HPE
POINTNEXT

Best Practices for Multi-site and Disaster-Tolerant OpenVMS Clusters

OpenVMS Boot Camp 2017, SID 303

Keith Parris, Engineer

Hewlett Packard
Enterprise

What things can I do to maximize performance in a multi-site cluster?

What things can I do to maximize performance in a multi-site cluster?

- The distance between sites adversely affects performance primarily in two areas:
  - Remote disk writes to keep cross-site shadowsets in synchronization
  - Remote lock operations
    - Including remote directory lookups
- Estimating inter-site latency:
  - See Excel spreadsheet to calculate latency due to speed of light through fiber optics at https://sites.google.com/site/keithparris/Latency_due_to_Speed_of_Light.xls
  - Rule of thumb: 1 millisecond round-trip for each 100 kilometers (roughly 62 miles)
  - Actual circuit path length may be (significantly) longer than physical inter-site distance
  - LOCKTIME.COM tool from http://encompasserve.org/~parris/ may be used to measure actual round-trip times in an existing cluster

What things can I do to maximize performance in a multi-site cluster?

- Speeding Remote Shadowset Writes:
  - If access to the disks at the remote site is via MSCP Serving or most Fibre Channel SAN Extension methods, remote writes will take **two round trips** between sites to complete
  - Remote Writes can be done in only **one round trip** between sites with either:
    - Cisco I/O Acceleration, or
    - Brocade Fast Write
  - Reads typically take **one round trip**

What things can I do to maximize performance in a multi-site cluster?

- Ensure inter-site network is:
  - Clean (no packet loss)
  - Lowest latency possible, given the inter-site distance
    - This may mean avoiding LAN-within-IP encapsulation methods or MPLS which tend to add latency
  - Low as possible in jitter of packet latency
  - Sufficient in bandwidth to perform a Volume Shadowing Full-Copy operation to restore redundancy of all the cross-site shadowsets in a “reasonable” amount of time (say, overnight)
- With longer inter-site distances, the SYSGEN parameter CLUSTER_CREDITS may need to be raised to avoid SCS credit waits on the VMS$VAXcluster SYSAP connection
- If MSCP Serving is in place, the MSCP_CREDITS and/or MSCP_BUFFER parameters may need to be raised on the VMS$DISK_CL_DRVR → MSCP$DISK SYSAP connection
- Performance of MSCP Serving and Lock Tree Remastering can be better if the network supports Jumbo Packets

What things can I do to maximize performance in a multi-site cluster?

– With a high-quality inter-site network, PEDRIVER transmit and receive window sizes may need to be increased for greater throughput with longer inter-site distances
  – Use SCACP> CALCULATE WINDOW_SIZE /SPEED=*megabits* /DISTANCE=*units=distance*

```
SCACP> help calculate example

CALCULATE

  Example


     SCACP> CALCULATE WINDOW_SIZE /SPEED=1000/DISTANCE=KILOMETERS=500

         The command in this example calculates the window size to be
         used between two nodes that are 500 kilometers apart, connected
         by a 1 Gigabit/Second line speed. The command produces output
         similar to the following:

            Calculate Window Size  2-JUN-2006 17:49:18.41:
                   Inter-node link DISTANCE:          500 KILOMETERS
                   Inter-node link SPEED:            1000 Mb/s
                   ------------------------        -----------------
                   Recommended WINDOW SIZE:          1024 frames
```

How many sites can I have in a single cluster?

How many sites can I have in a single cluster?

- The OpenVMS Cluster design limit for number of nodes in a cluster is 256
  - So with one node per site, that implies an absolute limit of 256 sites
- The OpenVMS Cluster Software SPD indicates support for a maximum of 96 nodes
- Largest known OpenVMS cluster was 151 nodes
- Host-Based Volume Shadowing supports a maximum of 6 shadowset members, so if Volume Shadowing is used to keep data replicated between sites, that implies a maximum of 6 sites with storage, and maybe a quorum node at a 7th site

How far apart can my sites be?

“The maximum system separation is 150 miles. With proper consulting support via HP Services Disaster Tolerant Consulting Services, the maximum system separation is 500 miles.”

-- OpenVMS Cluster Software Product Description
http://h41379.www4.hpe.com/doc/spdclusters.pdf

How far apart can my sites be?

- OpenVMS Clusters have no inherent design limit on distance
- Host-Based Volume Shadowing has been tested out to a simulated distance of over 60,000 miles successfully
- During development of IP as a Cluster Interconnect (IPCI), OpenVMS Engineering successfully formed and tested a cluster with nodes on several continents simultaneously
- So why the limit in the SPD?
  - Concerns about application performance, because of the latency due to the speed of light between sites
- Increased distance primarily affects performance in two areas:
  - Remote disk writes to keep shadowsets in synchronization
  - Remote lock operations (including directory lookups)

How far apart can my sites be?

– We advise that before the datacenter sites are chosen, you first test application performance at the proposed inter-site distance using a network emulator to introduce network latency to simulate the proposed inter-site circuit path length (distance)

– We know of a number of customer sites running OpenVMS Clusters with circuit path lengths in the 600-900 mile range and one with 3,000 miles between sites, but “your mileage may vary” depending on your applications. If you need an official support statement, send e-mail to OpenVMS.Programs@hpe.com

How far apart can my sites be?

Case Study

- Customer question: Can I spread my OpenVMS cluster across a 1,000 mile distance for disaster tolerance?
- Investigation: 1,000 miles between datacenter sites implies a 16 millisecond round-trip time. With MSCP Serving for access to remote disks, each disk write would take 2 x 16 = 32 milliseconds.
- We asked the customer questions about their application performance and expectations:
  - Questions: How many disk writes are needed for your customer transactions, and what response time expectations do your users have for these transactions?
  - Answers: Transaction requires 4 disk writes, and the response time requirement is 1 second.
  - Analysis: 4 x 32 milliseconds = 128 milliseconds, or about 1/8 second, so this would appear to easily meet the 1 second maximum response time requirement.
- This gave us a rough indication that the application would likely still work acceptably in a 1,000-mile cluster. Testing of the actual application performance with latency simulated via a network emulator box would still be advisable, to confirm this initial prediction.

How should I choose nodenames?

How should I choose nodenames?

- Avoid picking names descriptive of the hardware, platform, architecture or model
  - Nodenames tend to stay forever, but may be applied to different hardware boxes over time. Many customers still have nodenames like VAX01 yet are Integrity Server boxes now.
- Avoid choosing nodenames based on the company name, because of buyouts, mergers, name changes, etc.
- Choose a scheme that allows easily identifying the datacenter site based on the nodename
- Choose a scheme that is scalable, in terms of:
  - Number of nodes
  - Number of sites
- Examples:
  - Fixed prefix, and range of node numbers indicates site
    - e.g. XYZ1nn for nodes at Site 1; XYZ2nn for nodes at Site 2; XY3nn for nodes at Site 3
  - Prefix indicates site, suffix differentiates nodes within that site
    - e.g. DEN001, DEN002 in Denver, ATL001, ATL002 in Atlanta

How should I choose nodenames?

– Case studies:
  – Poor foresight: First two nodes at the first site were Node 1 and Node 2. Second site had Node 3 and Node 4. Then a node was added to each site, becoming Node 5 at the first site and Node 6 at the second site. Very confusing.
  – Not scalable: Even-numbered nodes are at one site; odd-numbered nodes are at opposite site: Problem: How do you name a quorum node at a 3rd site? How do you expand from two sites to 3 sites if you need to?

How should I number my disk devices?

How should I number my disk devices?

- Choose disk unit numbers so the site at which the disk is located can be easily identified from the unit number
  - e.g. 1DGA1xxx disks are at Site 1; 1DGA2xxx are at Site 2, etc.
- Leave room for expansion
- Number shadowsets logically as well
  - e.g. DSA0 through DSA999 for cross-site shadowsets with members located at multiple sites
  - e.g. DSA1xxx for shadowsets located only at Site 1, such as a local system disk shadowset for the site
- Choose disk unit numbers and shadowset device names so it is also easy to identify which shadowset the disk is a member of
  - e.g. Shadowset DSA14 has members 1DGA1141 and 1DGA1142 at Site 1 and 1DGA2141 and 1DGA2142 at Site 2 and 1DGA3141 and 1DGA3142 at Site 3

How should I choose allocation classes?

How should I choose allocation classes?

– It is no longer necessary for the node doing MSCP Serving to have the same allocation class as the served disk(s).

– If you have server nodes with similar hardware, simpily pick a unique allocation class per node. This prevents device name conflicts for devices like DVDs as DQA0 or such.

When should I use IPCI (IP as a Cluster Interconnect)?

When should I use IPCI (IP as a Cluster Interconnect)?

- Primary reason:
  - When the network cannot provide at least the illusion of a bridged LAN between sites
- Secondary reasons:
  - When the methods available to encapsulate LAN packets within IP to simulate bridging introduce so much latency, jitter and/or packet loss that using the IP network directly using IPCI will provide better performance and availability
  - As a backup path to provide higher availability in the event of a problem on the LAN, e.g. ride through Spanning Tree reconfigurations
- Why do you prefer LANCI over IPCI in general?
  - IPCI has slightly lower maximum packet payload size than LANCI
  - IPCI has slightly higher host CPU overhead than LANCI, since the code path includes TCP/IP Software as well as PEDRIVER
  - Because IP Multicast is seldom set up for IPCI at most customer sites, the Hello packets used for path discovery and path status must use Unicast rather than Multicast packets, adding a small amount of overhead to the hosts, which is greater as the number of nodes in the cluster goes up
- Market shift: When IPCI was invented, the threat was that all extended or bridged LANs would go away. But because Virtual Machines can now migrate between servers, transparent to the VM, and do so even between sites, for purposes of DR, the market has demanded the ability to support for the same IP segment at different sites. Network vendors have developed technology to meet this demand, one that OpenVMS Engineering had no way to predict. This same technology can support OpenVMS Clusters.

Should I use different Site Numbers in my multi-site OpenVMS cluster?

Should I use different Site Numbers in my multi-site OpenVMS cluster?

- Site Numbers are a shorthand way of setting Read Costs for shadowset members
- SHADOW_SITE_ID may be set to a positive, non-zero value to designate a node as being located at a given site
  - Alternatively, you could use $ SET DEVICE/SITE=n DSAnnn: for each and every shadowset to set the server Site ID, but this is an older method entailing a lot more work
- With different Site ID values, the read costs for member disks are set to default values:
  - Local DECram disk: 1
  - Local SAS or Fibre Channel magnetic disk: 2
  - Remote Fibre Channel Disk: 42 (difference of 40 above local disk)
  - MSCP-Served disk: 501 (difference of 499 above local disk)
- The decision of whether or not to use different Site numbers will then depend on whether these default Read Cost settings are acceptable, or whether it is considered worthwhile to override them all manually to still retain the Site number setting and yet have customized Read Cost values, set at system startup time via a set of $ SET DEVICE /READ_COST commands.

How should I set Read Costs for Host-Based Volume Shadowing?

How should I set Read Costs for Host-Based Volume Shadowing?

- Because Shadowing keeps members disks identical, a read operation could be satisfied from any one of the member disks
- Idea of differing Read Costs is to send reads from servers within a site to disks at the same site, to avoid the inter-site latency
- When a shadowset is read, Shadowing adds the local queue length for each member disk to the Read Cost and directs the read to the member with the lowest total. For disks with equal Read Costs, Shadowing sends reads to the member disks in “round-robin” order.
- Sometimes a node experiences a burst of reads.
  - With equal Read Cost values, the reads can be equally distributed across all the member disks for lower average response times.
  - With different Read Cost values, the reads will be distributed first to member(s) with lower Read Cost value(s), and only after the local queue length to those disks rises, to members with higher Read Cost values.
  - With small differences read costs, when the local queue depth reaches the difference in read costs, Volume Shadowing will start sending some of the reads to the remote disk, allowing them to be satisfied faster.
  - With large differences in read costs, the local queue depth has to get very high before any reads start to be directed to the remote disk. The local disk must handle more of the burst of reads.

How should I set Read Costs for Host-Based Volume Shadowing?

Example:

– Local random read disk response time is 6 milliseconds

– Circuit path length between sites is 200 kilometers (about 124 miles), so round-trip time between sites is 2 milliseconds.

– Remote disks are MSCP-served, so remote reads take 2 round trips, or 4 milliseconds, plus the regular 6 millisecond disk response time, for a total of 6 + (2 x 2) = 10 milliseconds (assuming the disk is idle).

– So when the local disk queue length rises to 1, a local disk read would have to be placed in the queue behind the first request, so it would take 6 x 2 = 12 milliseconds for the 2nd read request. At this point, if the local disk is likely to be idle, it becomes better to send a read to the remote disk (10 millisecond response time) than to add it to the queue for the local member disk (12 millisecond response time).

– So relatively small differences in Read Cost values allow reads to be biased toward the local disk, but retain the option to start sending some of the reads across to the disk at the opposite site if the local queue length gets too long.

How should servers in my disaster-tolerant cluster start up?

How should servers in my disaster-tolerant cluster start up?

– Divide node startup into 3 separate and distinct phases:

1. Booting
2. Mounting cross-site shadowsets
3. Starting applications and then allowing access to users

– Set all systems up to boot “Conversational” by default, either at the EFI (Integrity) or SRM (Alpha) level, so that instead of rebooting automatically after a crash or temporary power failure, they will always stop at the SYSBOOT> prompt, allowing manual intervention and control.

– System startup should be either strictly manually controlled or at least human-directed (i.e. use a spare SYSGEN parameter like USERD1 or something to indicate whether or not, and if so, how cross-site shadowsets should be mounted, and if applications should then be started up or not)

– Depending on the circumstances, the cross-site shadowsets might need to be:
  – Left unmounted for further manual troubleshooting, followed by appropriate manual action by the system manager
  – Mounted using member disks from both (all) sites at once
  – Mounted using only member disks from Site 1 (perhaps that site has the only valid copy of the data)
  – Mounted using only member disks from Site 2
  – Similarly, in a 3-site cluster, options to mount shadowsets using only member disks from Site 3, or a pair of sites: 1&2, 2&3, or 1&3

– Because preservation of data on disk is so crucial in a disaster-tolerant cluster, and it is important to avoid a “wrong-way” shadow copy, use the qualifier /POLICY=(REQUIRE_MEMBERS,VERIFY_LABEL) on all the MOUNT commands for cross-site shadowsets.

Questions?

HPE
POINTNEXT

Thank you

keith.parris@hpe.com

Hewlett Packard
Enterprise