WHITE PAPER

sgi®

Performance-per-Watt Best Practices on Intel® processor-based SGI® Rackable™ Servers

April 2011

Performance-per-Watt Best Practices on Intel® Processor-based SGI® Rackable™ Servers

The term performance-per-watt is a measure of the energy efficiency of a computer architecture or a computer hardware. It can be represented as the rate of transactions or computations or a certain performance score that can be delivered by a computer for every watt of power consumed. While performance-per-watt is useful, absolute power measurements are also important. Some newer generation of server architectures may provide better performance-per-watt, but continued performance increases can negate the gains in power efficiency. Also, benchmarks that measure power under heavy load may not adequately reflect typical efficiency. Power draw of some components like power supplies, video cards, etc are temperature dependent, so temperature dependence should also be taken into account while measuring energy efficiency of a system. To help address these various challenges, a standard benchmark like SPECpower®, has been designed to measure power at a series of load levels which also takes into account temperature dependency. This paper refers to SPECpower benchmarks executed on SGI® Rackable™ half-depth C2005-TY6 server and SGI® Rackable™ 2U standard-depth C2112-4TY14 server to demonstrate their high energy efficiency and to derive best practices for maintaining such efficiency for continued reduction of operational costs in data centers.

1. Need for Energy Efficiency Benchmarks

With an increasing demand of power rating of servers to comply with the data center environments, energy-smart designs for servers are being implemented industry-wide. Apart from measuring direct consumption of electric power for an application workload, it is important to consider additional draw of energy for cooling and for any auxiliary equipment that is temperature dependent and is required to run the overall system in a standard data center. Also, energy efficiency figures from varied sources may not be comparable due to differences in workload, configuration, test environment, etc. Factors like choice of processor SKUs and their power saving technologies, memory type, size and number, power supplies, system motherboard design and BIOS settings for power management, number of fans, fan size and fan speed control and other features like RAID on board, number of HBAs, etc have an effect on the overall power consumption. So an industry-standard benchmark is required so it can provide reliable power and performance metrics considering all factors starting from system infrastructure to power consumption at different application load levels and can mimic an ideal data center environment. The benchmarking exercise leads IT Management to derive certain facts for a server class as follows:

- Realize system hardware optimization features to minimize power consumption.
- Select the processor type and system configuration for performance, keeping power consumption in mind. Once determined, optimize both performance and power to achieve the peak performance-per-watt for a server class.
- Assist users with decisions who want to run similar workloads with the choice of power savings with minimal performance tradeoffs.
- Perform comparative analysis of application workloads and servers – Idle power vs. application power draw; which applications consume least or max power in what mode and why.
- Realize thread placement and memory placement policies on the CPU sockets for a balanced workload, optimal performance and minimal power consumption.
- Understand server configurations that provide maximum power savings over others. Get vendor-competitive metrics with std. benchmarks.
- Realize reasons of higher power consumption for specific configurations and enhance the next generation of servers.

2. The SPECpower® Benchmark

SPECpower (SPECpower_ssj®2008) is an industry-standard benchmark that provides a means to measure power (at the AC input) in conjunction with a performance metric. This helps IT managers to consider power characteristics along with other selection criteria to increase the efficiency of data centers. The benchmark exercises the CPUs, caches, memory hierarchy and the scalability of shared memory processors (SMPs) as well as the implementations of the JVM (Java Virtual Machine), JIT (Just-In-Time) compiler, garbage collection, threads and some aspects of the operating system. The benchmark runs on a wide variety of operating systems and hardware architectures and should not require extensive client or storage infrastructure.

The benchmark workload is a Java application on a number of warehouses or threads/cores. It generates and completes a mix of transactions, and the throughput is the number of transactions completed over a fixed period.

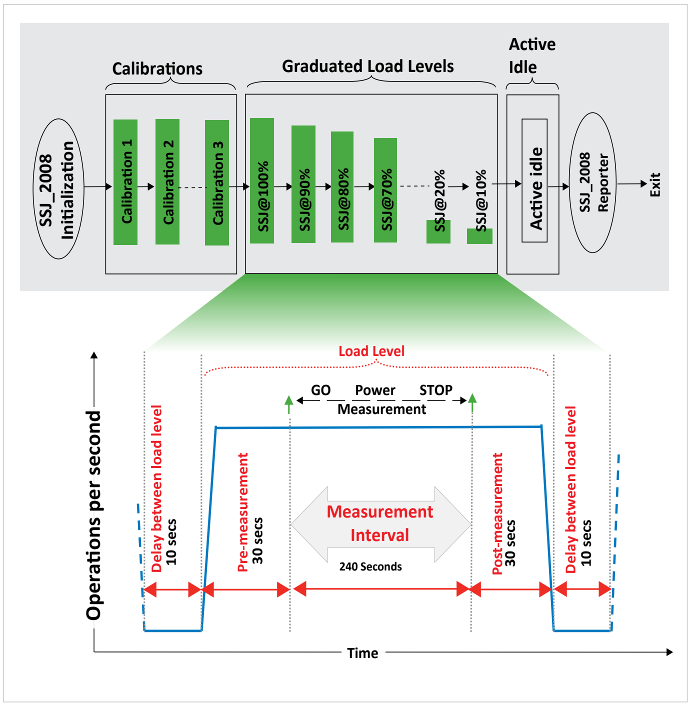


Figure 1: Graduated Load levels in a SPECpower_ssj2008 benchmark

Benchmark Phases involve the Calibration phase and the Graduated load phase. The Calibration phase of the benchmark determines the target throughput at 100% CPU load level. In the Graduated load phase, each load level is a 240 second "measurement interval" plus a ramp up "pre-measurement" interval and a ramp down "post-measurement" interval along with a "delay between load level", to ensure settle time and proper synchronization between the load levels for consistent power and performance measurement. The average power for each distinct measurement interval must be reported. This reporting is important because different customer environments have different interests with regard to system utilization. A pool of single-purpose servers may operate at very low utilization most of the time, where Idle and 10% throughput intervals are most interesting. A large system designed to consolidate through virtualization could well run near the 80% interval for much of the time and at idle for some of the time. For compute intensive environments, the target may be to use the system at 100% capacity almost all of the time. The throughput and power measurements should be reported for every measurement interval, so that users can interpret any collective metric and can draw specific conclusions at utilization points that are of interest to them.

The Performance-per-Watt metric for the benchmark is computed, as follows:

1. The total performance metric for each of the distinct measurement intervals is computed and these totals are summed;
2. The average power measured for each benchmark interval including the Active-Idle measurement is added together;
3. The quotient of the summed performance metric and the summed power measurements is reported as the Performance-per-Watt value.

As temperature has a substantial affect on power characteristics of a system, it is measured on a regular basis throughout the benchmark. The minimum acceptable temperature is also restricted to an appropriate level. A minimum of 20 degrees Celsius (68 degrees Fahrenheit) is a good starting point. A minimum higher than 20 degrees could be specified, such as 23 or 25 degrees Celsius (73-77 degrees Fahrenheit), although this could be problematic, as many data centers are maintained in the 22-23 degrees Celsius range.

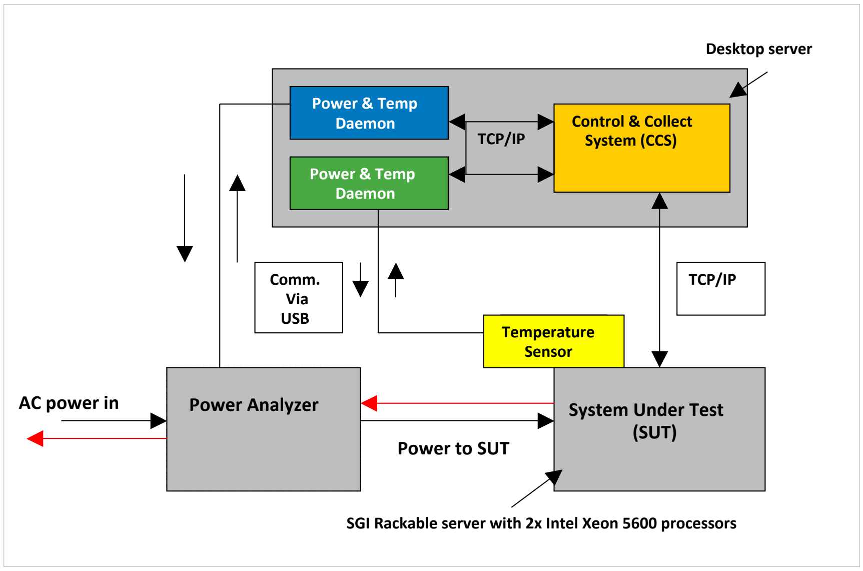


Figure 2: SPECpower_ssj2008 Benchmark Flow Diagram

Figure 2 shows the basic design of SPECpower benchmark workflow where it explains how the benchmark is configured and run. The Control and Collect System (CCS) controls the workload in the System under Test (SUT) and collects the data for the power consumed and the temperature changes during the workload run, as defined in SPEC script[1]. For information on benchmark setup and recommended devices please refer to the SPECpower Benchmark Setup Guide[2].

3. SPECpower Benchmarks on SGI® Rackable™ Servers

This section describes the hardware and software configurations used to run the SPECpower_ssj2008 benchmarks on Intel® Xeon® 5600 processor based SGI® Rackable™ half-depth C2005-TY6[3] and 2U standard-depth C2112-4TY14[4] server platforms, as well as the benchmark results.

[1] http://www.spec.org/power/docs/SPECpower-Power_and_Performance_Methodology.pdf
[2] http://www.spec.org/power/docs/SPECpower-Measurement_Setup_Guide.pdf.

3.1. Benchmark Configurations and Tuning

The benchmark configuration comprised of:

Servers:

- 1x SGI® Rackable™ half-depth C2005-TY6 server with
  - 2x Intel® Xeon® 5670 processors 2.93 GHz, 12x 2GB 1.35/1.5V DIMMS 1333 MHz, HT enabled;
  - 1 x SLC 32GB Intel SSD drive;
  - 1 x 250 W Power Supply;
- 1x SGI® Rackable™ 2U standard-depth C2112-4TY14 server with 4-nodes, each node comprised of
  - 2x Intel® Xeon® L5640 2.27 GHz processors, 4x 4GB DIMMS 1333 MHz, HT enabled;
  - 1 x 64GB 2.5" OCZ SSD drive;
  - 1 x 1400W Supermicro Shared Power Supply

OS:

- Microsoft Windows Server 2008 R2 Enterprise Edition 64-bit version

JVM Version and Options:

- SGI® Rackable™ C2005-TY6 server:
  - IBM J9 VM (build 2.4, JRE 1.6.0 IBM J9 2.4 Windows Server 2008 amd64-64 jvmwa6460sr7-20091214_049398 (JIT enabled, AOT enabled);
  - Each JVM instance was affined to four logical processors on a socket;
  - -Xms1500m -Xmx1500m -Xmn1100m -Xaggressive -Xcompressedrefs -Xgcpolicy:gencon -XlockReservation -Xnoloa -XtlhPrefetch –Xlp
- **SGI® Rackable™ C2114-4TY14 system, per node:**
  - IBM J9 VM (build 2.4, JRE 1.6.0
  - IBM J9 2.4 Windows Server 2008 amd64-64 jvmwa6460sr7-20091214_49398 (JIT enabled, AOT enabled);
  - Each JVM instance was affined to 4 threads on a single socket;
  - -Xmn1600m -Xms2000m -Xmx2000m -Xaggressive -Xcompressedrefs -Xgcpolicy:gencon -XlockReservation -Xnoloa -XtlhPrefetch –Xlp

Controller System HW/SW:

- Dell Precision 590 Intel Xeon CPU 5140@2.33 GHZ, 8GB memory
- Windows Server 2003 EE CCS 1.2.4
- Oracle JRockit(R) (build R27.5.0-110-94909-1.6.0_03-20080204-1558-windows-ia32, compiled mode)

Measurement Devices:

- Power Analyzer: Yokogawa Electric Corporation WT210;
- Temperature Sensor: Digi International, Inc, Watchport/H Model

Tuning:

- Display was turned off after 1 minute
- Run was started via Remote Desktop
- Enabled "Power Saver" power scheme
- Each JVM affined to four logical processors on a socket (on C2005-TY6 server and per-node of C2112-4TY14 server);
- Using the local security settings console, "lock pages in memory" was enabled for the user running the benchmark.

BIOS Settings:

- SGI® Rackable™ C2005-TY6 server:
  - Hardware Prefetcher Disabled
  - Adjacent Cache Line Prefetcher Disabled
  - QPI Speed set 4.8GT/s

[3]http://www.sgi.com/pdfs/4193.pdf[2]http://www.bitmover.com/lmbench/
[4]http://www.sgi.com/pdfs/4193.pdf

  - DCU Prefetcher Disabled
  - DCU Data Prefetcher and DCU Instruction Prefetcher Disabled
  - MLC Streamer and MLC spatial Prefetcher Disabled
  - Intel Turbo Boost Technology Disabled
- SGI® Rackable™ C2112-4TY14 system, per node:
  - Hardware Prefetcher Disabled
  - DCU Prefetcher Disabled
  - Data Reuse Optimization Disabled
  - Intel Virtualization Tech Disabled
  - Execute Disable bit Technology Disabled
  - Turbo Mode Disabled
  - C3 state Disabled
  - Serial ports Disabled
  - USB 2.0 Controller Mode Full Speed
  - Active State Power Management Enabled
  - Memory Frequency set to 1066MHz
  - QPI Frequency set 4.8GT/s
  - QPI L0s and L1s Enabled
  - Fan Mode - Energy Saving / ES

3.2. Power Management Features on SGI® Rackable™ C2005-TY6 and C2112-4TY14 Server Platforms

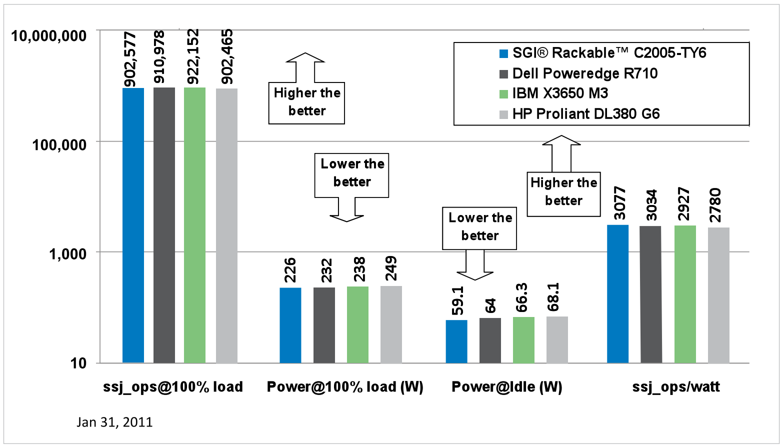
- The SGI® Rackable™ C2005-TY6 and C2112-4TY14 servers have Flash-enabled technology that reduces overall power consumption up to a factor of 8 compared to 2.5-inch and 3.5-inch HDDs;
- SGI® Rackable™ C2005-TY6 and C2112-4TY14 servers support low-voltage processors from Intel which run at the same clock rates as their higher-voltage counterparts but consumes less energy;
- On SGI® Rackable™ C2005-TY6 and C2112-4TY14 servers, the CPUs and DIMMs support active power management that allows for significant power savings at idle and moderate loads, but can run at peak frequency when needed;
- P-states and C-states on SGI® Rackable™ C2005-TY6 and C2112-4TY14 servers allow the CPU frequency and power to dynamically change based on the workload;
- Unused memory channels can be dynamically placed into a low power state on SGI® Rackable™ C2005-TY6 and C2112-4TY14 servers;
- SGI® Rackable™ C2005-TY6 server has options of disabling unused serial, USB and Ethernet ports, whereas SGI® Rackable™ C2112-4TY14 has options of disabling unused serial and USB ports;
- The Intel-based SGI® Rackable™ C2005-TY6 server is also optimized at the motherboard level for performance-per-watt by incorporating spread-core processor/memory layout, high efficiency VRs, comprehensive temperature monitoring, and extensive characterization and tuning across hardware and software;
- SGI® Rackable™ C2112-TY14 server uses cooling zones or isolated zone cooling, also prevents overcooling by allowing particular fans to run faster in order to cool a specific zone that is heating up without having to run all the fans faster. Optimal placement of multiple fans also lends itself to cooling particular set of nodes;
- SGI® Rackable™ C2112-TY14 server provides various options to set the speed of the fans. (fan modes can be set to 'balanced', 'Energy Savings', 'Full Speed', or 'Performance).
- SGI® Rackable™ C2005-TY6 and C2112-4TY14 servers are supported by SGI Management Center which, on servers with Intel® Intelligent Power Node Manager, provides power and thermal monitoring and policy based power management.

3.3. Benchmark Results

Results as of the date this paper was written (Jan 31, 2011) show that a single SGI® Rackable™ C2005-TY6 half-depth server with 2x Intel® Xeon® 5670 processors 2.93 GHz (6 cores/12 threads), 12x 2GB 1.35/1.5V 1333 MHz memory DIMMs has a low power of only 59.1 watts@idle, the lowest power of 226 watts@100% load and highest performance-per-watt metric of 3,077 compared to competitive single server platforms based on Intel® Xeon® 5670 processors (Figure 3). One SGI® Rackable™ 2U standard-depth C2112-4TY14 server with 4-nodes, each node having 2x Intel® Xeon® L5640 processors 2.27 GHz (12 cores/24 threads), 4x 4GB 1.35/1.5V 1333MHz memory DIMMs has the 3rd lowest power of 255 watts@active idle, 3rd lowest power of 724 watts@100% load and the 3rd highest ssj_ops/watt of 2,788 compared to competitive 4-node servers based on Intel® Xeon® 5600 processors (Figure 4). Figures 5 and 6 show the energy efficiency of the servers at variant CPU load levels. The red bars represent the efficiency (throughput per watt) for each measurement interval. The blue line shows the average power requirement at each of the 11 measurement intervals, including the Active Idle measurement, where no throughput is shown. Figures 7 and 8 show the variation of power of the servers at variant CPU load levels.

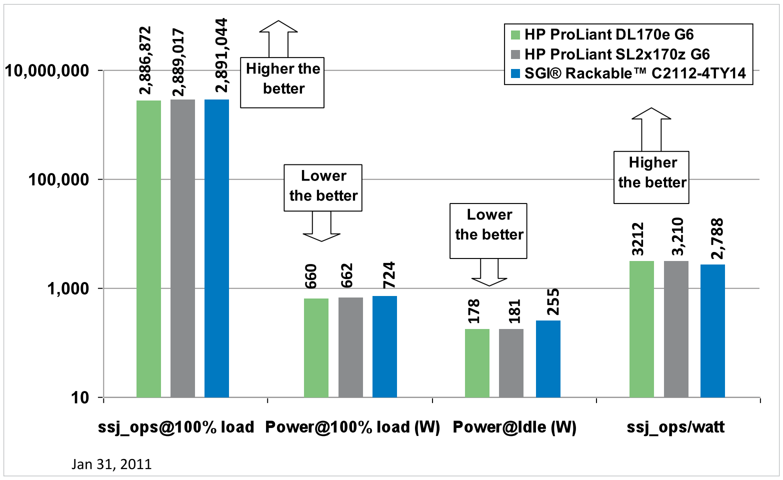
The efficient chassis, fan and power infrastructure of SGI® Rackable™ servers, low voltage but high performance Intel® processors, support for low voltage but fast bandwidth of memory, Flash technology and power management features - all contributed to high energy efficiency of the servers in their class. For full SGI® disclosures, please refer to:

http://www.spec.org/power_ssj2008/results/power_ssj2008.html



Jan 31, 2011

Figure 3: SPECpower®: 2X Intel Xeon 5670 processors



Jan 31, 2011

Figure 4: SPECpower®: ssj2008 4-node systems with Intel® Xeon® L5640 processors

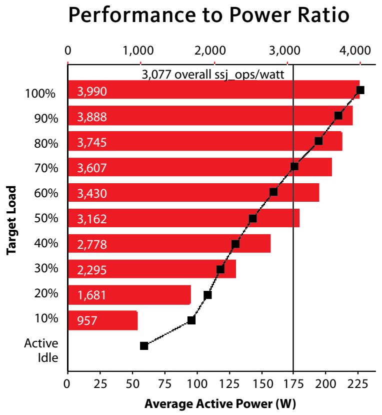


Figure 5: SPECpower_ssj2008 Results: Energy Efficiency at various CPU load levels SGI® Rackable™ C2005-TY6 server with Intel® Xeon® 5670 processors

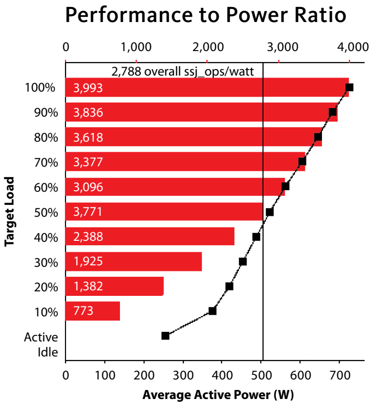


Figure 6: SPECpower_ssj2008 Results: Energy Efficiency at various CPU load levels SGI® Rackable™ C2112-4TY14 server with Intel® Xeon® L5640 processors

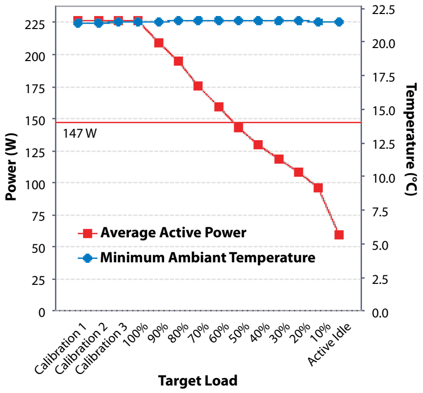


Figure 7: SPECpower_ssj2008 Results: Variation of power at various CPU load levels SGI® Rackable™ C2005-TY6 server with Intel® Xeon® 5670 processors

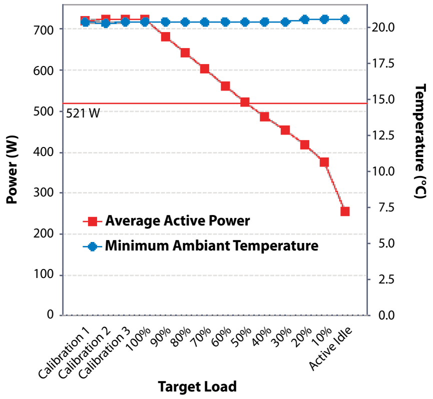


Figure 8: SPECpower_ssj2008 Results: Variation of power at various CPU load levels SGI® Rackable™ C2112-4TY14 server with Intel® Xeon® L5640 processors

4. Case Study

This section describes some tests performed internally to evaluate the influence of several system-level components on the overall power consumption. Case study with variant number of power supplies and CPU SKUs has been done to show how they affect the ultimate performance-per-watt metric.

4.1. Number of Power Supplies

This section describes the tests performed on a SGI® Rackable™ 2U standard-depth 4-node C2112-4TY14 server with one and two 1400 W power supplies in order to show how they effect idle power, power@100% load and the ultimate performance-per-watt metric. Each node comprised of 2x Intel® Xeon® L5640 processors 2.27 GHz (12 cores/24 threads), 4x 4GB 1.35/1.5V 1333MHz memory DIMMs.

Results show that two power supplies contribute most to a higher power consumption on a system between idle and 50% load levels, compared to that with one power supply, the idle power with two power supplies being 12% higher (Figure 9). This leads to effect mostly the workloads within an envelope of 50% CPU load.

This exercise shows that it is important to choose the optimal number of the power supplies and their power rating on a multi-node system, depending on the CPU load levels that the workload will usually run at.

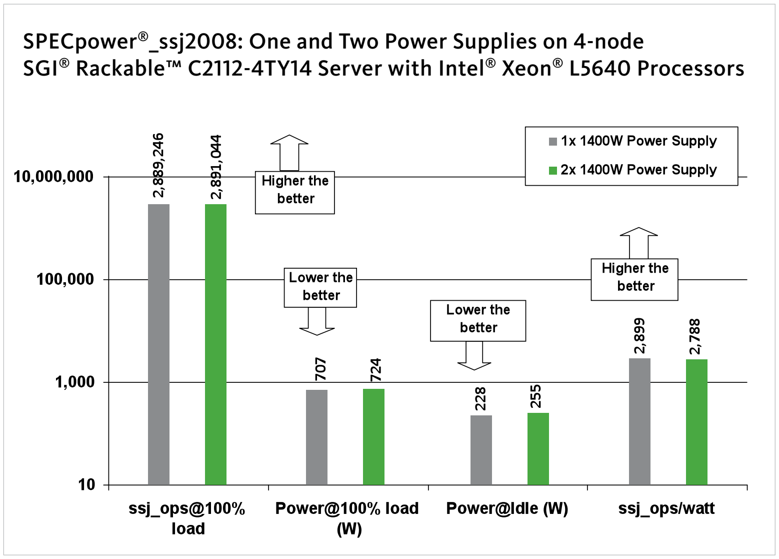


Figure 9: Effect of the number of power supplies on system power consumption

4.2. Choosing the Right Processors

This section describes the tests performed on a single node out of a 4-node SGI® Rackable™ 2U standard-depth C2112-4TY14 server using two types of processors, in order to show their effect on idle power, power@100% load and the ultimate performance-per-watt metric. In the two tests, the node comprised of 2x Intel® Xeon® L5640 processors 2.27 GHz (12 cores/24 threads) and of 2x Intel® Xeon® X5670 processors 2.93 GHz (12 cores/24 threads), 4x 4GB 1.35/1.5V 1333MHz memory DIMMs.

Results show that power consumption increases with the increase in CPU load as usual, the increase being 26% higher @ 100% load with Intel® Xeon® X5670 processors (2.93 GHz) compared to Intel® Xeon® L5460 processors (2.27 GHz) (Figure 10). The performance, however is 21% higher with Intel® Xeon® X5670 processors (2.93 GHz) (Figure 11) resulting in almost similar or slightly better performance-per-watt metric with Intel® Xeon® L5640 processors (2.27 GHz) (Figure 12). The notable difference in power consumption and performance between the two processors in the range of 50-100% CPU load shows that there is a marked tradeoff of performance vs. power consumption for applications utilizing the CPU load in this range. So the appropriate choice of processors depends on the goal that one wants to achieve within a specific range of CPU load level – low power or high performance.

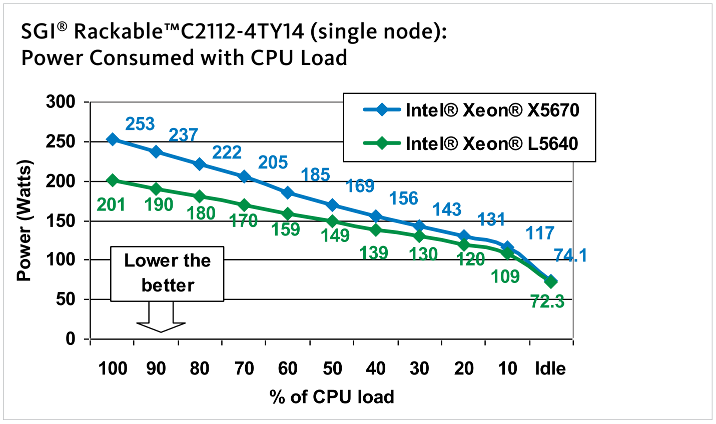


Figure 10: Effect of power consumption with CPU load: Intel® Xeon® L5640 and Intel® Xeon® X5670 processors

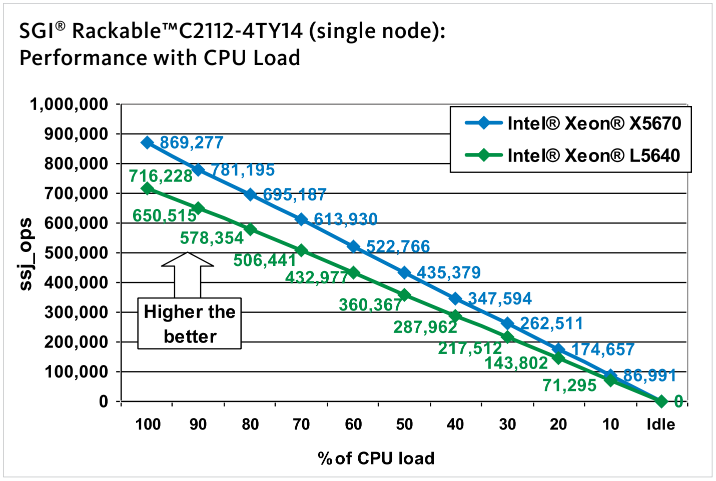


Figure 11: Effect of performance with CPU load: Intel® Xeon® L5640 and Intel® Xeon® X5670 processors

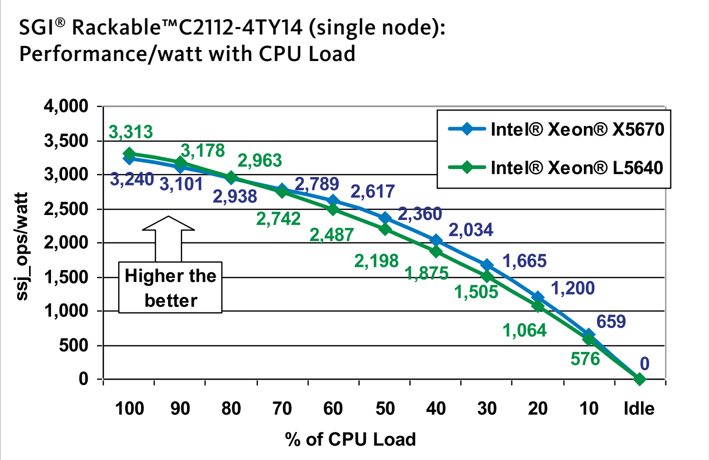


Figure 12: Effect of performance-per-watt with CPU load: Intel® Xeon® L5640 and Intel® Xeon® X5670 processors

5. Performance-Per-watt: Best Practices

This section describes some of the best practices derived from the SPECpower_ssj2008 benchmarks on Intel® Xeon® 5600 processor based SGI® Rackable™ half-depth C2005-TY6 server and 2U standard-depth C2112-4TY14 server. These recommendations are subject to change while moving to a different platform supporting a different motherboard and processor SKU with variations of BIOS options, disk drives or while using different chassis or network hardware/interface. It is recommended to derive a new set of best practices on each platform under test.

The best practices for achieving the highest performance-per-watt metric on Intel® Xeon® 5600 processor based SGI® Rackable™ servers, can be described as follows:

- Performance-per-watt is a ratio that is best achieved with the highest performance out-of-the-box with the lowest power consumption.
- Using low voltage (1.35V-1.5V) memory DIMMs and optimal DIMM count helps to reduce power consumption. Allocating 3x 2GB or 2x 4 GB per socket may be sufficient to run an application optimally but with reduced power.
- Through BIOS options, it is recommended to reduce the memory speed to 1066MZ and QPI to 4.8GTs to get a higher performance-per-watt. This is more visible in a multi-node power benchmark.
- Using processors with a lower clock-rate (< 3.0GHz) and with lower watts (50-60 W or 95 W) often lead to a slightly lower performance, but yield a higher Performance-per-watt metric. This is more visible in a multi-node benchmark.
- For Internet workloads:
  - Applying proper JVM options (for Java Applications) helps to achieve optimal performance;
  - Running the right number of Java instances is recommended (e.g., 3 for IBM J9 VM and 1 for Oracle HotSpot JVM, per six-core socket);
  - Allocating number of cores/threads per JVM such that each JVM instance is affined to a few logical processors or per-socket is important for process and memory locality.
- Using BIOS options on the motherboard (e.g., disable serial ports and USB 2.0 Controller, regulate fan speed, etc) helps to lower the overall power consumption by the system.
- Using a single local disk (preferably a single low power flash device) helps to minimize power without sacrificing I/O performance.
- The Windows 2003 EE operating system power management options works differently from Linux. The memory footprint on Windows is lower. Using the supported BIOS options on the OS for lowering power is recommended.
- Using system-specific power management software (where applicable) with enabled options for server balanced power and performance helps to increase Performance-per-watt.
- Understanding customer goals is important – whether it is a) the performance of the application at certain CPU levels or b) the overall minimization of power for the system hardware to minimize energy cost within an acceptable trade-off in performance.
- It is important to understand the CPU load level(s) the application(s) will execute on a system of choice. Notable differences in power consumption and performance can be seen at certain CPU load levels with variant processor SKUs, number of power supplies or system configuration, whereas the difference may not be prominent at other load levels. So it is recommended to come up with the right configuration to minimize the trade-off as much as possible. For more details on tuning, please refer to SGI® SPECpower publications at:

http://www.spec.org/power_ssj2008/results/power_ssj2008.html.

6. Summary

As depicted in this paper, SGI® Rackable™ half-depth C2005-TY6 server and 2U standard-depth C2112-4TY14 server are compact combined with power and cooling efficiency, thus providing both high performance and low power. This enables adding more of these server hosts into a rack to minimize energy usage and maintain an optimal performance level in a Data Center.

The recommendations for best practices mentioned in this paper are based on results derived from tests on the above mentioned SGI® Rackable™ servers. These are subject to variations with motherboard, BIOS, processor SKUs, and memory options supported for other Intel® Xeon® 5600 processor-based SGI® servers. Additionally, performance may vary due to different storage devices or network hardware/interface.

SGI® Rackable servers are best suitable for Internet/Cloud, Financial Trading and Commercial markets, where energy efficiency for server consolidation in large data centers with minimal performance overhead, is the primary goal.

Contact Information

Contact: Sanhita Sarkar
Performance Engineering, SGI, 46555 Landing Parkway Fremont, CA 94538

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

Global Sales and Support
North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.927.8000
Asia Pacific +61 2.9448.1463