Running Linux on HPE Integrity Superdome X

Configuring Linux on HPE Integrity Superdome X

Contents

Executive summary

The purpose of this white paper is to facilitate the successful deployment and configuration of Linux® operating systems on an HPE Integrity Superdome X with HPE Superdome BL920s Gen8 and Gen9 Server Blades. It includes a description of recommended HPE hardware and software components for HPE Integrity Superdome X and best practices you may use when deploying Linux—specifically Red Hat® Enterprise Linux and SUSE Linux Enterprise Server.

Overview of HPE Integrity Superdome X with HPE BL920s Gen8 and BL920s Gen9 Server Blades

The following section provides an overview of the hardware and software requirements for the HPE Integrity Superdome X with HPE Superdome BL920s Gen8 and Gen9 Server Blades.

Hardware requirements and features

HPE BladeSystem Superdome Enclosure is the building block for Superdome X. Each compute enclosure supports 15 fans, 12 power supplies, associated power cords, and four HPE Crossbar Fabric Modules (XFMs). Configurations of one-to-eight mission-critical, scalable x86 blades can be populated in an enclosure with support for hard partitions (HPE nPars) containing one, two, three, four, six, or eight blades. Multiple nPars of different sizes are supported within a single enclosure. This partitioning flexibility allows the servers to be right sized for any workload.

Note

The HPE Integrity Superdome X QuickSpecs is available at: hpe.com/h20195/v2/GetDocument.aspx?docname=c04383189.

Hardware specifications

Table 1. Supported processor configurations for both the BL920s Gen8 and Gen9 Server Blades

	Processor	Number of cores per processor	Frequency	Cache	Power
BL920s Gen8	Intel® Xeon® Processor E7-2890 v2	15c	2.8 GHz (3.4 GHz Turbo)	37.5 MB	155W
	Intel Xeon Processor E7-2880 v2	15c	2.5 GHz (3.1 GHz Turbo)	37.5 MB	130W
	Intel Xeon Processor E7-4830 v2	10c	2.2 GHz (2.7 GHz Turbo)	20 MB	105W
	Intel Xeon Processor E7-8891 v2	10c	3.2 GHz (3.7 GHz Turbo)	37.5 MB	155W
	Intel Xeon Processor E7-8893 v2	6c	3.4 GHz (3.7 GHz Turbo)	37.5 MB	155W
BL920s Gen9	Intel Xeon Processor E7-8890 v3	18c	2.5 GHz (3.3 GHz Turbo)	45 MB	165W
	Intel Xeon Processor E7-8880 v3	18c	2.3 GHz (3.1 GHz Turbo)	45 MB	150W
	Intel Xeon Processor E7-4850 v3	14c	2.2 GHz (2.8 GHz Turbo)	35 MB	115W
	Intel Xeon Processor E7-8891 v3	10c	2.8 GHz (3.5 GHz Turbo)	45 MB	165W
	Intel Xeon Processor E7-8893 v3	4c	3.2 GHz (3.5 GHz Turbo)	45 MB	140W
	Intel Xeon Processor E7-8890 v4	24c	2.2 GHz (2.6 GHz Turbo)	60 MB	165W
	Intel Xeon Processor E7-8894 v4	24c	2.4 GHz (3.4 GHz Turbo)	60 MB	165W
	Intel Xeon Processor E7-8880 v4	22c	2.2 GHz (2.6 GHz Turbo)	55 MB	150W
	Intel Xeon Processor E7-8855 v4	14c	2.1 GHz (2.4 GHz Turbo)	35 MB	140W
	Intel Xeon Processor E7-8891 v4	10c	2.8 GHz (3.3 GHz Turbo)	60 MB	165W
	Intel Xeon Processor E7-8893 v4	4c	3.2 GHz (3.4 GHz Turbo)	60 MB	140W

Memory

Table 2. Supported memory configurations for Superdome X

	BL920s Gen8	BL920s Gen9 (v3)	BL920s Gen9 (v4)
Double data rate type	DDR3	DDR4	DDR4
Max. memory per complex	12 TB	24 TB	24 TB
DIMM slots per server blade	48	48	48
DIMM sizes	16 GB, 32 GB	16 GB, 32 GB, 64 GB	32 GB, 64 GB

I/O adapters

✓ = Supported
X = Not supported

Table 3. I/O adapters

Type	Product	BL920s Gen8	BL920s Gen9 (v3)	BL920s Gen9 (v4)
Mezzanine Ethernet adapter	HPE Ethernet 10GB 2-port 560M Adapter	✓	✓	✓
Mezzanine FlexFabric Ethernet adapters	HPE FlexFabric 10GB 2-port 534M Adapter (700748-B21)	✓	X	X
	HPE FlexFabric 20Gb 2-port 630M Adapter (700076-B21)	X	✓	✓
	HPE FlexFabric 20Gb 2-port 650M Adapter (700767-B21)	X	X	✓
LOM Ethernet adapter	HPE Ethernet 10Gb 2-port 560FLB Adapter (655639-B21)	✓	✓	✓
LOM FlexFabric Ethernet adapters	HPE FlexFabric 10Gb 2-port 534FLB Adapter (700741-B21)	✓	X	X
	HPE FlexFabric 20Gb 2-port 630FLB Adapter (700065-B21)	X	✓	✓
	HPE FlexFabric 20Gb 2-port 650FLB Adapter (700763-B21)	X	X	✓
Mezzanine Fibre Channel Adapter	HPE QMH2672 16GB Fibre Channel Host Bus Adapter (710608-B21)	✓	✓	✓
InfiniBand Controller	HPE InfiniBand FDR 2-port 545M Adapter (702213-B21)	X	✓	✓

Note

The 650M and 650FLB are supported currently only for use in NIC, NIC+RoCE, and iSCSI mode on the BL920S Gen9 (v4 only). No FCoE deployments should be attempted. When installing with RHEL 7.2 and SLES 12 SP1 and prior releases, please install the operating systems with the be2iscsi driver blacklisted.

- For RHEL 7.2 or below, install adding "`rdblacklist=be2iscsi modprobe.blacklist=be2iscsi`" to the boot command line.

- For RHEL 7.3, install adding "`rd.driver.pre=be2iscsi rd.iscsi.firmware=1`" to the boot command line when booting from iSCSI LUN.

- For SLES 11 SP4, install adding "`brokenmodules=be2iscsi`" to the boot command line.

- For SLES 12, SLES 12 SP1, install adding "`modprobe.blacklist=be2iscsi`" to the boot command line.

For RHEL 7.5, add the "nopti" option to the kernel command line using the boot loader. Once the system finishes the installation and reboots, add the "nopti" option to the kernel command line again. Once the system boots up, upgrade the kernel to 3.10.0-862.1.1 or its later version. Once system boots to this kernel, "nopti" can be safely removed from the command line.

RoCE is supported with the 650FLB and 650M network adapters on the BL920s Gen9 (v4 only) with RHEL 7.2 and higher releases.

iSCSI is only supported with the 650FLB on the BL920s Gen9 (v4) with RHEL 7.3 with firmware recipe 4.62 or higher. Ensure that the firmware on your SDX enclosure is at this revision level before enabling iSCSI on the systems. Refer to the HPE Integrity Superdome X IO Firmware and Windows® Drivers Image section of this paper for more details on firmware updates.

A 16-socket (eight-blade) Superdome X system may be configured with 1 to 24 dual-port Fibre Channel mezzanine controllers. There may also be between 8 and 16 dual-port network LOM adapters; one is required in LOM slot 1 while an optional one may be utilized in LOM slot 2. It is also possible to add up to 24 additional network mezzanine controllers supported in any available mezzanine slot. It should be noted to not configure different brands of I/O cards to the same interconnect in the same blade. An eight-socket system (four-blade) configuration has similar constraints. In all case, each server blade supports a maximum of two LOMs and three mezzanine cards.

Interconnect modules

Table 4 outlines the supported interconnect modules for Superdome X with BL920s Gen8 and Gen9 Server Blades.

✓ = Supported

X = Not supported

Table 4. Interconnect modules

Type	Product	BL920s Gen8	BL920s Gen9 (v3)	BL920s Gen9 (v4)
Ethernet interconnect	HPE 6125XLG Ethernet Blade Switch (711307-B21)	✓	✓	✓
	HPE 6125G Ethernet Blade Switch (658247-B21)	✓	✓	✓
	HPE 6127XLG Ethernet Blade Switch (787635-B21)	X	✓	✓
Ethernet interconnect (pass-thru)	HPE 10GbE Ethernet Pass-Thru Module for c-Class BladeSystem (538113-B21)	✓	✓	✓
Fibre Channel interconnect	Brocade 16Gb/16 SAN switch for BladeSystem c-Class (C8S45A)	✓	✓	✓
	Brocade 16Gb/28 SAN Switch Power Pack+ for BladeSystem c-Class (C8S47A)	✓	✓	✓
Fabric Extender	HPE Cisco B22HP Fabric Extender with 16 FET for BladeSystem c-Class (657787-B21)	✓	✓	✓
	HPE Cisco B22HP Fabric Extender for BladeSystem c-Class (641146-B21)	✓	✓	✓
InfiniBand interconnect	HPE 4X FDR InfiniBand Switch (648312-B21)	X	✓	✓
	HPE 4X FDR Managed Switch Module for c-Class BladeSystem	X	✓	✓

One to four Fibre Channel interconnects may be installed in bays three through eight in the back of the enclosure, dependent upon the mezzanine slots being utilized. One of the Ethernet interconnects must be installed in bay one; a second Ethernet interconnect of the same type may be installed in bay two if desired. If additional Ethernet adapters are used, the corresponding interconnect modules of the same type may be installed in bays three through eight—again dependent on the mezzanine slots utilized.

See the configuration options table in the HPE Integrity Superdome X QuickSpecs.

Server configuration

Server partitioning (HPE nPars)

The HPE Integrity Superdome X may be partitioned into different mixes of HPE nPars. The configurations specified herein allow the customer to hard partition their HPE Superdome X into the configuration of 2-socket, 4-socket, 6-socket, 8-socket, 12-socket, or 16-socket configurations that best suits their computing needs.

Table 5. Supported HPE nPars configurations

Supported HPE nPars	BL920s Gen8	BL920s Gen9 (v3)	BL920s Gen9 (v4)
2s	✓	✓	✓
4s	✓	✓	✓
6s	X	✓	✓
8s	✓	✓	✓
12s	X	X	✓
16s	✓	✓	✓

Maximum configurations

Table 6 outlines the maximum configurations for the BL920s Gen8 and BL920s Gen9 Server Blades.

Table 6. Maximum configurations

BL920s Gen8	16-socket	12-socket	8-socket	6-socket	4-socket	2-socket
Processor cores (15 cores per socket)	240	N/A	120	N/A	60	30
Logical processors (Hyper-Threading—enabled)	480	N/A	240	N/A	120	60
Memory capacity (32 GB DIMMs)	12 TB	N/A	6 TB	N/A	3 TB	1.5 TB
Mezzanine slots (current supported limit)	24 (24)	N/A	12 (12)	N/A	6 (6)	3 (3)
LOM	16	N/A	8	N/A	4	2
BL920s Gen9	**16-socket**	**12-socket**	**8-socket**	**6-socket**	**4-socket**	**2-socket**
Processor cores (24 cores per socket)	384	288	192	144	96	48
Logical processors (Hyper-Threading—enabled)	768	576	384	288	192	96
Memory capacity (64 GB DIMMs)	24 TB	18 TB	12 TB	9 TB	6 TB	3 TB
Mezzanine slots (current supported limit)	24 (24)	18 (18)	12 (12)	9 (9)	6 (6)	3 (3)
LOM	16	12	8	6	4	2

It is recommended for Linux customers to review the following guidelines when configuring the HPE Integrity Superdome X BL920s Gen8 and Gen9 Server Blades:

Firmware-first PCIe error containment and recovery
HPE Integrity Superdome X facilitates the containment of errors in the I/O subsystem, through the firmware-first PCIe error handling, PCI Advanced Error Reporting (AER), and Live Error Recovery (LER) mechanisms. I/O errors that might cause data corruption on other systems are safely isolated by the Superdome X hardware and firmware. In cases of transient I/O errors, the error containment feature allows Linux to continue operation. Hewlett Packard Enterprise is continually working with I/O adapter partners to increase the number of error cases that may be successfully recovered.

Live Error Recovery
Live Error Recovery (LER) is a hardware feature supported on high-end x86 Intel® processors that enable error containment and recovery of PCIe errors that would otherwise cause the system to crash and potentially corrupt data. Recovery from PCIe uncorrectable errors is important to the Superdome X because these systems support many more PCIe devices than most Linux servers do. The potential for hardware errors increases proportionally to the number of PCIe devices. Online recovery from such errors is critical to Hewlett Packard Enterprise customers, especially those who support large I/O deployments.

Containment
The biggest concern when an I/O error occurs is preventing data corruption. This may occur whether or not the device recovers from the error. If a device fails in the middle of a large file transfer and an error occurs, the device could potentially write corrupt values to the file before the driver is notified of the problem. With firmware-first mechanism and Intel's LER processor feature, errors are immediately contained before a corruption occurs; all reads and writes are halted until the driver is notified that an appropriate action may be taken.

Recovery
Our goal is for the system to handle PCIe I/O errors with no intervention from the user and to allow workloads to continue without interruption. If a large file is being transferred over the network device via FTP and an error occurs on the network device, the file transfer will pause momentarily as the device recovers from the error and then resumes operations. The completed file transfer will be intact and free of corruption.

Recovery of a specific I/O device will largely be dependent upon the capability of its driver. It should be noted that all supported drivers on the Superdome X have some level of PCIe I/O error recovery enabled. A large percentage of PCIe I/O errors will be recovered and normal operations will resume. There is a subset of errors and configurations where recovery will not occur.

In these cases, the system may panic or the I/O device will become deconfigured. In general, PCIe I/O errors are unlikely and the situation where they are not recoverable is even less likely to occur, the odds of the server going down or losing functionality is relatively low. Hewlett Packard Enterprise is working closely with our I/O partners and the greater Linux community to continue PCIe I/O error recovery enablement. As drivers continue to mature, more PCIe I/O errors will be recoverable.

Error logging
Even though PCIe I/O errors rarely happen, it is important for the user to be notified when they occur. If a specific device continually experiences errors, it may be an indication that the device is faulty and needs to be replaced.

Linux AER trace events
Hewlett Packard Enterprise has worked with the Linux open source community to provide an enhanced mechanism to report errors using kernel trace events. When a PCIe error occurs, an event is triggered to notify any user space applications programmed to intercept the event. Trace events also offer a user-readable log providing details about the error.

Syslog
The user may examine Linux system logs to determine that an error occurred and diagnose the problem. The syslog has output from the kernel and drivers regarding the error.

OA live logs
If the system goes down and the OS is not available, the user may use the OA system event log (SEL) and forward progress log (FPL) to see if an error occurred. These errors are also sent to the remote management systems such as HPE Insight Remote Support (IRS).

Software overview

Operating systems

HPE Superdome X supports the following Linux distributions:
✓ = Supported
X = Not supported

Table 7. Operating system details

Linux distribution	BL920s Gen8	BL920s Gen9 (v3)	BL920s Gen9 (v4)	Technical exceptions
RHEL 6.5 with maintenance update kernel (24 June 2014 or later)	✓	X	X	It is required to use RHEL 6.5.z kernel-2.6.32-431.20.3.el6 (or later version). See RHSA-2014-0771 for details
RHEL 6.6	✓	✓	X	
RHEL 6.7	✓	✓	✓	
RHEL 6.8	✓	✓	✓	
RHEL 6.9	✓	✓	✓	
RHEL 7.0	✓	X	X	
RHEL 7.1	✓	✓	X	
RHEL 7.2	✓	✓	✓	It is required to use RHEL 7.2.z kernel-3.10.0-327.18.2.el7 (or later version). See RHSA-2016:1033 for details
RHEL 7.3	✓	✓	✓	
RHEL 7.4	✓	✓	✓	
RHEL 7.5	✓	✓	✓	It is required to use RHEL 7.5.z kernel-3.10.0-862.2.3 (or later version) RHSA-2018:1318
SLES 11 SP3	✓	✓	X	kISO required drivers.suse.com/hpe/hpe-Integrity-superdome-X/
SLES 11 SP3 for SAP®	✓	✓	X	
SLES 11 SP4	✓	✓	✓	
SLES 12	✓	✓	X	kISO required drivers.suse.com/hpe/hpe-Integrity-superdome-X/
SLES 12 SP1	✓	✓	✓	kISO required drivers.suse.com/hpe/hpe-Integrity-superdome-X/
SLES 12 SP2	✓	✓	✓	
SLES 12 SP3	✓	✓	✓	

Note

Refer to the HPE OS support matrices and minimum OS requirements matrices for Red Hat and SUSE for the latest supported OS and technical requirements.

HPE/Red Hat matrices

OS

Minimum OS requirements

HPE/SUSE matrices

OS

Minimum OS requirements

HPE ConvergedSystem 900 with SAP HANA

The HPE ConvergedSystem 900 is exceptionally well suited to run SAP HANA® due to its impressive compute power and generous memory capacity. For HANA deployments, the version of Linux that is needed is SLES for SAP or RHEL for SAP. HANA is supported on the Superdome X configured as an HPE ConvergedSystem 900 in Scale-up, Scale-out deployment model or Tailored Datacenter Integration (TDI). More information is available at hpe.com/us/en/product-catalog/detail/pip.hpe-convergedsystem-900-for-sap-hana-scale-up-configurations.7311041.html.

Customers using the CS900 product for RHEL for SAP or SLES for SAP should be aware of the following information:

1. RHEL for SAP Business Applications

2. Information and Installation guides for SAP HANA on RHEL

3. SLES 11 for SAP with Installation Kit

4. Information and Installation guides for SAP HANA on SUSE Linux Enterprise Server

5. SAP HANA fully supports Superdome X nPars partitioning; see SAP note 2103848 for details.

WBEM Providers

Linux Web-Based Enterprise Management (WBEM) Providers are based on the SMX 3.0 technology for RHEL 6/7 and SLES 11/12 running on HPE Integrity Superdome X servers. Current releases include the WBEM Providers, tools, required registration files, and installation scripts bundled into several RPM packages for easy installation on Linux systems.

The latest version of the WBEM Providers for SLES and RHEL OSs are available from the HPE Software Delivery Repository (SDR) site. These providers will run on both the BL920s Gen8 and Gen9 servers. Review the operating systems information in Table 7 to ensure the correct supported OSs for each platform.

Important note

WBEM Providers are delivered on the Software Delivery Repository site only. The Service Pack for ProLiant (SPP) does not support Superdome X. The SPP firmware is not intended to be used with the HPE Integrity Superdome X servers as well as software like hp-snmp-agents, hp-health, hp-diags, hpssa. Please do not install the SPP on BL920s Gen8 or BL920s Gen9 server blades.

RHEL:

- RHEL 6.5 WBEM Providers

- RHEL 6.6 WBEM Providers

- RHEL 6.7 WBEM Providers

- RHEL 6.8 WBEM Providers

- RHEL 6.9 WBEM Providers

- RHEL 7.0 WBEM Providers

- RHEL 7.1 WBEM Providers

- RHEL 7.2 WBEM Providers

- RHEL 7.3 WBEM Providers

- RHEL 7.4 WBEM Providers

- RHEL 7.5 WBEM Providers

- RHEL 6.6/7.1 (or later releases) InfiniBand WBEM Providers

Installation instructions for RHEL: downloads.linux.hpe.com/sdr/repo/bl920-wbem/rhel/install_guide_rhel.txt.

SLES:

- SLES 11 SP3 WBEM Providers

- SLES 11 SP4 WBEM Providers

- SLES 12 WBEM Providers

- SLES 12 SP1 WBEM Providers

- SLES 12 SP2 WBEM Providers

- SLES 12 SP3 WBEM Providers

- SLES 11 SP4/SLES 12 SP2 (or later releases) InfiniBand WBEM Providers

Installation instructions for SLES 11 and SLES 12: downloads.linux.hpe.com/sdr/repo/bl920-wbem/suse/install_guide_sles11_and_sles12.txt.

Linux WBEM providers of HPE Integrity Superdome X Release Notes: downloads.linux.hpe.com/sdr/repo/bl920-wbem/integrity%20 superdome %20x%20linux%20wbem%20providers%20release%20notes.pdf.

Note
The InfiniBand WBEM Provider is only supported on the BL920s Gen9 for RHEL 6.6/7.1 or later and SLES 11 SP4/SLES 12 SP2 or later installations. Please see Table 7 for detailed information on supported OSs for your server blade.

HPE Integrity Superdome X IO Firmware and Windows Drivers Image
Customers should utilize the HPE Integrity Superdome X IO Firmware and Windows Drivers Image for updating firmware on the HPE BL920s Gen8 and Gen9 Server Blades. The HPE Integrity Superdome X IO Firmware and Windows Drivers Image is available at: h20565.www2.hpe.com/hpsc/swd/public/readIndex?sp4ts.oid=7595549, under "Software-CD ROM." Despite the name of the image, this package may be used to update firmware regardless of any installed OS.

Updated I/O drivers for Superdome X supported I/O are available in the Linux distributions and are not contained in the I/O image. Note that some OS versions require a special installation kit provided by the OS vendor noted in the OS table (Table 7).

Virtualization solutions
Kernel Virtual Machine
Kernel Virtual Machine (KVM) is a kernel-integrated full virtualization solution available with RHEL and SLES distributions. The HPE BL920s Gen8 and Gen9 Server Blades may be used with KVM. For more information regarding KVM, review the virtualization section of the release notes for your preferred operating system.

Red Hat Enterprise Virtualization
Red Hat also offers Red Hat Enterprise Virtualization (RHEV), a complete virtualization management solution for virtualized servers and desktops. For more information on RHEV, visit Red Hat's website at: redhat.com/en/resources/red-hat-enterprise-virtualization. Hewlett Packard Enterprise is investigating support of RHEV on the HPE BL920s Gen8 and Gen9 Server Blades. Updates will be provided when merited.

Serviceguard for Linux
HPE Serviceguard for Linux (SGLX) is a high-availability and disaster recovery solution that provides business continuity for your critical applications. SGLX is designed to address the financial, business, and operational impacts of planned and unplanned downtime in demanding x86 or Linux environments. Visit hpe.com/us/en/product-catalog/detail/pip.376220.html for the latest information on SGLX releases from Hewlett Packard Enterprise.

The HPE Serviceguard for Linux certification matrix, available at support.hpe.com/hpsc/doc/public/display?docId=emr_na-a00018054en_us reflects support for HPE Superdome X products and supported OS releases.

Recommended hardware configuration

Hyper-Threading

The Intel Xeon processor E7 v4, v3, and v2 families implement the Intel Hyper-Threading technology feature. When enabled Hyper-Threading allows a single physical processor core to run two independent threads of program execution. The operating system may use both logical processors, which often results in increased system performance. However, the two logical processors incur some overhead while sharing common physical resources, so the pair neither has the same processing power, nor consumes the same electrical power as the two physical cores. Beyond that, the increased parallelism may cause increased contention for shared software resources, so some workloads may experience a degradation when Hyper-Threading is enabled.

On the HPE BL920s Gen8 and Gen9 Server Blades, the Hyper-Threading state is an attribute of the partition, therefore, it is controlled through interaction with the Onboard Administrator (OA). Hyper-Threading is enabled by default. The current state may be ascertained through the OA **parstatus** command:

```
Parstatus -p <partition identifier> - V
Hyper Threading at activation: Enabled
Hyper Threading at next boot: Enabled
```

Note

Both the current state and the state after the next boot are shown; any change to the Hyper-Threading state will not take effect until the partition is rebooted.

The OA parmodify command may be used if it is desired to change the Hyper-Threading state:

```
parmodify -p <partition identifier> -T [y|n]
```

As a general thumb rule, it may be desirable to change Hyper-Threading from the disabled to the enabled state if the workload that is running is fully utilizing all of the physical cores, as evidenced by a high value for CPU utilization. Another case where Hyper-Threading would be beneficial is if the physical cores are spending a relatively large amount of time, waiting for cache misses to be satisfied.

If there are applications running that tend to consume the entire capacity of the processor cache, it may be beneficial to disable Hyper-Threading. The two logical processors share a common physical cache so there may be contention with each other if the application cache footprint is large.

Performance analysis tools available through the Hewlett Packard Enterprise support teams may be used to gain insight into the benefit or detriment of using Hyper-Threading. You may want to experiment running your workloads in both states to see which one works best for your particular application.

Memory

For performance reasons, it is highly recommended to have the same amount of memory on each blade in a partition.

Power management

The HPE Integrity Superdome X implements a variety of mechanisms to govern the trade-off between power savings and peak performance. A more thorough discussion of other power per performance mechanisms appears in these references:

- Red Hat Enterprise Linux 6.x Power Management Guide

- Red Hat Enterprise Linux 7.x Power Management Guide

- SUSE Linux Enterprise Server 11 SPx System Analysis and Tuning Guide—Chapter 11

- SUSE Linux Enterprise Server 12 System Analysis and Tuning Guide

I/O

It is recommended that all disks connected to the Superdome X over the two-port QLogic Fibre Channel adapters be set up multipathed for high reliability and hardware availability.

Recommended software configuration and tuning

Operating system

Linux will run perfectly well out of the box on most small servers, but some minor administrative adjustments are recommended to get the highest value from the considerable hardware resources available with the Superdome X.

Installing on a QLogic Fibre Channel device

To enable boot from a QLogic Fibre Channel device, the boot controller needs to be enabled from the HPE Device Manager. Hewlett Packard Enterprise recommends enabling only the primary and secondary boot controllers to avoid delays during system initialization.

1. When boot reaches the **EFI Boot Manager** menu, press **U** for **System Utilities**

2. Select **Device Manager**

3. For each HPE QMH2672 16Gb 2P FC HBA—FC:

    a. Select the **device**

    b. Select **Boot Settings**

    c. Toggle **Adapter Driver to <Enabled>**

    d. Toggle **Selective Login to <Disabled>**

    e. Toggle **Selective LUN Login to <Disabled>**

    f. Ctrl-W to **save**

    g. Ctrl-X to exit the **Device Manager**

Always enable multipath during the install

Hewlett Packard Enterprise recommends enabling the Linux device mapper (DM) multipath during the Linux operating system installation. The DM multipath provides redundancy, reliability, and improved I/O performance.

Check whether the chosen Linux version has inbuilt support for the chosen SAN storage solution. If not, it may be necessary to edit /etc/multipath.conf so it includes the appropriate parameters as advised by the storage vendor.

Don't use IPv6

Configuration of IPv6 networks on the Superdome X is not supported, and it is recommended that IPv6 be disabled during installation.

Trim EFI boot entries

Customers are advised to remove stale or unwanted boot options to save boot time. Use "`efibootmgr`" to remove the preboot execution environment (PxE) boot options, when the OS is installed. And remove the boot options of the old OS after installing a new one.

Configure serial console

To enable the serial console, add the following to the kernel command line:

```
Console=ttyS0,115200n8
```

Note

Enabling serial console will result in a small performance penalty at boot time.

Ensure the baud rate of the serial console on your partition is set correctly from the OA interface. The default setting for the baud rate for each blade when it is the first blade in a partition is set at the factory to 9600. Increase the baud rate to run at 115200.

In order to do this, determine the blade number of the lowest numbered blade in your partition. If you have made a partition consisting of blades 1, 3, 5, and 7 this would be blade 1. If the partition consists of blade 6 and blade 8, the lowest blade would be 6. Check the baud setting of the blade's serial console by typing at the OA prompt:

```
SET SCRIPT MODE ON

HPONCFG <blade # of lowest numbered blade in the partition> <<EOF
```

Then copy and paste in the following:

```
<RIBCL VERSION="2.0">

<LOGIN USER_LOGIN="username" PASSWORD="password">

<RIB_INFO MODE="read">

<GET_GLOBAL_SETTINGS/>

</RIB_INFO>

</LOGIN>

</RIBCL> EOF
```

If the user scrolls back after all the output completes, there is a dump of all the global settings:

```
<GET_GLOBAL_SETTINGS>

<SESSION_TIMEOUT VALUE="30"/>

…..

<SERIAL_CLI_SPEED VALUE="9600"/>

…..

<!-- ======== END RIBCL RESULTS ======== -->
```

Look for the SERIAL_CLI_SPEED VALUE in the list, and as shown earlier if it says, it's 9600, which is the factory default, you need to change it to 115200.

To do this, type from the OA prompt:

```
HPONCFG <blade#> << EOF

Then cut and paste in the following

<RIBCL VERSION="2.0">

<LOGIN USER_LOGIN="username" PASSWORD="password">

<RIB_INFO MODE="write">

<MOD_GLOBAL_SETTINGS>

<SERIAL_CLI_SPEED value="5"/>

</MOD_GLOBAL_SETTINGS>

</RIB_INFO>

</LOGIN>

</RIBCL> EOF
```

You may perform the read operation after the above operation completes to verify if the baud rate is now set to 115200.

The iLO baud rate setting is local to the blade and moves with the blade. Applying the same change to other blades is necessary if:

1. Blades are physically swapped, or

2. Partition is reconfigured to add or remove blades.

Enable timestamps for debugging

For RHEL 6.x (BL920s Gen8 only), add kernel parameter "`printk.time=y`" to the kernel command line in order to add kernel timestamps. This is useful for debug purposes.

Note

It is not necessary to enable this for SLES configurations as it is already enabled by default.

Adjust EFI memory map

For RHEL 6.5 (BL920s Gen8 only), add "`add_efi_memmap`" to the kernel command line.

Maintaining the real-time clock

The real-time clock (RTC) in a partition is synced to OA time when the partition is rebooted. If you change the time on the OA, it may affect the RTC on the partition.

It is recommended to configure network time protocol (NTP) on both the OA (as described in the OA User Guide) and the Linux partition. By configuring NTP on both, the time should be correct on initial boot of the nPars. The firmware sets it correctly, and it should stay correct as Linux runs by virtue of ongoing synchronization with the configured NTP servers.

For further information on configuring NTP:

*   RHEL 6.x

*   RHEL 7.x

*   SLES 11

*   SLES 12

Kernel tuning

Specific recommendations for tuning Linux installations may be found at the following websites:

*   RHEL 6.x

*   RHEL 7.x

*   SLES 11

*   SLES 12

NUMA considerations

The Linux operating system includes algorithms to place memory close to the processor requesting access. If the memory reference patterns of a particular application are well characterized, it may be possible to use specific tuning to achieve memory placement that is superior to the operating system's default behavior.

Given the NUMA characteristics of Superdome X, latency of accessing memory that is local to a given processors' NUMA node is the lowest, followed by accessing memory off a buddy NUMA node on the same blade, followed by accessing memory off any NUMA node on a remote blade. It is important for applications to have their tasks and memory placed closer to one another. Doing so shall enable lower latency for memory accesses and result in better and predictable performance.

The Linux operating system includes algorithms to place memory close to the processor that accesses it. However, the applications tasks or memory may migrate to other NUMA nodes over time, resulting in a drop (or variability) in performance. Linux provides various tools such as `numactl (8)`, `taskset (1)`, and `cpusets (7)` to help the user bind a given workload's tasks or memory to desired NUMA nodes. Tools such as `numastat (8)` may be used to monitor a given process or workload's memory usage across various NUMA nodes. For longer running workloads, user level tools such as numad (available in RHEL 6.5 and later releases) may be used either as an advisory service or as an executable to monitor the workload characteristics and try to move its tasks or memory closer to another.

NUMA locality of I/O adapters plays a crucial role in the I/O performance. The irqbalance daemon, which runs by default, has the tendency to spread interrupt requests (IRQs) without paying attention to NUMA locality of the I/O adapters. Depending on the workload's characteristics, this may lead to poor performance. For the highest I/O performance, it is important to keep all the IRQs associated with a given I/O adapter distributed amongst the CPUs belonging to the local NUMA node. The irqbalance daemon should be disabled.

From the Linux host OS point of view, KVM virtual machines are an example of a long-running multi-threaded application. By specifying bindings for the virtual machines' resources to a specific set of host resources, we enable better and predictable performance for the workloads in the virtual machine. There is support available in libvirt/virsh (and numad) to help accomplish the same.

Scale-up workloads need to pay special attention to their data structure design or layout, granularity of locking, and the design of any application-specific locking primitives. Care must be taken to avoid too much bouncing of cache lines that need to be exclusively accessed. Too much cache line bouncing on any large symmetric multiprocessing (SMP) or NUMA systems may severely impact application scaling.

AutoNUMA is another tool that can help in this area. The objective of AutoNUMA is to avoid the need for the administrator to set up hard bindings and to solve the more difficult problems faced by the user space NUMA job managers. This helps in achieving optimal performance on NUMA systems for all types of workloads. When AutoNUMA is enabled, the Linux kernel is capable of reaching NUMA convergence dynamically and incrementally based on the current load, in turn handling NUMA placements automatically.

Dump and crash dump analysis tools
HPE Integrity Superdome X hardware and software are highly robust; however, sometimes failures may occur. In the event of a failure, it is important to recover from the failure completely and resume normal operations as soon as possible. Having a good crash dump may help in getting data to diagnose the failure. Here are the recommended configuration steps to enable crash dump:

General guidelines
1. Ensure there is enough space in the file system containing /var/crash to save the dump. Ideally, this file system should be equal to the memory size of the system if a full dump is specified. In practice, for the default dump type that excludes unneeded pages and with compression, 500 GB to 1 TB is generally enough to store a compressed dump from a system with 12 TB of memory. For idle systems, 100 GB of free space is generally enough, but it is not possible to guarantee that a given size will be sufficient.

2. The time for the kdump utility to copy a dump for a 12 TB system with default dump settings (dump level-1) and lzo compression is an average of three hours. Time may vary from one to five hours on SLES 11 SP3 and all RHEL 6.x versions. On RHEL 7.x, SLES 11 SP4 and SLES 12, there have been substantial performance improvements in dump and a 12 TB dump will complete in an average of one hour.

SLES 11 SP3/SP4 guidelines
1. Correctly set the crashkernel size. Superdome X requires a larger crashkernel size than the default to produce a successful dump. Always set crashkernel to 512 MB on all configurations of the BL920s Gen8 and Gen9 Server Blades. On a system installed with defaults, use yast2-kdump to set the crashkernel to 512 MB and then reboot for this new value to take effect. It is best not to edit the kernel command line directly in bootloader config file. If you do, the kernel parameter line should be crashkernel=1024M-:512M.

2. After the reboot, to check if dump is enabled with the new crashkernel size issue the command `chkconfigboot.kdump` and `serviceboot.kdump` status.

   For SLES 11 SP3 only, the following dump RPMs need to be installed after the initial install to fix server dump-related issues. Assuming the customer has registered for support or updates from SUSE with the purchase of the appropriate SLES 11 SP3 software for the BL920s Gen8, go to download.suse.com/index.jsp and access Patch Finder to download the crash 918 patch release from 20 May 2013. The following RPM need to be downloaded and installed (These are the earliest revisions that fix the problems; later versions are also acceptable.):

   `Crash-6.0.7-0.16.1.x86_64.rpm Kdump-0.8.4-0.39.2.x86_64.rpm`

   `Makedumpfile-1.5.1-0.15.1.x86_64.rpm`

   `SLES 11 SP4` comes with the correct RPMs so this update is not required.

3. Modify the kdump config files to boot the crashkernel with multiple CPUs.
   To make sure dumps will work on all BL920s Gen8 and Gen9 configurations, follow these steps: Edit the /etc/sysconfig/kdump file to set

   `KDUMP_CPUS="4" KDUMPTOOL_FLAGS="NOSPLIT"`

   Then rebuild kdump initrd by executing service boot.kdump restart.

4. Leave compression and dump level at their defaults, which are lzo and 31 (smallest dump size).

SLES 11 SP3/SP4 large I/O configurations

For large I/O configurations, Hewlett Packard Enterprise recommends setting the crashkernel=832M.

For systems with the 650M and/or 650FLB adapter, blacklist the be2iscsi and fclp modules, since neither iSCSI nor FCoE are supported currently with these cards. Note that on SLES 11 SP4, effectively blacklisting these modules from the capture environment may require removing them from the INITRD_MODULES list in /etc/sysconfig/kernel.

Furthermore, on SLES 11 SP4, significant amounts of memory may be consumed configuring multipath LUNs that are not needed in the crash capture environment. The unneeded LUNs can be blacklisted in the capture environment by the following steps, which temporarily modify the /etc/multipath.conf file on the system while constructing a new kdump initrd.

1. Determine the wwids of the set of multipath LUNs that are needed in the crash capture environment. Minimally, this will be device containing the root file system and the directory in which crash dumps are collected; as an example, use 3600c0ff0001fa1d43f25685604000000 and 3600c0ff0001fa1d43f25685605000000.

2. If the system already has a multipath.conf, save a copy:

```
# test -f /etc/multipath.conf && cp -p /etc/multipath.conf /etc/multipath.conf.restore
```

3. Based on the example wwids above, append these lines to /etc/multipath.conf:

```
blacklist {
    wwid ".*"
}
blacklist_exceptions {
    wwid "3600c0ff0001fa1d43f25685604000000"
    wwid "3600c0ff0001fa1d43f25685605000000"

}
```

4. As described elsewhere in this document, edit these settings in /etc/sysconfig/kdump:

```
KDUMP_CPUS="4" KDUMPTOOL_FLAGS="NOSPLIT"
MAKEDUMPFILE_OPTIONS=" --cyclic-buffer 96000 -d"
```

5. Generate a new dump initrd, which will include the modified multipath.conf:

```
# mkdumprd -f
```

6. Restore the original multipath.conf

```
# test -f /etc/multipath.conf.restore && mv -f /etc/multipath.conf.restore /etc/multipath.conf
```

7. This procedure must be repeated in the future when the kdump initrd is regenerated (for example, after installing SUSE maintenance updates).

SLES 12/SLES 12 SP1-SP3 guidelines

1. Starting with SLES 12 the size of the crashkernel may be split between low and high memory. On large I/O configurations 4096M may be needed, we recommend 384M, low and 2048M, high to insure being able to capture a dump in the unlikely event of a crash, with the crashkernel=4096M, high on large I/O configurations.

   To set the crashkernel size and configure kdump on SLES 12:

   a. Edit in **/etc/default/grub** the kernel command line in

   ```
   GRUB_CMDLINE_LINUX_DEFAULT=
   ```

   ```
   and substitute
   ```

   ```
   crashkernel=384M,low crashkernel=2048M,high
   ```

   ```
   for any other crashkernel= entries that may be in this line
   ```

   b. Execute `grub2-mkconfig -o /boot/grub2/grub.cfg`

   c. Edit the **/etc/sysconfig/kdump** file to set

   ```
   KDUMP_CPUS="4"
   ```

   ```
   KDUMPTOOL_FLAGS="NOSPLIT"
   ```

   d. Reboot and then the new crashkernel size that is now 384 MB in low memory and 2 GB at high memory will be in effect. If a larger size is needed, increase the crashkernel=<size>high, but leave the crashkernel=384M, low value always the same

2. After the reboot, to check if dump is enabled with the new crashkernel size, execute the following step:

   ```
   systemctl status kdump
   ```

3. Modify the kdumpconfig files to boot the crashkernel with multiple CPUs.

   To ensure dumps will work on all Superdome X configurations, follow these steps:

   a. Edit the **/etc/sysconfig/kdump** file to set

   ```
   KDUMP_CPUS="4"
   ```

   ```
   KDUMPTOOL_FLAGS="NOSPLIT"
   ```

   b. Leave compression and dump level at their defaults, which are Izo and 31 (smallest dump size)

RHEL 6.5–6.9 and RHEL 7.x guidelines

1. For RHEL 6.5 (BL920s Gen8) only, install RHEL 6.5 Maintenance Release Kernel kernel-2.6.32-431.20.3.el6.x86_64.rpm from the access.redhat.com/downloads/ website. This is needed to fix dump defects that affect the BL920s Gen8 system. See Red Hat kernel errata RHSA-2014-0771 for details.

2. Edit the **/boot/efi/efi/redhat/grub.conf** file to change the crashkernel=auto in the kernel command line to `crash kernel=512M`. RHEL 6.5 (BL920s Gen8) installations default to `crashkernel=auto`, which has a formula that sizes the `crashkernel` based on the memory size of the system. However, the values it chooses generally will not work with the BL920s Gen8 due to the systems larger I/O configurations.

Note

For very large I/O configurations, a 512 MB crashkernel size will not be adequate. For RHEL 6.x, Hewlett Packard Enterprise recommends setting crashkernel=832M and for RHEL 7.x crashkernel=2048M or higher if needed.

3. To get a large enough file system space to save a dump, it is good practice to create a separate file system for /var/crash to avoid running out of memory in the root file system. To save dumps in a file system other than root, one must modify **/etc/kdump.conf** to change this parameter:

```
# <fs type> <partition> - Will mount -t <fs type> <partition> /mnt and copy

# /pro/vmcore to /mnt/<path>/127.0.0.1-%DATE/.

# NOTE: <partition> can be a device node, label or uuid.
```

For example, if /var/crash is itself a mounted volume, one must first identify it:

```
[root@dhd1 sysconfig] # mount|grep /var/crash

/dev/mapper/mpathap4 on /var/crash type ext4 [rw] [root@dhd1 sysconfig] # blkid|grep mpathap4

/dev/mapper/mpathap4: UUID="6e0f0dbd-c68b-4623-8d07-5f7e571d4f22" TYPE="ext4"
```

Now that you know the universal unique identifier (UUID), edit these lines into **/etc/kdump.conf:**

```
path /

ext4 UUID=6e0f0dbd-c68b-4623-8d07-5f7e571d4f22
```

Note
Failure to allocate sufficient file system space may result in saving dumps underneath a normally mounted volume. This situation may be difficult to reclaim space under the mount point.

---

4. Modify the **kdump config** files to boot the crashkernel with multiple CPUs. Edit **/etc/sysconfig/kdump** file to change

```
nr_cpus=1 to nr_cpus=4 in the line labeled KDUMP_COMMANDLINE_APPEND
```

5. **Only** for RHEL 6.5 (BL920s Gen8): Edit the **/etc/sysconfig/kdump** file to add `disable_cpu_apicid=X`, where X is the apicid of CPU 0, which may be obtained by executing: `cat /proc/cpuinfo | grep apicid | head -1`

Unless there has been a failure in first processor in the first socket that deconfigures it, CPU 0 on all BL920s Gen8 Server Blade that have 15-core processors will have an apicid of "0" and CPU 0 on all BL920s Gen8 Server Blades with 10-core processors will have an apicid of "4".

The edited line when done should look like the following example for 15-core processor sockets:

```
KDUMP_COMMANDLINE_APPEND="irqpoll nr_cpus=4 disable_cpu_apicid=0 reset_devices
cgroup_disable=memory mce=off"
```

```
For 10-core processor sockets: KDUMP_COMMANDLINE_APPEND="irqpoll nr_cpus=4 disable_cpu_apicid=4
reset_devices cgroup_disable=memory mce=off"
```

Note
RHEL 6.6 and higher versions kexec-tools initialization scripts automatically determine the correct CPU 0 apicid and adds it to the `kdump` command line automatically. Step 5 is only performed when running RHEL 6.5.

---

6. For better dump compression or copy performance, it is recommended to edit the /etc/kdump.conf file to change the core collector line from its default of

```
core_collector makedumpfile -c --message-level 1 -d 31
```
to
```
core_collector makedumpfile -l --message-level 1 -d 31
```

This switches the dump from using zlib compression to lzo compression, which is much faster. On RHEL 7/7.1, kdump.conf already sets the default to L compression.

7. After all kdump configuration changes have been made, rebuild the kdump initrd to make them take effect by utilizing the following commands: Execute "touch /etc/kdump.conf" to ensure that the kdump will rebuild the kdump initrd when kdump is restarted.

```
For RHEL 6.x: service kdump restart
```

```
For RHEL 7.x: systemctl restart kdump
```

Software advisory notes

Finding customer advisories

For the most up-to-date information about issues on the Superdome X, visit the following link: hpe.com/support/hpesc. Search for Superdome X, select your product and click on the **Top issues & solutions** tab then **Advisories, bulletins,** and **notices** on the right. This will provide the list of the most recent advisories.

General known issues

Boot messages

The following messages will be seen while booting a BL920s Gen8 and Gen9 Server Blades:

```
lpc_ich 0000:20:1f.0: I/O space for ACPI uninitialized lpc_ich 0000.20:1f.0: I/O space for GPIO
uninitialized lpc_ich 0000:20:1f.0: No MFD cells added
```

and

```
iTCO_wdt: failed to et TCOBASE address, device disabled by hardware/BIOS
```

These messages indicate hardware on auxiliary blades is not initialized. Only the hardware on the monarch blade is initialized, and if a blade is auxiliary, then the hardware is disabled. These messages are expected and may be ignored. There will be one set of these messages for every auxiliary blade. There are seven sets of them on an eight-blade system.

Boot message (ptp_clock_register failed)

The following message may be seen while a BL920s Gen8 and Gen9 Server Blade is booting or a network driver is loaded after boot:

```
ixgbe 0000:85:00.0: ptp_clock_registered failed
```

Only eight Precision Time Protocol (PTP)—capable network devices may be registered. After the limit of eight is reached, this error message is written to the message log as each additional network device is configured. Each NIC port counts as another network device.

Network devices registered for PTP allow the system to participate in IEEE 1588 PTP, which enables systems on the network to synchronize their clocks, thus allowing more precise time operations across systems.

The Linux ethtool utility may be used to query the status of PTP on a particular device. If the system does not participate in PTP satisfactorily, you may reconfigure the network to work around the eight-device limitation.

Large I/O configuration boot messages

During boot on the Superdome X, several messages such as the following may be seen:

```
pci 0000:71:00.1: BAR 2: can't assign io (size 0x20)
```

These messages are due to the large I/O card configurations of this system, which exceeds the I/O port space the firmware has reserved. This I/O port space is not needed by most of the cards, and the few that do need it get their space allocated first in the initialization process. The OS still attempts to allocate I/O port space for all the cards, regardless, so these messages are repeated as the cards on the higher numbered nodes are initialized. Ignore these messages.

During boot, the following message during Advanced Configuration and Power Interface (ACPI) thermal management initialization will be seen and may be ignored as it causes no issues:

```
ACPI Exception: AE NOT FOUND, No or invalid critical threshold (20090903/thermal-386)
```

Suspend/resume

The hibernate/suspend/resume functionality is not supported on Superdome X. If by accident a hibernation operation is initiated on a Superdome X, power cycle the system from the OA and reboot the system in order to recover.

Clocksource failover

Clocksource tsc unstable/Enable clocksource failover

This message may be seen occasionally on the host when running virtual machine guests. No action is necessary. Disregard the recommendation in the message to enable clocksource failover. Hewlett Packard Enterprise specifically recommends against enabling clocksource failover on the Superdome X.

Some ipmitool commands not supported

The system firmware only includes Intelligent Platform Management Interface (IPMI) support for HPE tools and applications. For example, sensor data record (SDR) status is not supported, and you will see the following error message when executing the `ipmitool` command:

```
# ipmitool sdr
```

```
Error obtaining SDR info: Invalid command Unable to open SDR for reading
```

Configuring KVM guest with SR-IOV on HPE FlexFabric 20Gb 2-port 650FLB/650M

KVM guests require that the allocated memory is equal to the maximum memory limit when Virtual Functions (VFs) are configured on the FlexFabric 650FLB/650M Adapter. You can manually change the allocated memory using the virsh CLI or using the virt-manager. Hewlett Packard Enterprise highly recommends RHEL 7 or SLES 12 host OS for use with KVM guests supported with SR-IOV.

SLES 11 SP3 known issues

Long boot time

At certain points while booting SLES 11 SP3, several minutes will elapse without output to the console. This is normal for configurations with more I/O and memory.

Expected diagnostic message during boot

This diagnostic is expected when booting SLES 11 SP3. You may ignore this message.

```
FATAL: Error inserting mgag200 (/lib/modules/3.0.101-0.15-
default/kernel/drivers/gpu/drm/mgag200/mgag200.ko): Invalid argument
```

Out-of-sync cursor when using iLO remote console

By default, SLES 11 SP3 creates a xorg.conf file that incorrectly sets up the iLO virtual mouse cursor as a relative pointing device. So, we have two permanently out-of-sync cursors on the screen. A fix is not available in SLES 11, but is present in SLES 12. For more information, ask your SUSE support representative to review Bug 715607. To work around this issue, the iLO virtual mouse cursor may be easily reconfigured as an absolute pointing device with the following edits to the xorg.conf file:

1. Locate the mouse InputDevice section in /etc/X11/xorg.conf

2. Change the value of the Driver line from mouse to evdev

3. Change the value of the Device line from /dev/input/mice to /dev/input/by-id/usb-HP_Virtual_Keyboard-event-mouse

4. Optionally, comment out the Protocol line, since the evdev driver does not require a protocol. A revised InputDevice section:

```
Section "InputDevice" Driver    "evdev" Identifier  "Mouse [1]" Option  "Buttons" "9"

Option    "Device" "/dev/input/by-id/usb-HP_Virtual_Keyboard-event-mouse" Option    "Name"
"Avocent HP 336047-B21"

Option    "Protocol" "explorerps/2" Option "Vendor" "Sysp" Option      "ZAxisMapping" "4 5"
EndSection
```

5. After saving these changes, restart X:

```
service xdm restart
```

Install multipath-tools update to be able to mount virtual removable media

SUSE has released an application update for multipath tools, available from download.suse.com/patch/finder/ website (use Patch Finder) for customers who have support contracts with SUSE for updates. Without this fix, you will not be able to mount anything over the virtual USB media from the HPE iLO Integrated Remote Console (IRC) while the SLES OS is booted on Superdome X.

The following official RPMs correct this problem and another issue involving segmentation faults running the multipath command:

```
kpartx-0.4.9-0.97.1.x86_64.rpm multipath-tools-0.4.9-0.97.1.x86_64.rpm
```

These patches or later will resolve the issues. Due to dependencies, these RPMs must be installed together.

EDD not supported

During installation, the following message may be seen, because Superdome X does not support EDD:

```
[    26.767479] BIOS EDD facility v0.16 2004-Jun-25, 0 devices found [   26.774131] EDD
information not available.

Insmod: error inserting "/modules/edd.ko": -1 No such device
```

Disregard this message; it causes no issues and is due to the Linux kernel always trying to load the EDD module whether the hardware supports it nor not.

SLES 12 SP2 known issues

Kdump service fails on first boot after install.

kdump.service fails causing "systemctl status" to appear degraded. The following error message is seen in the output of the command "systemctl status kdump.service" kdump service succeeds though.

SLES 12 SP3 known issues
System crash dump fails invoking "makedumpfile":

Case I: System crash dumps have been observed to fail occasionally with an error message similar to the following:

```
Can't allocate memory for parallel_info.

Commandline parameter is invalid.

Try 'makedumpfile --help' for more information.

makedumpfile Failed.

Running makedumpfile --num-threads 5223023572093018873  -d 31 -l /proc/vmcore failed (1).
```

There is no known resolution for this issue at this time.

Case II: Crash dump is seen to fail sometimes with the default KDUMP_FREE_DISK_SIZE of 64M on Btrfs setup, with the following error message:

```
makedumpfile Failed.

Dump too large. Aborting. Check KDUMP_FREE_DISK_SIZE.

Running makedumpfile  -d 31 -l /proc/vmcore failed (1).
```

Workaround solution:
Dump gets collected if KDUMP_FREE_DISK_SIZE=0 on a Btrfs setup.

Word of caution with this workaround: If the dump is part of the root file system and free space is a concern then it is suggested to configure dump onto a different file system to avoid filling up the root file system.

RHEL 6.5–6.9 and 7.0–7.5 known issues
Configuring large KVM guests on RHEL 6.5 (BL920s Gen8 only)

For large KVM guests (greater than 30 virtual CPUs), the following additional setup instructions are useful to avoid soft lockups, loss of time stamp counter (TSC) sync, and other issues on the hosts and guests:

1. Install tuned package and use the profile virtual-host on the host tuned-adm profile virtual-host

2. Disable **Pause Loop Exit (PLE)** on the host by executing the following commands (repeat each time the host is rebooted and prior to starting any virtual machine [VM] guests)

```
modprobe -r kvm_intel modprobe kvm_intel ple_gap=0
```

3. Before starting any VM guest, pin all the vCPUs of the VM to the host CPUs of choice by using the numactl utility. When running large KVM guests, messages like the following will be seen and may be ignored on the host:

```
hrtimer: interrupt took 4389495 ns
```

Configuring crashkernel on RHEL 6.8 on large memory and large IO systems

Crashkernel initialization can fail when RHEL 6.8 boots on systems where the memory per blade equals or exceeds 1 TB and the crashkernel=auto or crashkernel=832M initialization parameters are used in the boot command line. The following message is seen in the console logs that indicates a crashkernel memory reservation failure:

```
"crashkernel reservation failed. Found area cannot be reserved: start=0x4000000, size=0x34000000."
```

For such large memory systems that have large IO configurations, a solution is to use a crashkernel size less than 832M in the boot command line, e.g., "crashkernel=816M".

GNOME System Monitor

If Superdome X has more than 240 CPUs (such as an eight-blade system with Hyper-Threading enabled), the GNOME System Monitor utility will report incorrect numbers of CPUs generally in the range of 200–255 and sometimes accompanied by glibtop errors. This is due to a scaling issue in the glibtop software and will be fixed in a future version of GNOME System Monitor and glibtop.

RHEL 7.3 anaconda installer hangs occasionally

The installer dialog for customizing elements like date/time, software selection, installation source, installation destination and security policy fails to enable these options. This has been observed on systems with LVM configurations and is often accompanied by the lvmetad service failing with a segfault error. This can be worked around by temporarily removing/unpresenting the LVM disks.

RHEL 7.4 QLogic driver error message with stack trace during boot

The error message "disabling already-disabled device" appears in the console and dmesg during system boot along with a stack trace similar to the following:

```
pcie_aer_get_firmware_first+0x2a/0xd0

pci_release_region+0x8a/0xe0

pci_disable_device+0x84/0x90

qla2x00_disable_board_on_pci_error+0x135/0x160 [qla2xxx]
```

Disregard this message as it does not affect the normal functioning of the system.

RHEL 7.5 installation may result in kernel panic on HPE BL920s Gen9 (Superdome X) systems

A known issue related to the fix for the Meltdown vulnerability causes a kernel panic with a NULL pointer dereference during the installation of RHEL 7.5 on HPE BL920s Gen9 (Superdome X) systems with Intel Xeon Processor E7-8800 v4 Series. When the problem appears, the following error message is displayed on the console:

```
WARNING: CPU: 576 PID: 3924 at

kernel/workqueue.c:1518__queue_delayed_work+0x184/0x1a0
```

Then the system reboots, or enters an otherwise faulty state.

Workaround and resolution for this problem:

During fresh install of RHEL 7.5 or upgrade from RHEL 7.4 to RHEL 7.5, add the "nopti" option to the kernel command line using the boot loader. Once the system finishes the installation and reboots, add the "nopti" option to the kernel command line again. Once the system boots up, upgrade the kernel to 3.10.0-862.1.1 or its later version. Once system boots to this kernel, "nopti" can be safely removed from the command line.

Summary

Hewlett Packard Enterprise is committed to providing a long road map of solutions for mission-critical computing based on the Linux operating system. The solutions are grounded on solid hardware platforms, augmented by HPE system software. We are working with the open source community to make continuous improvements in the Linux operating system and application stacks.

Further innovations will be released to the market on a regular basis. This white paper will be updated as further enhancements are announced.
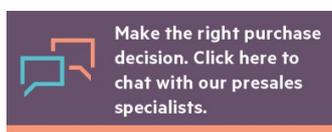
Resources

- HPE Integrity Superdome X QuickSpecs

- Red Hat Enterprise Linux from HPE

- RHEL OS certification matrix

- RHEL Technical Notes

- SLES OS certification matrix

- SUSE Linux Enterprise Server from HPE

Learn more at
hpe.com/servers/superdomex

Make the right purchase decision. Click here to chat with our presales specialists.

Sign up for updates