WHITE PAPER

SGI® UV™ for SAP HANA®

Scale-up, Single-node Architecture Enables Real-time Operations at Extreme Scale and Lower TCO

TABLE OF CONTENTS

1.0 Introduction

SGI® UV™ for SAP HANA® is a purpose-built, in-memory computing appliance for large or growing environments running on SAP HANA. Developed by SGI—the trusted leader in high performance computing—the system combines Intel® XEON® E7 processors with SGI NUMAlink® ASIC technology. Currently SAP-certified as a four-, eight-, twelve*-, or sixteen*- socket system, SGI's new appliance for SAP HANA® is designed to scale to 32 sockets and 24TB of shared memory as a single node. This paper describes the future-ready, modular architecture of SGI UV for SAP HANA that enables enterprises to achieve real-time operations at extreme scale and lower cost of ownership.

2.0 SGI UV for SAP HANA

Building on SGI's proven in-memory computing technology and unique scale-up architecture, SGI UV for SAP HANA enables large enterprises to confidently leverage the power of SAP HANA for mission-critical applications and heavy, multi-engine analytics that require single-node systems, and to reduce overhead and raise service levels for cluster-supported environments that become a struggle. Using a scale-up single-node architecture with breakthrough coherent shared memory, SGI UV for SAP HANA enables running SAP Business Suite and other SAP applications in large enterprises. The system easily combines OLTP and OLAP workloads and eliminates the time-consuming extract, transform, and load (ETL) process to generate real-time reports on demand. Users can perform very complex joins at massive scale and run multiple analytic engines simultaneously to include text, geo-spatial, and live data streaming. Consolidation of applications and infrastructure can be achieved, while also eliminating costly silos that impede productivity.

As shown in Figure 1, the single-node architecture of SGI UV for SAP HANA allows enterprises to run applications free from the complexity and overhead of clustered appliances. There are no cluster nodes, cluster network, or storage area network to configure and administer. In addition, there is no need for database partitioning or re-balancing I/O when increasing the size of the SGI appliance, as performance scales near linearly and automatically.

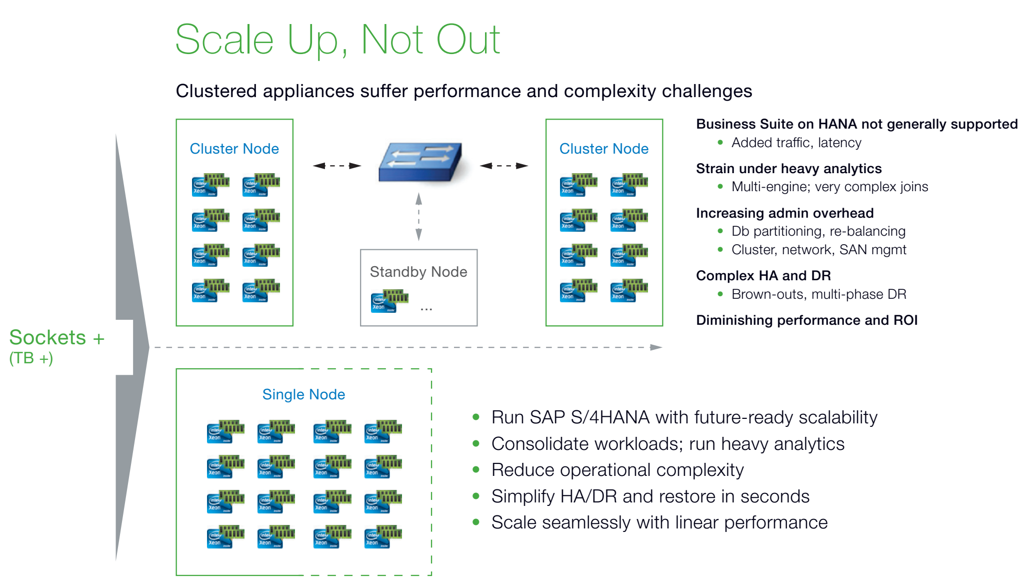

Figure 1. By simply adding sockets and memory, SGI's single-node architecture scales seamlessly.

**under controlled availability for >8 sockets and >6TBs*

3.0 Architectural Overview

The SGI UV for SAP HANA appliance features the SGI UV 300H, an advanced symmetric multiprocessing (SMP) system designed to scale from four to 32 sockets and up to 24TB of cache-coherent shared memory as a single node. The modular chassis architecture of the SGI UV 300H enables users to grow the single-node system in four-socket increments without adding complexity. The chassis are interconnected using 7th Generation SGI NUMAlink 7 (NL7) ASIC technology and an All-to-All topology, delivering extreme network bandwidth with ultra-low latency. The appliance also features:

- Intel® Xeon® E7 processors for high density memory
- SAP HANA® on SUSE® Linux® Enterprise Server for SAP Applications
- NetApp® E2700 RAID for data and log backup
- Enterprise-class reliability, availability, and serviceability (RAS)

The appliance is delivered pre-racked, pre-configured, and pre-tested in a standard 19-inch 42U rack. A single Rack Management Controller (RMC) is provided per system.

3.1 SGI UV 300H

The basic building block for the appliance is the SGI UV 300H, a model of the SGI UV 300 supercomputer specifically designed for SAP HANA and part of the SGI UV server line for in-memory computing. The SGI UV 300H features a 5U modular chassis that hosts four processors with up to 120 threads and integrated NUMAlink ASICs to interconnect modules, processors, and shared memory. By combining additional chassis (up to eight per standard 19-inch rack), SGI UV 300H is designed to scale up to 32 sockets and 960 threads (with hyper threading enabled). All of the interconnected chassis operate as a single system running under a single operating system instance.

Each SGI UV 300H modular chassis features the following:

- Four Intel Xeon E7 8890 v2 15-core processors internally connected in a ring by Intel Quick Path Interconnect (QPI) links
- Eight memory risers, each with two SGI Jordan Creek ASICs and support for 12 DDR3 memory DIMMs
- Up to 3TB of memory per chassis using 32GB DIMMs
- Two SGI HARP ASICs to connect the processors to the SGI NUMAlink 7 network fabric
- One BaseIO card (one per system)
- Four 2.5-inch SSD drives (four-socket system only)
- Support for a maximum of eight full-height, 6/7-length (10.5-inch maximum length) Gen3 x8 PCIe slots
- Support for a maximum of four full-height, double-wide, 6/7-length (10.5-inch maximum length) Gen3 x16 PCIe slots
- Four 1600 watt power supplies
- Eight 80mm x 38mm cooling fans
- Two 36mm x 28mm cooling fans in each power supply (four per chassis)
- Rackmountable 19-inch form factor

3.2 Intel Xeon E7 Processors

Intel Xeon E7 8890 v2 processors utilize Intel® QuickPath Interconnect 1.1 Technology (QPI) to provide high-speed point-to-point connections between the four processors within the SGI UV 300H chassis. Key features of the Intel E7 processor include:

- 15 processor cores per socket
- Three full-width Intel QPI links per processor (maximum transfer rate is 8.0GT/s, aggregate bandwidth of 25.6GB/s per QPI link)
- Hyper-threaded cores, with two threads per core
- 64-bit computing with support for 48-bit virtual addressing and 46-bit physical addressing
- 32KB Level-1 instruction cache with single bit error correction; 32KB Level-1 data cache with error correction on data and detection on TAG
- 256KB of Level-2 instruction/data cache, ECC protected (SECDED)
- 37.5MB instruction/data last level cache (LLC), ECC protected (Double Bit Error Correction, Triple bit Error Detection (DECTED), and SECDEC on TAG
- Up to 2.5MB per core instruction/data LLC, shared among all cores
- 32 lanes of PCIe 3.0
- DDR3 memory

For more information about the Intel Xeon processor E7 v2 product family, visit:
www.intel.com/content/www/us/en/processors/xeon/xeon-e7-v2-family-details.html

3.3 SAP HANA on SUSE Linux Enterprise Server for SAP Applications

To accelerate time to value, the appliance arrives pre-configured with a single instance of SAP HANA running on SUSE Linux Enterprise Server for SAP Applications. SAP provides licensing for SAP HANA and complete first-line appliance support. To learn more about how to combine database, data processing, and application platform capabilities in-memory with the SAP HANA platform and run your business in real-time, visit hana.sap.com/abouthana.html. To learn why SUSE Linux Enterprise is the leading platform for SAP solutions on Linux and choice for SAP HANA, visit www.suse.com/products/sles-for-sap/

3.4 NetApp E2700 RAID

To protect against power loss to the volatile DDR3 memory, log files and data are also written synchronously to persistent storage using NetApp® E2700 RAID arrays. For each SGI UV 300H chassis, an E2700 array equipped with 21.6TB of SAS storage (900GB SAS drives x 24) is directly attached via 6Gb SAS. The RAID storage has one or more XFS file systems written across it. Logical Volume Manager (LVM) with multipathing provides volume management and RAID 6 is utilized to protect against drive failure. To learn more about the streamlined performance of NetApp E2700 RAID, visit www.netapp.com/us/products/storage-systems/e2700/index.aspx

3.5 Rack Management Controller

The Rack Management Controller provides the top layer of system control for the SGI UV 300H system. This controller is a standalone 1U rack-mount chassis. Through the use of an internal 24-port Ethernet switch, a single Rack Management Controller can provide system control for a 32-socket (8-chassis) SGI UV 300H system configured in one or two racks.

3.6 Custom-Designed Rack

The custom-designed 42U rack can hold up to five SGI UV 300H chassis, five NetApp E2700 RAID arrays, and the Rack Management Controller. The rack is designed to support both air or water-assisted cooling. For larger systems, up to eight SGI UV 300H chassis can be configured in a single rack, with a second rack utilized for NetApp E2700 RAID storage. Remaining rack space can be utilized for other 19-inch rack-mount equipment. SGI UV for SAP HANA appliances are pre-configured and pre-racked at the SGI factory for standard deployments. Organizations can also elect to have SGI UV for SAP HANA installed on-site by SGI support engineers in existing standard 19-inch racks.

4.0 Scale-Up Design with SGI NUMAlink 7

The inherent single-node scalability and performance of SGI UV for SAP HANA is facilitated by integrated SGI NUMAlink 7 interconnect technology.

4.1 Four to 32-Socket Scalability

Using advanced SGI HARP ASICs contained in each SGI UV 300H chassis and 7th-generation SGI NUMAlink 7 network interconnects, SGI UV for SAP HANA is designed to scale up as a single-node server by simply adding more chassis. For each SGI UV 300H chassis, a NetApp E2700 RAID array is added to the appliance to provide greater backup capacity. Only one Rack Management Controller is needed no matter how large the system grows. As shown in Figure 2, SGI's future-ready architecture is designed to scale up to to 32 sockets and 24TB of shared memory in four socket increments. As of this writing, SAP has certified four- and eight- socket configurations under general availability and twelve-and sixteen-socket configurations under controlled availability.

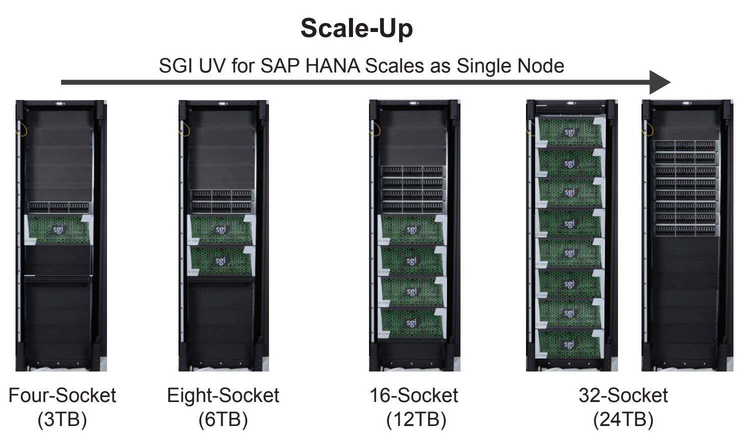

Figure 2. SGI UV for SAP HANA can easily scale from four-socket to 32-socket configurations in increments of four sockets.

4.2 SGI HARP-Based Motherboard

Innovative SGI HARP ASICs are the heart of the SGI UV 300H chassis (Figure 3). These links connect the processors across multiple chassis to form an extreme bandwidth, ultra-low latency SGI NUMAlink 7 (NL7) network fabric and a single-node system. Each SGI UV 300H has two SGI HARP ASICS, each with two QPI channels that connect to two of the four processors within the chassis. The SGI HARP ASIC exposes 16 four-lane NL7 channels. A HARP interface board double connects the two HARP ASICs, leaving 14 links per ASIC that are exposed at the bulkhead for cabling to HARP ASICs in other chassis. Each link has a peak bi-directional transfer rate of 14GT/s and 7.47GB/s to provide extreme throughput for large volume HANA databases and hosted applications.

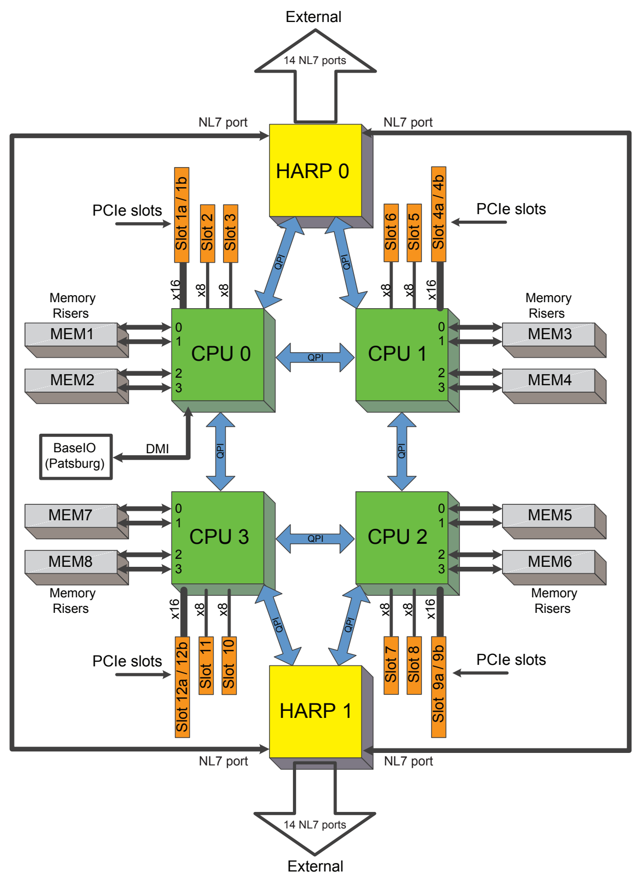

Figure 3. SGI HARP ASICs connect multiple chassis into a single system while Intel QuickPath Interconnects connect processors within a chassis.

As shown in the figure above, Intel QPI links provide communication between the four Intel processors within each SGI UV 300H chassis. The links connect the processor sockets in a ring, resulting in a maximum of two QPI hops for a processor socket to communicate with the other three processor sockets in the chassis. Intel QPI features include:

- Cache coherency
- Fast/narrow uni-directional links and concurrent bi-directional traffic
- Error detection via CRC with error correction via link level retry
- Packet based protocol

Intel QPI also has extensive RAS features, including:

- Self-healing via link width reduction
- Link-level retry mechanism
- 8-bit CRC or 16-bit rolling CRC
- Error reporting mechanisms including data poisoning indication and viral bit
- Support for lane reversal as well as polarity reversal at the QPI links
- High-bandwidth ECC protected crossbar router with route-through capability

4.3 ccNUMA Memory Architecture

Memory is physically distributed both within and among the SGI UV 300H chassis, and is accessible to and shared by all processors connected to the NUMAlink fabric. SGI NUMAlink 7 provides memory cache coherency, referred to as ccNUMA.

Non-uniform Memory Access (NUMA)

In distributed shared memory systems, memory is physically located at various distances from the processors. As a result, memory access times (latencies) are different, or non-uniform. For example, it takes less time for a processor to reference its locally-installed memory than to reference remote memory. The total memory within the NUMAlink fabric is referred to as global memory, but a number of different memory sub-types are present within SGI UV for SAP HANA:

- **Local memory.** If a processor accesses memory that is directly connected to a processor socket, the memory is referred to as local memory.
- **Off-processor socket memory**. Memory managed by another socket but local to the chassis has a maximum of two QPI hops.
- **Remote memory**. If processors access memory located in other chassis, the memory is referred to as remote memory. This path could have a maximum of two QPI hops and one NL7 hop.

Cache Coherency

SGI UV 300H uses caches to reduce memory latency. Although data exists in local or remote memory, copies of the data can exist in various processor caches throughout the system. Cache coherency keeps the cached copies consistent. To accomplish this feat, ccNUMA technology uses a directory-based coherence protocol in which each 64-byte block of memory has an entry in a table (directory). Like the blocks of memory that they represent, the directories are distributed among the chassis. A block of memory is also referred to as a cache line.

Each directory entry indicates the state of the memory block that it represents. For example, when the block is not cached, it is in an 'unowned' state. When only one processor has a copy of the memory block, it is in an 'exclusive' state, and when more than one processor has a copy of the block, it is in a shared state. A bit vector indicates which caches may contain a copy. When a processor modifies a block of data, the processors that have the same block of data in their caches are notified of the modification. In general, SGI UV systems use an invalidation method to maintain cache coherence. The invalidation method flushes all cache copies of the block of data, and the processor that wants to modify the block receives exclusive ownership of the block.

4.4 All-to-All NUMAlink 7 Topology

SGI UV for SAP HANA features an All-to-All network topology in which all SGI HARP ASICS are connected to all other HARP ASICs. The topology is based on the NL7 high-speed interconnect channel and industry standard cables. The All-to-All topology scales from one to eight chassis in one-chassis increments with a maximum latency of under 500ns. Figure 4 illustrates the All-to-All topology of a full 32-socket system. In the illustration, all NL7 ports are occupied in each SGI UV 300H chassis, with red lines representing internal connections. The system as depicted features eight SGI UV 300H chassis and 112 NL7 cables.

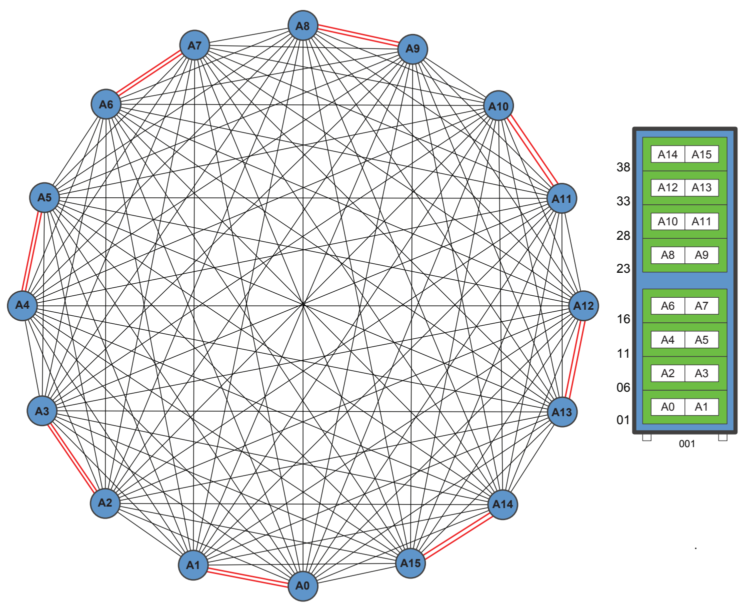

Figure 2. All-to-All topology for an eight chassis single-node system.

4.5 Adaptive Routing

NUMAlink 7 also provides adaptive routing around congested networks and failed links to achieve high bandwidth and low latency across the appliance. As a means for determining network congestion, the HARP ASIC monitors traffic on its NL7 links and knows which links have the highest amount of use. It also monitors how long a packet has waited to be sent. There is a primary path and up to three secondary paths for routing packets between any two chassis.

The primary path is the shortest path, representing the lowest number of hops. The secondary paths can have more hops and therefore more latency. Using both the primary and secondary paths increases the total available bandwidth between the two nodes. Prior to sending a packet, the HARP ASIC selects the best path for the packet to take based on the following criteria:

- Shortest path
- Path with the least congestion
- The length of time the packet waits to be sent (also known as wait time)

5.0 Enterprise-Class Reliability, Availability, and Serviceability

Like the Intel Xeon E7 processors and NetApp E2700 RAID arrays used in the appliance, SGI UV 300H also brings extensive RAS features including memlog, hot pluggable redundant components, and overall system design.

SGI's memlog utility helps overcome errors on memory DIMMs that can lead to application performance issues and unplanned downtime. Corrected memory errors are logged and analyzed. If a DIMM page is deemed defective, an attempt is made to transparently relocate data to a new page and retire the old page, enabling applications to continue running without interruption. Administrators are also alerted to failing DIMMs so that they can be replaced during planned maintenance windows.

Component redundancy includes N+1 hot pluggable fans and N+N or N+1 hot pluggable power supplies with online fault detection. All components are serviced from the front of the chassis for easy access.

The SGI UV 300H system leverages 20 years of SGI in-memory computing expertise. To deliver reliable, stable, scalable systems, SGI has invested heavily in meticulous engineering practices. These include design practices for interconnect controller hardware, high speed interconnects, high-speed printed circuit board (PCB) design, and platform software development. For a detailed look at SGI UV 300H RAS, please see the ***SGI® UV™ 300** System Reliability, Availability and Serviceability* white paper.

6.0 Conclusion

The ability to combine database, data processing, and application platform capabilities in-memory with the SAP HANA platform is truly game-changing. Imagine if your operations, financial, research, or marketing teams could operate in real time. Now imagine leveraging SAP HANA in your enterprise at extreme scale and lower total cost of ownership (TCO). SGI and the future-ready architecture of SGI UV for SAP HANA makes this possible.

7.0 About SGI

SGI is a global leader in high performance solutions for compute, data analytics and data management that enable customers to accelerate time to discovery, innovation, and profitability. Visit sgi.com for more information.

Global Sales and Support: **sgi.com/global**