ETSI TR 104 004 V1.1.1 (2024-09)

TECHNICAL REPORT

Environmental Engineering (EE); Processor power management functionality of servers

Reference

DTR/EE-EEPS60

Keywords

server

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00   Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Important notice

The present document can be downloaded from the ETSI Search & Browse Standards application.

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format on ETSI deliver.

Users should be aware that the present document may be revised or have its status changed, this information is available in the Milestones listing.

If you find errors in the present document, please send your comments to the relevant service listed under Committee Support Staff.

If you find a security vulnerability in the present document, please report it through our Coordinated Vulnerability Disclosure (CVD) program.

Notice of disclaimer & limitation of liability

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.
No recommendation as to products and services or vendors is made or should be implied.
No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.
In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

Copyright Notification

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Environmental Engineering (EE).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Executive summary

CPU Power Management functions (P-states and C-states or their equivalent) offer hardware available, software initiated functions to reduce CPU voltage and frequency when workload demands are low or absent to reduce server energy use. Depending on the amount of time a server operates below 25 % utilization or is idle, with longer periods generating higher energy savings, power demand can be reduced by up to 50 % with the full implementation of P-states and C-states. These power demand reductions improve SERT overall efficiency scores by up to 30 %. Power management functions come at a cost of reduced server performance and increased response time (exit latency). Data from SERT measurements on a single configuration with power management turned on resulted in performance reductions at the 100 % utilization level in the Hybrid ssj worklet of 6 % for an EPYC™ CPU and 14 % for a Power9™ CPU compared to power management off. For response time, the literature shows that P-state transition times and C-state exit latencies cause response delays that are problematic for some workloads and applications. As an example, Operating System (OS) managed power management profiles can interfere with virtualization programs like VMware® impeding performance rates on the virtualized images. Power management can be a benefit or a problem depending on the data centre operations, workloads and applications. While the implementation of a power management profile can be beneficial in many instances, the optimum P-state and C-state settings and the choice of controlling software - BIOS, hypervisor or operating system - will depend on the specific use case. In some cases, such as high-speed financial trading, network providers or high-performance computing, power management functions will need to be turned off to ensure the performance and response times demanded by those workloads.

Introduction

Data centre energy consumption and IT equipment efficiency are increasingly the focal point of data centre customers and operators, non-governmental organizations and government regulators with an interest in data centre operations. There are a host of options and approaches to reduce data centre energy consumption that can focus on the individual IT equipment, the data centre IT system, or the facility equipment. As servers are the majority of the IT equipment in a data centre and have a validated energy efficiency metric, the SPEC SERT suite, server energy efficiency requirements have attracted the most attention of regulatory efforts to improve data centre energy efficiency.

Server processor power management features, which can moderate power demand based on factors such as CPU utilization, provide a means for server manufacturers and/or data centre operators to reduce server power consumption and improve server energy efficiency. In many CPU architectures, the primary means of server processor power management is through implementation of CPU P-states and C-states. Enablement of power management states can reduce performance and increase response time. The present document will focus on the characteristics, benefits and limitations of CPU power management features in deployment situations.

Server power management features are also available for system fans, memory, storage, Graphic Processing Unit and I/O components. Their combined implementation introduces greater complexity and requires a higher level of integration with the server's platform firmware and operating system. Component power management is beyond the scope of the present document. More specifically the present document will provide a brief overview of CPU P-states and C-states, SERT test data detailing the power demand reductions achieved by different power management implementations, the SERT measured active state energy efficiency improvements, and the performance and exit latency impacts of server power management features. The present document only addresses the current state of these technologies.

1 Scope

The present document is focused on addressing the characterization of the process power management functionality of servers.

The processor power management of servers is limited to those within scope of Commission Regulation (EU) 2019/424 [i.1].

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] Commission Regulation (EU) 2019/424 of 15 March 2019 laying down ecodesign requirements for servers and data storage products pursuant to Directive 2009/125/EC of the European Parliament and of the Council and amending Commission Regulation (EU) No 617/2013.

[i.2] C. Chou, L. N. Bhuyan and D. Wong: "µDPM: Dynamic Power Management for the Microsecond Era", 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), Washington, DC, USA, 2019, pp. 120-132, doi: 10.1109/HPCA.2019.00032.

[i.3] I. Curtis: "Hot Chips 2020 Live Blog, Next Gen Intel Xeon Ice Lake-SP", Hot Chips Conference, August 17, 2020.

[i.4] D. Molka, R. Schone, M. Werner: "Wake-up latencies for processor idle states on current x86 processors", Computer Science - Research and Development, Vol. 30, doi: 10.1007/s00450-014-0270-z; 2014/04/01.

[i.5] S. Kanev, K. Hazelwood, G. Wei and D. Brooks: "Tradeoffs between power management and tail latency in warehousescale applications", 2014 IEEE International Symposium on Workload Characterization (IISWC), Raleigh, NC, 2014, pp. 31-40, doi: 10.1109/IISWC.2014.6983037.

[i.6] Broadcom®: "Virtual machine application runs slower than expected in ESXi".

[i.7] Lenovo™: "Tuning VMware for Increased Performance - Lenovo ThinkSystem and x86 Servers".

[i.8] Huawei: "Huawei V5 Server Best Practice with VMware ESXi System 03".

[i.9] HPE: "Workload profiles and performance options | UEFI System Utilities User Guide for HPE Compute servers".

[i.10] The Green Grid: "SERT™ Active Efficiency: Demonstrating How SERT™ Active Efficiency Testing Includes Server Idle".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

Basic Input/Output System (BIOS): firmware that provides basic boot capabilities for a platform

C-state: processor power consumption and thermal management state within the global working state

Hybrid ssj: one of the 7 CPU worklets used in SPEC® SERT® suite

P-state: power consumption and capability state within the active/executing states, C0 for processors

NOTE: Performance states allow Operating System direct Power Management (OSPM) to make tradeoffs between performance and energy conservation.

power management profile: combination of P-state and C-state settings designed to reduce power consumption while addressing the specific operating needs of a customer environment or workload

Server Efficiency Rating Tool (SERT): performance and power software measurement tool created by the SPEC benchmark standards consortium

NOTE 1: SERT was specifically designed for use in government sponsored server energy efficiency programs.

NOTE 2: SERT has components that run on the system under test and a controller system, and interfaces with a power analyser connected between the electrical socket and server power supply.

NOTE 3: Detailed performance and power data is collected while running server worklets at different load level, and these measurements are combined into an overall weighted server energy efficiency score.

worklet: parts of a *workload* consisting of specific code sequences which are executed during testing

workload: group of *worklets* which share common attributes and are combined into an overall result

NOTE: SERT includes CPU, Memory and Storage workloads.

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

BIOS	Basic Input/Output System
CPU	Central Processing Unit
DPM	Dynamic Performance Mode
HDD	Hard Disk Drive
OEM	Original Equipment Manufacturer
OS	Operating System
OSPM	Operating System direct Power Management
QPS	Queries Per Second
SERT	Server Efficiency Rating Tool
SKU	Stock Keeping Unit
SPS	Static Power Savings

4 CPU Power Management

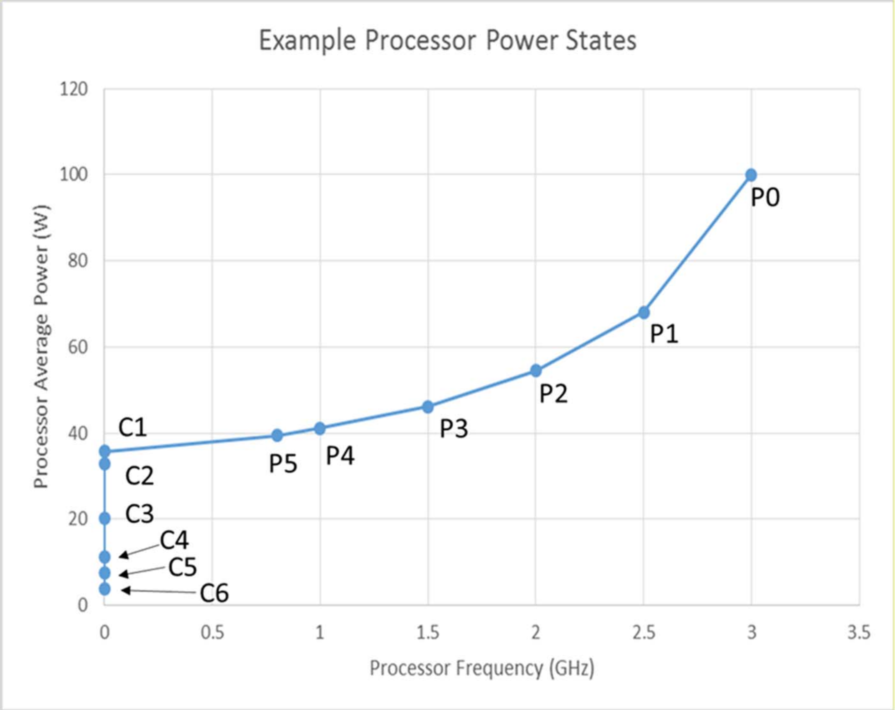
4.1 Overview of CPU P-states and C-states

CPU power management is executed through implementation of P-states and C-states. P-states modulate the power consumption of the CPU as a function of workload demand. An accepted method for implementing P-states is to reduce the voltage and frequency of the individual cores or groups of cores while the processor is executing instructions. P-states consist of predetermined operating frequencies and associated operating voltages which are selected to match the processor work capability to the current workload applied to the server. C-states significantly reduce the processor voltage during periods where the processor has no work to perform. During no work periods, C-states sequentially reduce the power demand in time defined steps.

For the purposes of the present document, a specific power management profile is a combination of P-state and C-state settings designed to reduce power consumption while addressing the specific operating needs of a customer environment or workload. Power management profiles implemented in production servers may contain any number of processor or system power management settings. The primary method of reducing power during idle periods is turning off portions of the CPU which is realized with a penalty of increasingly longer latencies before the processor can begin executing instructions (Figure 8).

Figure 1 shows the power and frequency curve for different levels of P-state and C-state in an Intel® Xeon® x86 processor. AMD x86 processors have similar P-states and C-states with power levels specific to their architecture. Servers using the IBM® POWER9™ processor and Linux® operating system have a similar profile, while IBM POWER9 and PowerVM systems have a different profile. See Annex A for details.

NOTE 1: Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.



NOTE: This is for illustration only. Actual processor power will vary by processor SKU.

Figure 1: Processor Power Graph representing the available power states for each individual core for typical processor architectures

Since P-states and C-states are independent features, their usage policies can be combined in an infinite number of combinations. Server manufacturers typically provide 2 or more standard power management policy implementations (profiles) which can be selected in the OS or the BIOS: maximum performance, maximum power savings and one or more intermediate implementations designed to balance power savings against increased response time and latency. Data centre operators can, if they desire, 'tune' their P-state/C-state policies to fit their specific workload profile(s) through a testing routine similar to performance optimization tuning for their workload(s).

NOTE 2: The Green Grid Server published a report in March 2018, "SERT™ Active Efficiency: Demonstrating How SERT™ Active Efficiency Testing Includes Server Idle" [i.10], which can be consulted for a more detailed discussion of the CPU power management function capabilities and benefits.

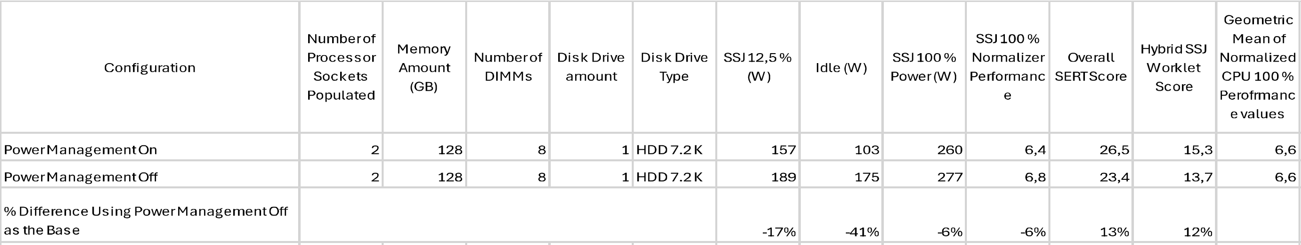
4.2 Impact of P-state and C-state enablement on server power demand

To evaluate the impact data was collected on SERT power and performance with processor power management turned on and off for three sets of server configurations: one with an AMD EPYC™ 7262 processor; in addition to two other similar configurations from two manufacturers, an Intel® Xeon® x86 Silver 4109T processor; and one with an IBM® POWER9™ processor with PowerVM®. For purposes of comparison between configurations, the Hybrid ssj normalized performance and power demand measurements are graphed for each configuration set.

To evaluate the performance/power benefits of power management, the overall SERT score for the EPYC and Intel Xeon x86 processors-based servers and the Hybrid ssj worklet efficiency score for the EPYC, Silver 4109T and POWER9 processor-based server are provided. The POWER9 SERT test only ran the Hybrid ssj worklet. As a result, the Hybrid ssj worklet score will be used as a proxy for the overall SERT score. The data for the three configuration sets shows that the implementation of power management features reduces the power demand and the performance and increases the overall SERT score.

This clause provides information on the server configuration, key SERT measurements and calculated values, and Hybrid ssj power to performance graphs for the three configuration sets followed by the 4 key conclusions that can be drawn from a review of the data. A more detailed analysis of the data is provided in Annex A.

Table 1: Configuration details and key power, performance and efficiency score metrics for the EPYC server



Configuration	Number of Processor Sockets Populated	Memory Amount (GB)	Number of DIMMs	Disk Drive amount	Disk Drive Type	SSJ 12,5 % (W)	Idle (W)	SSJ 100 % Power (W)	SSJ 100 % Normalizer Performance	Overall SERT Score	Hybrid SSJ Worklet Score	Geometric Mean of Normalized CPU 100 % Perofrmance values
Power Management On	2	128	8	1	HDD 7.2 K	157	103	260	6,4	26,5	15,3	6,6
Power Management Off	2	128	8	1	HDD 7.2 K	189	175	277	6,8	23,4	13,7	6,6
% Difference Using Power Management Off as the Base						-17%	-41%	-6%	-6%	13%	12%	

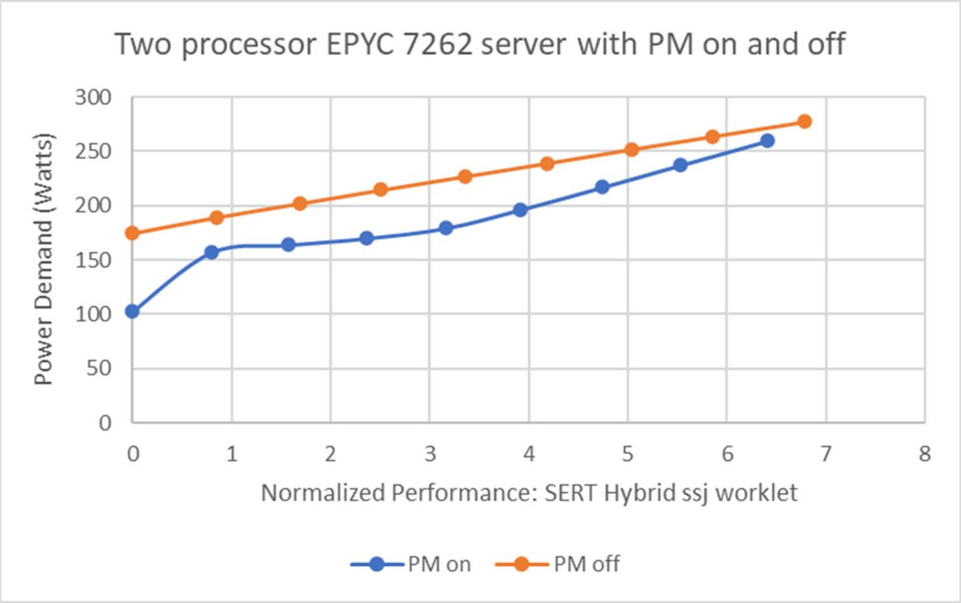
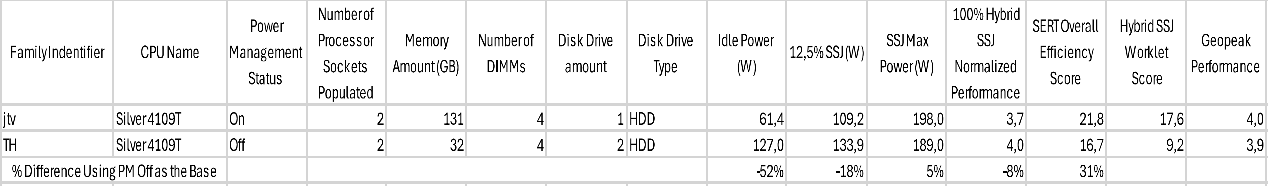


Figure 2: Hybrid ssj interval normalized performance vs. power demand for an EPYC based server

Table 2: Configuration details and key power, performance and efficiency score metrics for two server configurations with Silver 4109T CPU



Family Indentifier	CPU Name	Power Management Status	Number of Processor Sockets Populated	Memory Amount (GB)	Number of DIMMs	Disk Drive amount	Disk Drive Type	Idle Power (W)	12,5% SSJ (W)	SSJ Max Power (W)	100% Hybrid SSJ Normalized Performance	SERT Overall Efficiency Score	Hybrid SSJ Worklet Score	Geopeak Performance
jtv	Silver 4109T	On	2	131	4	1	HDD	61,4	109,2	198,0	3,7	21,8	17,6	4,0
TH	Silver 4109T	Off	2	32	4	2	HDD	127,0	133,9	189,0	4,0	16,7	9,2	3,9
% Difference Using PM Off as the Base								-52%	-18%	5%	-8%	31%		

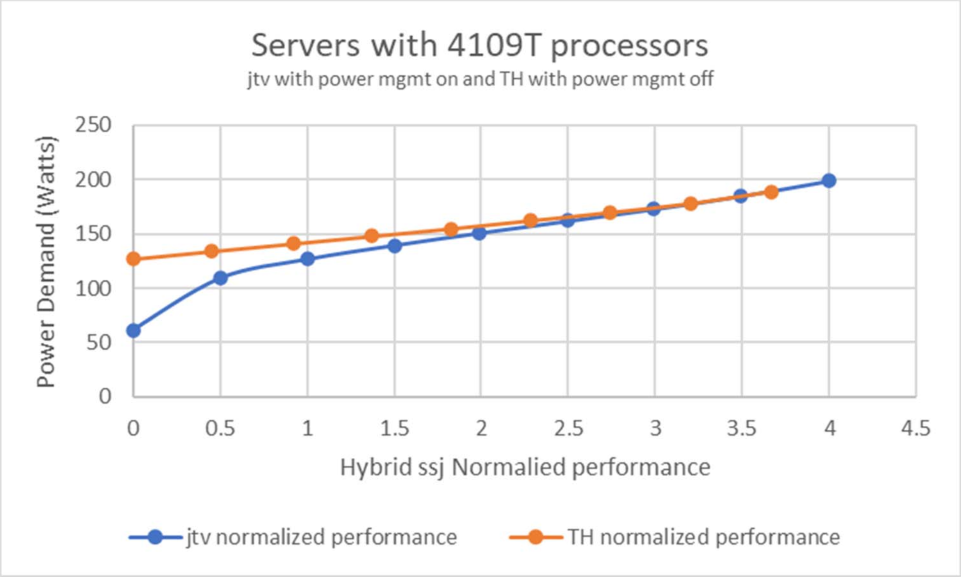
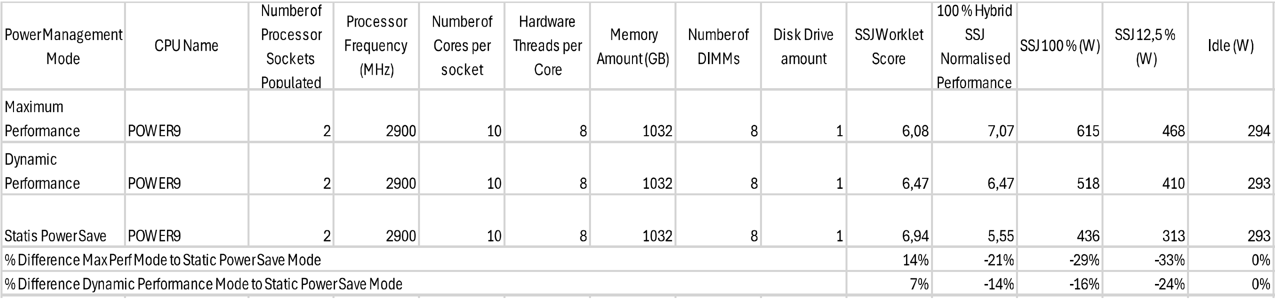


Figure 3: Hybrid ssj interval normalized performance score vs. power demand for two Silver 4109T based Servers with different power management settings

Table 3: Configuration details and key power, performance and efficiency score metrics for the Power9 Server



Power Management Mode	CPU Name	Number of Processor Sockets Populated	Processor Frequency (MHz)	Number of Cores per socket	Hardware Threads per Core	Memory Amount (GB)	Number of DIMMs	Disk Drive amount	SSJ Worklet Score	100 % Hybrid SSJ Normalised Performance	SSJ 100 % (W)	SSJ 12,5 % (W)	Idle (W)
Maximum Performance	POWER9	2	2900	10	8	1032	8	1	6,08	7,07	615	468	294
Dynamic Performance	POWER9	2	2900	10	8	1032	8	1	6,47	6,47	518	410	293
Statis Power Save	POWER9	2	2900	10	8	1032	8	1	6,94	5,55	436	313	293
% Difference Max Perf Mode to Static Power Save Mode									14%	-21%	-29%	-33%	0%
% Difference Dynamic Performance Mode to Static Power Save Mode									7%	-14%	-16%	-24%	0%

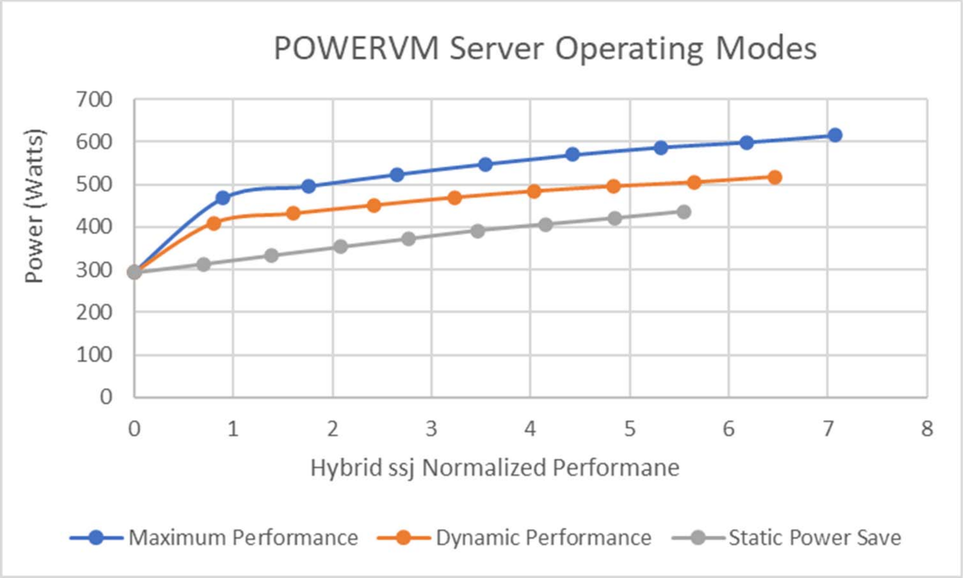


Figure 4: Hybrid ssj interval normalized performance score vs. power for a POWER9 PowerVM Server

A review of the measurement data and calculations provided above indicates the following benefits of enabling power management on a server:

1) Enablement of power management results in a higher overall SERT score or Hybrid ssj worklet efficiency scores of 6 % to 30 % for the three processor types (Figure 5). The reduction in power achieved through the P-state and C-state enablement results in an increase in the SERT score over and above any score reduction caused by lower maximum performance values.

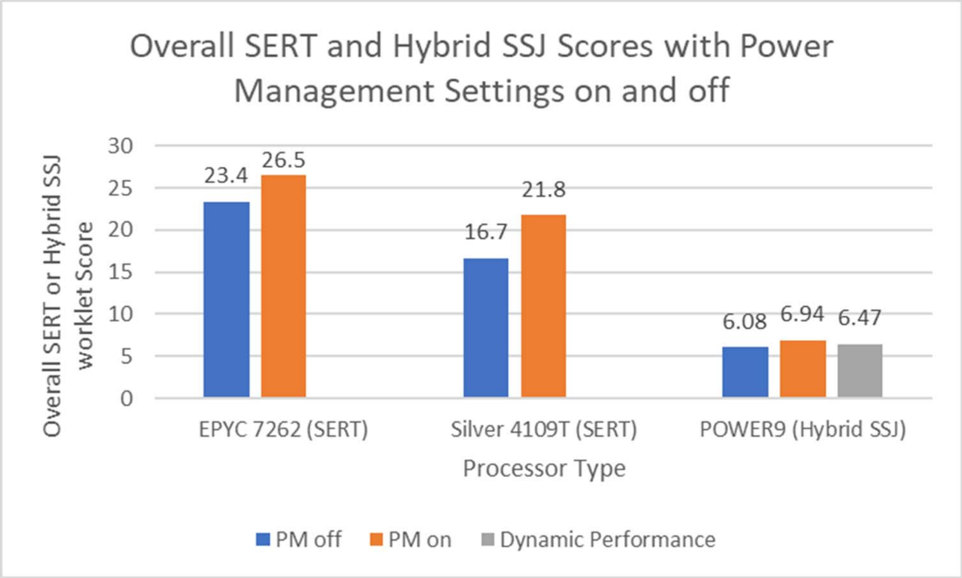


Figure 5: Impact of power management settings on Overall SERT and Hybrid ssj worklet score

2) Enablement of C6 state results in a significant reduction in the idle power. For the EPYC 7262 and Silver 4109T processors, use of the C6 state reduces idle power by 41 % and 52 % respectively. These lower idle power values come at the expense of higher latencies and lower response times when recovering from the idle state (see next clause).

3) The data for the two Silver 4109T processor configurations, which are designed and manufactured by two OEMs using the same general set of components, demonstrate that a manufacturer's firmware and method of integrating components can enhance system performance. Even with power management on, one tested server configuration (jtv) has higher maximum worklet performance values than another tested server configuration (TH) (with the exception of the SOR worklet, Table A.1, Annex A). The overall SERT score is a robust metric capable of differentiating the energy efficiency of server designs and system differences.

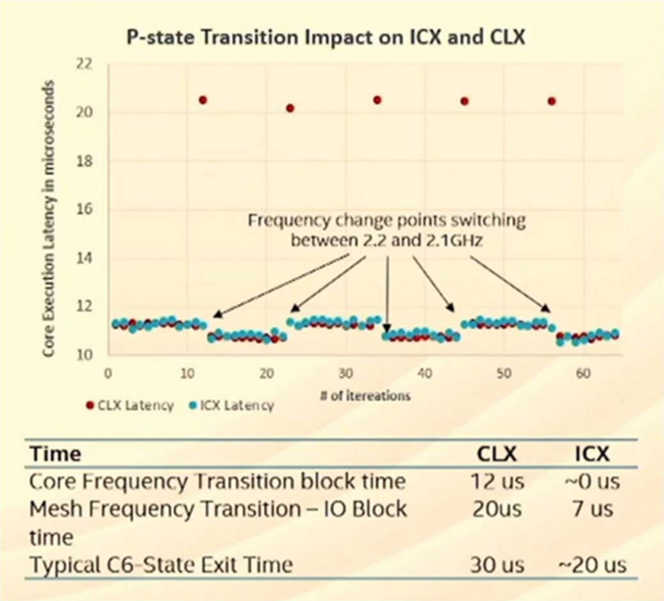
4) The POWER9 and EPYC server measurement data illustrates that voltage and frequency control can result in reduced system performance, but the overall SERT or available worklet efficiency score is increased. In the case of the POWER9 processor installed in a PowerVM server, the DPM and SPS modes make specific voltage and frequency reductions reducing the 100 % hybrid ssj normalized performance by 14 % and 21 % respectively. The EPYC based server shows a 6 % reduction in the 100 % hybrid ssj normalized performance with power management on. The POWER9 hybrid ssj worklet score increases by 7 % and 14 % compared to the Maximum Performance Mode for the DPM and SPS modes respectively, and the overall SERT score on the EPYC server increases by 13 %.

4.3 Latency impacts resulting from power management enablement

Power management enablement results in latency impacts and slowed response time due to the transition times between P-states and the time required to enter and exit C-states [i.2]. For data centre operators and users, exit latencies on the order of 5 microseconds are the maximum response delay that sensitive customers, such as high-performance computing, network providers (Telco), banking and trading, can tolerate. The exit latency associated with the C6 power management setting is the absolute limit for almost all applications, which is why C-states of C7 or higher are not typically deployed on server systems.

P-state transition time is a function of the CPU design and may be further modified by the OEM manufacturer where a standard power management profile is selected for a server, or by the data centre operator where a purpose-built profile is used. Figure 6 provides a comparison of 2nd Generation Intel® Xeon® Scalable Processor (aka Cascade Lake or CLX) and 3rd Generation Intel® Xeon® Scalable Processor (aka Ice Lake or ICX) P-state and C-state latency.

In Figure 6, the two processors experience a core frequency switch at approximately every 12 iterations. The frequency shift represents a change in P-state, where the frequency and/or voltage will change to match power demand to changes in workload. For the Cascade Lake core, the frequency shift causes a longer core execution latency time of approximately 21 microseconds (the higher valued red data point) as compared to the Ice Lake core which experiences no additional latency in the transition. The discontinuity in the latency measurement points indicate a change in the core frequency from 2,2 GHz to 2,1 GHz and back again. The data in table below the graph provides additional information about mesh frequency transitions and C6 exit time which are not shown on the graph.

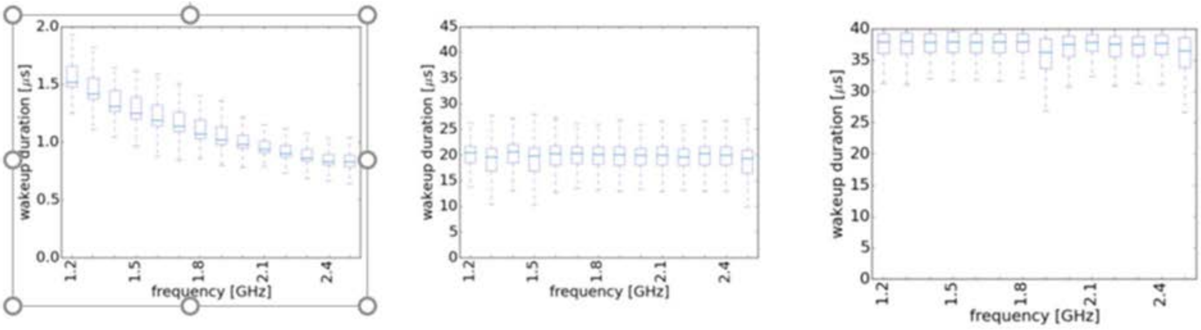


Time	CLX	ICX
Core Frequency Transition block time	12 us	~0 us
Mesh Frequency Transition – IO Block time	20us	7 us
Typical C6-State Exit Time	30 us	~20 us

Figure 6: P-state transition time on Intel® Xeon® Cascade Lake and Ice Lake CPUs [i.3]

Figure 6 shows that the P-state transition (graphic) and C6 exit latency (table only) times are reduced when moving from the second to third generation Intel® Xeon® x86 scalable processors. Figure 7 shows the first generation Intel® Xeon® Scalable Processor (aka Sandy Bridge) C6 exit latency times (wake-up duration).

These latency times are longer than the Cascade Lake times, further illustrating the reduction in C-state latency times achieved with newer CPU technologies. It should be noted, the exit latency times show in Figures 6 and 7 are for the CPU or core only. The exit latency for the overall server system will be longer, as shown in Figure 8, due to the additional time needed to energize or start-up system devices not part of the CPU.



Sandy Bridge EP Remote Active CC1 Halt

Sandy Bridge EP Remote Idle C3

Sandy Bridge EP Remote Idle C6

Figure 7: Measured C-state latencies for a Sandy Bridge X86 processor [i.4]

Figure 8 shows the exit latencies for the C1 to C6 states illustrating the relative time delays between the exit latency for the C1 to C6 states. Though the source does not specifically identify the processor type, the target residency and exit latency values are representative values for server CPUs. Where C-states are applied, production servers typically use some combination of CC1 (C1), C1.1, C3 and C6.

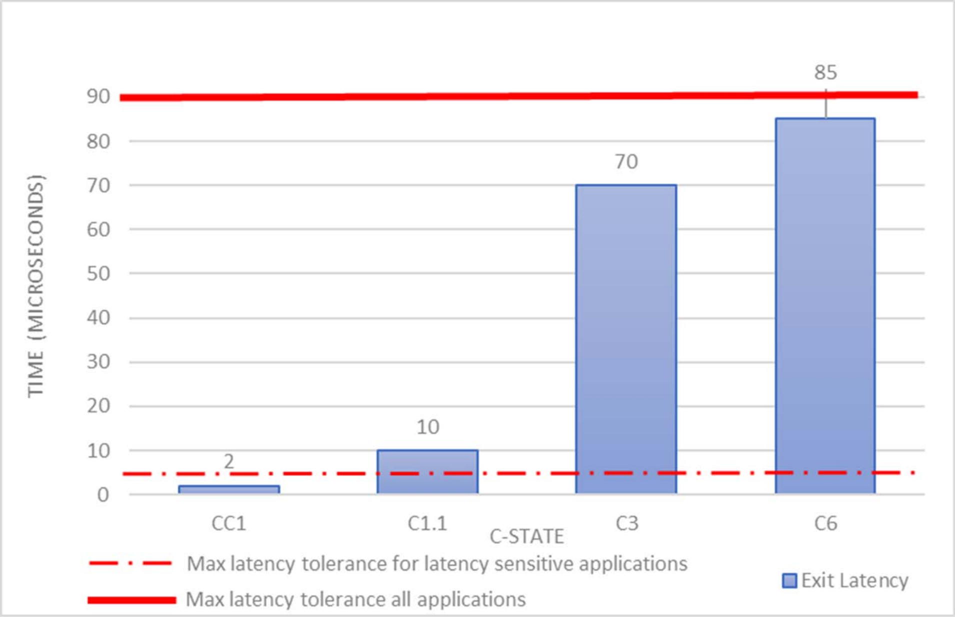


Figure 8: System Exit Latency of Server C-states [i.5]

Figure 9 provides some further insights into the impacts of the exit latency. When there are a small number of transactions or queries, the CPU will reside in C6 more often, increasing the average exit latency. As the transaction or query rate increases, the CPU cores spend progressively less time in C6 and more time in active, C1, C1.1 and C3 states, decreasing the average exit latency. Figure 9 illustrates two key points:

1) Moving into progressively deeper C-states, requires progressively longer periods when there is no work. If the core or CPU goes into C6 state, there is a significant work execution recovery time should the workload demands suddenly escalate. Some or all of these latency delays are unacceptable for certain workloads [i.5].

2) For workloads with very short bursts of less than 100 microseconds, the opportunity to enter into C3 and C6 will be limited, in turn limiting the power management benefits.

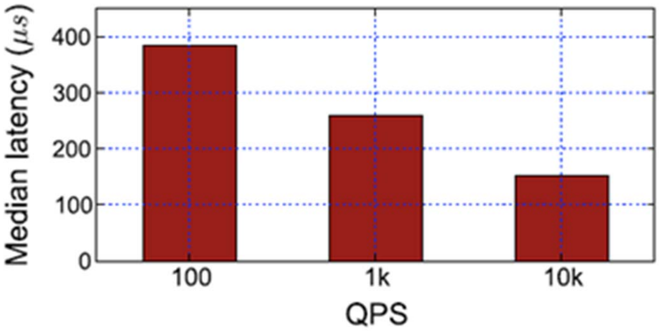


Figure 9: Roundtrip Latency degradation for RPC transport layer for varying Queries Per Second (QPS) [i.5]

The data presented in this clause illustrates that while power management functions can reduce server power demand and improve the SERT active efficiency score, there is also latency/response time and performance degradation penalty associated with power management enablement. Some data centre operations and applications cannot tolerate a latency delay around 5 microseconds while other operations and applications cannot tolerate the exit latency time associated with the C6 state. In these cases, P-state transition times and C-state exit latency are serious concerns for data centre operators and their impact can necessitate leaving power management functions disabled. For many applications, the C6 exit latency times are acceptable and can be implemented.

As an example, power management enablement in the BIOS or OS has been shown to inhibit virtual machine application performance. The BIOS or OS settings interfere with hypervisor settings and the management of multiple instances on the server. Several hardware and software manufacturers [i.6], [i.7], [i.8] and [i.9] recommend disabling BIOS or OS based power management functions to eliminate performance and latency impacts on virtualized servers. Power management settings can be implemented in the hypervisor, but the settings should be managed by the hypervisor and tuned to the requirements of the virtualized environment. Virtualization allows a single server to run many workloads, increasing the utilization of each server in the data centre and reducing the number of servers and energy consumption required to deliver a specific set of workloads. OEMs need the freedom to be able to deploy and ship the power management profile needed to meet their individual customer's data centre operating requirements and workloads. Mandating power management settings that interfere with server Virtualization functions will not further data centre energy efficiency.

4.4 Conclusion/Recommendations

Server processors can achieve significant power savings through the enablement of P-states and C-states, as shown by the data for the EPYC, Xeon Silver 4109T and POWER9 processors. P-state and C-state enablement improves server energy efficiency by up to 25 % as measured by the overall SERT score and reduces idle power by up to 50 %. These power demand reductions and improved SERT overall efficiency scores come at the cost of server performance and response times. While some data centre operators can use the manufacturer's default power management profile, other operators will require purpose-built or performance-oriented power management profiles to meet their operational objectives.

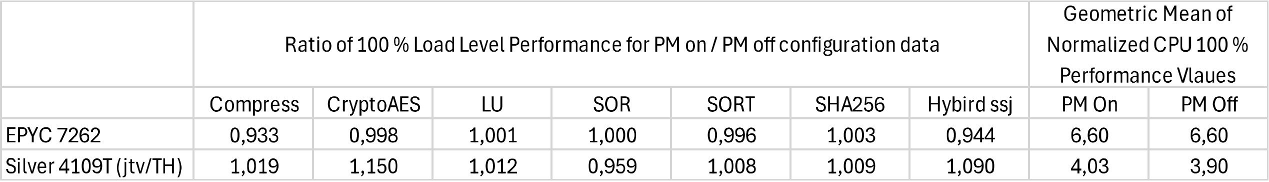
Power management is not a one size fits all solution and different customers have traditionally been empowered to enable and configure specific power management capabilities as needed to best meet their operating objectives and requirements. For some data centre operators or applications, such as high frequency stock trading, banking, supply chain management, active retail artificial intelligence or management, or highly virtualized systems, a performance-oriented setting is likely to be specified because the maximum performance level and/or minimum response time should be available or power management is managed by another software program such as the hypervisor. Other operators or users can accept the lower performance and/or longer response time at lower utilization rates associated with a default power management as these are not material to their workloads.

Annex A: Detailed discussion of Hybrid ssj performance and power measurements and key indicators for server configurations with power management turned on and off

Following the collection of SERT® power and performance data with power management turned on and off for three server configurations: one with an AMD EPYC™ processor, one with an Intel® Xeon® x86 processor and one with a POWER9™ processor. The two configurations for the Intel® Xeon® x86 were manufactured by 2 OEMs and have slightly different component loads, but the two configurations provide representative data for the discussion. For purposes of comparison between configurations, the Hybrid ssj worklet interval power data is graphed for each configuration sets (Figures 2 to 4) and the overall SERT score for the EPYC and Intel® Xeon® x86 processors-based 14servers and the Hybrid ssj worklet efficiency score for the POWER9 processor based server are provided (Tables 1 to 3). The Hybrid ssj worklet was chosen because it measures 8 performance intervals providing power demand data at 12,5 % utilization of the server CPU.

The data for the three configuration pairs shows that the implementation of power management features consistently reduces the power demand resulting in an improvement of the overall SERT or Hybrid ssj SERT score when power management is turned on. The impact of power management on performance is more varied, with minor increases or decreases in performance occurring in the different worklets (Table A.1). The variations in maximum performance on the EPYC system can be primarily attributed to experimental error while the better performance data for the Silver 4109T system with power management turned on is likely due to a better design for server jtv.

Table A.1: Impact of Power Management on Worklet Performance



	Ratio of 100 % Load Level Performance for PM on / PM off configuration data							Geometric Mean of Normalized CPU 100 % Performance Vlaues	
	Compress	CryptoAES	LU	SOR	SORT	SHA256	Hybird ssj	PM On	PM Off
EPYC 7262	0,933	0,998	1,001	1,000	0,996	1,003	0,944	6,60	6,60
Silver 4109T (jtv/TH)	1,019	1,150	1,012	0,959	1,008	1,009	1,090	4,03	3,90

AMD EPYC processor

The configuration and key metrics data for the AMD EPYC processor configuration are shown in Table 1. Power management enablement does create significant reductions in the server power demand during lower utilization load periods of the test: a 17,1 % reduction in power at the 12,5 % hybrid ssj power load level and 41,3 % reduction in the idle power portion of the test.

The net power reductions at the lower utilizations result in SERT score improvements due to reduced power demand at each testing interval. For this system, enabling CPU power management resulted in a 11,9 % increase in the Hybrid ssj worklet efficiency score and 13,2 % increase in the overall SERT active efficiency score, a significant improvement for this equipment.

The Geopeak performance, the geometric mean of the six 100 % CPU worklet performance values, are roughly equivalent for the two tests (Table A.1) - the power management implementation did not seriously affect the CPU performance during periods of maximum workload demand (i.e. sustained 100 % utilization with no idle periods.) This is expected due to the way the SERT test sends work to the server. At the 100 % workload utilization level the server is so busy that C-states will generally not be utilized in the CPU cores and P-states would be utilized for an insignificant portion of the test if at all.

Intel® Xeon® Silver 4109T Based Server

The configuration and key metrics of two comparable Intel® Xeon® 4109T based servers made by two different manufacturers are shown in Table 2 and Figure 3. The Geomean performance and the 100 % normalized worklet performance (Table A.1) varies between the two servers; this is attributed to the differences in configurations and manufacturers. This data demonstrates that server power and performance for the same CPU and a comparable configuration can be affected by the server architecture and the firmware. In turn, the SERT score - 30,3 % higher for server jtv - directly reflects and substantiates these OEM based system differences and the benefits of enabling power management. Server TH has one more Hard Disk Drive (HDD), which adds 5 to 10 W of additional power across all load levels, and server jtv has 99 GB more of memory representing between 7 and 10 W of additional power across all load levels. The additional power demand of these components largely offset between the two server configurations and do not materially affect the analysis and comparison of the power demand and overall SERT scores.

Looking at the power vs performance achieved curve in Figure 3, server TH does not have P-states and C-states functioning (either not configured or misconfigured.) The server TH performance-power curve matches the curve for the power management off configuration in Figure 3 (orange curve). Server jtv appears to have both P-states and C-states enabled, as exhibited by the curve shape. The P-states manifest differently in the EPYC and Silver 4109T servers, but the net outcome is a non-linear reduction in power demand and an overall SERT score increase are the same.

There are significant reductions in the server power demand between server jtv and server TH: 18,5 % reduction in the 12,5 % hybrid ssj power load level and 51,6 % reduction in the idle power. Maximum power is 4,8 % higher for jtv with power management on which likely results from differences in the product configurations. The power reductions at the lower utilizations and idle power for configuration jtv translate into a 30,3 % improvement in the overall SERT score.

IBM POWER9 Processor

The configuration and key metrics for the POWER9 PowerVM server shown in Table 3 and Figure 4. All three operating profiles have implemented idle power savings. The PowerVM system implements different operating profiles than the power management profiles offered on an x86 processor and Windows operating system. Servers offering POWER9 with Linux will offer power management modes similar to those to an x86 processor and Windows OS through the Linux OS. Data for a Power Linux system is not provided.

The maximum power and normalized maximum performance values (Figure 4) are different for each of the three operating modes because each mode sets the voltage and frequency at different levels specific to the operating mode, with Maximum Power mode having the highest settings, Dynamic Performance mode with lower settings and Static Power Saving mode having the lowest settings.

Only the Hybrid ssj worklet was run for the POWER9 configuration, so the Hybrid ssj worklet score will be used as a proxy for the overall SERT score. The Maximum Performance mode has the lowest overall SERT score and the Static Power Save mode the highest. As the operating mode becomes more restrictive, the worklet efficiency score improves because the power reduction achieved by the voltage/frequency reductions outweighs the attendant loss of performance when calculating the overall SERT score.

This highlights a very important point. While power management may reduce power demand and result in a higher overall SERT efficiency score, it does not necessarily ensure that the server will be able to meet the performance demands of a given data centre operator or user(s). For some data centre operators or applications, such as high-speed stock trading or banking, supply chain management or active retail artificial intelligence or management systems, the maximum performance setting (no power management) is likely to be specified because the maximum performance level should be available at all times. Other operators or users can accept the lower performance rates and power demand offered by Dynamic Performance at lower utilization rates because computational capacity can be captured from idle cores. There are very few, if any, operators or users who will accept the constraints imposed by the Static Power Save mode - yet it offers the best SERT efficiency score.

History

Document history		
V1.1.1	September 2024	Publication