



SiCortex

DATASHEET

SC5832



- Up to 8.2 Tflops on a single back-plane
- Up to 5832 900 milliwatt processors matched to memory pipeline
- Up to 8 terabytes main memory
- Microsecond MPI ping-pong
- 20 kilowatts single cabinet
- Straight Linux: just type "make"

SC5832

The SC5832 Hardware Environment

The SiCortex SC5832 offers a complete 8.2 TFlops (peak) computing system packaged in a single cabinet. The cabinet holds 12, 20, 28 or 36 circuit modules. Each circuit module (CM) accommodates 27 SiCortex node chips and associated memory. There are two memory sockets per node, which can be configured with:

- 4 GByte per node, for a total of 4 TBytes on 36 modules

- 8 GByte per node, for a total of 8 TBytes on 36 modules

Each circuit module accommodates up to three (3) PCIe modules, each operating at a peak speed of 2 GBytes/sec, plus 2 Gigabit Ethernet ports.

Each node contains all of the switch fabric circuits necessary to support interprocessor communications. Switch fabric circuits connect directly to the high performance backplane interconnect. The Kautz graph topology of the backplane assures that each node can get to every other node in a minimal number of steps.

The SiCortex hardware platform also includes a System Service Processor (SSP) with its own integrated control network for configuration and monitoring as well as for centralized storage and distribution of configuration parameters and operating system images. The SiCortex nodes contain no disk or non-volatile storage, thereby reducing the issues of "node personalities."

The SiCortex power system distributes 48 Volt DC power to the circuit modules from a redundant and hot swappable central supply. This approach is more reliable and energy efficient than the conventional method of distributing AC power to every node.

The SiCortex Software Environment

The SiCortex software environment provides a powerful and fully integrated set of tools to support applications, combining maximum user-level performance with minimum system management overhead. All components of the environment have been tested together and pre-loaded into each SiCortex system prior to shipment, eliminating a major source of user integration effort and potential error.

Linux On SiCortex

The SiCortex system is rooted in the Linux 2.6 kernel, from linux-mips.org, with more recently-released updates ported as needed. Changes to the kernel are minimal: they enable features such as native support of the Lustre file system (see below), application performance monitoring, and extremely fast MPI. The standard Linux EDAC framework has been extended to reach the SiCortex hardware's pervasive use of ECC, which includes the L1 and L2 caches.

The SiCortex system includes many standard Linux utilities, such as bash, CVS, Emacs, GNU make, openSSH, tcsh, and vim. Within the parameters established by the system manager, each SiCortex node is fully capable of running as an interactive timesharing system. End users access the system by launching jobs from their workstations via the scheduler, by logging into a time-shared partition on the system, or by logging into the head node of a partition.

MPI and Networking

The Message Passing Interface (MPI) is the most common programming interface for writing parallel applications in Linux. The SiCortex implementation is based on MPICH-2 and supports all MPI-1 operations and most MPI-2 operations, including MPI I/O.

SiCortex MPI is optimized to provide extremely fast communication at all levels. Small message ping-pong is approximately one microsecond. The fabric can support large message bandwidth of 6 GBytes/sec between nodes. Large message bandwidth simultaneously into all nodes in a communicator exceeds 600 MBytes/sec. Even with a maximum number of processes, applications startup time is typically measured in seconds.

In addition to using MPI, all nodes can use standard TCP/IP over the fabric.

The microcode and hardware primitives that support MPI communication are also an excellent foundation for other methods. For example, the fabric hardware enforces reliable, in-order message delivery that benefits the performance of TCP sockets as much as it benefits MPI.

For programs that use multithreading facilities such as pthreads or openMP, each SiCortex node is a cache-coherent SMP.

Compilers and Development Tools

All SiCortex systems are delivered with the SiCortex Pathscale compiler suite, which includes compilers for C, C++, and Fortran 77/90/95. The Pathscale tool chain is fully interoperable with libraries and object files compiled by gcc. Both native and cross-compile versions are included.

SiCortex Linux includes gcc for C and C++. These tools function in both native and cross-development forms, and generate code for either 32 or 64 bits. Support for gmake, glibc (the GNU C library), binutils (the GNU binary utilities) and gdb (the GNU debugger) is included.

SiCortex Linux also includes basic installations of perl, python, tcl, tk, and other utilities.

Performance Tools and Debuggers

Since SiCortex systems are focused on sustained application performance, the software suite includes an exceptionally powerful array of open source and commercial analysis tools. They include:

- PAPI and papiex to provide a consistent interface to the rich array of hardware performance registers.
- perfmon and pfmon, the kernel and user-level infrastructure for performance analysis tools.
- MPIP, the summary tool for MPI performance.
- HPCToolkit, tools to gather statistical profiles based on hardware performance registers.
- oprofile, the system-wide profiler.
- TAU for tracing and performance analysis.

Additional well-known tools are available commercially.

- Vampir-NG for visualizing performance data.
- TotalView for parallel debugging on SiCortex systems.

Math and Science Libraries

The SiCortex libraries include many specific high-performance packages for common functions such as linear system solvers, FFTs, and grid partitioning. These packages extend and optimize the base libm (part of glibc) library that ships with all systems.

I/O and File Systems

The standard SiCortex I/O interface is PCIeexpress, and also includes gigabit Ethernet. The systems support specific third-party modules for Ethernet, FibreChannel, and Infiniband.

NFS (version 3) and Lustre may be used for networked file systems. Typically, Lustre will provide higher performance and simplified management of a large-scale file system. In addition, ramfs and tmpfs are available for in-memory file systems.

SiCortex Linux includes standard Linux server facilities for NFS version 3. Most nodes use only the client code. I/O nodes may use the server code if they have storage systems connected via PCIeexpress.

The base cluster file system in SiCortex Linux is Lustre. Every node can be a Lustre applications client if so configured. I/O nodes (those with direct PCIeexpress connection) serve as Lustre servers for the other nodes. A Lustre system may have up to 512 OSTs (Object Storage Targets), each supporting four terabytes for a total storage capacity of two petabytes.

System Management

The SiCortex system includes a System Service Processor (SSP) which is the focal point for user account management, system management, activity monitoring, environmental monitoring, job scheduling, logging, and booting. It is also the central storage and distribution point for operating system images and communications parameters. The SSP provides a convenient unified view for managing and monitoring the entire system. It is visible to system administrators but not to users.

SLURM is the job scheduler for SiCortex systems.



SiCortex, the first company to engineer a cluster computer from the silicon up, is dedicated to the proliferation of multi-teraflops computing to a wide variety of users. The SiCortex approach represents a sea change in cluster computer design, where reduced energy consumption and increased performance are synergistic, not antagonistic.

sicortex.com

Specifications

SIZE: 72" H x 60" W x 60" D

CLEARANCE: left and right, 18"; front and rear, 36"; above 22"

WEIGHT: 1500 lb

HEAT: 70,300 BTU/hr

VOLTAGE: 208VAC, three-phase (3P+E);
230/400VAC, three-phase (3P+N+E)

CURRENT: 64A (208V); 34A (400V)

FREQUENCY: 50-60Hz

POWER: 20,600W

OPERATING TEMPERATURE: 10-35C (50-95F), less 1C/300m above sea level (1.8F/1000') to a maximum of 3000m (10,000')

RELATIVE HUMIDITY: 20-80%, non-condensing

NON-OPERATING TEMPERATURE: -40-70C (-40-158F)

- The SC5832 is classified as information technology equipment (ITE). It carries the CSA, CE, FCC, and VCCI marks.
- The SC5832 is compliant with CSA 22.2 60950-1 and EN 60950-1.
- The SC5832 is a Class A computing device compliant with FCC Part 15 Subpart B, EN 55022, EN 55024, ICES-003, VCCI, CNS 13438, and AS/NZS 3548.

Revised: 07/24/2008



SiCortex

SICORTEX INC. THREE CLOCK TOWER PLACE, SUITE 210, MAYNARD, MASSACHUSETTS 01754