lightbits

SDC 20
SNIA EMEA

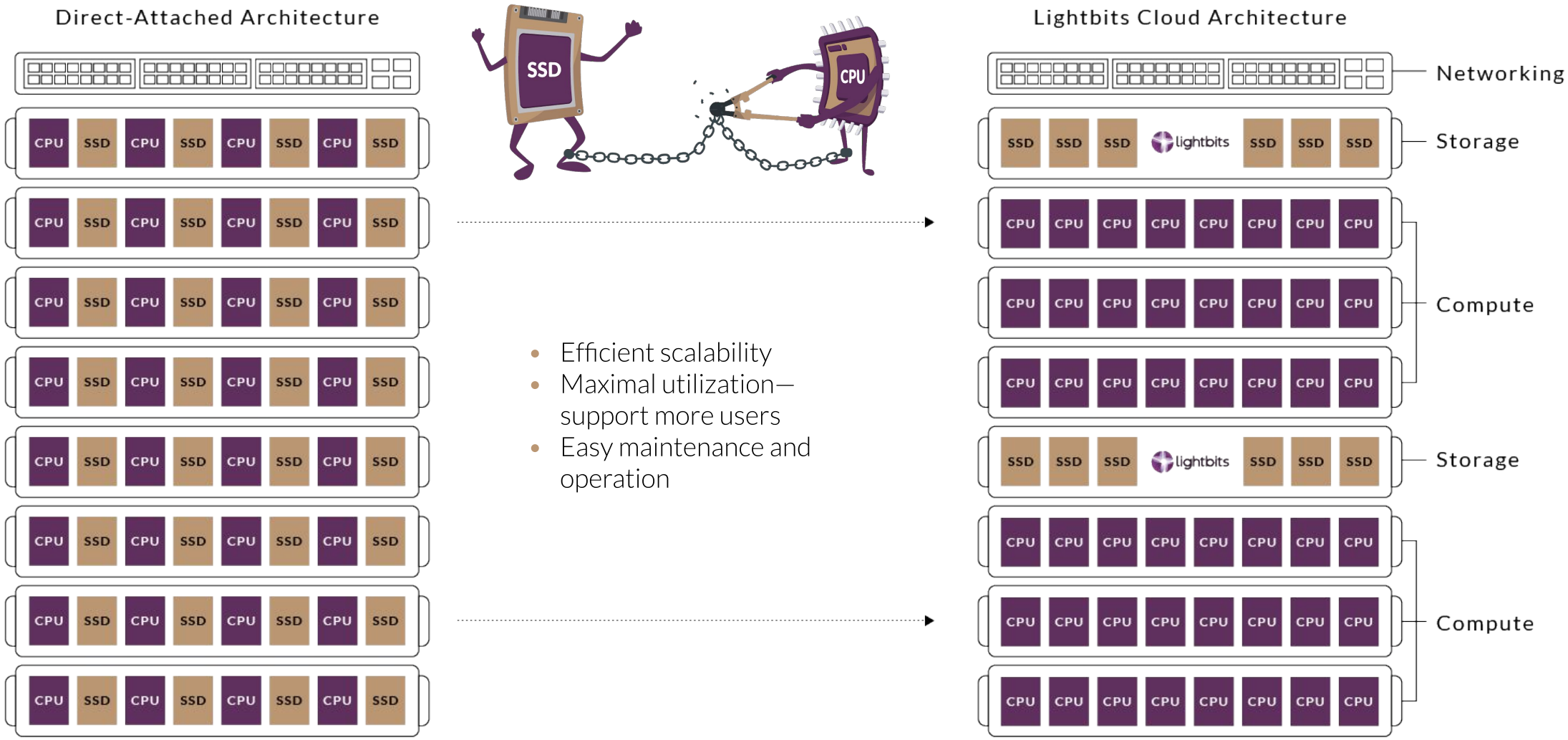
NVMe/TCP Standards-Based, Fault-Tolerant Clustered Storage with LightOS

Alex Shpiner
System Architect
Lightbits Labs
alex@lightbitslabs.com

About Lightbits Labs

- **Founded** in Q1 2016
- **Key milestones:** NVMe/TCP, LightOS 1.0, LightOS 2.0
- **80 Employees** (90% Engineering):
  - EX: MLNX, PMCS, IBM, EMC, APPLE
  - FIrst NVMe SSD and NVMe-oF products
- **Locations**:
  - Israel: Kfar Saba, Haifa
  - US: San Jose , NYC
  - Europe
  - China
- **Funding**: $55M , two rounds
  - Strategic investors: Cisco Investments, Dell Technologies Capital, Micron
  - Angel investors: Avigdor Willenz, Lip-Bu Tan, Marius Nacht and others
  - VCs: SquarePeg Capital, Walden International

From direct-attached storage to disaggregated storage servers

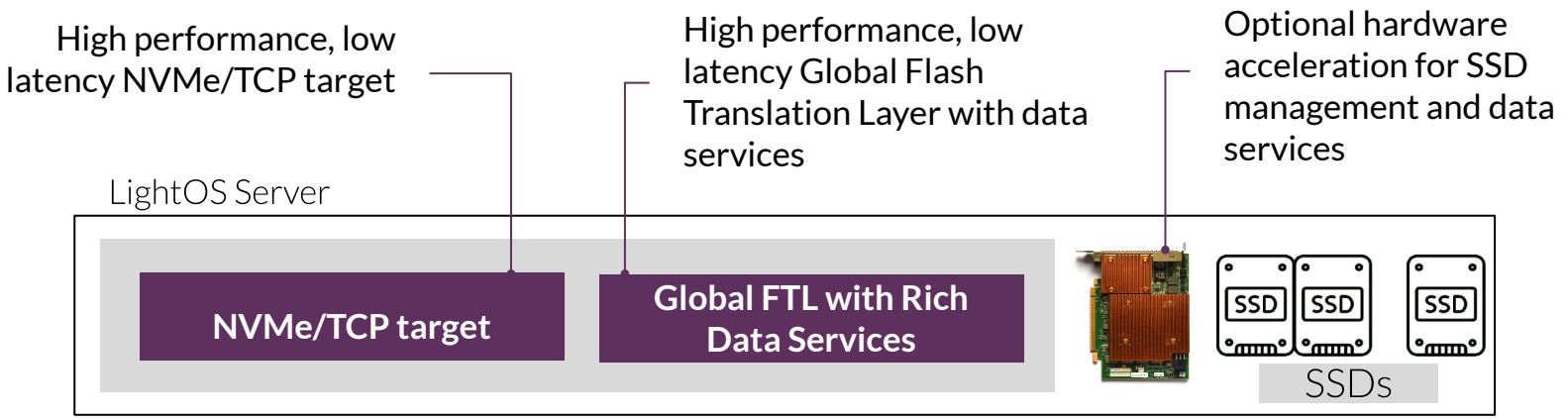


- Efficient scalability
- Maximal utilization—support more users
- Easy maintenance and operation

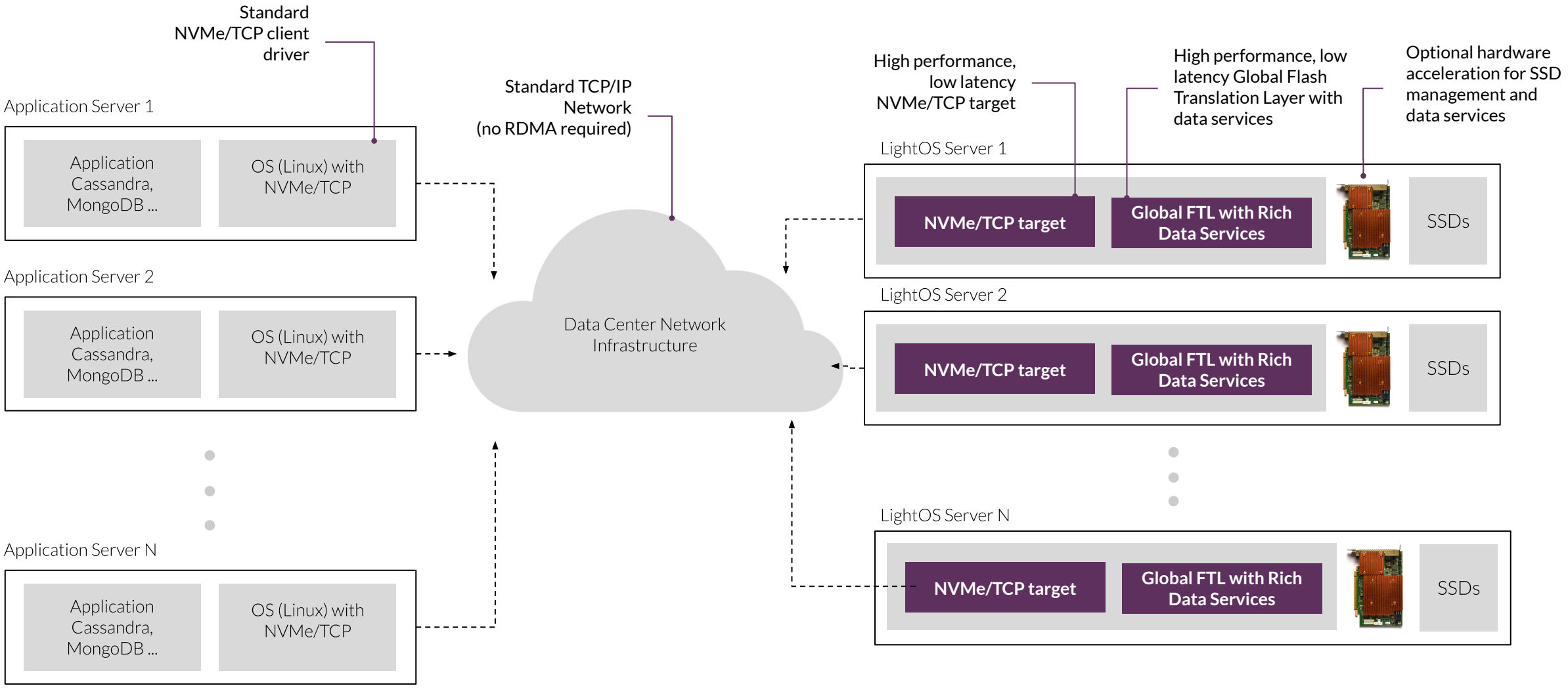
LightOS v1.x

NVMe/TCP target with data services

- First commercially-available, standards based NVMe/TCP
- Software-defined storage
- Standard servers, SSDs and networking
- High throughput, consistent low latency
- Data protection: drives fault tolerance
- Data services including compression and thin provisioning



LightOS-based Disaggregated Storage



Increased Blast Radius with Disaggregated Storage

- LightOS is a great alternative to DAS.
- But, it increases the blast radius from a single compute node to multiple compute nodes.
- For some applications this is a non-issue since the application itself replicates.
  - Although some still prefer a fault tolerant infrastructure.
- For those applications that want fault-tolerant storage:
  - How can we provide the economics, flexibility, and performance of disaggregation with durable disaggregated storage?

	With Application Replication	No Application Replication
LightOS™ v1.x	 Applications that **do replicate** still prefer a fault tolerant infrastructure solution.	
		

Surviving Any Blast

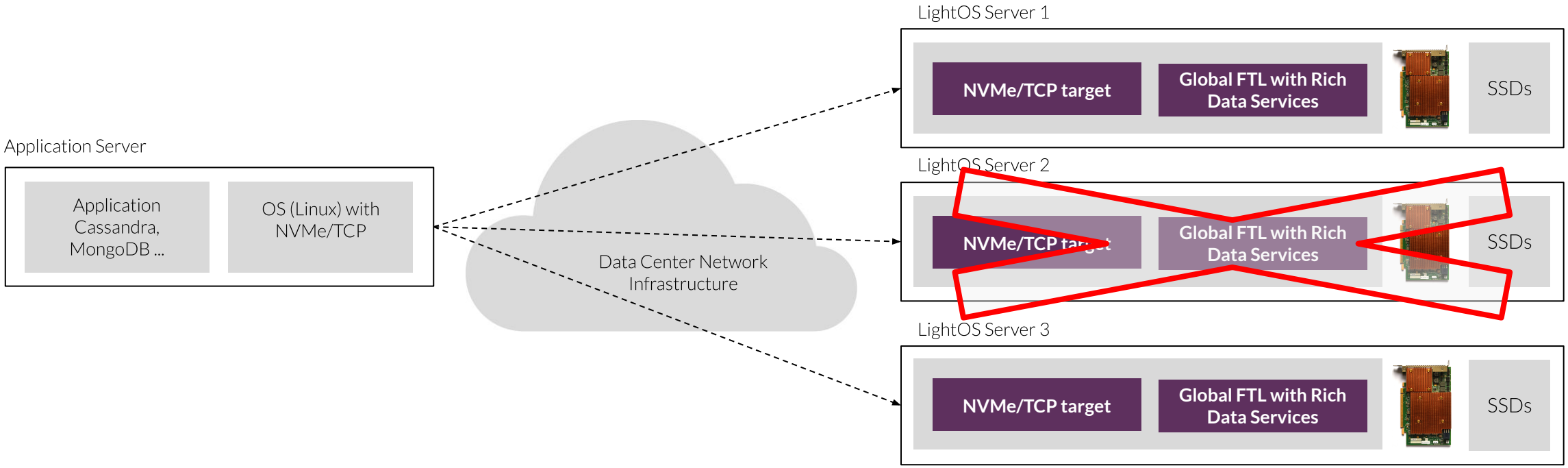
- With LightOS v2.x:
  - In case of server failure, computation nodes (clients) continue working ("business as usual").

	With Application Replication	No Application Replication
LightOS™ v1.x	✓ Applications that **do replicate** still prefer a fault tolerant infrastructure solution.	✗
LightOS™ v2.x	✓	✓

Extending the Scope of Failure Protection

Surviving Any Blast with LightOS Clustering

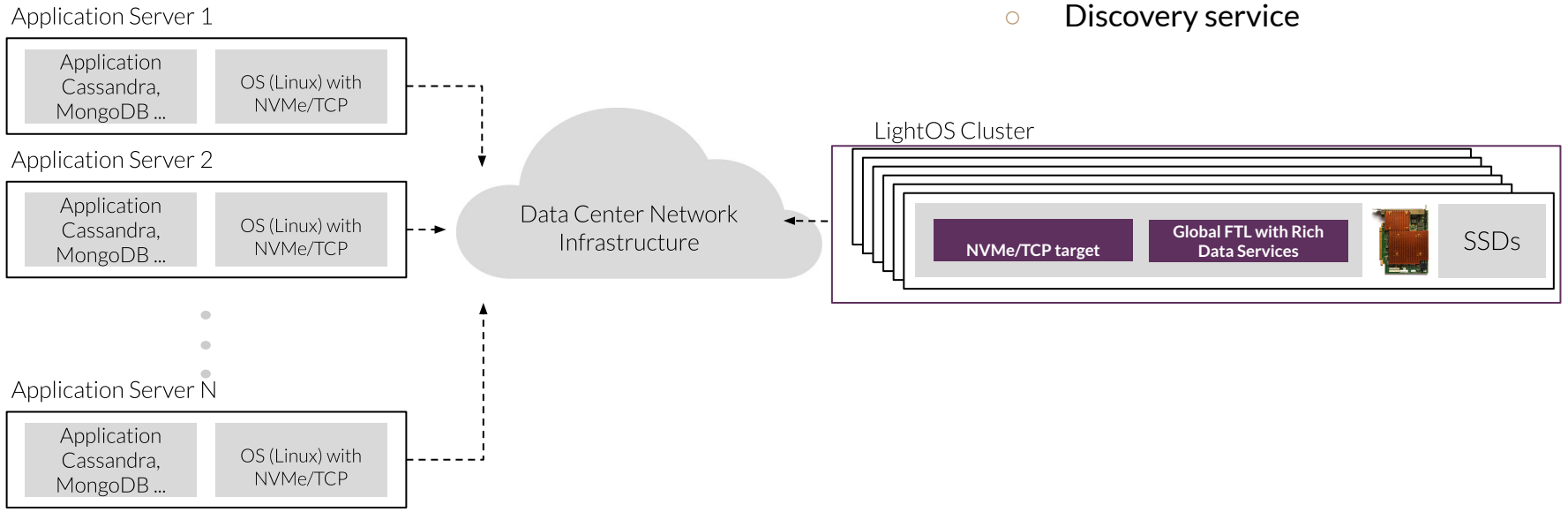
- Clients are connected to multiple storage servers.
- In case of server failure, the service continues from another server.
  - During failover some clients might suffer from performance hit.
  - **All clients continue working!**



LightOS Clustering

Key Features

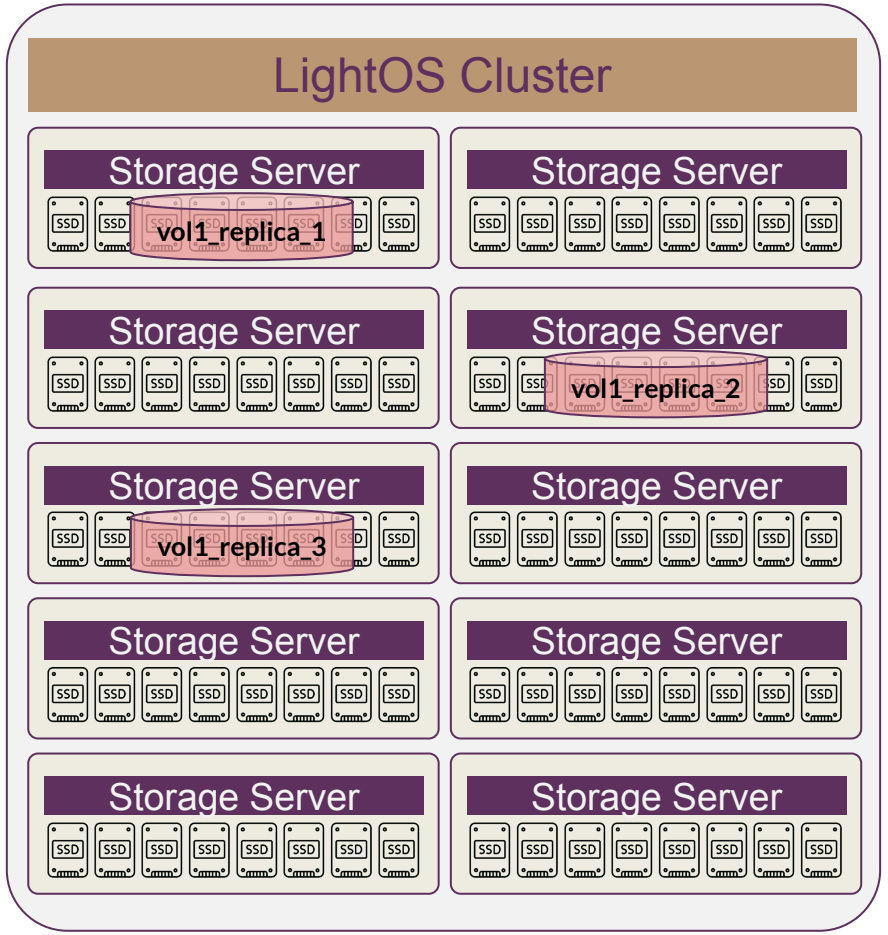
- **Inherit storage services from LightOS 1.x**
- **High performance and low latency**
  - Single hop reads
  - Two hop writes (user + replications)
- **Standard unmodified clients and network**
  - Leveraging standard NVMe-1.4 and NVMeoF 1.1
  - Transparent failover via multipath with Asymmetric Namespace Access (ANA)
- **Distributed and fault tolerant storage servers**
  - Automatic volume assignment
  - Failure domains
  - Management
  - Discovery service



Volume Assignment to Storage Servers

- Multi-replica volumes

- Each replica is stored on a separate storage server

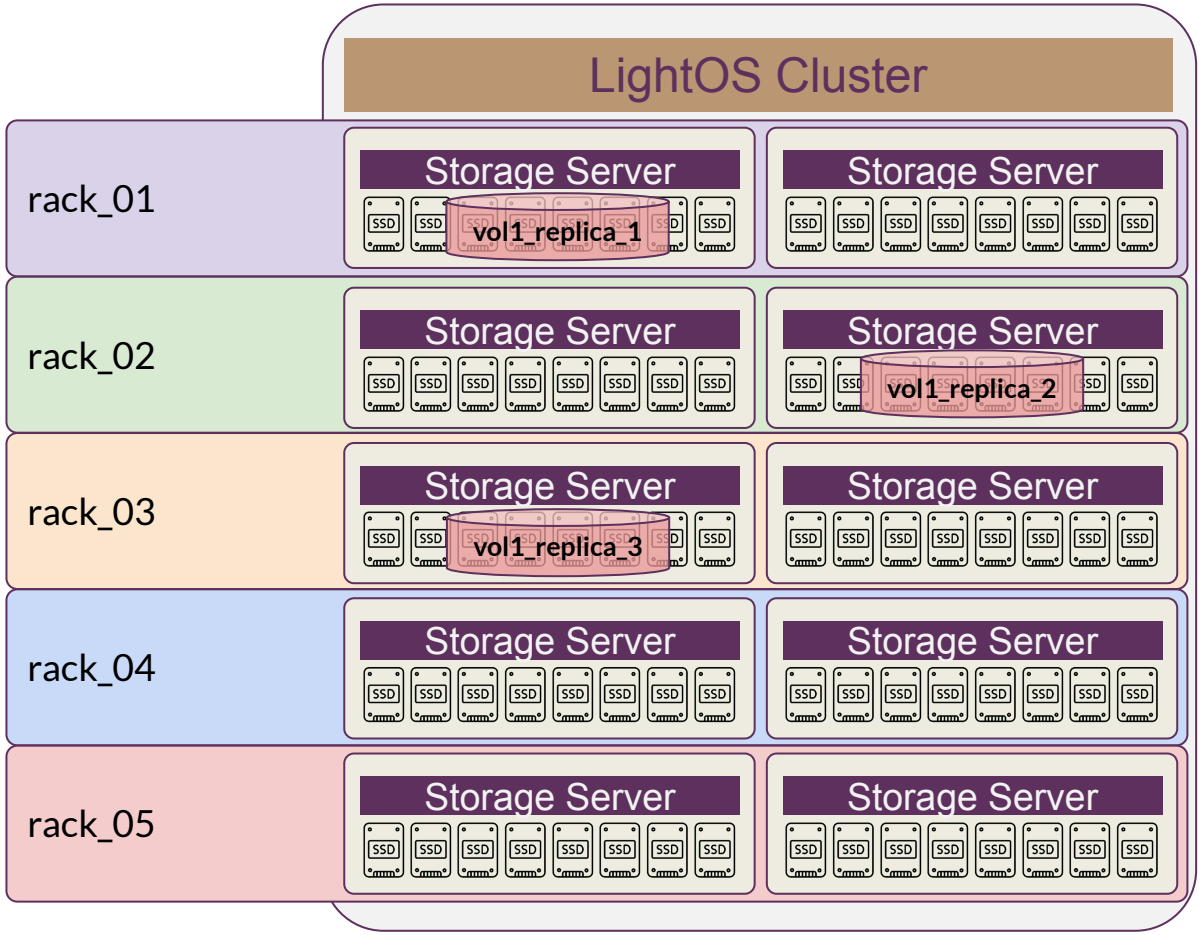


Failure Domains

- Different groups of storage servers can be impacted by common elements that share a point of failure:
  - Network
  - Power
  - Geographical

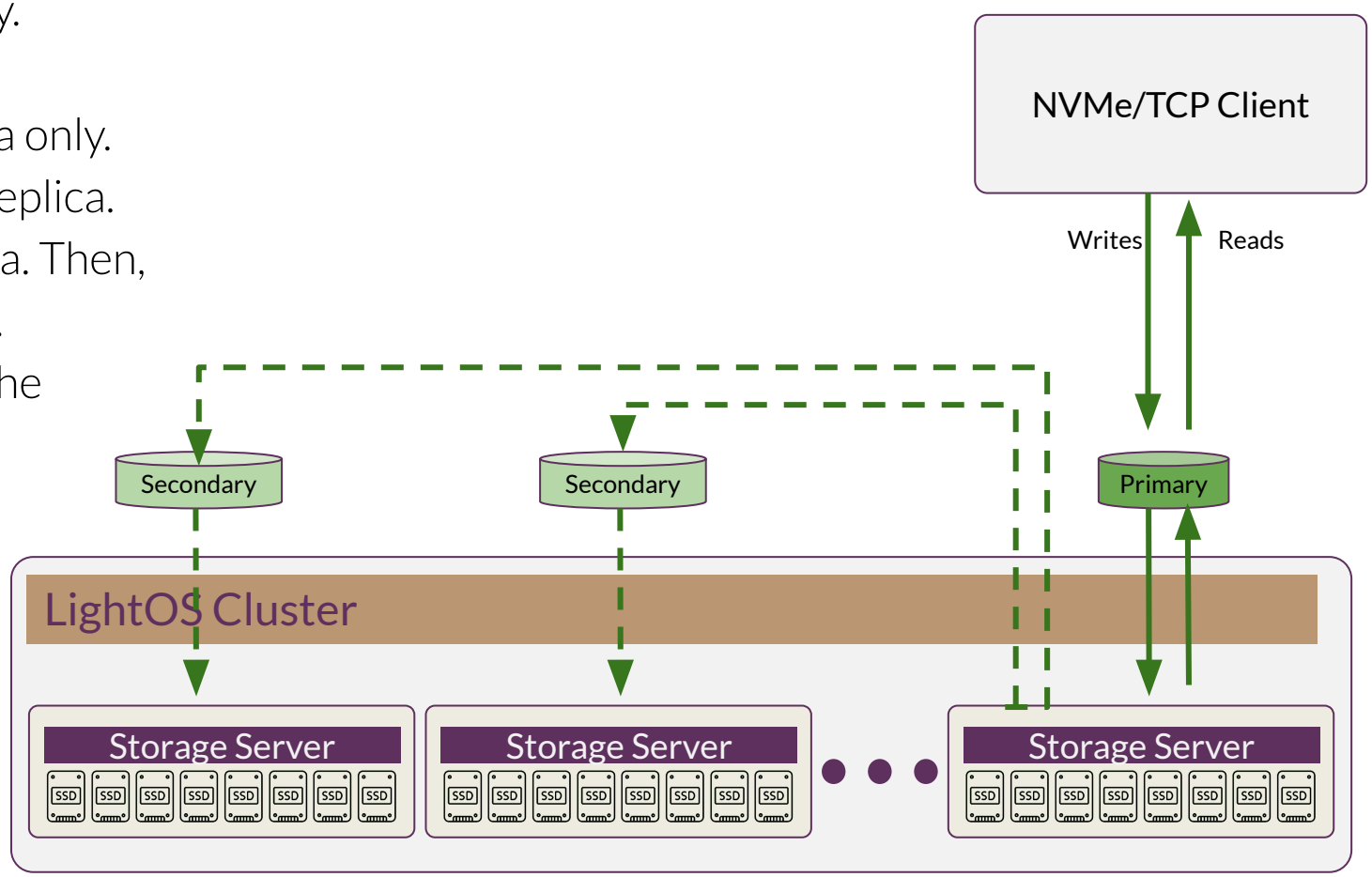
Failure Domains Aware Volume Placement

- User defined server assignments to specific failure domain groups.
- Configured via labels assigned to servers, reflecting common dependencies.
  - rack_01, rack_02, …
  - power_0, power_1, ...
- Replicas are placed in different failure domains.



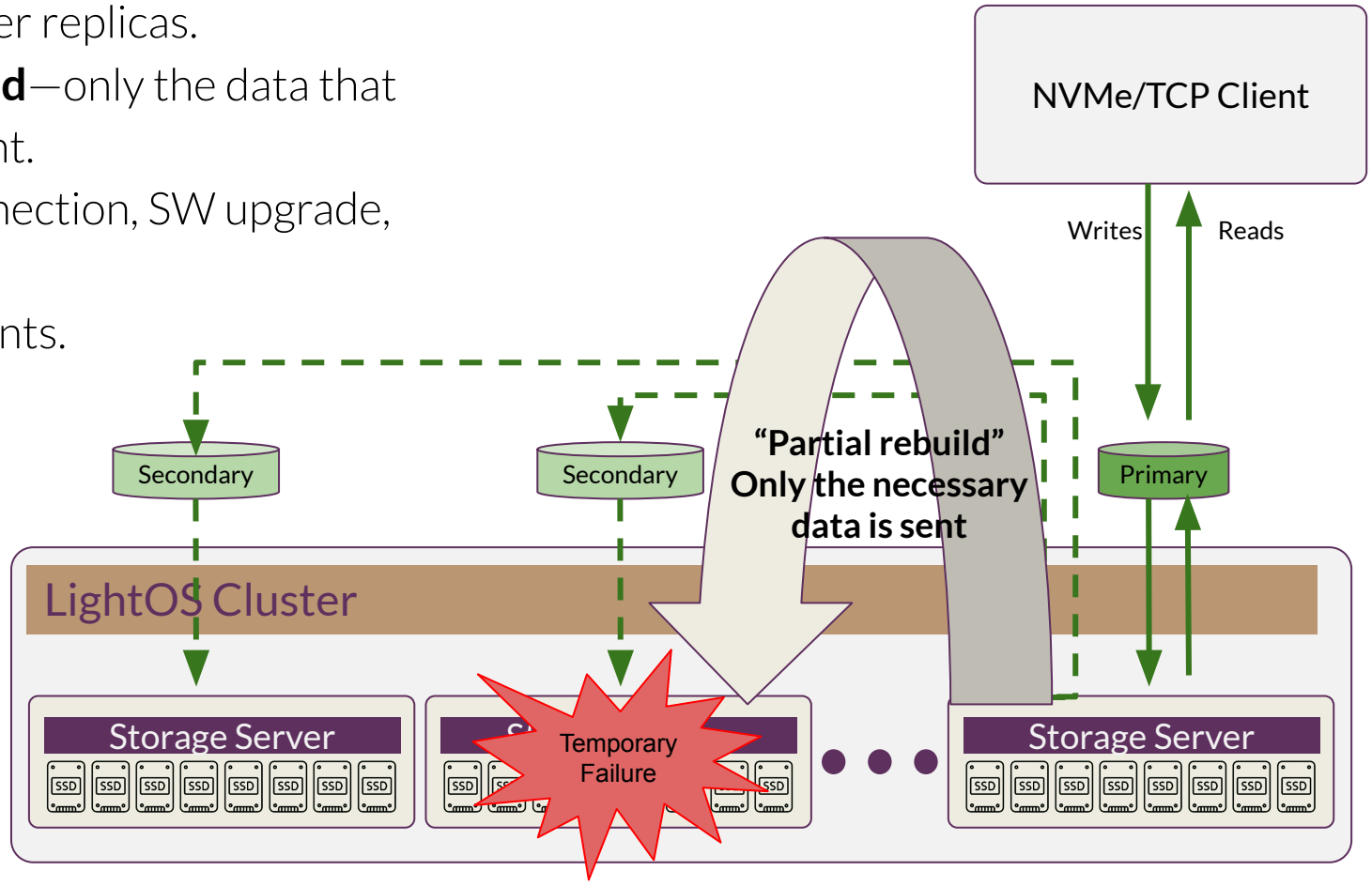
Data Flow

- One of the replicas is defined as a primary.
- The remaining replicas are secondary.
- Client communicates with primary replica only.
- Read requests are served from primary replica.
- Write requests are sent to primary replica. Then, primary replica replicates to secondaries.
- Write requests are acknowledged after the request was replicated.



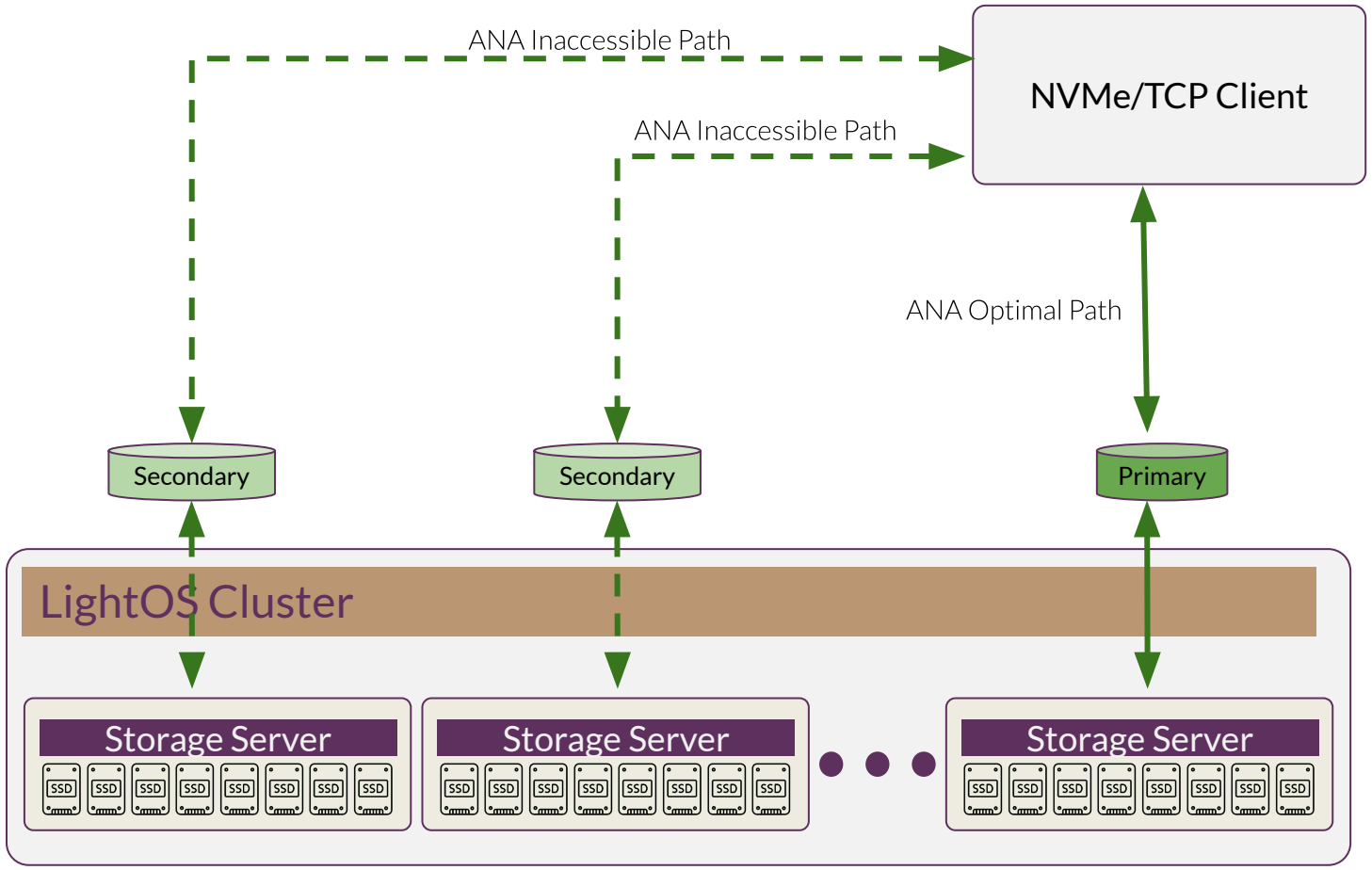
Server failure handling

- If a server fails, its data is rebuilt from other replicas.
  - Temporary failures use **partial rebuild**—only the data that was changed during a failure is re-sent.
  - Temporary Failure = network disconnection, SW upgrade, FW upgrade, etc.
- Rebuild operations are transparent to clients.



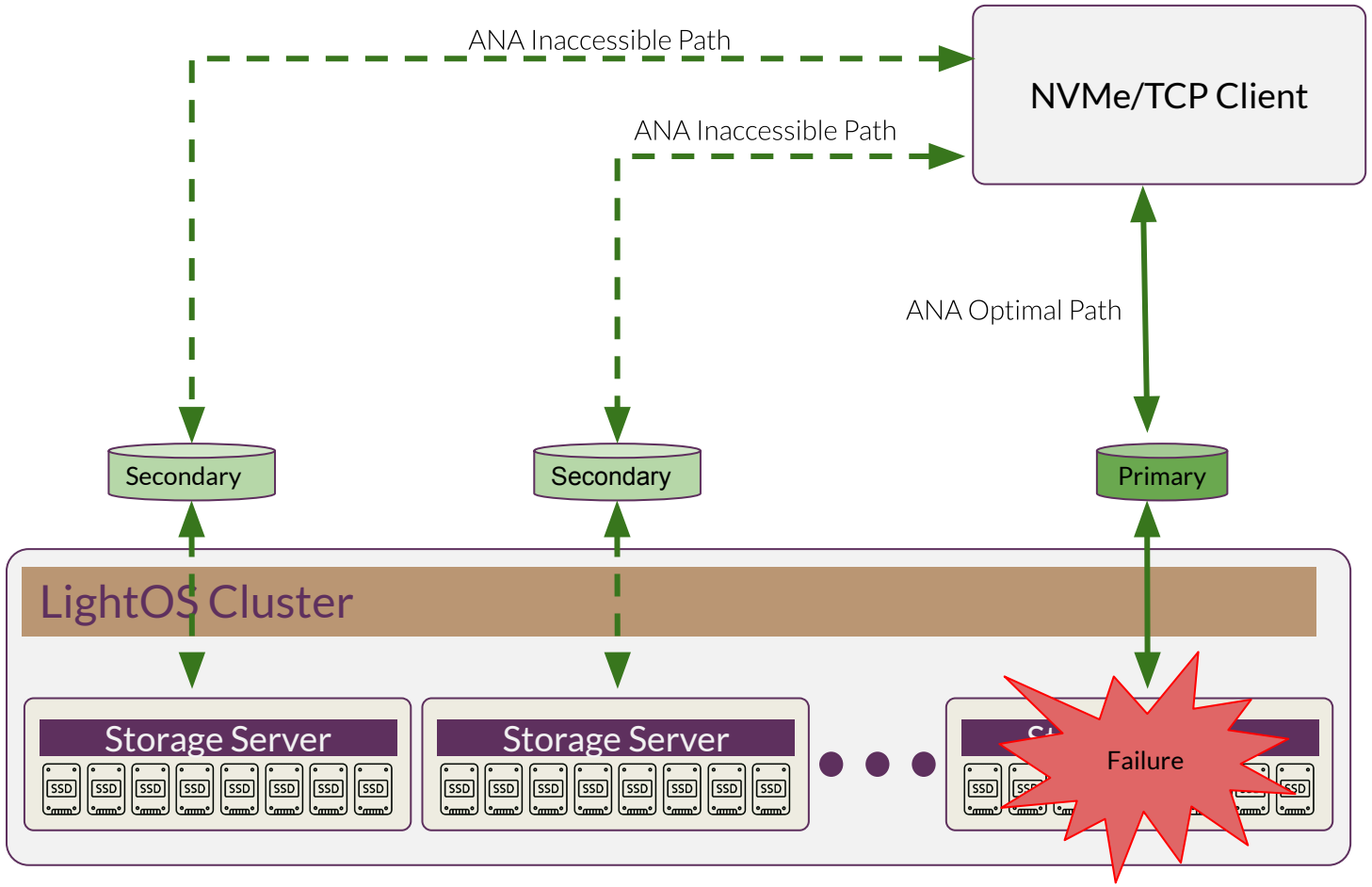
NVMe Asymmetric Namespace Access (ANA)

- NVMe Multipath IO defines access to NVMe namespace across two or more NVMe controllers (represent network paths).
- Multipath Namespace Access schemes:
  - **Symmetric**: All paths are equal
  - **Asymmetric**: Path state informs on access semantics
    - Optimized: Preferred accessible path
    - Non-Optimized: Non-preferred accessible path
    - Inaccessible
- **LightOS leverages NVMe ANA for Clustering**
  - Primary server reports "Optimized" ANA State
  - Secondary servers report "Inaccessible" ANA State
  - **Failure Handling**: Controller changes state in ANA report



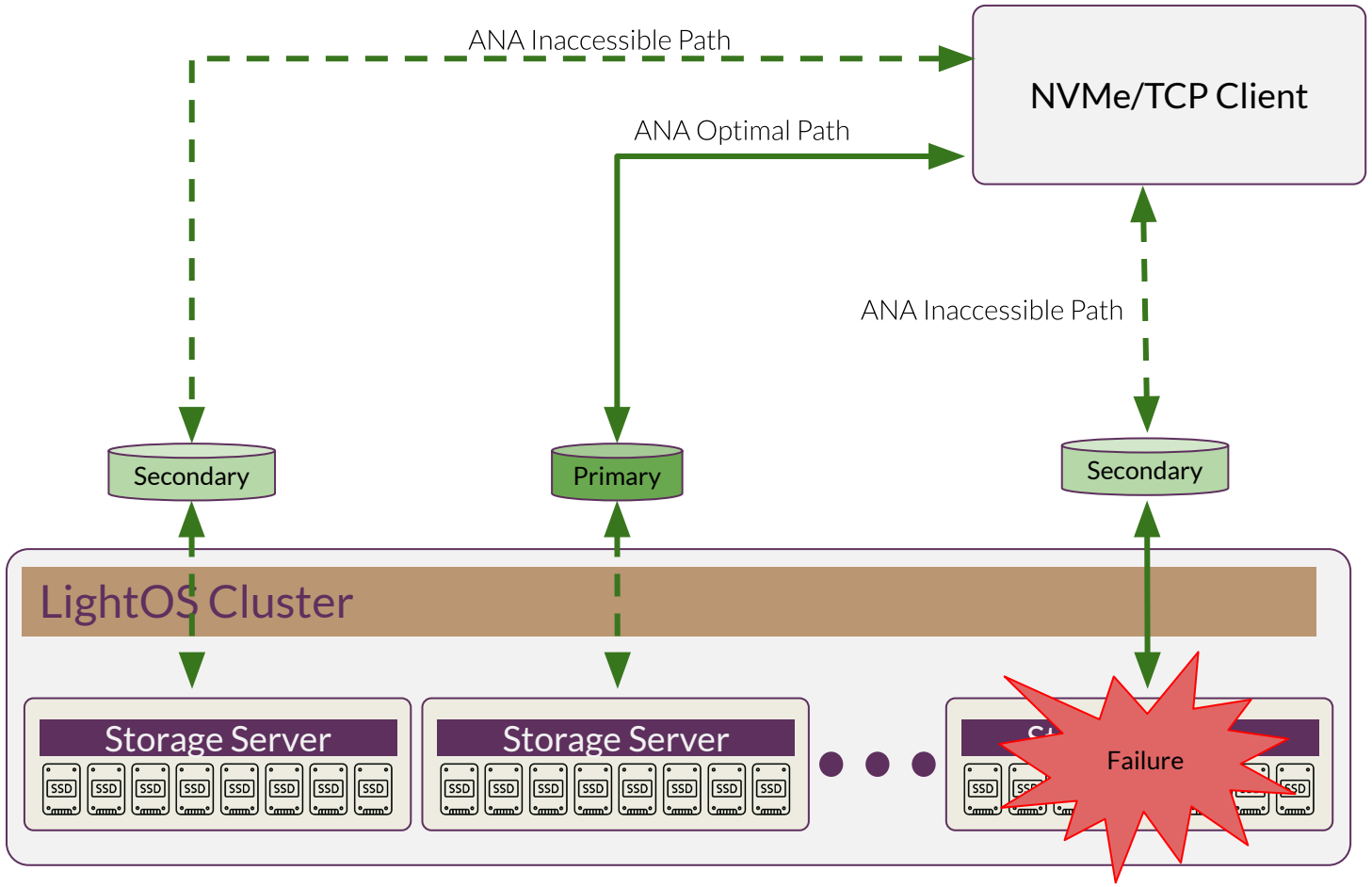
NVMe Asymmetric Namespace Access (ANA)

- NVMe Multipath IO defines access to NVMe namespace across two or more NVMe controllers (represent network paths).
- Multipath Namespace Access schemes:
  - **Symmetric**: All paths are equal
  - **Asymmetric**: Path state informs on access semantics
    - Optimized: Preferred accessible path
    - Non-Optimized: Non-preferred accessible path
    - Inaccessible
- **LightOS Leverages NVMe ANA for Clustering**
  - Primary server reports "Optimized" ANA State
  - Secondary servers report "Inaccessible" ANA State
  - **Failure Handling**: Controller changes state in ANA report



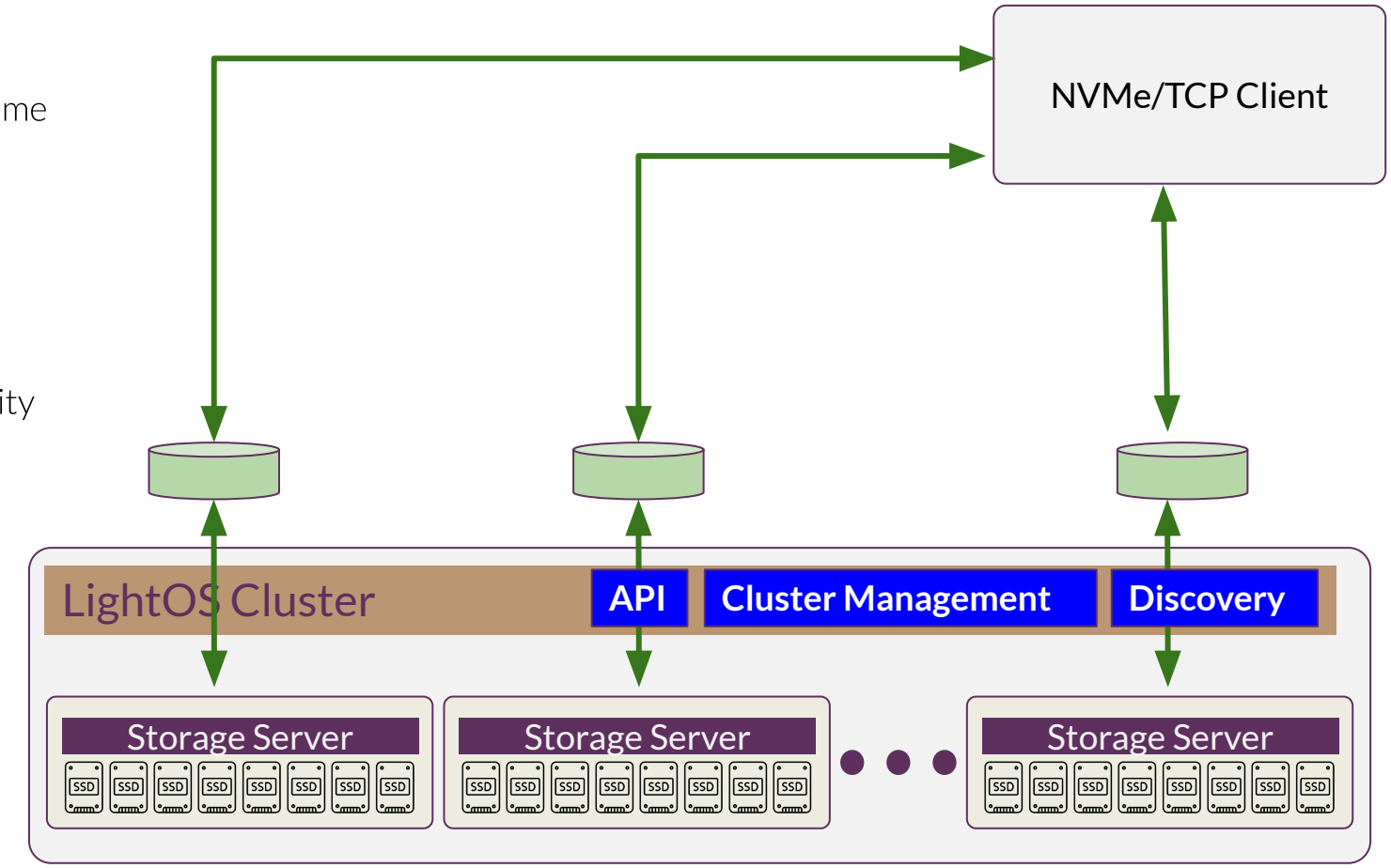
NVMe Asymmetric Namespace Access (ANA)

- NVMe Multipath IO defines access to NVMe namespace across two or more NVMe controllers (represent network paths).
- Multipath Namespace Access schemes:
  - **Symmetric**: All paths are equal
  - **Asymmetric**: Path state informs on access semantics
    - Optimized: Preferred accessible path
    - Non-Optimized: Non-preferred accessible path
    - Inaccessible
- **LightOS Leverages NVMe ANA for Clustering**
  - Primary server reports “Optimized” ANA State
  - Secondary servers report “Inaccessible” ANA State
  - **Failure Handling**: Controller changes state in ANA report



Clustering Services

- API Service
  - REST API service for cluster control and volume definitions
  - lbcli - command line utility
- Cluster Management
  - Managing cluster operation, configuration.
  - Monitoring, failure recovery orchestration
  - Replicas management, volume placing, capacity balancing, primary and secondary servers allocation.
- Discovery Service
  - Informs clients on accessible cluster volumes
  - NVMeoF standard
- All services are fault-tolerant



Demo: Cluster Status

- Status of storage servers:
  - state, IP address, failure domains, rebuild status

```
-bash-4.2# lbcli list nodes
Name         UUID                                   State    NVMe endpoint      Failure domains          Local rebuild progress
server00-0   174d6fcd-42f2-4c7a-834d-899045d2c7dc   Active   10.17.124.4:4420   [rack06-server52-vm05]   None
server01-0   c162151f-3869-4468-bb79-d5f08385c3d9   Active   10.17.124.5:4420   [rack03-server69-vm07]   None
server02-0   fdd672fc-394e-4080-bf5a-af2bcbc6aae7   Active   10.17.124.7:4420   [rack01-server64-vm07]   None
```

- Status of NVMe devices:
  - name, size, serial number, model, server

```
[root@rack12-server03 ~]# lbcli list nvme-devices
NAME      SIZE           NUMA-ID   SERIAL              MODEL               SERVER-UUID
nvme0n1   1000204886016  0         PHLF728500U21P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme4n1   1000204886016  0         PHLF728500UD1P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme8n1   1000204886016  0         PHLF728500UP1P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme6n1   1000204886016  0         PHLF728500VL1P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme5n1   1000204886016  0         PHLF728500WS1P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme9n1   1000204886016  0         PHLF728500XT1P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme1n1   1000204886016  0         PHLF728500Z41P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme7n1   1000204886016  0         PHLF728500ZS1P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme3n1   1000204886016  0         PHLF728501131P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme2n1   1000204886016  0         PHLF7285015J1P0GGN  INTEL SSDPE2KX010T7  9b62f6af-3983-55e4-b6b0-7757b7358897
nvme7n1   1000204886016  0         PHLF7253005S1P0GGN  INTEL SSDPE2KX010T7  d9f0fdf6-37b1-536b-8cf6-d95a13baeaf7
nvme1n1   1000204886016  0         PHLF725300AL1P0GGN  INTEL SSDPE2KX010T7  d9f0fdf6-37b1-536b-8cf6-d95a13baeaf7
```

Demo: Creating Volume

- Volume creation command:
  - Name, compression, replicas, size, ACL

```
-bash-4.2# lbcli create volume --name=demo_vol --compression=true --acl=demo --replica-count=3 --size=1Mib
Name       UUID                                 State      Protection State   NSID      Size      Replicas   Compression   ACL              Rebuild Progress
demo_vol   9b1aaf6a-e7d6-40ba-a8c6-1081af1ffb77   Creating   NotAvailable       0         1.0 MiB   3          true          values:"demo"
```

- Listing volumes:
  - Name, protection state, size, replicas, rebuild progress

```
-bash-4.2# lbcli list volumes
Name                                          UUID                                 State     Protection State   NSID      Size      Replicas   Compression   ACL                Rebuild Progress
default_volume_name_1566e9a7-8d54-4cdb-ba63-cfe70d650931   18cde8fc-2cfb-42c3-8ce7-259f66f4349c   Created   FullyProtected     1         4.0 GiB   3          false         values:"hostnqn1"   None
demo_vol                                      9b1aaf6a-e7d6-40ba-a8c6-1081af1ffb77   Created   FullyProtected     2         1.0 MiB   3          true          values:"demo"      None
```

Demo: Secondary Replica Disconnection

- Disconnecting server storing secondary replica
  - Automatically detected in several seconds
- Server state marked as “inactive”
- Path state updated

Client:

```
[root@rack12-server51 demo]# nvme list-subsys /dev/nvme4n2

nvme-subsys4 -
NQN=nqn.2014-08.org.nvmexpress:NVMf:uuid:00000000-0000-0000-0000-000000000000

\
 +- nvme4 tcp traddr=10.23.20.1 trsvcid=4420 live optimized
 +- nvme5 tcp traddr=10.23.20.2 trsvcid=4420 live inaccessible
 +- nvme6 tcp traddr=10.23.20.3 trsvcid=4420 live inaccessible


[root@rack12-server51 demo]# nvme list-subsys /dev/nvme4n2

nvme-subsys4 -
NQN=nqn.2014-08.org.nvmexpress:NVMf:uuid:00000000-0000-0000-0000-000000000000

\
 +- nvme4 tcp traddr=10.23.20.1 trsvcid=4420 live optimized
 +- nvme5 tcp traddr=10.23.20.2 trsvcid=4420 live inaccessible
 +- nvme6 tcp traddr=10.23.20.3 trsvcid=4420 connecting inaccessible
```

Initial state

Missing

Cluster:

```
[root@rack12-server03 demo]# lbcli list peers
NAME       UUID                                   State      NVME-Endpoint     Failure-Domains                 In-Local-Rebuild   Local-Rebuild-Progress
node00-0   14cff509-96a5-5e01-98bd-1a5751156fcf   Active     10.23.20.1:4420   [u'rack12-server03' u'nod...]   false              0
node01-0   4e21ff8a-d39a-522b-9401-06025c630f1e   Active     10.23.20.2:4420   [u'rack02-server59' u'nod...]   false              0
node02-0   8a8ab8cb-24d5-5966-b33a-00a3c3bca81e   Inactive   10.23.20.3:4420   [u'rack09-server94' u'nod...]   false              0
```

Demo: Primary Replica Disconnection

- Disconnecting server storing primary replica
  - Automatically detected in several seconds
- Server state marked as “inactive”
- Path state updated

Client:

```
[root@rack12-server51 ~]# nvme list-subsys /dev/nvme4n1
nvme-subsys4 - NQN=nqn.2014-08.org.nvmexpress:NVMf:uuid:00000000-0000-0000-0000-000000000000
\
 +- nvme4 tcp traddr=10.23.20.1 trsvcid=4420 live inaccessible
 +- nvme5 tcp traddr=10.23.20.2 trsvcid=4420 live inaccessible
 +- nvme6 tcp traddr=10.23.20.3 trsvcid=4420 live optimized

[root@rack12-server51 ~]# nvme list-subsys /dev/nvme4n1
nvme-subsys4 - NQN=nqn.2014-08.org.nvmexpress:NVMf:uuid:00000000-0000-0000-0000-000000000000
\
 +- nvme4 tcp traddr=10.23.20.1 trsvcid=4420 live inaccessible
 +- nvme5 tcp traddr=10.23.20.2 trsvcid=4420 live optimized
 +- nvme6 tcp traddr=10.23.20.3 trsvcid=4420 connecting optimized
```

Cluster:

```
[root@rack12-server03 demo]# lbcli list peers
NAME       UUID                                   State      NVME-Endpoint     Failure-Domains                  In-Local-Rebuild   Local-Rebuild-Progress
node00-0   14cff509-96a5-5e01-98bd-1a5751156fcf   Active     10.23.20.1:4420   [u'rack12-server03' u'nod...]   false              0
node01-0   4e21ff8a-d39a-522b-9401-06025c630f1e   Active     10.23.20.2:4420   [u'rack02-server59' u'nod...]   false              0
node02-0   8a8ab8cb-24d5-5966-b33a-00a3c3bca41e   Inactive   10.23.20.3:4420   [u'rack09-server94' u'nod...]   false              0
[root@rack12-server03 demo]#
```

Demo: Replica Re-connection and Catch-up

- Replica is “behind”
- Partial rebuild - only missing data (that was written during the failure) is resent
- Replica catches up

Cluster:

```
[root@rack12-server03 demo]# lbcli list peers
NAME      UUID                                  State       NVME-Endpoint     Failure-Domains                 In-Local-Rebuild   Local-Rebuild-Progress
node00-0  14cff509-96a5-5e01-98bd-1a5751156fcf  Active      10.23.20.1:4420   [u'rack12-server03' u'nod...]   false              0
node01-0  4e21ff8a-d39a-522b-9401-06025c630f1e  Active      10.23.20.2:4420   [u'rack02-server59' u'nod...]   false              0
node02-0  8a8ab8cb-24d5-5966-b33a-00a3c3bcae1e  Activating  10.23.20.3:4420   [u'rack09-server94' u'nod...]   false              0
```

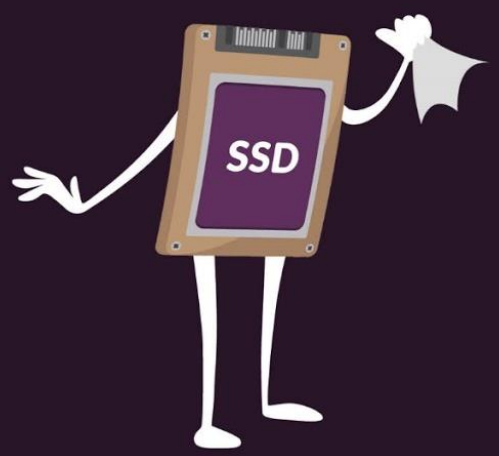
```
[root@rack12-server03 demo]# lbcli list volumes
Name           UUID State     NSID      Total Capacity   Number Replications   Minimum Replications   Compression Disabled   ACL          Rebuild
Progress
server51_vol1  b    Created   1         1.0 TiB          3                     1                      false                  values:"1"   None
server51_vol2  c    Created   5         1.0 TiB          3                     1                      false                  values:"1"   20
```

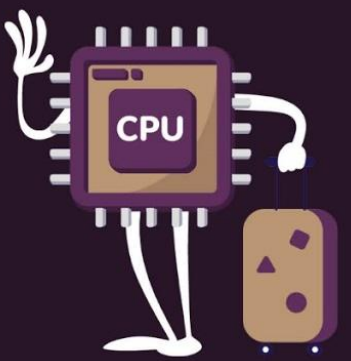
Summary

- Direct attached storage to disaggregated storage architecture.
- LightOS v1.x: first NVMe/TCP based storage server with data services.
- LightOS v2.x: first NVMe/TCP based clustering solution for fault-tolerance in a storage server level.
- Volume replicas assignment considering failure domains.
- Data flow and server failover using NVMe multipath with ANA.
- Clustering services and management capabilities.

Contact information:
www.lightbitslabs.com
alex@lightbitslabs.com

Thank you!





Backup slides

Client Configuration

- Standard upstream kernel
  - Available in RHEL/CentOS 8.1 and Ubuntu 19.10
  - Available via ELREPO and HWE
- Standard ‘nvme-cli’
- Client ID - hostnqn, similar to iSCSI IQN
- Automation of connection during boot is possible

```
[root@rack12-server51 ~]# rpm -qa | grep nvme
nvme-cli-1.9-1.x86_64
[root@rack12-server51 ~]# uname -r
5.4.2-1.el7.elrepo.x86_64
[root@rack12-server51 ~]#
```

```
[root@rack12-server51 ~]# cat /etc/nvme/hostnqn
nqn.2014-08.org.nvmexpress:uuid:00b84736-0694-4924-903a-41fea1f0ad4a
[root@rack12-server51 ~]#
```

Client Connections

- Connecting to all cluster nodes
  - Discovery service allows connection using a single command

```
#!/bin/bash
IPS="10.23.20.1 10.23.20.2 10.23.20.3"
for IP in ${IPS}
do
     nvme connect -t tcp -s 4420 -a ${IP} \
     --ctrl-loss-tmo -1 \
     -n nqn.2014-08.org.nvmexpress:NVMf:uuid:00000000-0000-0000-0000-000000000000 \
     -q nqn.2014-08.org.nvmexpress:uuid:00b84736-0694-4924-903a-41fea1f0ad4a
done
```

Client Connection Status

- Device is visible
  - via ‘`lsblk`’ and ‘`nvme list`’
- Connections, active (`optimized`) and backup (`inaccessbile`) are visible
  - via ‘`nvme list-subsys <dev>`’

```
[root@rack12-server51 demo]# nvme list
Node             SN                   Model                                    Namespace Usage                      Format           FW Rev
---------------- -------------------- ---------------------------------------- --------- -------------------------- ---------------- --------
/dev/nvme4n1     6659d524cbb88a8a     LightBox                                 3           1.10  TB /   1.10  TB      4 KiB +  0 B   1.0
```

```
[root@rack12-server51 demo]# nvme list-subsys /dev/nvme4n1
nvme-subsys4 - NQN=nqn.2014-08.org.nvmexpress:NVMf:uuid:00000000-0000-0000-0000-000000000000
\
 +- nvme4 tcp traddr=10.23.20.1 trsvcid=4420 live inaccessible
 +- nvme5 tcp traddr=10.23.20.2 trsvcid=4420 live inaccessible
 +- nvme6 tcp traddr=10.23.20.3 trsvcid=4420 live optimized
[root@rack12-server51 demo]#
```

Write Traffic

- Running random-write from a single client
- One storage server shows in/out traffic
  - 23 Gbs in, 46 Gbs out
- Two storage servers show only incoming traffic
  - 22 Gbs in, 42 Mbs out

In: 23.35 GBit/s
Out: 46.53 GBit/s
Active Volume

In: 22.73 GBit/s
Out: 40.75 MBit/s
Replica

In: 22.33 GBit/s
Out: 42.33 MBit/s
Replica

Read Traffic

- Running random-read from a single client
- One storage server shows outgoing traffic
  - 23 Gbs outgoing
- Two storage servers with secondary replicas are idle