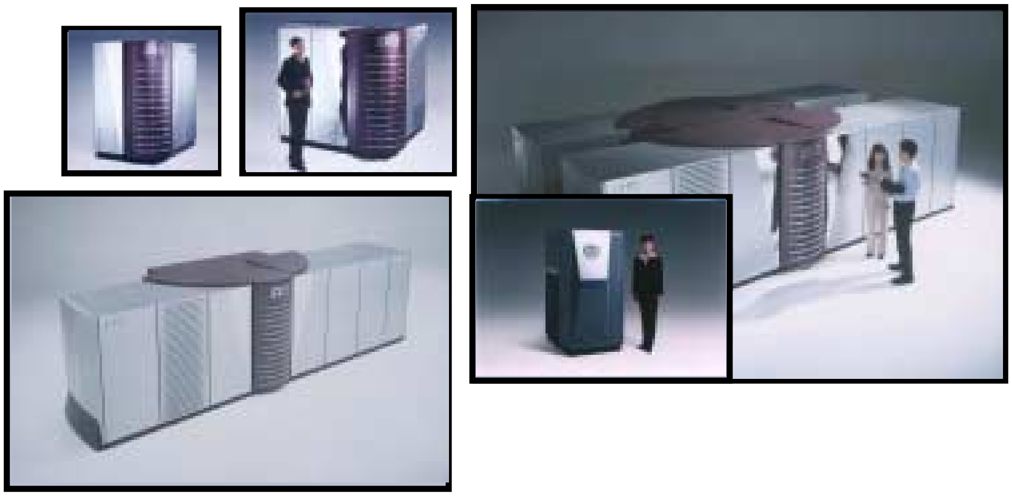
CRAY SX-5 Family

Parallel Vector Supercomputing



General Description

Cray Inc.
Updated Release
June 11, 2001

Contents

1.Introduction

The SX Series of vector supercomputers has provided innovative, high performance supercomputer products since 1985. Today the elements of SX architecture are the de facto standard for virtually all high-performance computing products. SX systems differentiate themselves by being designed specifically for technical users needing very high performance for modeling and simulation applications such as weather/climate, computational fluid dynamics, seismic processing, and other numerically intensive and numerically complex disciplines.

SX architecture focuses on the benefits derived from the highest possible sustained performance from each individual processor. To achieve this a high bandwidth, low latency memory subsystem is an integral part of the architectural equation, as are powerful, independent I/O processors.

SX architecture is a solutions-based custom approach to design. As such, it provides proven vector technology coupled to a very high bandwidth main memory. Compare this to the aggregate performance approach required by using commodity off-the-shelf microprocessor parts. The difference is in mapping software solutions to a hardware platform; it is important to recognize applications limitations. For example, most third party applications are not available, or are not mature, for distributed memory systems.

Most industrially relevant applications tend to be either a) not parallelized, b) parallelized for shared memory only, c) available with limited capability for distributed memory parallel systems, or d) have limited intrinsic parallelism that restricts how many processors can be applied. Four to eight processors are a common practical limit for most full-feature parallelized applications today. Thus a 4 processor 32 GFLOPS shared memory system can provide substantially better solutions capability than a 2048 processor 1 TFLOPS distributed memory system on certain critical industrial applications. Some examples of this type of application include finite element analysis and crash simulations.

Cray Inc. along with its partner, NEC, has taken the lead in putting PVP systems back into the mainstream of supercomputing. Among the firsts provided by the SX Series were the use of economical all-CMOS technology in a high-end supercomputing product, a high performance and high capacity SDRAM shared main memory, and the high speed IXS internode crossbar switch that supports instruction extensions and global addressing.

2. Scalable Computing

Today, so-called “scalable systems” are a popular category. The goal of a scalable system is to have both hardware and application software that will execute on a small platform, perhaps even a single processor computer, and then by adding more processors and more memory be able to solve larger, more complex, or more data intensive applications problems.

The concept began with massively parallel systems (MPP) which evolved into SMP clusters. Why the change? MPP simply did not work viably for commercial acceptability, so evolution to suit applications requirements was necessary. Although SX architecture is usually categorized as a parallel vector processor (PVP), it contains all of the elements of an SMP node, and Multi Node SX architecture is a superset of SMP clustering.

2.1 Applications Considerations

Although software developer’s goals are to sell applications, scalable computing appears to be an efficient model for them, and should minimize applications support problems. However, there is no standard architecture for distributed memory systems; for performance, each application has to be specifically tuned for each model of distributed memory computer. Reality therefore forces truly scalable software to be either non-portable, or of low relative performance on alternate platforms.

Because of the inherent difficulties writing parallel applications software for distributed memory systems SX architecture continues to evolve around a shared memory building block called a Single Node. Shared memory enables the simplest applications development and is fully compatible with all programming paradigms.

The solution to the requirement for wide scalability is a hybrid system. A powerful SMP with shared memory provides performance and capacity that enables industrial applications to be supported at the capability level. Beyond shared memory capacity limits, a distributed type memory is necessary to support the specially written applications that can utilize tens or hundreds of processors.

This latter class of application is almost exclusively the realm of own-codes (locally developed and supported) as found in certain weather and climate, some seismic processing, molecular dynamics, nuclear simulation, and some crash analysis, as examples. However, even when an application might be suitable for a distributed memory system, it may still be problematic to implement. One issue is that some applications, seismic processing and weather forecasting being examples, have huge volumes of serialized input or output data that cannot be processed in a parallel manner. As an example, a seismic processing input data set can easily be hundreds of gigabytes of serial data.

2.2 Execution Platforms

The scalable SMP product lines offered by multiple vendors are based on commodity microprocessor technology and low cost, high latency, low bandwidth memory systems. Further, the interconnections between SMP nodes are substantially slower than access to main memory, which creates substantial idling time while messages complete; this reduces performance and is a main reason that most scalable systems exhibit maximum performance in the 5-15% of peak range.

The designers responded to the need for powerful, efficient scalability with the SX Multi Node configuration. It provides individually powerful processors, very large high bandwidth shared main memory, and the IXS (Internode Crossbar Switch) to enable a tightly coupled cluster-like system with shared and distributed memory characteristics. The distinguishing features of the Multi Node configuration are that it is scalable, it takes advantage of both shared and distributed memory programming models, and has bandwidth for distributed memory that is 4 times the nearest competitor system.

The SX-5 family thus provides real applications performance, real capability, real capacity, and real scalability. The SX-5 family of scalable parallel vector processors is leading the industry into the future.

3. SX Development

The SX-5 family of scalable parallel vector supercomputers provides a commercial quality balanced system capable of providing solutions for a broad range of application requirements involving intensive computation, very large high performance main memory, and high input-output rates.

Because of the demonstrated performance on a wide range of applications, vector processing is being reintroduced under various names and formats by virtually all classes of computer product from personal computers to supercomputers. The proven parallel vector processing technology used in the SX-5 family of systems has been enhanced with state-of-the-art scalar capability, a high capacity high performance main memory, and powerful I/O subsystem, to deliver a balanced supercomputer that can address a wide range of application requirements.

Unrivaled experience in technical computing, and long-term commitment, provide assurance that an SX-5 Series choice is a correct choice. Continuous upgrade paths will always be available for SX users.

3.1 Technology

CMOS was introduced to high end supercomputing with the introduction of the SX-4 Series in 1995. CMOS has resulted in supercomputer class products becoming price competitive with high-end workstations. As a result, there is resurgence in the use of parallel vector systems because they work.

Prior to the 1990's, individual gate delays were the single major components of total circuit delay, so the design balance focused on reducing the system clock cycle to minimize gate delays. The technology available essentially dictated the use of bipolar ECL VLSI to achieve stable, fast clock cycles.

Today, individual gate delays are inconsequential relative to inter-chip and inter-board delays. CMOS technology can now provide superior performance because of its very high integration density that allows more circuits to be included on each VLSI chip. The number of signals that must be transmitted off-board or off-chip are minimized, so total circuit delay can also be minimized.

The SX-5 family is designed using 0.25 micron CMOS processes. This level of technology represents the mainstream production available at system introduction in 1998. By use of standard processes the overall cost of the VLSI has been minimized.

Each logic VLSI has over 1600 pinouts on a 25 millimeter square carrier and contains approximately 15 million transistors. This large number of pinouts enables a substantial number of off-chip signals to be available for such important uses as memory paths. This interconnect technology is a key difference between an SX-5 processor and a commodity microprocessor which typically has only a few hundred pinouts, and hence numerous comparative limitations.

Because of the low power dissipation of CMOS circuits the entire system is totally air-cooled. The high gate density made possible by CMOS enables a significantly reduced chip count for each processor as compared to even previous generation CMOS based systems, which results in higher system reliability.

3.2 Shared Memory Architecture

Today all industrially relevant supercomputing applications continue to be designed for shared memory systems such as the SX-5 family. Shared memory architecture provides the highest performance options with the most flexibility for all commonly used programming paradigms because it enables automated compiler vectorization *and* parallelization, and it supports message passing models.

Shared memory enables the compilers to automatically generate parallel tasks. The programmer can rely on compiler decisions, or optimization can be controlled through use of directives.

Shared memory is also superior to distributed memory even for message passing models because the inter-process communications speed can be up to half of memory bandwidth, a significantly higher performance than negotiating through the special switches or protocol laden communications paths used in distributed memory products.

The shared memory architecture employed in SX-5 Series and SX-5e Single Node models, and SX-5S High Performance Servers, thus provides ease of programming and allows for advanced automated parallelization by the compilers while providing the potential for substantially higher performance.

Comparisons of real-world performance for distributed memory systems are most commonly 5-15% of peak rated performance. For shared memory systems 30-60% of peak is expected for comparable effort. On this basis a 256 GFLOPS shared memory system would typically provide equivalent performance to a 1.2 TFLOPS distributed system. This assumes sufficient parallelism to exploit both memory architectures to some advantage, which is not often the case when a large number of processors must be used as on popular distributed memory systems.

3.3 Scalable Architecture

Shared memory systems provide the most applications friendly hardware environment. Shared memory can be implemented as true shared memory – a uniform memory space directly accessed by all processors – or as globally addressed distributed memory – special hardware accesses non-uniform memory located in various nodes as if it were local.

True uniform shared memory, as found on SX-5 Single Node models, provides consistent high bandwidth and low latency accesses for all processors, across the entire memory space. In comparison, globally addressable NUMA provides programming benefits associated with shared memory models, but performance suffers because of the long latencies and low bandwidth for remote memory accesses.

True uniform shared memory has two restrictions. First, a very large number of interconnects are required for the memory subsystem to service multiple processors. Thus there is a physical limitation with packaging technology that restricts the number of interconnects possible.

Second, the response and recovery time of memory chips is long relative to processor speeds. The solution is to implement a large number of independent memory banks. This randomizes memory chip accesses and thereby minimizes repetitive access to any individual memory chip, thus avoiding chip busy waits. Large banking factors add cost to control circuits, increase the number of interconnects, and generally add complexity to the design.

Non-uniform (distributed) memory greatly simplifies the hardware design by placing a large burden on the programmer in the case of non-global addressing. For NUMA with global addressing programming becomes simplified but performance is reduced because of the large disparity of accessing local or remote memory cells.

SX-5 Multi Node configurations provide the best of both memory designs, and thereby provide large scale and effective scalability. Within each individual node a true shared-memory programming model is used. For execution of Multi Node jobs a message passing model is used, as is the case with all current products on the market today.

The SX-5 difference is that the non-uniform memory of the Multi Node cabinets is accessed through the IXS Internode Crossbar Switch, which provides global addressing, global domains for some instructions, and global synchronization features. The 8 gigabytes per second IXS channel has more bandwidth than most complete systems, and each node has a channel pair.

Multi Node configurations thus overcome the physical limitations that restrict shared memory capacity and processor count while providing a uniquely transparent NUMA environment that is treated using standard OpenMP directives for Single Node parallelization, and MPI message passing programming models for Multi Node parallelization.

The resulting scalable system requires only a fraction of the processors needed by other systems to achieve any specific level of performance. It is always easier to implement parallelization for a relatively smaller number of powerful processors because of parallelization ratios, wasted interprocessor communications time, and total system reliability.

Furthermore, the SX-5 scalable architecture has multiple powerful independent IOPs so that there is no critical I/O bottleneck as found on most systems.

4. Architecture

SX-5 Series architecture bridges traditional shared memory parallel vector designs in server and Single Node systems with the scalability of NUMA architecture in Multi Node systems.
Server and Single Node memory architecture is shown in Figure 1 where it can be seen that all processors have uniform and tight coupling to the main memory unit.

Figure 1: SX-5 Single Node Architecture

Figure 2: SX-5 Multi Node Architecture

Figure 2 illustrates Multi Node architecture. The independent shared memory nodes can be seen, as can the coupling to other nodes through the IXS Internode Crossbar Switch. The IXS provides effective, tight hardware coupling between nodes to make a unified system rather than a simple cluster, with a global address space.

A Multi Node configuration can include up to 32 SX-5 Single Node systems in either A or B cabinet models (see section 9). The maximum configuration SX-5/512M32 provides a 5 TFLOPS system with 8 TB of main memory having 32 TB/s of memory bandwidth and 402 GB/s of I/O bandwidth.

4.1 Central Processing Unit

Each SX-5 Series processor (figures 3 and 4) contains a vector unit, scalar unit with on-chip cache for instructions and scalar data, and the memory interface. There are two types of processor, a 10 GFLOPS type used in the SX-5 Series, and a 4 GFLOPS type used for the SX-5e and SX-5S.

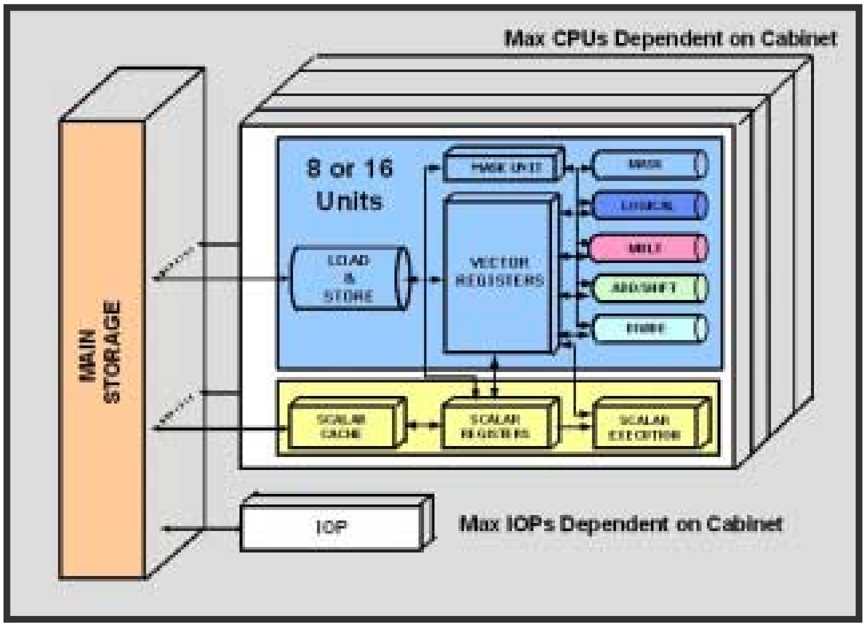


Figure 3: Basic Processor Architecture

The 10 GFLOPS type processors have 16 pipeline sets and are packaged one to a board. The 4 GFLOPS type processors have 8 pipeline sets and are packaged two per board. The scalar units in both processor types are the same.

Each logic chip is a flip-tab with 1600 pinouts. To ensure sufficient memory bandwidth there are more than 11,000 pinouts available on each processor board, most of which are memory paths.

Figure 4: 10 GFLOPS Type Processor Board (Heat Sink Removed)

4.1.1 Scalar Unit

The scalar unit is completely contained on a single VLSI. It includes scalar pipelines, scalar registers, an instruction buffer, a superscalar issue unit, a scalar operand cache, and a instruction cache.

The scalar pipelines include two independent floating point units each capable of multiplication, addition, and division, and two independent integer units.

There are 128 x 64 bit general purpose scalar registers that can be used for such operations as scalar arithmetic and logic, address calculations and indexing, holding vector sums, or holding scalar operands used by vector operations.

The instruction issue unit has an 8-kilobyte instruction buffer that is loaded from operand cache. Up to 4 instructions can be issued per clock. The issue unit has branch prediction and out-of-ordering logic for scalar, vector, and memory referencing instructions.

The scalar operand and instruction caches are each 64-kilobytes capacity.

4.1.2 Vector Unit

SX-5 Series processors are available in two variations. The SX-5 Series has a processor with a 10 GFLOPS vector unit. SX-5e and SX-5S models have a processor with a 4 GFLOPS vector unit.

The vector instruction set includes enhanced instructions with functions such as first order recursion to enhance the range of vectorizable programming constructs. Techniques such as vector instruction chaining are implemented. Intermediate results of complex vector operations (such as SAXPY, for example) can be concurrently stored in vector data registers while being reused for the continuing arithmetic operations without incurring additional overhead; those results can then be reused in other operations or stored to memory.

4.1.2.1 SX-5 Series 10 GFLOPS Type Vector Unit

The 10 GFLOPS type processor used in the SX-5 Series includes vector registers, vector pipelines, and a mask control unit.

There are 72 vector registers in the 10 GFLOPS processor. This is divided into 8 vector registers and 64 vector data registers. Each register is 64 bits wide. All vector instructions can be issued for vector registers, but a restricted set is available for vector data registers. The vector data registers are used as software controlled vector data cache area. Results from vector operations can be concurrently stored in both vector registers and vector data registers. This presence of vector data registers typically reduces required memory bandwidth by approximately 40% compared to vector systems lacking such a feature. Vector data can also be directly loaded and stored from vector data registers.

The vector registers operate into 16 wide vector pipeline sets. Each set includes a vector add-shift pipe, a vector multiply pipe, a vector divide pipe, and a vector logical pipe. During vector operation 16 data elements are operated on in parallel during each clock cycle.

The mask control unit enables vector compress-expand and vector controlled-memory operations such as results of vectorized IF constructs. The mask control unit has one operational register and 7 backing registers.

4.1.2.2 SX-5e and SX-5S 4 GFLOPS Type Vector Unit

The 4 GFLOPS type processor used in the SX-5e and SX-5S includes vector registers, vector pipelines, and a mask control unit.

There are 72 vector registers in the 4 GFLOPS processor. This is divided into 8 vector registers and 64 vector data registers. Each register is 64 bits wide. All vector instructions can be issued for vector registers, but a restricted set is available for vector data registers. The vector data registers are used as software controlled vector data cache area. Results from vector operations can be concurrently stored in both vector registers and vector data registers. This presence of vector data registers typically reduces required memory bandwidth by approximately 40% compared to vector systems lacking such a feature. Vector data can also be directly loaded and stored from vector data registers.

The vector registers operate into 8 wide vector pipeline sets. Each set includes a vector add-shift pipe, a vector multiply pipe, a vector divide pipe, and a vector logical pipe. During vector operation 8 data elements are operated on in parallel during each clock cycle.

The mask control unit enables vector compress-expand and vector controlled-memory operations such as results of vectorized IF constructs. The mask control unit has one operational register and 7 backing registers.

4.1.3 Numeric Support

The SX-5, SX-5e and SX-5S processors support IEEE 32 bit and 64 bit data in the vector and scalar units. The scalar unit also supports extended precision 128 bit data. The vector and scalar units support 32 and 64 bit fixed point data. The scalar unit can also operate on 8 and 16 bit signed and unsigned data.

4.1.4 Synchronization Support

Each processor has a dedicated 128 x 64-bit register set for synchronization of parallel processing tasks at the program level. Additionally each cabinet has a 128 x 64-bit privileged register set for the operating system

The synchronization registers have instructions such as load, store, test, test-and-set, store-add, etc. to implement the most efficient programming synchronization logic possible.

4.1.5 IXS Internode Crossbar Switch Processor Extensions

Multi Node models can be configured using either SX-5 Series or SX-5e models in either A or B size cabinets. The IXS Internode Crossbar Switch (IXS) includes an adapter unit housed in each mainframe cabinet and an external cabinet for the IXS itself.

The purpose of the IXS is to provide a tightly integrated scalable system configuration. It has functions for processor instructions and memory functions. This section addresses processor functions.

For Multi Node job classes using resources in more than one SX-5 or SX-5e cabinet, the IXS provides a global synchronization register set similar to that in each processor, but of global scope. There are 8 x 64 bit global synchronization registers available for each node.

The IXS also implements a restricted set of global instructions targeting the communications and synchronization of processes in Multi Node jobs. An example of a globally implemented instruction is *interrupt-many* whereby a single instruction issue can interrupt an address list of processors located throughout the Multi Node configuration.

Each hardware component (such as processors and IOPs) has its node number as part of its unique hardware identifying address. This facilitates globalizing certain instructions.

4.2 Memory Unit

The SX-5 Series provides the highest bandwidth and highest capacity memory available on any shared memory system designed to date. It is built with cost effective SDRAM components and implemented with a very high banking factor of up to 16,384 banks per cabinet with board-localized reordering logic.

The bandwidth available *per processor* is 80 gigabytes per second for 10 GFLOPS types, and 32 gigabytes per second for 4 GFLOPS types. In each case there is sufficient bandwidth available to each processor for its peak performance rating, and the rating is substantially higher than that provided by competing systems. Up to 256 gigabytes capacity can be configured in an A cabinet system.

A main memory board is shown in figure 5. Each board holds 8 gigabytes of 128 megabit SDRAM chips plus SECDED error correction coding bits. The memory modules, shown in Figure 6, plug into the main memory board.

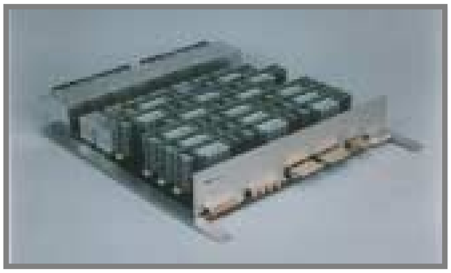


Figure 5: 4 GB Memory Board

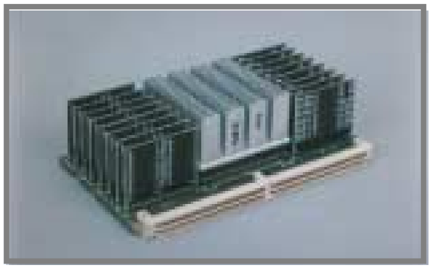


Figure 6: 256 Mb Module

4.2.1 Shared Main Memory

The shared memory architecture within each SX-5, SX-5e, and SX-5S cabinet is a non-blocking crossbar that provides uniform high-speed access. This architecture constitutes a symmetric multiprocessor system (SMP) in a parallel vector processor (PVP) format. Figure 1 logically illustrates how the processors all work on a shared uniform memory space.

All models are real memory mode machines that utilize page mapped addressing. The page mapped architecture allows program modules to be non-contiguously loaded, eliminating the need for periodic memory compaction procedures by the operating system and enabling the most efficient operational management techniques. As with all real memory mode machines, an entire program must be resident in memory for execution.

4.2.2 IXS Enabled Distributed Main Memory

SX-5M and SX-5Me configurations consist of multiple A or B cabinet systems interconnected through an IXS Internode Crossbar Switch (IXS). While each cabinet, referred to as a *node*, provides local shared memory, an access to memory in another node is through the IXS, which provides a globalized non-uniform memory architecture (NUMA) for programming purposes. A logical Multi Node configuration is shown in Figure 2 where the individual processor nodes and crossbar interconnects are represented..

For access to other nodes, the IXS provides cross-node page translation tables, synchronization registers, and enables global data movement instructions as well as numerous cross-node instructions. Memory addresses include the node number (as do CPU and IOP identifiers).

Processors in each node have full performance access to their entire local memory, and consistent but reduced performance access to the memory of all other nodes. All memory, both local and remote, is protected by lock-and-key mechanisms under control of the SUPER-UX Multi Node operating system. Even cross-node memory access is protected because of specialized hardware facilities in the IXS system.

Latencies for internode NUMA memory access are less than most workstation technology NUMA implementations, and the 8 gigabyte per second bandwidth of just a single IXS channel exceeds the entire memory bandwidth of most SMP class systems. Even so, because the internode startup latencies are greater than local memory accesses, internode NUMA access is best utilized by block transfers of data rather than by the transfer of single data elements.

This architecture, introduced with the SX-4 Series, has been popularized by many products and lends itself to a combination of traditional parallel vector processing (microtasking) combined with message passing (MPI). Message passing alone is also highly efficient on the architecture.

4.3 Input-Output Processor (IOP)

Each SX-5 type IOP supports channel cards including HIPPI-800, FC-AL, and Ultra SCSI for peripherals. HIPPI-800, ethernet (through 1000 Base T), and FDDI are offered for networking. ATM is also available. The bandwidth and total number of channels is cabinet dependent. Refer to section 9 for specifications.

The IOPs operate asynchronously with the processors as independent I/O engines so that central processors are not directly involved in reading and writing to storage media as is the case of workstation technology based systems.

4.3.1 HIPPI Channels

The 100 megabyte per second HIPPI-800 channel is a popular high performance standard used for both peripherals and network connectivity. Typical HIPPI-800 *sustained* production data rates to either the NEC N7764 RAID or Gen5 XLE Storage Systems[1] are typically over 75 megabytes per second.

4.3.2 FC-AL Channels

The 100 megabytes per second FC-AL channel supports connection of the most modern, cost-effective peripheral devices such as the NEC PoleStar RAID. FC-AL devices are providing performance equivalent to HIPPI-800 devices, but at prices reflecting the commodity market level. Because various FC-AL RAID products are specifically optimized for technical computing or commercial computing careful selection is required.

[1] Gen5 XLE is a trademark and product of Sun Microsystems/MAXSTRAT Division, Milpitas, CA.

4.3.3 SCSI Channels

Native Ultra and FWD SCSI channels are available. Direct SCSI support enables a wide choice of storage options. SCSI channels are most useful for connecting to tape devices. Most tape devices easily maintain their maximum data rate possible using SCSI channel interfaces.

Even modern SCSI RAID does not provide the performance required for supercomputing use, so although they can be configured, SCSI RAID is not recommended except for online bulk storage.

4.4 IXS Internode Crossbar Switch

The IXS Internode Crossbar Switch (IXS) is a proprietary device that connects SX-5 Series nodes using proprietary channels having 8 gigabyte per second bandwidths. The physical IXS is housed in its own cabinet. The IXS connection into each node's crossbar is through an IXS interface unit. Only SX-5 systems in A, B, and Be type cabinets can be configured with an IXS interface unit. The IXS is a non-blocking crossbar that provides up to 256 gigabytes per second of bandwidth for a full 32 node system.

The IXS provides very tight coupling between nodes enabling a single system image for both hardware and software. Examples of the facilities provided include internode addressing and page mapping, remote unit control, internode data movement, and remote processor instruction support as described in sections 4.1.5 and 4.2.2.

Each IXS adapter has memory address translation tables that are loaded by the operating system on a per-multi-node-job basis. The IXS page tables provide a means for processes in a Multi Node job to "see" data at common addresses. This greatly facilitates messaging between distributed processes since messaging becomes analogous to a memory reference rather than a network block.

The IXS supports both synchronous and asynchronous data transfers. Synchronous transfers are limited to 2 kilobytes, and asynchronous to 32 megabytes, per IXS-cross-node instruction.

5. Operating System Software

All models in the SX-5 family use the SUPER-UX operating system. It is UNIX based with substantial enhancements to support supercomputing requirements, especially for I/O, large scale job scheduling, accounting, and security.

Some of the major SUPER-UX enhancements include:

- Multi Node operation
- global shared memory software
- enhanced NQS batch subsystem
- configuration and logical partitioning options
- high performance file systems
- checkpoint/restart
- multilevel security option
- file archiving management
- unattended operation and control facilities

5.1 Multi Node Facility

SX-5M and SX-5Me Multi Node systems provide special hardware features in the IXS Internode Crossbar Switch, which enables efficient use of the total system. Many Multi Node Facility features are also available through channel connected clustered systems. The Global File System (section.5.4) provides a global file directory and access system for all jobs executing anywhere on the Multi Node system.

The Multi Node SUPER-UX kernel is enhanced to recognize a *Multi Node job class*. When a Multi Node job (i.e. a job using processors on multiple nodes) enters the system, the kernel will sequence all of the processes across the nodes, initialize the IXS page translation pages for the job, and provide specialized scheduling commensurate with the resources being used. Once initialization is complete the distributed processes included in the Multi Node job class can execute and communicate without significant operating system involvement.

Because of the resources used by such large-scale jobs, processes are typically not eligible for swapping and are given a high scheduling priority to move them through the system as quickly as possible. To allow more typical timesharing work to coexist with large Multi Node jobs, SUPER-UX has enhanced configuration and logical partitioning features (see section 5.4) to allow significant flexibility in overall system operation.

5.2 Global Memory

SX-GM (Global Memory) is a software facility that improves performance of MPI message passing within Single Node systems, and which provides program level single system image of common data in Multi Node systems. Using SX-GM results in a message simply being copied from memory location to memory location rather than traversing a network-like software path, greatly reducing the overhead normally experienced with message transmission.

Using SX-GM, a message is sent by two memory copies rather than through network-like software. Because of memory security, one copy through system space is performed followed by a copy to the destination user space when performed within Single Node shared memory space. On multinode transfers standard memory space protection validation is performed.

5.3 NQS Batch Subsystem

SUPER-UX NQS provides a batch scheduling environment. It is enhanced to add substantial user control over work in progress. Extended command capabilities include such features as *qcat,* which snapshots stdout and stderr files from an executing batch script and makes the copy available to the user.

NQS queues, queue complexes, and the full range of individual queue parameters and accounting facilities are supported. The NQS Queues have substantial scheduling and resource limit parameters available including time slicing, cpu time limits, maximum memory limits, etc.

Queue scheduling options such as banded and preemptive categories are supported to efficiently handle operations such as those found at mission critical facilities such as weather forecasting services.

5.4 Configuration and Logical Partitioning Options

SUPER-UX provides Resource Grouping that allows the system administrator to define logical and discrete scheduling groups that are mapped onto the processors.

Each Resource Group (RSG) has a maximum and minimum processor count, memory limits, and scheduling characteristics so that the system can be defined as multiple logical environments. For example, one portion of a system can be defined primarily for interactive work while another may be designated for non-swappable parallel processing scheduling using a FIFO scheme, and a third area can be configured to optimize a traditional parallel vector batch environment. In each case, any idle resources within an RSG used can be optionally “borrowed” by another RSG.

5.5 File Systems

SUPER-UX supports the native Supercomputing File System (SFS), a special performance file system (SFS/H), Multi Node Global File System (GFS), memory file system (MMF) and networked file systems (NFS and DFS).

All non-NFS file systems take advantage of high performance I/O caching functions that utilize MFF (see 5.5.4). The caching functions can be optimized on a file system basis. Further, file systems can be created on the MFF to provide exceptionally high file system performance.

Networked file systems such as NFS and DFS utilize standards based transfers to ensure operational compatibility with other systems.

5.5.1 Supercomputing File System

The native file system is Supercomputing File System (SFS). SUPER-UX can manage up to 8,192 terabytes of SFS on-line storage.

SFS has a flexible per-file- system caching facility that utilizes MFF. Numerous parameters can be set such as MMF cache size, write-through threshold limits, basic allocation unit size, and basic read/write unit size.

Other advanced features include device overflow and file systems that span multiple physical devices.

5.5.2 Supercomputing File System, Special Performance

The special performance Supercomputing File System, SFS/H, provides reduced overhead, even compared to the highly efficient SFS. SFS/H is designed to provide the absolute highest performance possible for large-scale data files.

5.5.3 Global File System

Multi Node systems offer the SX Global File System (SX-GFS) that enables the entire Multi Node configuration to view a single coherent file system. Keeping a transparent NFS-like user interface, SX-GFS provides performance approaching that of native disk subsystems – a much higher transfer rate than is possible with NFS or DFS file systems.

5.5.4 Memory File Facility

SX-MFF (Memory File Facility) provides a site defined high performance file system area that is resident in main memory. MFF space is also typically allocated for SUPER-UX high performance I/O caching features supporting SFS and SFS/H file systems.

SX-MFF space is defined at system startup and is locked into main memory to ensure the highest performance.

5.6 Checkpoint/Restart

Checkpoint/restart is valuable for interrupting very long executions for preventative maintenance, or to provide a restart mechanism in case of catastrophic system failure, or for recovery of correctable data errors. NQS batch jobs can be checkpointed by the owner, operator, or NQS administrator. No special programming is required for checkpointing.

5.7 Archiving Management

SUPER-UX provides SXBackStore archiving management software. Like typical archiving packages, SXBackStore provides control and decision functions to migrate specified file classes from disk residency to remote file or tape systems. The user can list all files as if they were local, and SXBackStore will transparently and automatically restore any archived file that is opened.

Primary archival devices include IBM, Sony, and Storage Technology robotic library systems.

SXBackStore provides flexibility so that the SX file archiving can be to a front-end file system managed by an archival host, or the entire archival process can be controlled on SUPER-UX to directly connected library systems.

In the future SX-GFS File systems hosted by a Linux server can also take advantage of UniTree as the archival manager.

5.8 Multilevel Security

The Multilevel Security (MLS) option is provided to support site requirements for classified projects or systems requiring restricted and controlled access. Security levels are site definable as to both names and relationships. Security access levels and permissions are tagged on an object level basis.

In addition to access control to objects, accesses and access attempts can be logged for audit traces.

5.9 Automatic Operation Facility

SX systems provide hardware and software options to enable operator-less environments. The system can be preprogrammed to power on, boot, enter multi-user mode, and shutdown-power off. Any event that can be made visible to software and responded to by closing a relay or executing a script can be serviced.

The automatic operation system includes a hardware device called the Automatic Operation Controller (AOC). The AOC serves as an external control and monitoring device for the SX system. The AOC can perform total environmental monitoring, including earthquake detection.

Cooperating software executing in the SX system communicates system load status and enables the automatic operation system to execute all UNIX functions necessary for system operation.

6 Compilers

FORTRAN90, C++, and HPF languages are supported on the SX-5. An optimized MPI message passing library is available. OpenMP is supported.

The compilers provide advanced automatic optimization features including automatic vectorization, automatic parallelization, partial and conditional vectorization, index migration, loop collapsing, nested loop vectorization, calculation conversions, common expression elimination, code motion, exponentiation optimization, optimization of masked operations, loop unrolling, loop fusion, inline subroutine expansion, conversion of division to multiplication, and instruction scheduling.

Compiler options and directives provide the programmer with considerable flexibility and steering of various optimizations.

6.1 FORTRAN90/SX

FORTRAN90/SX is offered as a native compiler as well as a workstation based cross development system that includes full compile and link functionality. FORTRAN90/SX offers automatic vectorization and parallelization applied to standard, portable FORTRAN90 codes. In addition to the listed advanced optimization features, FORTRAN90/SX includes data trace analysis and a performance data feedback facility

6.2 HPF/SX

HPF development on SUPER-UX is targeted toward SX Series Multi Node systems. NEC participates in the HPF forums working in the United States and Japan with a goal of further developing and improving the HPF language. HPF2 was defined, and the Japan Association for HPF (JAHPF) is supporting additional enhancements to the HPF2 language.

6.3 C++/SX

C++/SX shares its “back end” with FORTRAN90/SX, and as such provides comparable automatic vectorization and parallelization features. C++ is used for both C and C++ program compilation.

It should be noted that C++ object constructs are often written in a scalar context, and are thus not suitable for obtaining the potential performance available from vector systems such as the SX-5 family. Objects that operate on large vector data objects are conducive to obtaining good performance

Furthermore the discrete object level operators used in typical C++ programming often results in substantial software overhead in the codes themselves; specifically function entry, exit, and parameter passing often becomes overwhelming relative to the actual computation time.

7 PSUITE Integrated Development and Tuning Environment

PSUITE is the integrated development environment for SUPER-UX. It is available as a native environment hosted on the SX-5 Series with X-term displays, and as a cross environment for most popular workstations. PSUITE supports FORTRAN90, C, and C++ applications development.

7.1 Editing

The vi editor is the default used by the PSUITE Source Browsing function. Most editors are compatible with PSUITE windowing and can be substituted for vi at user discretion. The browsing function is displayed in an integrated X-Window.

7.2 Compiling

FORTRAN90, C, and C++ are supported by PSUITE. All major compiler options are available through pull downs and X-Window style boxes. Commonly used options can be enabled with buttons, and free format boxes are available to enter specific strings for compilation and linking.

7.3 Project Management

An automatic makefile can be generated and if necessary, tailored through editing. Therefore, the entire programming project can be maintained and rebuilt using simple point and click selections.

7.4 Debugging

Debugging is accomplished using the xdbx parallel multiple windowing debugger. Enhanced capabilities include the graphical presentation of data arrays in various 2 or 3 dimensional styles.

Figure 7 shows a graphical data display of a selected array section as it exists at the breakpoint. The array of interest is simply highlighted with the mouse cursor for selection.

The Totalview debugger is also offered.

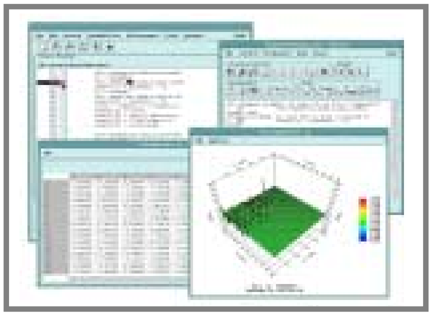


Figure 7: Debug Data Browser

7.5 Application Tuning

Integrated tuning is accomplished by use of parallelization tools, tuning tools, and integrated message-passing traces.

A windows in Figure 8 shows the parallelization assistance window that displays a flow diagram of procedural relationships. This analysis displays dependencies between potentially parallel code sections.

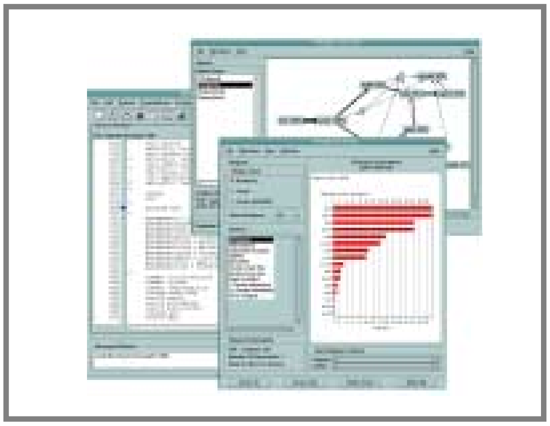


Figure 8: Calling Tree and 2D Profile

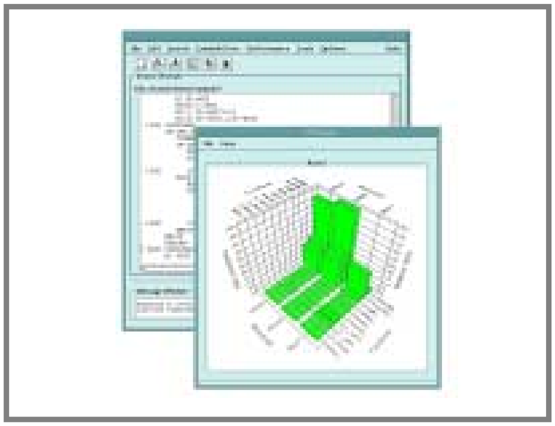


Figure 9: 3D Profile

Figures 8 and 9 show profiling displays in 2 and 3 dimensional graphics. These displays quickly point an applications tuner to the areas of code using the most execution time.

8. Networking

TCP/IP communications protocol is supported. HIPPI-800 has demonstrated specifically high performance and efficiency in local area networks. Other options include ATM through adapters and FDDI. Ethernet through 1000 Base SX is available.

SUPER-UX supports the Distributed File System (DFS) as well as NFS versions 2 (32 bit file system)and 3 (64 bit file system).

9 SX-5 Family Models Overview

9.1 SX-5 Series Capability Models

SX-5 Series capability models span a performance range from the 8 GFLOPS SX-5/1 in a "D" cabinet to the largest Multi Node configuration, the SX-5/512M. Specification for SX-5 Series Single Node Supercomputers is shown in Table 1. Table 2 shows representative SX-5M Multi Node models.

Table 1: SX-5 Series Model Overview

	SX-5 Series Cabinet Configurations			
Cabinet	**D**	**C**	**B**	**A**
Peak Performance	10 - 20 GF	20 - 40 GF	40 - 80 GF	80 - 160 GF
CPUs	1 / 2	2 / 3 / 4	4 to 8	8 to 16
CPU Type	10 GF	10 GF	10 GF	10 GF
Memory				
Capacity	8 - 32 GB	16 - 64 GB	32 - 128 GB	64 - 256 GB
Max Bandwidth	80 - 160 GB/s	160 - 320 GB/s	320 - 640 GB/s	640-1280 GB/s
IOP (max channels)	14 x 100 MB/s	30 x 100 MB/s	62 x 100 MB/s	126 x 100 MB/s

9.1.1 SX-5 Series Cabinets

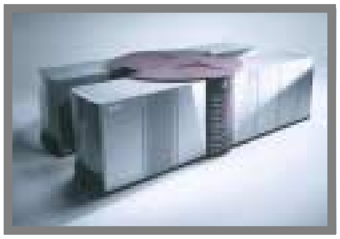


Figure 10: A Cabinet

SX-5 cabinets are available to enable right-sized configurations with appropriate expandability. There are no other major cabinets associated with the mainframe. Because the systems are 100% air cooled there are no external chilling units except for the requisite room air conditioning systems. No motor generating equipment is required, although a UPS system is recommended.

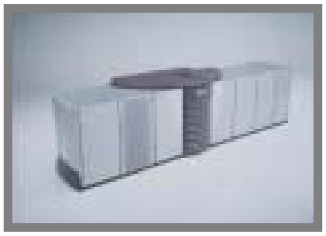


Figure 11: B Cabinet

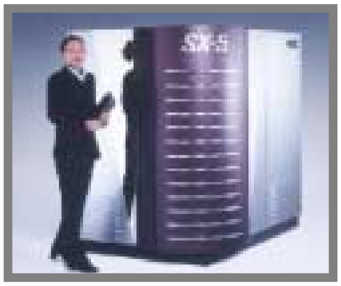


Figure 12: C Cabinet

The various SX-5 Series cabinets are shown in figures 10 through 13. All model cabinets are designed for installation in a computer room environment. A and B cabinets require raised flooring. C and D cabinets do not require raised floors if the room air conditioning is sufficient.

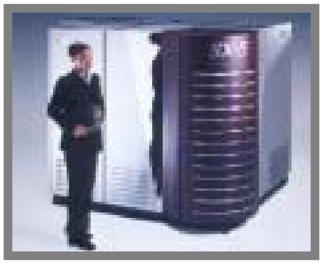


Figure 13: D Cabinet

9.1.2 SX-5M Multi Node Configurations

Multi Node models, which provide up to 5 TeraFLOPS of performance from 512 processors, are constructed using a proprietary high-speed crossbar linking multiple A or B cabinet Single Node systems (architecture shown in figure 2). The high speed internode cabling provides 8 gigabytes per second bi-directional transfers and the crossbar, called the IXS, supports a 256 gigabytes per second bisection bandwidth for the maximum 32 nodes.

A and B cabinet SX-5 Series systems can be equipped with an IXS to make a Multi Node configuration. Table 2 includes specifications for representative SX-5M Multi Node models. The first number after the slash denotes total processor count. The number after the M denotes the number of system cabinets. Multi Node configurations can be configured in various ways with no requirement that each node be identical.

	SX-5/32M2	SX-5/64M4	SX-5/128M8	SX-5/256M16	SX-5/512M32
Processors-Total					
CPUs	32	64	128	256	512
Peak GFLOPS	320	640	1,280	2,560	5,120
Memory-Total					
Capacity	512 GB	1 TB	2 TB	4 TB	8 TB
Bandwidth	2.56 TB/sec	5.12 TB/sec	10.24 TB/sec	20.48 TB/sec	40.96 TB/sec
IOP-Total					
Bandwidth	25 GB/sec	50 GB/sec	100 GB/sec	200 GB/sec	400 GB/sec
IXS					
Bandwidth	16 GB/sec	32 GB/sec	64 GB/sec	128 GB/sec	256 GB/sec

Table 2: Representative SX-5M Multi Node System Specifications

9.2 SX-5e Capacity Models

SX-5e models span a performance range from the 4 GFLOPS SX-5S/1 High Performance Server to the Largest Multi Node "e" configuration, the SX-5/512Me. Although the SX-5S and SX-5e are logically a continuum in a performance range, and have the same processing characteristics, they are divided into two products because of differences in their I/O configurability. Specifications for SX-5e Single Node Supercomputers are shown in Table 3. Table 5 shows the SX-5S High Performance Server models, and Table 4 shows representative SX-5Me Multi Node models.

Table 3: SX-5e Model Overview

SX-5e Model Cabinet Configurations		
Cabinet	**Ce**	**Be**
Peak Performance	16 - 32 GF	32 - 64 GF
CPUs	4 - 8	8 - 16
CPU Type	4 GF	4 GF
Memory		
Capacity	32 - 64 GB	64 - 128 GB
Max Bandwidth	128 - 256 GB/s	256 - 512 GB/s
IOP (max channels)	30 x 100 MB/s	62 x 100 MB/s

Figure 15: Ce Cabinet

9.2.1 SX-5e Single Node Cabinets

Figures 14 and 15 show the “e” type cabinets. No external cabinets are required except for the room air conditioning units.

Both model cabinets are designed for installation in a computer room. The Be cabinet requires raised flooring. The Ce cabinet does not require raised flooring if the room air conditioning is sufficient.

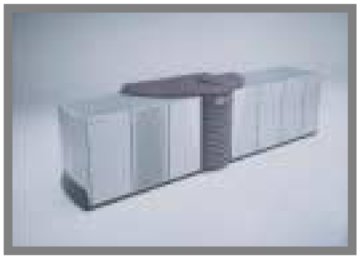


Figure 14: Be Cabinet

9.2.3 SX-5Me Multi Node Configurations

SX-5e Multi Node models, which provide up to 2 teraFLOPS of performance from 512 processors, are constructed using an NEC proprietary high-speed crossbar linking multiple SX-5 Be Single Node cabinets (architecture shown in figure 2). The high speed internode cabling provides 8 gigabytes per second bi-directional transfers and the crossbar, called the IXS, supports a 256 gigabytes per second bisection bandwidth for the maximum 32 nodes.

Only SX-5e systems with a Be cabinet can be equipped with an IXS to make a Multi Node configuration. Table 4 includes specifications of representative SX-5Me Multi Node models. The first number after the slash in the model designation denotes the number of processors. The number after the Me denotes the number of cabinets. The Multi Node configuration is flexible in that all nodes do not have to be identical, so equivalent performance levels can be attained by numerous specific configurations.

	SX-5/32Me2	SX-5/64Me4	SX-5/128Me8	SX-5/256Me16	SX-5/512Me32
Processors-Total					
CPUs	32	64	128	256	512
Peak GFLOPS	128	256	512	1,024	2,048
Memory-Total					
Capacity	256 GB	512 GB	1 TB	2 TB	4 TB
Bandwidth	1 TB/sec	2 TB/sec	4 TB/sec	8 TB/sec	16 TB/sec
IOP-Total					
Bandwidth	12.5 GB/sec	25 GB/sec	50 GB/sec	100 GB/sec	200 GB/sec
IXS					
Bandwidth	16 GB/sec	32 GB/sec	64 GB/sec	128 GB/sec	256 GB/sec

Table 4: SX-5Me Representative Systems Specifications

9.3 SX-5S High Performance Servers

SX-5S High Performance Server models provide cost effective supercomputing performance to the department level or smaller enterprise. Two cabinet types are available, the E cabinet which can hold up to 4 SX-5e 4 GFLOPS processors and 16 Gbytes of main memory, and the F cabinet which can hold up to 2 SX-5e 4 GFLOPS processors and 8 Gbytes of main memory.

Specifications for SX-5S High Performance Servers are shown in Table 5. The F cabinet is shown in Figure 16. The E cabinet is 430 mm wider and 615 mm deeper but otherwise identical.

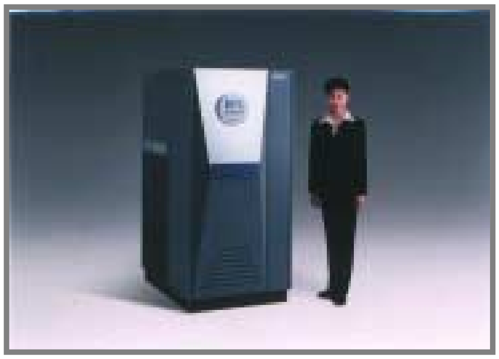


Figure 16: SX-5S HPC Server

	SX-5S/1	SX-5S/2	SX-5S/4
Cabinet	F	E - F	E
CPUs	1	2	4
System Peak	4 GF	8 GF	16 GF
Memory			
Max. Capacity	16 GB	16-32 GB	32 GB
Max Bandwidth	32 GB/sec	64 GB/sec	128 GB/sec
IOP (max channels)	6 x 100 MB/s	6-14 x 100 MB/sec	14 x 100 MB/sec

Table 5: SX-5S High Performance Server Model Overview

For Further Information on the SX-5 Series contact:

Cray Inc.
411 1st Avenue South Suite 600
Seattle WA 98104
206-701-2000

email: crayinfo@cray.com

Visit us at www.cray.com on the World Wide Web.

Specifications for all SX-5 Series models are subject to change without notice.
Environmental information contained herein is approximate and should not be used for planning purposes.