# HP XC3000 Cluster

## Reference guide



# HP XC3000 Cluster overview

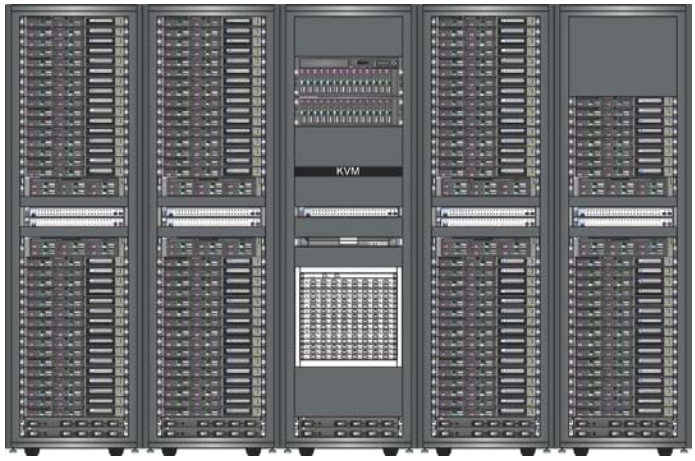
## At a glance

The Linux®-based HP XC3000 Cluster is a parallel supercomputer that scales from 17 to 256 nodes, and the architecture is designed to scale to higher node counts. It is based on 2-way HP ProLiant DL380 G3 servers, acting as dedicated service nodes and performing management and administrative functions within the cluster, and 2-way HP ProLiant DL360 G3 servers acting as application nodes. The HP ProLiant DL380 G3 and DL360 G3 servers are based on Intel® Xeon™ processors. These clusters are available in standard 17-, 34-, 68-, 128-, and 256-node configurations and utilize the high-bandwidth, low-latency Myricom Myrinet XP interconnect and the HP ProCurve 2650 Ethernet switches. These 10/100 Ethernet switches are used for an administrative/boot network. In addition to the standard configurations, HP XC3000 clusters can also scale in increments of four application nodes—starting at a minimum of 16 application nodes—to meet customer compute needs.

Each Linux-based HP XC3000 Cluster is composed of these elements:

- HP ProLiant DL380 G3 servers—dual-processor systems that serve as service nodes. One service node is required for every 16 compute/application nodes, and a service node is designated as the administrative/boot node for the cluster. Service nodes also take on other administrative roles and act as I/O nodes for file access.
- HP ProLiant DL360 G3 servers—dual-processor systems that serve as dedicated compute/application nodes
- HP ProCurve Switch 2650—a 48-port 10/100 Ethernet interconnect
- High-speed, low-latency Myricom Myrinet XP communications interconnect
- Local embedded service node storage plus SAN storage, using HP StorageWorks Modular SAN Array 1000 (MSA1000) or HP StorageWorks Enterprise Virtual Array 3000 (EVA 3000)
- HP Rack 10000 Series 42U rack enclosures
- One rackmount keyboard/monitor is required on XC3000 Cluster systems. On systems with two or more service nodes, one 4-port KVM (console/keyboard/video monitor) switch is required per configuration.

**Figure 1.** An HP XC3000 Cluster based on the dual Intel Xeon processor architecture with 128 nodes (256 processors), Myricom Myrinet XP high-speed interconnect, and MSA1000 SAN storage array is shown below. The configuration consists of 4 compute building block (CBB) racks containing 120 compute nodes and 8 service nodes. A utility building block rack contains the Myrinet switch, MSA1000 SAN storage array, KVM switch, and TFT display.




# HP XC3000 system architecture

The XC3000 Cluster has a parallel system architecture that supports the management and execution of multiple parallel and serial applications on the XC systems. The XC system architecture presents the user and the administrator with key single-system traits. Users see the XC system as a single system for login access, resource access, and job execution. Administrators control the system from a single service node, using it to perform tasks such as system management, performance monitoring, and hardware diagnostics, thus avoiding the complexity and difficulty of standard cluster administration. The XC system operates with a single root file system (the exception is the administration node, which for reliability reasons is managed independently). Because the system files are not scattered across multiple locations, there is only one set of configuration data to manage. Problems normally associated with managing version skew within a cluster cannot occur.

The XC3000 Cluster is a full-service system. Unlike many other clustered systems, the system services are organized to enhance the performance of tactical applications that require tightly coupled, synchronous cooperation between processes running on multiple nodes. This enhancement is accomplished by offloading the system services to specialized nodes, called service nodes, which provide all of the operating, administrative, and file services that are normally expected from a full-service system.

In the XC3000 architecture, applications are executed on one or more specialized application/compute nodes. Operations that are not of direct benefit to the application are migrated to other system components external to the application node. Within the application node, single processors are dedicated to the execution of an application's individual process. Virtually all services provided on the application node are dedicated to the application, which reduces or eliminates the context switching and resource contention normally found on general-purpose systems.

HP XC3000 Cluster systems are designed to support the XC cluster architecture as well as the XC cluster software. In addition, XC cluster systems are easy to configure, build, install, and support. To facilitate configuration and scaling, they are made up of modular building blocks: compute building blocks (CBBs), interconnect building blocks (IBBs), utility building blocks (UBBs), and storage building blocks (SBBs).

## Compute building blocks

The XC3000 Cluster system compute building block (CBB) rack contains up to 32 application nodes, two service nodes, and up to two HP ProCurve subnet 10/100 switches. Multiple CBB racks are connected together to expand the size of the cluster to up to 255 nodes. The CBBs consist of the following:

- An HP Rack 10000 Series 42U rack with server slide kits
- Three power distribution units (PDUs)
- Up to 32 dual-processor HP ProLiant DL360 G3 application nodes (minimum of four application nodes, except in the first CBB rack)
- Up to two dual-processor HP ProLiant DL380 G3 service nodes (one service node for every 16 application nodes)
- Up to two HP ProCurve Switch 2650 network switches with rackmounting kit (one for every 16 application/service nodes)
- Network cabling to connect the nodes to the HP ProCurve switches

The application nodes and service nodes are wired internally to the CBB rack to reduce external cabinet cabling. Each of the application and service nodes is also connected to the high-speed Myrinet XP interconnect switch in the interconnect building block or utility building block rack with a Myrinet PCI adapter card. Each application node iLO (Integrated Lights-Out maintenance processor) and administration LAN port is connected to the 10/100 48-port HP ProCurve Switch 2650 in the cabinet to form a subnet. Each service node in the rack will have on administration LAN connection to its HP ProCurve sub-network. There are up to two HP ProCurve Switch 2650s in an XC3000 CBB rack. The switches are linked together with the Gigabit link to form a sub-network of up to 32 application nodes and two service nodes. Each application node requires a single disk for swapping, and the disks in all the application nodes must be the same.

The service nodes in the compute rack are also connected to an administration switch that is in the utility rack. The iLO and LAN ports of the service nodes are connected to an HP ProCurve 2650 administration switch. Each application node can have up to two disks, which must be the same size and speed, and each application node must have one disk for swapping. Service nodes have up to six disks, which must be the same size and speed, and each service node must have one disk for swapping. Each service node will also have a Gigabit LAN PCI interface for external user communication.

One service node in the XC cluster system is designated the master administrative node and is responsible for booting the system and other administrative functions. Select service nodes are identified to provide I/O and file system access. These nodes will have additional disks or an optional Fibre Channel host bus adapter to connect to an external storage subsystem (MSA1000 or EVA 3000). Application nodes are added four at a time in a CBB rack, with up to 32 application nodes per CBB rack and one service node for every 16 application nodes.

## Interconnect building blocks

The XC3000 Cluster system interconnect building block (IBB) is used only in cluster configurations that have more than 128 nodes requiring a Myrinet switch configuration with more than 128 ports. Myrinet switch configurations with more than 128 ports require multiple node and link-level chassis and switch line cards and spine cards. For XC3000 configurations of 128 nodes or less, the utility building block with a single Myrinet 128-port chassis and Myrinet line cards is used. The IBB consists of an HP 10000 Series 42U rack, two power distribution units, and rackmount kits with up to four Myrinet XP switches (128-port chassis, switch cards) used to create an interconnect switch for clusters of up to 256 nodes. The application and service nodes in the CBB racks are connected to the Myrinet switch cards in the IBB rack. The IBB racks are connected to form a large switch fabric that allows up

to 256 nodes to be connected together. Up to two link chassis are mounted in an HP 10000 Series rack. Up to four node-level chassis are mounted in an HP 10000 series rack. Myrinet 8-port line cards (up to 16 cards) are added to provide the necessary switch ports to connect the nodes. A Myrinet SNMP monitor card is mandatory for each 128-port chassis. Myrinet 8-port spine cards are also added to provide the necessary switch fabric in a Clos switch configuration.

## Utility building blocks

The XC3000 Cluster system utility rack, or utility building block (UBB), is used in configurations with 128 nodes or less to house the 17-slot Myrinet chassis, the system TFT console, the administrative HP ProCurve network switch, and the optional MSA1000 SAN storage array. One UBB is required for each XC3000 system. (For 256-node cluster systems, node 256, a service node, is mounted in the UBB rack.) The UBB consists of these components:

- An HP 10000 Series 42U rack
- A Myrinet 128-port chassis (in smaller XC3000 Cluster systems—those with 128 nodes or less—only one Myrinet 128-port switch chassis is required, and this chassis is mounted in a utility rack)
- Two power distribution units
- An HP ProCurve 2650 system administration LAN switch (for configurations with two or more service nodes)
- An optional MSA1000 storage subsystem
- A rackmount keyboard/TFT monitor or KVM switch

## Storage building blocks

HP XC3000 Cluster system storage building blocks (SBB) are optional EVA 3000 storage subsystems that come in their own HP Rack 10000 Series 42U racks. The EVA 3000 storage subsystem is connected to the I/O service nodes in the cluster. Rules for configuring the MSA1000 and EVA 3000 storage subsystems are dependent on the file system requirements.

## System options

Placement of the CBB, IBB, UBB, and optional SBB is dependent on the layout of the customer site, and inter-cabinet cabling must be taken into account. The factory will build to a default layout if site-specific information is not available. It is recommended that site layout information be provided at the time of system purchase. System options for the XC3000 Cluster consist of application- and service-node memory content, application- and service-node disk content, TFT console with a KVM switch, SAN storage, and EVA 3000 storage. These options are supported by the XC3000 Cluster system software. Options not specified in this reference guide will need to be reviewed on a case-by-case basis for integration into the system. Equipment that is not approved will be shipped as non-integrated material.

## Node count

HP XC3000 Cluster systems have a minimum of 16 application nodes and one service node. Both are configured with a minimum of 1 GB of memory per CPU. Application nodes can be added in a CBB rack, four at a time, up to a total of 32, with one service node for every 16 application nodes. An XC3000 Cluster system can scale to up to 256 total nodes (240 application nodes and 16 service nodes). Larger XC3000 Cluster systems (beyond 256 nodes) can be designed and configured; contact HP for more information.

## Integration, services, and ordering information

HP XC3000 Cluster systems are fully integrated into racks at the factory and tested to assure proper system operation before being shipped to the customer site. XC3000 Cluster software is loaded and tested. Inter-cabinet cabling is labeled to facilitate installation at the customer site. XC3000 Cluster systems also come with hardware and software documentation as well as the necessary HP XC System Software licenses and software media kit. Special considerations in the XC hardware design have been made so that there is proper cooling of the major components in the cabinets. Special cable and cable management assemblies are included for proper inter- and intra-cabinet cabling as well as to promote adequate air flow and maintenance of the equipment. Special mounting brackets for the application nodes, network switches, interconnect chassis, and cooling baffles are also included to allow the most efficient use of the rack space. These considerations increase system reliability, allow the system to be factory integrated and shipped to the customer site without damage, and facilitate installation of the system at the customer site.

HP XC3000 Cluster systems are installed by HP Services. They come with both installation services and HP Consulting and Integration (C&I) system startup services as part of the solution. These services must be included in the initial system order. A three-year onsite service, parts, and labor warranty comes with each cluster. The service provided by the warranty provides coverage Monday–Friday from 9:00 a.m. to 5:00 p.m. Enhanced services are available to meet customers' needs. HP Services Consulting and Integration support is required to help customers with system startup. Optional training and application/solution development is available.

The XC3000 Clusters can be ordered through the High-Performance Technical Computing (HPTC) Competency Centers within each region. These centers have comprehensive menus and tools for developing customer-specific configurations based on the XC cluster architecture. In early 1Q04, the XC cluster menu and configuration rules will be active on Watson, the standard HP ordering and quoting tool.

This reference guide provides descriptions of some typical configurations based on total node count: 17, 34, 68, 128, and 256 nodes. As noted earlier, other configurations between 17 and 256 nodes are supported as standard offerings and are configured by determining the number of necessary compute building blocks and adding application nodes in increments of four.

# HP XC3000 Clusters: options

## 17-node cluster

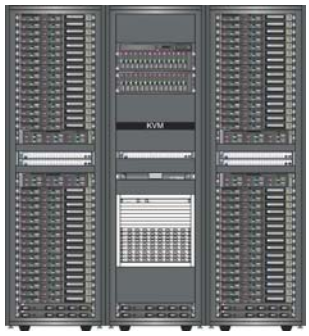
**Figure 2.** The 17-node Linux cluster based on the Intel Xeon processor consists of one TFT rackmount keyboard/monitor, one HP ProLiant DL380 G3 service node server, 16 HP ProLiant DL360 G3 application node servers, an HP ProCurve Switch 2650 Ethernet switch, and a 24-port Myrinet XP high-speed network interconnect (17-slot chassis) in two HP Rack 10642 (42U) cabinets. The configuration shown here consists of one partially filled compute building block rack with 17 nodes and a utility building block.

## 34-node cluster

**Figure 3.** The 34-node Linux cluster based on the Intel Xeon processor consists of one TFT rackmount keyboard/monitor, one KVM switch, two HP ProLiant DL380 G3 service node servers, 32 HP ProLiant DL360 G3 application node servers, two HP ProCurve Switch 2650 Ethernet switches, and a 40-port Myrinet XP high-speed network interconnect (17-slot chassis) in two HP Rack 10642 (42U) cabinets. The configuration shown here consists of one fully populated compute building block with 34 nodes and a utility building block.

## 68-node cluster

**Figure 4.** The 68-node Linux cluster based on the Intel Xeon processor consists of one TFT rackmount keyboard/monitor, one KVM switch, four HP ProLiant DL380 G3 service node servers, 64 HP ProLiant DL360 G3 application node servers, five HP ProCurve Switch 2650 Ethernet switches, and a 72-port Myrinet XP high-speed network interconnect (17-slot chassis) in three HP Rack 10642 (42U) cabinets. The configuration shown here consists of two fully populated compute building blocks with 34 nodes each and a utility building block.

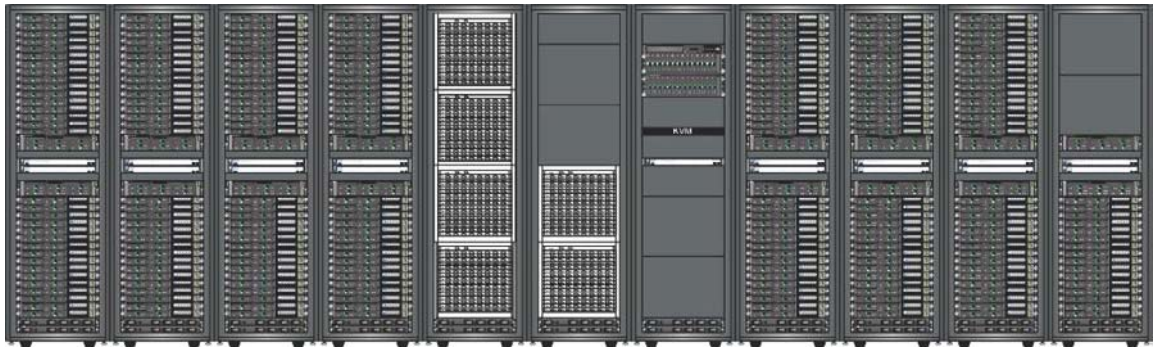


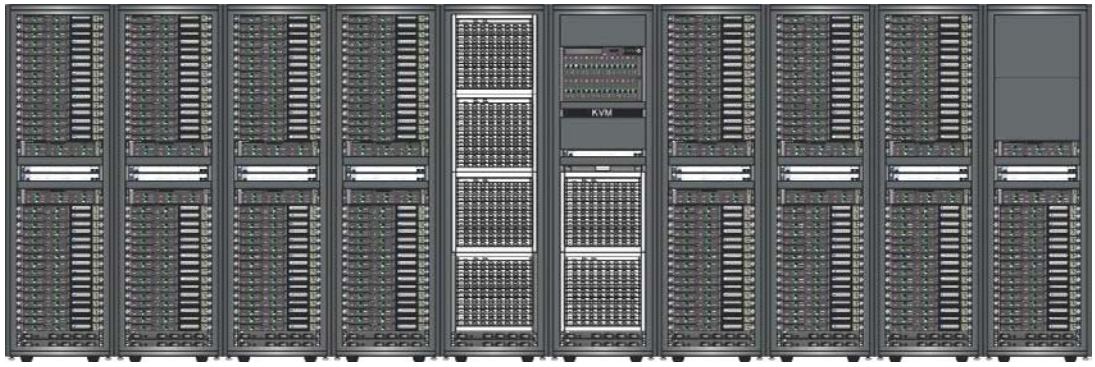
## 128-node cluster

**Figure 5.** The 128-node Linux cluster based on the Intel Xeon processor consists of one TFT rackmount keyboard/monitor, one KVM switch, eight HP ProLiant DL380 G3 service node servers, 120 HP ProLiant DL360 G3 application node servers, nine HP ProCurve Switch 2650 Ethernet switches, and a 128-port Myrinet XP high-speed network interconnect (17-slot chassis) in five HP Rack 10642 (42U) cabinets. The configuration shown here consists of one partially filled compute building block rack with 26 nodes, three fully populated compute building blocks with 34 nodes each, and a utility building block.

# 256-node cluster

**Figure 6.** The 256-node Linux cluster based on the Intel Xeon processor consists of one TFT rackmount keyboard/monitor, one KVM switch, 16 HP ProLiant DL380 G3 service node servers, 240 HP ProLiant DL360 G3 application node servers, 17 HP ProCurve Switch 2650 Ethernet switches, and a 256-port Myrinet XP high-speed network interconnect in 10 or 11 HP Rack 10642 (42U) cabinets. Note that the cabinet alignments shown are optimized for cabling and can be realigned to fit the computer room layout. The first configuration below consists of one partially filled compute building block rack with 18 nodes, 7 fully populated compute building blocks with 34 nodes each, a utility building block, and two interconnect building blocks. In the second configuration, an alternative cabinet layout, the UBB cabinet is combined with the second IBB rack.





# HP ProLiant DL380 G3 server—service node

**Figure 7.** The HP ProLiant DL380 G3 server

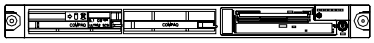
The HP ProLiant DL380 G3 server based on the Intel Xeon processor functions as a dedicated service node, directing the management and administrative functions within the XC3000 cluster. One HP ProLiant DL380 G3 server is needed for every 16 application nodes in the 17-, 34-, 68-, 128-, and 256-node configurations. Each service node includes these components:

- HP ProLiant DL380 G3 2U system chassis with two Xeon processors (a two-processor system is required). Two processor choices are available: 3.06 GHz or 3.2 GHz. Both feature a 533 MHz Front Side Bus.
- Up to 6 GB DDR 200 MHz maximum memory, advanced ECC protection (minimum 2 GB interleaved memory per node—requires two banks to be populated with same size memory)
- Three available PCI-X slots—two hot-swap 64-bit/100 MHz and one 64-bit/100 MHz
  - Slot 0 reserved for mandatory Myrinet XP PCI network interface card (rev. D)
  - Gigabit Ethernet LAN interface
  - SAN host bus adapter—QLA 2214 (optional for SAN storage, maximum of one per node)
- Six hot-swap drive bays (**note: one hard drive is required; drives must all be the same size and speed)**
- Integrated Lights-Out (iLO) for boot and server management
- Two embedded 10/100/1000 Ethernet network interface cards
- CD optical drive load device
- N+1 redundant power supply (required)
- Keyboard and video port

| |
|---|
| **Dual-processor HP ProLiant DL380 G3 server 3.2 or 3.06 GHz/533 MHz with 1 MB cache memory** |
| **HP ProLiant DL380 G3 node options** |
| **Memory (minimum 2 GB memory per application node, supporting memory interleaving—2 banks of equal-size memory)** |
| Add-on main memory—1024 MB PC2100 registered DDR SDRAM DIMM memory kit (2 x 512 MB) |
| Add-on main memory—2048 MB PC2100 registered DDR SDRAM DIMM memory kit (2 x 1024 MB) |
| DVD/CD drive |
| **Up to 6 hard disk drives (note: one hard drive mandatory for swapping; if multiple drives are selected, they must all be the same size and speed)** |
| Hard drive—36.4 GB Wide Ultra3 SCSI 10,000 rpm (1") hard disk drive |
| Hard drive—36.4 GB Wide Ultra3 SCSI 15,000 rpm (1") hard disk drive |
| Hard drive—72.8 GB Wide Ultra3 SCSI 10,000 rpm (1") hard disk drive |
| N+1 power supply (required) |
| **Host bus adapter—FC 2214 HBA (SAN storage interface)** |
| Gigabit Ethernet PCI adapter |
| Myrinet XP PCI adapter—rev. D |

## HP ProLiant DL360 G3 server—application node

**Figure 8.** HP ProLiant SL360 G3 server



The HP ProLiant DL360 G3 server based on the Intel Xeon processor functions as a compute node within the XC3000 cluster. (It is also used as a head/management node for 4- and 8-node configurations.) Each compute node includes these components:

- HP ProLiant DL360 G3 1U system chassis with two Xeon processors (a two-processor system is required). Two processor choices are available: 3.06 GHz or 3.2 GHz, both featuring 533 MHz Front Side Bus.
- 4 GB DDR 266 MHz maximum memory advanced ECC protection (using 1 GB DIMMs)
  - Two memory banks (two DIMM slots per bank)
  - Minimum 1 GB per CPU with two banks populated with the same memory type to allow interleaving
- Two available PCI-X slots—two hot-plug 64-bit/200 MHz
  - Slot 0 reserved for required Myrinet PCI network interface card (rev. D)
- Two disk drive bays **(note: one hard drive is required for swapping; additional drives must be the same size and speed; drive configurations in the other compute or application nodes in the cluster must be the same)**
- Integrated Lights-Out (iLO)
- Two embedded 10/100/1000 Ethernet network interface cards
- No CD drive
- No N+1 power

| |
|---|
| **Dual-processor HP ProLiant DL360 G3 server (either 3.06 or 3.2 GHz/533 MHz) with 1 MB cache memory** |
| **Node options** |
| **Memory (minimum 2 GB memory per application node, supporting memory interleaving—2 banks populated)** |
| Add-on main memory—1024 MB PC2100 registered DDR SDRAM DIMM memory kit (2 x 512 MB) |
| Add-on main memory—2048 MB PC2100 registered DDR SDRAM DIMM memory kit (2 x 1024 MB) |
| **Up to 2 hot-swap drives (note: one hard drive required for swapping; if 2 drives are selected, they must be the same size and speed)** |
| Hard drive—36.4 GB Wide Ultra3 SCSI 10,000 rpm (1") hard disk drive (optional) |
| Hard drive—36.4 GB Wide Ultra3 SCSI 15,000 rpm (1") hard disk drive (optional) |
| Hard drive—72.8 GB Wide Ultra3 SCSI 10,000 rpm (1") hard disk drive (optional) |
| Myrinet XP PCI adapter—rev. D |

## Myricom Myrinet XP high-speed interconnect

The Myricom Myrinet XP high-speed interconnect consists of these components:

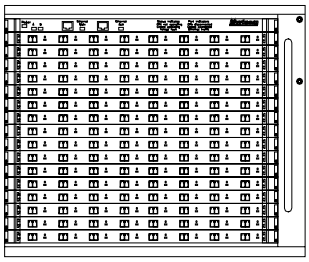
- Minimum of three (and up to 16) 8-port line cards with Myrinet 2000 fiber ports (8): each 2.0+2.0 Gb/s at a front-panel LC connector to a 50/125 fiber pair up to 200 meters in length
- Flow control, error control, and “heartbeat“ continuity monitoring on every link
- Low-latency, cut-through, crossbar switches, with SNMP monitoring card for high-availability applications

Select the appropriate number of high-speed interconnect 8-port line cards based on the number of nodes in the XC3000 Linux cluster (one per service node and application node). 8-port spine switch cards are not used in a Myrinet switch with 128 ports or less. The Myrinet XP switch chassis also includes an SNMP monitoring card for each switch frame.

For XC3000 Linux clusters with between 129 and 256 nodes, use a federated Myrinet XP switch configuration utilizing 8-port line cards, 8-port spine cards, and up to six Myrinet 17-slot chassis. Multiple Myrinet XP chassis are used and are linked by interconnected switch cards. Six chassis are required to support 256 ports, two link-level chassis, and four node-level chassis. The two link-level chassis are mounted in a separate IBB cabinet from the four node-level chassis. Rack assemblies are provided to support mounting and cooling of chassis. One node-level 8-port line card is needed for every eight nodes. One 8-port spine card is needed for every line-level switch card for the node-level chassis. 8-port spine cards are installed in the link-level chassis to connect to the spine card ports in the node-level chassis.

Fiber-optic cables are used to connect the switch ports to the Myrinet PCI adapter card and the switch cards in switch configurations requiring multiple chassis. Fiber-optic cable lengths are determined by placement of the switch and nodes in the cluster.

**Figure 9.** Myrinet XP 128-port, high-speed network interconnect with 16 8-port line cards and SNMP card



| **Myrinet XP high-speed interconnect chassis, 128 ports—8-port fiber switch card with crossbar (line) and without crossbar (spine)** (select one card for every eight processors, including service nodes; service nodes are connected on switch cards separately from application nodes) |
|---|
| SNMP monitoring card required |
| 17-slot switch frame (3-, 5-, and 9-slot switch frames not used in XC3000 Clusters) |
| 8-port fiber line switch card (with crossbar) |
| 8-port fiber spine switch card (without crossbar) |
| Fiber cable (lengths as required) |
| Myrinet XP PCI interface—fiber, rev. D, 2 MB (note: each node must contain one PCI interface card) |
| Myrinet rackmount and baffle kit |

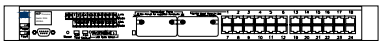
## Management network and console interconnect

The HP ProCurve Switch 2650 functions as the network administration/sub-network switch for the XC3000 Cluster. These 48-port Ethernet switches are reliable network components that are highly available. The HP ProCurve Switch 2650 includes these features:

- 48 10/100 ports
- 2 Gigabit link ports
- 10/100 auto-sensing per port automatically detects and sets the speed for any 10Base-T or 100Base-TX device.
- MDI/MDI-X cascade port makes it easy to add other hubs to the network.
- A comprehensive LED display with per-port indicators provides an at-a-glance view of status, activity, and speed.
- Automatic polarity correction and auto-partitioning on all ports helps find and fix common network problems.
- Built-in bridge automatically connects 10 MB/s and 100 MB/s devices without requiring additional products, modules, or stacking to operate correctly.
- Special 10000 rack-mounting kit to allow integration into rack for shipment (not included and must be ordered separately). Rack-mounting kit allows the HP ProCurve switch to be shipped mounted and cabled into the 10642 rack from the factory.
- An HP ProCurve switch is connected to a second HP ProCurve switch (sub-network switch) to form a sub-network consisting of 32 application nodes and two service nodes. Large clusters will have multiple sub-networks for each set of 34 nodes (located in CBB rack).

- One HP ProCurve switch (administration switch) is used in configurations with two or more service nodes to connect the MP and administration port of the service nodes in the cluster (located in the UBB rack).

**Figure 10.** HP ProCurve switch 2650



## Monitors and keyboards

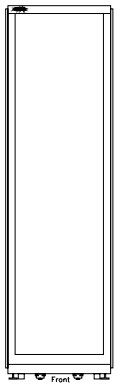
**Figure 11.**

A rackmount keyboard/monitor (RKM) is required with the XC3000. The RKM with TFT display is attached to the HP ProLiant DL380 G3 service node that is designated as the cluster system administration node. The TFT display is mounted in the utility rack (UBB) and provides local user access to the cluster. In systems with two or more service nodes, a KVM switch is also required to connect the RKM to multiple service nodes to allow for failover connectivity to other service nodes. The TFT display is connected directly to the single service node in the system when a KVM switch is not required.

## Cabinet and power distribution unit

| **HP Rack 10642 (42U)—rack cabinet with front and rear doors mounted on a shock pallet** |
|---|
| Blanking panels (graphite) |
| Side panel 42U HP Rack 10000 Series |
| Power distribution unit (PDU)—3 per CBB and 2 per IBB and UBB |
| Cable management kits |
| Power cord, terminated, 3-conductor, SPT-2, IEC320-C13 |

Figure 12. **Power distribution unit (PDU) with 12 IEC-320-C13 receptacles—PDU, 24 amp-High, North America, Japan.** Three PDUs are required for each compute/service node cabinet; two PDUs are required in the utility building block rack and interconnect building block rack.



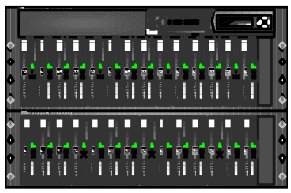

## SAN storage

### HP StorageWorks Modular SAN Array 1000

The HP StorageWorks Modular SAN Array 1000 (MSA1000) is offered as optional external SAN storage for the XC3000 Cluster. Select one host bus adapter for each designated service I/O node. The required system I/O bandwidth will determine the number of required I/O service nodes connected to the SAN storage subsystem. The MSA1000 storage subsystem is mounted in the UBB cabinet with the Myrinet chassis in XC clusters with 128 nodes or less. In XC clusters with more than 128 nodes, the MSA1000 storage subsystem is mounted in the utility cabinet. Up to 4 TB of storage is supported per MSA1000 subsystem. The MSA1000 supports up to two shelves.

If larger capacity or higher bandwidth is required, two MSA1000 storage subsystems can be configured. A ratio of one MSA for one I/O node is recommended for balanced performance in RAID 5 configuration.

**Figure 13.** HP StorageWorks Modular SAN Array 1000



The HP StorageWorks Modular SAN Array 1000 includes these components:

- One MSA1000 controller with 256 MB cache
- One MSA1000 Fibre Channel I/O module with 2 Gb SFP short wave transceiver
- Redundant hot-pluggable power supply/blower assemblies
- Universal rackmounting kit

- MSA1000 support CD and documentation
- Serial cable
- Two power cables
- Two 3-foot very high density cable interconnect (VHDCI) to VHDCI SCSI cables

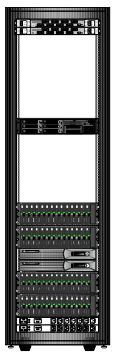
Options:

- MSA1000—one controller, 14 disk slots
- MSA1000 SAN switch 2/8 (integrated; includes four 2 Gb SFPs)
- 2 Gb SFP short-wave transceiver kit
- HP StorageWorks Enclosure Model 4454R (one maximum per MSA1000)
- 36.4 GB pluggable Ultra320 universal hard drive, 10,000 rpm (1")
- 72.8 GB pluggable Ultra320 universal hard drive, 10,000 rpm (1")
- 146.8 GB pluggable Ultra320 universal hard drive, 10,000 rpm (1")
- 36.4 GB pluggable Ultra320 universal hard drive, 15,000 rpm (1")
- 72.8 GB pluggable Ultra320 universal hard drive, 15,000 rpm (1")
- PCI-X single channel 2 Gb Fibre Channel host bus adapter—FCA2214
- 2 m LC-LC multimode Fibre Channel cable
- 5 m LC-LC multimode Fibre Channel cable
- 15 m LC-LC multimode Fibre Channel cable
- 30 m LC-LC multimode Fibre Channel cable
- 50 m LC-LC multimode Fibre Channel cable

**HP StorageWorks Enterprise Virtual Array 3000**

The HP StorageWorks Enterprise Virtual Array 3000 (EVA 3000) is offered as optional external SAN storage for the XC3000 Cluster. Select one host bus adapter for each designated service I/O node. The required system I/O bandwidth will determine the number of required I/O service nodes connected to the SAN storage subsystem. The EVA 3000 storage subsystem comes mounted in its own HP Rack 10000 Series cabinetry. Up to 8 TB of storage is supported per EVA 3000.

**Figure 14.** HP StorageWorks Enterprise Virtual Array 3000



The EVA 3000 storage array has these components:

- One 3U controller assembly with two HSV100 controllers that have redundant power supplies
- Two M5114 3U dual-redundant FC loop 14-bay disk enclosures
- 42U graphite storage cabinet with appropriate mounting rails and power

- Virtual controller software for HSV100 dual controllers
- Eight disk drives

Requirements:

- HP factory integration (required with each EVA storage subsystem)
- HP field installation—separate from XC3000 Cluster cabinet installation

Options:

- 42U EVA cab 60 Hz (based on Series 10000 rack system)
- 42U EVA cab 50 Hz (based on Series 10000 rack system)
- 2C2D EVA 3000-C/8 x 36 GB/15 HDD with Foundation Service solution
- 2C2D EVA 3000-C/8 x 72 GB/10 HDD with Foundation Service solution
- 2C2D EVA 3000-C/8 x 72 GB/15 HDD with Foundation Service solution
- 2C2D EVA 3000-C/8 x 146 GB/10 HDD with Foundation Service solution
- M5314 FC drive enclosure
- 36 GB/15K dual-port 2 Gb FC-AL 1" drive
- 72 GB/10K dual-port 2 Gb FC-AL 1" drive
- 146 GB/10K dual-port 2 Gb FC-AL 1" drive
- 72 GB/15K dual-port 2 Gb FC-AL 1" drive
- SAN switch 2/8-EL
- SAN switch 2/16
- 2 Gb SFP transceiver kit
- 2 m LC-LC Multimode Fibre Channel cable
- 5 m LC-LC Multimode Fibre Channel cable
- 15 m LC-LC Multimode Fibre Channel cable
- 30 m LC-LC Multimode Fibre Channel cable
- 50 m LC-LC Multimode Fibre Channel cable
- HP StorageWorks Virtual Controller Software v3.0 media kit for dual HSV100 controllers
- Storage Management Appliance III
- FCA2214 2 Gb FC host adapter (QLA2340)

## Hardware documentation

HP XC3000 Cluster systems come with two hardware documentation kits. Each hardware documentation kit includes both a CD (contains a soft copy of the *User System and Installation Guide*) and a hard copy of the *User System and Installation Guide*. Additional copies of the XC3000 Cluster hardware documentation kits can be ordered separately.

## HP XC System Software and documentation

HP XC System Software v1.0 is a fully integrated and supported cluster management software solution developed by HP. It consists of a Linux kernel, cluster management software, platform LSF scheduling software, license management software, message-passing interface, MLIB tools, RPMs, and system management modules. The HP XC System Software kit includes binaries, sources, and documentation on CD; hard-copy documentation; and a software use license.

- HP XC System Software v1.0 base license kit—comes with a base unlimited software license, one software media kit, administrator documentation kit, and user documentation kit

- HP XC System Software v1.0 CPU license—one license per CPU is required (must be ordered in addition to base license kit); comes in 1, 2, 16, 64, 128, 256, or 512 CPU license packages

Additional software media kits, administrator documentation kits, and user documentation kits may be ordered separately.

## Development tools for XC3000

- **Fortran and C/C++ compilers** from Intel are provided at added cost from Intel or a reseller to use with the HP XC System Software. The toolkits include the compiler license, documentation, and release notes.
- **TotalView** parallel debugger from Etnus, Inc. **is** provided at added cost from the vendor or a reseller.
- **Vampir/Vampirtrace** MPI performance analysis tool from Pallas **is** provided at added cost from the vendor or a reseller.

## HP factory rack integration

The factory rack integration provided by HP for the XC3000 Cluster includes the following:

- Factory integration must be ordered for each XC3000 Cluster system
- Staging and integration of the Intel Xeon processor–based HP servers, storage devices, and peripheral devices
- Review of customer order for required licenses, cables/cable lengths, software revisions, and hardware
- Thorough test of the system, running extensive diagnostics and system exercisers for an extended period of time to reasonably facilitate a problem-free installation and ongoing reliability
- Configuration of Linux cluster includes these services:
  - Consolidation and routing of all material across the customer purchase order for integration; technical edit and verification of the configuration against the order
  - Verification of hardware configuration
  - Cabling of entire cluster, labeling of all inter-cabinet cabling
  - Configuration of cluster interconnect—Myrinet XP
  - Configuration of each node's iLO port and network port
  - Configuration of service and administration nodes
  - Loading of Linux software on the system
  - Configuration of disk partitions
  - Setup of IP addresses for systems and all network equipment
  - Loading of configuration files
  - Verification of cluster configuration
- System labeling, including color-coding or point-to-point labels of all cables, device labels, and system node name labels
- Configuration of external storage, SAN storage, or local embedded file system storage

## Mandatory HP field installation

All HP XC3000 customers are required to have onsite customer installation of the system provided by HP technicians. This includes the unpacking of the equipment, inspecting the equipment for damage, positioning and joining the cabinets together as required, cabling up the system (power and data cables), and verifying proper operation of the equipment and running diagnostics. There is a charge for the XC3000 Cluster cabinets as well as a separate charge for EVA storage cabinet installation

(using the normal EVA installation charges). The field installation activities are focused on setting up the equipment and not on the software installation. HP Consulting and Integration (C&I) activities cover the setup and startup of the XC3000 Cluster management software. The installation for each cluster is ordered at the time of the initial system purchase and must be included on the order.

## Mandatory HP Consulting and Integration Services

All HP XC3000 customers are required to have onsite customer integration management and systems software knowledge transfer sessions provided by HP technicians. This must be ordered at the time of the initial system purchase. Following are the required startup services and optional services available for each size of XC3000 Cluster system. Optional XC Cluster user training is also available.

- Required startup services
  - One-day onsite systems software knowledge transfer for less than 18 nodes
  - Two-day onsite systems software knowledge transfer for 18–34 nodes
  - Three-day onsite systems software knowledge transfer for 35–69 nodes
  - Five-day onsite systems software knowledge transfer for 70–136 nodes
  - Ten-day onsite systems software knowledge transfer for 137–272 nodes

- Optional services—recommended for less than 34 nodes
  - Cluster Integration Management
  - Cluster Systems QuickStart
  - Cluster Applications QuickStart

  - Optional services—recommended for 34–135 nodes
  - Cluster Integration Management
  - Cluster Systems QuickStart
  - Cluster Applications QuickStart

  - Optional services—recommended for 136–272 nodes
  - Cluster Integration Management
  - Cluster Systems QuickStart
  - Cluster Applications QuickStart

- Optional training
  - Cluster systems administration course
  - Cluster applications migration course

Contact HP C&I Services (Frank Pietryka—**Pietryka@hp.com**) for more details and information on configurations for additional C&I Service offerings.

## HP customer support

HP customer support provides onsite hardware break/fix support and remote, remedial software call-center support. Customer support offsite software services include level 1 and 2 support:

- Level 1 is defined as everyday user/system administration issues.
- Level 2 is defined as problems related to installation and configuration, along with other problems not solvable by following the vendor-supplied documentation.

The HP software support team will work in parallel with the appropriate vendor and development groups to address level 3 and 4 support elevations:

- Level 3 elevations typically require that patches and modifications be generated by the vendor to resolve deficiencies in the product.
- Level 4 elevations deal with enhancements in the functionality of the product that will typically be included in future releases.

Software contracts can be tailored to meet any customer needs, including remedial break/fix, migration and upgrade planning, and a full suite of proactive deliverables. These services are available through HP Care Packs—to cover all needed levels of services for a limited number of instances—or through software contracts, from the Bronze up to the Platinum and Custom levels of service.

# Technical specifications

## HP ProLiant DL380 G3 server—service node

| | | |
|---|---|---|
| **Dimensions (h x w x d)** | | 3.4 x 19.0 x 25.75 in./<br>8.64 x 48.26 x 65.41 cm |
| **Weight** | Maximum | 60 lb./27.22 kg |
| | No drive | 47.18 lb./20.40 kg |
| **Input requirements (per power supply)** | Range line voltage | 90 to 132 VAC/180 to 265 VAC |
| | Nominal line voltage | 100 to 120 VAC/220 to 240 VAC |
| | Rated input current | 6 A (110 V) to 3 A (220 V) |
| | Rated input frequency | 50 to 60 Hz |
| | Rated input power | 600 W |
| **BTU rating** | Typical heat dissipation/hr. | 1,475 BTU/hr. |
| **SCSI connectors** | 3 HD68 connectors (2 internal, 1 external) supporting Ultra2 SCSI | |
| **Power supply output power (per power supply)** | Rated steady-state power | 400 W |
| | Maximum peak power | 400 W |
| **Temperature range** | Operating | 50° to 95° F/10° to 35° C |
| | Shipping | –40° to 158° F/–40° to 70° C |
| **Relative humidity (non-condensing)** | Operating | 10% to 90% |
| | Non-operating | 5% to 95% |
| **Maximum wet bulb temperature** | | 82.4° F/28° C |
| **Acoustic noise** | Idle<br>(fixed disk drives spinning) | |
| | $L_{WAd}$ (bels) | 6.6 |
| | $L_{pAm}$ (dBA) | 49 |
| | Operating<br>(random seeks to fixed disks) | |
| | $L_{WAd}$ (bels) | 6.8 |
| | $L_{pAm}$ (dBA) | 50 |

# HP ProLiant DL360 G3 server—application node

| | | |
|---|---|---|
| **Dimensions (h x w x d)** | | 1.65 x 19.0 x 25 in./4.19 x 48.26 x 63.5 cm |
| **Weight** | Maximum | 28.6 lb./13 kg |
| | No drives | 26 lb./11.80 kg |
| **Input requirements (per power supply)** | Range line voltage | 100 to 240 VAC |
| | Nominal line voltage | 100 to 120 VAC/220 to 240 VAC |
| | Rated input current | 2.66 A (110 V) to 1.33 A (220 V) |
| | Rated input frequency | 50 to 60 Hz |
| | Rated input power | 292 W |
| **BTU rating** | Heat dissipation/hr. | 966 BTU/hr. |
| **SCSI connectors** | | |
| **Power supply output power (per power supply)** | Rated steady-state power | 170 W |
| | Maximum peak power | 190 W |
| **Temperature range** | Operating | 50° to 95° F/10° to 35° C |
| | Shipping | –22° to 122° F/–30° to 50° C |
| **Relative humidity (non-condensing)** | Operating | 8% to 90% |
| | Non-operating | 5% to 95% |
| **Maximum wet bulb temperature** | | 101.1° F/38.7° C |
| **Acoustic noise** | Idle (fixed disk drives spinning) | |
| | $L_{WAd}$ (bels) | 6.5 |
| | $L_{pAm}$ (dBA) | 51 |
| | Operating (random seeks to fixed disks) | |
| | $L_{Wad}$ (bels) | 6.6 |
| | $L_{pAm}$ (dBA) | 52 |

## Myricom Myrinet XP high-speed interconnect

| Myricom 128-port chassis | |
|---|---|
| **Maximum BTUs per hour (with line cards in all slots)** | 3360 |
| **Power [(max.)?]** | 960 W |
| **Maximum input current (100 to 127 V)** | 9.6 A |
| **Maximum input current (200 to 240 V)** | 4.8 A |
| **Height (1U = 1.75 inches = 44.45 mm)** | 9U |
| **Width** | 17.18 in./43.6 cm |
| **Depth** | 17.5 in./44.5 cm |
| **Weight without line cards** | 46.3 lb./21 kg |
| **Weight with line cards in all slots** | 68.3 lb./31 kg |
| **Temperature (operating)** | 41° to 104° F/5° to 40° C |
| **Temperature (storage)** | –40° to 158° F/–40° to 70° C |
| **Relative humidity** | 90% at 65° C |
| **Shock and vibration** | Conforming to EN 60068 (IEC 68) |
| **Fan MTTF** | 500,000 hours (each) |
| **Power input voltage** | 100 to 127 V/200 to 240 V 50/60 Hz |
| **Power factor** | .99 |

## HP ProCurve Switch 2650 (48-port 10/100 switch)

| | |
|---|---|
| **Dimensions (h x w x d)** | 17.4 x 8.0 x 1.8 in./44.2 x 20.3 x 4.6 cm |
| **Weight (single power system)** | 6.0 lb./ 2.7 kg |
| **Temperature (operating)** | 32° to 131° F/0° to 55° C |
| **Temperature (storage)** | N/A |
| **Humidity** | Operating—15 to 95% @ 104° F/40° C, non-condensing<br>Non-operating—90% @ 149° F/65° C, non-condensing |
| **Voltage** | 100 to 127 VAC/200 to 240 VAC 50/60 Hz |
| **Power** | 36 W |
| **Heat dissipation** | 123 BTU/hr. |

## HP Rack 10642 (42U)—base cabinet (empty)

| Height | Total cabinet<br>78.7 in./199.90 cm | Shipping<br>86.22 in./219 cm |
|---|---|---|
| **Depth** | 39.69 in./100.82 cm | 48 in./121.92 cm |
| **Width** | 24 in./60.96 cm | 32 in./81.28 cm |
| **Weight** | 253 lb./114.8 kg | 325 lb./147.4 kg |
| **PDU (qty. varies with cabinet)** | NA/Japan—220 to 240, 24 amp, 50/60 Hz | Int—220 to 240, 32 amp, 50/60 Hz |
| **Color** | Doors: graphite metallic; frame: carbon | |

## HP StorageWorks Modular SAN Array 1000

Refer to MSA1000 specifications on the HP StorageWorks Web site.

The MSA1000 storage array is mounted in the XC3000 utility rack when configured with an XC3000 Cluster.

## HP StorageWorks Enterprise Virtual Array 3000

Refer to EVA 3000 specifications on the HP StorageWorks Web site.

The EVA 3000 comes configured and mounted in its own standalone 10000 series cabinet when configured with an XC3000 Cluster.

## HP XC3000 compute building block rack (preliminary)

The compute building block has one 10642 rack, 32 DL360 nodes, two DL380 nodes, two HP ProCurve 2650 switches, three PDUs, and cabling (maximum configuration).

| Height | Total cabinet<br>78.7 in./199.90 cm | Shipping<br>86.22 in./219 cm |
|---|---|---|
| **Depth** | 39.691 in./100.82 cm | 48 in./121.92 cm |
| **Width** | 24 in./60.96 cm | 32 in./81.28 cm |
| **Weight (est.)** | 1800 lb./816 kg | 1872 lb./849 kg |
| **Power connection—3 PDUs** | NA/Japan—220 to 240, 24 amp, 50/60 Hz (each PDU) | Int—220 to 240, 32 amp, 50/60 Hz (each PDU) |
| **Power consumption (est.)** | 11,000 W (prelim.) | |
| **Heat dissipation (est.)** | 35,000 BTU/hr. (prelim.) | |
| **Temperature range (est.)** | 50° to 95° F/10° to 35° C operating (prelim.) | –22° to 122° F/–30° to 50° C non-operating (prelim.) |
| **Relative humidity (est.) (non-condensing)** | 10% to 90% operating (prelim.) | 5% to 95% non-operating (prelim.) |

## HP XC3000 utility building block rack (preliminary)

The utility building block has one 10642 rack, one Myricom 17-slot chassis, TFT display, one HP ProCurve 2650 admin switch, one MSA1000 storage array, two PDUs, and cabling (maximum configuration).

| **Height** | **Total cabinet**<br>78.7 in./199.90 cm | **Shipping**<br>86.22 in./219 cm |
|---|---|---|
| **Depth** | 39.691 in./100.82 cm | 48 in./121.92 cm |
| **Width** | 24 in./60.96 cm | 32 in./81.28 cm |
| **Weight (est.)** | 600 lb./272 kg | 672 lb./305 kg |
| **Power connection—2 PDUs** | NA/Japan—220 to 240, 24 amp, 50/60 Hz (each PDU) | Int—220 to 240, 32 amp, 50/60 Hz (each PDU) |
| **Power consumption (est.)** | 2500 W (prelim.) | |
| **Heat dissipation (est.)** | 7200 BTU/hr. (prelim.) | |
| **Temperature range (est.)** | 50° to 95° F/10° to 35° C operating (prelim.) | –22° to 122° F/–30° to 50° C non-operating (prelim.) |
| **Relative humidity (est.) (non-condensing)** | 10% to 90% operating (prelim.) | 5% to 95% non-operating (prelim.) |

## HP XC3000 interconnect building block rack (preliminary)

The interconnect building block has one 10642 rack, four Myricom 17-slot chassis, and two PDUs (maximum configuration).

| **Height** | **Total cabinet**<br>78.7 in./199.90 cm | **Shipping**<br>86.22 in./219 cm |
|---|---|---|
| **Depth** | 39.69 in./100.82 cm | 48 in./121.92 cm |
| **Width** | 24 in./60.96 cm | 32 in./81.28 cm |
| **Weight (est.)** | 700 lb./318 kg | 772 lb./350 kg |
| **Power connection—2 PDUs** | NA/Japan—220 to 240, 24 amp 50/60 Hz (each PDU) | Int—220 to 240, 32 amp 50/60 Hz (each PDU) |
| **Power consumption (est.)** | 4000 W (prelim.) | |
| **Heat dissipation (est.)** | 14,000 BTU/hr. (prelim.) | |
| **Temperature range (est.)** | 50° to 95° F/10° to 35° C operating (prelim.) | –22° to 122° F/–30° to 50° C non-operating (prelim.) |
| **Relative humidity (est.) (non-condensing)** | 10% to 90% operating (prelim.) | 5% to 95% non-operating (prelim.) |

# Appendix—HP XC3000 product menu

Refer to the HP XC Web site at **www.hp.com/techservers/clusters/xc_clusters.html** for the latest XC3000 product menu, and consult the HPTC Competency Center in your region for ordering information.



5982-1779EN Rev. 2, 03/17/2004