Distributed Storage Systems

John Leach
john@brightbox.com
twitter @johnleach

Brightbox Cloud
http://brightbox.com

Brightbox™

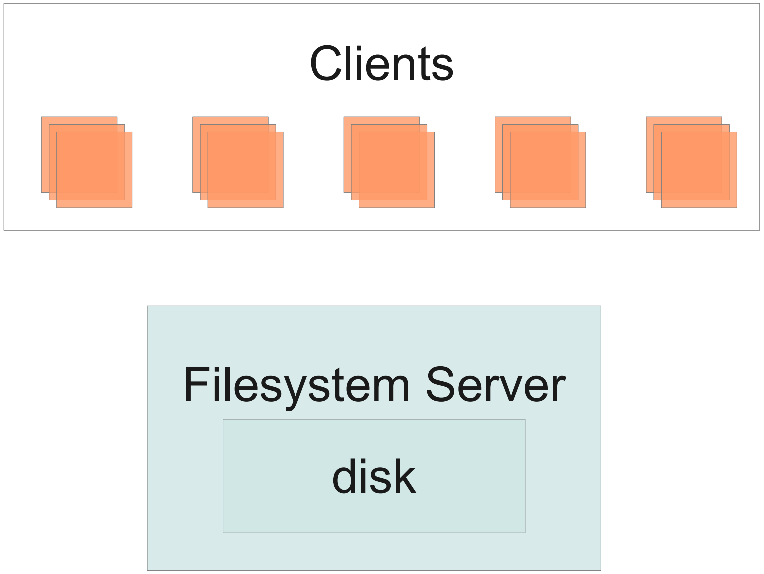
Our requirements

- Bright box has multiple zones (data centres)
- Should tolerate a zone failure
- Scale smoothly as data size grows
- Should use exciting unproven technology
- Libre software license

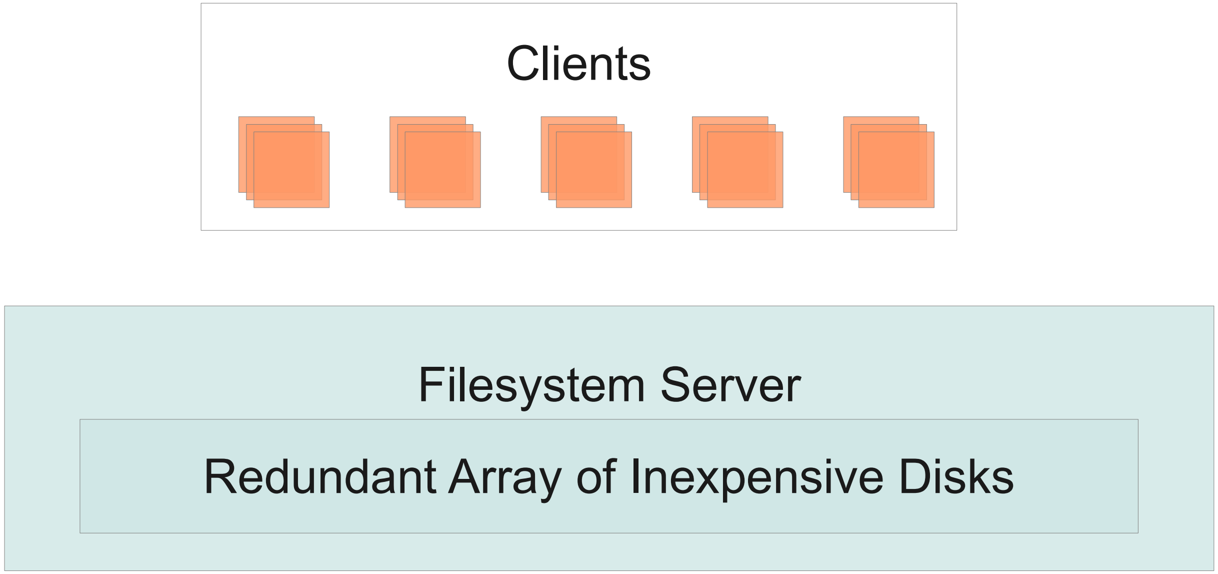
Brief history of file access

Brightbox™

Scaling NFS: One disk



Scaling NFS: RAID

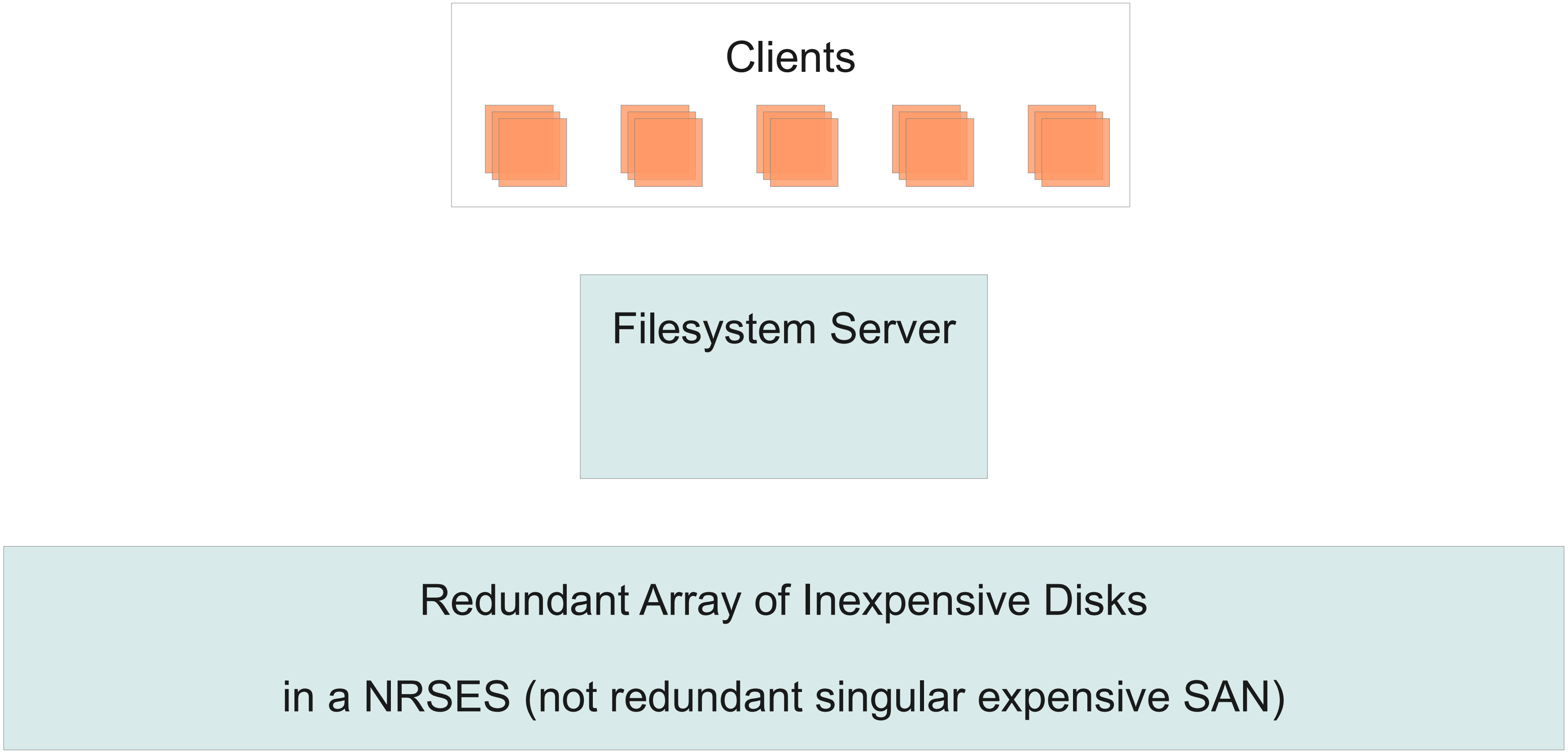


Clients

Filesystem Server

Redundant Array of Inexpensive Disks

Scaling NFS: SAN



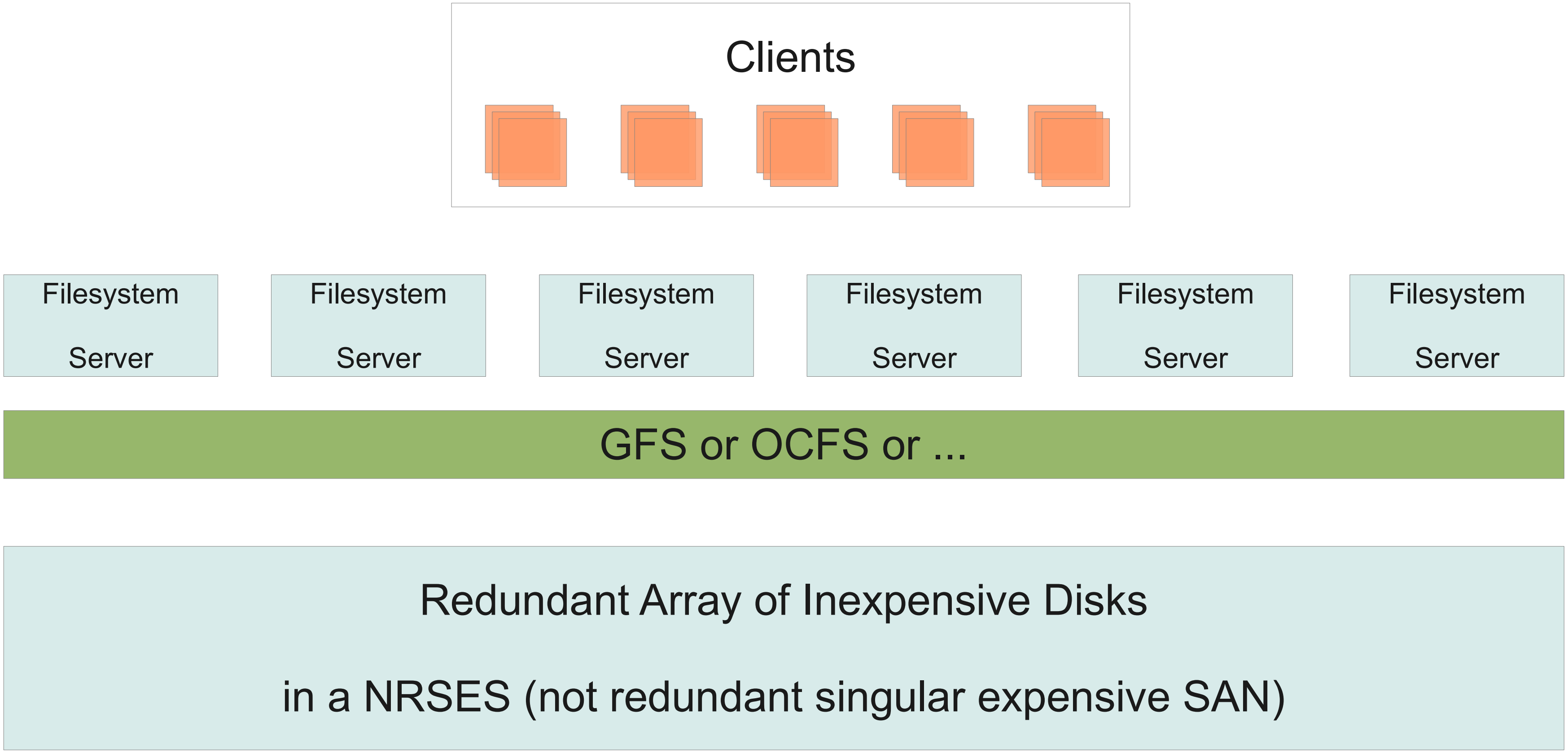
Clients

Filesystem Server

Redundant Array of Inexpensive Disks

in a NRSES (not redundant singular expensive SAN)

Scaling NFS: Shared disk fs



Clients

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

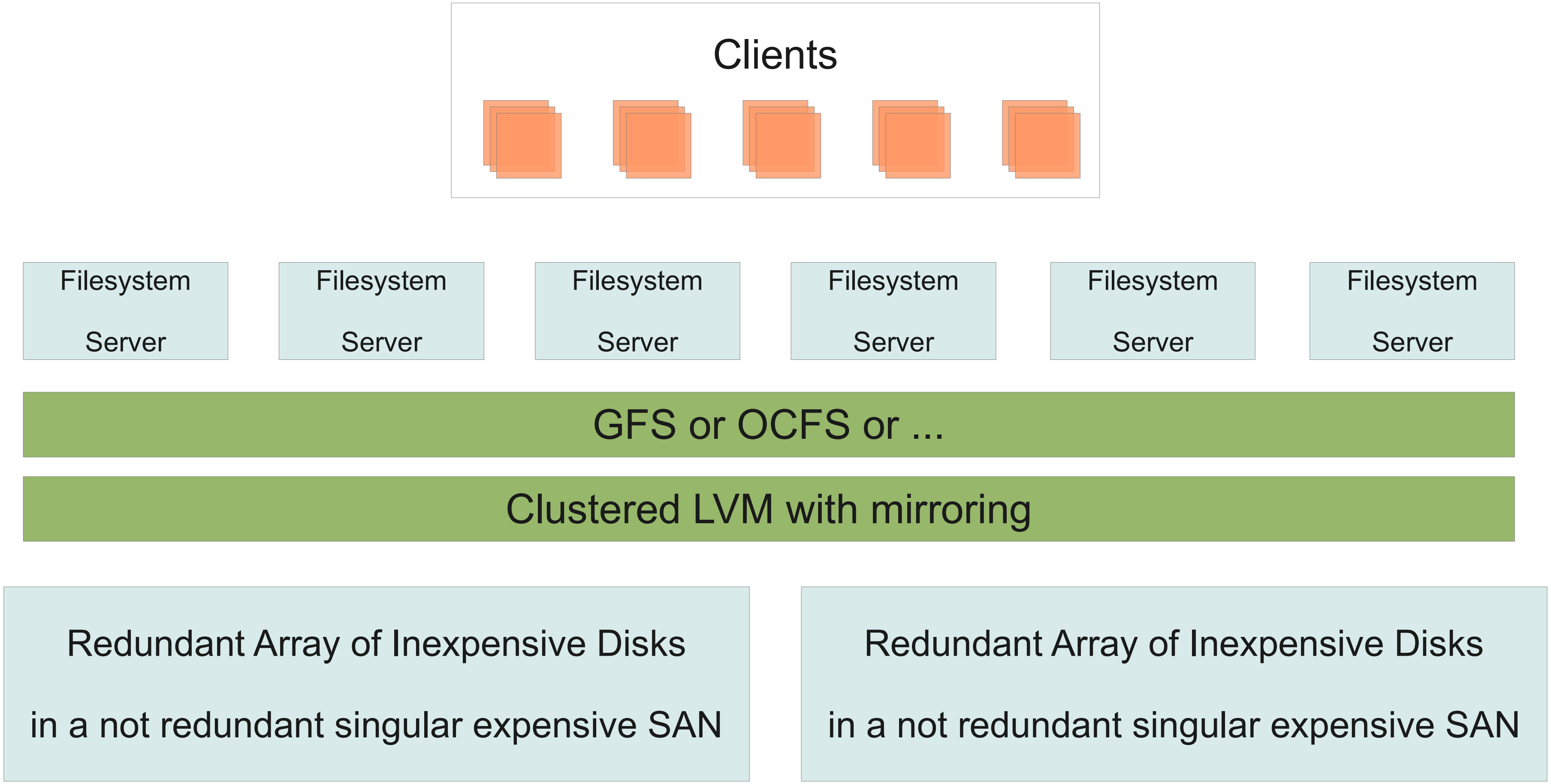
Filesystem Server

GFS or OCFS or ...

Redundant Array of Inexpensive Disks

in a NRSES (not redundant singular expensive SAN)

Shared disk fs: Replication



Clients

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

GFS or OCFS or ...

Clustered LVM with mirroring

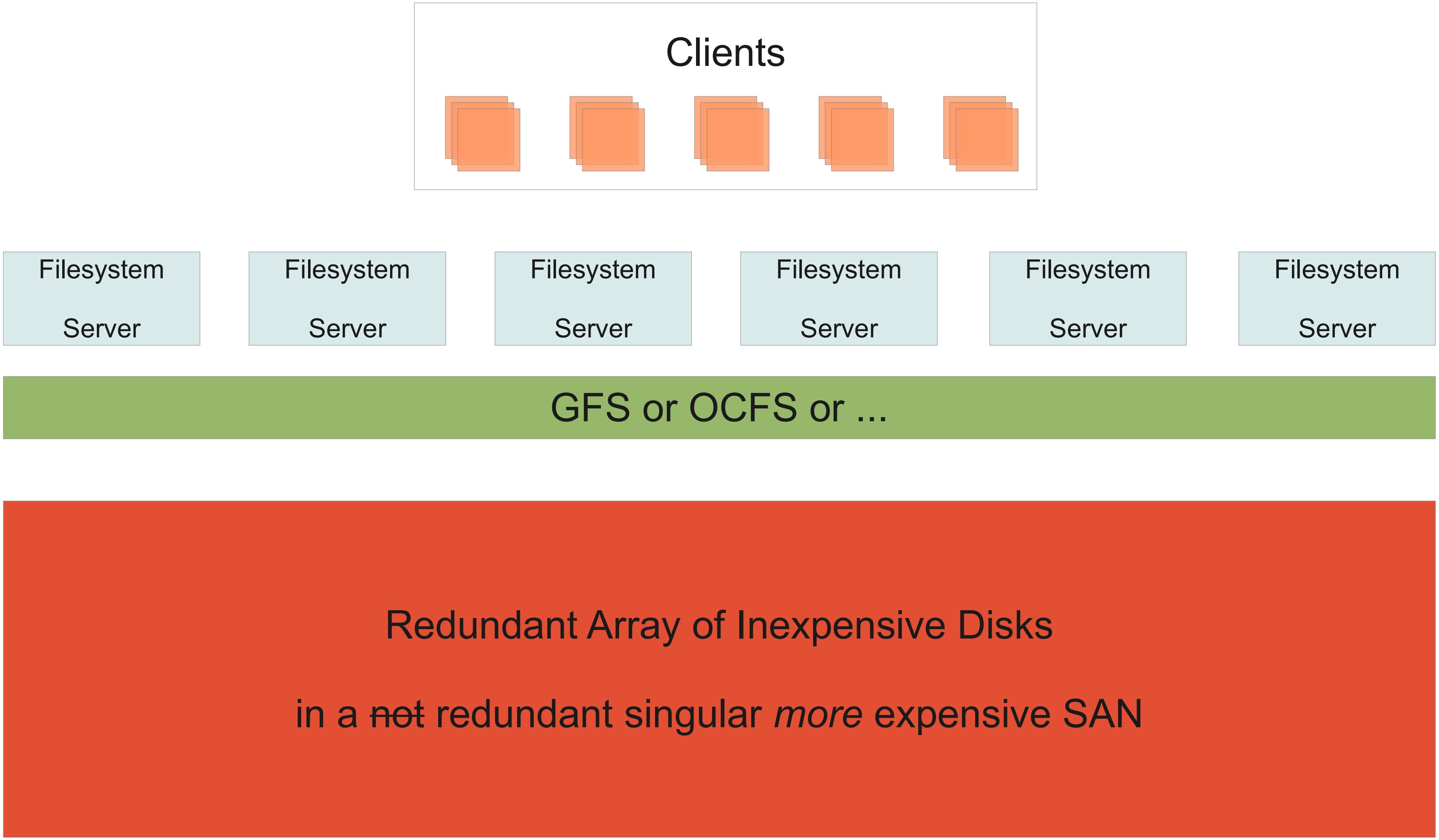
Redundant Array of Inexpensive Disks

in a not redundant singular expensive SAN

Redundant Array of Inexpensive Disks

in a not redundant singular expensive SAN

Shared disk fs: Replication



Clients

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

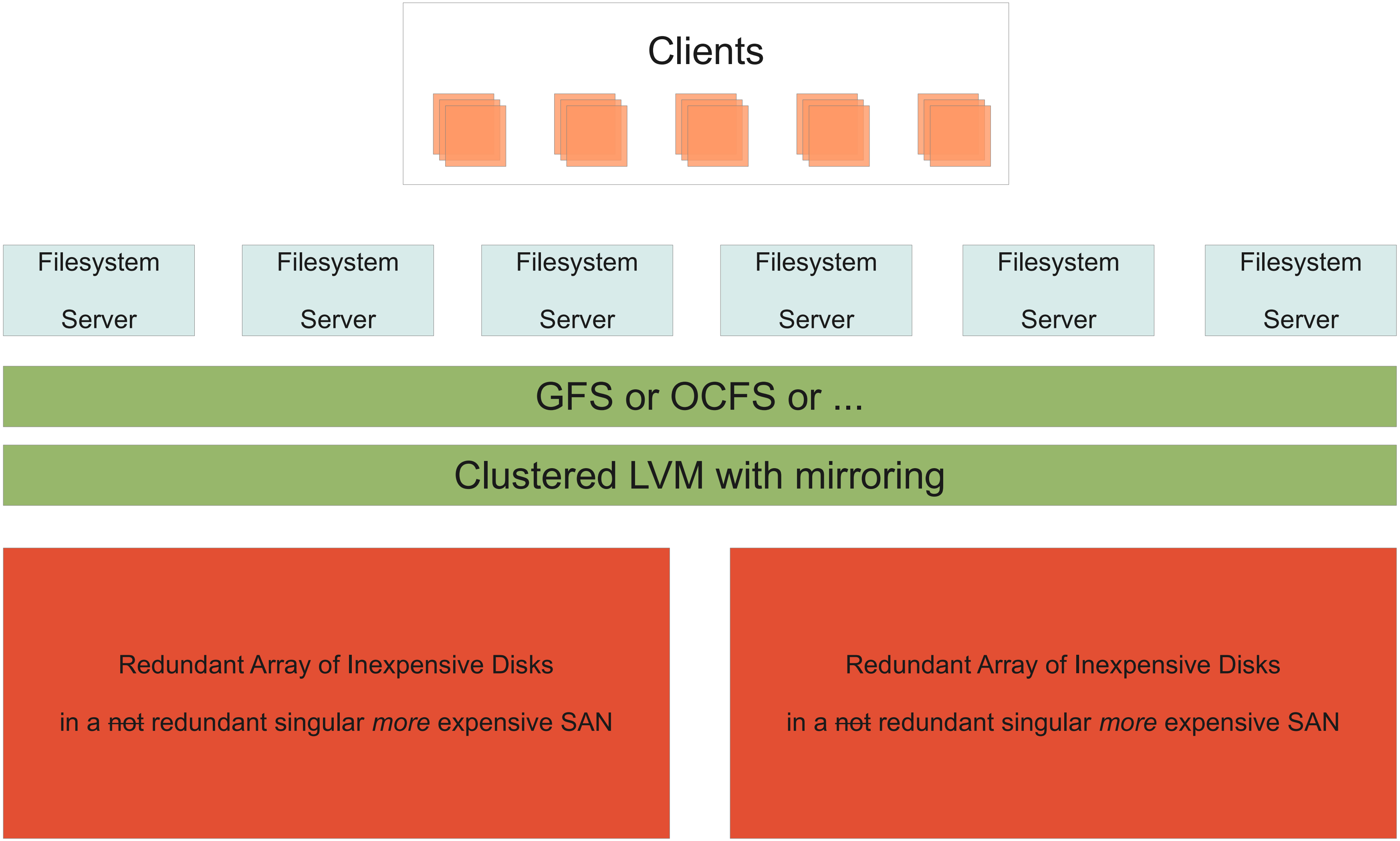
Filesystem Server

GFS or OCFS or ...

Redundant Array of Inexpensive Disks

in a ~~not~~ redundant singular *more* expensive SAN

Shared disk fs: Replication



Clients

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

Filesystem Server

GFS or OCFS or ...

Clustered LVM with mirroring

Redundant Array of Inexpensive Disks

in a ~~not~~ redundant singular *more* expensive SAN

Redundant Array of Inexpensive Disks

in a ~~not~~ redundant singular *more* expensive SAN

Old techniques

- Hot or warm standby servers
- Expensive SAN hardware
- Shared block devices
- Moving IP addresses
- Server side replication
- Scales mostly vertically
- Manual partitioning to scale horizontally

New techniques

- Shared nothing
- Clever clients
- Automatic partitioning
- Automatic replication
- Clever stuff: DHT, Vector clocks, PAXOS, Mapreduce, Merkle trees, Unicorn hooves
- ~~POSIX~~

New Problems

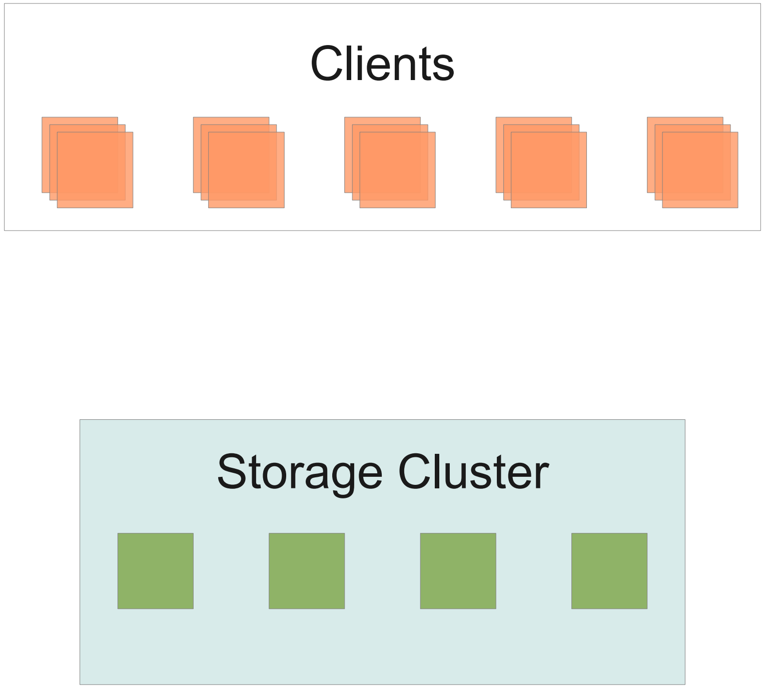
- Locating your data
- Ensuring consistency
- Something has to give

Brightbox™

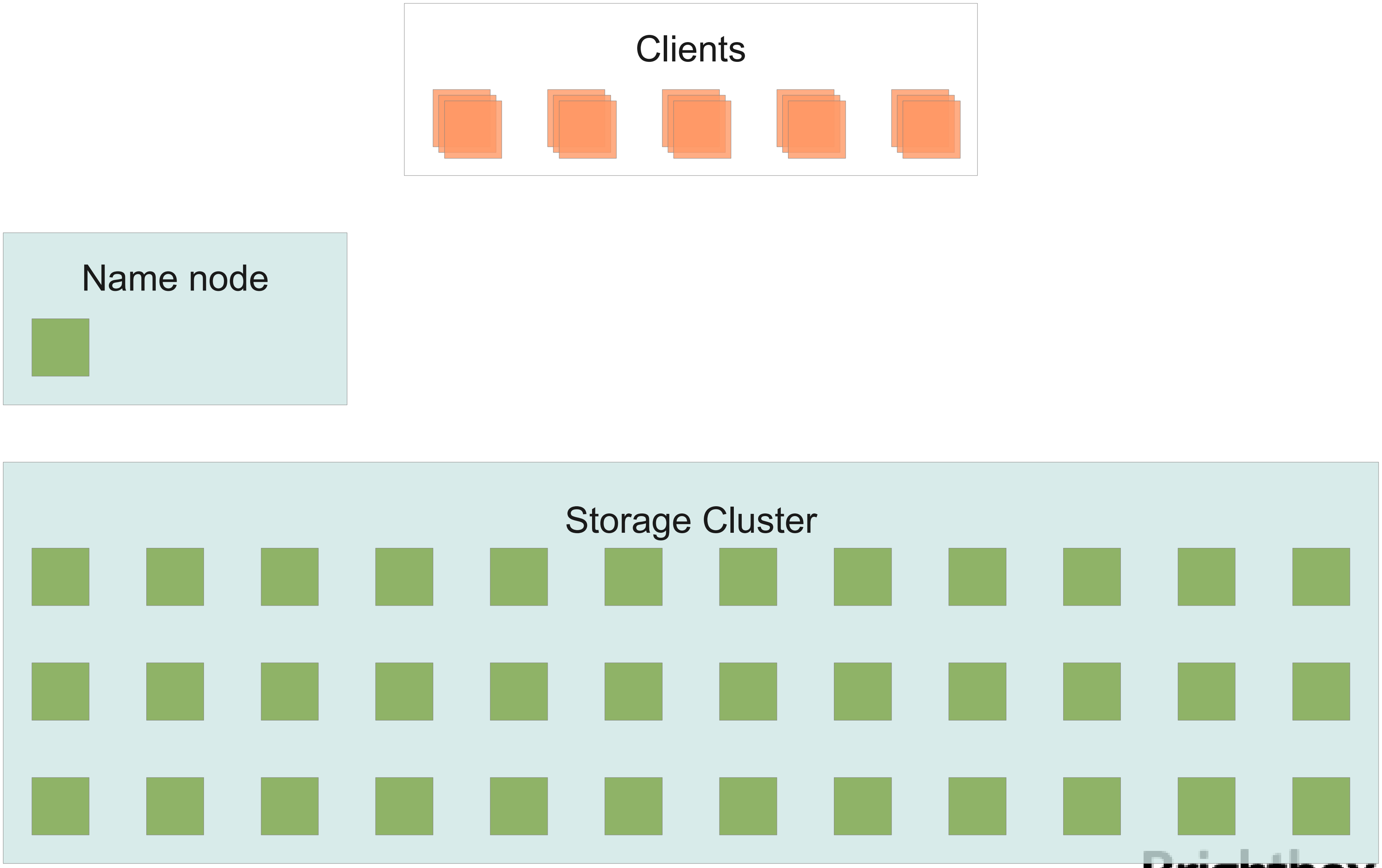
Brewers CAP theorem

- Consistency
- Availability
- Partition tolerance

GlusterFS



Hadoop File System



Clients

Name node

Storage Cluster

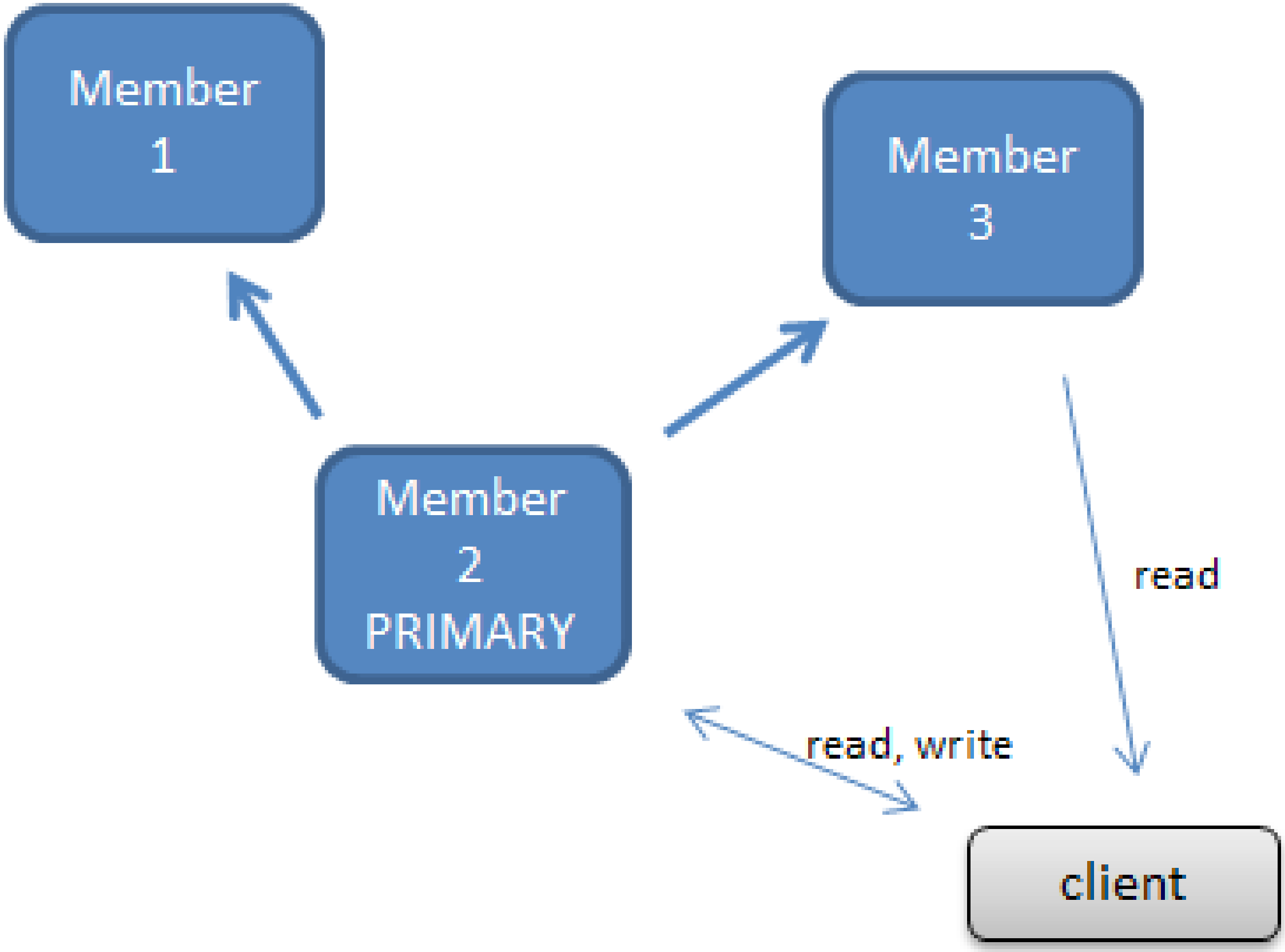
Hadoop File System

- Hot failover patches in Feb
- Batch processing, not interactive
- High throughput, not low latency
- Map Reduce
- Namenode SPOF
- Multi-data centre
- Consistent

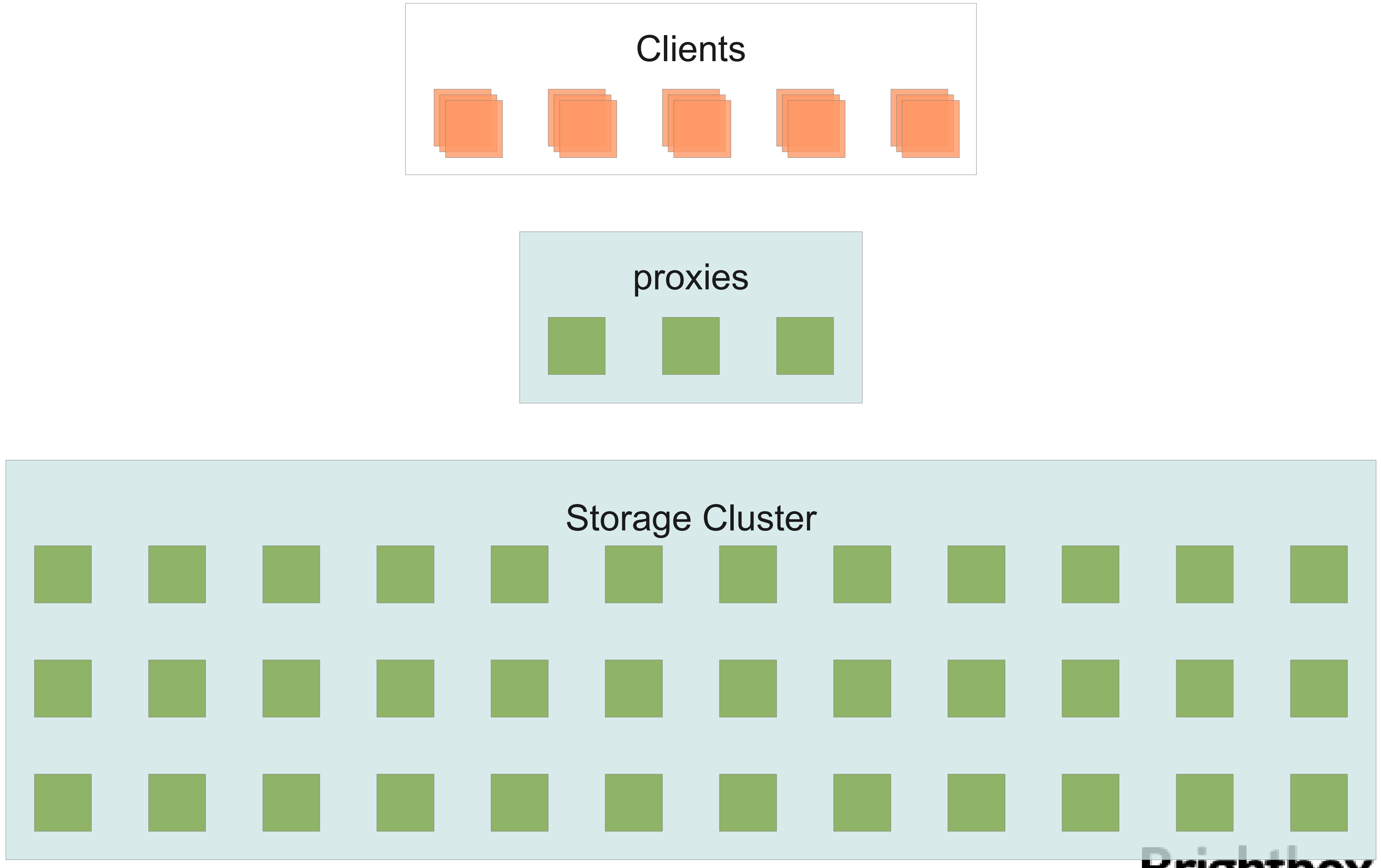
MongoDB

- Document store, dynamic schema
- Async replication
- Primary server for writes
- Automatic sharding
- Map Reduce
- GridFS for large files
- Multi-datacentre, but not partition tolerant
- Mostly consistent

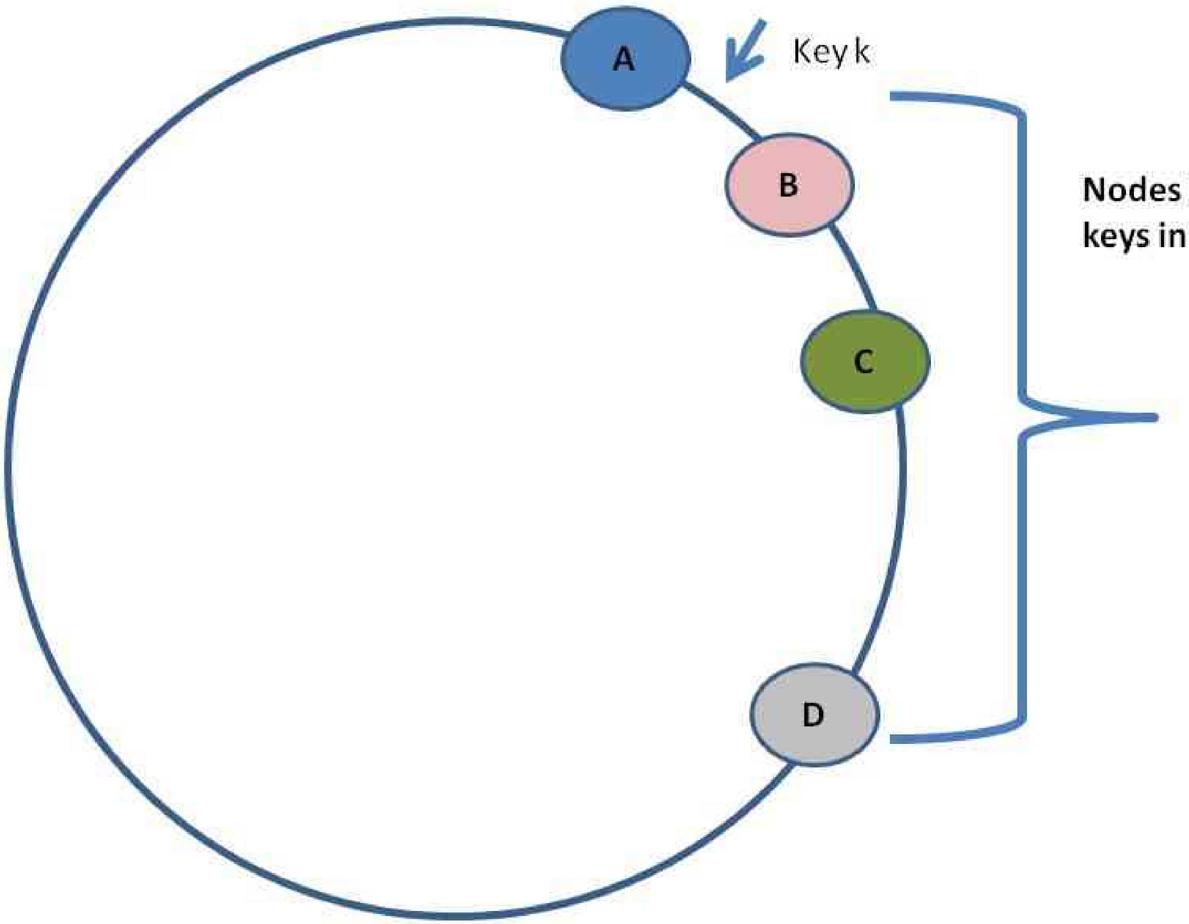
MongoDB



Openstack Swift



Openstack Swift



Nodes B,C,D store keys in the range A-B

Cassandra

- P2P, DHT, Gossip, Hinted Handoff
- Column orientated. Data ordered.
- Design schema for types of queries
- Very fast highly available writing
- Per request consistency. Multi-data centre
- Thrift API

Riak

- Key value store.
- DHT, Gossip, Vector Clocks
- Map reduce
- Luwak for large files

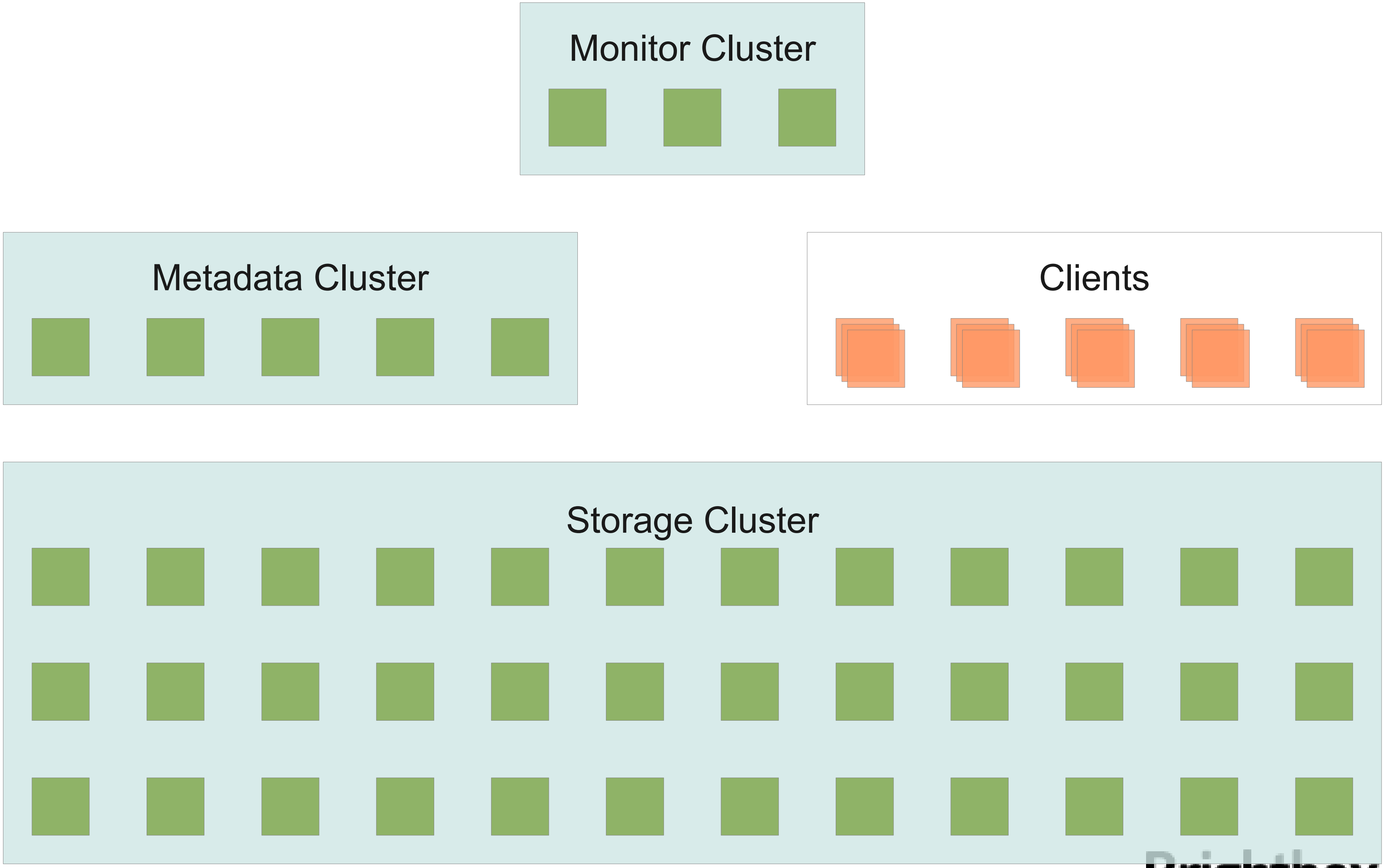
Zookeeper

- PAXOS like consensus protocol
- Read scales up with more servers
- Writes slow down with more servers
- Always consistent
- In-memory
- Strict ordering
- Small data

Ceph

- Object store
- Full POSIX file system on top
- PAXOS for cluster state
- CRUSH rather than DHT
- Multi-datacenter.
- Strongly consistent, not partition tolerant
- RBD, S3-alike, plus POSIX

Ceph


Monitor Cluster
Metadata Cluster
Clients
Storage Cluster

Distributed Storage Systems

John Leach
john@brightbox.com
twitter @johnleach

Brightbox Cloud
http://brightbox.com