WHITE PAPER

SGI Prism™ XL
The Open Accelerator Platform for Delivering a Scalable Petaflop per Cabinet

November 2010

Executive Summary

The SGI Prism XL, SGI's latest high-end platform, is designed to deliver up to a petaflop per cabinet of raw compute power. It is built on an infrastructure to leverage 100s, up to 1000s, of high-bandwidth PCIe x 16 slots, in an accelerator-agnostic architecture, that is supported by an open software stack for management and development. This paper is intended to describe the high-level technical capabilities of the product to customers who might use or manage the system.

Accelerators Offer Superior Flops/Dollar/Watt for Many Applications

We have occasionally seen Top 10 supercomputing sites based upon the Top 500 list (www.top500.org) utilize accelerators to achieve their top rankings. For example the Tsubame system used ClearSpeed processors to vault itself to the number seven spot in the June, 2006 list. With the wider adoption of GPU acceleration technology, motivation from GPU vendors, and greater efficiencies achieved on GPU code we expect this trend to not only continue, but to accelerate. Indeed, for some applications GPU and other accelerators provide the best flops/dollar/watt versus typical x86 microprocessors.

The Challenge Confronting Buyers of Accelerator Systems

The use of accelerators for high-performance computing (HPC) has grown from experimental and hobbyist to production-level over the past couple of years. A handful of small systems in a departmental lab or office environment are being expanded into whole data centers full of accelerator-augmented servers that are taking on significant parts of a organizations's workflow in application spaces such as seismic processing, image processing, and rendering. This expansion leads to the challenges seen in many large HPC datacenters of system and data management as well as power and cooling provisioning.

At the same time the budding success of these platforms have led to many more options in the marketplace as well, from accelerators based on GPU technology (such as NVIDIA's Tesla series) to System on a Chip (SoC) options from Tilera. Many of these options have their own infrastructure options (such as the NVIDIA "S" boxes) that lead to infrastructure 'sprawl' as an organization experiments with different options or deploys several options covering different workflows in its environment. The options have different interfaces, different management schemes, making having more than one or two types a challenge.

SGI Prism XL as an Accelerator-Agnostic Platform

Over the years SGI has deployed a number of accelerator-augmented systems including its Origin® line with the Tensor Processing Unit (TPU) and its Altix® line with FPGA-based RASC™ Technology. These earlier accelerator systems were highly proprietary in nature and with limited flexibility to adopt both newer and alternate accelerator technologies.

With SGI Prism XL, though, SGI introduces a platform with flexibility being one of its core values. Given that PCIe bandwidth and features have continued a steady increase over the past few years, and have an interesting roadmap going forward, we designed our latest accelerator platform with PCIe as the interface to the accelerator card.

SGI Prism XL uses the PCIe gen2 bus to access the accelerator, leading to greater flexibility among possible accelerators and future-proofing the platform for new accelerators becoming available.

Best-of-Breed Platform for 100s to 1000s of PCIe Slots

SGI Prism XL's innovative platform leverages an optimized PCIe infrastructure and is uniquely designed starting with the PCI slot and then crafted with enough I/O and memory for a complete system. This sets the platform apart from other accelerator and GPU-capable systems that are designed first with a two-socket server motherboard in mind, with one or more PCIe slots populated with GPU's, or those that use an external chassis plugged into one or more standard rackmount servers.

SGI Prism XL is designed for customers that want to commit to an accelerator-focused architecture for either research or are already in production using accelerator technologies. The motherboard functions almost solely as a high-speed path to off-accelerator memory, I/O and networking.

STIX™ Architecture

The SGI Prism XL is designed with SGI's new innovative STIX Architecture, with the core architectural element being the stick. A stick is a rectangular element, 5.78 inches wide, 3.34 inches high and 37.125 inches deep, with a maximum weight of 21 pounds.

Part of the simplicity of the STIX Architecture is that the stick is a self-contained element, with its own cooling and power solution. It can be mounted in various enclosures in racks, pulled out, put on a tabletop and plugged into a local power outlet (it has an auto-sensing power supply) for local development work, demonstrations or trouble-shooting.

Anatomy of a Stick

Each SGI Prism XL stick contains two slices that are identical in configuration, mirrors of each other. Each slice consists of a full-length, full-height PCIe gen 2 x 16 slot and a single-socket AMD 4100-powered motherboard. Each slice has up to two 2.5 inch SATA drives allowing up to 4 Terabytes of storage on a single stick.

The stick includes a 1 kW power supply that easily accommodates today's 200W to 225W cards but has the capability of powering up to 300W cards which are on the roadmaps of some accelerator vendors.

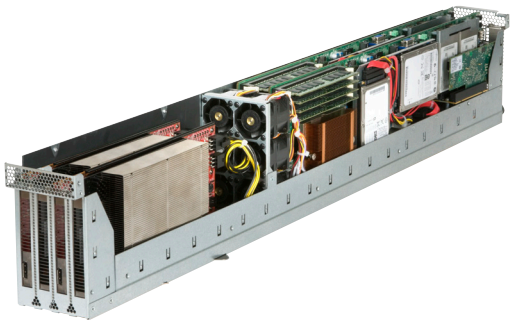

Figure 1: An SGI Prism XL stick showing the two slices, cut longitudinally down the middle. Rear of the stick is, in this case, at the lower left of the photo.

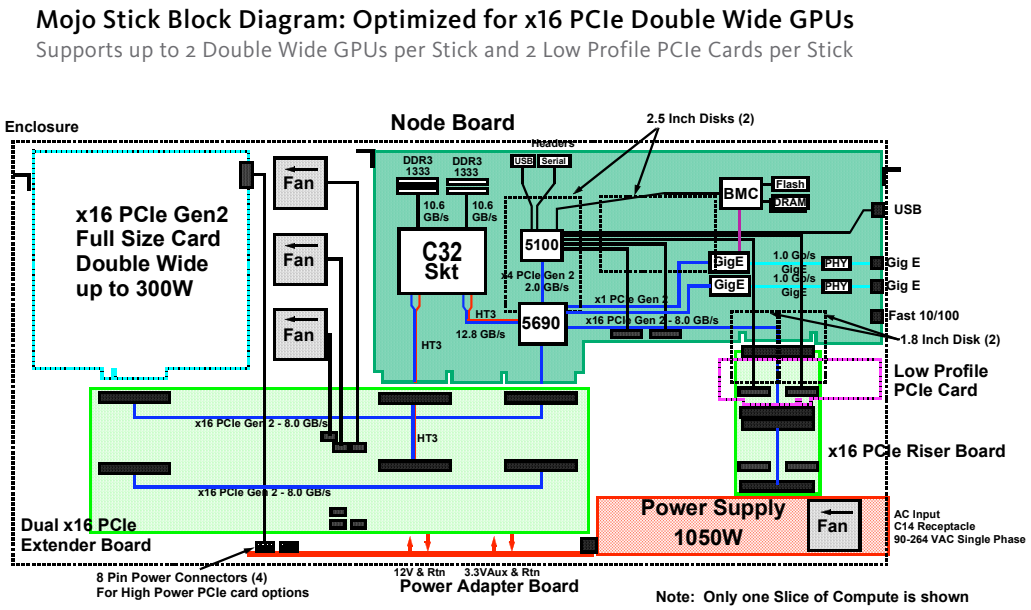

Figure 2: Block Diagram of a slice.

The front of the stick contains the two PCIe gen 2 x 16 slots with accelerators installed. The middle of the stick contains the two motherboards, and the rear of the stick contains networking and I/O options.

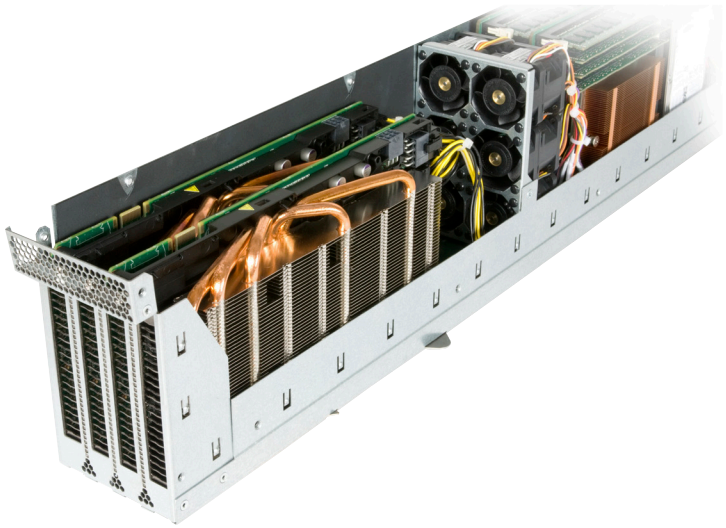

Figure 3: Showing detail of accelerator cards, in the case of this stick, two NVIDIA M2050 Tesla-20 cards, with Fermi-based accelerators.

Accelerator Options

SGI Prism XL will support the latest accelerator cards from NVIDIA, AMD, and Tilera.

- For optimal passive cooling, ECC support and leading double-precision Floating Point processing capability the NVIDIA M2050 and M2070 are both supported.
- For maximum single-precision Floating Point processing capability, the AMD Firestream 9370 is supported.
- For low-wattage applications Tilera TILEncore TLB-26400-7-PCIe-2X10-4-GC and TLB-36400-7-PCIe-2S10-2S1-4-GC will be supported.

SGI Prism XL also has the capability of adding other accelerators and technologies based upon customer demand and accelerator roadmaps. As mentioned previously, the platform can support up to a 300W PCIe card, future-proofing for the top accelerators on the market in the future.

Networking and I/O Options

Each SGI Prism XL slice contains two GigE ports, two USB ports, and a PCIe gen 2 x 8 slot for networking and I/O connectivity. By default an SGI Prism XL machine comes with a fat-tree GigE network setup among its slices, including the switches and cables required. Supported additional networking options include a Mellanox HBA with QDR Infiniband capability that can be plugged into the PCIe x 8 slot.

Either a single- or dual-port card can be selected, leading to either a single-plane or dual-plane fat-tree IB network, with the required Mellanox switches and cables.

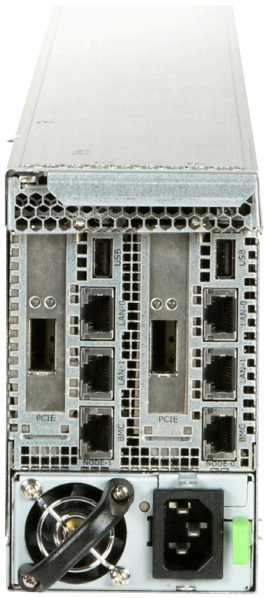

Figure 4: Front of stick showing Ethernet, USB and IB connections. This stick is configured with single-port Infiniband HBAs.

SGI Prism XL System Management

The SGI Prism XL system is configured like a typical HPC cluster, with each slice of a stick being a node in the cluster. As mentioned previously, by default the system has a fat-tree GigE network, with optional Infiniband networks added.

A four-socket capable AMD Opteron 6100 head-node is the brains of the cluster and where all the management software resides. The head-node can have either two or four sockets populated with processors and up to 32 DIMM slots populated with DIMM's. To populate all the DIMM slots requires populating all the processor sockets.

SGI Management Center (SMC) is the management software of choice, and this software allows the administrator to administrator not only the SGI Prism XL system but also other SGI systems in the environment whether Rackable™, Altix UV or Altix ICE systems. A single GUI can be used to manage all these types of SGI systems in an environment.

At the board level a Board Management Controller (BMC) chip gathers all relevant data and transports it across the GigE network to the head-node where it can be monitored and acted upon via SMC.

SGI Prism XL Software Development

Software development on the SGI Prism XL is aided by SGI's Accelerator Execution Environment™ (AEE). AEE is an accelerator-specific package of drivers, the SDK, and tools that SGI has bundled together for customer convenience and installed on the platform. There's an AEE version for each supported accelerator in the platform: for example AEE-N-1 is the first package for NVIDIA GPU accelerators, the Tesla-20 series.

AEE will support CUDA, OpenCL, and MDE programmers, depending upon the accelerator being used. In addition to AEE, the following development packages will support SGI Prism XL program development: Allinea DDT, TotalView Debugger, Portland Group PGI Accelerator, and CAPS Enterprise HMPP.

SGI Prism XL for Exascale Computing

For scientists and engineers willing to dive head-first into accelerator-driven computing, the SGI Prism XL provides an ideal platform to manage and deploy the variety of accelerators likely to appear over the coming years as exascale computing becomes the goal of research over the next decade. A petaflop-sized system used to require hundreds of racks and a huge datacenter. With SGI Prism XL able to deploy a petaflop into a single cabinet, 10s or 100s of petaflops, out to an exaflop, becomes an easy target.

Corporate Office
46600 Landing Parkway
Fremont, CA 94538
tel 510.933.8300
fax 408.321.0293
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.912.7500
Asia Pacific +61 2.9448.1463