The Design and Development of the Server Efficiency Rating Tool™ (SERT)

Michael G. Tricker
Microsoft Corporation
Mike.Tricker@microsoft.com

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

ABSTRACT

According to the United States Environmental Protection Agency (US EPA) almost 3% of all electricity consumed within the US in 2010 goes to running datacenters, with the majority of that powering servers and the associated air conditioning systems dedicated to eliminating the heat they produce. The EPA launched the ENERGY STAR® Computer Server program in May 2009, intended to deliver information to better enable server purchasing decisions based on projected power consumption.

The Server Efficiency Rating Tool (SERT)™ has been developed by the Standard Performance Evaluation Corporation (SPEC) SPECpower committee to address the EPA requirements for Version 2 of the ENERGY STAR server program. Unlike many tools sourced from the SPEC organization the SERT is not intended to be a benchmark, and for Version 2 does not offer a single score model. Instead it produces detailed information regarding the influence of CPU, memory, network and storage I/O configurations on overall server power consumption.

This paper describes the design and development of the SERT, including discussion of the collaborative nature of working with the EPA and the various industry stakeholders involved in the design, review and development process. Many of the core ideas behind SERT were derived from the SPECpower_ssj2008 and other SPEC-developed benchmarks, and this paper illustrates where ideas and code were shared, as well as where new thinking resulted in entirely new solutions. It also includes thoughts for the future, as the ENERGY STAR server program continues to evolve and the SERT will evolve with it.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Reliability, Standardization

Keywords

SPEC, Benchmark, Energy Efficiency, Power Analysis, Server, Datacenter, Energy Star, Environmental Protection Agency (EPA)

ICPE'11, March 14–16, 2011, Karlsruhe, Germany.
Copyright 2010 ACM 978-1-4503-0519-8/11/03…$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 software. It has grown to include 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 structured in four major groups: the Open Systems Group (OSG), the High Performance Group (HPG), the Graphics and Workstation Performance Group (GWPG) and the SPEC Research Group (RG). The OSG includes subcommittees covering major areas of desktop, workstation and server performance and benchmarking. These include CPU, Java, Virtualization, Web and Power, specifically addressed by the SPECpower subcommittee, which has been responsible for creating the SPECpower_ssj2008 benchmark (ssj2008), and now the SERT for the EPA ENERGY STAR program for computer servers.

The EPA has been tracking computer power consumption for several years, and in January 2006 hosted the Conference on Enterprise Servers and Data Centers: Opportunities for Energy Savings. At the end of that year the EPA announced its intention to develop an ENERGY STAR for Enterprise Computer Servers program, with broad industry support and participation. This eventually led to the ENERGY STAR Computer Server specification, which was launched in May 2009. This recommended the use of the ssj2008 benchmark to provide data required to complete the EPA Power and Performance Data Sheet.

Ssj2008 was developed to be the first industry-standard cross-platform benchmark for evaluating the power and performance characteristics of volume and multi-node class servers. It is based on server-side Java workloads, exercising CPUs, memory hierarchies including caches, general Symmetric Multiprocessing (SMP) scalability and many aspects of the Java implementations used in the testing.

2. How the SERT Differs from Existing SPEC Benchmarks

The SERT was never intended to be a benchmark, and although developed by a team with a strong history in benchmarking was always intended to be a tool that delivered a broad set of data derived from the individual tests, rather than a single metric with many digits of precision. Benchmarks encourage the fine tuning of system configurations and parameters with the intention of creating results that can be used by marketing organizations to promote one system at the expense of others. This has created an entire sub-culture of performance experts devoted to the ultimate tuning of their employers configurations, resulting in configurations (and benchmark scores) that in some cases bear little resemblance to anything a real-world customer workload could expect to run on, or reliably and consistently deliver.

The SERT deliberately does not attempt to simulate specific end user workloads, but instead provides a set of focused synthetic worklets that exercise specific aspects of the Server (or System) Under Test (SUT). These worklets have been developed to exercise the processor, memory, network I/O and storage I/O subsystems, and may be combined into various configurations to also run serially or in parallel to provide "system" tests integrated across the different subsystems.

Since the SERT is not a benchmark and is targeted at "as shipped" (or "out of the box") configurations it eliminates the perceived need for fine tuning, which aligns with the objectives laid out by the EPA. The ENERGY STAR Computer Server program specifically requests that test systems be "as shipped" by the manufacturer or system builder for both hardware and software configurations. This allows for only minimal (and comprehensively documented) configuration options that would typically be performed by a system administrator on initial "out of the box" configuration. Such customization might include applying the latest firmware updates, patches or service packs and performing network configuration as recommended by the hardware and software vendors.

The SERT is not intended to provide a single numerical score accurate to several digits. Instead the aim of Version 2 of the ENERGY STAR computer server program is to provide a detailed Power and Performance Datasheet that is available from the server vendor describing each product or product family they offer, and may be used to compare specific aspects of different servers from both the same and different manufacturers. By providing the results for each major subsystem it is easier for potential purchasers to consider how each system could best serve their different workload requirements, e.g., a very good CPU score would be of interest to an HPC workload, but possibly of less interest to a Web server.

Since one easy to configure and use tool cannot hope to address all server architectures, usage models and configurations SPEC has worked closely with the EPA and industry stakeholders to define the basic server classes and configurations that can be tested with the initial release. Subsequent releases of the SERT and the ENERGY STAR computer server program will likely widen the range of supported configurations, e.g., beyond four processors and into more specialized workload-specific configurations.

3. The SERT Development Model

The core of the SERT development methodology is compliance with the well-established SPEC model of industry-wide collaboration, where anyone who is willing to contribute actual effort being welcome to join the SPECpower subcommittee that is responsible for SERT development. The strength of this model derives from the participation of company representatives from varied backgrounds spanning hardware (computers and devices) and software (operating systems, device drivers and programming environments). It is however sometimes difficult to get sufficient development resources to support all the desired features, since the majority of company representatives contribute to SPEC in addition to their "real" paying jobs!

The ongoing architecture and design evolution of the SERT is captured in the SERT Design Document, which is regularly updated and shared with the EPA and thence with the ENERGY STAR Computer Server program industry stakeholders, who can provide feedback and input to SPEC via regular conference calls and reviews organized and mediated by the EPA.

The Design Document includes a subset of the full SERT development schedule as it relates to the public test phases. This includes the Alpha and Beta milestones and related test phases (with the different classes of volunteer requested for each phase), and leading up to the final release or "finalization" milestone. The related Process Document describes the support, servicing and update models for the SERT, and has also been developed in collaboration with the EPA.

Many of the EPA's ENERGY STAR stakeholders are already represented within SPEC, but the EPA is also targeting smaller providers of server systems as well as the large (and well-known) Original Equipment Manufacturers (OEMs). It is therefore important that the smaller Value Added Resellers (VARs) and "white box" builders also get the opportunity to contribute their feedback to the overall design process. Their input regarding the eventual testing is particularly important since they typically have far fewer resources at their disposal than the major OEMs, and thus are considerably more constrained in the configurations they can test and the dedicated test hardware (and actual test engineers) they can afford to provide.

Although the development and review process is intended to be as open as possible it was quickly decided that distributing the source code for the SERT would be inadvisable. Were source code to become generally available there is a risk of individuals accidentally (or perhaps even deliberately) providing "customized" versions that could replace those shipped as part of the SERT package.

Such customized derivatives might lead to very different results being reported back to the EPA, leading to additional work on the part of both the EPA and SPEC in ensuring that all reported results come only from officially sanctioned SERT distributions that are available to all stakeholders. The decision was taken early in SERT development that anyone could request a code review of any part of the SERT source, but that the code itself would not be shared outside the SPECpower subcommittee and the EPA itself.

It was also decided early on that a clearly defined set of hardware and operating system platforms would be supported, based on membership of the SPECpower subcommittee and the willingness of those members to actually commit resources to enable porting, testing and on-going support. It was decided that only 64-bit operating systems would be supported, as this deliberately restricts how many releases of operating systems that were available prior to the completion of SERT would be supported.

Such constraints were put in place due only to limited resources availability, and not due to any technical constraints on the SERT framework or any of the worklets themselves. SPEC remains open to adding further support if the resources for development, testing and ongoing support were to become available thanks to other companies joining the SPECpower committee.

4. The SERT in Detail

4.1 Hardware Configuration

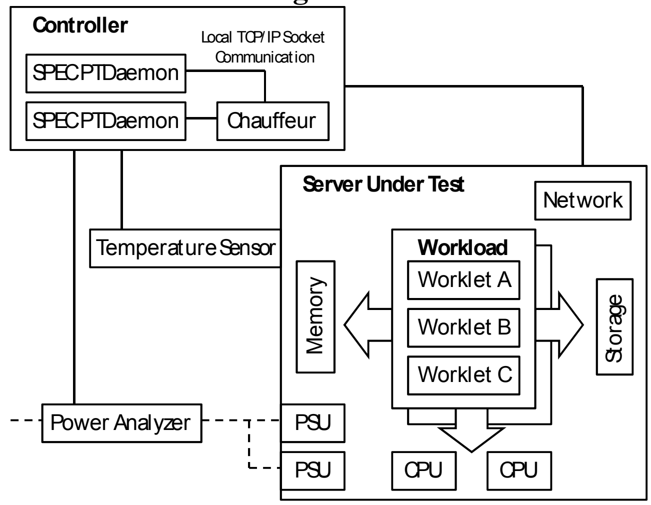

Figure 1. Configuration Layout

4.2 SERT Components

Because the SERT has been developed by the original creators of ssj2008 they share some concepts and even source code, whilst differing in other key areas. For example the Power and Temperature Daemon (PTDaemon or PTD) was developed for ssj2008 to coordinate between power analyzers, temperature probes and the ssj2008 test harness itself. It is used unchanged by the SERT for the same purpose.

The fundamental building block of the SERT is a **worklet**, which comprises a set of one or more **transactions,** which may be initiated or executed by a **user**. A worklet may also be considered to be a sub-workload, or possibly a very small workload with extremely limited interactions. In benchmark terminology it would be described as a micro-benchmark.

A user represents an external agent that can request or initiate work, such as a human. Each user may include unique identifying information, and can also maintain and record state across the life of one or more transactions. There may be multiple types of user, more than one of which may be associated with a single worklet. For example an online retailer would handle many unique users, but may also have different classes of users depending on their purchasing history or willingness to share financial data to qualify for discounts.

A transaction is an operation that takes input such as a user and some initial parameters, performs some processing and delivers an output or result. Some transactions may be able to perform validation on some combination of their inputs and outputs for auditing purposes. For example a transaction may simulate a database update, or a page request from a Web site.

A worklet is a set of transactions associated with a particular type or instance of user. A worklet may represent a sequence of user interactions, such as requesting and updating a database record, as part of selecting and purchasing an item from an on-line retailer via their Web site. Operations spanning multiple transactions may have "think time" gaps in between transactions, which will alter the load patterns on the SUT.

The SERT supports the concept of an **interval**, in which worklets are scheduled for execution by a user. Each interval includes pre-measurement, recording and post-measurement time periods. Every worklet is scheduled for subsequent execution, at which point the worklet iterates through all its component transactions, submitting each in turn to a per-Java Virtual Machine (JVM) thread pool. As one transaction completes, the next is submitted to the thread pool.

In SERT terminology a **workload** is defined by a set of worklets and their associated users. Workload execution comprises several phases, from warm-up through calibration to one or more measurement phases. Each phase is a sequence of measurement intervals, and multiple measurement phases may be used for different mixes of users, transactions and so on.

There is one other special case worklet that needs to be described, which is the **modifier**. In some cases it may be impossible to test a piece of hardware, or the hardware in question may not vary its power consumption (or has such low consumption) that developing a test is not strictly warranted. In such a case a software modifier may be substituted, which effectively simulates the power consumption of the hardware without ever actually touching that hardware.

4.3 Configuration

The SERT inherits many concepts from the ssj2008 benchmark, including the use of the PTDaemon, a Reporter and a cross-machine Director. As mentioned above the PTDaemon is used to coordinate the inputs from one or more power analyzers, and temperature sensors used to measure server input air temperature. It has grown beyond its original purpose as part of ssj2008 and is also used by other SPEC benchmarks, including the SERT.

The component equating to the ssj2008 Director has become part of Chauffeur, which runs on both the Controller and the SUT. Chauffeur coordinates operations between the various components running on both systems, and the Reporter is used to format the raw results produced by the worklets (including the original input from the GUI) into XML, HTML or even (eventually) PDF for publication via the EPA's Power and Performance Datasheet.

The SERT GUI allows the test engineer to input configuration information describing both the SUT and the Controller. At the same time the engineer runs an information gathering agent on the SUT, querying and reporting system configuration data such as the number, core count and speed of the processors, cache hierarchy and sizes, amount of system memory and number of DIMMs present in which slots, the number and types of various IO devices installed and so forth.

This system information is combined with the configuration information entered manually into the GUI by the test engineer, together with the worklets selected for execution and their initial configuration parameters (as appropriate) and exported as an XML file. This file may be passed directly to the Chauffeur controller process. It may also be saved to a file for subsequent test runs with the same hardware configuration, or may be shared across multiple SUTs as the basis for their configuration files. This helps eliminate some of the simple but annoying input errors that have been observed from users of ssj2008.

Once the configuration information has been passed to Chauffeur each worklet is run based on its configuration and initial parameters, and the sequence of serial or parallel operations specified by the configuration file. Chauffeur supports the affinitization of JVMs to specific sockets or cores, and provides the option of running all the worklets in a single JVM or of starting up a new JVM for each worklet, which is then deleted on completion.

Each worklet provides its results to the Reporter for correlation and formatting. Worklets also typically do a certain amount of initial condition validation, to ensure they are not executing with invalid parameters or an inappropriate configuration.

To ensure valid results there is a warm-up phase at the start of each worklet execution phase to allow the system to settle and for power consumption to stabilize. This is then followed by one or more calibration phases, in which the maximum performance of the worklet is calculated to provide the 100% load point, which is used in the subsequent measurement phase. Measurements are then run, the results are gathered and finally there is a cool down phase after which the worklet terminates and the JVM may (optionally) be deleted.

4.4 Worklets

4.4.1 CPU

The CPU worklets include a mixture of integer and floating point mathematical calculation tasks, together with some text-oriented operations such as the XML validation. The intention is to provide tests that are effectively cache resident, minimizing the amount of IO bus traffic to main memory. By minimizing the number of main memory reads and writes once the worklet is running it effectively forces the majority of the work onto the actual CPU (or more than one CPU as appropriate), so providing realistic power consumption values.

The worklets are multi-core friendly without having to make explicit use of more than one hardware thread. Since the majority of the worklets are written in Java being able to affinitize a JVM to a specific set of cores or sockets is important.

4.4.2 Storage IO

Storage IO worklets are targeted at either solid state or rotating media that is present within the server enclosure. There is no intention of supporting external media in the first release of the SERT, so Fiber Channel, iSCSI and Fiber Channel over Ethernet are not addressed. The worklets are designed to prevent the data being read or written from becoming cache resident, by forcing all IOs to actually touch the actual media.

Storage IO is a highly critical aspect of modern server workloads, with as many (or even more) opinions as there are vendors selling storage-related products. Rather than become mired in diverse opinions the SERT team decided to focus on the basics: mixtures of random and sequential reads and writes of various sized buffers to and from large numbers of small files up to small numbers of enormous files. This sets a strong baseline of functionality that future versions of the SERT can build on.

4.4.3 Memory

The memory worklets are derived from CPU worklets, but using as much physical memory as possible, and minimizing the opportunity for the test data to become cache resident. A typical example is for a processing intensive worklet to use a large in-memory look-up table instead of calculating a result on the fly. Although this requires an initial “fill memory with data” phase in the worklet it is a very effective way of using memory, so long as the queries are broadly distributed and thus do not continually hit the cache.

4.4.4 Network IO

The Network IO worklet poses a particular challenge due to the EPA’s desire to minimize the amount of hardware needed for testing purposes, over and above the actual SUT and controller systems and the power analyzer(s) and temperature sensor(s). To realistically stress network IO requires (especially with multiple and multi-port NICs now commonplace on a broad range of servers) considerable external hardware, including one or more other servers at least as powerful (and fully featured) as the SUT, together with one or more switches and potentially even more.

It was also observed in initial prototyping that a NIC used surprisingly little more power when running at 100% capacity than when it was actually idle. It was observed that the CPU used more power due to the device driver but the actual NIC consumption only varied by less than a watt. After this was explained to the EPA they agreed that a modifier could be used to handle the networking IO load in the first release of the SERT.

4.4.5 System (or Integration) Tests

Since every class of worklet has been stand-alone, focused on testing one aspect of the server hardware it was suggested that some form of combined test be developed in which multiple worklets could be run serially or in parallel to more closely approximate to a real-world workload. There has also been a requirement from the EPA to get industry input on idle state power consumption before the SERT was completed, so a system test has been built derived from ssj2008 which can run under the Chauffeur test harness. This test is likely to remain in the final SERT distribution.

4.5 Supported Hardware

Version 1 of the ENERGY STAR program supports servers with up to four processors, and for Version 2 the EPA decided to extend the program. The SERT has therefore been designed to support arbitrarily large servers, but for the first release is again targeted at up to four processors and has also added support for Blade Servers and Multi-Node servers, which include two or more independently booted nodes (each with an independent operating system instance) within a single physical enclosure.

The SERT does not explicitly support virtualization, meaning that it expects to be executed on physical (rather than virtualized) hardware. Although virtualization is becoming more commonplace it was felt that the first release of SERT will do a good job of simulating the sorts of loads imposed on the hardware when running at high load. Given additional development resources this is likely to be revisited in subsequent releases.

4.6 Implementation Languages

One of the primary considerations in the development of the SERT was cross-platform support. Many hardware and software companies participate in SPEC, and most of the major OEMs are represented within the SPECpower subcommittee, so it is critical that the SERT support as many operating systems and hardware architectures as possible. Java is a popular choice because many

SPEC benchmarks have some or all components developed in Java, so there is a considerable body of code that may be incorporated into new tests as appropriate.

However, Java is not appropriate for all tasks, so it was decided that C and C++ should also be supported. Chauffeur is capable of supporting worklets written in C or C++, and some Java worklets call out to libraries developed in C via the Java Native Application (JNA) interface to access lower level operating system APIs that do not map onto Java classes.

There are also a few cases where worklets make use of the libraries supplied with JVMs to avoid having to implement some standard functionality such as compression and encryption algorithms. Such algorithms may have legal ramifications in some countries, so by using whatever libraries are appropriate in those markets the level of legal review required is significantly reduced.

5. Conclusions

The SERT is still evolving, and is expected to go live as part of the EPA's ENERGY STAR Computer Server version 2 program in 2011. Using the team and the experience gained during the development of ssj2008 has enabled SPEC to develop a tool that is unique in the industry today. By working collaboratively with the EPA and ENERGY STAR industry stakeholders the SERT has targeted all sectors of the server market, from the OEMs with international presence to small VARs and white box builders.

By not trying to emulate real-world customer workloads and avoiding the specialization that can accompany a benchmark the SERT offers the range of power usage data that server buyers actually require to support environmentally conscious purchasing decisions. The level and quality of industry participation in specification and design reviews and actual development enables the SERT to avoid suggestions of favoritism, producing a tool that builds on the international credibility and track record of the SPEC organization.

As the power footprint of servers and datacenters becomes an increasing environmental and political issue in more countries, tools like the SERT will be required by more agencies, with more customers looking to them for business-critical data. The SERT will continue to evolve, supporting more classes of server and types of hardware, positioning both the SPEC organization and the EPA for international leadership roles in power usage reduction and enabling customers to make well informed purchasing decisions.

6. Acknowledgement

The authors would like to acknowledge the current members of the SPECpower subcommittee who have contributed to the design, development, testing and overall success of the SERT: Paul Muehr, Greg Darnell, Hansfried Block, Christian Koopmann, Karl Huppler, Jeremy Arnold, Sanjay Sharma, and David Ott. SPEC and the tool and benchmark names SERT and SPECpower_ssj2008 are registered trademarks of the Standard Performance Evaluation Corporation.

7. References and Links

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

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

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

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

[5] ENERGY STAR Computer Specification Version 1.0 Power and Performance Data Sheet: http://www.energystar.gov/ia/partners/prod_development/new_specs/downloads/servers/Final_Datasheet.xls

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

[7] ENERGY STAR Computer Servers Draft 1 Version 2.0 Power and Performance Datasheet: http://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer_servers/Draft1Version2PowerPformanceDatasheet.pdf