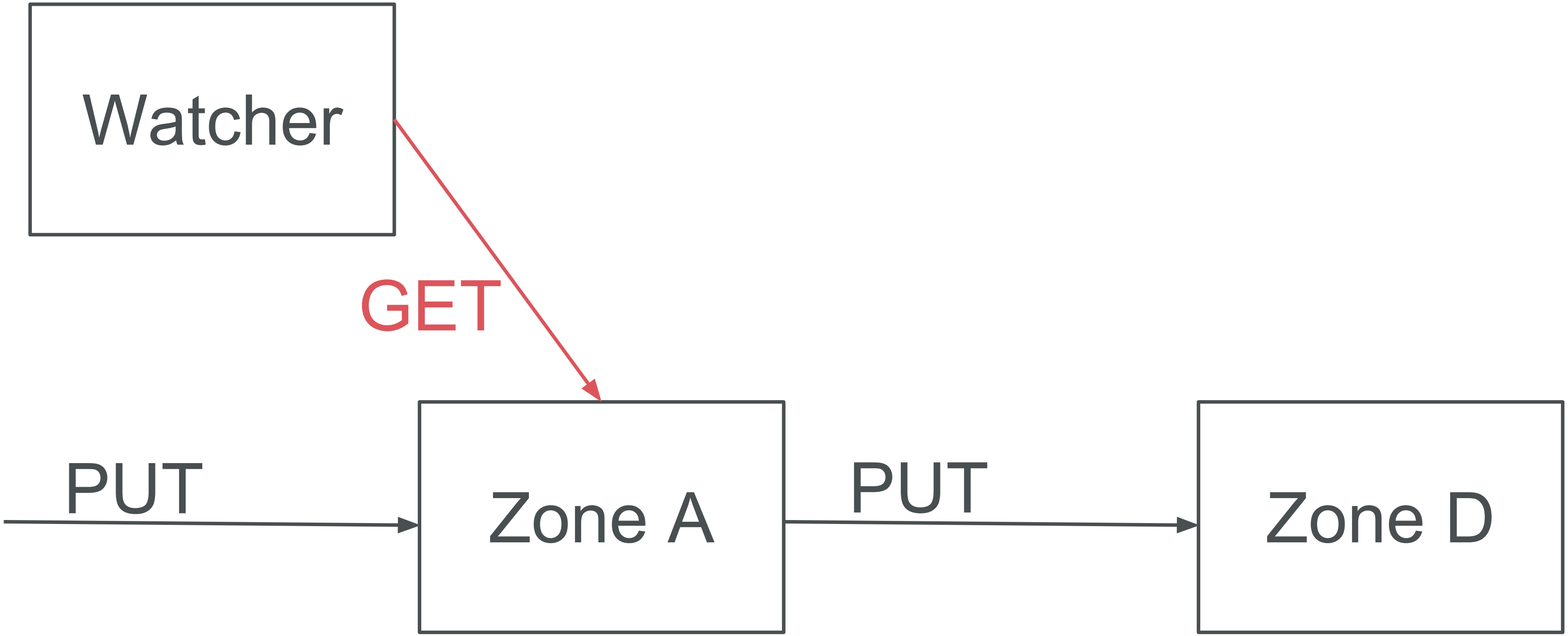
Watcher

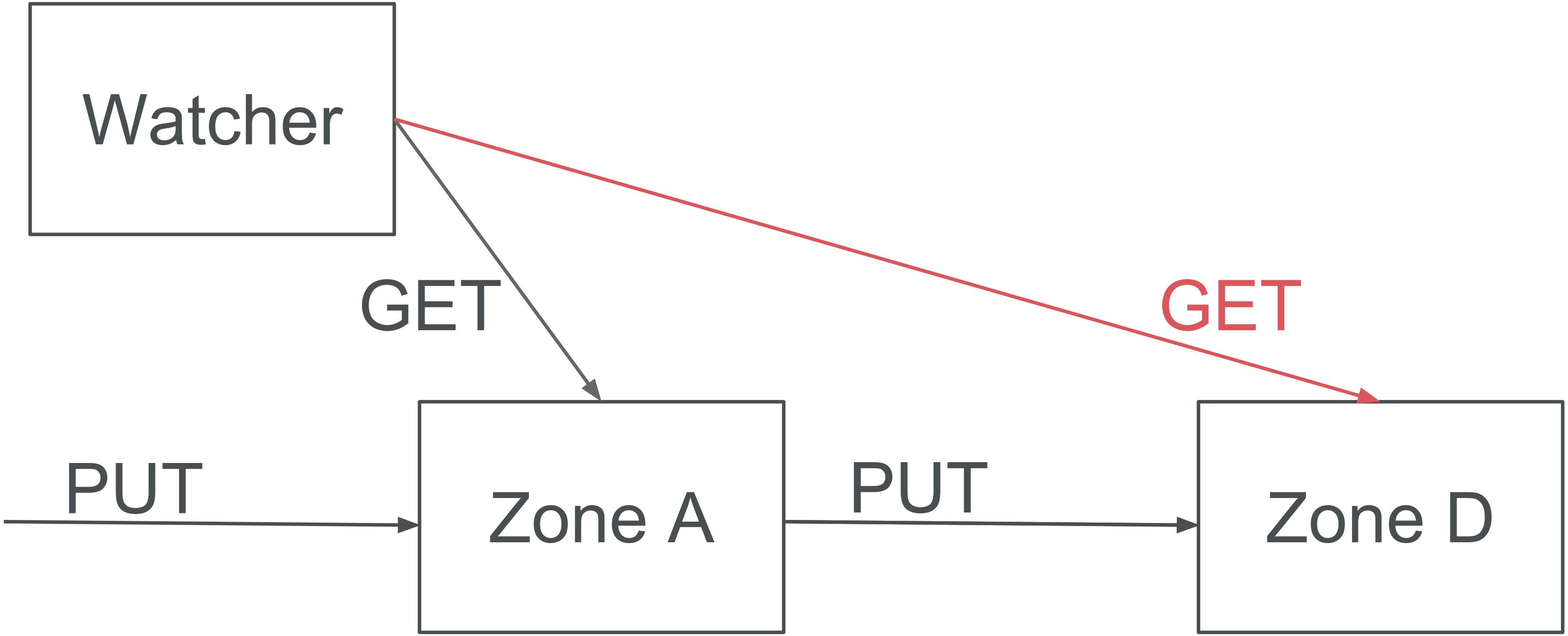
PUT

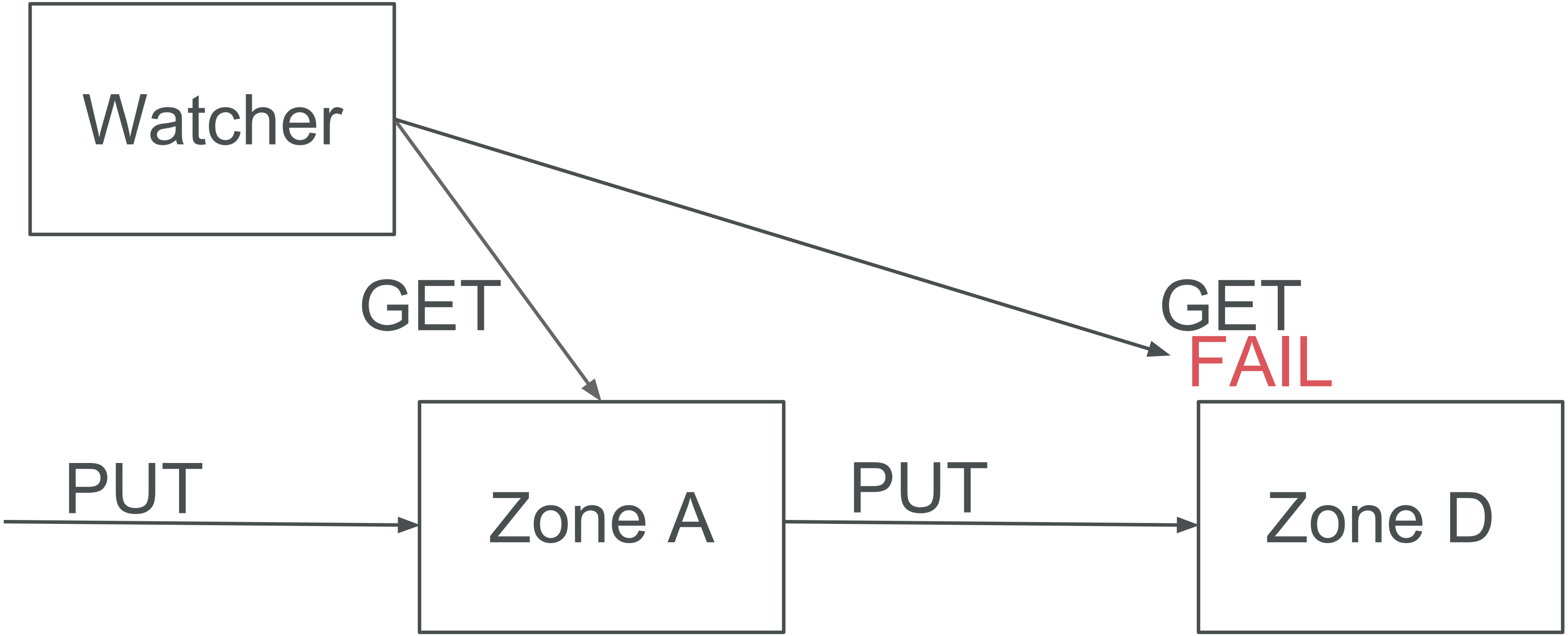
Zone A

PUT

Zone D

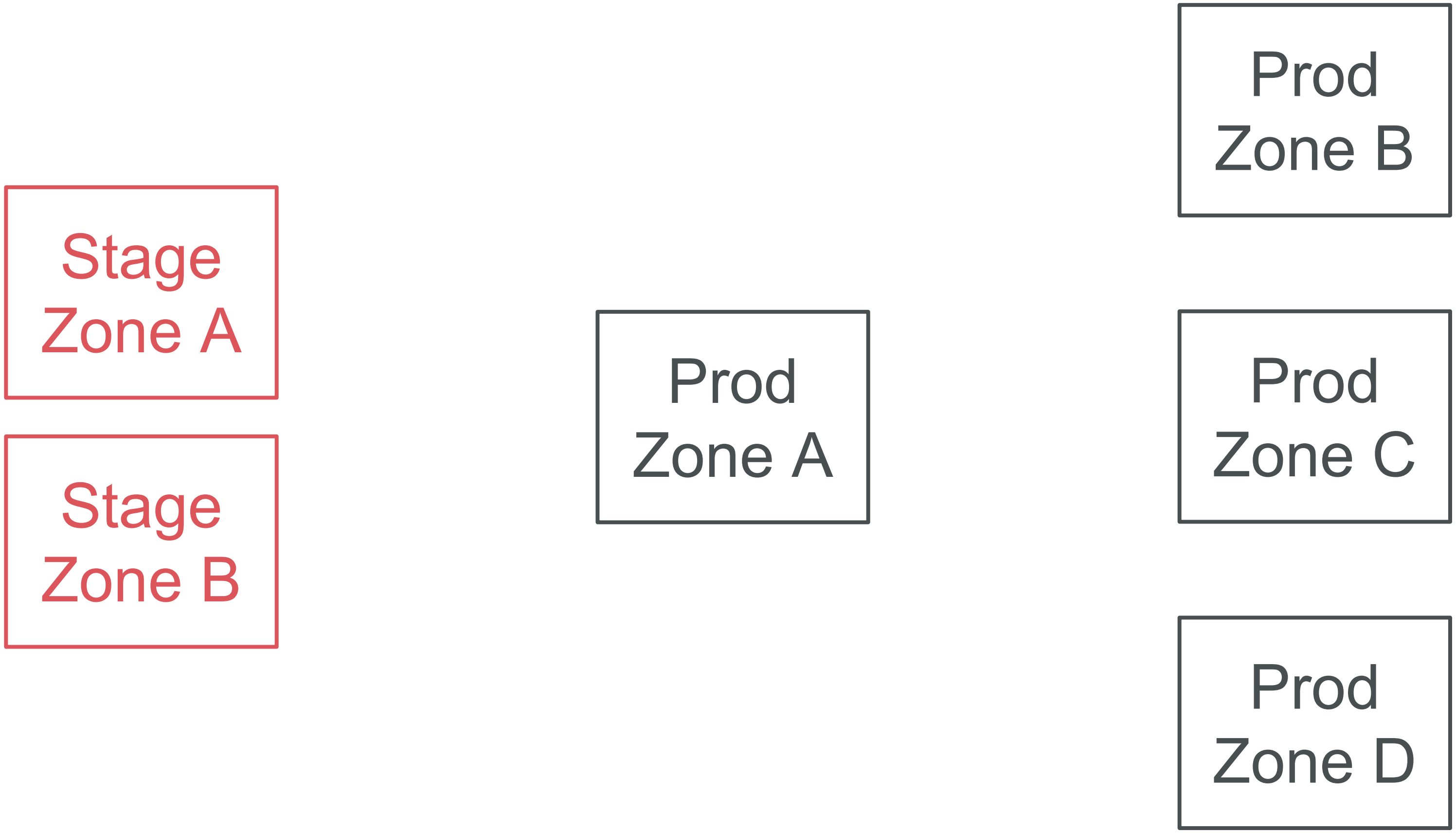

Watcher
GET
PUT
Zone A
PUT
Zone D


Watcher
GET
GET
PUT
Zone A
PUT
Zone D


Watcher
GET
GET
FAIL
PUT
Zone A
PUT
Zone D

Software bugs? Protect!

Gradual Release Process



Stage Zone A

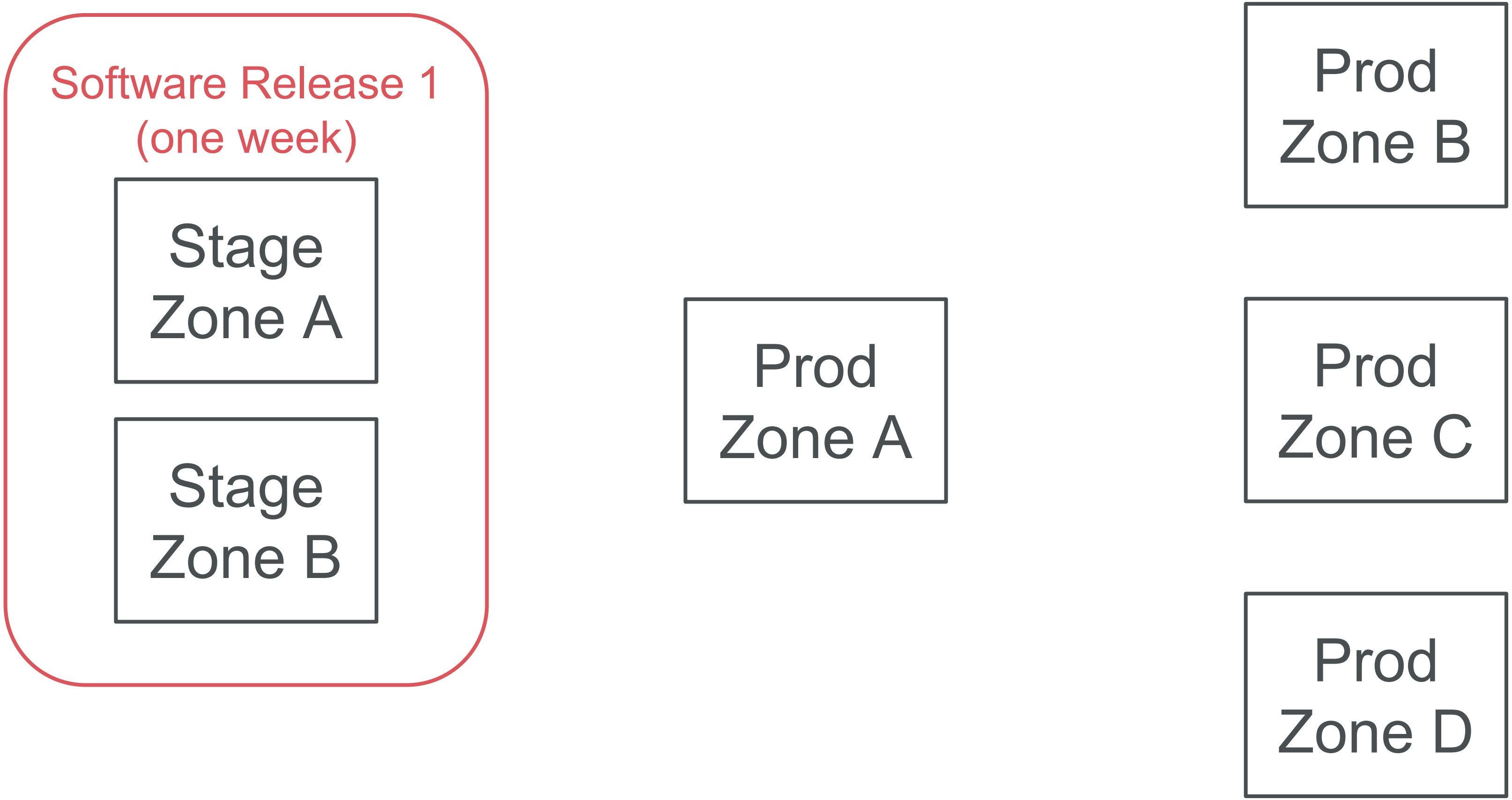
Stage Zone B

Prod Zone A

Prod Zone B

Prod Zone C

Prod Zone D



Software Release 1
(one week)

Stage
Zone A

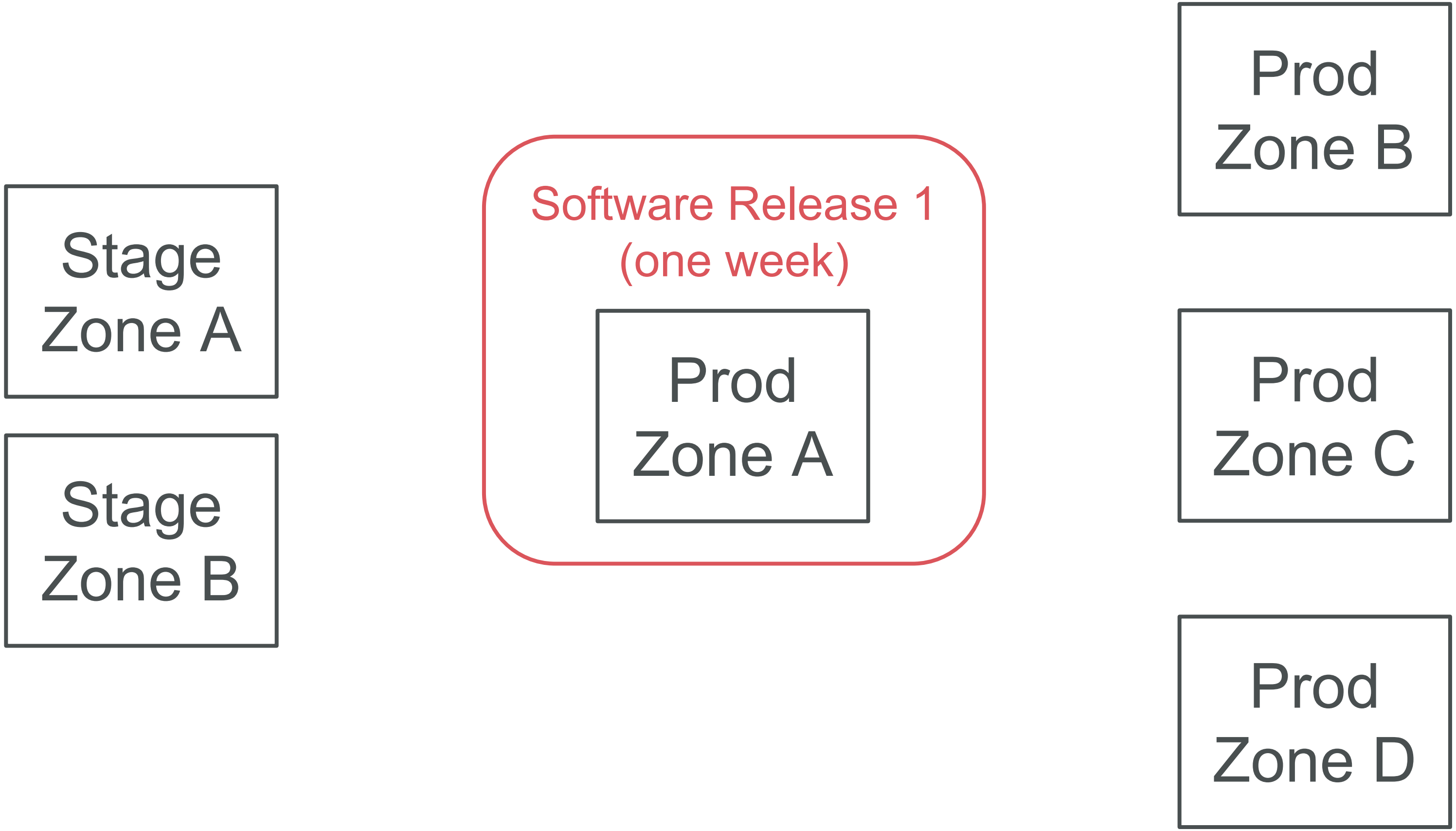
Stage
Zone B

Prod
Zone A

Prod
Zone B

Prod
Zone C

Prod
Zone D



Stage Zone A

Stage Zone B

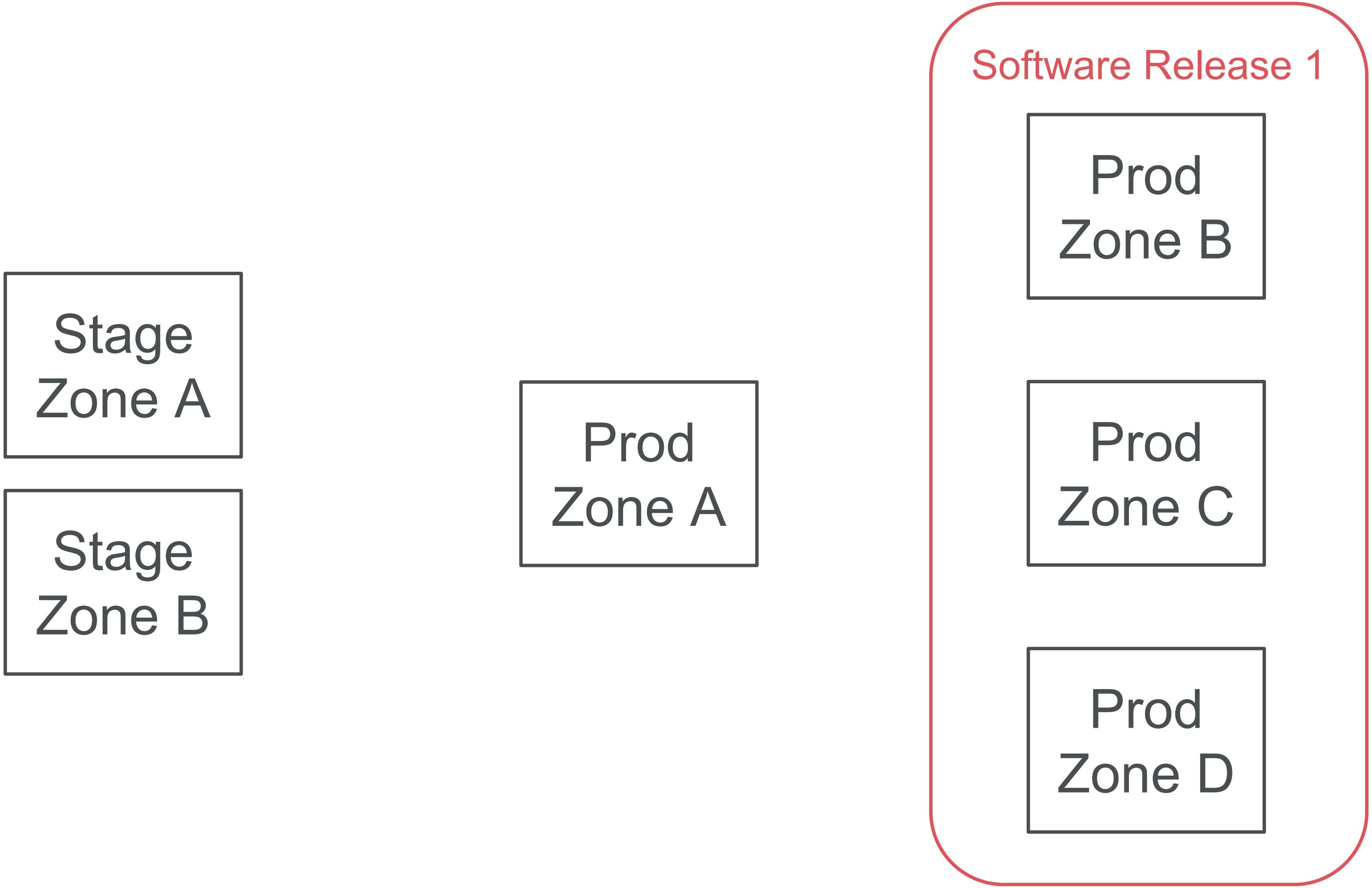
Software Release 1 (one week)

Prod Zone A

Prod Zone B

Prod Zone C

Prod Zone D



Software Release 1

Stage Zone A

Stage Zone B

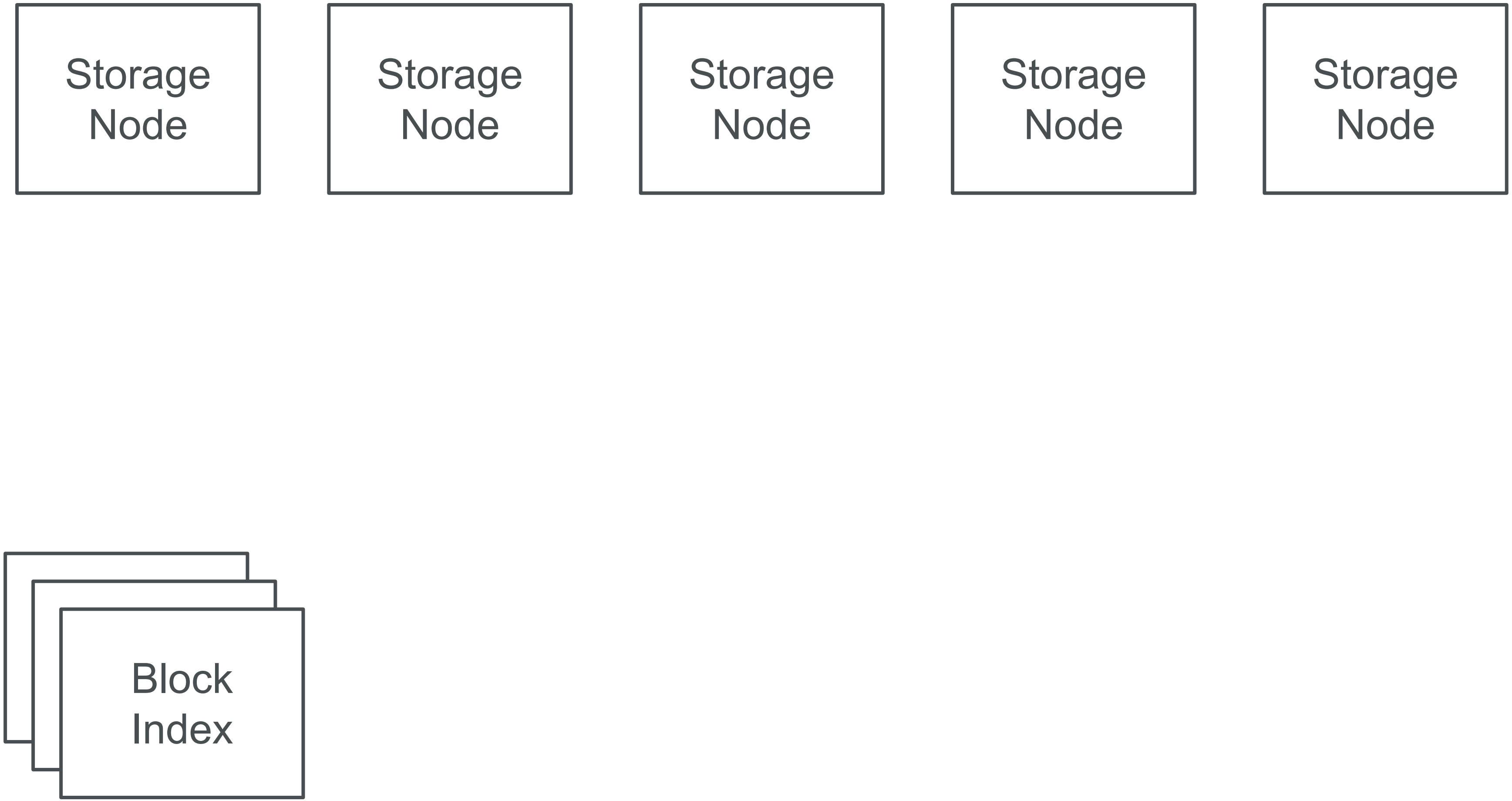
Prod Zone A

Prod Zone B

Prod Zone C

Prod Zone D

Story Time!



Storage Node

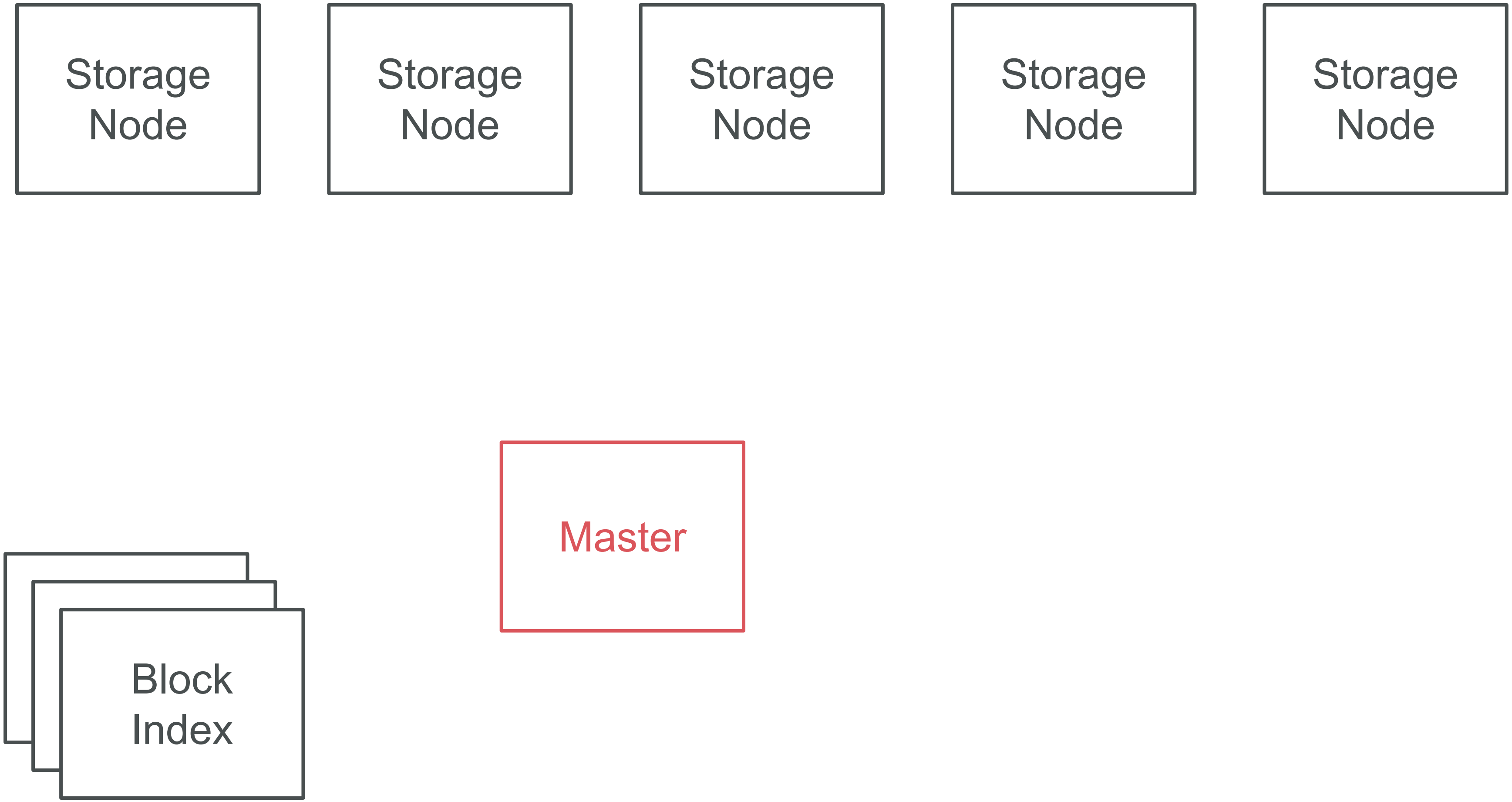
Storage Node

Storage Node

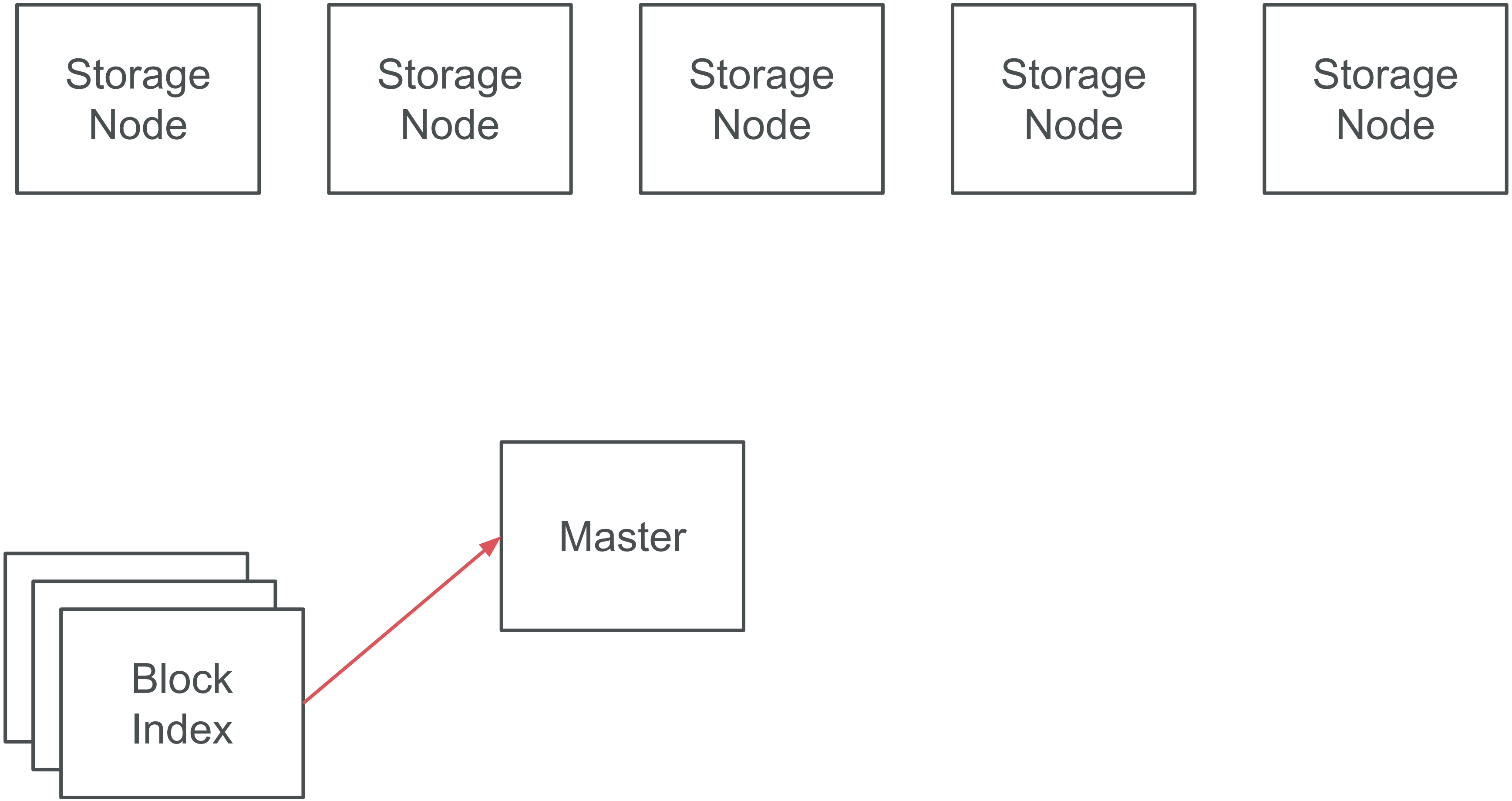
Storage Node

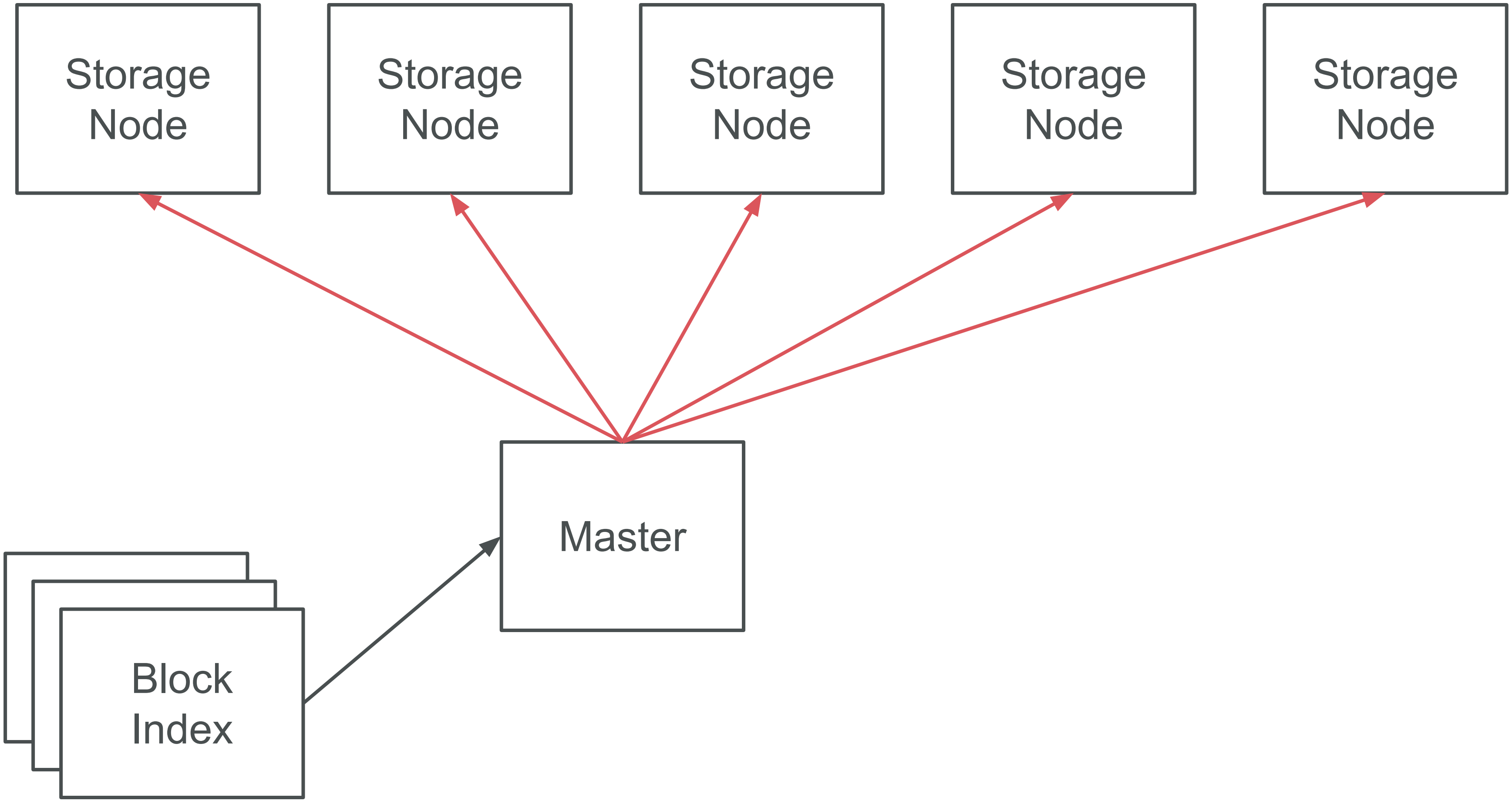
Storage Node

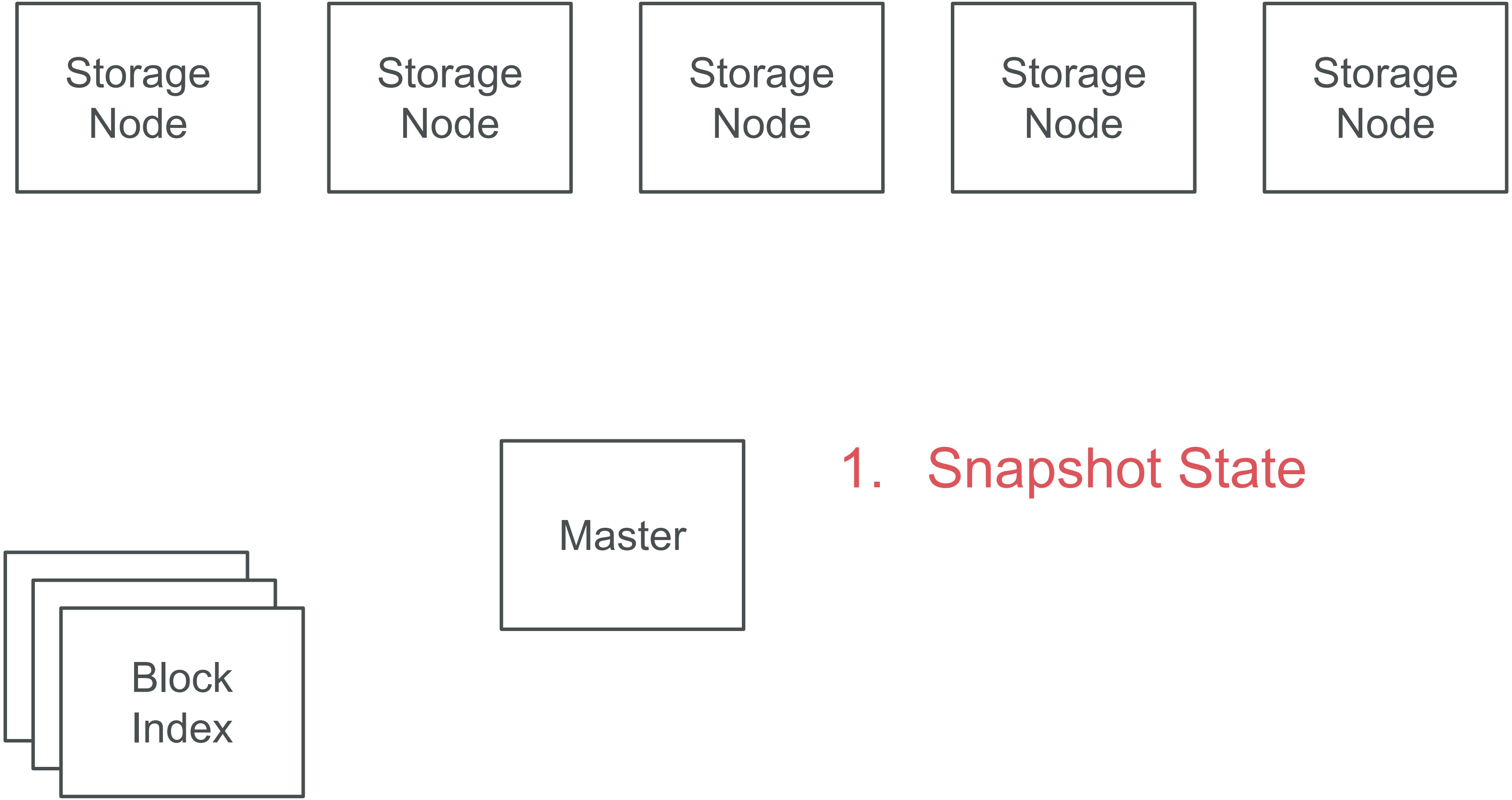
Block Index



Storage Node
Storage Node
Storage Node
Storage Node
Storage Node
Master
Block Index


Storage Node
Storage Node
Storage Node
Storage Node
Storage Node
Master
Block Index


Storage Node
Storage Node
Storage Node
Storage Node
Storage Node
Master
Block Index



Storage Node

Storage Node

Storage Node

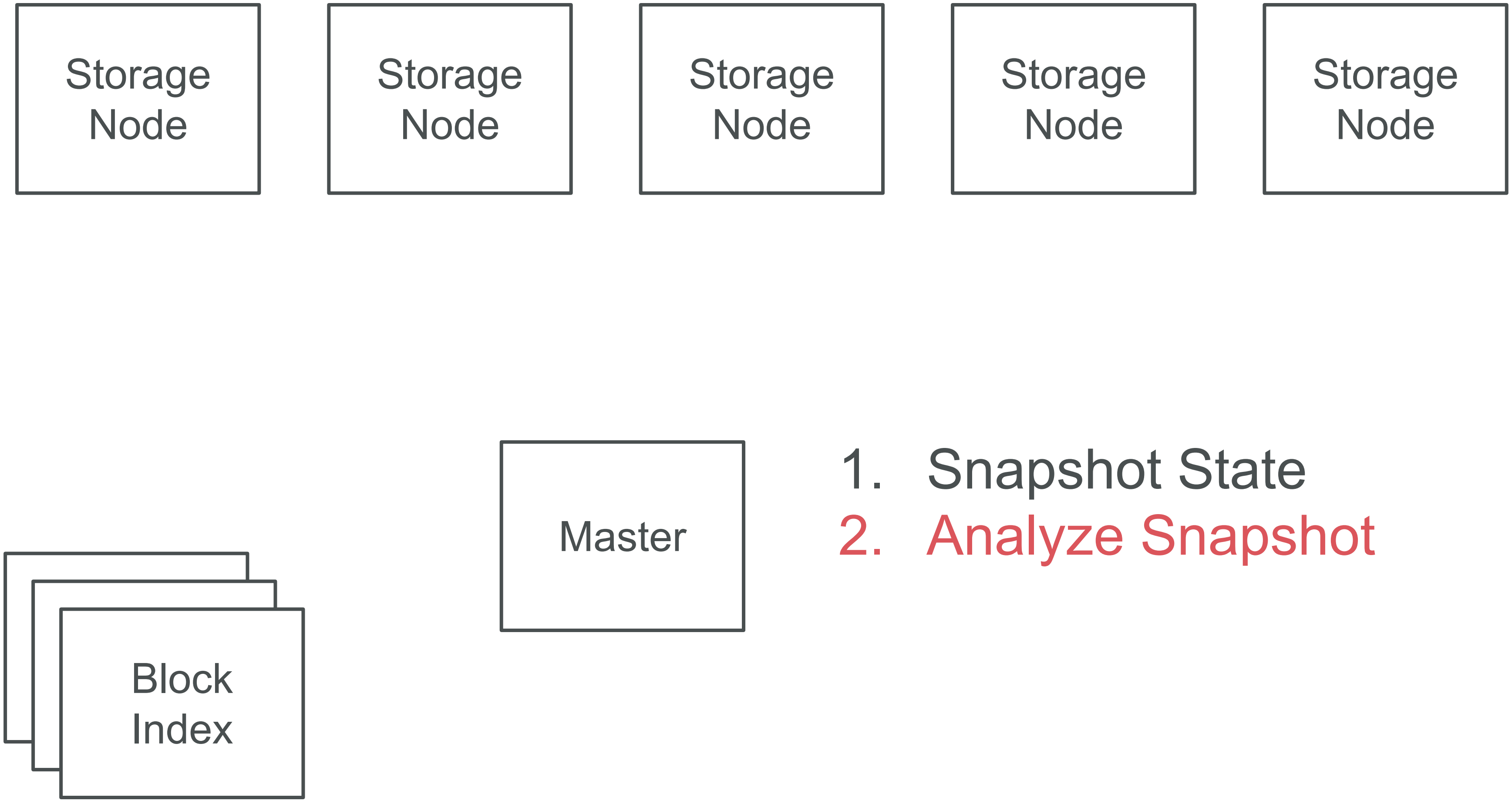
Storage Node

Storage Node

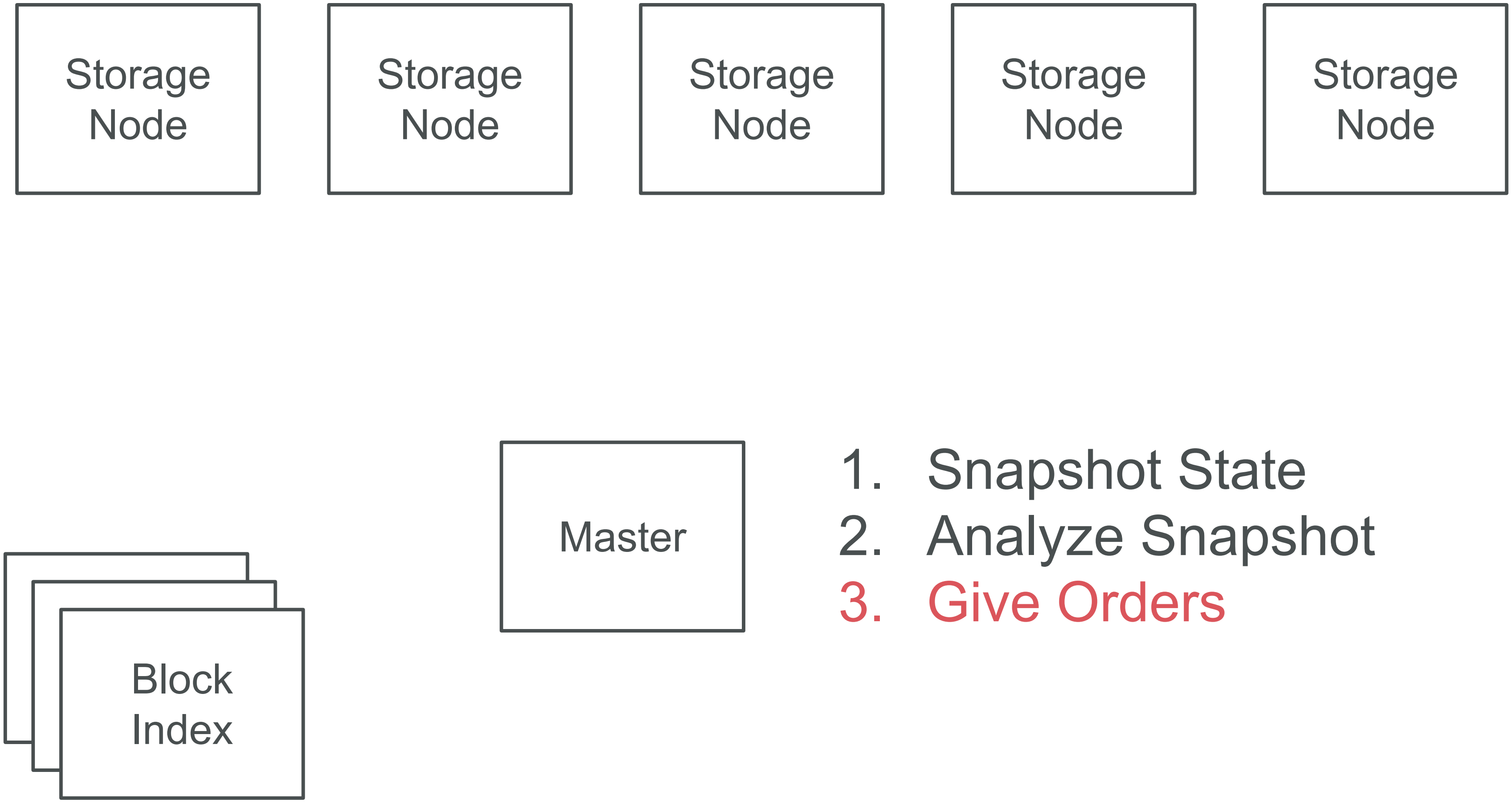
Master

Block Index

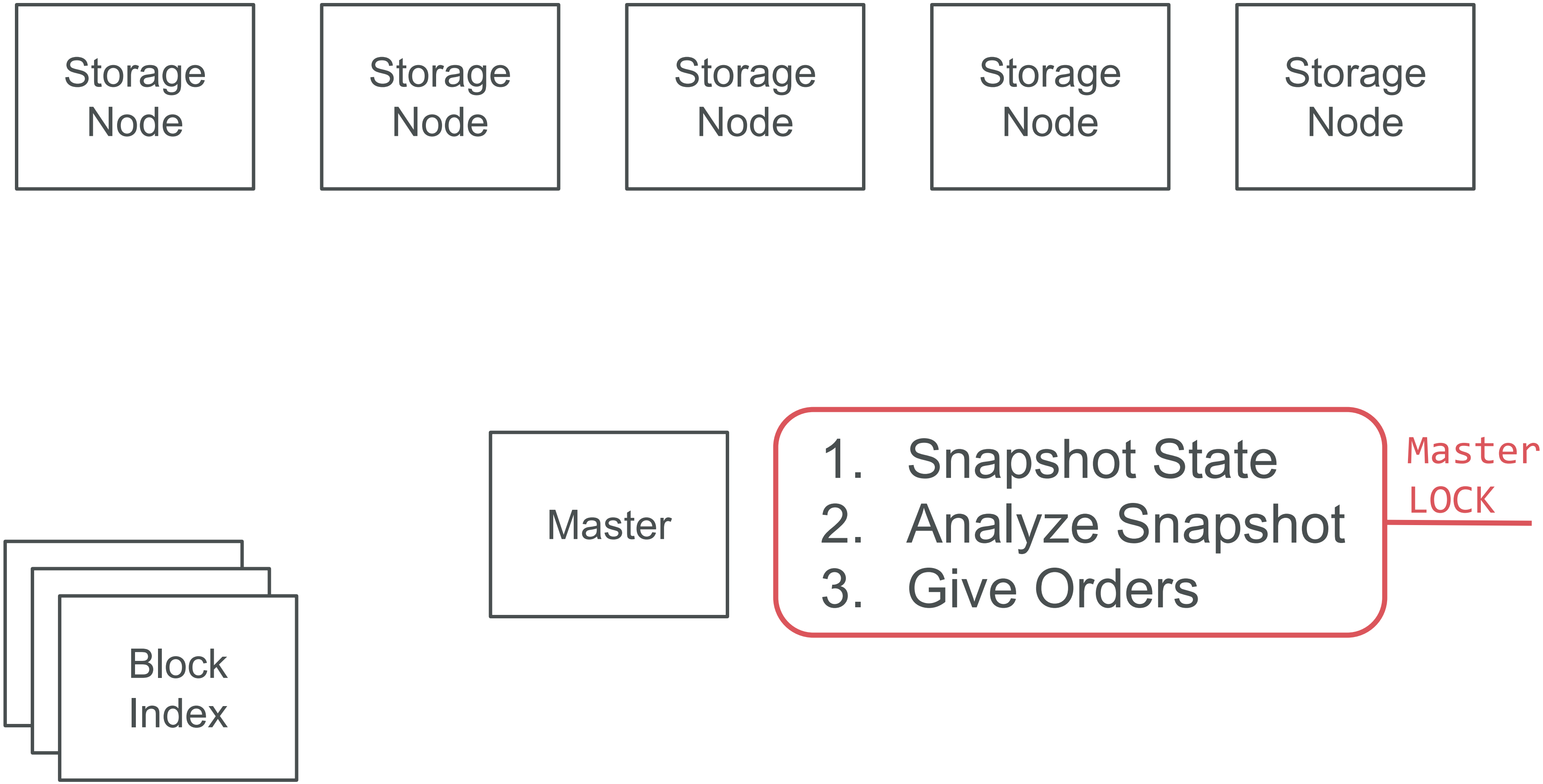
1. Snapshot State



1. Snapshot State
2. Analyze Snapshot



1. Snapshot State
2. Analyze Snapshot
3. Give Orders



Storage Node

Storage Node

Storage Node

Storage Node

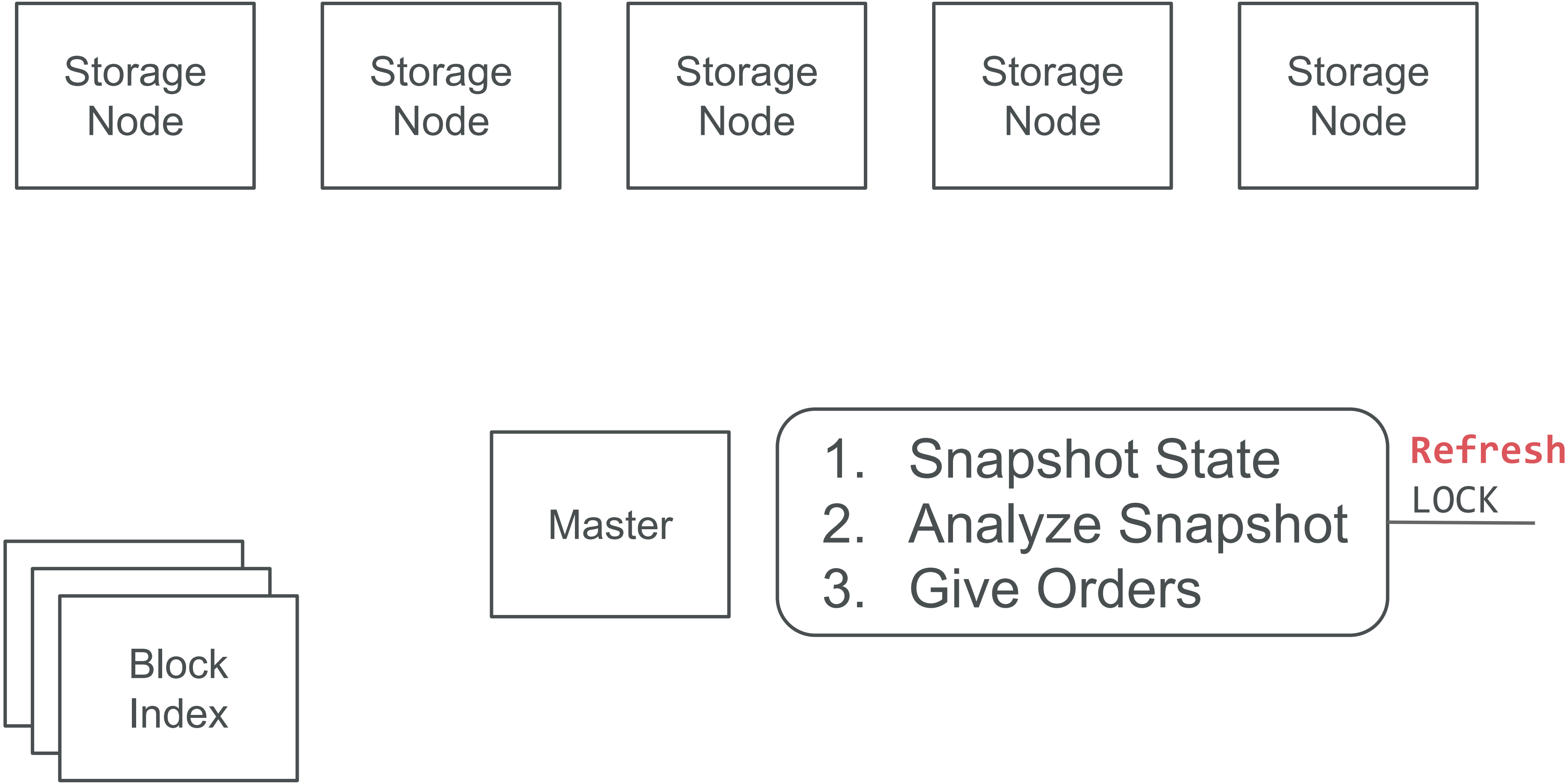
Storage Node

Master

1. Snapshot State
2. Analyze Snapshot
3. Give Orders

Master LOCK

Block Index



Storage Node

Storage Node

Storage Node

Storage Node

Storage Node

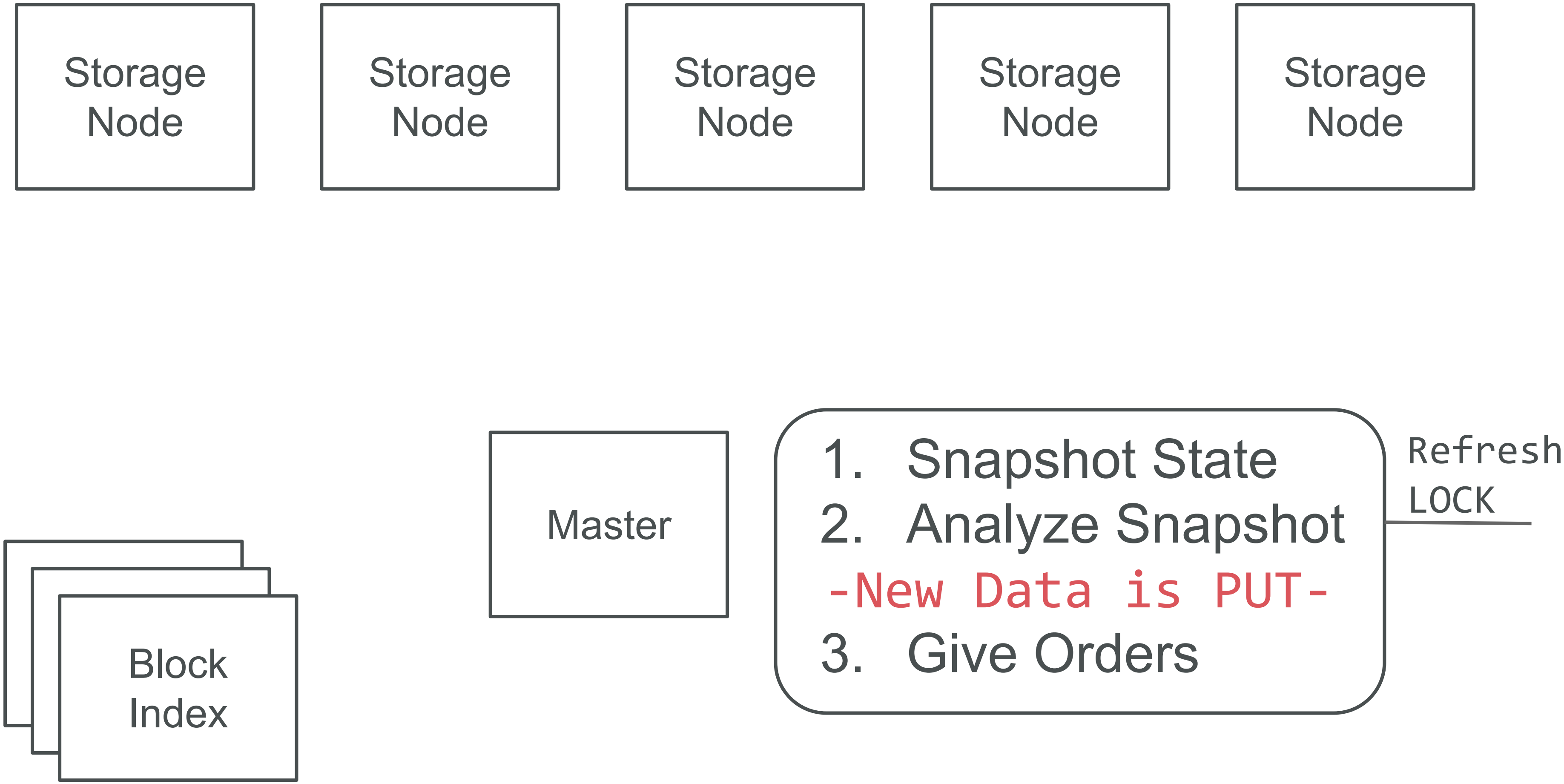
Master

1. Snapshot State
2. Analyze Snapshot
3. Give Orders

Refresh

LOCK

Block Index



Storage Node

Storage Node

Storage Node

Storage Node

Storage Node

Master

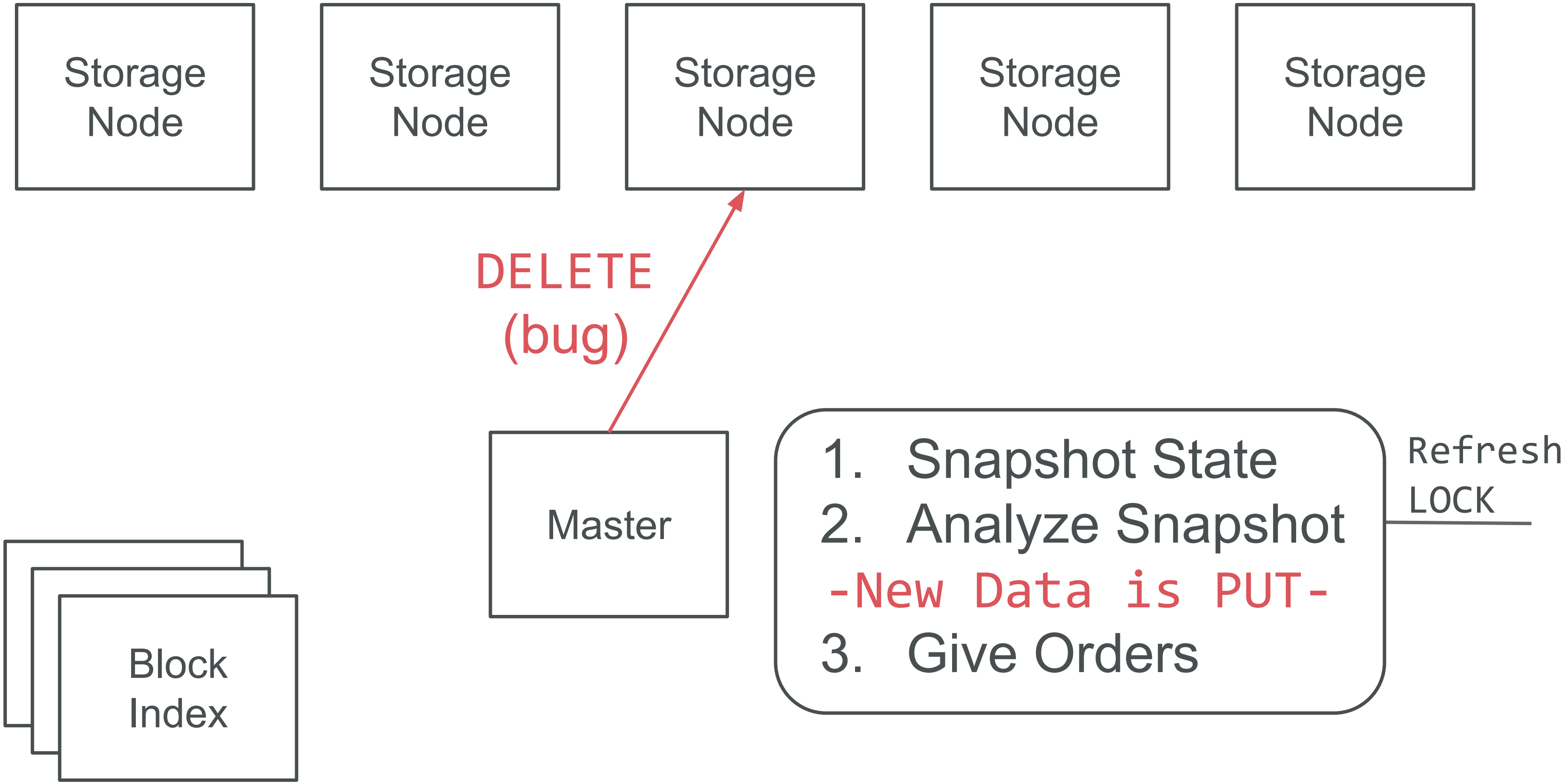
Block Index

1. Snapshot State
2. Analyze Snapshot

-New Data is PUT-

3. Give Orders

Refresh
LOCK



Storage Node

Storage Node

Storage Node

Storage Node

Storage Node

DELETE
(bug)

Master

Block Index

1. Snapshot State
2. Analyze Snapshot

-New Data is PUT-

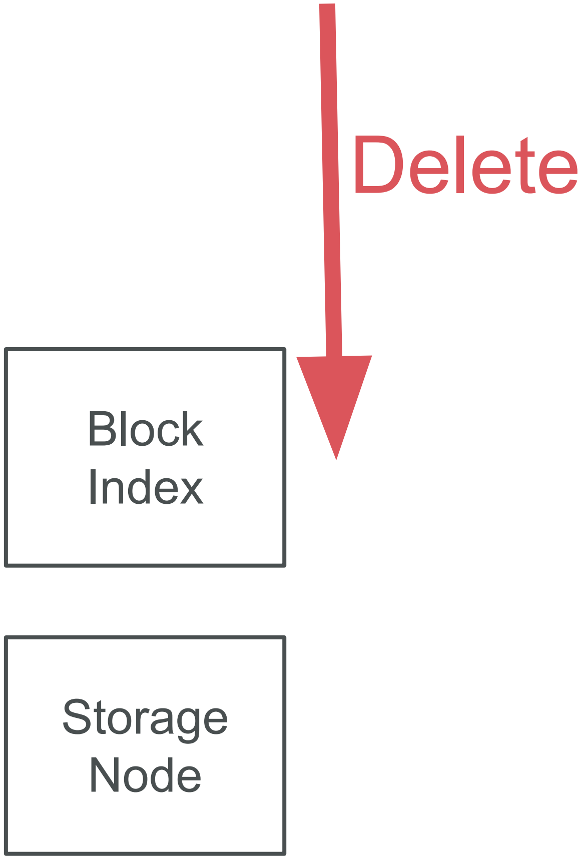
3. Give Orders

Refresh
LOCK

Purgatory and Trash

Block Index

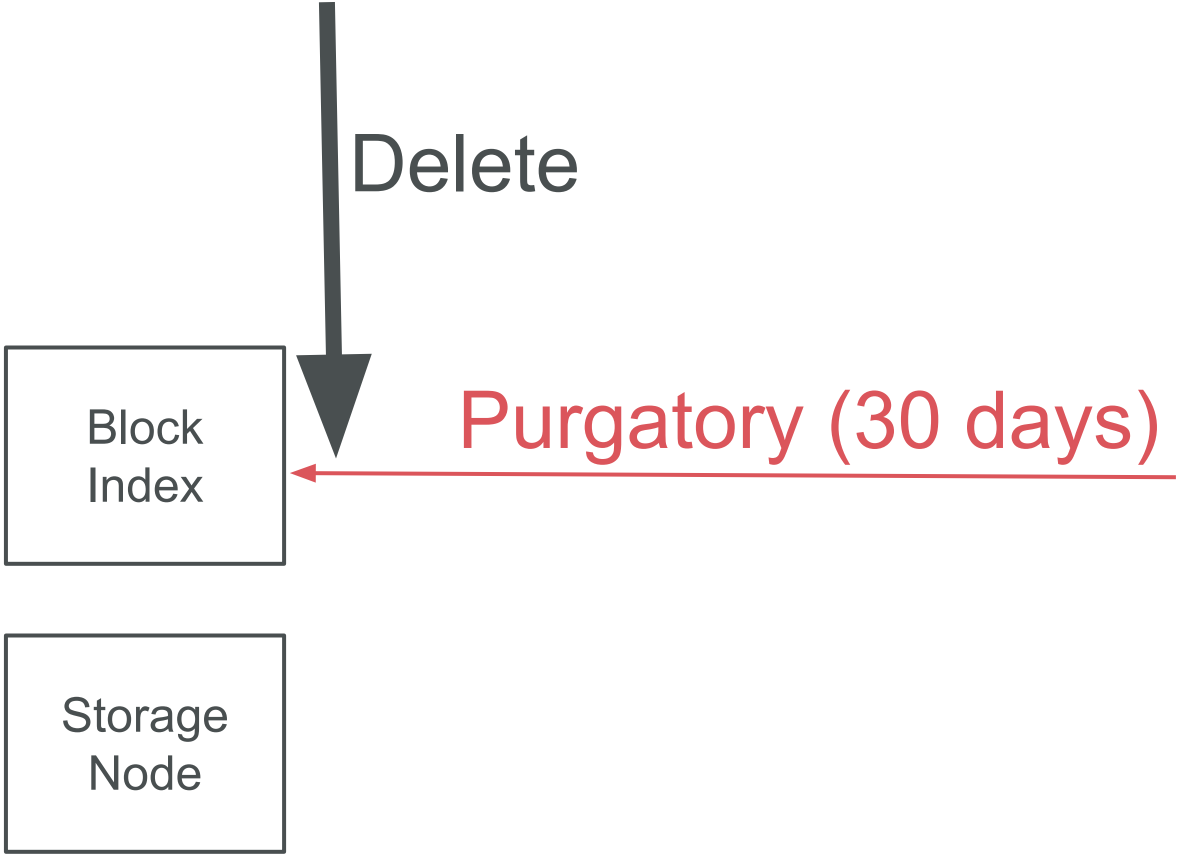
Storage Node



Delete

Block Index

Storage Node

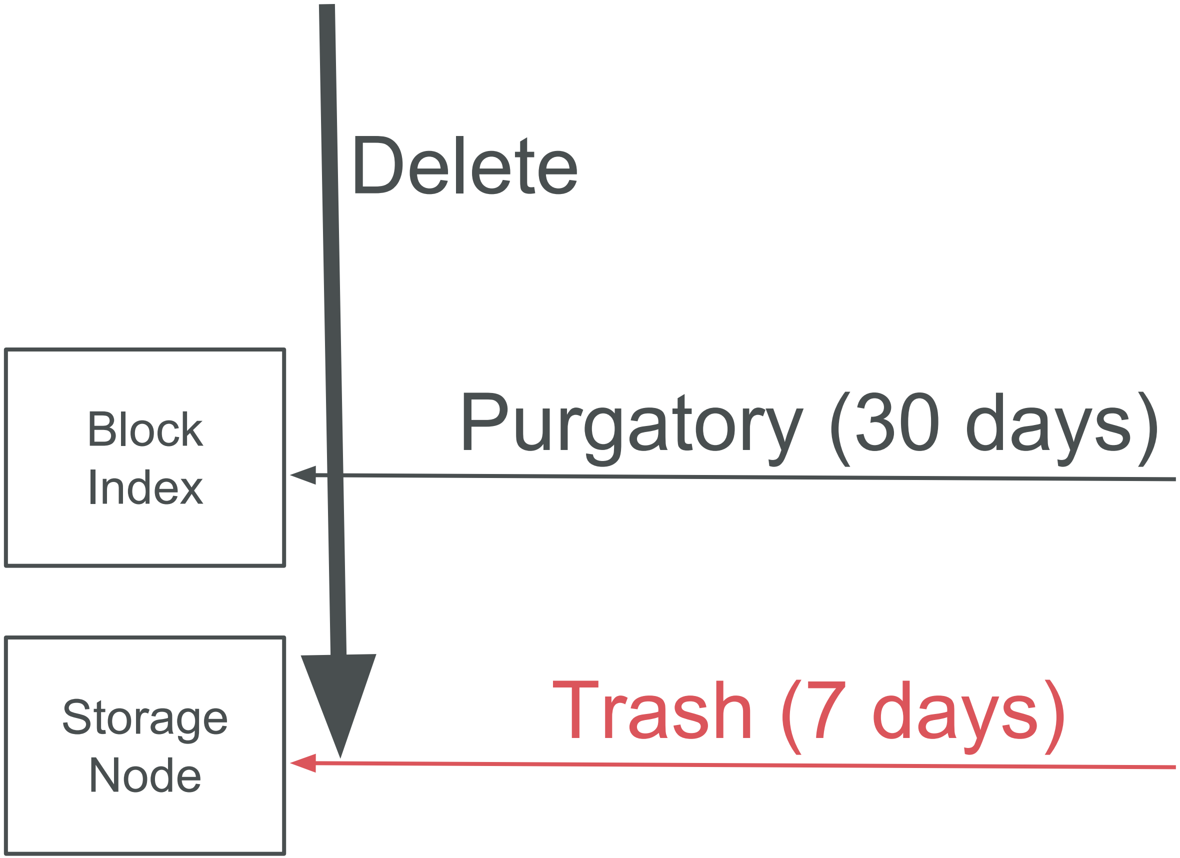


Delete

Block Index

Purgatory (30 days)

Storage Node



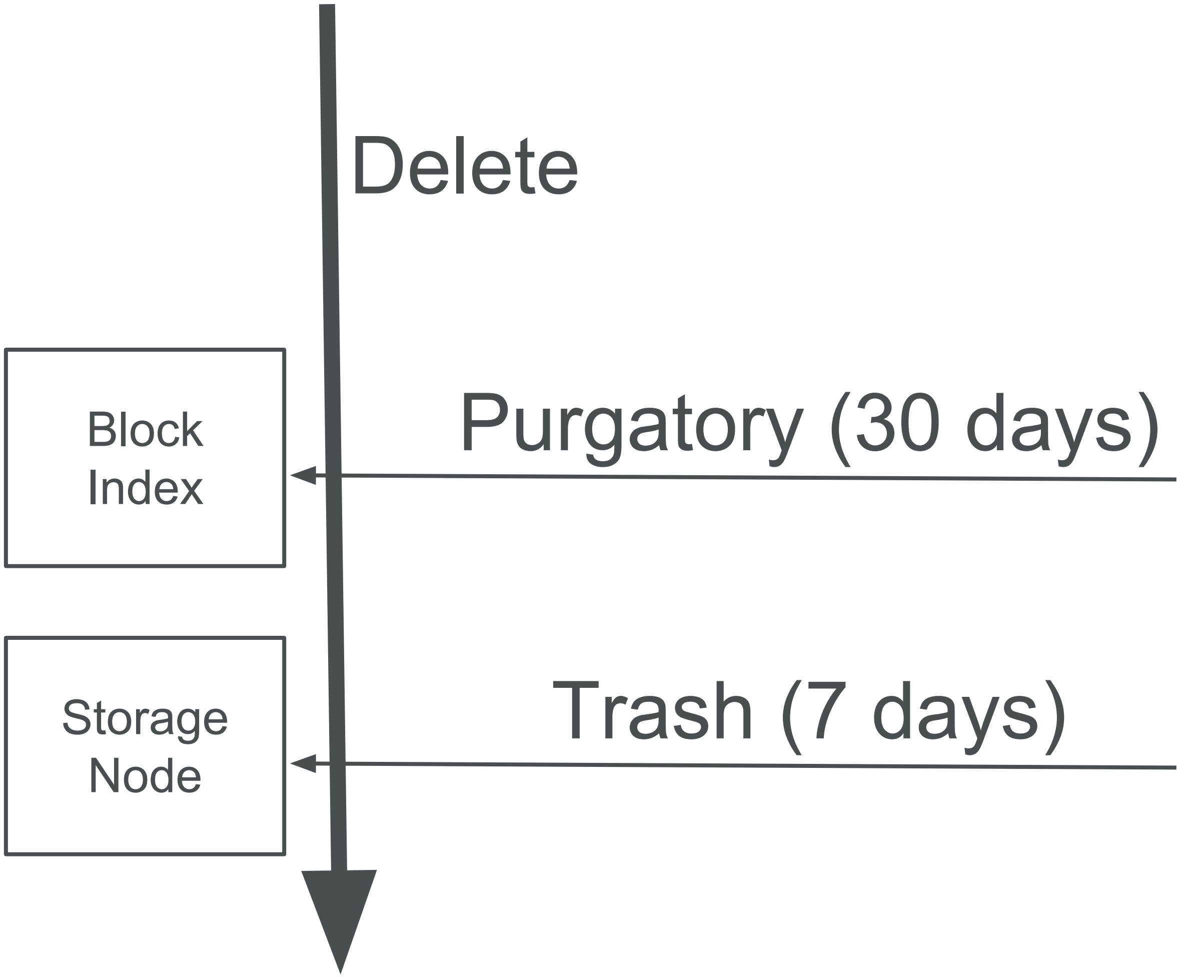
Delete

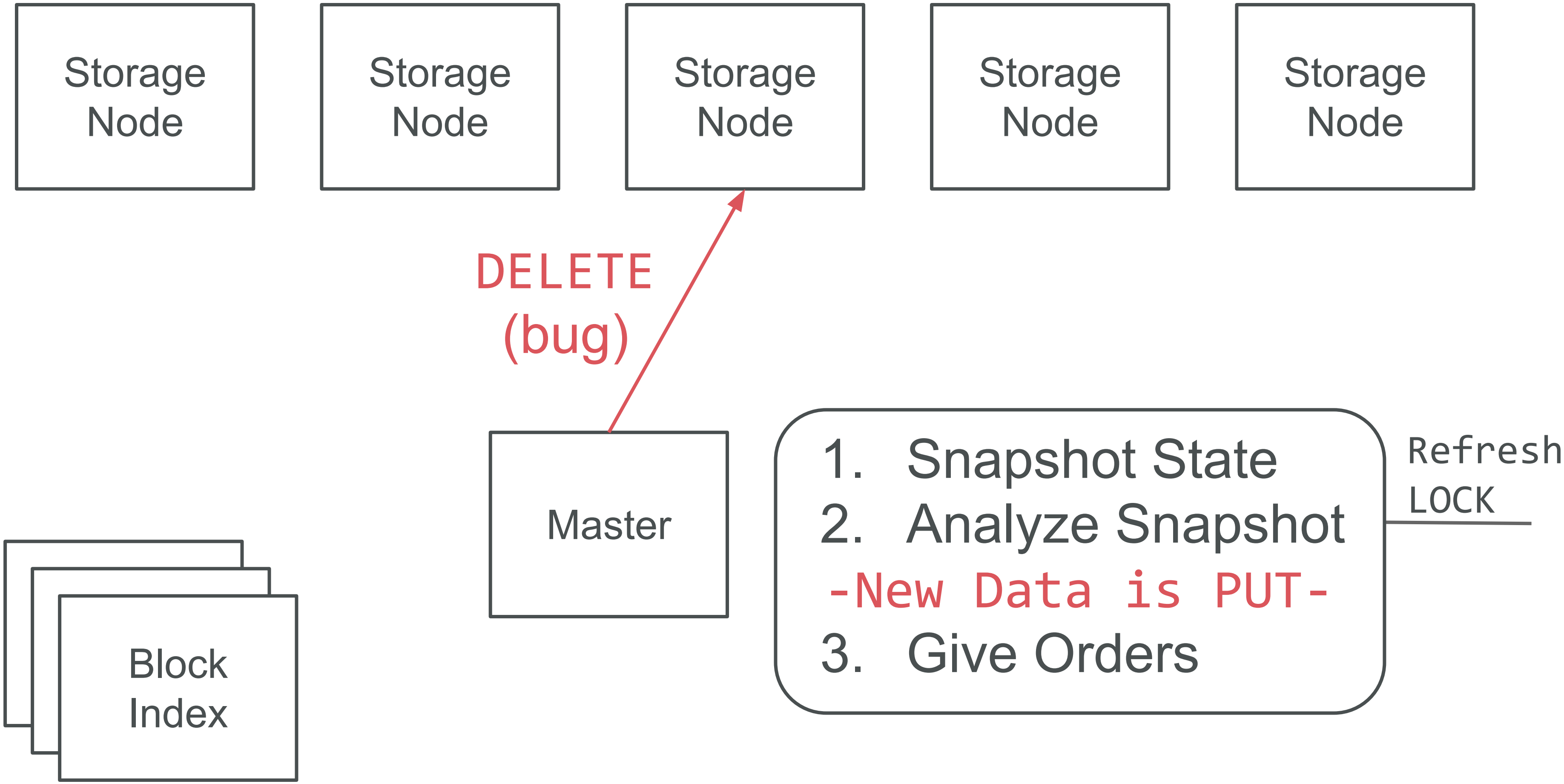
Block Index

Purgatory (30 days)

Storage Node

Trash (7 days)


Delete
Block Index
Purgatory (30 days)
Storage Node
Trash (7 days)



Storage Node

Storage Node

Storage Node

Storage Node

Storage Node

DELETE (bug)

Master

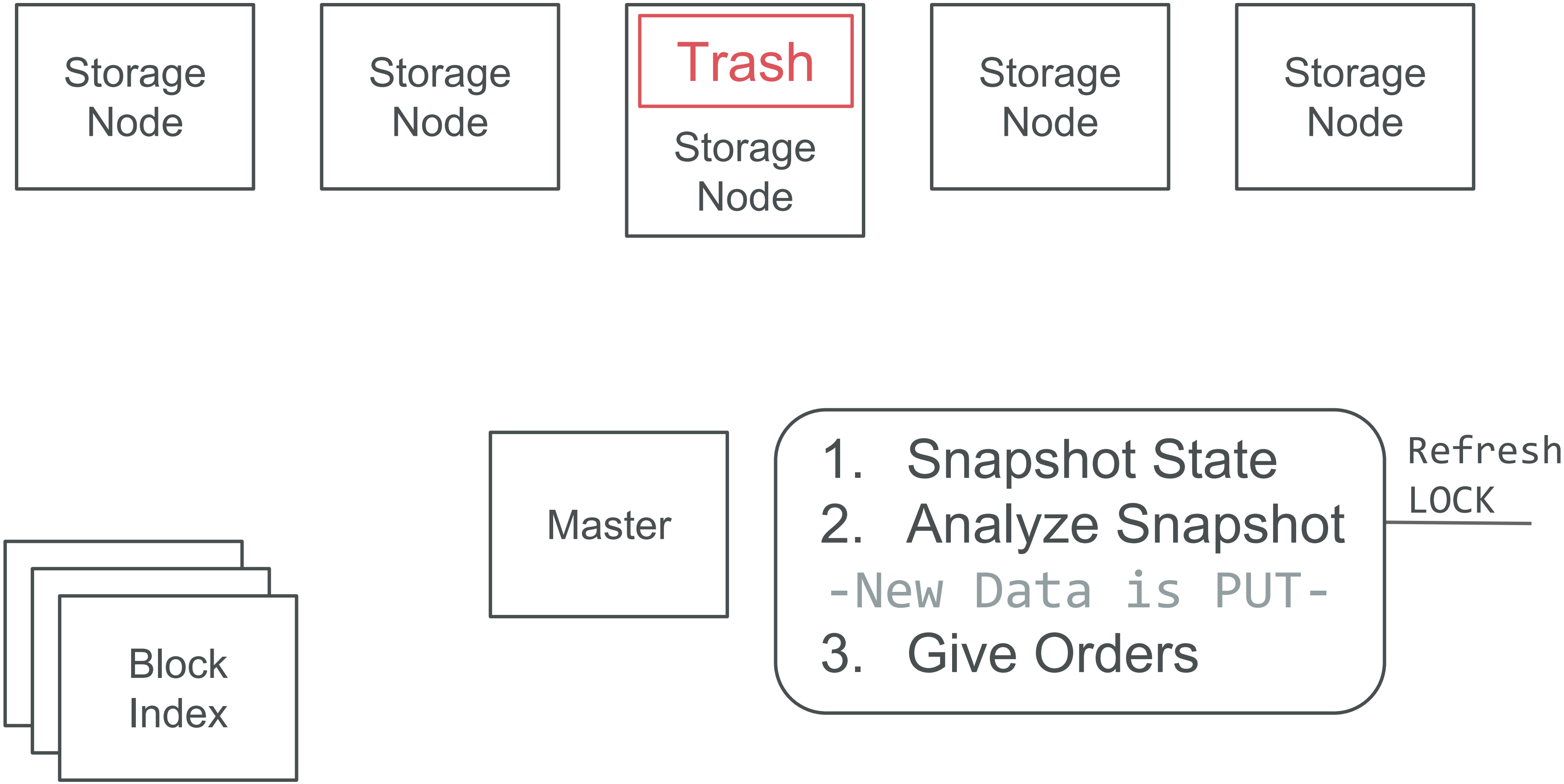
1. Snapshot State
2. Analyze Snapshot

-New Data is PUT-

3. Give Orders

Refresh LOCK

Block Index



Storage Node

Storage Node

Trash

Storage Node

Storage Node

Storage Node

Block Index

Master

1. Snapshot State
2. Analyze Snapshot

-New Data is PUT-

3. Give Orders

Refresh

LOCK

Software Bugs?

Verify!

Tests

Checkers

Protect!

Gradual Release

Purgatory+Trash

Biggest Fears

Software

Hardware

Humans

Tooling/Automation

Protections

Fear of Hardware

It will crumble underneath me

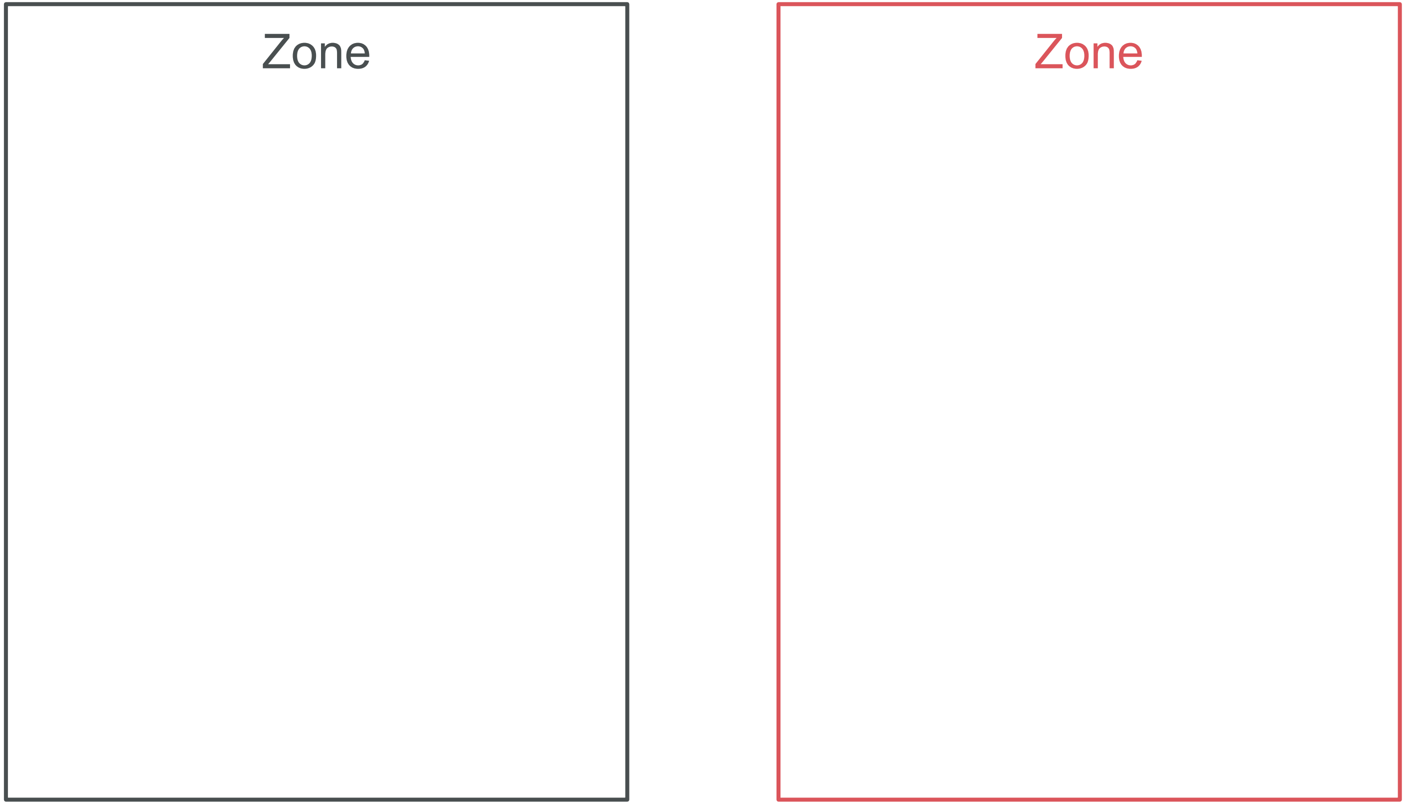
Hardware Failure? Verify!

Pre Production Qualification

Production
Machine Checkers

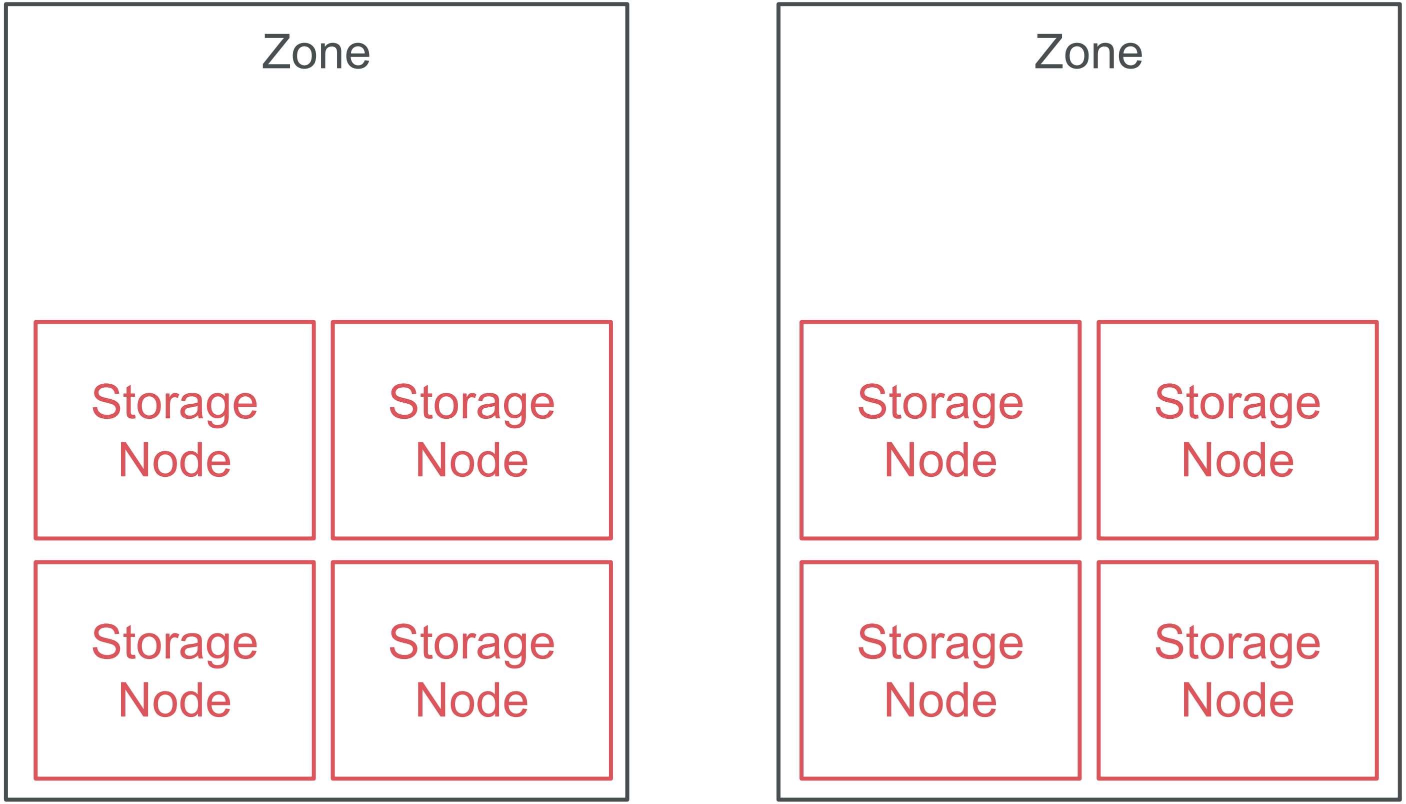
Hardware Failure? Protect!

Redundancy



Zone

Zone



Zone
Storage Node
Storage Node
Storage Node
Storage Node
Zone
Storage Node
Storage Node
Storage Node
Storage Node

Hardware Failure?

Verify!

Pre Production Qualification
In Production Checks

Protect!

Redundancy

Biggest Fears

Software

Hardware

Humans

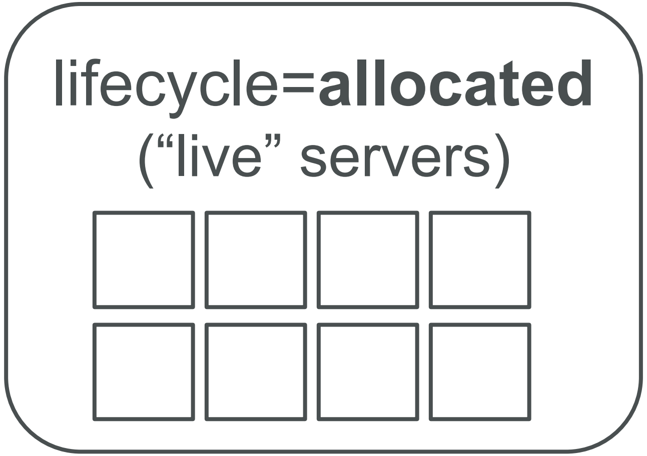
Tooling/Automation

Protections

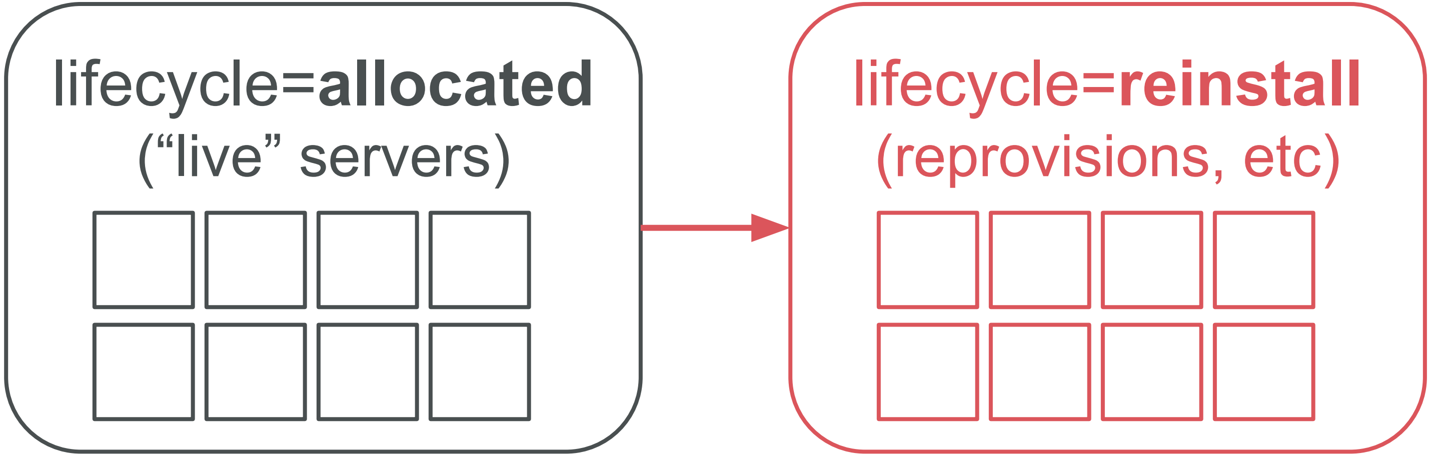
Fear of Humans

Accidents

Story Time!

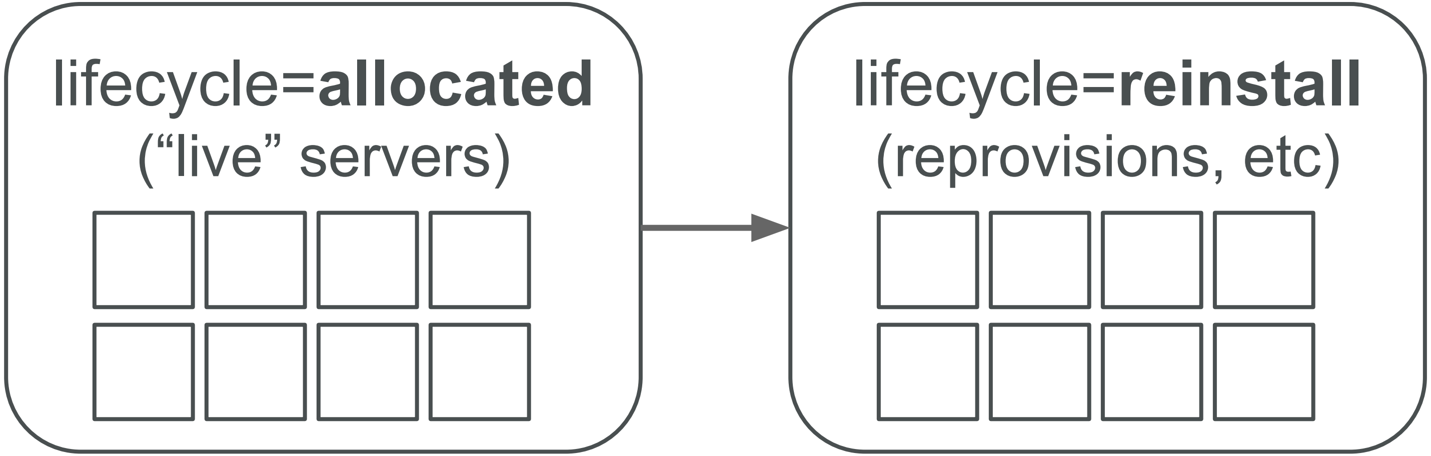


lifecycle=**allocated**
("live" servers)



lifecycle=**allocated**
(“live” servers)

lifecycle=**reinstall**
(reprovisions, etc)

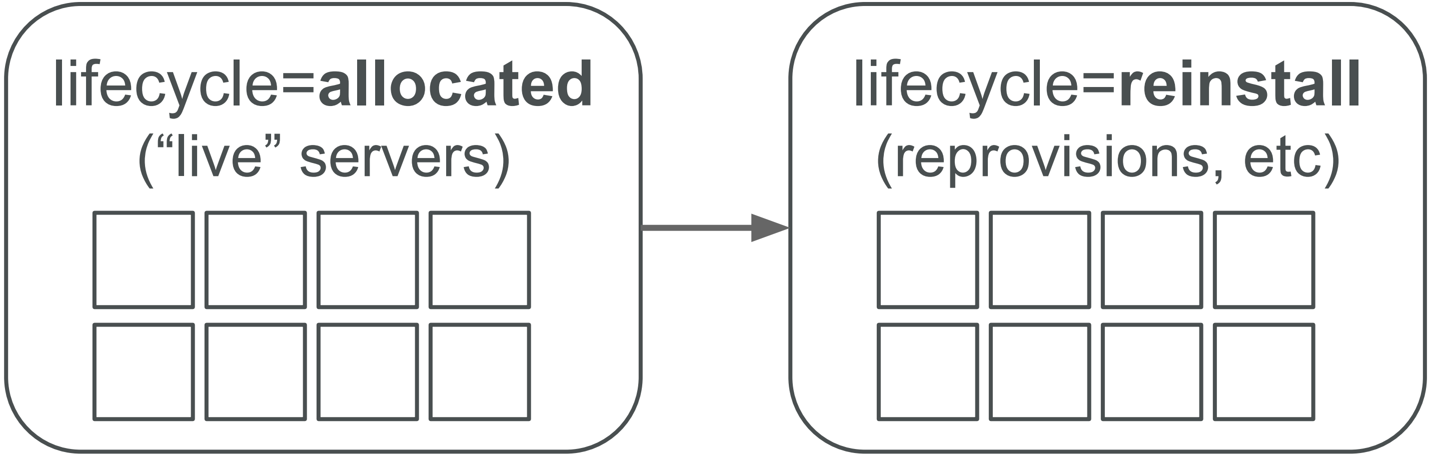


lifecycle=**allocated**
(“live” servers)

lifecycle=**reinstall**
(reprovisions, etc)

To rush upgrades:

```
gsh -q cache lifecycle=reinstall upgrade-host.sh
```



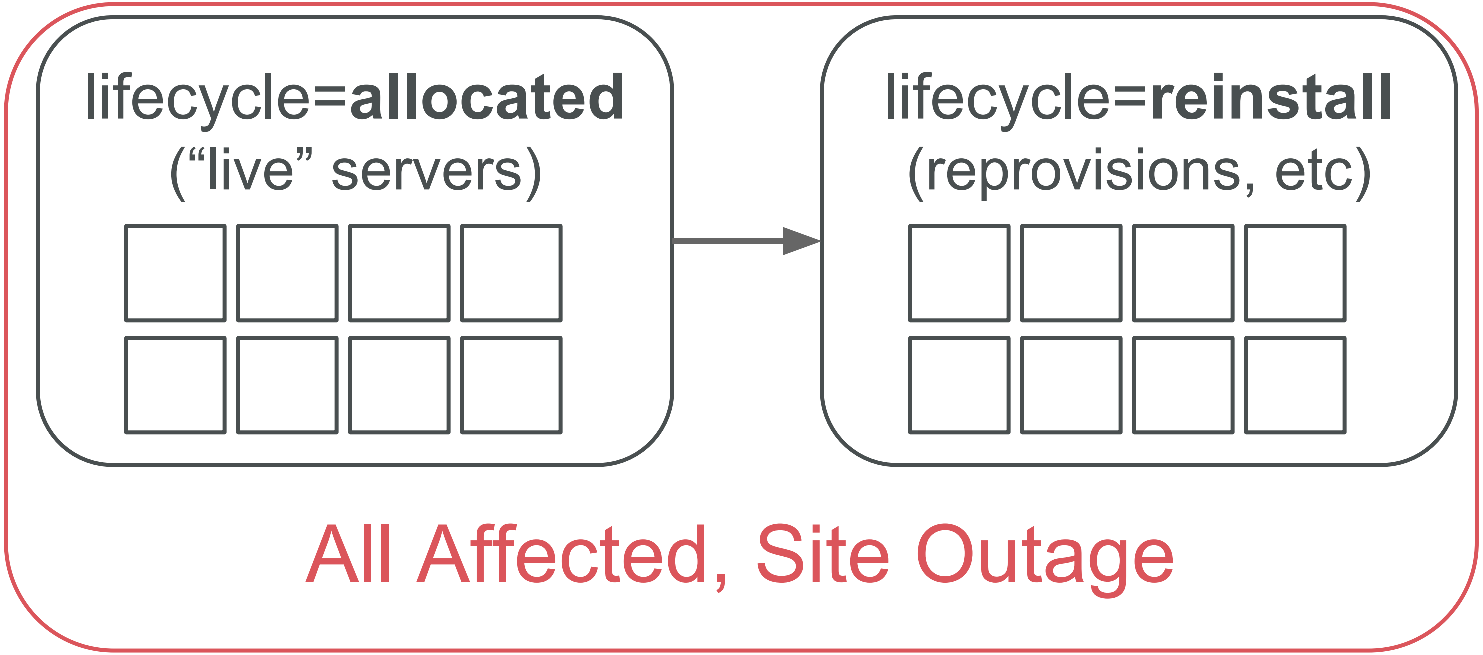
lifecycle=**allocated**
(“live” servers)

lifecycle=**reinstall**
(reprovisions, etc)

To rush upgrades:

```
gsh -q cache lifecycle=reinstall upgrade-host.sh
```

Parsed (and ignored) as environment variable

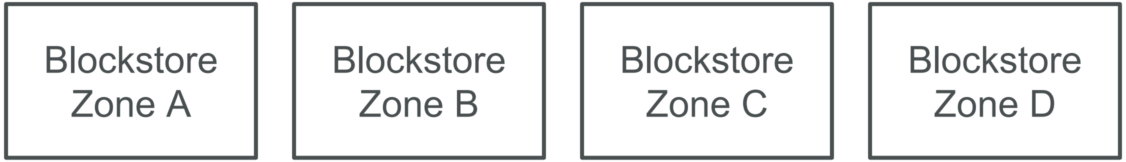


To rush upgrades:

```
gsh -q cache lifecycle=reinstall upgrade-host.sh
```

Human Error? Protect!

Distributed Shell Gating



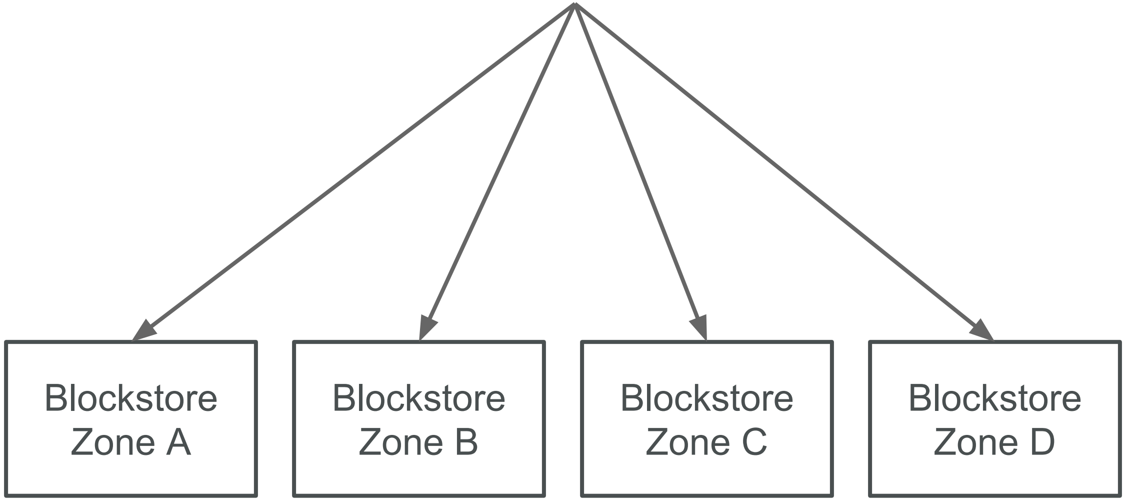
Blockstore
Zone A

Blockstore
Zone B

Blockstore
Zone C

Blockstore
Zone D

```
dsh -q “storage-node” “upgrade-host”
```



Blockstore Zone A

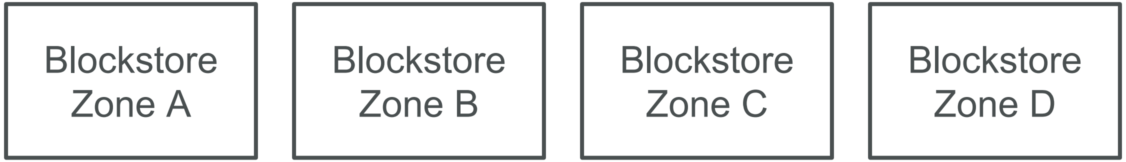
Blockstore Zone B

Blockstore Zone C

Blockstore Zone D

```
dsh -q “storage-node” “upgrade-host”

    ERROR:root:Exiting early
    because: Can't run DSH across
    multiple Blockstore zones.
```



Blockstore Zone A

Blockstore Zone B

Blockstore Zone C

Blockstore Zone D

Sudo Passwords

```
[/home/mah] ➜ rm -rf /mnt
```

```
[/home/mah] ➜ rm -rf /mnt
[sudo] password for mah:
```

Tomoyo

Tomoyo?

Linux Kernel module

System call access controls

```
[/mnt/blocks] ➜ sudo rm block-12345
```

```
[/mnt/blocks] ➜ sudo rm block-12345
rm: cannot remove `block-12345’: Operation not permitted
```

```
[/mnt/blocks] ➜ sudo rm block-12345
rm: cannot remove `block-12345’: Operation not permitted

(strace output)
unlinkat(AT_FDCWD, "block-12345", 0)
= -1 EPERM (Operation not permitted)
```

Human Error?

Protect!

Distributed Shell Gates
Sudo Passwords
Tomoyo

Biggest Fears

Software

Hardware

Humans

Tooling/Automation

Protections

Fear of Automation

It will decide on its own to go reformat all of the hard drives

Tooling Bugs? Verify!

Tool/Automation Reports

Lifecycle of Tooling

Manual Labor
Scripts
Self driven automation

Tooling Bugs? Protect!

Lifecycle of Tooling

Manual Labor

Scripts

Human Authorized Execution

Self driven automation

Diagnosis

Diagnosis

```
PartitionTableInputOutputError
* Host: abc-de11-9f
* Reading /dev/sdam1’s partition table.
* Encountered IO error.
* read(3, 0xe3b600, 512) = -1 EIO (Input/output error)
```

Prescription

Prescription

```
This disk is unusable. Thus, DecommissionDisk
> bsctl osd decommission_disk abc-de11-9f 7037
```

Human Authorization

Human Authorization

```
If every diagnosis above checks out as reasonable..
Type 'yep, that evidence seems legit'
to run these commands:
```

Replication-Dependent Gating

```
[/home/mah/] ➜ bsctl deallocate abc-de11-9f
```

```
[/home/mah/] ➜ bsctl deallocate abc-de11-9f
abc-de11-9f has 2917234 imperfectly replicated blocks.
Aborting.
Please wait for replication before trying again.
[/home/mah/] ➜
```

Respecting Isolation

Automation Bugs?

Verify!

Automation Report

Protect!

Human Authorization

Replication Dependent Gating

Respect Isolation

Biggest Fears

Software

Hardware

Humans

Tooling/Automation

Protections

Fear of Protections

They’re going to brick the servers
They’re going to not actually work!

Protections bug? Verify!

Disaster Recovery Testing

Protections Bug? Protect!

Override Capabilities

```
[/home/mah/] ➜ override_tomoyo_policy.py --no_enforce
```

Failure in Protections?

Verify!

Disaster Recovery Testing

Protect!

Override Capabilities

Recap?

Find your fears

Verify! Protect!

Embrace your paranoia

Verify! Protect!

David Mah
mah@dropbox.com