WHITE PAPER

SGI® UV™ 300RL for Oracle® Database In-Memory

Single-system Architecture Enables Real-time Business at Near Limitless Scale with Mission-critical Reliability

TABLE OF CONTENTS

1.0 Introduction

SGI® UV™ 300RL is an advanced, in-memory computing system for large or growing Oracle Database In-Memory environments. Developed by SGI—the trusted leader in high performance computing—the system combines Intel® Xeon® E7 processors with SGI NUMAlink® ASIC technology. Certified with Oracle Linux, UV 300RL scales from 4 sockets up to 32 sockets, and from 1TB up 24TB of cache-coherent shared memory as a single system. This paper describes the future-ready, modular architecture of SGI UV 300RL that enables enterprises to accelerate analytics and achieve real-time business at near limitless scale with mission-critical reliability.

2.0 SGI In-Memory Computing for Oracle Database In-Memory

Building on SGI's globally proven in-memory computing technology and a unique future-ready architecture, SGI enables large and growing enterprises to run Oracle Database In-Memory on a single Intel-based system with unparalleled scale-up capacity, and seamless scale-up simplicity.

SGI UV 300RL allows enterprises to implement the innovative 'dual-format' architecture of Oracle Database In-Memory at greater scale to identify business trends in seconds, make faster, smarter decisions, and gain competitive advantage. You can speed analytics by orders of magnitude. You can accelerate mixed workload OLTP. And you can run transactional and analytic workloads concurrently, eliminating the expense and time-delay of ETL processes, to get real-time answers on demand.

As shown in Figure 1, the modular single-system architecture of SGI UV 300RL enables you to grow your Oracle Database In-Memory environment without adding overhead. There are no cluster nodes or cluster network to configure and administer. No additional software or licensing is required. And there's no need for data segmentation or re-balancing I/O when increasing the system's size, as performance scales near linearly and automatically.

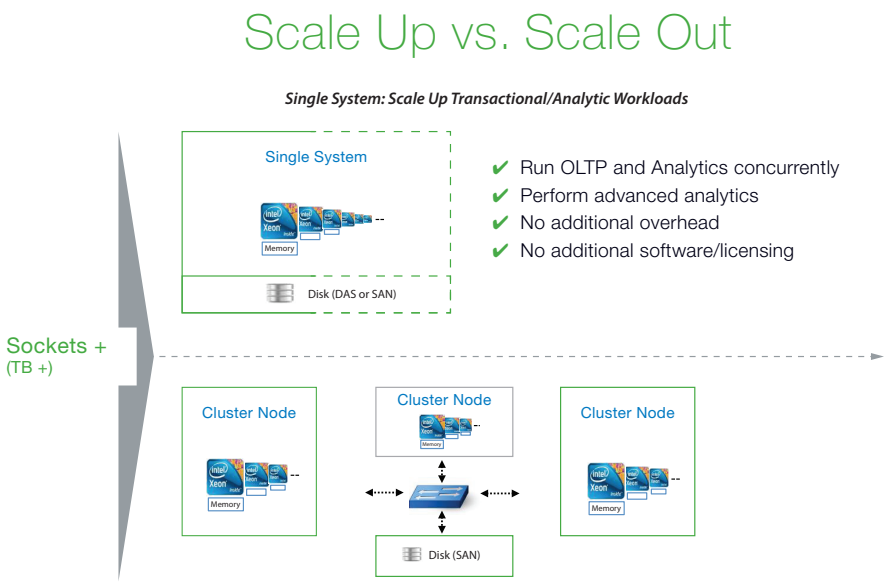

Figure 1. By simply adding sockets and memory, SGI's single-system architecture scales seamlessly.

3.0 Architectural Overview

SGI UV 300RL is an advanced symmetric multiprocessing (SMP) system designed to scale from four to 32 sockets and 1 to 24TB of cache-coherent shared memory as a single system. The modular chassis architecture of the UV 300RL enables users to grow the single 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 system also features:

- Intel® Xeon® E7 processors for high density memory
- Oracle Linux 7 with Unbreakable Enterprise Kernel (UEK)
- Rack Management Controller (RMC; one per system)
- Enterprise-class reliability, availability, and serviceability (RAS)

The system is built for mission-critical applications, delivering four nines (99.99%) system uptime. To provide maximum data protection and Five Nines availability, you can full leverage the Oracle Maximum Availability Architecture, such as RMAN for Backup, Oracle Application Cluster (RAC) for High Availability, and Oracle DataGuard for Disaster Recovery. And multiple PCIe Gen3 slots provide fast I/O to your existing enterprise storage or SGI's world-class storage offerings.

3.1 SGI UV 300RL

Part of the SGI UV server line for in-memory computing, SGI UV 300RL is a model of the SGI UV 300 supercomputer specifically designed for Oracle Database In-Memory. Users can also leverage the UV 300RL to run Oracle 11g. The system features a 5U modular chassis that hosts four processors with up to 144 threads and integrated NUMAlink ASICs to interconnect modules, processors, and shared memory. By combining additional chassis (up to eight per standard 19-inch rack), UV 300RL is designed to scale up to 32 sockets and 1,152 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 300RL modular chassis features the following:

- Four Intel® Xeon® E7 8890 v3 processors (4-, 10-, 16- or 18-core; 2.3 - 3.2GHz), 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 DDR4 memory DIMMs (choice of 8GB, 16GB, 32GB, or 64GB up 2133MT/s)
- Up to 6TB of memory per chassis
- Two SGI HARP ASICs to connect the processors to the SGI NUMAlink 7 network fabric
- One BaseIO card (one per system)
- 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 v3 processors utilize Intel® QuickPath Interconnect Technology (QPI) to provide high-speed point-to-point connections between the four processors within the SGI UV 300RL chassis. Key features of the Intel® E7 processor include:

- 4, 10, 16, or 18 processor cores per socket
- Three full-width Intel® QPI links per processor delivering a maximum transfer rate of 9.6GT/s per 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)
- 45MB 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
- DDR4 memory

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

3.3 Oracle Linux with UEK

To fully optimize and scale Oracle Database In-Memory, the SGI UV 300RL is certified and factory-installed with Oracle Linux 7 with Unbreakable Enterprise Kernel (UEK). To learn more about the Oracle Linux operating system and the latest innovations, tools, and features, visit http://www.oracle.com/us/technologies/linux/overview/index.html

3.4 Rack Management Controller

The Rack Management Controller provides the top layer of system control for the SGI UV 300RL 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 4- to 32-socket (1 to 8 chassis) single system.

3.5 Custom-Designed Rack

The system's custom-designed 42U rack can hold up to eight UV 300RL chassis and the Rack Management Controller. The rack is designed to support both air or water-assisted cooling. Remaining rack space can be utilized for other 19-inch rack-mount equipment. SGI UV 300RL systems are pre-configured and pre-racked at the SGI factory per type of deployment. Organizations can also elect to have UV 300RL installed on-site by SGI support engineers in existing standard 19-inch racks.

3.6 Storage Flexibility

SGI UV 300RL is equipped with industry-standard PCIe Gen3 expansion slots (up to 12 per chassis) provide optimum flexibility for persistent storage with fast I/O. Connect directly or by network to your existing enterprise-class 3rd party storage or select from the entire SGI InfiniteStorage line.

- UV 300RL is pre-configured with Oracle Linux with UEK and ready for installation of Oracle 12c with the Oracle Database In-Memory option (or standalone Oracle 12c or Oracle 11g installation)
- UV 300RL is pre-racked and pre-tested prior to shipment and installed by an SGI services engineer
- SGI provides support for the UV 300RL. Oracle provides support for Oracle Linux and Oracle Database In-Memory. Other solution components (e.g., storage and storage network) are supportive by respective vendors.
- To learn about enterprise-class SGI InfiniteStorage offerings, visit http://www.sgi.com/products/storage/

4.0 Scale-Up Design with SGI NUMAlink 7

The inherent single-system scalability and performance of SGI UV 300RL is facilitated by integrated SGI NUMAlink 7 interconnect technology.

4.1 Four to 32-Socket Scalability

Using advanced SGI HARP ASICs contained in each chassis and 7th-generation SGI NUMAlink 7 network interconnects, UV 300RL is designed to scale up as a single-node server by simply adding more chassis. 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.

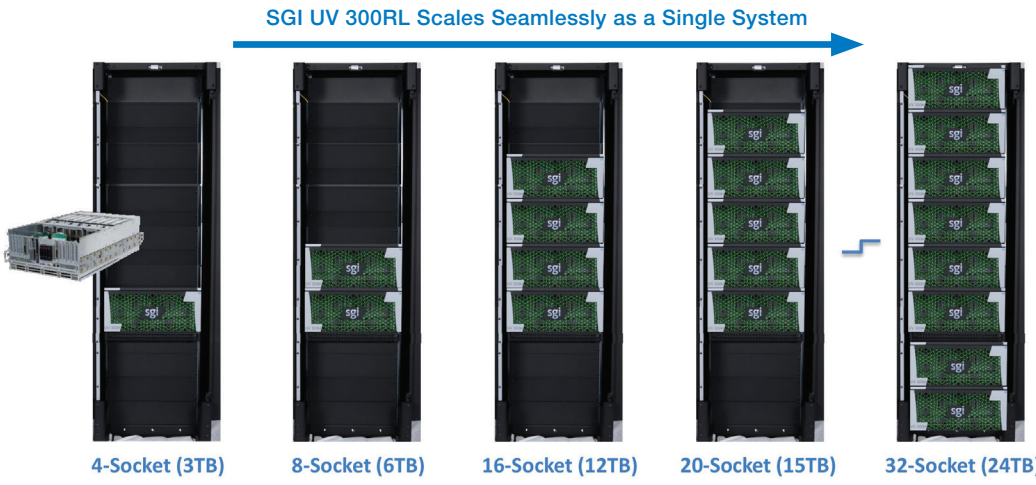

Figure 2. SGI UV 300RL scales up seamlessly in four-socket increments

4.2 SGI HARP-Based Motherboard

Innovative SGI HARP ASICs are the heart of the SGI UV 300RL 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 system. Each UV 300RL 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 14.94GB/s to provide extreme throughput for large volume Oracle databases and hosted applications.

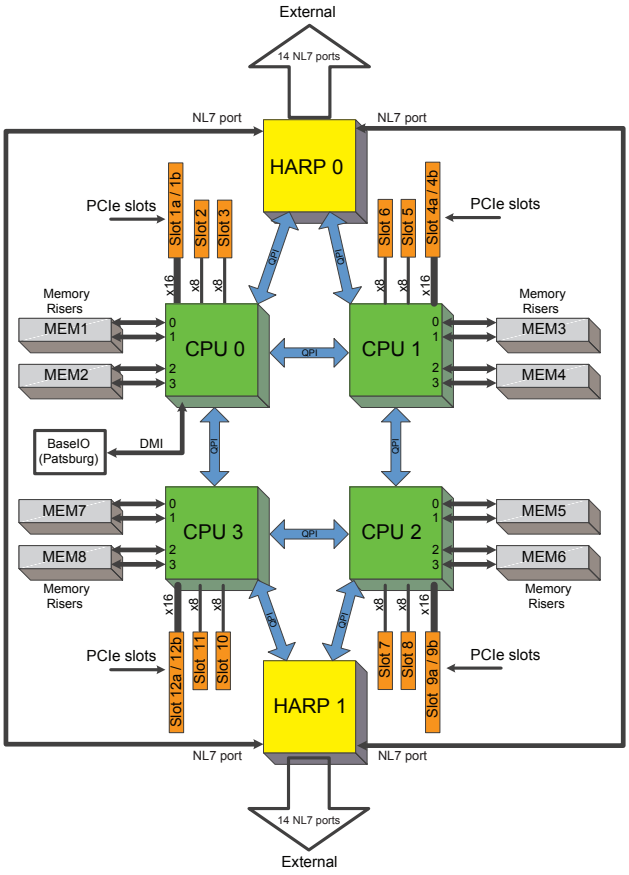

Figure 3. SGI HARP ASICs connect multiple chassis into a single system while Intel® QuickPath Interconnects connect processors within a chassis.

As shown in the figure above, Intel® QPI links provide communication between the four Intel® processors within each SGI UV 300RL 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 300RL 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 300RL:

- **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 300RL 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 300RL 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 round trip latency (any processor reading data located on any DIMM) 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 UV 300RL chassis, with red lines representing internal connections. The system as depicted features eight UV 300RL chassis and 112 NL7 cables.

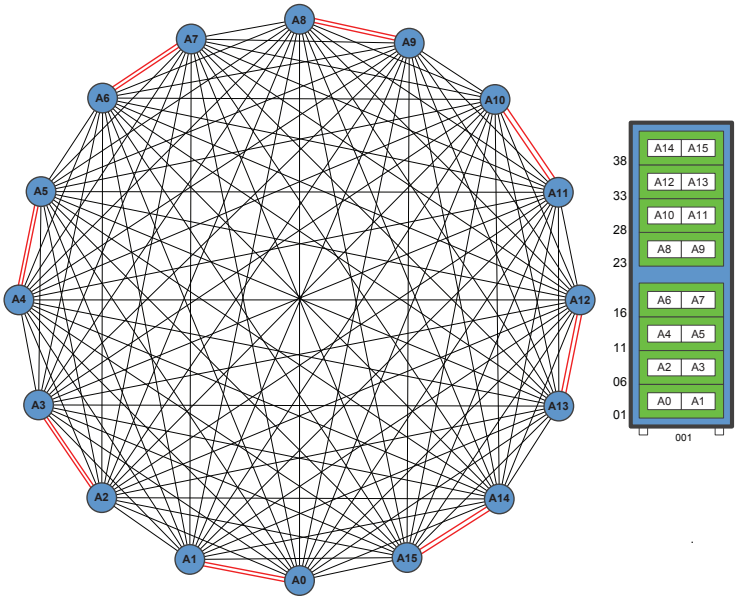

Figure 2. All-to-All topology for an eight chassis single system.

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 system. 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 used in the system, SGI UV 300RL also has 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. Chassis are equipped with slides to enable easy access to components from the front and rear at time of service.

The SGI UV 300RL 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 300RL RAS, please see the ***SGI® UV™ 300 System Reliability, Availability and Serviceability*** white paper.

6.0 Conclusion

The ability to accelerate Oracle 12c to the speed of memory with the Oracle Database In-Memory option is game changing. Imagine if your financial, manufacturing, and marketing teams could access vital information with up-to-the-moment accuracy, get answers instantaneously, and operate in real time. Now imagine leveraging Oracle Database In-Memory across your enterprise with near limitless scale and mission-critical reliability. SGI and the future-ready architecture of SGI UV 300RL make 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