Extended distance Oracle RAC: zero downtime, high speed, commodity hardware only

Artem Danielov
CTO, FlashGrid

NYOUG 2016 Summer General Meeting

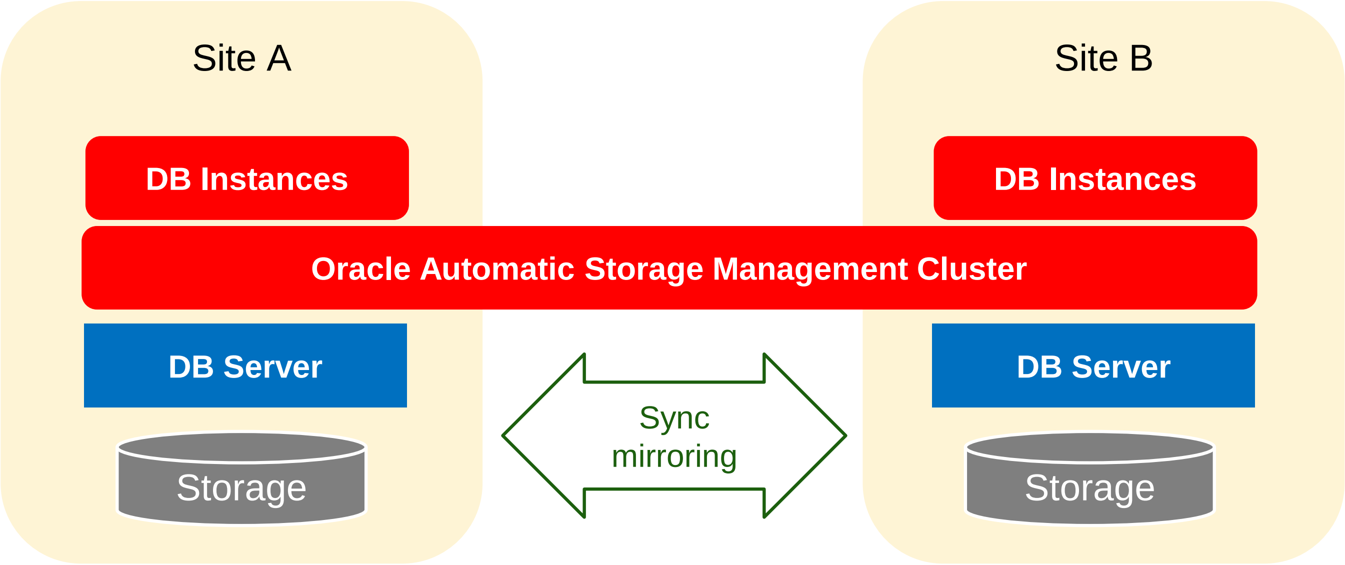
Agenda

- Why extended distance database cluster?
- Traditional approaches to storage in extended distance clusters
- Distributed storage without SAN
- NVMe SSDs vs. SAN
- Server hardware options
- Network options
- Impact of distance on performance
- ASM fail groups and quorum disks

Why Extended Distance Database Cluster?

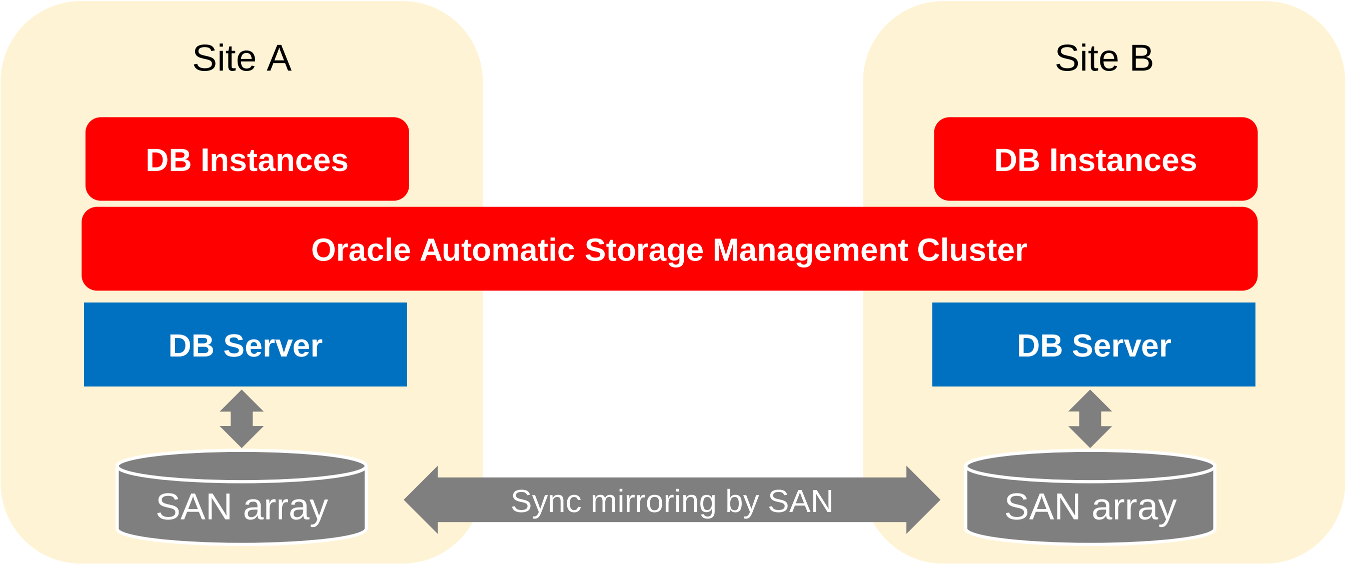
flashgrid

- Metro-scale disaster recovery
- Full copy of the data on each site
- Extremely fast recovery from site failure
- RAC, RAC One Node, or Single Instance with manual failover



Storage Mirroring by SAN

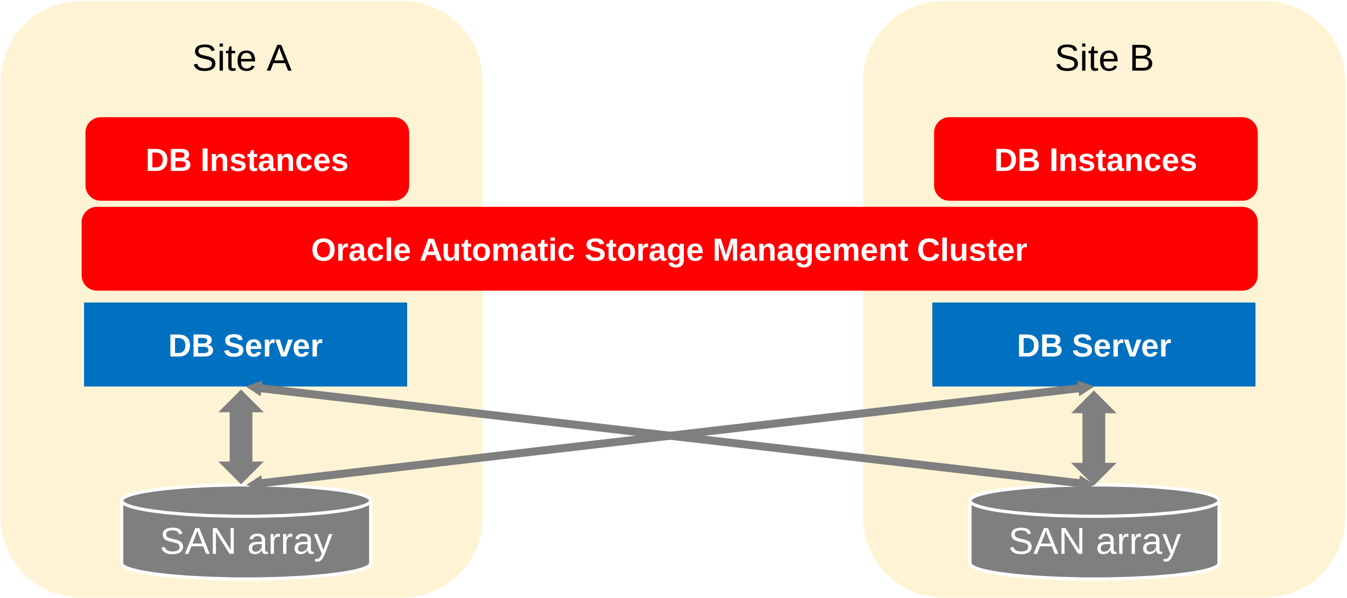
- SAN array on each site
- Sync mirroring done at the SAN array level
- DB servers connected to local SAN array only



Storage Mirroring by ASM

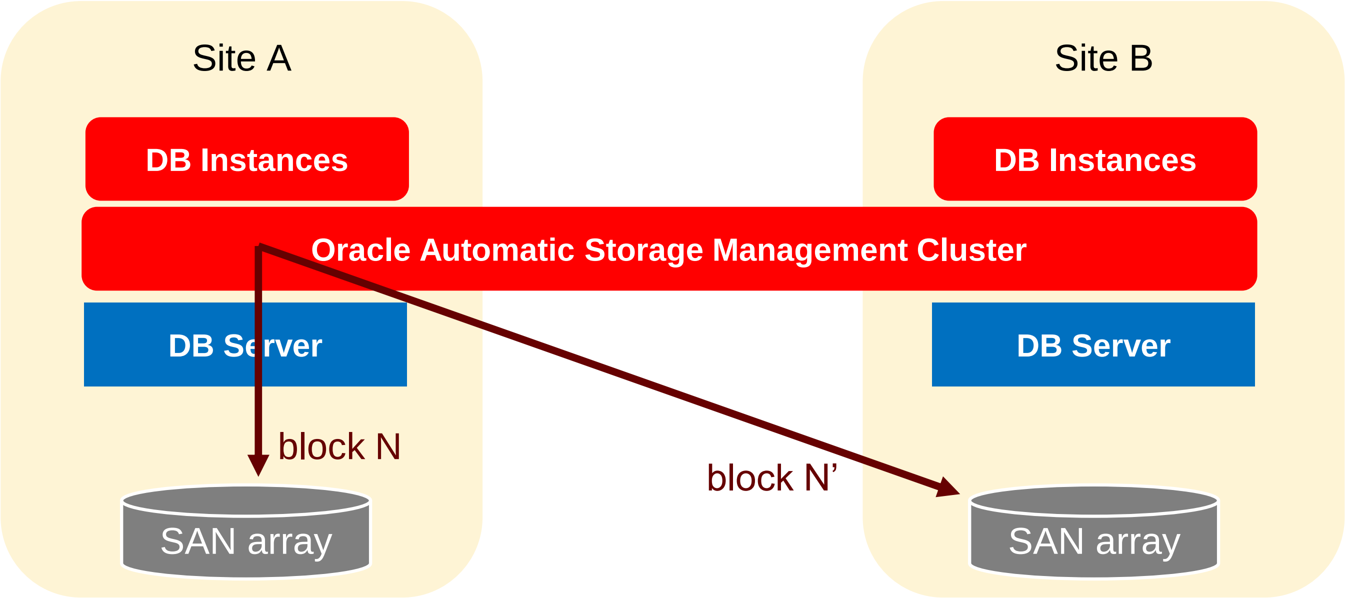
flashgrid

- SAN array on each site
- DB servers connected to both SAN arrays
- Sync mirroring done by ASM – each block of data written to both arrays



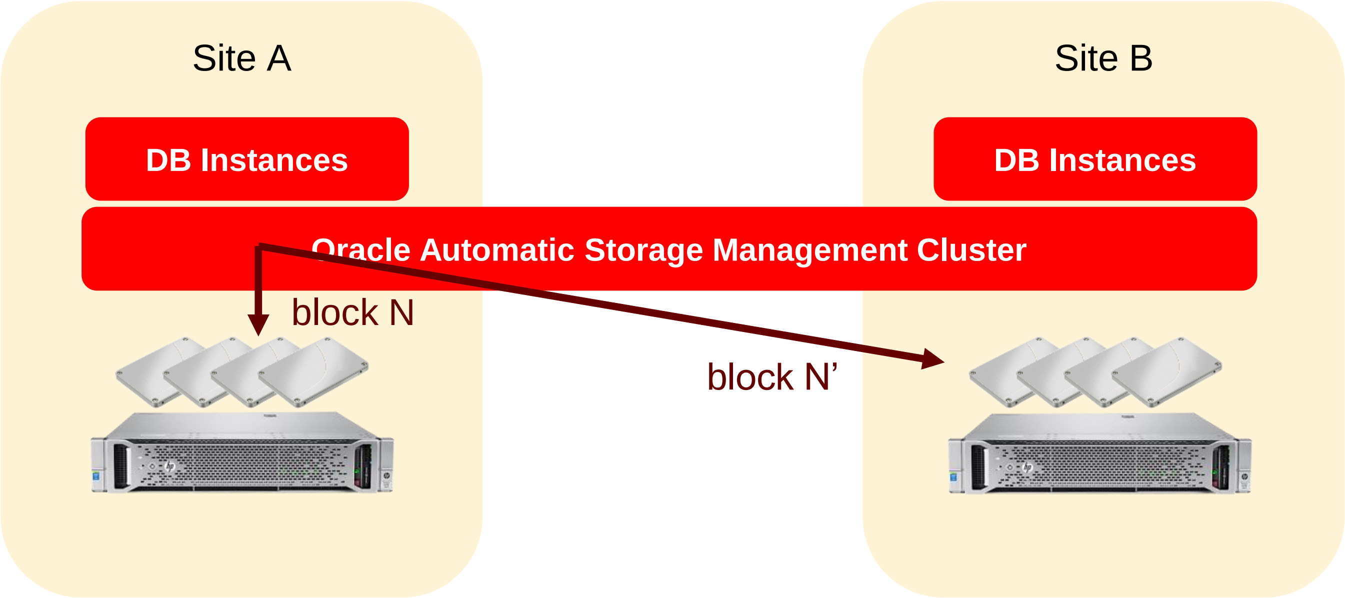
Storage Mirroring by ASM

- SAN array on each site
- DB servers connected to both SAN arrays
- Sync mirroring done by ASM – each block of data written to both arrays



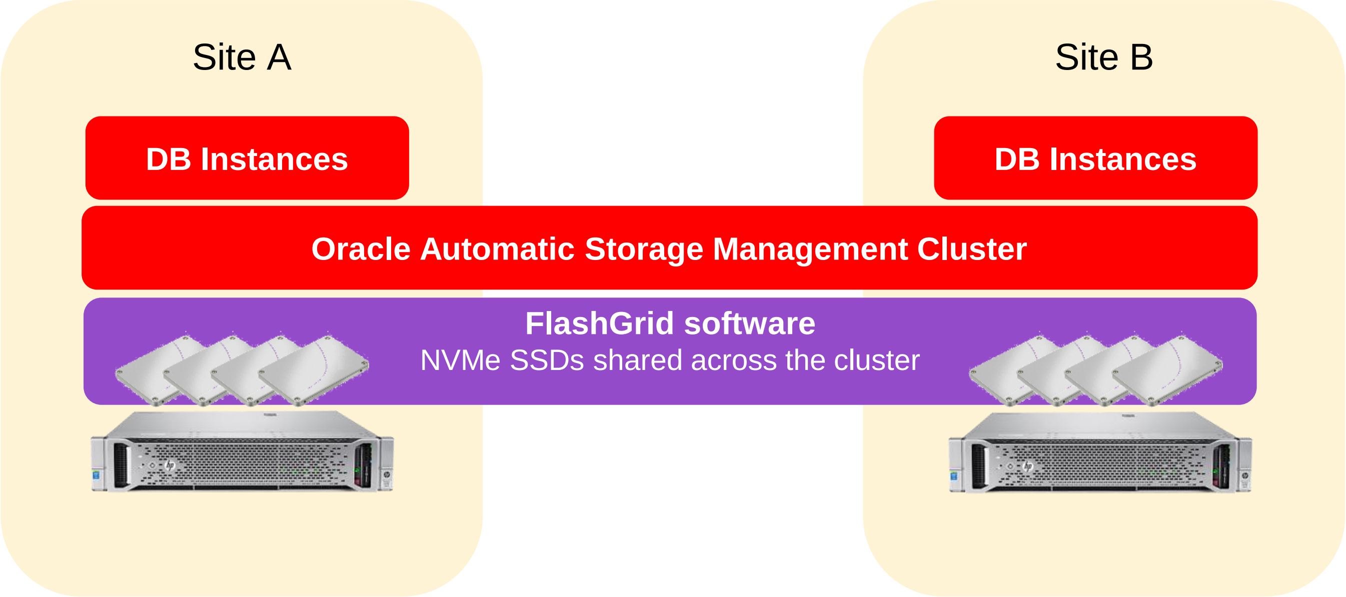
Storage Mirroring: FlashGrid software

- SSDs inside the servers
- Sync mirroring done by ASM – each block of data written to servers at both sites
- SSDs shared across sites by using FlashGrid software



Shared Access Using FlashGrid Software

- SSDs inside the servers
- Sync mirroring done by ASM – each block of data written to servers at both sites
- SSDs shared across sites by using FlashGrid software



Site A

Site B

DB Instances

DB Instances

Oracle Automatic Storage Management Cluster

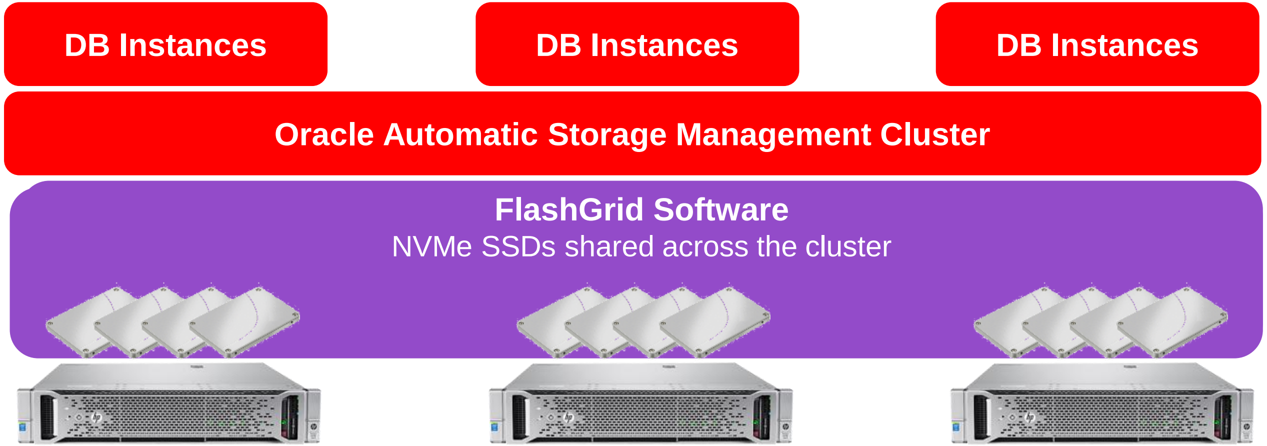
FlashGrid software

NVMe SSDs shared across the cluster

FlashGrid: Open Storage Software for Oracle Clusters

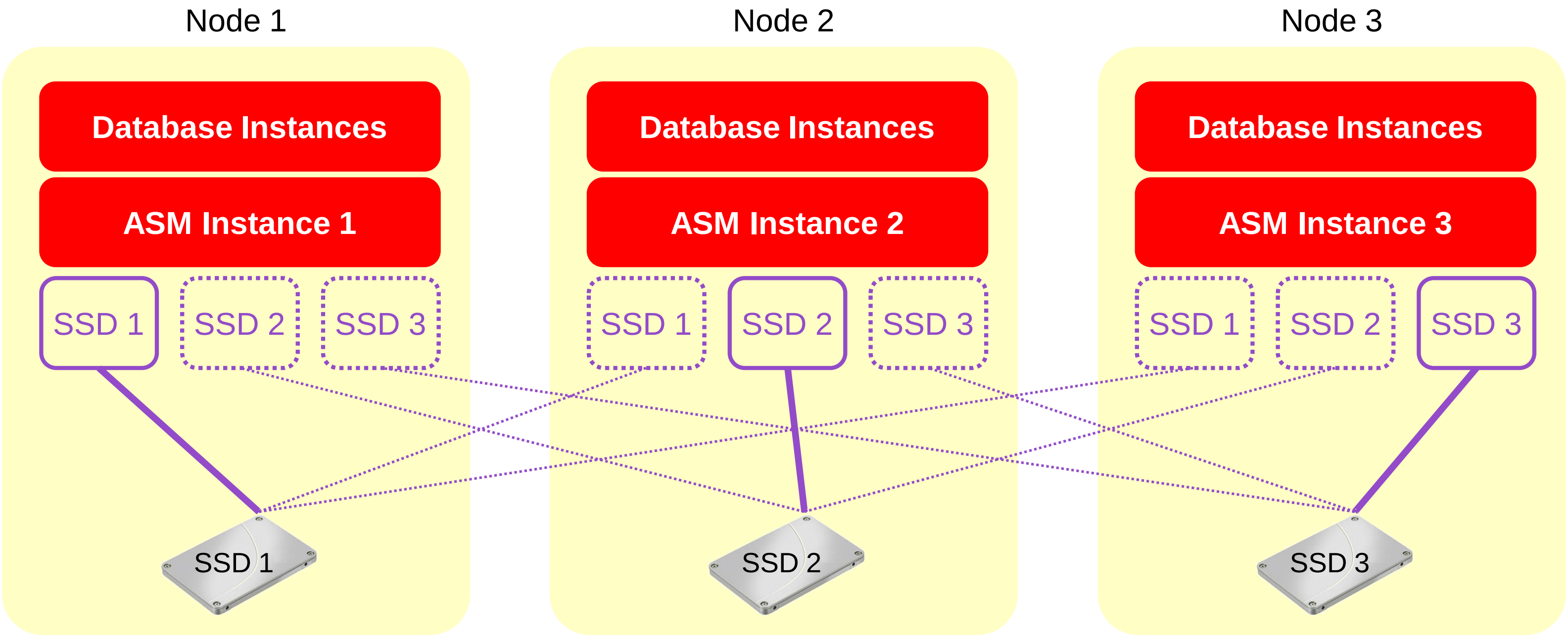
flashgrid

- Turn NVMe PCIe SSDs inside standard servers into scalable shared storage
- 0.4 to 76 TB per node; 2 to 100 nodes
- Leverage proven Oracle ASM for high availability and data mirroring
- Maximize database performance with FlashGrid Read-Local™ Technology
- RAC, RAC One Node, Single Instance, Enterprise or Standard Edition


DB Instances
DB Instances
DB Instances
Oracle Automatic Storage Management Cluster
FlashGrid Software
NVMe SSDs shared across the cluster

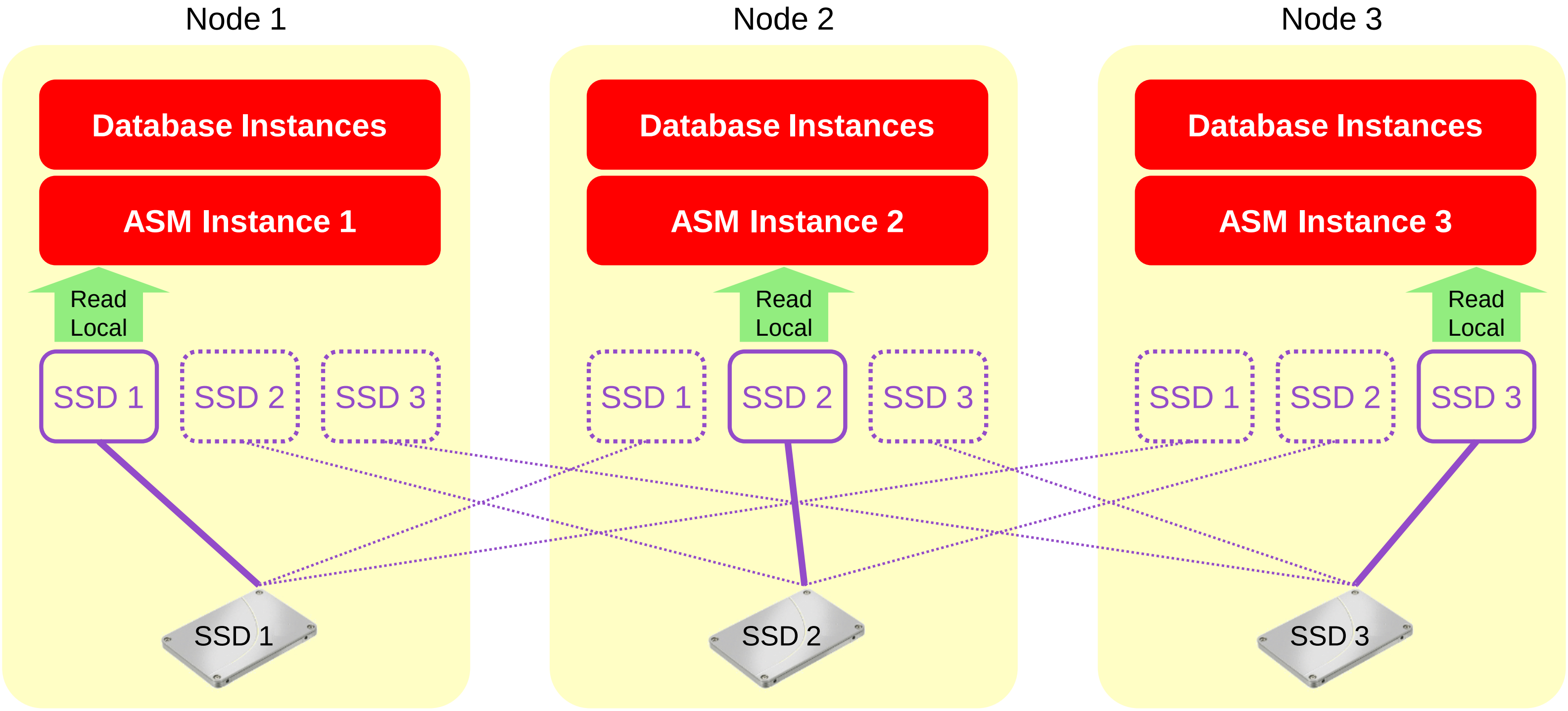
FlashGrid Architecture

flashgrid



- Oracle ASM manages data, volumes, mirroring, snapshots
- FlashGrid manages SSD devices and connections

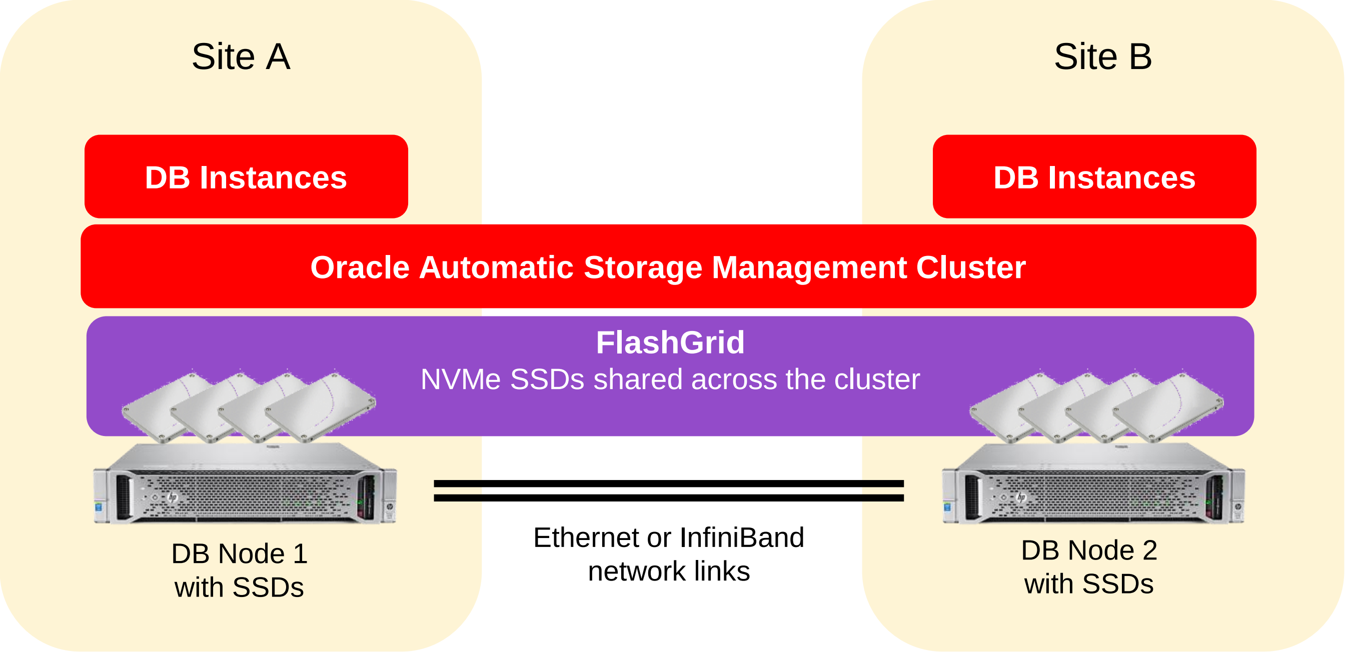
Prevent Network Congestion with FlashGrid Read-Local™ Technology



- Minimize network overhead by serving reads from local SSDs at the speed of PCIe
- Accelerate both reads and writes

Two Sites with FlashGrid Software

- HA/DR solution at metro scale
- Synchronous data mirroring across sites



Site A

Site B

DB Instances

DB Instances

Oracle Automatic Storage Management Cluster

FlashGrid

NVMe SSDs shared across the cluster

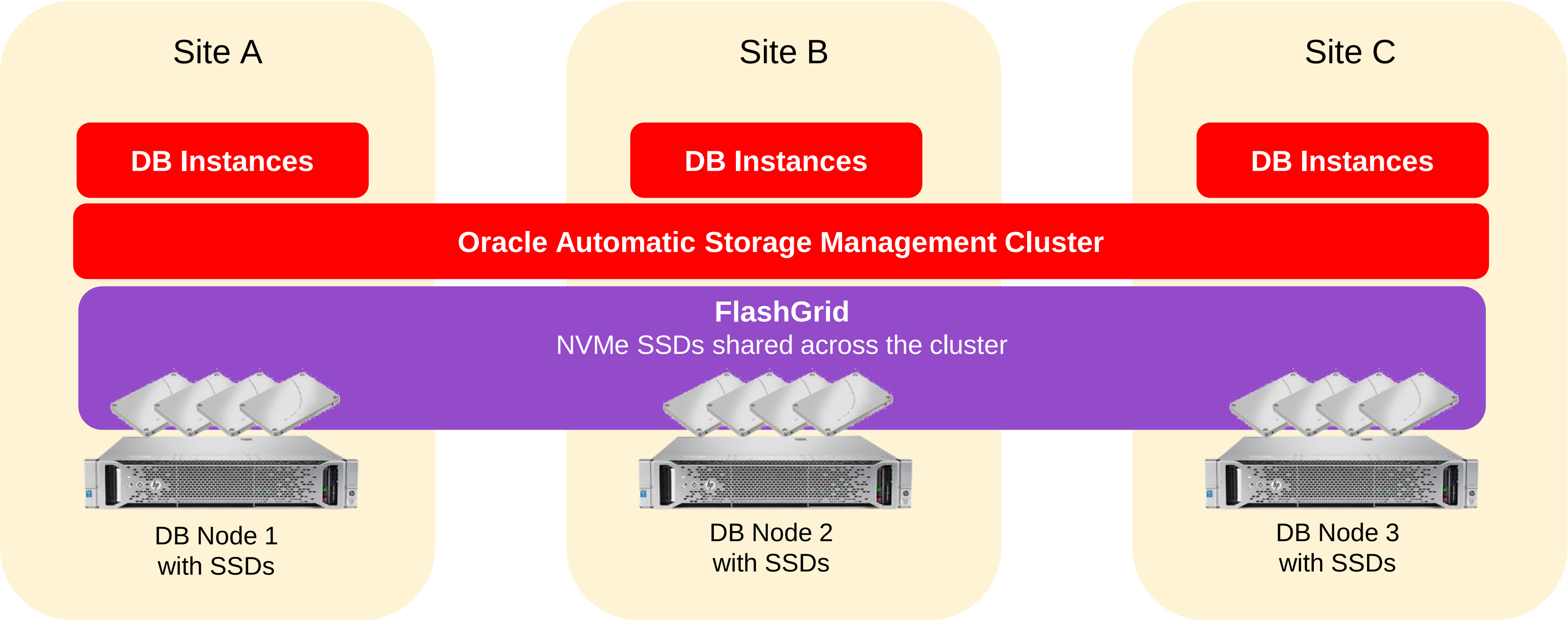
DB Node 1
with SSDs

Ethernet or InfiniBand
network links

DB Node 2
with SSDs

Three Sites with FlashGrid Software

- HA/DR solution at metro scale
- Synchronous data mirroring across sites



Ethernet or InfiniBand network links between each pair of sites

High-Availability Architecture

- Distributed storage with no single point of failure
- Data mirroring across sites with proven Oracle ASM
- Reduce risk of data corruption with only standard Linux components in data path
- Reduce backup and restore times with all-flash storage
- Reduce risks by using standard server hardware

NVMe – The New High-Performance Storage Standard

- Highly efficient replacement for legacy SCSI (FC, SAS, SATA) stack
- High IOPS, low latency, low CPU consumption
- Avaialble from all server and SSD vendors
- 2.5” hot-plug and add-in PCIe card form-factors
- Used in Oracle Exadata*

Performance of one NVMe SSD similar to a flash array

3 GB/s
400K IOPS

3 GB/s
250K IOPS

More bandwidth for full table scans and backups

24 GB/s

with 8 NVMe SSDs
inside 2 database servers

+ FlashGrid software

3 GB/s

NVMe-Optimized Server Options

	2.5” hot-plug NVMe SSDs			Add-in PCIe card NVMe SSDs		Max total NVMe flash capacity per server
Server model	# slots	Max capacity per SSD	Max capacity per server with 2.5” NVMe SSDs	# PCIe slots available for NVMe SSDs	Max flash capacity per server with 6.4TB add-in card SSDs	
Oracle Server X6-2L	9	3.2 TB	28.8 TB	5	32 TB	60.8 TB
Oracle Server X6-2	4	3.2 TB	12.8 TB	3	19.2 TB	32 TB
Dell PowerEdge R730xd	4	3.2 TB	12.8 TB	5	32 TB	44.8 TB
Dell PowerEdge R930	8	3.2 TB	25.6 TB	9	57.6 TB	83.2 TB
Dell PowerEdge R630	4	3.2 TB	12.8 TB	2	12.8 TB	25.6 TB
HPE ProLiant DL380 Gen9	6	2 TB	12 TB	5	32 TB	44 TB
HPE ProLiant DL560 Gen9	6	2 TB	12 TB	6	38.4 TB	50.4 TB
HPE ProLiant DL580 Gen9	5	2 TB	10 TB	8	51.2 TB	61.2 TB
Supermicro 1028U-TN10RT+	10	3.2 TB	32 TB	2	12.8 TB	44.8 TB
Supermicro 2028U-TN24R4T+	24	3.2 TB	76.8 TB	2	12.8 TB	89.6 TB
Supermicro 2028R-NR48N	48	3.2 TB	153.6 TB	2	12.8 TB	166.4 TB

Network Options for Extended Distance Clusters

- 10 GbE recommended
- 1 GbE possible in low load use-cases
- Long-haul InfiniBand available, up to 50 miles

- RAC One Node or Single-Instance: storage and Oracle Private Network can share the network links
- RAC: separate network links recommended for storage and Oracle Private Network

- Redundant connectivity between sites
- Multicast required
- No routing

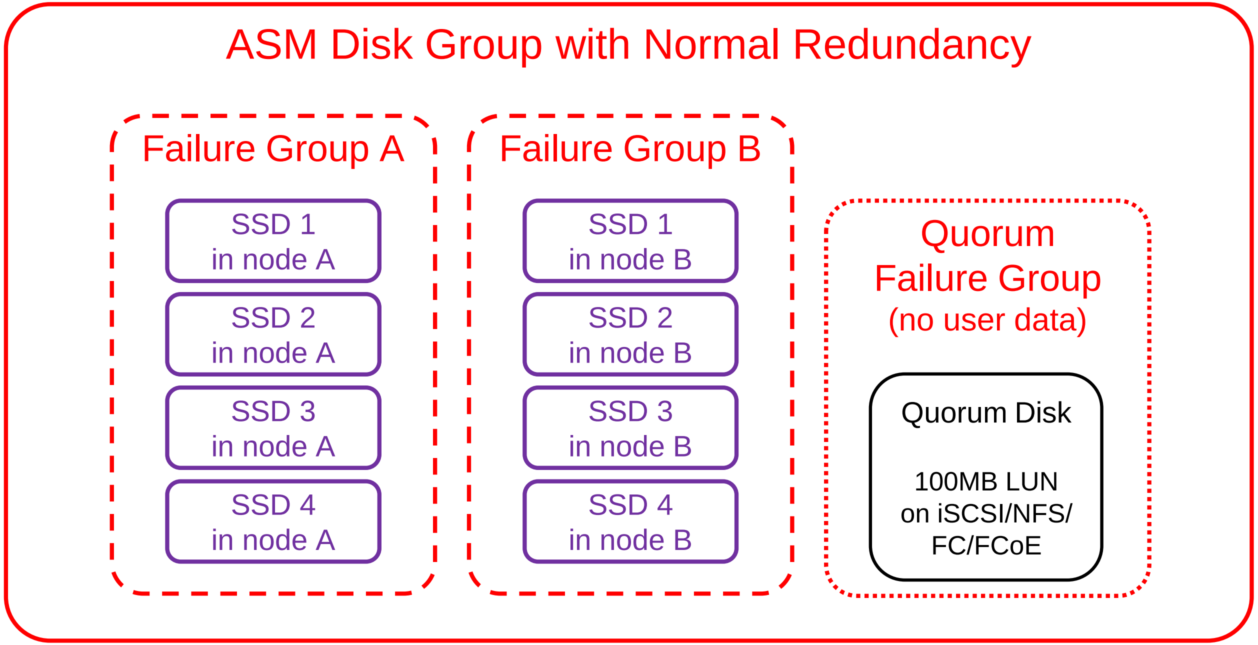
Impact of Distance on Performance

- Extra storage latency: 0.005 ms / km
- 100 km: 0.5 ms extra latency – smaller than typical SAN latency
- Only write latency increased, reads are local

- RAC One Node or Single-Instance: no Cache Fusion to worry about
- RAC: extra latency has impact on Cache Fusion, more careful assessment needed

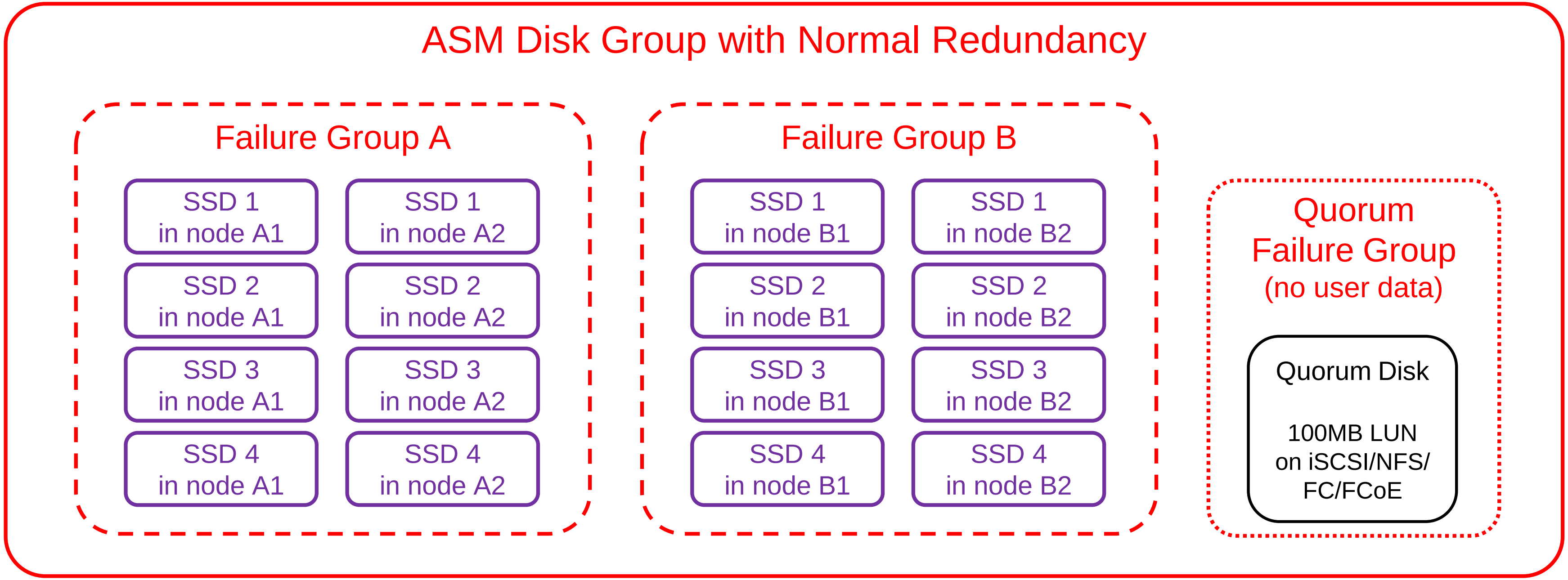
- Higher storage write latency increases transaction completion time, but little impact on the number of concurrent transactions

ASM Fail Groups: 2 sites



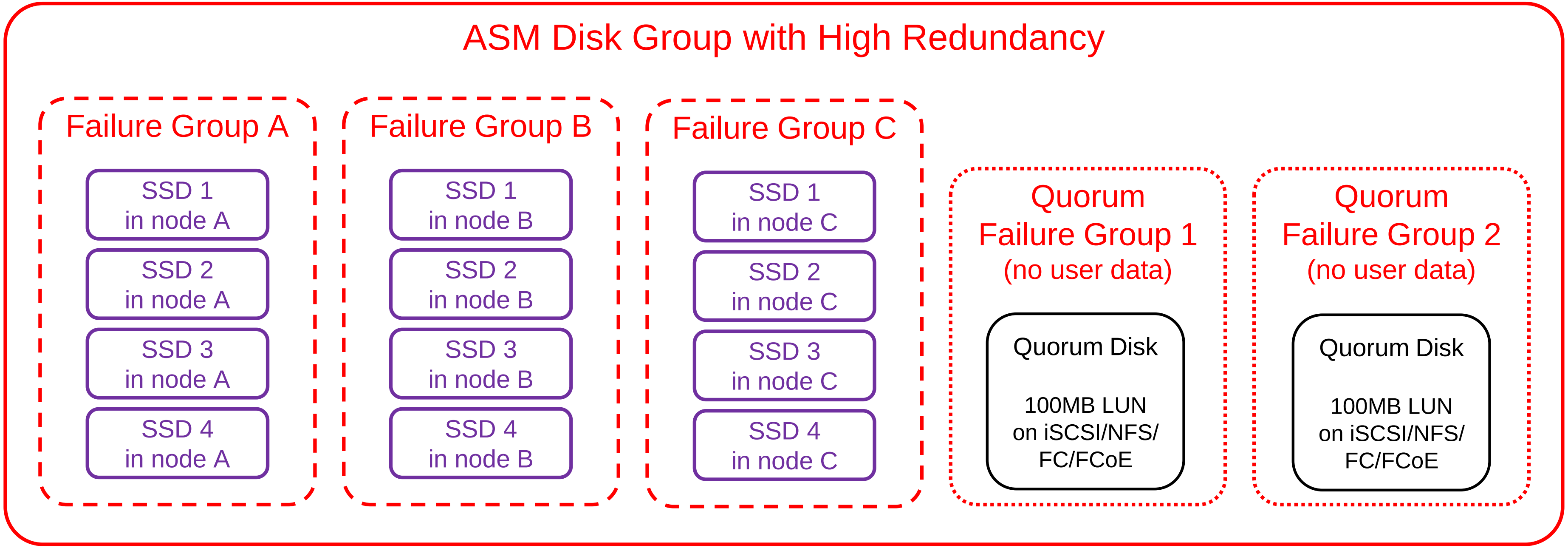
- One fail group per site
- One quorum disk per disk group required at a 3rd site
- For extra HA, two quorum disks at two independent sites possible
- Quorum Disk: 100 MB on iSCSI/NFS/FC/FCoE storage, very low performance needed (~1 IOPS)

ASM Fail Groups: 2 sites, 2 nodes / site



- One fail group per site
- One quorum disk per disk group required at a 3rd site
- For extra HA, two quorum disks at two independent sites possible
- Quorum Disk: 100 MB on iSCSI/NFS/FC/FCoE storage, very low performance needed (~1 IOPS)

ASM Fail Groups: 3 sites



ASM Disk Group with High Redundancy

Failure Group A

SSD 1 in node A

SSD 2 in node A

SSD 3 in node A

SSD 4 in node A

Failure Group B

SSD 1 in node B

SSD 2 in node B

SSD 3 in node B

SSD 4 in node B

Failure Group C

SSD 1 in node C

SSD 2 in node C

SSD 3 in node C

SSD 4 in node C

Quorum Failure Group 1 (no user data)

Quorum Disk

100MB LUN on iSCSI/NFS/ FC/FCoE

Quorum Failure Group 2 (no user data)

Quorum Disk

100MB LUN on iSCSI/NFS/ FC/FCoE

- One fail group per site
- Two quorum disks per disk group required at two independent sites
- Quorum Disks: 1GB LUN on iSCSI/NFS/FC/FCoE storage, very low performance needed (~1 IOPS)

Questions?

www.flashgrid.io/solutions_stretched_clusters/

artem AT flashgrid DOT io