



Disk Load Balancing, Fault Tolerance, and Configuration Limits for NonStop Systems

Technical brief

Table of contents	
Summary.....	2
Introduction	2
Processors and memory	3
ServerNet.....	4
Adapters, SACs, CLIMs, and configuration limits.....	5
I/O buses	7
Examples	7
Glossary	9
References.....	10

Summary

A frequently asked question about disk load balancing is "how can I balance the load of FCSA-attached disk I/O across the X and Y ServerNet fabrics?" The short answer is that it is done automatically by SCS (T8456), which is used for all I/O to FCSA-attached devices. SCS also automatically balances the ServerNet I/O to CLIM-attached devices.

The remainder of this document provides more complete discussion of several topics related to load balancing, including:

- Processor fault tolerance and load balancing
- ServerNet fault tolerance and load balancing
- Adapter, SAC, and CLIM fault tolerance and load balancing
- I/O bus fault tolerance and load balancing
- Maximum device configuration limits

Introduction

The underlying principles of fault tolerance and load balancing are always the same. They:

- Don't allow any single failure in any hardware or software component to make data inaccessible.
- Spread the workload evenly across all available system components to get the most capacity and responsiveness at the lowest cost.

Application of these principles to different generations of NonStop systems can result in different advice because each system generation packages the hardware components differently, resulting in different potential points of failure and different potential performance bottlenecks. There are 3 significantly different generations of NonStop systems in common use today:

- S-series systems, characterized by S-series processor enclosures and running G-series RVUs. The ServerNet topology is Tetra-8 or Tetra-16.
 - S-series I/O consists of S-series I/O enclosures containing IOMF adapters and SNDA adapters with S-PIC (SCSI) SACs. The I/O bus is SCSI and the disks are internal SCSI disks in the S-series enclosures.
 - S-series systems support a backward-compatible I/O generation with SNDA adapters with F-PIC (fiber) SACs connected to 45xx disk modules.
 - S-series systems support a forward-compatible I/O generation by replacing some S-series I/O enclosures with IOAME enclosures.
- NS-series systems, characterized by a P-switch ServerNet topology and running H-series RVUs. Note: Neoview systems running N-series RVUs are architecturally similar to NS-series P-switch systems, but Neoview system configuration is tightly constrained and optimized for the application suite which runs on Neoview systems, so this document does not discuss Neoview systems.
 - NS-series I/O consists of IOAME enclosures containing FCSA adapters connected to ESS (HP XP Enterprise Storage Systems) and FCDM disk enclosures. The I/O bus is Fibre Channel. VIO systems integrate parts of an NS-series system and IOAME together into shared packaging.
 - NS-series systems support a backward-compatible I/O generation by replacing some IOAME enclosures with S-series I/O enclosures.
 - NS-series systems support a forward-compatible I/O generation by replacing some IOAME enclosures with CLIMs.

- NB-series systems, characterized by processors and ServerNet switches in a c-7000 enclosure and running J-series RVUs.
 - NB-series I/O consists of CLIMs connected to ESS (HP XP) and MSA70 disk enclosures containing SAS disks. The I/O buses are Fibre Channel and SAS.
 - NB-series systems support a backward-compatible I/O generation by replacing some CLIMs with IOAME enclosures.
 - NB-series systems support another backward-compatible I/O generation by replacing some CLIMs with S-series I/O enclosures.

Processors and memory

For fault tolerance, an IOP should run in two processors which don't share any packaging. In S-series systems running G06.22 or older RVUs, the halves of an IOP had to run in processors in the same S-series processor enclosure. Starting with the G06.23 RVU, this constraint was eliminated. Best fault tolerance is achieved when the loss of one processor enclosure cannot prevent access to both halves of the IOP, but this can increase the ServerNet path length and affect performance, so a reasonable compromise is to split each IOP between two processor enclosures which are directly connected to each other. S-series processor enclosure connectivity is described in the following ServerNet section.

In an NS16x00 system, the processors are packaged in processor complexes. Logical processors 0, 1, 2, and 3 share a processor complex. Processors 4-7 share a complex, processors 8-11 share a complex, and processors 12-15 share a complex. In other NS-series systems, processors are rack mounted pairwise in processor modules. Processors 0 and 2 share a module, as do processors 1/3, 4/6, 5/7, 8/10, 9/11, 12/14, and 13/15.

In all system types, the \$SYSTEM IOP must run in processors 0 and 1 irrespective of any common packaging.

In S-series systems, disks, adapters, and processors are packaged together, so the SCF ADD DISK command provides load balanced default values when processors are not specified. In NS-series and NB-series systems, processor packaging is decoupled from I/O packaging so it is necessary to specify processors in the SCF ADD DISK command to spread the IOPs across the available processors.

For load balancing, half of the IOPs configured to use a processor should use it as a PRIMARYCPU and the other half should use it as a BACKUPCPU. If the configuration of PRIMARYCPUs and BACKUPCPUs in the CONFIG file is unbalanced and it is not convenient to stop the IOPs and change the configuration, but it is possible to swap the current primary and backup IOP halves online with SCF PRIMARY. SCF PRIMARY only affects the running IOP. It does not change the configured PRIMARYCPU or BACKUPCPU.

The size of physical memory imposes a practical limit on how many IOPs can run in a processor. Each DP2 IOP half, whether primary or backup, requires a minimum amount of memory to support any function. There is also a practical minimum which is the amount of memory needed to efficiently support the set of applications using the volume. This latter characteristic is a performance issue and is not covered here, except to say that the optimal memory configuration is typically obtained by varying DP2 attributes until the desired performance is reached.

For the purpose of estimating memory needs, one can first assume that each DP2 process pair requires the same amount of memory in each processor. This is not always the case, but is a good approximation. This assumption also addresses the memory needs should the primary processor fail and all disks are then primaried to the remaining processor.

Note that the best processor loading is typically achieved by configuring half for primary and half for backup. In some environments a different mix might yield the best performance. Many customers alter the primary/backup mix during the day to address different loads.

The amount of memory each volume consumes is application-dependent, but can be limited by user-controllable disk attributes such as: CBPOOLLEN, AUDITRAILBUFFER/SQLMXBUFFER, LKIDLONGPOOLLEN, LKTABLESPACELEN, and cache size. In general, the largest consumer of memory is cache.

At a minimum, DP2 memory is required for stack space, internal buffers for both DP2 FM and I/O Driver, control blocks, lock tables, work pools, etc. The minimum consumption for these areas in most S-series systems is about 31 MB. For S70000 systems, NS-series systems, and NB-series systems, the minimum is 47 MB.

There is typically a direct relationship between memory and processor consumption. Since this is a processor performance issue, it will not be addressed here except to note that there is likely a minimum memory working set for each volume below which performance is unacceptable. Also, consider that the minimum cache configuration requires 18 MB.

The lock tables can be configured through SCF through the LKTABLESPACELEN and LKIDLONGPOOLLEN attributes. There is one LKID Long pool, so the maximum consumption is limited by this attribute. The default is 8 MB. The LKTABLESPACELEN attribute is applied to 4 separate tables, so multiply this value by 4 for the total (default = 15 MB each). These tables grow as needed, and at a minimum they consume about 1 MB.

The ADP/SQLMXBUFFER maximum size is also configurable. If the volume is not an ADP nor has SQL/MX tables, these areas can be ignored. Otherwise, add the corresponding size to what the volume requires.

With this in mind, the maximum number of volumes configurable in a processor is determined by memory consumed by cache and user-controllable attributes. The absolute minimum requirement is the minimum (31 MB) + cache (18 MB) + lock tables (1 MB) = 50 MB. From the physical memory size, subtract the size of the working set for memory used by all other non-disk system and customer processes to determine how much memory is left for disks.

For example, for an 8 GB processor, assuming system + customer processes, consume 2 GB, there is 6 GB available = 120 disks. For a 4 GB processor with 3 GB available for disks, the maximum is about 60 disks. An S70000, NS-series, or NB-series system has a higher minimum memory usage (47 MB vs. 31 MB), so the equivalent calculations show that 6 GB of available memory supports 90 disk IOPs and 3 GB of available memory supports 45 disk IOPs.

Note that the values used in this example will be different between hardware platforms, and may change significantly from one RVU to another.

For some system configurations, a calculation based on path count (see the Adapters, SACs, CLIMs, and Configuration Limits in the following section) will produce a smaller maximum number of volumes than this calculation based on memory limitations. The effective limit is the smaller of the two limits.

ServerNet

All generations of NonStop systems contain two separate ServerNet fabrics, named X and Y. In any system, adapters in S-series I/O enclosures are single-fabric so it is important to configure a disk's -P (primary) and -MB (mirror backup) paths through adapters using one fabric and configure the -B (backup) and -M (mirror) paths through adapters on the other fabric. Adapters in slots 50, 51, and 53 are on the X fabric. Adapters in slots 52, 54, and 55 are on the Y fabric. The separation of -P from -B and -M from -MB provides fault tolerance. The separation of -P from -M provides load balancing because a write to a mirrored volume goes to both drives and the -P and -M paths are the active paths by default.

In any system with IOAMEs or CLIMs, I/O traffic to any FCSA or CLIM is automatically distributed across both ServerNet fabrics by SCS (T8456). This provides both fault tolerance and load balancing. SCS automatically rebalances the load after failure and restoration of a ServerNet fabric.

However, the module number of an FCSA adapter does imply an affinity for a specific ServerNet switch (module 2 = X fabric, module 3 = Y fabric) for FCSA reset processing, so the four paths to a disk volume should be configured to use both fabrics as described above for S-series I/O. The convention for identifying the location of a CLIM always uses a ServerNet connection point with module=2. This does not indicate an affinity between CLIMs and the X fabric.

The connection points of a CLIM to an NS-series P-switch are identified by group, module, slot, and port. The slot number identifies a FRU which holds four port connectors. For best fault tolerance, the -P and -B paths to a disk volume should not use CLIMs connected to the same slot number. The same is true for the -M and -MB paths and the -P and -M paths.

The connection points of a CLIM to an NB-series ServerNet switch are identified by group, module, slot, port, and fiber. The port number identifies a connector which has a 1-to-4 splitter cable connected to it. For best fault tolerance, the -P and -B paths to a disk volume should not use CLIMs connected to the same port number. The same is true for the -M and -MB paths and the -P and -M paths.

In an S-series system, the four low-numbered processor enclosures (groups 1 through 4) are directly connected to each other and each I/O enclosure is directly connected to only one processor enclosure. The four high-numbered processor enclosures are only connected to one other processor enclosure each (1-5, 2-6, 3-7, and 4-8). The loss of a low-numbered processor enclosure causes the isolation of its directly connected I/O enclosures and the directly connected high-numbered processor enclosure and its directly connected I/O enclosures.

Adapters, SACs, CLIMs, and configuration limits

In S-series systems running G06.04 or older RVUs, all the adapters used by an IOP had to be in the same S-series enclosure and that enclosure had to be in the same topology branch as the IOP. A "topology branch" is an S-series processor enclosure and all of its directly connected I/O enclosures. Starting with the G06.05 RVU, this constraint was relaxed so that the adapters for an IOP could be spread across multiple enclosures within one topology branch. Starting with the G06.23 RVU, and in all H-series and J-series RVUs, the adapters and IOP can be spread across the entire system. But spreading the adapters across multiple topology branches in an S-series system increases the ServerNet path length, so a reasonable compromise is to put the adapters in the same topology branch(es) as the IOP.

In systems running any G-series RVU and in systems running H06.07 or older RVUs, only two processors can share access to a SCSI SAC, so all of the devices connected to a SCSI SAC must have their IOPs running in the same two processors. Starting with the H06.08 RVU, and in all J-series RVUs, this constraint was relaxed so that eight processors can share access to a SCSI SAC at the same time.

With NS-series I/O connected to any system type, an IOP in any processor can use any FCSA adapter. But there are some maximum configuration limits. In S-series systems running the G06.23 or older RVUs, a total of 256 device paths can be configured to use an FCSA. Starting with the G06.24 RVU, and in all H-series and J-series RVUs, this constraint was relaxed so that each processor can have up to 125 paths configured through each FCSA. Paths to all device types (disk, tape, and Open-SCSI) are included in this limit.

To maximize the number of IOPs that can be configured in a processor, each IOP should use 4 different FCSAs for its 4 paths (-P, -B, -M, and -MB). Then each IOP only has one path through each FCSA and 125 IOPs can be configured in the same processor. If more than 4 FCSAs are used, then even more IOPs can be configured in the same processor, but available memory also limits the

number of IOPs which can be launched in a processor. See the explanation in the preceding "Processors and memory" section.

Ignoring memory limitations, the approximate maximum number of FCSA-attached mirrored disk volumes which can be configured on a system is:

The number of processors → Divided by 2 (since an IOP has 2 halves and each half uses all 4 paths) → Times the number of FCSA adapters (number of SACs doesn't matter) → Times the 125 limit → Divided by 4 (because each volume has -P, -B, -M, and -MB paths)

For many system configurations, memory limitations will produce a smaller maximum number of volumes than this calculation based on path count. The effective limit is the smaller of the two limits.

For best load balancing, distribute the IOPs which use an FCSA or CLIM across all available processors and configure the IOPs which run in the same processor to distribute their paths across all available FCSAs or CLIMs. This will balance the load on the processors and also spread I/O traffic across the multiple ServerNet paths leading to each FCSA or CLIM. The I/O paths through an adapter should be distributed across the SACs on the adapter.

In addition to distributing all configured device paths evenly across all FCSAs, SACs, and CLIMs, all paths which are active by default (that is, -P and -M) should be distributed evenly across all FCSAs, SACs, and CLIMs. If the configuration of -P, -B, -M, and -MB paths in the CONFIG file is unbalanced and it is not convenient to stop the IOPs and change the configuration, it is possible to swap the current active and inactive paths online with SCF SWITCH. SCF SWITCH only affects the running IOP. It does not change the configured -P, -B, -M, and -MB paths.

CLIMs connected to any system type have a limit of 1000 paths from each processor through each CLIM. This limit cannot be reached at the moment because the CLIM itself only allows the configuration of a total of 512 device paths through it and available memory will limit the number of IOPs which can be launched in a processor.

A CLIM can be connected to a maximum of 8 ESS ports, with a maximum of 500 ESS LDEVs per ESS port. Both maxima cannot be reached at the same time due to the limit of 512 total paths per CLIM. To avoid exceeding these limits, extraneous ESS ports and LDEVs should be hidden from a CLIM by using SAN fabric zoning and ESS LUN masking.

In an S-series adapter (PMF, IOMF, or SNDA) or FCSA adapter, multiple SACs are packaged on one adapter. It is possible to configure a mirrored disk volume using only 2 adapters, with different SACs on one adapter carrying the -P and -MB paths and different SACs on the other adapter carrying the -B and -M paths. But using 4 different adapters provides better fault tolerance.

For CLIM-connected disk volumes, it is possible to configure 4 paths to a mirrored volume using only 2 CLIMs, with the -P and -MB paths through one CLIM and the -B and -M paths through the other CLIM, but using 4 separate CLIMs for the 4 paths to a mirrored disk volume is more fault tolerant.

I/O buses

The most common issue at this level in the architecture is bandwidth. For all bus types (SCSI, Fibre Channel, and SAS domains), multiple devices share the bandwidth of a single backplane, cable, or expander. For this reason, it is best to spread the devices evenly across the available I/O buses.

An I/O bus is also a potential point of failure, so the primary and mirror drives of a mirrored disk volume should not be configured on the same I/O bus. All of the disks in a 45xx disk module share one I/O bus. All of the disks in an FCDM disk module share two Fibre Channel loops. All of the disks in an MSA70 enclosure share two SAS domains. Half of the disks (in the odd numbered disk slots) in an S-series processor or I/O enclosure share one SCSI bus. The even numbered disk slots share the other SCSI bus.

A Fibre Channel loop through a daisy chain of FCDM disk modules or a SAS domain through a daisy chain of MSA70 disk enclosures should also be considered as a potential point of failure, so the primary and mirror drives of a mirrored disk volume should not be configured on the same daisy chain.

A failure or service procedure in one member of a daisy chain (either FCDMs or MSA70s) can isolate all of the enclosures further down the daisy chain, so the alternate path (that is, the other Fibre Channel loop or SAS domain) should be connected to the opposite end of the daisy chain so that failure or removal of one disk enclosure does not eliminate both paths to any other disk enclosure.

A Fibre Channel SAN fabric should be considered as a potential point of failure because the entire fabric can pause long enough to cause disk I/O to time out during fabric reconfiguration. For best fault tolerance, all disk -P and -MB paths should use a SAN fabric which is separate from the SAN fabric used by all disk -B and -M paths.

Examples

The first example shows 16 mirrored FCSA-attached disk volumes load balanced in every way.

- Across 4 processors
 - 4 IOPs primary in each processor
 - 4 IOPs backup in each processor
- Across 2 ServerNet fabrics (based on FCSA affinity)
 - 32 configured paths through each fabric
 - 16 active paths through each fabric
- Across 8 SACs (4 FCSAs*2 SACs on each FCSA)
 - 8 configured paths through each SAC
 - 4 active paths through each SAC
- Across each combination of processor and SAC
 - 4 configured paths from each processor to each SAC
 - 2 active paths from each processor to each SAC
 - 1 active path from each primary processor to each SAC

Example 1

Volume	CPUs	-P	SAC	-B	SAC	-M	SAC	-MB	SAC
\$D1	0,1	(110,2,1)	1	(110,3,1)	1	(110,3,2)	1	(110,2,2)	1
\$D2	1,0	(110,2,1)	1	(110,3,1)	1	(110,3,2)	1	(110,2,2)	1
\$D3	2,3	(110,2,1)	1	(110,3,1)	1	(110,3,2)	1	(110,2,2)	1
\$D4	3,2	(110,2,1)	1	(110,3,1)	1	(110,3,2)	1	(110,2,2)	1
\$D5	0,1	(110,2,1)	2	(110,3,1)	2	(110,3,2)	2	(110,2,2)	2
\$D6	1,0	(110,2,1)	2	(110,3,1)	2	(110,3,2)	2	(110,2,2)	2
\$D7	2,3	(110,2,1)	2	(110,3,1)	2	(110,3,2)	2	(110,2,2)	2
\$D8	3,2	(110,2,1)	2	(110,3,1)	2	(110,3,2)	2	(110,2,2)	2
\$D9	0,1	(110,3,1)	1	(110,2,1)	1	(110,2,2)	1	(110,3,2)	1
\$D10	1,0	(110,3,1)	1	(110,2,1)	1	(110,2,2)	1	(110,3,2)	1
\$D11	2,3	(110,3,1)	1	(110,2,1)	1	(110,2,2)	1	(110,3,2)	1
\$D12	3,2	(110,3,1)	1	(110,2,1)	1	(110,2,2)	1	(110,3,2)	1
\$D13	0,1	(110,3,1)	2	(110,2,1)	2	(110,2,2)	2	(110,3,2)	2
\$D14	1,0	(110,3,1)	2	(110,2,1)	2	(110,2,2)	2	(110,3,2)	2
\$D15	2,3	(110,3,1)	2	(110,2,1)	2	(110,2,2)	2	(110,3,2)	2
\$D16	3,2	(110,3,1)	2	(110,2,1)	2	(110,2,2)	2	(110,3,2)	2

The second example shows 8 mirrored CLIM-attached disk volumes load balanced in every way.

- Across 4 processors
 - 2 IOPs primary in each processor
 - 2 IOPs backup in each processor
- Across 4 CLIMs
 - 8 configured paths through each CLIM
 - 4 active paths through each CLIM
- Across each combination of processor and CLIM
 - 4 configured paths from each processor to each CLIM
 - 2 active paths from each processor to each CLIM
 - 1 active path from each primary processor to each CLIM

Example 2

Volume	CPUs	-P	-B	-M	-MB
\$D21	0,1	C100251	C100261	C100263	C100253
\$D22	1,0	C100251	C100261	C100263	C100253
\$D23	2,3	C100251	C100261	C100263	C100253
\$D24	3,2	C100251	C100261	C100263	C100253
\$D25	0,1	C100261	C100251	C100253	C100263
\$D26	1,0	C100261	C100251	C100253	C100263
\$D27	2,3	C100261	C100251	C100253	C100263
\$D28	3,2	C100261	C100251	C100253	C100263

Glossary

Acronym	Meaning
CLIM	Cluster I/O Machine.
DP2	Disk Process 2. The IOP used for disk devices.
DP2 FM	The File Management part of the DP2 process.
ESS	Enterprise Storage System.
F-PIC	Fiber Plug-In Card. Used with an SNDA adapter.
FCDM	Fibre Channel Disk Module. Contains disk drives.
FCSA	Fibre Channel ServerNet Adapter.
FRU	Field Replaceable Unit.
HP XP	The model name of a Hewlett Packard ESS product.
IOAME	I/O Adapter Module Enclosure. Contains adapters like the FCSA.
IOMF	I/O Multi-Function adapter.
IOP	I/O Process.
MSA70	Modular Smart Array 70. Contains disk drives.
PMF	Processor Multi-Function adapter.
RVU	Release Version Update.
S-PIC	SCSI Plug-In Card. Used with an SNDA adapter.
SAC	ServerNet Addressable Controller.
SAS	Serial Attached SCSI. An I/O protocol.
SCF	Subsystem Control Facility.
SCS	ServerNet Connection Services. Product number T8456.
SCSI	Small Computer System Interface. An I/O protocol.
SNDA	ServerNet Device Adapter.

References

Configuration:

S-series system (pre-G06.05) with S-series I/O enclosures

Document:

Support Note S97086A on G-series Disk Load Balancing

Configuration:

S-series system (G06.05-G06.22) with S-series I/O enclosures

Documents:

Support Note S99056 on G-series Enclosure Interleaving

Manual 529937, SCF Reference Manual for the Storage Subsystem

Configuration:

S-series system with IOAME connected to FCDM and ESS disks

Documents:

Manual 528254 on FCSA Installation and Support

Manual 528764 on Modular I/O Installation and Configuration

Configuration:

NS-series system with IOAME and S-series I/O enclosures

Documents:

Manual 529567 on Integrity NonStop NS-Series Planning

Flexible Disk Configuration (FDC) on HP Integrity NonStop Systems

Configuration:

NS-series or NB-series system with CLIMs and MSA70 enclosures

Document:

Manual 544999 on NonStop CLIM Installation and Configuration

Configuration:

NB-series system with CLIMs and IOAMEs

Document:

Flexible Disk Configuration (FDC) on HP Integrity NonStop BladeSystems

Technology for better business outcomes

© Copyright 2009 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Linux is a U.S. registered trademark of Linus Torvalds. Microsoft and Windows are U.S. registered trademarks of Microsoft Corporation. UNIX is a registered trademark of The Open Group.

4AA2-3322ENW, December 2008
576166-001 Assigned April 2009

