SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

Synchronous DR over IP fabrics with NetApp MetroCluster-IP

Vijay Singh
NetApp

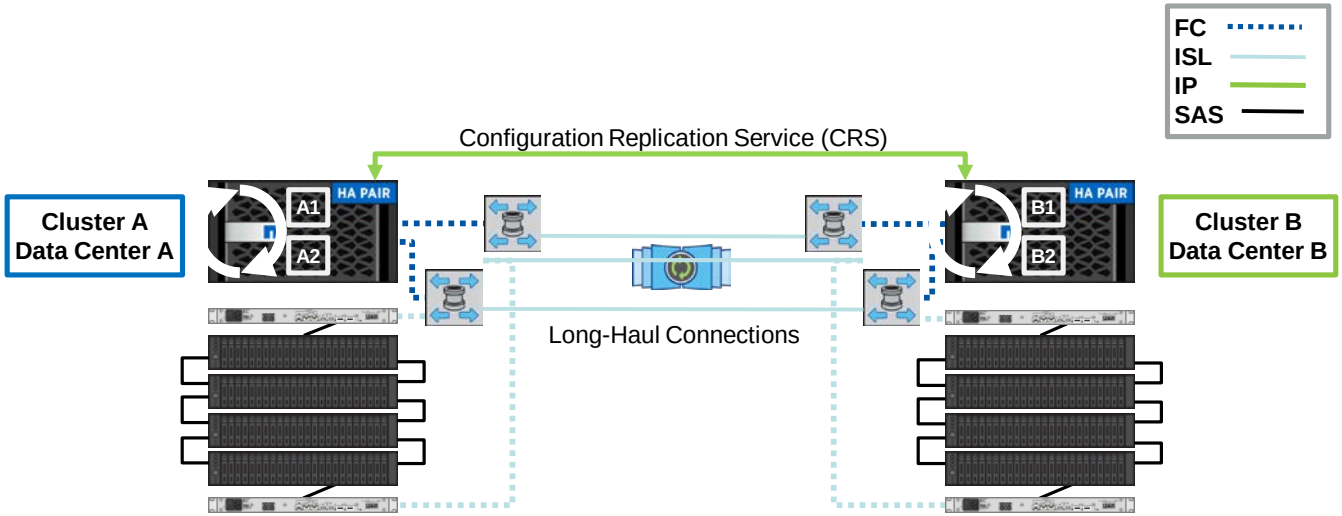
Agenda

1) What is MetroCluster
2) MetroCluster FC architecture
3) Why MetroCluster IP
4) New Architecture
   - RDMA Details
   - Storage Details
5) Network Topology
6) Security
7) Performance

What is MetroCluster

1) NetApp is a data storage provider of Primary & Secondary Storage
2) NetApp does custom, ODM & OEM hardware, as well as software defined
3) ONTAP operating system on everything
4) ONTAP provides integrated synchronous disaster recovery (DR)
5) Clustered ONTAP MetroCluster (MCC) has local HA, remote DR
6) 2-cluster solution, each cluster 1 - N nodes (symmetric config today)
7) Peer Clusters, Mirror Storage and Go
8) Single button Switch-Over, Switch-Back

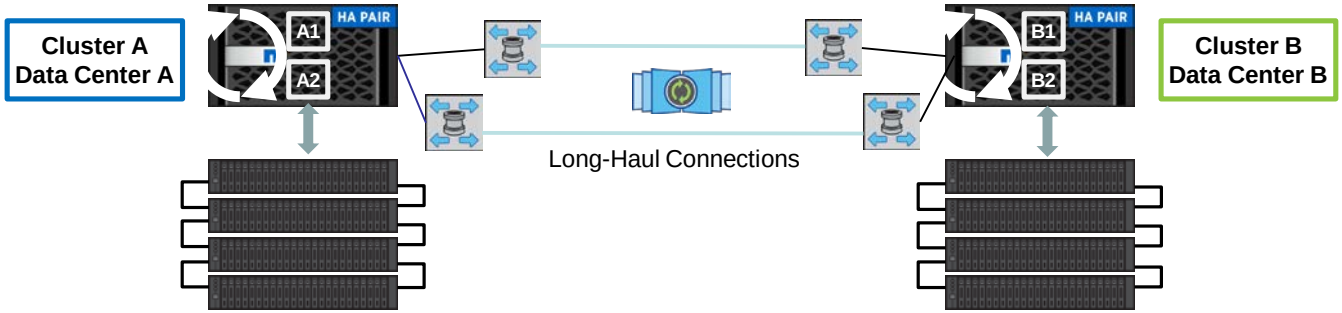
Fibre-Channel MetroCluster



Why IP / Ethernet

1) Ethernet speeds vs FiberChannel speeds
2) Port limited platforms, multiple applications sharing ports
3) Integrated storage platforms, no physical remote access available
4) Ease of deployment, number of wires, cost
5) Cost of operations with shared inter-site links
6) Allows to do software defined
7) Cloud ready, DR to Cloud, DRaaS

IP MetroCluster



MetroCluster IP: Major Architectural Parts

1) Ethernet Switches, use the switches for Clustering
2) Inter-Site Links (ISLs) – sized for workload, needed redundancy
3) Two isolated fabrics – use VLANs from the switch
4) Ethernet / IP RDMA for synchronous mirroring of in-flight data
   - RoCE or iWARP
5) Head-as-the-Target storage
   - Present disks/namespaces/slices as LUNs behind iSCSI target
6) RDMA & iSCSI sharing the same network interface(s)

MCC-IP Architecture – Components

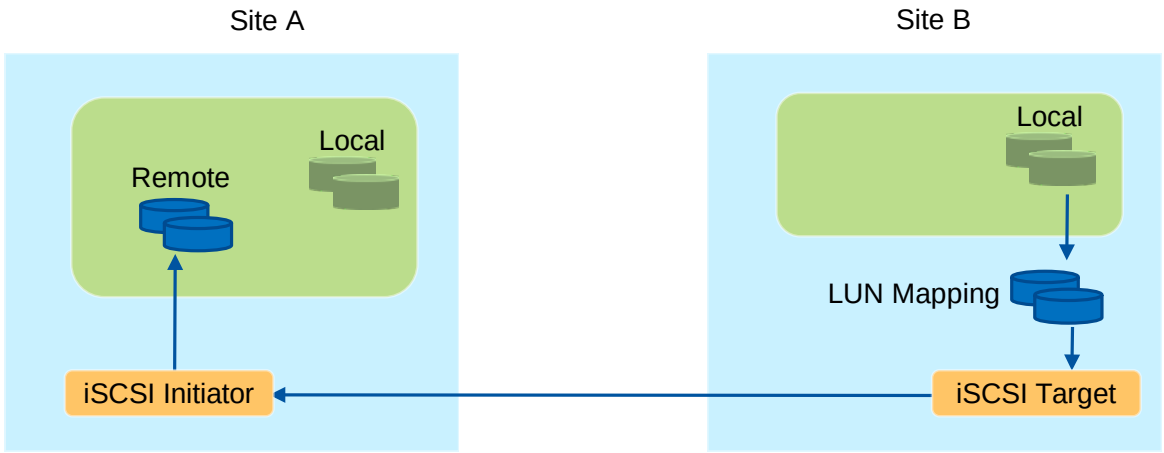
- All Replication Planes over TCP/IP
  1. CRS / TCP
  2. NVLOG / TCP : iWARP
  3. Raid SyncMirror / TCP: iSCSI

- Head acts as Proxy to IP Fabric
  1. iSCSI Initiator and Target
  2. NVMe over Fabric possible in the future

MetroCluster: RDMA

1) MCC architecture ensures client writes are logged to DR nodes
   - HA & DR sync replication – multi-mirror design
2) MCC-IP supports distances up to 700 km, client latency limited
   - FC only 300km
   - Needs end-to-end, hop-by-hop buffer credit management
3) We used FC-VI, VIA implementation of RDMA for FiberChannel
   - Limited choice of vendors
   - No software implementation
4) Debugging for FC is hard
   - Register dumps, FC analyzers

MetroCluster IP: Head-as-target



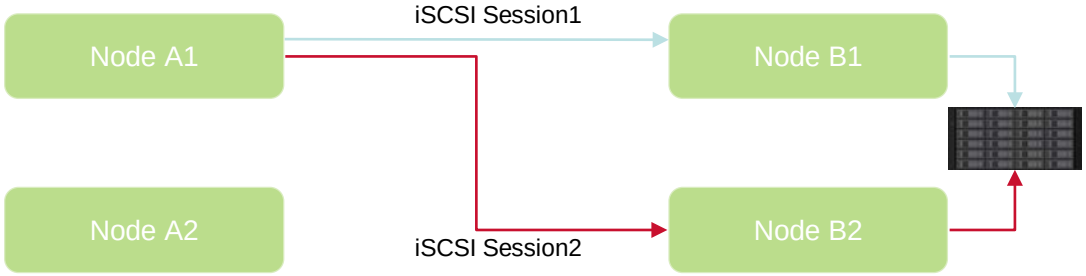
Source: Placeholder

- Storage agnostic – FC, SAS or NVMe, wire-protocol remains iSCSI
- Allows asymmetric configurations
- Namespace level reservations
- Superset of features

MetroCluster IP: Storage

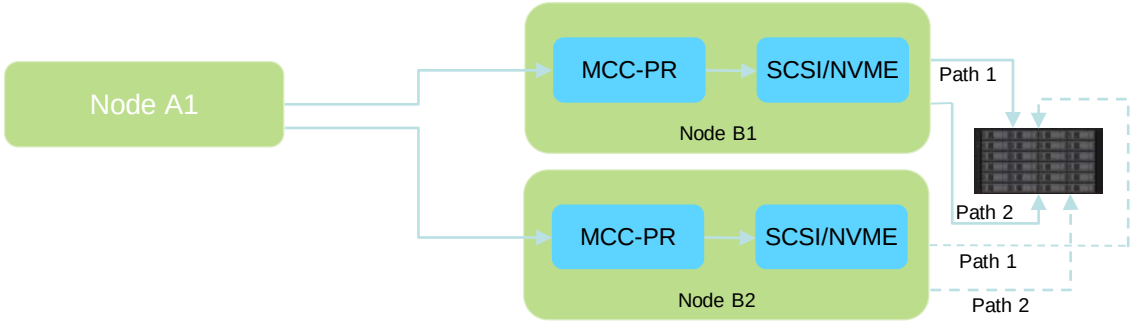
1) No separate cabling needed, head as a target
2) Leverage Ethernet Cluster switches, FC switches not needed
3) No need for FC to SAS conversion
4) Local storage is presented as a LUN, could be anything (cloud container)
5) Ethernet L2 CoS and IP QoS used for traffic prioritization
6) CPU involvement needed, but bulk data + TCP stateless offload makes this cheap
7) Easy to utilize bandwidth by adding additions iSCSI sessions with LUN striping across sessions
8) Reservations can be implemented in software if not available

MetroCluster IP: Storage Multipathing



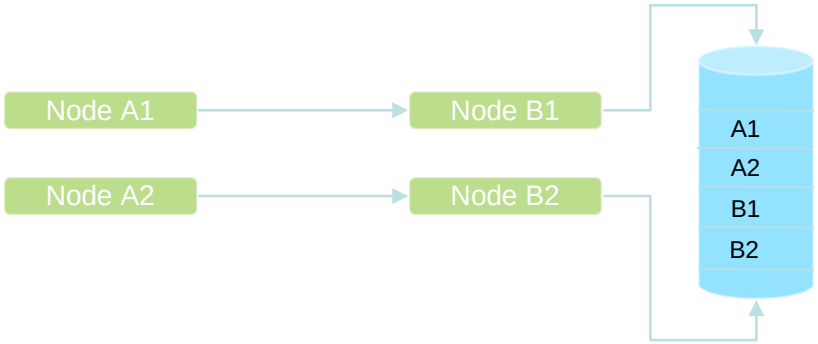
- Each node exports all the disks directly attached to it
- Since the HA-partners have shared-storage both the nodes export the same view of disks via the iSCSI target
- From the initiator each node establishes iSCSI sessions to both the nodes of the DR-site
- The 2 iSCSI connections show up as different paths to same disks (uuid based)
- On takeover all the IO on the victim node failover to the session on the node in takeover

MetroCluster IP: Persistent Reservations



- ONTAP uses Write Exclusive reservations to fence victim nodes on Takeover/Switchover
- The reservation commands arriving on the target will be passthrough with the key belonging to the initiator
- Any reservation conflict on the target will be returned back to the initiator node and since the key still belongs to the initiator it will resolve the conflict
- Global fencing initiated on a DR event (fence all IOs)

Advanced Disk Partitioning



- Very dense drives, requiring slicing of containers
- PRs not supported natively for some devices at namespace granularity
- MCC-IP can support disk-slicing without requiring group-reservations
- On a DR event, the remote cluster nodes are fenced by the MCC-PR layer avoiding the need for reservations

MetroCluster IP: Network Topology

1) The simplest network configurations is a stretched L2 using dedicated ISLs (leased lines, dark fiber, dedicated lambda)
2) Next up is stretched L2 with shared ISLs, network SLA needed for latency, bandwidth and jitter
3) Then we have L3, i.e. routed configurations, network SLA needed
4) Finally we can operate over the Internet, to the cloud
   - Some interesting dynamics between the storage & network admins
   - This opens up the door for product innovations
   - TCP & IP telemetry provide insight into DR network
     - TTL can give hop information, traceroute for L3 nodes

MetroCluster IP: Security

1. IPSec available for end-to-end security
2. The DR module operates in an isolated virtualized network stack
3. In most configs, link-local IP addresses (v4 or v6) can be used
4. The host firewall is used for additional security
5. RDMA uses memory registration, WRITEs with tags
6. iWARP runs with a strong CRC on top of TCP

MetroCluster IP: Performance

1. Mainly 2 micro-benchmarks used:
   - Small (4KB/8KB) random write
   - Large (64KB) sequential write
2. Random write is MUCH better with iWARP (credits vs. window)
3. Sequential write is ON PAR with FC
4. Note: watch out for iWARP RDMA WRITES
   - Completions are locally generated
   - Read-after-write needed to confirm Placement
   - Need to tune IRD/ORD values (use MPAv2)

Final Thoughts

1. High velocity of development (PoC with software iWARP)
2. Domain knowledge widely available (compared to FC)
3. New & interesting things possible with shared fabrics