The Implementation of the Server Efficiency Rating Tool

Mike G. Tricker
Microsoft Corporation
mike.tricker@microsoft.com

Klaus-Dieter Lange
Hewlett-Packard Company
klaus.lange@hp.com

Jeremy A. Arnold
IBM Corporation
arnoldje@us.ibm.com

Hansfried Block
Fujitsu Technology Solutions GmbH
hansfried.block@ts.fujitsu.com

Christian Koopmann
University of Paderborn
koop-chris@hotmail.de

ABSTRACT

The Server Efficiency Rating Tool (SERT) [1] has been developed by Standard Performance Evaluation Corporation (SPEC) [2] at the request of the US Environmental Protection Agency (EPA) [3], prompted by concerns that US datacenters consumed almost 3% of all energy in 2010. Since the majority was consumed by servers and their associated heat dissipation systems the EPA launched the ENERGY STAR Computer Server [4] program, focusing on providing projected power consumption information to aid potential server users and purchasers. This program has now been extended to a world-wide audience.

This paper expands upon the one published in 2011 [6], which described the initial design and early development phases of the SERT. Since that publication, the SERT has continued to evolve and has entered the first Beta phase in October 2011 with the goal of being released in 2012. This paper describes more of the details of how the SERT is structured. This includes how components interrelate, how the underlying system capabilities are discovered, and how the various hardware subsystems are measured individually using dedicated worklets.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness)

General Terms

Design, Experimentation, Measurement, Performance, Reliability, Standardization

Keywords

SPEC, SERT, Rating Tool, Benchmark, Energy Efficiency, Power, Server, Storage, Datacenter, ENERGY STAR, Environmental Protection Agency, EPA

ICPE '12, April 22–25, 2012, Boston, Massachusetts, USA.
Copyright 2012 ACM 978-1-4503-1202-8/12/04…$10.00.

1. INTRODUCTION

SPEC was founded in 1988 as a nonprofit organization dedicated to the creation of industry standards for measuring the performance of various aspects of computers and their associated software. It now includes representatives from more than 80 member companies and organizations and has released more than 30 industry-standard benchmarks, which have been used to create more than 20,000 peer-reviewed published performance reports.

SPEC is composed of four major groups: the Open Systems Group (OSG), the High Performance Group (HPG), the Graphics and Workstation Performance Group (GPWG) and most recently the newly created Research Group (RG). The OSG comprises groups covering the major areas of desktop, workstation and server benchmarking and performance evaluation. These groups are responsible for benchmarks characterizing CPU, Java, SFS, Virtualization, and Power. The latter is specifically addressed by the SPECpower Committee, which is responsible for creating and updating the SPECpower_ssj2008 benchmark (ssj2008) [7]. This industry standards committee is currently developing the SERT for the EPA's next generation of ENERGY STAR for Servers program.

Ssj2008 was developed as the first industry-standard cross-platform benchmark for evaluating the combined power and performance characteristics of volume and multi-node server systems. It is based on primarily transactional server-side Java workloads, which exploit many aspects of commercially available Java implementations while exercising processors (CPUs), memory hierarchies (including caches), and the general Symmetric Multiprocessing (SMP) scalability of the systems under test.

The EPA has been tracking the growth in computer (and more specifically server) energy consumption for several years, hosting the Conference on Enterprise Servers and Data Center: Opportunities for Energy Savings in January 2006. Later that year the EPA announced its intention to develop an ENERGY STAR for Enterprise Computer Servers program with broad industry participation and support. This resulted in the ENERGY STAR Computer Server specification launched in May 2009, which recommended the use of ssj2008 to provide the data required to complete the EPA Power and Performance Data Sheet [8].

The SERT has been developed specifically to address the EPA requirements for Version 2 of the ENERGY STAR server [5] program. Unlike most SPEC products, it is not a benchmark having a single score model for use in comparison or marketing.

Instead, it is an evaluation tool that produces detailed information regarding the influence of CPU, memory, and storage IO configurations on overall server power consumption. This resulting information is intended to educate and enable informed purchasing decisions across a broad spectrum of potential customer types and technical backgrounds.

To provide an example of potential usage patterns the Storage IO worklets included in the SERT have been used for an extensive series of experiments on various storage device configurations, including different numbers and models of SATA and SAS HDD and SSD storage devices. The tests were executed on two different computer server models with different maximal storage device capacities under the Microsoft Windows Server 2008 R2 operating system. A modified version of SPEC PTDaemon was employed for the power measurements of total system power in parallel with RAID controller power and storage device power.

This paper also outlines some of the thoughts on how to best describe these subsystem capabilities in ways usable by the broadest range of potential consumers. The authors also intend to provide a follow-up paper with the experimental evaluation once a stable set of SERT worklets is finalized.

2. MOVING BEYOND SSJ2008

When the EPA began to develop Version 2 of the ENERGY STAR for Computer Servers program, they decided that more detailed information regarding the relationship between power consumption and performance for servers was needed. This decision in turn led to the initial requirements for the SERT, which differs from previous SPEC projects in a number of significant ways.

The first and most important difference is that the SERT is not intended to be a benchmark, a fact reflected by the “Rating Tool” aspect of its name. Benchmarks relating to performance and energy efficiency typically focus on the capabilities of servers in addressing specific application areas or business models, often by simulating typical workloads such as Web, File & Print or Database Servers. In contrast, the SERT focuses on providing a first order approximation of energy efficiency across a broad range of application environments.

Unlike most benchmarks, there is no single absolute score as the final outcome of a measurement sequence. This is combined with the EPA requirement for the SERT to be run in an “as shipped” or “out of the box” system configuration, with minimal configuration changes allowed to the system firmware/BIOS, operating system (OS) or middleware such as the Java Virtual Machine (JVM) of the System Under Test (SUT). This deliberate lack of opportunity for optimization is intended to move the focus onto delivering results for the major power-drawing subsystems within a server (CPU, memory, network and storage IO) that will be of use to prospective purchasers and users of servers who need to support multiple workloads with differing performance and IO characteristics.

The distribution of SERT will be similar to that of existing SPEC Benchmarks and the Full Disclosure Report (FDR) produced by each SERT test run will include all setup and tuning details sufficient to enable others to re-create the result(s). The goal is to present the customer the real raw data without a company’s marketing “spin”, which may unduly benefit companies with greater resources to draw from. It is also intended (in agreement with the EPA) that absolute scores may not be used in marketing materials. The aim is to increase participation from smaller companies in the program, specifically looking beyond the traditional multi-national Original Equipment Manufacturers (OEMs), such as smaller Value Added Resellers (VARs) and local system integrators who may manufacture their owns systems from widely available motherboards and components and are a critical part of worldwide emerging markets.

3. AN OVERVIEW OF THE SERT

The SERT is designed to be scalable to a maximum of 64 nodes (limited to a set of homogenous servers or blade servers) and to support multiple power analyzers and temperature sensors. The simplest SERT hardware measurement configuration requires four main hardware components; one **Power Analyzer**, one **Temperature sensor,** a **SUT** and the **Controller**.

The SERT is composed of several elements, starting with the test harness, named **Chauffeur,** which handles the logistical side of measuring and recording the power consumption and inlet temperature of the SUT. It also controls the software installed on both the SUT and Controller, communicating via the TCP/IP transport protocol.

Chauffeur communicates with the **Director**, which instructs the SUT to execute the **suite**, comprising a set of workloads. The **workload** comprises a set of worklets, which exercise the SUT while Chauffeur collects the power and temperature data. The **worklets** are the actual code designed to stress a specific system resource or resources, such as the CPU, memory or storage IO.

The temperature sensor must be placed no more than 50mm in front of (upwind of) the main airflow inlet of the SUT. The SERT will measure the inlet temperature of the SUT and marks the results “valid” only if the temperature measured is 20^{o}C or higher, in order to discourage the “gaming” of the test environment. A stable temperature value is not required during warm-up or measurement phases.

The power analyzer must be located between the AC Line Voltage Source and the SUT. Both are connected to the Controller via their device specific interfaces, as shown in Figure 1. Each analyzer and sensor interacts with its dedicated instance of the **SPEC PTDaemon,** which gathers their readings while the worklets are executed.

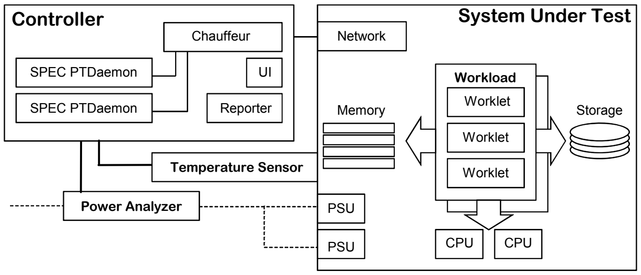


Figure 1. The SERT Overview

The **Reporter**, executed after all measurements phases are completed, compiles all of the environmental, power, and performance data for a complete test run into an easy to read report. The output format will be HTML, plain text, extensible markup language (XML), and comma-separated values (CSV); the HTML report includes a graphical visualization of the results.

4. ARCHITECTURAL CONSIDERATIONS

When the SERT was being designed, a number of decisions were made to produce a comprehensive (and extensible) test in a timely manner to meet the EPA's ENERGY STAR requirements. As SPEC is primarily a volunteer organization that relies on the resources made available by the participating members, it imposes constraints on the development phases.

Consequently, the SERT is currently targeted at 64-bit hardware and OSs only, as this limits the amount of integration testing that is required. Likewise, the majority of code has been developed in Java, partly to ease the cross-platform porting and also reflecting the expertise of SPECpower as ssj2008 was developed in Java. Some C code for certain low-level operations is implemented, as this is also relatively easy to port across platforms.

The SERT goes beyond the hardware goals of Version 2 of the ENERGY STAR server program and is intended to support servers with up to eight processors (also referred to as sockets) and up to 64 nodes, which may be blades or a set of homogeneous multi-node servers. Multi-node servers are defined as having shared infrastructure such as power supplies or backplanes that prevent the servers from operating independently. For example, blade servers are installed in a common enclosure, which usually includes shared power supplies, fans, storage devices, and IO infrastructure such as a backplane or switch.

A primary design goal for the SERT was to scale system performance in proportion to the system configuration. As more components are added (CPU, memory and storage) to a server, the workloads included in the SERT needed to use those resources efficiently, resulting in higher performance when compared against the same basic server design with a less rich hardware configuration. Likewise, if faster components are used instead of the default ones, then the performance needs also increase to reflect that change. This is very important, as adding more or faster components will typically increase the power consumed by a server, affecting the overall efficiency reported. The SERT also supports multiple workload levels (currently idle, 33%, 67% and 100%) that show the overall power/performance characteristics for the server under varying degrees of load, as typically observed in data centers across varying workloads and usage scenarios.

It is also important that the SERT not unnecessarily penalize servers that are not designed to be expandable, but at the same time credits those with greater expandability. Many higher-end servers include highly desirable reliability features such as redundant power supplies and fans, so it is important that such servers not be unduly penalized by the SERT. To ensure that all sorts of server vendors could afford to use the SERT it was agreed with the EPA that the only hardware to be tested will be included within the primary server enclosure. This eliminates the need for complex and expensive external storage devices and network hardware, which greatly simplifies the configuration and use of the SERT. At the same time it includes the server components that draw the most power, including storage devices such as solid state disks or rotating media.

5. SERT: DEFINITION AND EXECUTION

The SERT is composed of a suite of worklets, each of which exercises the SUT in a specific way. For example, the XmlValidate worklet performs validation on a randomly generated XML document, while the Sequential IO worklet performs sequential IO operations on all storage devices included in the SUT. These worklets are grouped into workloads according to the component of the SUT that they are intended to stress: CPU, memory, and storage IO. In addition, a Combined workload consists of application-focused worklets that stress the components of the SUT in a more balanced manner. Figure 2 shows the relationship between the overall suite, workloads, and worklets.

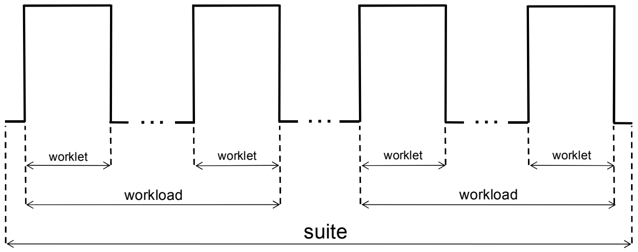


Figure 2. Suite Overview

During a SERT run, each of the worklets is executed consecutively. Each worklet is run in its own set of JVMs or processes in order to minimize interactions between different worklets. Chauffeur automatically launches these client JVMs and coordinates the work among them. Most worklets use multiple client JVMs on the SUT and Chauffeur automatically uses operating system-specific affinity commands to pin each JVM to specific processors in order to avoid artificial limits to scaling. In this context, "client JVM" refers to the client side of a client-server communication pattern and is the JVM that does all of the real work. JVM command-line options are set by Chauffeur (with configurable overrides), allowing for self-tuning of heap sizes and ensuring that the command-line options are reported accurately.

The use of multiple JVMs for running a single worklet is primarily to avoid software bottlenecks (whether in the JVM implementation or in the SERT worklets) from limiting scalability since SERT is intended primarily for measuring the energy efficiency of the hardware and not the software stack. SERT is quite capable of running each worklet in a single JVM, but performance results are likely to be better when using multiple JVMs, e.g., each JVM can be affinitized to a specific processor and therefore all memory accesses will be local to that processor.

Worklets designed for concurrent execution may also be combined into a co-mingled worklet where the individual component worklets run simultaneously rather than consecutively. This introduces more realistic task switching, which is especially useful for IO load simulation. In the current implementation, there is no direct support for the parallel composition of worklets. Instead, a co-mingled worklet can be implemented by creating a new worklet that consists of transactions taken from other worklets, e.g., a processor-intensive transaction and a disk access transaction. As in any other worklet, these transactions could be specified to execute in whatever ratios are desired, e.g., 70% processor intensive, 20% disk reads, 10% disk writes. A future version of SERT/Chauffeur may include support for directly running multiple worklets in parallel.

Most worklets use a "Graduated Measurement" execution sequence (Figure 3). These worklets begin by executing a short warm-up phase (30 sec.), and then run two calibration phases (120 sec.) to automatically determine the maximum throughput each worklet can run on the SUT. Then the worklet runs at multiple load levels, such as 100%, 67%, and 33% of the maximum throughput, generating independent scores for each load level. Each interval of execution includes a pre-measurement (15 sec.) and post-measurement (15 sec.) period in addition to the actual measurement period; each of these periods run for a fixed amount of time. Between each load level a sleep phase (10 sec.) is observed.

Performance and power are reported for the measurement phase (120 sec.) only. This ensures that the worklet is running at steady-state in all client JVMs at the time performance and power are measured.

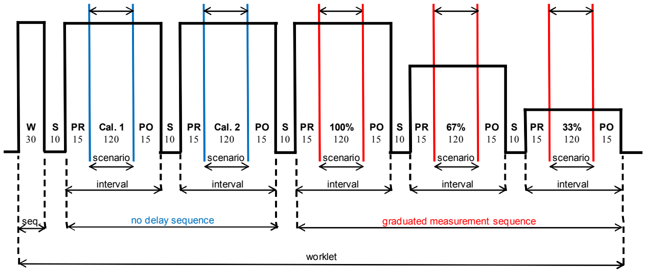


Figure 3. Phases of a Graduated Measurement Sequence

An alternative “Fixed Iteration” execution sequence (Figure 4) is used for worklets that do not support multiple load levels. These worklets run a fixed number of test iterations rather than for a fixed period of time. They optionally include some number of pre- and post-measurement iterations, similar to the pre- and post-measurement periods in a Graduated Measurement sequence.

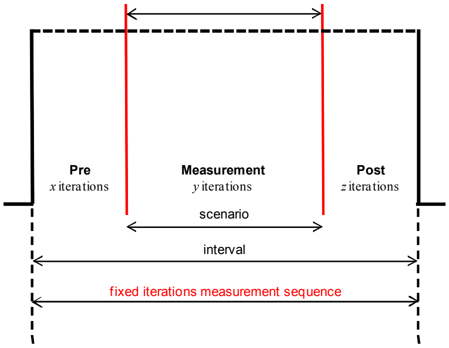


Figure 4. Phases of a Fixed Iteration Sequence

All of the time intervals are configurable in Chauffeur (though the SERT run rules will probably disallow users from changing the interval lengths) and the interval lengths can be adjusted separately for each worklet. The current SERT builds use a two minute warm-up period for the Storage IO worklets since testing has shown that a longer warm-up provides more consistent results for these worklets. Warm-up intervals of 30 seconds are working well for most other worklets, but additional adjustments to the interval lengths will be made if necessary as the SERT is finalized. While a 30 second interval may be needlessly long for some worklets, this constitutes less than 5% of the total worklet run time, so it is unlikely that the warm-up periods will be shortened.

One challenging design goal was that the SERT should thoroughly test the SUT, but at the same time not take so long to complete a test pass that multiple runs in a normal working day became impossible. A complete pass is currently taking between four and five hours depending on SUT hardware configuration and this will be further tuned during the Beta program.

The results from individual worklets are reported individually and can also be combined into higher-level metrics at the workload level to summarize the performance for a particular subcomponent.

5.1 Target Load Levels

Since servers frequently run at less than 100% utilization, it is important for the SERT to assess energy efficiency at multiple load levels. The Chauffer test harness runs each worklet in a calibration mode to determine the maximum transaction rate that the worklet can achieve on the SUT. For each Target Load Level (100%, 67%, 33%), Chauffeur calculates the target transaction rate and the corresponding mean time from the start of one transaction to the start of the next transaction. During the measurement interval, randomized delays are inserted into the worklet execution; these delays follow an exponential distribution that statistically converges to the desired transaction rate. As a result, lower target loads consist of short bursts of activity separated by periods of inactivity. Figure 5 shows a 67% and 33% target load distributions.

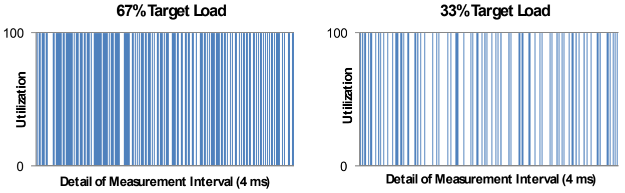


Figure 5. Load Distribution at different Target Loads

6. WORKLET CANDIDATES

SERT worklets were designed under a set of public guidelines [10] to ensure consistent results across a broad spectrum of technologies. For example, each workload must automatically calibrate itself to report the maximum performance available in that specific hardware configuration, and must then be adjustable to target load levels from 100-0% of the maximum performance. Each worklet also needs to scale with the available hardware resources which the execution model deemed “important”, e.g., a CPU worklet needs to scale with the number of processors, cores, hardware threads and the clock frequency.

The SERT Design Document [1] offers a detailed breakdown of what each worklet does and how it works. Currently 16 worklets are under evaluation and categorized in Table 1.

The workloads can be summarized as:

CPU: Data compression, encryption/decryption, complex number arithmetic, matrix factorization, floating point array manipulation, sorting algorithm, string manipulation, and XML document validation;

Memory: XML document manipulation and validation using pre-computed and cached data lookup, and array manipulation with read/write operations across four major classes of data transformation;

Storage IO: Four individual transaction pairs combining sequential/random read/write and a mixed transaction which combines all four;

Combined: The concept of CSSJ is derived from ssj2008, which simulated an on-line Transaction Processing workload in which customers order and pay for goods from warehouses that handle delivery and stock replenishment;

Active Idle: A steady state in which the server is ready to execute any worklet but is not actually doing so, leading to a measure of efficiency for a fully functional but otherwise idle state.

Table 1. Worklet Candidates

Workload	Worklet	Sequence Execution	Metric
CPU	Compress	Graduated	Transactions/sec
	CryptoAES	Graduated	Transactions/sec
	SOR	Graduated	Transactions/sec
	SORT	Graduated	Transactions/sec
	SHA256	Graduated	Transactions/sec
	FFT	Graduated	Transactions/sec
	LU	Graduated	Transactions/sec
	XmlValidate	Graduated	Transactions/sec
Memory	XmlValidate1	Graduated	Transactions/sec*cache size*cache scaling factor
	XmlValidate2	Graduated	Transactions/sec*cache size*cache scaling factor
	Flood	Fixed	Memory bandwidth (GB/sec)*memory size (GB)
Storage IO	Random	Graduated	Transactions/sec
	Sequential	Graduated	Transactions/sec
	Mixed	Graduated	Transactions/sec
Combined	CSSJ	Graduated	Transactions/sec
Idle	Active Idle	N/A	N/A

There are no worklets related to **Network IO**, which will be handled by a "configuration modifier" that simulates the steady state efficiency of a network device. After testing a variety of network interface cards (NICs) across a range of workloads it was observed that the power consumption of the actual devices approximated very closely to a constant (including in the case of NICs that perform offloading from the host processor), with CPU and memory power consumption being the biggest factors influencing overall system efficiency. Combined with the extensive set of external hardware required to effectively test network bandwidth and performance, it was agreed with the EPA that a modifier would be applied to simulate the network IO contribution to overall server efficiency.

7. STORAGE IO WORKLETS

The Storage IO worklets developed for SERT generate synthetic loads on server storage devices mimicking basic access patterns from real world usage models. The tests described in this paper were performed to check the suitability of the implementation for the designed purpose, especially testing whether the design goals given in the SERT Design Document section 2.6.1 [1] are met.

7.1 Test Configurations

The experiments described in this paper are based on prerelease versions of SERT. The results may not be representative for the final release.

In order to show the scaling capabilities of the Storage IO worklets the tests were executed on two different server models:

1.) Fujitsu PRIMERGY TX300 S6 tower server with up to 20 internal 2.5" disk drive bays was selected for showing scale out properties using many devices;

2.) the rack server model PRIMERGY RX300 S6 with up to 12 internal 2.5" disk drive bays was used for the experiments with different device technologies and for separate measurements of the controller and storage device power.

7.1.1 Power Measurement Set-Up

Each controller and each storage device backplane requires its own power analyzer for the internal measurements. The high end configuration in the tower server includes up to two controllers and two backplanes, which exceeded the limits of available power analyzers. Therefore internal measurements were performed for the rack server experiments only.

Temperature sensors were used in all test scenarios to measure the ambient temperature and ensure that it always stays above the required minimum of 20°C, which has been selected as a realistic data center temperature and ensures that testing is not "gamed" by the use of artificially low temperatures. The temperature sensors are omitted in the following configuration pictures for better readability.

7.1.1.1 Tower Server Measurement Set-Up

For this test series the server Power Supply Unit (PSU) was connected to a ZES LMG450 multichannel power analyzer as shown in Figure 6. The default version of SPEC PTDaemon as included in the SERT Beta 1 kit was used for this configuration.

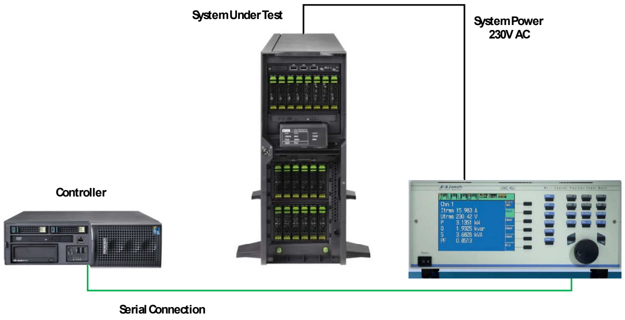


Figure 6. Tower Server Measurement Set-Up

7.1.1.2 Rack Server Measurement Set-Up

Besides an Infratek 107A-1 power analyzer measuring the overall server power consumption at the system's PSU (230V AC), two high precision ZES LMG95 single phase power analyzers were added for measuring the RAID controller and storage device power (12V DC). One of these was connected to a PCI Express (PCIe) adapter card inserted between the PCIe main board slot and the RAID controller. The other one was connected to the storage device backplane. Figure 7 shows the general set-up.

The RAID controller requires two voltages: 3.3V for standby and 12V for active mode. Separate measurements have shown that the standby power does not change with the load. Therefore, only the 12V power was measured for these tests. A fixed amount of 2.1W standby power was added to all controller power measurements for result evaluation. The storage device backplane includes a

SAS expander in order to support 12 device ports using two SAS 4x connectors provided by the RAID controller. The expander power is included in the device power measurements.

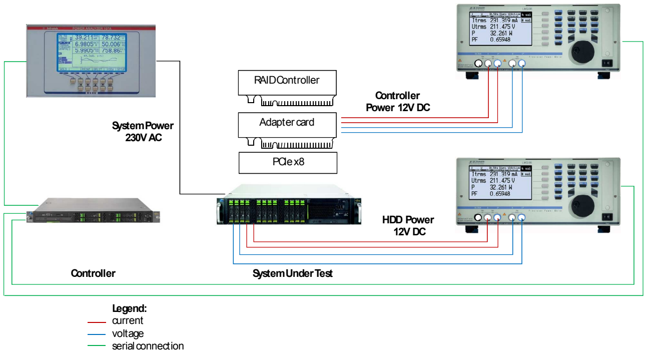


Figure 7. Rack Server Measurement Set-Up

A modified version of SPEC PTDaemon, which supports DC measurements, was implemented for the test series. The uncertainties of DC measurements are significantly higher than those of AC measurements, specifically with lower voltages. In order to stay below the 1% uncertainty threshold required for SPEC power measurements, high precision power analyzers had to be used. This special version of PTDaemon is for internal use only. Currently there are no plans to release this version with the final SERT kit.

7.1.2 The SERT Storage IO Worklets

The SERT includes three Storage IO worklets implementing the basic storage access patterns (sequential, random and mixed) using the characteristics given in Table 2.

Table 2. Storage Worklet Characteristics

Worklet	Access Pattern	Block Size	Read / Write Ratio
Sequential	100% Seq.	128kB	9 / 1
Mixed	50% Seq.	128kB	9 / 1
	50% Rand.	8kB	7 / 3
Random	100% Rand.	8kB	2 / 1

The code requires the storage test devices being formatted to a standard file system, e.g., NTFS (Windows), ext4 (Linux). For optimal performance this file system should store no other files but the test files created by the storage worklets. SERT starts one client instance per storage device, each running four user threads in parallel. Each user thread creates two test files of 1GB size, i.e., there are eight 1GB test files per device. These test files are generated consecutively per device in order to ensure largely sequential layout on the physical storage media. Existing test files will be reused for subsequent tests and will not be recreated for every test run. Because of the four parallel users per device, the sequential access is not completely sequential on the physical media. However, the four user threads are required to guarantee sufficient outstanding IO operations in the device queue to keep them constantly busy, even for high performance storage devices, e.g., SSDs.

Table 3 shows the test files and their corresponding users. The file name format is: <Client-ID>-<User-ID>-<File_Number>.dat; where Client-ID uniquely identifies the storage device.

All storage worklets share the same basic code. The default access pattern and test file definitions shown above are specified in configuration files and can be modified for research purposes without changing the code, but must be used unchanged for valid SERT results.

Table 3. Test Files

File Name	User	Size
001-0001-001.dat	1	1GB
001-0001-002.dat	1	1GB
:	:	:
001-0004-002.dat	4	1GB

As specified in the SERT Design Document [1] the Storage IO worklets should give credit to higher performance storage devices independently of any controller or main memory caching features. In order to achieve this goal the storage worklet code uses basic OS File IO routines configured to provide unbuffered access to the physical devices circumventing any caching mechanisms. This kind of IO routines is not directly available via integrated Java classes. Instead, the Java Native Access (JNA) interface is used to call the native OS File IO routines from within the SERT Java code.

Finally, it should be mentioned that the sequential code completely walks through one test file per user before changing to the next one. The random code switches position randomly between the two test files and within these files.

7.1.3 The Tested Configurations

The basic configuration of the two test systems as described below was nearly identical for most of the test cases:

- CPU: 2 x Intel Xeon X5675
- RAM: 12 x 2GB (Rack Server) / 8GB (Tower Server) PC3-10600R DIMMs
- RAID Controller: 1 x LSI 2108 SAS
- PSU: 1 x 800W
- OS: Microsoft Windows Server 2008 R2
- File System: NTFS
- JVM: Oracle HotSpot 1.6.0_27-b07 (Rack Server) Oracle HotSpot 1.7.0_02-b13 (Tower Server)
- Storage Devices: different types and numbers (see below)

The tower server was tested with 146GB 2.5” SAS 10krpm HDDs only, which were used as boot devices on both servers, too.

A second partition was created on these boot devices and used for some of the storage worklet test cases.

The tower server test cases are described using the following notation: **OS + (*x*, *y*)**. Where *x* denotes the number of disks connected to the first RAID controller in addition to the OS boot device (maximum eight included OS) and *y* denotes the device count for the second RAID controller (maximum 12). The following configurations have been tested:
OS + (7, 0), OS + (4, 0), OS + (2, 0), OS + (1, 0), OS + (0, 0), OS + (7, 12), OS + (7, 9), OS + (6, 6), OS + (4, 4), OS + (3, 4)

As the rack server includes a single RAID controller only, the description is given as OS + *x*, *x* = number of storage devices (maximum 12 including OS).

The rack server test configurations:

OS + 8, OS + 4, OS + 2, OS + 1, OS + 0 (= 2nd boot dev. partition)

This sequence of test configurations was executed using the following storage devices from different manufacturers (all 2.5" form factor):

- 120GB SATA 5.4krpm
- 500GB SATA 7.2krpm
- 146GB SAS 10krpm
- 146GB SAS 15krpm
- 64GB SATA SSD

SATA and SAS disks have significant price/performance differences. One of the key differences relates to their respective densities, with SAS offering significantly better performance; while SATA offers much better density. It is not the intention of this paper to analyze the respective benefits of the competing technologies, but to make use of the different storage access attributes across varying devices speeds.

Each of these configurations on both servers was tested with five consecutive SERT runs in order to examine the run to run variations. The SERT test configuration file (config-all.xml) was modified to execute the Storage IO and Idle worklets only, resulting in a reduced execution time of about one hour per test run.

7.2 Tower Server Test Results

This section presents the results of the experiments executed on the tower server with up to 19 tested storage devices: detailed description of the Storage IO worklet scaling capabilities, comparison of the power consumption of the three worklets, and results of the hardware configuration changes.

After finishing the first set of tests using the SERT Beta 1 kit, problems with the seeding of the Random Number Generator in the worklet code were detected. These problems have been fixed in a subsequent internal SERT release. The results presented below are from a second test series using this internal release.

7.2.1 Storage Device Scaling - Sequential Access

Table 4 and Table 5 show the throughput and power results of the sequential access worklet for an increasing number of devices at all three load levels, starting with a second partition on the OS boot device up to 19 additional SAS 10krpm HDDs.

Observations:

- Throughput for the second partition on the boot device is close to a single separate device.
- Throughput scales almost linearly with the number of storage devices.
- Throughput for the two configurations with seven HDDs is about the same. Power consumption for the OS + (3,4) configuration is higher because a second RAID controller was added and the 12 HDD backplane includes a SAS expander, which consumes additional power. The basic power difference is clearly visible looking at Idle power values.
- The three non-zero load levels get to the expected throughput. Power difference between 33% and 67% is higher than between 67% and 100% due to active power management at lower load levels.
- Processor time as shown in Table 4 and 5 is very low and scales with the number of storage devices, except for the low end configurations with one to four devices, which all cause a similar base load on the CPU.

Table 4. Storage Device Scaling Results – Part 1

Devices **SAS 10krpm**	**OS+** **(0,0)**	**OS+** **(1,0)**	**OS+** **(2,0)**	**OS+** **(4,0)**	**OS+** **(7,0)**
100% seq. (MB/s)	40.9	44.2	88.3	175.0	303.8
67% seq. (MB/s)	27.6	29.6	58.9	117.3	203.8
33% seq. (MB/s)	13.7	14.8	29.5	58.8	101.6
100% seq. (W)	120.6	126.3	140.4	162.4	195.5
67% seq. (W)	118.8	124.5	137.5	157.1	188.3
33% seq. (W)	115.7	121.5	131.8	148.4	176.1
Idle (W)	109.3	115.3	121.4	132.5	154.8
% Processor Time	0.7%	0.5%	0.4%	0.7%	1.4%

Table 5. Storage Device Scaling Results – Part 2

Devices **SAS 10krpm**	**OS+** **(3,4)**	**OS+** **(4,4)**	**OS+** **(6,6)**	**OS+** **(7,9)**	**OS+** **(7,12)**
100% seq. (MB/s)	305.0	349.7	523.9	692.9	823.3
67% seq. (MB/s)	204.7	234.4	351.6	464.8	552.5
33% seq. (MB/s)	102.4	117.1	175.6	232.2	276.3
100% seq. (W)	211.2	223.5	258.3	298.0	323.6
67% seq. (W)	204.9	216.6	250.4	287.9	311.8
33% seq. (W)	192.0	202.7	233.5	268.9	291.9
Idle (W)	172.5	180.5	204.7	230.9	255.0
% Processor Time	1.2%	1.4%	2.1%	2.6%	3.0%

7.2.2 Worklet Power Comparison

The power and performance differences for all three worklets and all tower server test configurations are shown in Table 6 and 7. For this comparison only the results of the 100% load level are shown.

Mixed access throughput is roughly 85% and random throughput is about 75% of pure sequential throughput for most configurations.

There are only minor differences in power consumption between the three worklets with sequential at the top and random at the bottom.

Table 6. Storage IO Power Comparison Results – Part 1

Devices **SAS 10krpm**	**OS +** **(0, 0)**	**OS +** **(1, 0)**	**OS +** **(2, 0)**	**OS +** **(4, 0)**	**OS +** **(7, 0)**
100% Seq. (MB/s)	40.9	44.2	88.3	175.0	303.8
100% Mix. (MB/s)	19.2	20.3	39.2	79.2	137.0
100% Rnd. (MB/s)	2.1	2.1	4.1	8.2	14.3
100% Seq. (W)	120.6	126.3	140.4	162.4	195.5
100% Mix. (W)	118.8	123.4	136.4	158.3	190.6
100% Rnd. (W)	116.7	120.5	133.3	153.0	184.2
Idle (W)	109.3	115.3	121.4	132.5	154.8

Table 7. Storage IO Power Comparison Results – Part 2

Devices **SAS 10krpm**	**OS +** **(3, 4)**	**OS +** **(4, 4)**	**OS +** **(6, 6)**	**OS +** **(7, 9)**	**OS +** **(7, 12)**
100% Seq. (MB/s)	305.0	349.7	523.9	692.9	823.3
100% Mix. (MB/s)	137.4	157.6	235.7	313.2	371.9
100% Rnd. (MB/s)	14.3	16.4	24.6	32.7	38.9

Devices SAS 10krpm	OS + (3, 4)	OS + (4, 4)	OS + (6, 6)	OS + (7, 9)	OS + (7, 12)
100% Seq. (W)	211.2	223.5	258.3	298.0	323.6
100% Mix. (W)	207.2	218.7	253.3	291.9	316.5
100% Rnd. (W)	199.5	212.0	247.2	285.8	310.4
Idle (W)	172.5	180.5	204.7	230.9	255.0

7.2.3 JVM Comparison

For some selected configurations a second sequence of tests was executed using IBM J9 JVM instead of Oracle HotSpot. The results presented in Table 8 show that there are virtually no power or performance differences between these JVMs.

This is the desired behavior of all SERT worklets and specifically of the Storage worklets.

Table 8. Oracle HotSpot versus IBM J9 - Results

Devices @ 100% load	OS+ (1,0)	OS+ (4,0)	OS+ (3,4)	OS+ (7,12)
Seq. HotSpot (MB/s)	44.2	303.8	523.9	823.3
Seq. J9 (MB/s)	44.4	305.0	523.7	822.4
Rnd. HotSpot (MB/s)	2.1	14.3	24.6	38.9
Rnd. J9 (MB/s)	2.2	14.4	24.7	39.0
Seq. HotSpot (W)	126.3	195.5	258.3	323.6
Seq. J9 (W)	125.5	193.6	260.4	328.8
Rnd. HotSpot (W)	120.5	184.2	247.2	310.4
Rnd. J9 (W)	122.9	182.2	248.5	313.6
Idle (W)	115.3	154.8	204.7	255.0

7.2.4 RAM Comparison

Another experiment compares power consumption for two main memory configurations: 12 x 8GB and 6 x 1GB. The DIMM technology for both configurations was the same.

Both configurations perform about the same, i.e., the smaller memory capacity is still sufficient to exercise the full number of HDDs unrestricted. Previous tests have shown that each Storage IO client instance requires less than 256MB of heap space, so even much smaller memory configurations are able to support the Storage IO worklets.

Table 9. 96GB versus 6GB - Results

Devices @ 100% load	OS+ (1,0)	OS+ (7,12)
Seq. 96GB (MB/s)	44.2	823.3
Seq. 24GB (MB/s)	44.1	823.3
Rnd. 96GB (MB/s)	2.1	38.9
Rnd. 24GB (MB/s)	2.1	38.9
Seq. 96GB (W)	126.3	323.6
Seq. 24GB (W)	118.8	313.8
Rnd. 96GB (W)	120.5	310.4
Rnd. 24GB (W)	113.3	302.5
Idle 96GB (W)	115.3	255.0
Idle 24GB (W)	107.5	253.3

Although the number and capacity of DIMMs was cut by half, the system power was only reduced slightly as shown in Table 9. The base difference can be inferred comparing the Idle rows. Memory power is only a minor part of the overall power, which is dominated by the storage device power at Idle and by CPU power at 100% load.

7.2.5 PSU Comparison

For the following test series in the tower server a second PSU was added. This is a typical configuration for many data centers which require PSU redundancy. As expected this has no influence on the IO performance. However, there is a significant rise in power, mainly because PSU efficiency is very poor below 20% of nominal power, e.g., below 20% of 2 x 800W = 320W.

Table 10. 1 PSU versus 2 PSUs - Results

Devices @ 100% load	OS+ (1,0)	OS+ (7,12)
Seq. 1PSU (MB/s)	44.1	823.3
Seq. 2PSUs (MB/s)	44.0	823.4
Rnd. 1PSU (MB/s)	2.1	38.9
Rnd. 2PSUs (MB/s)	2.1	38.9
Seq. 1PSU (W)	118.8	313.8
Seq. 2PSUs (W)	137.1	321.4
Rnd. 1PSU (W)	113.3	302.5
Rnd. 2PSUs (W)	132.7	309.7
Idle 1PSU (W)	107.5	253.3
Idle 2PSUs (W)	126.4	261.0

7.2.6 CPU Comparison

For the final test series in the tower server the tests were repeated using different CPU models: a top bin high performance unit and a low voltage model with significantly reduced performance. These experiments were executed on two storage configurations only with the minimal and maximal number of storage devices.

The CPU properties of the standard processor used for the majority of the experiments and the new ones added for this comparison are given in Table 11.

Table 11. Storage IO CPU Properties

CPU	Freq. (GHz)	Cores	Threads	Cache (MB)	TDP (W)
X5690	3.46	6	12	12	130
X5675	3.06	6	12	12	130
L5609	1.86	4	4	12	40

The throughput and power results of these experiments are presented in Table 12. All results are for the sequential access worklet at 100% load.

Table 12. Storage IO CPU Comparison - Results

Devices @ 100% load	OS+ (1,0)	OS+ (7,12)
Seq. X5690 (MB/s)	44.3	822.0
Seq. X5675 (MB/s)	44.2	823.3
Seq. L5609 (MB/s)	44.2	822.7
Seq. X5690 (W)	127.5	327.0
Seq. X5675 (W)	126.3	323.6
Seq. L5609 (W)	116.4	307.1
Idle X5690 (W)	114.4	258.5
Idle X5675 (W)	115.3	255.0
Idle L5609 (W)	103.7	243.8

The performance results confirm that the storage device throughput does not depend on the CPU capabilities. Even the low end processor is capable of saturating the highest number of storage devices. The two 130W CPUs only show minor power differences, whereas the low voltage CPU consumes significantly less power. This is the desired behavior and conforms to the SERT design goals.

7.3 Rack Server Test Results

The extended measurement set-up of the rack server gives a more detailed view of the server power consumption, especially showing the power drawn by the main Storage IO components, the RAID controller, and the storage devices themselves. All the test results presented below are from tests using the Beta 1 SERT release.

7.3.1 System-, Disk-, and Controller-Power

Table 13 displays the performance and power usage for one to eight SATA 5.4krpm HDDs using the sequential access pattern. Different from the previous tables, the power for the storage devices and the RAID controller are shown instead of overall system power.

Table 13. Storage IO Component Power - Results

Devices SATA 5.4krpm	Idle	100%	67%	33%
1 HDD (MB/s)	N/A	14.8	9.9	5.0
2 HDDs (MB/s)	N/A	29.9	20.1	10.0
4 HDDs (MB/s)	N/A	60.6	40.4	20.2
8 HDDs (MB/s)	N/A	120.2	80.5	40.3
1 HDD Disk (W)	13.8	15.8	15.5	15.3
2 HDDs Disk (W)	14.6	18.6	18.0	17.5
4 HDDs Disk (W)	16.1	24.0	22.7	21.6
8 HDDs Disk (W)	19.5	34.7	32.4	30.3
1 HDD Ctr. (W)	8.3	8.3	8.3	8.3
8 HDD Ctr. (W)	8.3	8.3	8.3	8.3

Same as with the previously described tower tests, the performance scales almost linearly with the number of HDDs and the three load levels are matched closely. The device power for the load levels however does not change much. Significant differences can be observed for higher device counts and between Idle and 100% only.

The controller power is completely independent of any load or the number of connected storage devices; it stayed at a constant level for all these experiments.

Another view showing the minimal and maximal configurations of these tests is presented in Table 14, emphasizing the difference between overall system power and component power.

The overall system power increase is much higher than the power increase of the Storage IO components, especially for higher number of devices. The two Delta rows in Table 14 confirm this observation. System power usage is dominated by CPU power, which significantly increases with higher throughput. The device power however is less dependent on the load. Particularly for rotating media, it is mainly determined by the basic power for spinning the platters.

Table 14. System Power versus IO Device Power - Results

Devices SATA 5.4krpm	Idle	100%	67%	33%
1 HDD Sys. (W)	137.7	154.6	147.1	142.9
8 HDDs Sys. (W)	144.0	213.6	200.7	183.9
Delta	6.3	58.9	53.7	41.0
1 HDD Disk (W)	13.8	15.8	15.5	15.3
8 HDDs Disk (W)	19.5	34.7	32.4	30.3
Delta	5.7	18.8	16.8	15.0
1 HDD Ctr. (W)	8.3	8.3	8.3	8.3
8 HDD Ctr. (W)	8.3	8.3	8.3	8.3

7.3.2 Comparing Storage Technologies

A global overview covering all types of storage devices included in the tests is given in Table 15.

The displayed scaling does not exactly match the real differences between the technologies due to a problem with the Storage IO worklet code. The SATA 5.4k, SAS 10k, and SSD tests were performed with code, which did include some unwanted debug code. This code generated an excessive number of log messages on the OS device, which effectively limited the throughput of these configurations. The other configurations have been tested using new binaries without the debug code. Due to the high number of tests, there was no time left to repeat the measurements. Some comparison tests indicate that the difference is less than 5%.

Remarks:

- The 8 SSD configuration is limited by the available bandwidth of the SATA connectors.
- System power increases with the number of devices and the rotational speed, where SSD technology shows the expected lower System power, especially in Idle mode.
- System power in 1 HDD configurations is dominated by CPU power, causing the irregular SSD behavior for this configuration.

Table 15. Storage Technology Comparison - Results

Devices @ 100% Seq.	SATA 5.4k	SATA 7.2k	SAS 10k	SAS 15k	SATA SSD
1 HDD (MB/s)	14.8	41.0	34.0	76.1	177.0
2 HDDs (MB/s)	29.9	61.0	66.7	151.0	365.5
4 HDDs (MB/s)	60.6	124.2	135.0	302.0	723.2
8 HDDs (MB/s)	120.2	252.7	270.9	588.0	844.7
1 HDD (W)	154.6	174.3	176.1	184.0	193.8
2 HDDs (W)	173.0	176.6	191.3	201.7	200.4
4 HDDs (W)	189.3	201.6	214.1	224.4	212.8
8 HDDs (W)	213.6	226.9	255.2	263.6	218.2
Idle 1 HDD (W)	137.7	145.9	147.8	150.4	145.9
Idle 8HDDs	144.0	169.0	189.3	197.5	147.1

Table 16 displays power consumption, adding power measured internally for the different devices. This data show that SSD device power is actually below device power for the other four technologies, even for the one device configuration. SSD system power for one device is higher though due to the increased CPU power caused by the higher SSD throughput.

Table 16. System-, Controller-, and Device-Power - Results

Devices @ 100% Seq.	SATA 5.4k	SATA 7.2k	SAS 10k	SAS 15k	SATA SSD
1 HDD Sys. (W)	154.6	174.3	176.1	184.0	193.8
2 HDDs Sys. (W)	173.0	176.6	191.3	201.7	200.4
4 HDDs Sys. (W)	189.3	201.6	214.1	224.4	212.8
8 HDDs Sys. (W)	213.6	226.9	255.2	263.6	218.2
1 HDD Dev. (W)	15.8	17.5	19.9	19.8	15.4
2 HDDs Dev. (W)	18.6	20.8	26.5	26.6	17.8
4 HDDs Dev. (W)	24.0	28.7	40.5	39.8	22.2
8 HDDs Dev. (W)	34.7	44.4	68.3	66.8	26.4
1 HDD Ctr. (W)	8.3	8.3	8.4	8.4	8.4
8 HDDs Ctr. (W)	8.3	8.5	8.5	8.7	8.6

7.4 Problems Observed

This paper primarily presents the results from the sequential access worklet. The reason is that the other two worklets, mixed and random access, have shown inconsistent results in several configurations, specifically with one or two storage devices only. This problem was caused by setting bad seeds for the random number generators in the storage worklet code of the Beta 1 kit. It has been fixed in a later internal SERT release, which was used for the repeated tower server test results presented above. There was no time left to repeat the rack server tests with this new SERT kit. However, the tower server experiments have shown that the sequential access worklet results of the Beta 1 kit are accurate.

In the critical configurations, observed are very high run to run variations for the Beta 1 test sequences, characterized by a Coefficient of Variation (CV) between 10% and 30%. The acceptable CV limit for SERT tests is defined as 3%. Most of the configurations typically show very low variations, e.g., CV < 0.5% for the five consecutive test runs in our experiments.

Also for the critical configurations in the Beta 1 kit the 100% load point was missed regularly, in some cases less than 70% of the calibrated throughput has been achieved.

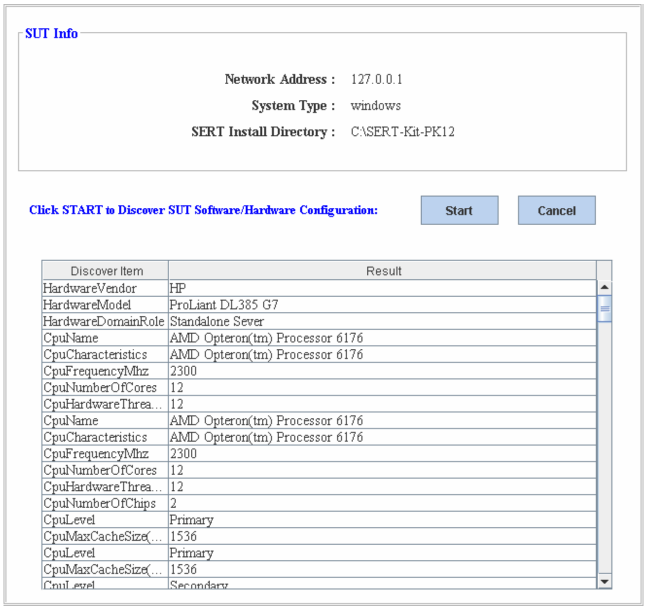
8. SERT UI

Users may configure the SERT by manually editing the various configuration files or utilize the newly designed SERT User Interface (SERT UI) in order to manage the behavior of each component.

During **Host Discovery** (Figure 8), the detailed hardware and software configuration of the SUT are gathered automatically by a remote task that uses the industry standard Common Information Model (CIM) definitions that are widely supported across hardware and OS platforms.

The SERT UI provides a graphical interface for gathering all the SUT hardware and software configuration data, configuring and running the SERT, as well as archiving the measured results and log files. It also supports the ability to save and re-import complete configurations to simplify repeated testing.

The default mode executes the entire SERT suite (all worklets) in sequence, each worklet in a new instance of the local JVM, in order to create an EPA compliant test record. The SERT UI also offers an advanced research mode allowing the selective execution of a subset of workloads and worklets.



SUT Info

Network Address : 127.0.0.1

System Type : windows

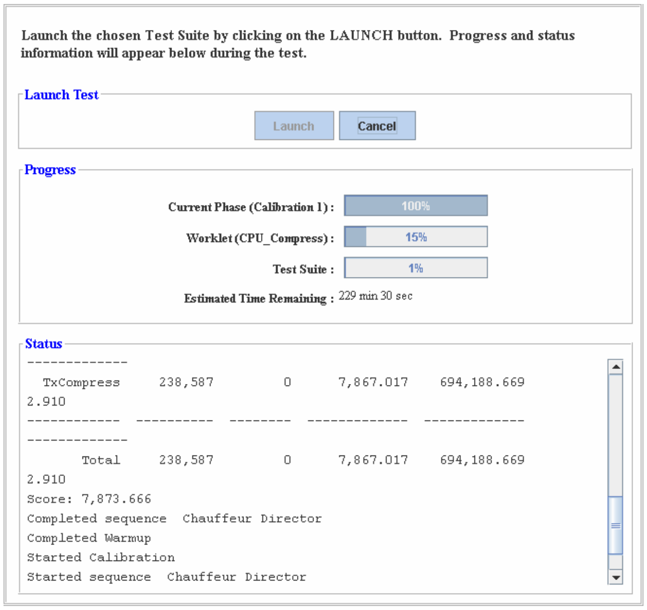
SERT Install Directory : C:\SERT-Kit-PK12

Click START to Discover SUT Software/Hardware Configuration:

Discover Item	Result
HardwareVendor	HP
HardwareModel	ProLiant DL385 G7
HardwareDomainRole	Standalone Sever
CpuName	AMD Opteron(tm) Processor 6176
CpuCharacteristics	AMD Opteron(tm) Processor 6176
CpuFrequencyMhz	2300
CpuNumberOfCores	12
CpuHardwareThrea...	12
CpuName	AMD Opteron(tm) Processor 6176
CpuCharacteristics	AMD Opteron(tm) Processor 6176
CpuFrequencyMhz	2300
CpuNumberOfCores	12
CpuHardwareThrea...	12
CpuNumberOfChips	2
CpuLevel	Primary
CpuMaxCacheSize(...	1536
CpuLevel	Primary
CpuMaxCacheSize(...	1536
CpuLevel	Secondary

Figure 8. SERT UI: Host Discovery

At the **Launch Test** (Figure 9) the progress of the entire suite can be observed, as well as the status of the currently executing worklet.



Launch the chosen Test Suite by clicking on the LAUNCH button. Progress and status information will appear below during the test.

Launch Test

Progress

Current Phase (Calibration 1) : 100%

Worklet (CPU_Compress) : 15%

Test Suite : 1%

Estimated Time Remaining : 229 min 30 sec

Status

```
-------------
  TxCompress     238,587         0      7,867.017     694,188.669
2.910
------------  ----------  --------  -------------  -------------
-------------
       Total     238,587         0      7,867.017     694,188.669
2.910
Score: 7,873.666
Completed sequence  Chauffeur Director
Completed Warmup
Started Calibration
Started sequence  Chauffeur Director
```

Figure 9. SERT UI: Suite Launch Menu

9. CONCLUSIONS

At the time of writing, the first Beta of the SERT has been delivered, with a second to follow soon, and a Release Candidate is targeted for the first half of 2012. The expectation is that the SERT will be released, together with Version 2 of the ENERGY STAR Computer Server program, in the first half of 2012.

By building on the knowledge gained during the development and on-going support of ssj2008, SPECpower was able to develop a tool that is easier to configure and use while offering a broader set of tests focusing on the major sub-components of servers. It provides the ability to support large systems with a high number of processors and server nodes, and with unlimited memory and on-board storage devices. The design is fundamentally extensible so that as new hardware types emerge additional worklets can easily be added.

With the growing worldwide interest in increasing server and data center efficiency it is anticipated that the SERT will be even more widely used than ssj2008. There are already plans in consideration for future enhancements that the highly modular architecture supporting various forms of serial and parallel test execution is designed to support. Additional platforms and architectures may be supported as industry resources are made available for test and development.

The results presented here demonstrate that the Storage IO worklets included in the SPEC SERT largely meet the intended design goals. They can be used for reliable measurements of storage device efficiency, extending the capabilities of currently available computer server efficiency benchmarks.

Experiments have shown that total system power increases significantly in proportion to the number of storage devices and the load levels, whereas the storage device power increases only marginal at higher load levels. The RAID controller power was almost stable under all test conditions.

Further experiments are planned comparing additional hardware configurations, other operating systems plus different JVM versions and parameters. A comparison of the most important RAID configurations is intended for testing the applicability of the Storage IO worklets to these device configurations.

By offering a detailed breakdown of subsystem efficiency, the SERT enables potential server purchasers to evaluate and compare aspects of different servers that relate most closely to a broad range of potential workloads. This range can include any environment, from small office users combining all their applications onto a handful of servers up to enterprise data centers supporting many tens of thousands of users and thousands of different workloads. It is anticipated that the SERT will be the first of a new class of system evaluation tools that will be widely used across the world in the years to come.

10. ACKNOWLEDGMENTS

The authors want to acknowledge the additional members of the SPECpower Committee who have contributed to the design, development, testing, and overall success of the SERT: Paul Muehr, Van Smith, Greg Darnell, John Beckett, Karl Huppler, Sanjay Sharma, and David Ott.

The authors would also like to acknowledge Karin Wulf for providing a modified version of SPEC PTDaemon supporting DC power measurements and Peter Klassen for executing numerous measurements for this paper.

11. REFERENCES

[1] Server Efficiency Rating Tool public Design Document (latest version): http://www.spec.org/sert/docs/SERT-Design_Doc.pdf

[2] Standard Performance Evaluation Corporation home page: http://www.spec.org

[3] US EPA ENERGY STAR Enterprise Servers home page: http://www.energystar.gov/index.cfm?c=archives.enterprise_servers

[4] US EPA ENERGY STAR Computer Specification Version 1.0: http://www.energystar.gov/ia/partners/product_specs/program_reqs/computer_server_prog_req.pdf

[5] US EPA ENERGY STAR Computer Servers Draft 1 Version 2.0: http://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer_servers/Draft1Version2ComputerServers.pdf

[6] K.-D. Lange and M. G. Tricker. The design and development of the server energy efficiency rating tool (SERT). In International Conference on Performance Engineering (Mar. 2011), 145-150. DOI= http://doi.acm.org/10.1145/1958746.1958769

[7] K.-D. Lange. Identifying Shades of Green: The SPECpower Benchmarks, IEEE Computer, V42 #3 2009, 95-97, DOI= http://dx.doi.org/10.1109/MC.2009.84

[8] US EPA ENERGY STAR Computer Servers Draft 1 Version 2.0 - http://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer_servers/Draft1Version2PowerPerformanceDatasheet.pdf

[9] Server Efficiency Rating Tool home page: http://www.spec.org/sert/