Sun Fire™ Midframe Server Configuration Best Practices

By James Hsieh - High End Services Engineering CTE (HES-CTE)

Sun BluePrints™ OnLine - September 2001

http://www.sun.com/blueprints

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303 USA
650 960-1300 fax 650 969-9131

Part No.: 816-2122-10
Revision 1.1, 09/26/01
Edition: September 2001

Sun Fire™ Midframe Server Configuration Best Practices

The Sun Fire™ 3800, 4800, 4810 and 6800 Midframe servers are highly configurable and provide new features, capabilities, and technologies. Proper planning during configuration and installation will increase overall system Reliability, Availability, and Serviceability (RAS). The emphasis of this article is to recommend “Best Practices” to achieve these goals, and aid in configuring a Sun Fire system for mission critical applications.

Specifically, this article covers the following topics:

- Power Configuration
- Platform Configuration (with special notes for Sun Fire 6800 server)
- Memory and I/O Configuration

Consider the function and requirements of the machine prior to following recommendations made in this article. While many recommendations made here apply to the majority of cases, not all recommendations will apply to every circumstance.

Power Configuration

Overview:

- Configure two RTS Modules per RTU.
- Separate power sources for each RTS.
- If separate power sources are not available, use only one RTS.
- Check the *Sun Fire 6800 System and Sun Fire Cabinet Rackmount Placement Matrix* (Sun Document #816-2062) when adding supported devices to the Sun Fire Cabinet.

Sun Fire servers that are either factory or field installed in a new Sun Fire rack take advantage of new power input and distribution modules unique to the Sun Fire product line. Properly configured, this new technology can provide power source failover capabilities to the rack and its components. Each Sun Fire rack has at least one Redundant Transfer Unit (RTU) that can contain either one or two Redundant Transfer Switch (RTS) modules. Sun Fire 6800 servers and servers with additional power requirements will have two RTUs. The function of the RTU is to distribute power.

A single RTS module is capable of supplying full power to a single RTU. The RTS module monitors incoming power, and, in the event of a brownout or power failure, attempts to switch incoming power to a second RTS module if it is present. At any given time, only one RTS will be active (indicated by two green LEDs on the RTS module) while the second RTS (if present) will be in a “stand-by” mode (indicated by a single green LED on the RTS module). RTS modules do not load share. If the active RTS loses power, failover to a second RTS will occur if one is available. The second RTS will then be supplying the RTU with power. If a second RTS module is not available, the active RTS will eventually shut itself and the RTU off. If the second RTS is unable to supply power to the RTU while the first RTS is in a “failed” state, the RTU will shut down. If the first RTS has recovered from its failure, the second RTS can switch back to the first RTS.

In order to provide failover capabilities to the systems and storage contained in a Sun Fire rack, each RTU in the rack should have two RTS modules. Each RTS should be connected to a power source separate from the other RTS. Separate power sources originate from a different commercial source or Uninterruptible Power Supply (UPS). Phase matching between the two power sources is not required.

Both RTS modules from a single RTU should never be plugged into the same power source. If a separate power source is unavailable, the second RTS unit in an RTU should be disconnected until a second separate power source is available. If both RTS modules are plugged into the same power source, brownout conditions can

cause the RTU to power off as a brownout may cause an active RTS to try to unsuccessfully switch to the second RTS module experiencing the same brownout. If the second RTS is not connected, the active RTS does not attempt a switch.

FIGURE 1 and FIGURE 2 illustrate power connections for Sun Fire Midframe servers.

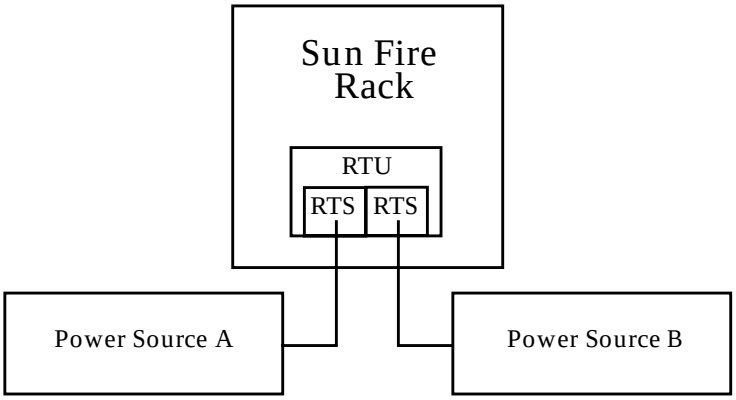

FIGURE 1 Sun Fire 3800/4800/4810 Rack With Single RTU Configured

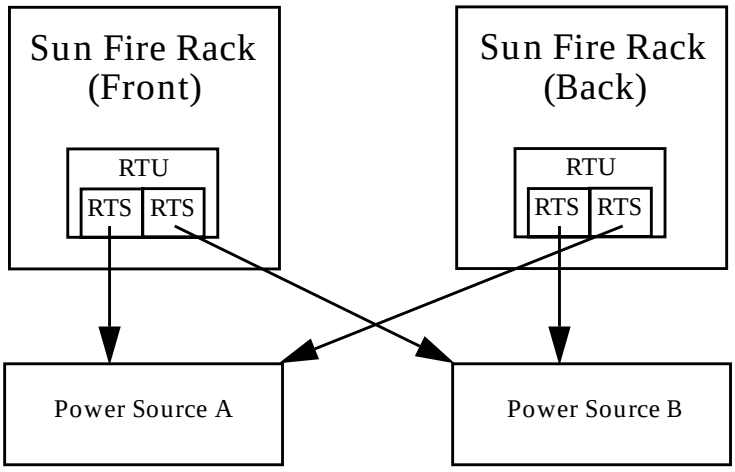

FIGURE 2 Sun Fire 6800 Rack With Two RTUs Configured

Refer to the *Sun Fire Cabinet Installation and Reference Manual* for more specific configuration and installation instructions.

For power failover capabilities, check that the racks for Sun Fire 3800, 4800, and 4810 systems are ordered with the Dual AC Input option which includes a second RTS for the RTU. The standard configuration may only have a single RTS in an RTU. As Sun Fire 6800 servers have higher power requirements and are internally divided into two separate power grids, these servers come standard with two RTUs and should also come configured with a full compliment of four RTS modules.

Be sure to consult the *Sun Fire 6800 System and Sun Fire Cabinet Rackmount Placement Matrix* document when adding supported devices to the Sun Fire Cabinet to ensure proper placement and power configurations.

Platform Configuration

This section covers system controller configuration, Sun StorEdge™ D240 Media Tray storage configuration, configuring segments and domains, and additional Sun Fire 6800 configuration.

System Controller Configuration

Overview:

- Configure the system with two system controllers.
- RS-232 Port should be accessible during the initial setup.
- Use a 100BaseT Ethernet connection.
- Put the System Controller on a switched, private network.
- Set up logging.
- Set SCPOST diag-level to “max.”
- Set platform and domain passwords.

All Sun Fire Midframe servers are controlled by a Sun Fire System Controller (SC). Configuration of the SC is the first step in the configuration of a Sun Fire server. A Sun Fire server requires only one SC, but can be configured with a second SC for capabilities such as redundant clock signals. In configurations which do not come with two SCs, a second SC is available through the Redundancy Kit option.

The SC can be accessed two ways—via the built-in RS-232 serial port or via its 10/100 Ethernet port. Access to the serial port should be available during the initial setup of the SC as this is the only connection where System Controller Power On Self Test (SCPOST) output can be viewed. The serial port can be accessed using a network terminal server or a serial port on an administration workstation. The port settings should be 9600 bps, 8 bits, no parity, 1 stop bit (9600 8N1).

Once the Ethernet port has been configured, it should be the primary access path to the SC. A telnet session is used to connect to the SC from the network. Access via the Ethernet port is faster than the serial port, and allows for multiple simultaneous connections to the SC. A 100BaseT link is strongly recommended for the SC Ethernet connection and required for use with Sun™ Management Center (SunMC). When in service, access to the serial port should be available to provide an alternate access path to the SC in the event of a network problem, or if the SC is rebooted or reset. Serial port access is also required to monitor certain SC and platform related errors as this is where these errors will be displayed. However, if only one connection is possible, the Ethernet port should be chosen as the primary connection path for the speed, multi-session access, and logging capabilities it provides.

For best performance, the SC should be configured on a switched, private network. FIGURE 3 illustrates a sample network topology. If configuring two SCs for the network, assign each SC a separate IP address so they do not conflict with each other on the network. A private network is also recommended in order to provide additional security for the SC.

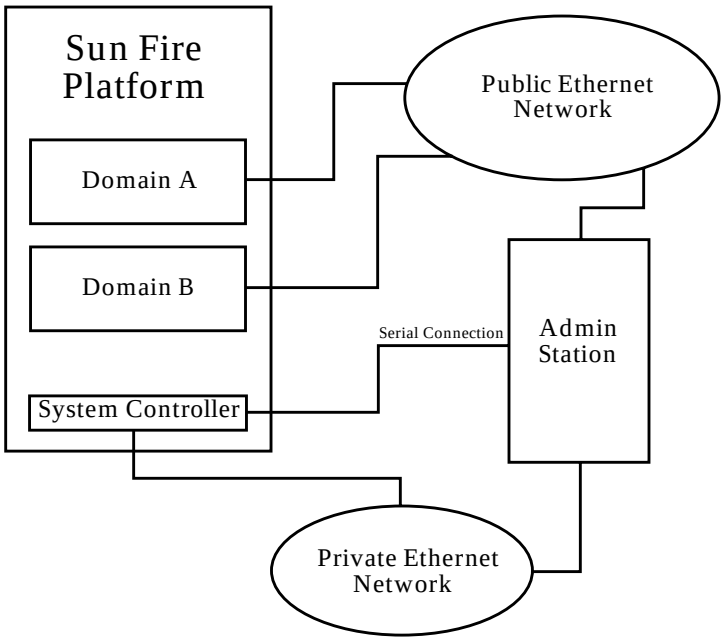

FIGURE 3 Simplified Network Topology

Once networking is established, enable logging to a `syslog` host. This is done using the `setupplatform` and `setupdomain` commands, and enables the SC to send error messages to another machine where it can be monitored and recorded.

```
heslab-12:SC> setupplatform -p loghost

Loghosts
--------
Loghost [129.146.63.105]: 129.146.63.105:local0

heslab-12:SC> console a

Connected to Domain A

heslab-12:A> setupdomain -p loghost

Loghosts
--------
Loghost [129.146.63.105]: 129.146.63.105:local1

heslab-12:A>
```

SNMP can also be enabled with the same commands for SunMC support as well.

```
heslab-12:SC> setupplatform -p snmp

SNMP
----
Platform Description [heslab-12.eng 4810]:
Platform Contact [james.hsieh@east]:
Platform Location [MPK12-2711]:
Enable SNMP Agent? [yes]:
Trap Hosts [129.154.221.220]:
Public Community String [P-public]:
Private Community String [P-private]:

heslab-12:SC> console a

Connected to Domain A
heslab-12:A> setupdomain -p snmp

SNMP
----
Domain Description [Domain A of heslab-12 (ME)]:
Domain Contact [james.hsieh@east]:
Trap Hosts [129.154.221.220]:
Public Community String [A-public]:
Private Community String [A-private]:

heslab-12:A>
```

Detailed setup information is discussed in the *Sun Fire Midframe Server Administration Best Practices* document and in the *Sun Fire 6800/4810/4800/3800 System Controller Command Reference Manual*

When setting up the SC, the SCPOST level should be set to maximum to ensure that the SC hardware is given a thorough testing before being placed into service. SCPOST output is available only through the serial port. The entry will look like:

```
heslab-12:SC> setupplatform -p sc

SC POST
-------
SC POST diag Level [max]: max

heslab-12:SC>
```

The platform and domain shells should be password protected using the password command in the SC.

```
heslab-12:SC> password

Enter new password:
Enter new password again:
heslab-12:SC>
```

From the platform shell, one has the ability to set or change the platform and domain shell passwords. From a domain shell, one can only change the password of the particular domain. SC security issues relating to SC setup and management are discussed in more detail in the Sun BluePrints™ OnLine article, S*ecuring the Sun Fire Midframe System Controller.*

Sun StorEdge D240 Media Tray Storage Configuration

Overview:

- Mirror boot drives.

The Sun StorEdge D240 Media Tray has been specially designed for use with systems within a Sun Fire rack. The Sun StorEdge D240 Media Tray is an Ultra-Fast-Wide Single Ended SCSI unit which contains two disk drives and space for two additional DDS-4 tape drives, DVD-ROM drives, or disk drives.

A typical configuration for a Sun StorEdge D240 Media Tray is two disk drives, a DVD-ROM drive, and a DDS-4 tape drive. The Sun StorEdge D240 Media Tray has two redundant power supplies and can also be set for a “split-bus” (two separate SCSI bus) configuration.

An ideal configuration utilizes two Sun StorEdge D240 Media Tray arrays with the boot drives mirrored between the two Sun StorEdge D240 Media Tray arrays and separate SCSI cards. This configuration is ideal since even in split-bus mode, the two

SCSI buses of a Sun StorEdge D240 Media Tray share a common centerplane, which could be considered a Single Point of Failure (SPOF). An example of an ideal setup is shown in FIGURE 4.

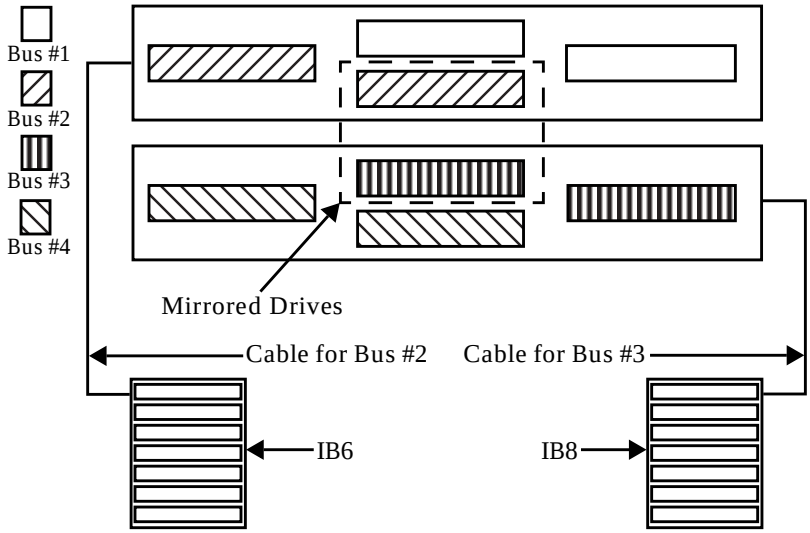

FIGURE 4 Mirroring Disks between two Sun StorEdge D240 Media Tray units and two controllers on two I/O Assemblies

If utilizing two Sun StorEdge D240 Media Tray units is not possible, the Sun StorEdge D240 Media Tray should be placed in “split-bus” mode, and the boot drive for the domain should be mirrored across the two halves of the Sun StorEdge D240 Media Tray and connected to separate SCSI cards. FIGURE 5 gives an example of this setup.

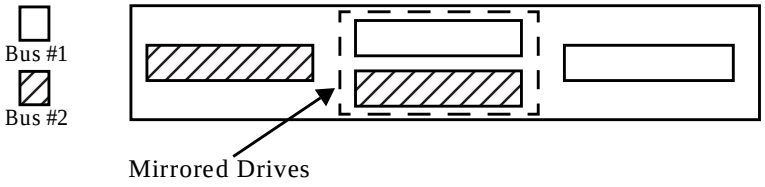

FIGURE 5 Mirrored Disks on Different Busses

In some configurations, it may not be possible to mount the Sun Fire server and all the desired boot devices in a single cabinet for power or space reasons. A separate rack and/or alternate storage choices may be required in these cases. Be careful to

not exceed maximum SCSI bus lengths if relocating a Sun StorEdge D240 Media Tray into a different cabinet as the Sun StorEdge D240 Media Tray uses Single-Ended UltraSCSI connections.

Configuring Segments and Domains

Overview:

- Configure each domain with component redundancy.
- For I/O, configure redundant paths across assemblies and I/O busses.
- Configure each domain in its own segment.

Once the SC has been set up, the next step is to configure the server into segments and domains. For the mission critical environment key considerations should be redundancy, the elimination of single points of failure, and the ability to quickly reconfigure around failures should a failure occur. To balance mission critical needs with application performance requirements, be sure that the operational requirements of the server are clearly understood before proceeding.

The Sun Fire server line incorporates the concept of “domains” or multiple independent copies of Solaris™ Operating Environment running on one system. Each domain requires a minimum of one System Board and one PCI or Compact PCI (cPCI) I/O Assembly in order to function. The 3800, 4800, and 4810 systems can have up to two domains while a 6800 system can have up to four domains. Configure each domain with as many redundant components (e.g. I/O paths, I/O assemblies, fans, etc.) to increase RAS. Availability considerations may thus limit the number of domains a system contains.

Each Sun Fire I/O assembly contains two separate I/O controller ASICs. Each ASIC provides one 66MHz bus and one 33MHz bus. Be sure I/O channels within a domain are distributed across I/O assemblies, the controllers, and their busses. TABLE 1, TABLE 2, and TABLE 3 illustrate the I/O Controller and bus distribution between the slots.

TABLE 1 Sun Fire 4800/4810/6800 8 Slot PCI I/O Assembly Layout

Slot	Slot Capacity	I/O Controller/Bus
0	33MHz 5 VDC (short)	0/B
1	33MHz 5 VDC (short)	0/B
2	33MHz 5 VDC	0/B
3	66/33 Mhz 3.3 VDC	0/A
4	33MHz 5 VDC	1/B

TABLE 1 Sun Fire 4800/4810/6800 8 Slot PCI I/O Assembly Layout *(Continued)*

Slot	Slot Capacity	I/O Controller/Bus
5	33MHz 5 VDC	1/B
6	33MHz 5 VDC	1/B
7	66/33MHz 3.3 VDC	1/A
<----Handle	Ejector levers---------->	

TABLE 2 Sun Fire 4800/4810/6800 4 Slot cPCI I/O Assembly Layout

Slot	3	2	1	0
Slot Capability	33 MHz 5 VDC	33 MHz 5 VDC	66/33 MHz 3.3 VDC	66/33 MHz 3.3 VDC
I/O Controller/Bus	1/B	0/B	1/A	0/A
Handle and Ejector levers------->				

TABLE 3 Sun Fire 3800 6 Slot cPCI I/O Assembly Layout

Slot	Slot Capability	I/O Controller/Bus
0	66/33 MHz 3.3 VDC	0/A
1	66/33 MHz 3.3 VDC	1/A
2	33MHz 5 VDC	0/B
3	33 Mhz 5 VDC	0/B
4	33MHz 5 VDC	1/B
5	33MHz 5 VDC	1/B

The latest *Sun Field Engineer Handbook* (available through Sun Enterprise Services) also contains diagrams showing which slots on an I/O Assembly are controlled by which bus.

The Sun Fire servers incorporate the concept of domains along with the new concept of “segments.” When a server is divided into two segments, the software in the SC logically isolates connections of one segment from the other, so that the failure of one domain within a particular segment will normally not affect the domain(s) running in the other segment (an unlikely exception where an error might not be contained would be in the case of a centerplane failure). Every Sun Fire Midframe server can be configured for a maximum of two segments. A Sun Fire 3800, 4800, or 4810 server has two Fireplane Switches (RP0 and RP2), and requires a minimum of one Fireplane Switch per segment. On the 3800, the Fireplane Switches are actually integrated with the centerplane. The diagrams in FIGURE 6 illustrate possible domain / segment combinations for these systems.

Sun Fire 3800/4800/4810 Single Segment Domain Allocation	
RP0	RP2
Segment 0	
Domain A	
Domain B	

Sun Fire 3800/4800/4810 Single Segment Domain Allocation[1]	
RP0	RP2
Segment 0	Segment 1
Domain A	Domain C (using Domain B MAC Addr)

1. As of June 2001, this is the behavior of 3800/4800/4810 systems when segmented.

FIGURE 6 Domain / Segment Combinations for Sun Fire 3800 / 4800 / 4810 and Two Fireplane Switches

A 6800 contains four Fireplane Switches (RP0, RP1, RP2, and RP3) and requires a minimum of two Fireplane Switches per segment. The diagrams in FIGURE 7 illustrate the possible domain/segment combinations.

Sun Fire 6800 Single Segment Domain Allocation			
RP0	RP1	RP2	RP3
Segment 0			
Domain A			
Domain B			

Sun Fire 6800 Single Segment Domain Allocation			
RP0	RP1	RP2	RP3
Segment 0		Segment 1	
Domain A		Domain C	
Domain B		Domain D	

FIGURE 7 Domain/Segment Combinations for Sun Fire 6800 server and Four Fireplane Switches

While it is possible for a single segment to contain two domains, the isolation of errors is the primary reason that segmentation is recommended for multiple domain configurations. In the case that customer requirements include multiple domains on any single Sun Fire system, the system should be divided into two segments so that each segment contains a minimum number of domains. If a Sun Fire system only has one domain, the system should remain in a single segment. System segments are configured on the SC by running setupplatform.

```
heslab-12:SC> setupplatform -p partition

Partition Mode

--------------

Configure chassis for single or dual partition mode? [dual]:

heslab-12:SC>
```

Multiple segments do not come completely without cost. Segmenting a system divides in half the theoretical maximum data bandwidth available to the domains within the segment, although it preserves snooping address bandwidth. However, even if performance is of the highest concern, dual segments may still be the preferable configuration. This is because the applications running on the domain may not be capable of exceeding the lowered bandwidth capacity, and may be affected more by the reduced snooping address bandwidth that two domains in a single segment will cause. For mission critical roles, two segments containing one domain within each segment is ALWAYS preferred over two domains within a single segment because of the error isolation features.

Segmenting a Sun Fire system into two segments also limits the reconfiguration options in case of a failure. In single segment mode, the failure of a Fireplane Switch set will cause the domain(s) in the segment to crash, but the system can be reconfigured around the failed switch set and brought back up on the surviving set. In a dual segment configuration, failure of a Fireplane Switch set will cause the domain(s) within that segment to fail, but since the Fireplane Switch sets in the system are already divided, there is no way to configure around the failed component. The domains associated with the failed Fireplane Switch will remain unavailable until the failed component is replaced.

In the case of the 6800, even though a single failed Fireplane Switch means there is a surviving switch in the segment, it is not possible to bring up the segment on one switch. The 6800 requires both Fireplane Switches to have a viable segment.

Additional Sun Fire 6800 Configuration

Overview:

- Keep all devices for a domain in the same power grid.

For the Sun Fire 6800 server, there is an additional consideration to configuring multiple domains in the system. The Sun Fire 6800 server differs from all other Sun Fire models in that it has two separate internal power grids, each supplied by a different RTU. The boards are separated as follows:

Power Grid 0:	Power Grid 1:
SB0	SB1
SB2	SB3
SB4	SB5
IB6	IB7
IB8	IB9
RP0	RP2
RP1	RP3

In a dual segment configuration, domains A and B are associated with RP0 and RP1, so boards for those domains should be chosen from Power Grid 0. Domains C and D are associated with RP2 and RP3, so boards for those domains should be chosen from Power Grid 1. Configuring domains to use boards from both power grids should be avoided so that in the unlikely event of an RTU failure, the effects of the power loss will be limited. If a domain depends on both power grids, there is twice the likelihood of a power failure causing the failure of the domain. This is illustrated by the following examples:

- Example 1:

  Domain A is created using RP0, RP1, SB0, and IB6. Domain C is created using RP2, RP3, SB1, and IB7. If the RTU supplying power grid 1 were to fail, domain C would fail, but there would be no effect on the operation of domain A. A failure of the RTU supplying grid 0 would cause domain A to fail, but domain C would remain available.

- Example 2:

  Domain A is created using RP0, RP1, SB1, and IB6. Domain C is created using RP2, RP3, SB0, and IB7. Because domain A was created using SB1 and domain C was created using SB0, the domain now has components residing in both power grids, so a failure of either RTU (or power grids) would cause all domains to fail.

The following `showboards -v` command output from a Sun Fire 6800 system demonstrates a proper division of boards between the two power grids for a dual segment system:

```
Slot       Grd    Pwr    Component Type              State        Status         Domain
----       --     --     -------------               ----         ----           -----
SSC0       -      On     System Controller           -            Passed         -
SSC1       -      -      Empty Slot                  -            -              -
ID0        -      On     Sun Fire 6800 Centerplane   -            -              -
PS0        0      On     A152 Power Supply           -            OK             -
PS1        0      On     A152 Power Supply           -            OK             -
PS2        0      On     A152 Power Supply           -            OK             -
PS3        1      On     A152 Power Supply           -            OK             -
PS4        1      On     A152 Power Supply           -            OK             -
PS5        1      On     A152 Power Supply           -            OK             -
FT0        0,1    On     Fan Tray                    Low Speed    OK             -
FT1        0,1    On     Fan Tray                    Low Speed    OK             -
FT2        0,1    On     Fan Tray                    Low Speed    OK             -
FT3        0,1    On     Fan Tray                    Low Speed    OK             -
RP0        0      On     Repeater Board              -            OK             -
RP1        0      On     Repeater Board              -            OK             -
RP2        1      On     Repeater Board              -            OK             -
RP3        1      On     Repeater Board              -            OK             -
/N0/SB0    0      On     CPU Board                   Assigned     Not Tested     A
/N0/SB1    1      On     CPU Board                   Assigned     Not Tested     C
/N0/SB2    0      On     CPU Board                   Assigned     Not Tested     A
/N0/SB3    1      On     CPU Board                   Assigned     Not Tested     c
/N0/SB4    0      On     CPU Board                   Assigned     Not Tested     A
/N0/SB5    1      On     CPU Board                   Assigned     Not Tested     c
/N0/IB6    0      On     PCI I/O Board               Assigned     Not Tested     A
/N0/IB7    1      On     PCI I/O Board               Assigned     Not Tested     C
/N0/IB8    0      On     CPCI I/O board              Assigned     Not Tested     A
/N0/IB9    1      On     PCI I/O Board               Assigned     Not Tested     C
```

Memory and I/O Configuration

Overview:

- Set memory interleave settings for “within-board” and “optimal.”
- Distribute additional memory evenly among CPUs.
- Balance I/O channels.

There are a number of items relating to performance which directly impact system availability. It is important to understand these issues, so proper decisions can be made in the configuration of the system. Overall performance of a customer system requires careful configuration of parts of the system other than the server (such as storage and network interfacing), which are beyond the scope of this article. This section addresses only issues related specifically to Sun Fire servers.

The best configuration for a mission critical Sun Fire 3800, 4800, and 4810 server is a single domain system. This allows for greater flexibility and redundancy in I/O layout and resiliency from failures. A single domain configuration also provides for the highest possible level of performance because both Fireplane Switches are available to the single domain to route address and data transactions across the system. All resources are dedicated solely for the use of the single domain and do not have to be shared. For a Sun Fire 6800 server, similar processing and I/O flexibility and redundancy can be achieved for two domains as long as each domain is in a separate segment; however, Fireplane Switch redundancy is lost by segmenting.

There are two memory interleave configuration options for a domain when the `setupdomain` command is run. They are set by the following two entries:

```
interleave-scope [within-board]: within-board
interleave-mode [optimal]: optimal
```

The default values should be “within-board” and “optimal,” respectively, and these are the values recommended for most configurations, including those which include multi-board domains. If all memory on a system board is comprised of the same sized memory DIMMs, memory access will be optimally distributed across all the available memory modules on the board. For optimal interleaving, one should populate an entire System Board with identically sized DIMMs (resulting in a 16-way interleave). If populating an entire System Board with memory is not possible, memory should be distributed as evenly as possible among the CPUs on all boards for best performance.

Setting interleave-scope to “across boards” should be avoided unless necessary, as it will prevent the domain from performing Dynamic Reconfiguration (DR) operations.

For I/O configuration, check that the capabilities of the cards that are being used match the capabilities of the slot into which the cards are being inserted as much as possible. Cards should be distributed between the two controller ASICs and their busses as much as possible for load balancing.

Conclusion

The design of the Sun Fire Midframe servers provide many features which allow for increased RAS. Proper installation and configuration planning should be done in order to take advantage of these features to prepare the Sun Fire Midframe server for its role in supporting mission critical applications.

References

- *Sun Fire 6800 System and Sun Fire Cabinet Rackmount Placement Matrix* (Sun Document #816-2062)
- *Sun Fire 6800/4810/4800/3800 System Controller Command Reference Manual* (Sun Document #805-7372)
- *Sun Fire 6800/4810/4800/3800 Systems Service Manual* (Sun Document #805-7363)
- *Sun Fire 6800 System Installation Guide* (Sun Document #805-7375)
- *Sun Fire 4810/4800/3800 Systems Installation Guide* (Sun Document #805-7370)
- *Sun Fire Cabinet Installation and Reference Manual* (Sun Document #806-2942)
- *Securing the Sun Fire Midframe System Controller* (Sun Blueprints OnLine September 2001)
- *Sun Field Engineer Handbook*

Author's Bio:

James Hsieh is a member of the Sun CPR Engineering Americas (formerly HES-CTE) group and is part of the team responsible for providing engineering support of Sun's midrange server line. Prior to his current role with CPR Engineering, James worked for Sun Enterprise Services supporting Mission Critical customers.

Prior to Sun, James worked for over ten years with UNIX and Sun systems as a software engineer and as a systems administrator for large groups of UNIX systems.