

Sun Fire™ 880 Workgroup Server

An Architectural Overview

Prepared by:

Alan Goldman
IDSG Business Unit
alan.c.goldman@east.sun.com
781 442-3560

Enabling Technology

UltraSPARC™ III Microprocessors

The Sun Fire™ 880 server is based upon Sun's second generation of 64-bit microprocessor and the SPARC™ V9 UltraSPARC™ architecture. This architecture will enable the performance of future microprocessors with cycle times in excess of 1 GHz to scale proportionately. The 750 MHz microprocessor is built by a 0.18 micron CMOS process utilizing seven metal layers and providing over sixteen million transistors.

Some of the more prominent features of the UltraSPARC III microprocessors which provide enhanced performance and scalability include:

- High clock rate with minimal latencies
- A deep pipeline
Generally the deeper the pipeline, the higher the penalty incurred from an incorrect branch prediction. Instructions being processed must be flushed, a new set of instructions must be accessed and started through the sequence of processing. The UltraSPARC III has a 90+% branch prediction rate using a 16K entry prediction RAM and branch correlation algorithm. In addition, there is a small amount of alternate path buffering. If a predicted branch is not taken, the buffering makes a few instructions immediately available, thereby minimizing the penalty.
- On-chip memory controller
 - Capable of handling numerous simultaneous accesses with out-of-order completion
 - The main memory bus is 512 bits wide and has a peak throughput of 3.2 Gbytes/sec.
- On-chip L2 cache controller with on-chip tag RAM
To reduce latency to the 8 Mbyte L2 (external) cache, both the L2 cache controller and tag RAM reside on the processor. Since the L2 tag RAM is operating at processor speeds and not the slower L2 cache speed, cache misses are detected earlier and memory fetch operations may be initiated sooner.
- 32 Kbyte, 4-way associative instruction cache
- 64 Kbyte, 4-way associative data cache
- Instruction prefetch into a 2 KB instruction prefetch buffer
- 4 instructions fetched per cycle
- 2 KB fully associative write cache
- This on-chip write cache eliminates up to 90% of the store activity to the L2 (external) cache. As a secondary benefit, cache coherency operations are accelerated for both the individual processor and the multiprocessor environment.

Since the on-chip L2 cache tags and write cache are both on chip, all operations are managed at chip speed, no external operations are required. External processors need make a single inquiry for cache coherency.



- Arithmetic and floating point optimizations
 - Up to two floating point loads issued per cycle
 - Three floating point units (one add/subtract, one multiply, one divide)
 - Low latency floating point divider
 - Two graphics units (one ALU, one multiply)
- Address Translation Buffer

These table entries enhance the efficiency of virtual to physical memory address translation. In the UltraSPARC III processor, the size of the address translation buffer has been geared for large databases such as Oracle™ and Sybase™, thereby offering an optimized database engine. By comparison, Intel™ based processors have a much smaller translation buffer.
- Visual Instruction Set (VIS)

The VIS is a set of extensions to the core instructions which accelerates multimedia, image processing, networking applications and Java™ performance. These instructions can also accelerate matrix operations typically found in intensive engineering applications.

Sun™ Fireplane Interconnect (System Bus)

The Sun™ Fireplane Interconnect (system bus) is a crossbar switch with six ports, one for each of the four dual processor/memory modules and two for the I/O subsystem, i.e. one for each PCI bridge chip. All ports may operate simultaneously. The total aggregate bandwidth is 9.6 GB/sec. of sustained throughput.

The system bus is located on the motherboard.

FC–AL Storage Controller

The Sun Fire 880 Server provides an internal storage subsystem with an integrated Fibre Channel Arbitrated Loop (FC–AL) controller supporting a maximum of twelve 1.0", 10,000 RPM, 18.2 GB, 36.4 or 72.8 GB disks. (The storage backplane for the first six drives is standard. An optional, second backplane provides support for an additional six drives.)

Fibre Channel is an industry standard, high–speed, serial data, transfer interface. In addition to strong performance characteristics, FC–AL provides powerful networking capabilities that allow switches and hubs to enhance availability and bandwidth between systems and storage controllers.

FC–AL is also a high–reliability interconnect. The interface is robust enough to allow multiple devices to be removed from the loop at once without interruption to on–going services.

FC–AL has the following characteristics:

- Industry standard: FC–AL development effort is part of the ANSI/ISO accredited SCSI–3 standard, helping to avoid the creation of non–conforming, incompatible implementations.
- Broadly supported: All major system and storage vendors are implementing FC–AL, thereby insuring a wide variety of choices and inter–operability.
- Facilitates failover: Upon failure of a controller or FC loop, devices are capable of quickly disassociating from the failure and re–associating with an alternate controller available via a secondary FC loop. This feature provides the capability of masking the outage from applications and users. Clearly, an alternate hardware route/path and optional software are generally required in order to provide complete transparency to users and applications.

On the Sun Fire™ 880 Server, a FC–AL controller is integrated on the motherboard. An alternate path to either the internal disks or external storage arrays may be provided by the inclusion of an optional PCI to



FC–AL controller and software which masks the failure of an individual controller to the application. When both paths are operational, improved performance and throughput may also be possible.

Remote System Control (RSC) and System Service Processor (SSP)

The Sun Fire™ 880 Server features a System Service Processor (SSP) and Remote System Control (RSC) software, enabling complete console access, monitoring and control from remote locations via any client device on the network, a serial line or modem.

The System Service Processor is a fully independent processor card that resides on the system motherboard. The SSP allows administrators to remotely query the status of the system, diagnose faults and initiate a system power on/off or reboot. Because it operates independently from the server, the SSP can constantly monitor a variety of conditions perform the following:

- View the server's front panel including key switch position and LEDs
- Run diagnostic tests and configure the server remotely
- Monitor and report errors including output from power–on, self–test (POST) and OpenBoot Diagnostics
- Reboot, reset, power–on and power–off on demand
- Notify of server problems and enter detailed log entries of RSC events

The RSC card plugs into a dedicated slot on the system I/O board and includes integrated PCMCIA modem, serial and Ethernet interfaces; it does not occupy a PCI slot. The RSC firmware runs independently from the host and uses standby power drawn from the host system. The RSC module is powered by an independent battery which operates for up to 40 minutes after a complete failure.

The RSC hardware and software continue to run even when the server is off–line. It can send notification of hardware failures or other events to administrators via pager or e–mail.

The card provides three ports that are accessible through an opening in the rear panel of the system:

- 10 Mbps Ethernet port via an RJ45 twisted pair Ethernet (TPE) connector
- 56 Kbps modem port via an RJ11 connector
- EIA–232D serial port via an RJ45 connector



System Architecture

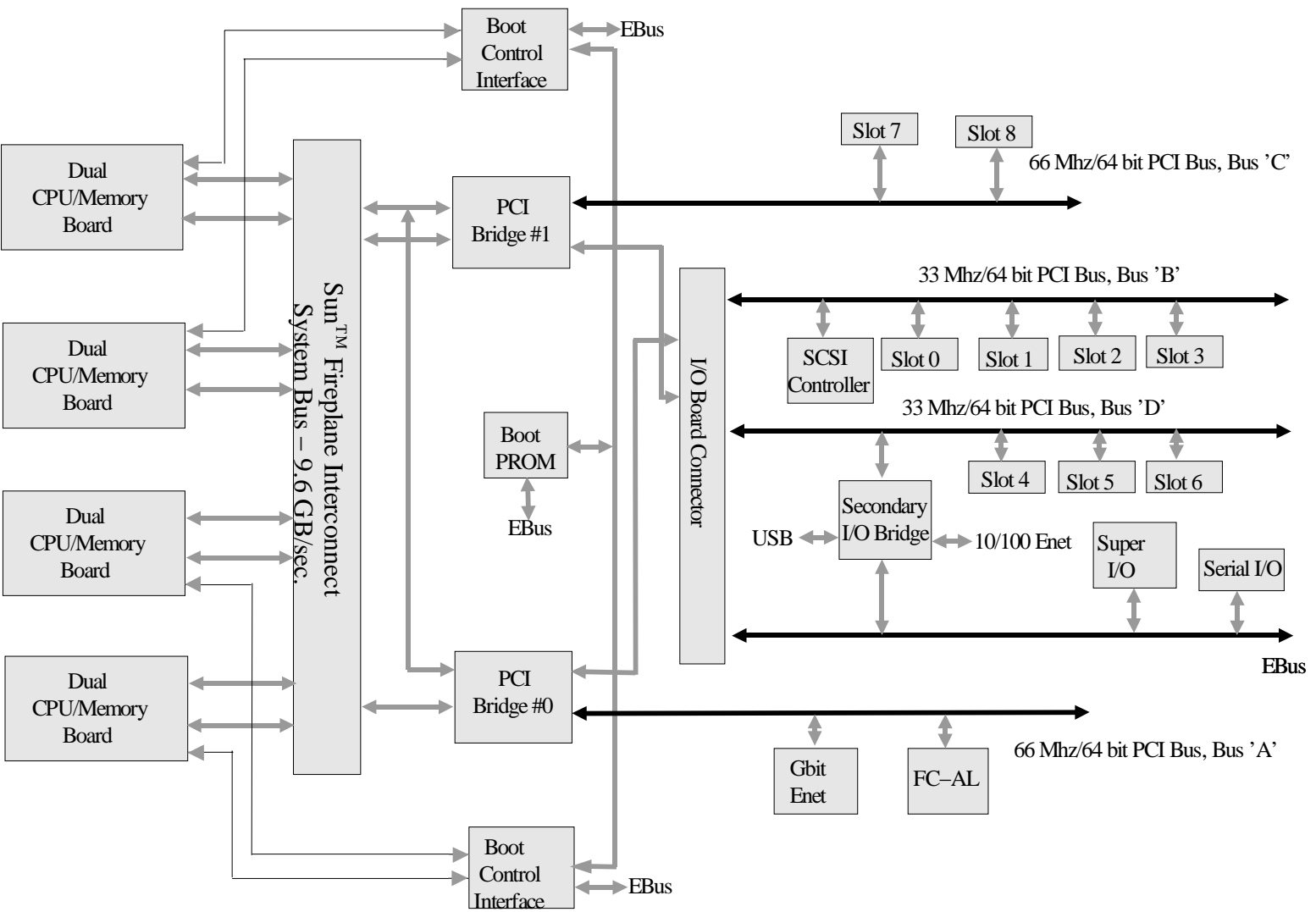


Figure 1, Sun Fire™ 880 Server, Schematic

Dual Processor/Memory Modules

The Sun Fire™ 880 server system is comprised of one or more dual-processor/memory modules which mount perpendicular to the motherboard, the alternate side of which is reserved for the I/O subsystem. The minimum configuration consists of a single dual processor/memory module; the maximum configuration has four. Single processor modules are not available, consequently, a system will always contain an even number of UltraSPARC III processors.

Each processor has an 8 MB external (L2) cache implemented with eight SRAM devices.

The processor slots are hot pluggable, thereby providing a higher level of system availability. Dynamic reconfiguration of processor modules requires firmware and Solaris™ support that will be available after revenue release of the server.

Memory Subsystem

Memory within the Sun Fire™ 880 Server is distributed across the dual CPU/memory modules with each processor controlling a portion of the total memory. In order to minimize latency, the memory controller is integrated on the processor chip. Access to the memory associated with the alternate processor on the same dual processor/memory module incurs a minimal latency penalty. Latency to non-local memory is substantially less than that incurred by more traditional approaches which utilize a single, external memory controller for all accesses.

The main memory data bus provides 512 bits of data (64 bytes) and 36 bit of ECC in a single access. This size corresponds in size exactly to a full external cache block.

There are eight DIMM slots per processor; 16 slots per dual processor/memory module. These are arranged as two memory groups of four slots per processor. Each processor may have a different amount of memory. Configuration requirements state:

- DIMMs must be added in groups of four (no partially populated groups)
- All four DIMMs in any group must be of identical capacity
- The minimum memory configuration per dual processor/memory module is eight DIMMs, four in each memory subsystem (four per processor)

Memory interleaving is a function of the number of DIMMs and their capacity relative to the other memory groups on the dual processor/memory module. Interleaving is implemented on a 64 byte boundary to coincide with the width of the memory data bus. The level of interleaving is implemented as:

- 8 DIMMs 4-way if all DIMMs are identical;
otherwise 2-way
- 12 DIMMs 4-way between any two groups configured identically
2-way on each group, otherwise
- 16 DIMMs 8-way if all 16 DIMMs are identical;
otherwise, 4-way between any two groups configured identically;
otherwise, 2-way within any group that doesn't match any other group



If successive accesses alternate between two distinct logical groups, the sustainable bandwidth is 1.6 GB/sec. When the access pattern involves four logical groups, the sustainable bandwidth is 2.4 GB/sec.

In the event of a system failure resulting from an individual DIMM, Automatic System Recovery will attempt to identify and bypass only the failed component. In this event, memory interleaving for the associated dual processor/memory module is disabled.

The currently available memory options are:

- 0.5 GB Memory Expansion (4 x 128 MB DIMMs)
- 1.0 GB Memory Expansion (4 x 256 MB DIMMs)
- 2.0 GB Memory Expansion (4 x 512 MB DIMMs)
- 4.0 GB Memory Expansion (4 x 1 GB DIMMs), a future option subject to availability and qualification

The memory subsystem is capable of supporting up to 4 GB per memory group which would allow a total capacity of 64 GB with an 8-way system.

I/O Subsystem Architecture

The I/O subsystem interfaces to the system bus via two PCI bridge chips, each of which controls one 66 Mhz, 64 bit PCI bus and one 33 Mhz, 64 bit PCI bus. One 66 Mhz, 64 bit PCI bus is reserved for the integrated FC-AL and Gbit Ethernet controllers. PCI buses A and B are connected to PCI bridge #0; PCI buses C and D with PCI bridge #1. (Please refer to Figure 1, Sun™ Fire 880 Server, Schematic.)

There are a total of nine PCI slots available for user selected host bus adapters, all of which are hot pluggable. Two slots operate at 66 Mhz, 64 bits, 3.3 V; seven at 33 Mhz, 64 bits, 5 V. All slots accept universal (3.3V/5V) adapters, either 64 or 32 bits wide, and either full length (12") or short (7") cards.

However, if a 33 Mhz adapter is inserted into either of the 66 Mhz slots, it will cause that bus segment to operate at 33 Mhz. During a hot plug operation, a 33 Mhz adapter cannot be connected to a 66 Mhz slot.

The DVD and removable media device(s) are connected via an integrated SCSI controller from one of the 33 Mhz PCI buses. The serial I/O ports interface via an integrated controller to the other 33 Mhz PCI bus.

Sustainable throughput via the PCI bridge chip to the 66 Mhz bus utilizing 64 bit wide adapters, dual masters and streaming DMA writes of at least 512 bytes in length is approx. 390 Mbytes/sec. Throughput utilizing a 33 Mhz bus, dual masters and streaming DMA writes of at least 512 bytes in length is approx. 252 Mbytes/sec. These figures assume no other traffic on the bus.

Internal Disk Subsystem

The internal disk subsystem is managed by an integrated fibre channel, arbitrated loop (FC-AL) disk controller and will support one to twelve, hot pluggable, FC-AL disks arranged as two groups of six drives. The drives connect to a backplane which provides power, signal and data. The first storage backplane supporting the first six drives is standard; the second backplane is an option.

The storage backplanes are supported in a serial configuration where all twelve drives are on the same FC loop(s). Although physically possible, the backplanes are not intended to be used as independent entities, i.e. separate loops for each backplane. This decision was made in accordance with the intention to qualify and support the most commonly requested configurations. To use the backplanes



independently will likely reduce the availability to data and/or reduce the overall storage capacity.

The backplanes come with an embedded provision for an alternate, independent FC–AL loop in order to achieve a higher level of availability and potentially higher throughput to the internal disk storage. This alternate loop may be activated by the inclusion of the (X)6727A, PCI to dual FC–AL controller with an internal connector, the (X)6755, internal FC–AL cable, and software product(s). A software product, such as Veritas™ Volume Manager with Dynamic Mutipathing or Multipathing I/O, a tentative enhancement to Solaris 8 SU 4, will mask the failure of a FC loop or controller from the application, thereby allowing uninterrupted processing.

The (X)6727A, PCI to dual FC–AL controller with an internal connector, provides both an internal and an external connector for one loop and only an external connector for the alternate. Supported configurations preclude using both the internal and external connectors of the same loop simultaneously. In other words, the internal storage array may not be extended to include external storage. The second loop may be used independently from the usage of the first.

Currently supported disks include:

- 18.2 GB, 1.0", 10,000 RPM, FC–AL disk
- 36.4 GB, 1.0", 10,000 RPM, FC–AL disk
- 72.8 GB, 1.0", 10,000 RPM, FC–AL disk, a future option subject to availability and qualification

The 18.2 GB, 1.0", 10,000 RPM, FC–AL disk is being retired almost coincidentally as the Sun Fire 880 system is being introduced.

Control Panel

The control panel includes nine LED status indicators, a power button and a security keyswitch. At the top of the status and control panel, three general status LEDs provide a snapshot of the system status. Below the power button and security keyswitch, a graphical display provide six LED icons to indicate specific fault conditions and locations.

The power button is recessed to preclude inadvertent shutdown of the system. The operation of the power button may disabled by the security keyswitch.

The four position security keyswitch controls the power on modes of the system and prevents unauthorized users from powering off the system, reprogramming the system firmware and performing CPU/memory or PCI hot–plug operations.

Power Distribution System

The required power is provided by two 1100/1200 watt (output) power supplies inserted into dedicated slots located at the rear of the unit, along the bottom. An optional third power supply provides N+1 redundancy which will preclude a loss of server availability in the event of a failure of an individual power supply. The input power requirement is 1500 watts per supply.

Each power supply requires its own country specific, 15 Amp AC power cord. The sources of power may be on independent grids, thereby removing another potential point of failure for the system.

Each power supply provides a total of five DC output voltages (3.3V, 5.0V, 12V, 42 V, and 5.0V standby). Output current is shared equally between each of the supplies via active current sharing circuitry.

Three LEDs on the rear of each power supply provide status information for each of the power supply bays.



Environmental Monitoring and Control (EM&C) System

The environmental monitoring and control (EM&C) system protects the system against:

- Extreme temperatures
- Lack of adequate air flow
- Power supply problems

Monitoring and control capabilities resides at the operating system level as well as within the system's PROM firmware. Consequently, protection is assured even if the system is halted or unable to boot.

The EM&C system uses an industry standard I²C bus to monitor and control temperature sensors, fans, power supplies, status LEDs and the front panel keyswitch. Temperature sensors monitor ambient temperature throughout the system as well as that of each CPU.

The hardware and software components of the EM&C system insure that the temperature remains within a predetermined range for safe operation. Failure to remain within that range will result in either a 'warning' or 'critical' condition with the appropriate error message to the system console, if present, an entry in the system log file and illumination of the system and thermal fault LED indicators on the front status panel.

A critical condition will be followed by a graceful shutdown of the system.

The EM&C system will also detect a failure of the cooling fans. The base system configuration has three primary fan trays with a total of five individual fans. Systems with the redundant cooling option have twice as many trays and individual fans.

In the event of a failure of any fan, the monitoring system generates an error message, activates the system fault and thermal fault LEDs on the status and control panel, lights the appropriate fan fault LED inside the cabinet, and activates the appropriate secondary fan tray, if present.

The power subsystem is monitored in a similar manner. In the event of a problem with a power supply, an error message is displayed on the console, if present, a log file entry is made and the system fault and power fault LED indicators are activated on the status and control panel. LEDs on the back of each power supply will indicate the status and nature of the failure.

System Rackmounting Kit

The Sun Fire™ 880 Server is intended to operate either in a remote/branch office or within a data center. In a branch/remote office, it is anticipated that the system will reside on the floor. Data centers typically arrange equipment within racks for the most efficient use of floor space.

An optional rack mount kit is available, (X)9628A. The casters from the system need not be removed when mounting it within a rack. The server is 17 rack units (RU) high, including the kit itself. Within a rack, the Sun Fire™ 880 Servers may be mounted two high. Each unit weights from 194 lbs. minimal configuration to 299 lbs. fully configured, approximately.

For servicing, the unit slide forward from the rack. Side panels open for access to the dual processor/memory modules on one side or the I/O subsystem, including PCI adapters, on the other.

The Sun DeLorean Rack, SG-ARY030A or SG-XARY030A, may be used.

Other manufacturers of racks include, but are not limited to:

- Rittal (www.rittal.com),
- Pentair (www.pentair-ep.com),
- APW/Wrightline (www.wrightline.com)

All offer four post racks that are 78" (2,000 mm.) high and 39" (1,000 mm.) deep. None of these products have been tested; neither the companies nor the products are endorsed or recommended.



Reliability, Availability, and Serviceability (RAS)

Reliability, availability and serviceability are three aspects of a system's design and quality that contribute to continuous operation and consequently, minimize system downtime.

- Reliability
- Reliability refers to the capability of a system to operate continuously without failures and to maintain data integrity. Reliability influences MTBF.
- Availability
System availability measures the percentage of time that a system is accessible by users and is providing service.
- Serviceability
Serviceability measures the time to restore a system to operation once a failure has occurred. Serviceability influences MTTR.

Various metrics may be applied in calculating RAS, including:

- MTBF – mean time between failures. This measurement is influenced by quality, design, environmental considerations such as power and cooling and even operational errors, i.e. how well the hardware and software verifies the intention of staff.
- MTTR – mean time to repair. This measurement is influenced by overall design, pre-/anticipatory and post-failure diagnosis, ease of access and efficiency of FRU replacement. Dispatch and response time of service personnel are typically not included in this calculation as this factor is dependent upon service contracts (SLAs), geographical conditions, etc. which can radically influence the measurement. Nevertheless, these times will ultimately influence availability.

Simplistically stated, Availability = 100% – (MTBF + MTTR), where availability, MTBF and MTTR are measured in percentages relative to operational time.

The Sun Fire™ 880 Server's reliability, availability and serviceability features include:

- Error correction and parity check for improved data integrity for memory and on internal data paths
- Internal error detection and diagnosis capability, including via an internal 'back door bus'
- Hot pluggable disks, power supplies, fans, PCI cards and dual-processor/memory modules
- Easily accessible LED status indicator, especially for PCI cards
- Front panel LED display
- Remote System Console (RSC) for monitoring and administrative capability
- Automatic System Recovery (ASR) for isolation of failed components and automatic reboot capability
- Environmental monitoring and fault protection
- RAID capability for disks
- Watchdog timer

This hardware device is continually reset as long as the operating system is running. In the event of a system hang, the timer will expire and force a reset/reboot. The timer must be enabled and the system rebooted in order to activate the device.



Installation Data

Cabinet

Hardware Dimensions

	U.S.	Metric
Height <ul style="list-style-type: none">with castors – desksidewithout castors NOTE: Rackmounting does <u>not</u> require the removal of the casters	28.1" 27.6"	714 mm 701 mm
Width	18.9"	480 mm
Depth	32.9"	836 mm
Weight (minimum/maximum)	194/299 lbs.	88/136 kg.
Shipping Weight		

Environment

Power Requirements

	U.S.	International
Operating <ul style="list-style-type: none">2 power supplies required, 3 provides N+1, hot-pluggable redundancy3 supplies maximum	1500 W input; 1100W @ 120 VAC per supply	1500 W input; 1100W @ 240 VAC per supply
Tolerance	47 – 63 Hz	47 – 63 Hz

Temperature

	Fahrenheit	Celsius
Operating	41° – 95° F	5° – 35° C
Non-operating	-4° – 140° F	-20° – 60° C

Humidity (noncondensing)

Operating	20% – 80%
Non-operating	5% – 95%



Noise (in accordance with ISO 9296)

Operating acoustic noise	7.2 dB(A)
Idling acoustic noise	

Regulations

Meets or exceeds the following requirements

Safety	UL 1950, CSA C22.2 950 (from UL), DEMKO EN 60950, TUV 60950, U ENL CB scheme IEC 950 (CE mark)
RFI/EMI	Class B 47CFR15.B U.S. Code of Fed. Reg. Title 47, Part 15 Subpart B ICES-003; Industry Canada EM55022; 1998, CISPR 22: 1997 (European) AS/NZ 3548 (Australian/New Zealand), CNS 13438 (Taiwan)
X-ray	DHHS 21 Subchapter J; PTB German X-ray Decree