What's New in POWER9 Performance a102780

Ian Nash
iannash@au.ibm.com
IBM POWER Systems Architect

Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Credits

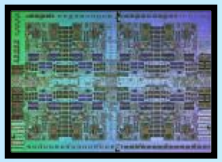
This session includes materials developed by Bret Olszewski, Ron Arroyo, Todd Rosedahl, Stephen Naspany and Nigel Griffiths of IBM


POWER9

http://www.fudzilla.com/news/processors/41449-ibm-shows-off-POWER9

*Actually, an IBM 704 introduced in 1954

POWER Processor Technology Roadmap

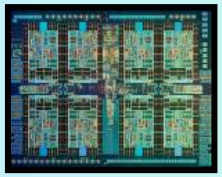


POWER7
45 nm

Enterprise

- 8 Cores
- SMT4
- eDRAM L3 Cache

1H10

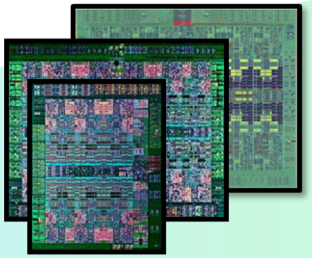


POWER7+
32 nm

Enterprise

- 2.5x Larger L3 cache
- On-die acceleration
- Zero-power core idle state

2H12

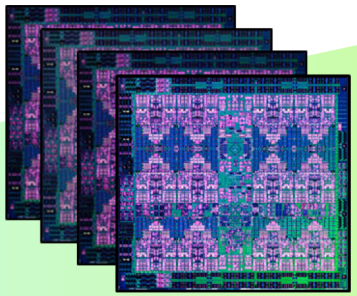


POWER8 Family
22nm

Enterprise & Big Data Optimized

- Up to 12 Cores
- SMT8
- CAPI Acceleration
- High Bandwidth GPU Attach

1H14 – 1H17



POWER9 Family
14nm

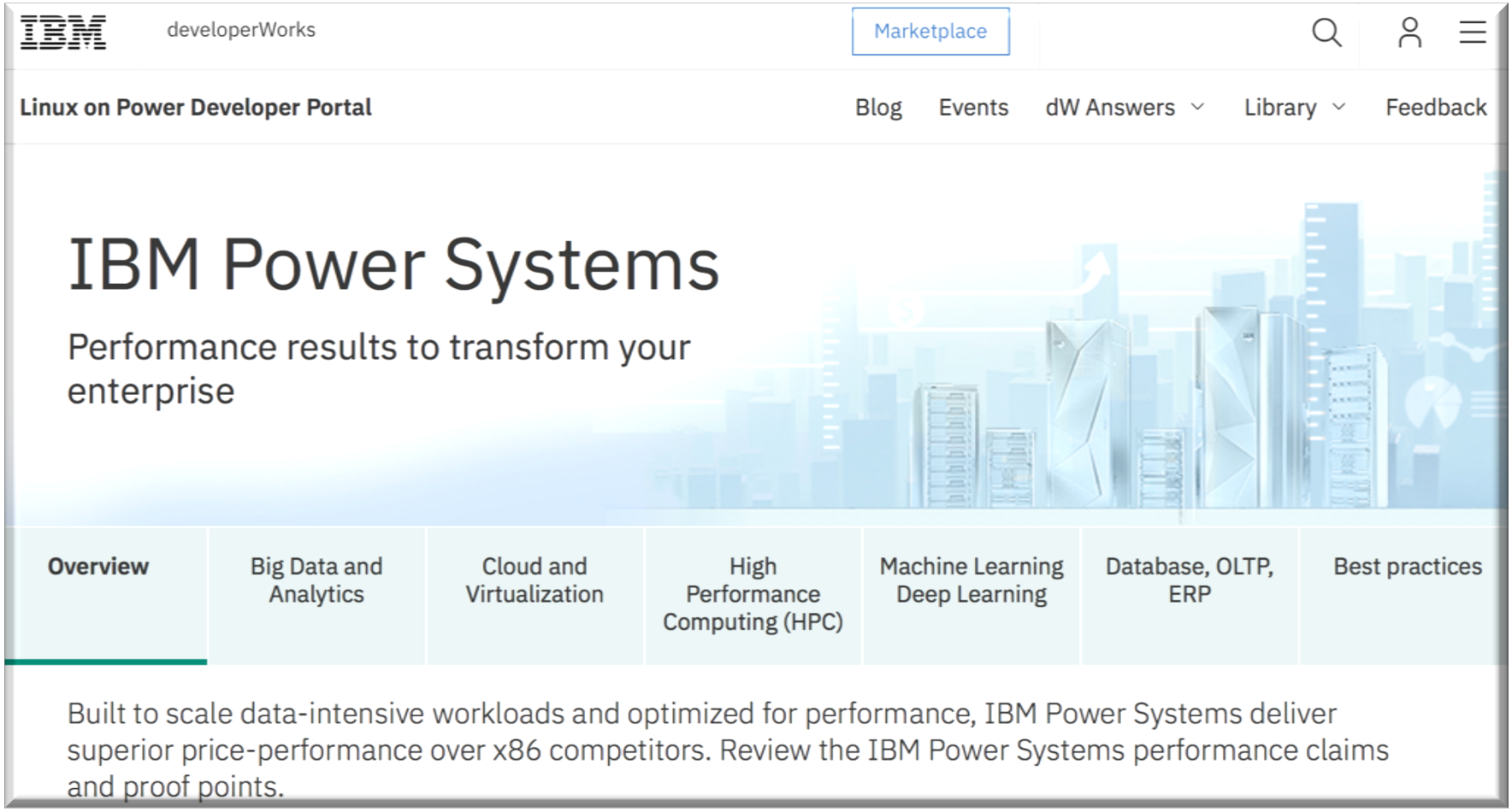
Built for the Cognitive Era

- Enhanced Core and Chip Architecture Optimized for Emerging Workloads
- Processor Family with Scale-Up and Scale-Out Optimized Silicon
- Premier Platform for Accelerated Computing

2H17 – 2H20

Power Systems Performance Collateral

https://developer.ibm.com/linuxonpower/perfcol/



POWER9 Performance Best Practices

A brief checklist

This document is intended as a short summary for customers on key items that should be looked at when planning a migration. For a more in-depth and more complete set of recommendations, please refer to the document links provided on the second page.

Description	Instructions
Ensure firmware is current	Fix Central provides latest updates. Latest F/W levels as of this writing : FW910 for POWER9 models S914, S922 and S924 Use the FLRT tool to obtain the recommended levels for a given platform. **NOTE**: Ensure required HMC level is installed when updating F/W.
Memory DIMMs	For optimal performance on workloads that are memory bandwidth sensitive follow these recommendations: **S914/S922/S924**: • Assign minimum 4 DIMMs per socket • DIMMs on same memory channel must have the same size **All POWER** : Follow proper memory plug-in rules
Ensure OS level is current	Fix Central provides the latest updates for AIX, IBM i, VIOS, Linux, HMC and F/W. In addition to that, the FLRT tool provides the **recommended levels** for each H/W model. Use these tools to maintain your system up to date.
SMT8	In order to take full advantage of the improved performance of the POWER9 CPU, we recommend customers evaluate the use of SMT8 in their environment. Although a change between SMT modes is dynamic (via smtctl), we recommend when moving to SMT8 to reboot the given partition to get the best performance of this change.
40GbE adapter	RHEL7: For network bandwidth sensitive workloads, we recommend increase the receive queue size from 1024 to 8192.
Sizing a system	• When migrating to POWER9, we recommend considering using SMT8, and size the LPARs based on the SMT8 rPerf values; in many instances, this will likely reduce the number of VPs required. • Use Workload Estimator (WLE) for sizing LPARs for CPU consumption as it provides better sizing results.
Right-size your Shared LPARs	• Assign entitled capacity (EC) to sustained peak utilization for LPARs with critical SLA requirements • Assign EC to average utilization and number of virtual CPUs to peak utilization(physical core consumption) for LPARs with non-critical SLA • Ensure the average LPAR utilization is equal or less than 75% of the entitled capacity
Java	• IBM JDK8 SR5 is the minimum level to exploit POWER9 • Open JDK 1.8 provides partial support for P9 ISA • Use of 64k size pages normally increases application performance
Partition Placement	Current FW levels ensure optimal placement of the partitions. However, if constant DLPAR operations are executed on partitions on the CEC, it is recommended the use DPO to optimize placement.

Description	Instructions
Compilers	• IBM xlC for Linux : 13.1.5 & 15.1.6 support for P9 ISA • Advanced Toolchain : 11.0-3 and later • gcc: Version 7 of gcc is recommended for P9 ISA support. Also includes support for "-mtune=power9"
IBMi	Ensure Technology Updates are current (see link below)
AIX Tunables/ VIOS Tunables	• Tuning a VIOS is not recommended unless directed by VIOS/AIX support. • **Restricted** tunables **should not** be modified (unless directed by AIX/VIOS development) • Tunables **should not** be migrated across AIX levels.
AIX CPU utilization	The AIX OS system is optimized for best raw throughput at higher CPU usage. If the customer requires to reduce CPU usage (pc), use the **schedo** tunable vpm_throughput_mode to tune the workload and evaluate the benefits of raw throughput vs. CPU usage.
VIOS configuration	• If configured with shared processors : ➢ Assign total entitlement of all VIOS partitions to be 10-15% of cores in shared pool and assign CPU ratio of 2:1 (vCPUs:ec). Refer to the PowerVM Best Practices for additional recommendations ➢ Assign uncapped mode and set variable weight capacity of VIOS partition higher than all client LPARs serviced by VIOS • For performance and flexibility, it is recommended to use IBM i to virtualize internal storage to IBM i. If you must use VIOS, follow the wiki at the following link. • For vFC, ensure no more than 64 client connections total per physical fcs adapter on the VIOS. Also, ensure no more than 64 storage ports configured per vFC adapter on the client. These are physical limits; practical limits may differ based on workload. • For vSCSI disks, ensure the queue_depth for virtual disks is less than or equal the queue_depth of the physical disk in the VIOS. • For vSCSI adapters, ensure you configure VTDs based on the following formula: Max VTDs = (512 -2) / (virtual_q_depth + 3) • Only enable the largesend attribute on the SEA (physical adapter backing the SEA) if all LPARs serviced by the VIOS are AIX partitions.
Virtual Ethernet adapters on AIX	• Increase the virtual Ethernet (vETH) device driver buffers if the partition is dropping packets on the virtual interface even when running with entitled CPU capacity. e.g., chdev –l ent# -a max_buf_xxx=NNNN **NOTE**: For desired buffer size adjustments, refer to "AIX on Power – Performance FAQ" link below • Set largesend on vETH adapter to improve performance (AIX): chdev –l en# -a mtu_bypass=on (or) ifconfig en# largesend

https://www14.software.ibm.com/webapp/set2/sas/f/best/home.html

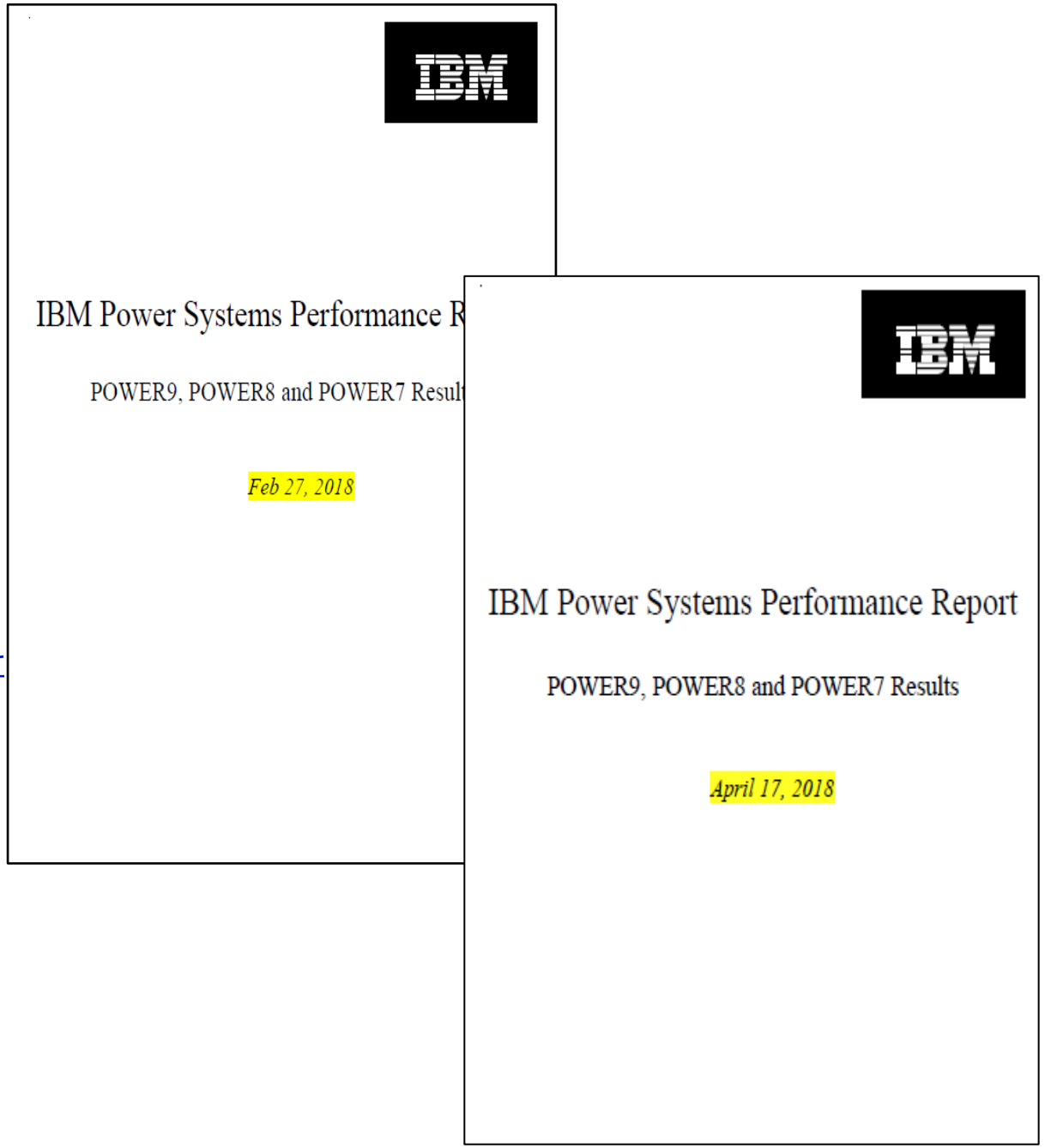
Feb 27th 2018

The Must Have Document

Google:

ibm power systems performance report

https://www.ibm.com/systems/power/hardware/reports/system_perf.html



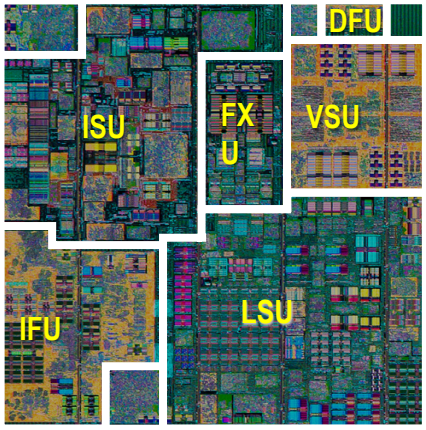
IBM POWER Architecture & Terminology Basics

thread	a hardware/software abstraction on a physical core
processor	collection of cores on the same physical die
chip	may be one or more processors packaged on a single socket.
SMT	Simultaneous Multi-threading
SCM	Single-Chip Module – one chip per socket
DCM	Dual-Chip Module – two chips packaged per socket
NUMA	Non-Uniform Memory Architecture
NUCA	Non-Uniform Cache Architecture

POWER8 core	POWER9 PowerVM core	POWER9 OpenPOWER core
Up to 12 cores/chip **SCM/DCM** **64 KB Instruction Cache** **32 KB Data Cache** **512 KB Level 2 cache** **8 MB L3 Cache** **Support for SMT8**	**Up to 12 cores/chip** **SCM** **64 KB Instruction Cache** **64 KB Data Cache** **512 KB L2 Cache** **10 MB eDRAM L3 Cache** **Support for SMT8**	**Up to 24 cores/chip** **SCM** **32 KB Instruction Cache (not shared)** **32 KB Data Cache (not shared)** **512 KB L2 Cache (shared)** **10 MB eDRAM L3 (shared)** **Support for SMT4**

POWER9 Cores

- SMT4 and SMT8 cores constructed from Modular “Execution Slice” architecture
- Uniform execution slices facilitate efficient resource utilization, low latency, and data-type sharing
- Chips manufactured with SMT4 or SMT8 cores (*not switchable)*

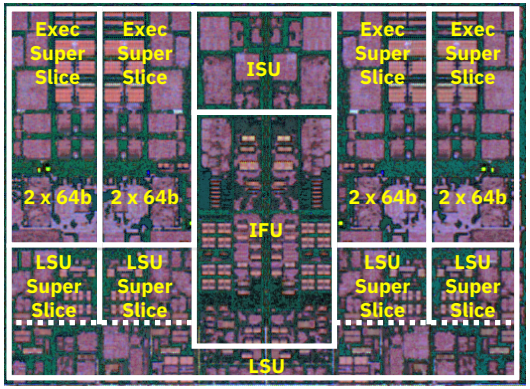


POWER8 SMT8 Core

Up to 12 / socket
96 threads max

1 or 4 partitions (mode)

4 x 128b
Super-slice



POWER9 SMT8 Core

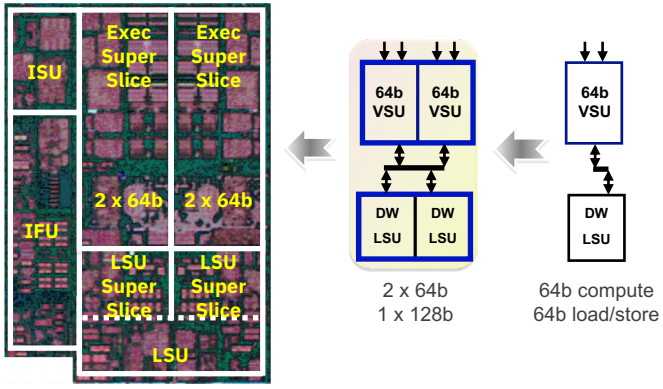
Up to 12 / socket
96 threads max

PowerVM Ecosystem
Optimized for single large partition
Tech. and migration continuity

2 x 128b
Super-slice

128b
Super-slice

64b
Slice



POWER9 SMT4 Core

Up to 24 / socket
96 threads max

Linux Ecosystem
Optimized for core / partition granularity
4 partitions (KVM)

rPerf Comparisons – SMT4 & SMT8

			rPerf							
Model				**S924** 24c/3.4-3.9		**S924** 20c/3.5-3.9		**S922** 20c/2.9-3.8	
	Cores/GHz	SMT		**System**	Per core	**System**	Per core	**System**	Per core
p750	32c/3.6	4	335	**1.38x**	1.84x	**1.18x**	1.89x	**1.01x**	1.62x
		8	-	**1.74x**	2.32x	**1.49x**	2.39x	**1.27x**	2.04x
p750+	32c/3.5	4	354	**1.30x**	1.74x	**1.12x**	1.8x	**0.95x**	1.53x
		8	-	**1.64x**	2.19x	**1.41x**	2.27x	**1.20x**	1.92x
	32c/4.0	4	397	**1.16x**	1.55x	**1.0x**	1.6x	**0.85x**	1.36x
		8	-	**1.47x**	1.95x	**1.26x**	2.01x	**1.07x**	1.72x
S824	24c/3.52	4	371	**1.25x**	1.25x	**1.07x**	1.28x	**0.91x**	1.09x
		8	397	**1.47x**	1.47x	**1.26x**	1.51x	**1.07x**	1.28x
	16c/4.15	4	284	**1.62x**	1.09x	**1.39x**	1.15x	**1.19x**	0.95x
		8	304	**1.91x**	1.28x	**1.64x**	1.32x	**1.40x**	0.89x
	12c/3.89	4	207	**2.23x**	1.12x	**1.91x**	1.15x	**1.63x**	0.98x
		8	221	**2.63x**	1.32x	**2.62x**	1.35x	**1.92x**	1.16x

rPerf Comparisons – S824 vs S924

Throughput Increase between SMT modes			
	ST -> SMT2	SMT2 -> SMT4	SMT4 -> SMT8
S824	45%	30%	7%
S924	**70%**	38%	**26%**

The real-world case for migrating PowerVM POWER8 workloads to POWER9 is from SMT4 to SMT8:

		S924			
		ST	SMT2	SMT4	SMT8
S824	ST	197	335	462	583
	SMT2	285	*1.18X*	*1.62X*	*2.04X*
	SMT4	371	*0.90X*	*1.24X*	***1.57X***
	SMT8	397	*0.84X*	*1.16X*	*1.47X*

rPerf earlier Architectures

Use the IBM Power Systems Performance Report for POWER8 to POWER9 sizings for SMT4 and SMT8.

https://www.ibm.com/systems/power/hardware/reports/system_perf.html

For older architectures, SMT breakdowns are not provided by the report. For reference these approximations are ‘roughly’ used:

POWER7/7+, SMT2 sizing is 83% of SMT4 rating

POWER7/7+, Single-Thread sizing is 56% of SMT4 rating

POWER6, Single-Thread sizing is 66% of SMT2 rating

IBM Power Systems Performance Capabilities Reference

https://www-03.ibm.com/systems/resources/systems_power_software_i_perfmgmt_pcrm_feb2018.pdf

POWER9 S914 & S924 Updates (April 4th & 17th)

Section 1 – AIX Multiuser SPEC CPU2017 Performance

All results in this table reflect performance with firmware and Operating System updates to mitigate Common Vulnerabilities and Exposures issue numbers CVE-2017-5715, CVE-2017-5753 and CVE-2017-5754 known as Spectre and Meltdown.

Model	Processor/ # Cores	GHz	Cache L1 (KB) Per core	Cache L2/L3/L4 (MB)/ System	SPEC int_ rate 2017	SPEC int_ rate_ base 2017	SPEC fp_ rate 2017	SPEC fp_ rate_ base 2017	OS Version
S924	p9/24	3.4 to 3.9	64/64	12/240/-	277	213	-	-	SLES 12 SP3

Section 2a – AIX Multiuser Performance (rPerf : POWER9) – Non-default Processor Power Mode Setting

All POWER8 and POWER9 results in this table reflect performance with firmware and Operating System updates to mitigate Common Vulnerabilities and Exposures issue numbers CVE-2017-5715, CVE-2017-5753 and CVE-2017-5754 known as Spectre and Meltdown.

Model	Processor / # Cores	Freq. GHz*	Cache L1 (KB) Per core	Cache L2/L3/L4 (MB)/ System	LPAR Size# cores	rPerf ST	rPerf SMT2	rPerf SMT4	rPerf SMT8	Non-default EnergyScale Power Mode Setting
S914	p9/4	2.3 to 3.8	64/64	2/40/-		32.3	54.9	75.7	95.4	Max performance*
S914	p9/6	2.3 to 3.8	64/64	3/60/-		47.3	80.4	110.9	139.8	Max performance*
S914	p9/8	2.8 to 3.8	64/64	4/80/-		68.3	116.1	160.2	201.8	Max performance*

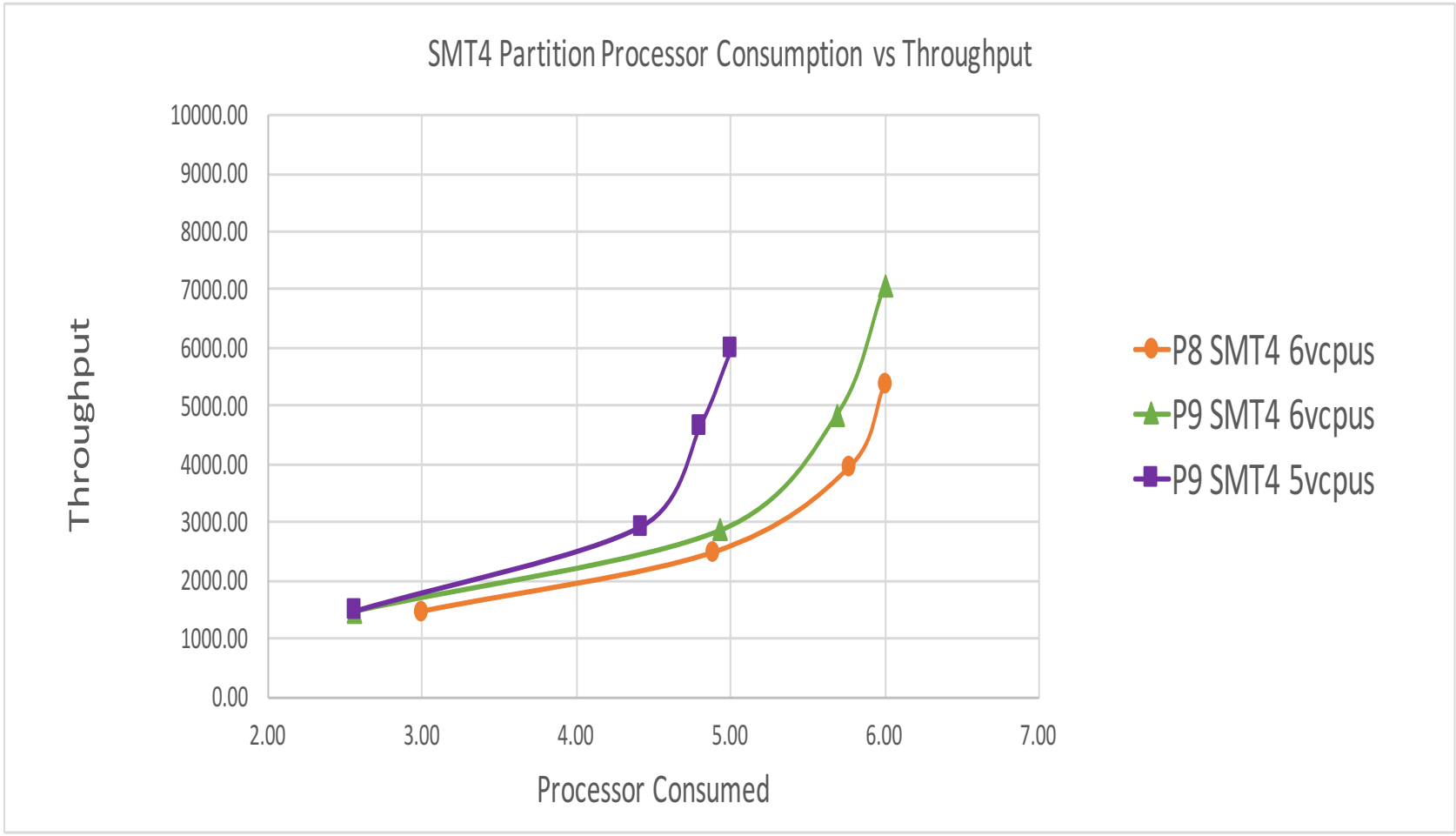
*S914 systems running in maximum performance mode may observe measurably higher sound levels under high utilization.

Section 3 - Java Benchmarks (SPECjbb2015 Published Results)

All results in this table reflect performance with firmware and Operating System updates to mitigate Common Vulnerabilities and Exposures issue numbers CVE-2017-5715, CVE-2017-5753 and CVE-2017-5754 known as Spectre and Meltdown..

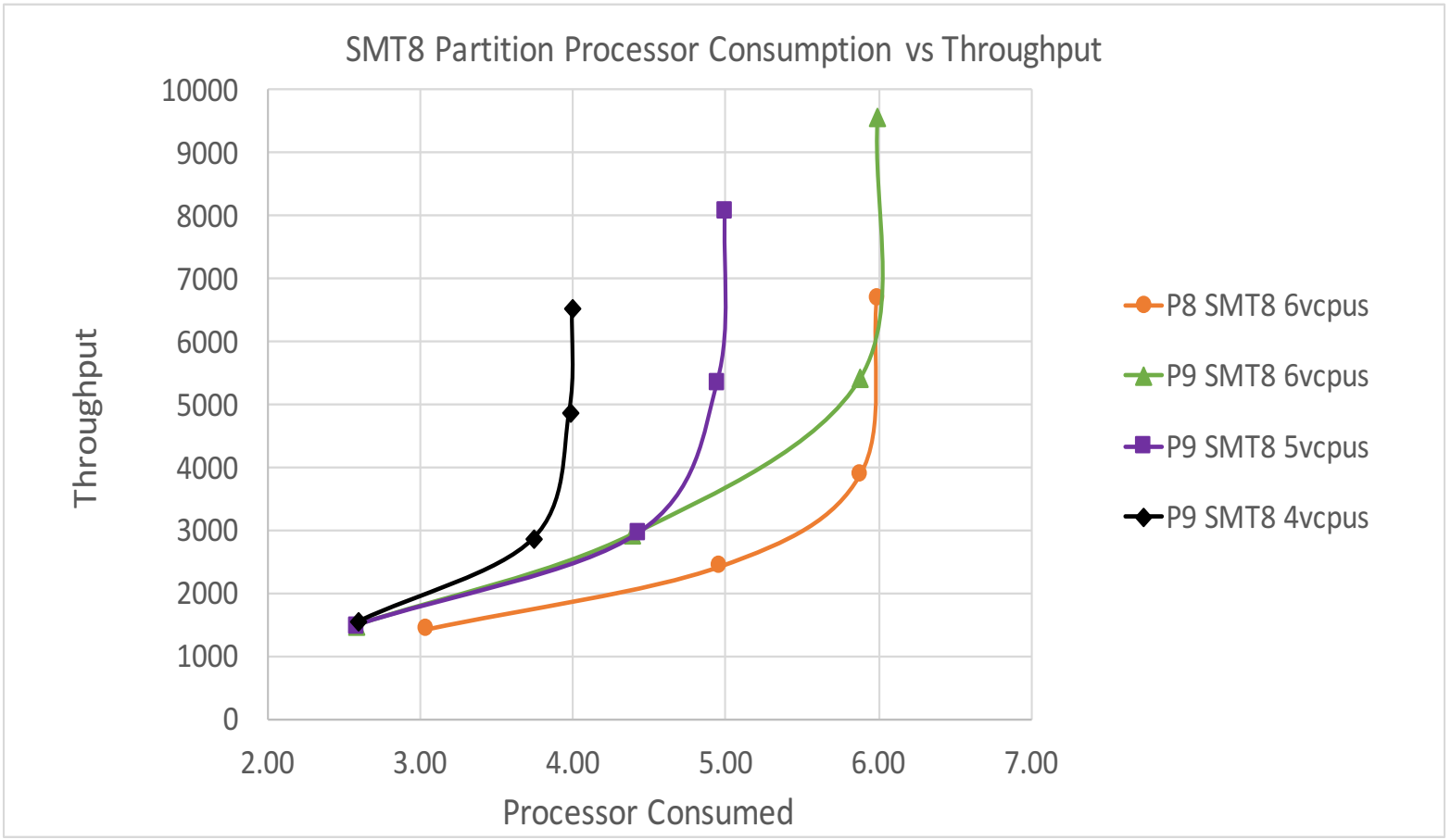
Model	Proc / # Cores	GHz	L1 Cache (KB)/core	L2/L3/L4 Cache (MB)/system	SPECjbb2015-MultiJVM max-jOPS	SPECjbb2015-MultiJVM critical-jOPS	OS Version
S924	p9/24	3.4 - 3.9	64/64	12/240/-	165,581	56,942	SLES 12 SP3

POWER8->POWER9 SMT4 to SMT4 (Transactional Workload)



- Migration with same VP count, improved utilization and response times
- 20% reduction in VP, similar response time and reduction in physical consumption

POWER8->POWER9 SMT8 to SMT8 (Transactional Workload)



- Migration same VP count: reduced utilization for same workload with similar or improved response time and higher throughput
- Migration to 5 vcpu partitions will observe similar response time for same workload with further reduced PC consumption
- 33% reduction in VP, better or equal response times for utilizations < 80%, higher throughput, lowered PC

Java & Websphere on POWER9

Best practices for Java and IBM WebSphere Application Server (WAS) on IBM POWER9

Workload	Throughput increase SMT8 vs SMT4
SPECjbb2015 max-jOPS	24.5%
SPECjbb2015 critical-jOPS	37.6%
DayTrader7 throughput	35%

https://www.ibm.com/developerworks/library/l-java-was-power9/index.html

Lab Example, POWER8->POWER9 right-sizing

The following migration analysis are estimated based on IBM internal measurements on the DayTrader7 workload

- Example S824/24c 3.52Ghz vs S924/24c
- rPerf P9/P8 ratios: SMT4 1.25x & SMT8 1.47x
- Customer migration experience may vary by workload
- Physc = AIX Physical Consumption

P8 SMT4 -> P9 SMT4 both 6 vcpus				
P8 Utilization	P8 Physc	P9 Utilization	P9 Physc	Est. Physc Improvement
20	2.06	17	1.94	6%
40	3.18	32	2.78	14%
60	4.32	46	3.63	19%
80	5.45	61	4.47	22%
P8 SMT8 6 vcpus -> P9 SMT8 4vcpus				
P8 Utilization	P8 Physc	P9 Utilization	P9 Physc	Estimated PC Improvement
20	2.29	17	1.6	43%
40	3.35	36	2.3	46%
60	4.41	56	3	47%
80	5.46	75	3.7	48%

Informal OSDB Micro-benchmark (SMT8)

Test	S824	S924	Relative Increase
Insert.SingleIndex.Contested.Rnd	96575	141943	47%
I.MIndex.Contested.Rnd	81808	120636	47%
I.MKeyIndex.Contested.Rnd	76032	109604	44%
I.DocVal.TwentyInt	67273	101930	55%
I.PartialIndex.FullRange	103715	144893	39%
Update.SetWithIndex.Rnd	71291	104519	46%
U.SetWithMIndex.Rnd	62483	94582	51%
U.DocVal.TwentyNum	51662	80280	55%
U.ManyElementsInArrary	2904	5091	75%
MultiUpdate.Contended.NoIndex	85654	124096	44%

POWER8 24c/3459 MHz vs POWER9 20c/3522 MHz
P8 results adjusted to P9 frequency
PowerVM LPAR, 4.0 Entitlement, 4 Virtual Processors, SMT8
RHEL 7.4 3.10.0-693.el7.ppc64le (No SPEC fixes)
MongoDB 3.6.2, no tunings, 32 DB threads. Mongo—perf microbenchmark

Informal OSDB Micro-benchmark

	ST	SMT2	SMT4	SMT8
TPS	54914	105530	161098	201966
Total Transactions	9886913	18998143	29001033	36353569
Average Latency	1.821 ms	0.948 ms	0.621 ms	0.495 ms

POWER9 24c/~3.5 GHz
PowerVM LPAR, 2.0 Entitlement, 6 Virtual Processors
RHEL 7.4 3.10.0-693.el7.ppc64le (No SPEC fixes)
Postgresql 9.6-3, no tunings, pgbench microbenchmark with 100 clients & 6 threads

All Migrations should consider moving to SMT8

The architectural changes in POWER9 see much higher improvements in capacity and latency with SMT8

- The lab chose to be conservative with POWER8 and SMT4
- Because AIX 7.2 had already been developed, there was resistance to changing default SMT mode from 4 to 8 when POWER8 shipped
- While the POWER8 architecture saw little improvement between SMT4 and SMT8 (typically < 7% for transactional workloads), gains are much more significant with POWER9
- Single- & two-socket POWER9 systems may not have greater peak bandwidth than scale-out POWER8 CDIMM systems, but they have enough
  - Even In-Memory workloads on POWER8 could rarely achieve ~50% of the total memory bandwidth before running out of CPU capacity
  - Standard DDR4 architecture provides better price/performance ratio for this class of systems
  - PCIe Gen4 will yield higher I/O performance and is not gated by bus limits

Assessing SMT on POWER9

- SMT is dynamic in AIX and Linux – it is trivial to test!
- Start with traditional products known to behave well with SMT (WAS, DB2, Oracle, SAP, some OSDB). The more current, the better.
- Check for software recommendations, but remember that most did not assess SMT8 fully because of lab's conservative approach with POWER8
- Older software levels migrating from older architectures should be reviewed more carefully. They will not suffer from SMT8, but they may benefit less.
- Open Source products that have never been tested with SMT8 should be assessed individually for scaling performance if using higher core counts
  - The concern is that some Open Source products may have never been tested with dozens of logical cpu instances (lock/latch contention)
  - Issues appear as non-linear context switch behavior as cores are added or SMT is increased

IBM has announced the intent to automatically update to SMT8 with AIX 7.2 TL3 on POWER9

Larger LPAR Migrations to POWER9 should review VP counts

- Focus on larger workloads where 20-33% reductions may be possible
  - A wide-range of performance results show this is possible
  - You can't reach higher frame utilization with SMT4 and high VP counts
  - Cost of software licensing by core warrants the effort
- Many organizations are slow to reconsider VP changes
  - Larger POWER6 to POWER7 migrations encountered "high physical consumption" complaints because AIX Dispatcher changes – requiring post-migration tuning/resizing to meet rPerf expectations
  - AIX more aggressively used Virtual Processors from POWER7 on. This algorithm is (mostly) consistent across POWER7, POWER8 and POWER9
  - Reducing VPs aids memory affinity for those customers nervous about spanning nodes
  - If you didn't assess VP sizings with POWER8 or are jumping architectures, now is the time to revisit

Migration - Other

- rPerf-level improvements can be expected with POWER8 or POWER9 modes
  - Both support SMT8, the technology is mature
  - Observationally, POWER9 mode will reduce physical consumption (covered later) due to dispatcher improvements
  - This is an early statement and may evolve as vendors exploit POWER9 capabilities in code, optimizations, compiler and JVM improvements
  - Lab/Support will prefer latest mode due to more current profiler tooling
  - Future feature developments may require POWER9 mode

Memory Speeds

DIMM / FC	Speed (Socket)	
	<= Half Populated	> Half Populated
16 GB / EM62	2.6 GHz	2.1 GHz
32 GB / EM63	2.4	2.1
64 GB / EM64	2.4	2.1
128 GB / EM65	2.4	2.1

- Memory Notes
  - Peak B/W up to 170 GB/s per socket
  - ½ population provides best memory bandwidth
  - Workloads sensitive to memory capacity should populate all slots
  - S914 does not support 128 GB DIMM

Spectre/Meltdown?

Section 2 – AIX Multiuser Performance (rPerf : POWER8 and up)

All POWER8 and POWER9 results in this table reflect performance with firmware and Operating System updates to mitigate Common Vulnerabilities and Exposures issue numbers CVE-2017-5715, CVE-2017-5753 and CVE-2017-5754 known as Spectre and Meltdown.

Model	Processor / # Cores	Freq. GHz*	Cache L1 (KB)	Cache L2/L3/L4 (MB)	LPAR Size# cores	rPerf ST	rPerf SMT2	rPerf SMT4	rPerf SMT8
S812	P8/4	3.00	32/64	2/32/128		31.3	45.3	58.9	63.0
S822	P8/4	3.00	32/64	2/32/128		31.3	45.3	58.9	63.0
S822	P8/6	3.80	32/64	3/48/128		56.4	81.9	106.4	113.8
S822	P8/8	4.15	32/64	4/64/128		77.5	112.4	146.1	NA
S822	P8/10	3.4	32/64	5/80/128		83.1	120.4	156.6	167.5
S822	P8/8	3.00	32/64	4/64/128		60.9	88.4	114.8	122.9
S822	P8/12	3.8	32/64	6/96/256		110.0	159.6	207.4	221.9
S822	P8/16	4.15	32/64	8/128/256		151.1	219.2	284.9	NA
S822	P8/20	3.4	32/64	10/160/256		161.9	234.8	305.2	326.6
S922	p9/4	2.8 to 3.8	64/64	2/40/-		30.4	51.6	71.2	89.8
S922	p9/8	3.4 to 3.9	64/64	4/80/-		68.4	116.3	160.5	202.3
S922	p9/16	3.4 to 3.9	64/64	8/160/-		133.4	226.9	313.1	394.5

While IBM will make no performance-related statement regarding any customer workload, rPerfs have published with ratings including the patches for AIX & PowerVM firmware.

POWER8 ratings have been reduced 5-7%. One could infer that POWER9 ratings for customers not implementing the patches would be higher than those published.

Testing

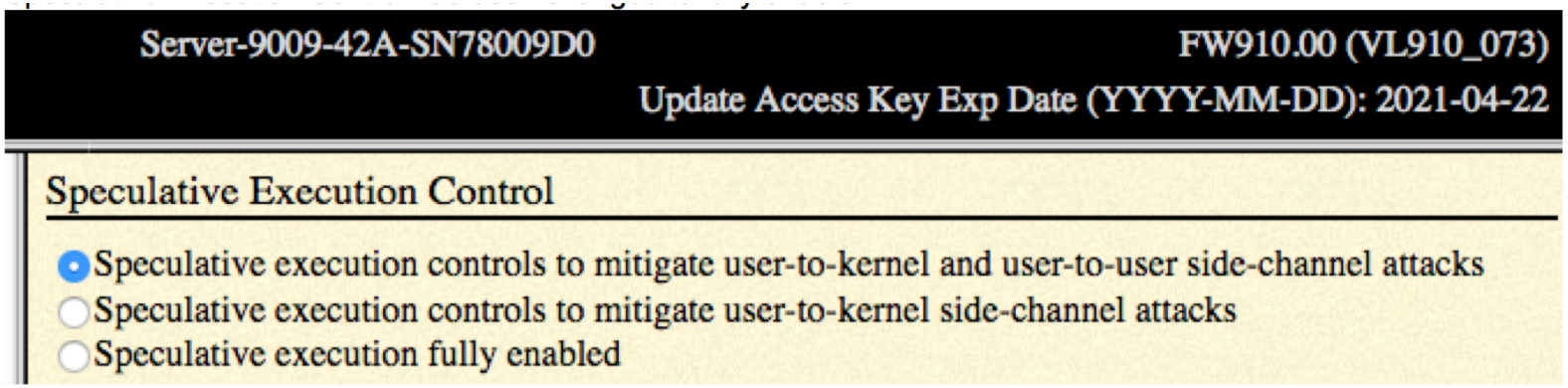
- AIX patches do not do anything unless PowerVM firmware contains patches
- Customers, instead of booting between firmware levels with and without security fixes should use LPM between systems to test their workloads

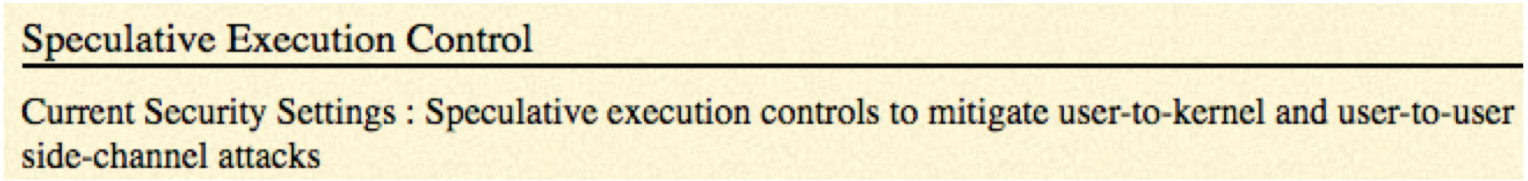
Security

AIX 7.2 code levels for POWER9 support what looks like an option to display speculation security settings (currently undocumented)

```
# lparstat -x

LPAR Speculative Execution Mode     : 2
```





Utilization Values with Simultaneous Multithreading (SMT)

POWER processor-based systems support up to 8 SMT hardware threads per core. Remember, tools will report each SMT thread as a vcpu (virtual), lcpu (logical) or cpu. These threads can be equally weighted by the Linux *Completely Fair Scheduler* (CFS).

While commands for monitoring CPU use are similar between Linux and a Unix derivative like AIX, the utilization numbers are different. AIX uses a calibration mechanism built into the POWER hardware to account for spare/idle capacity left in a core based on SMT threads used. Linux does not use this calibration.

Example: to reach 100% ”busy” in SMT4 Linux on a single core, four vcpus would have to consume 25% each (100/4)

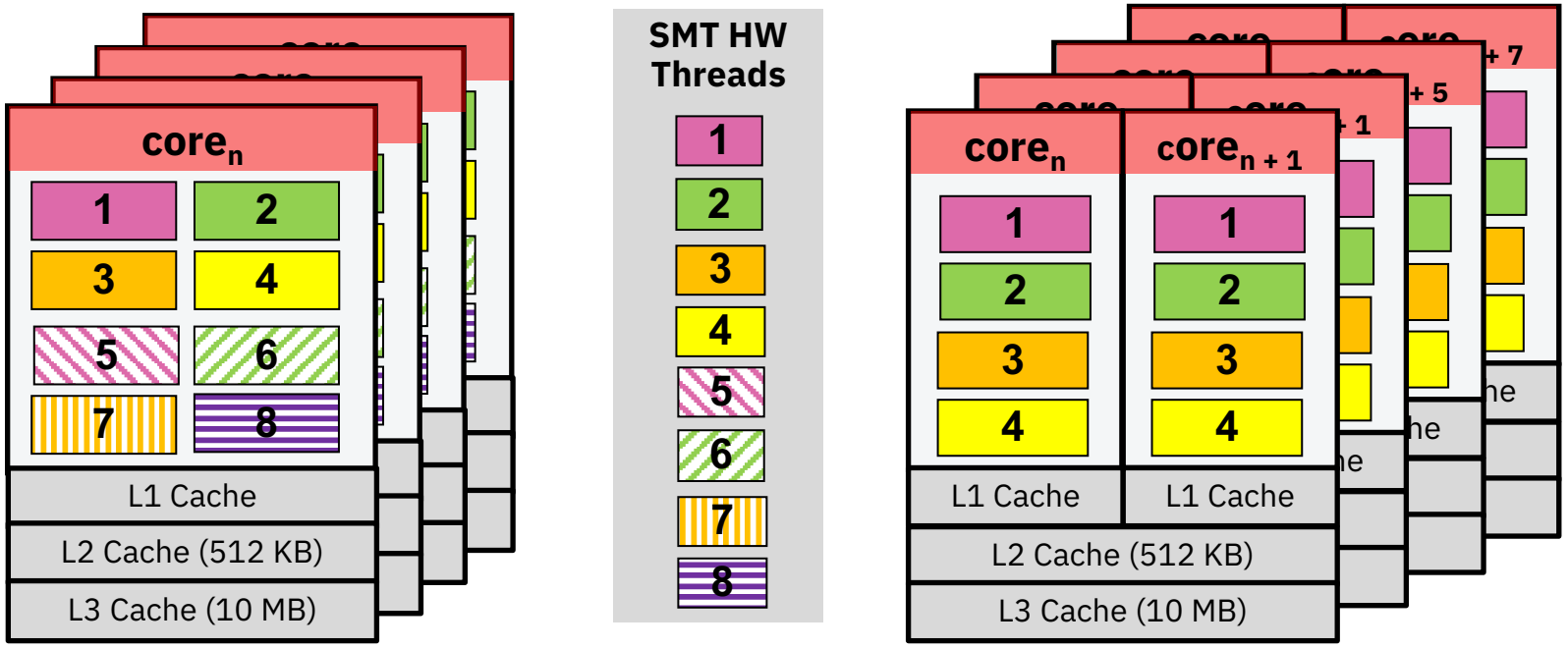
SMT Mode	Core utilization% 1 busy* thread (1 thread / vcpu)		
	Linux	**AIX POWER8**	**AIX POWER9**
Single Thread	90-99%	99%	99%
2	50%	77%	50%
4	25%	60%	44%
8	12.5%	56%	32%

*Single core, single VP

How does SMT work in POWER9?

For POWER9, SMT levels supported depend on virtualization layer

- OpenPOWER Linux cores support 1, 2 or 4 SMT threads only
- PowerVM Linux or AIX cores support 1, 2, 4 or 8 SMT threads

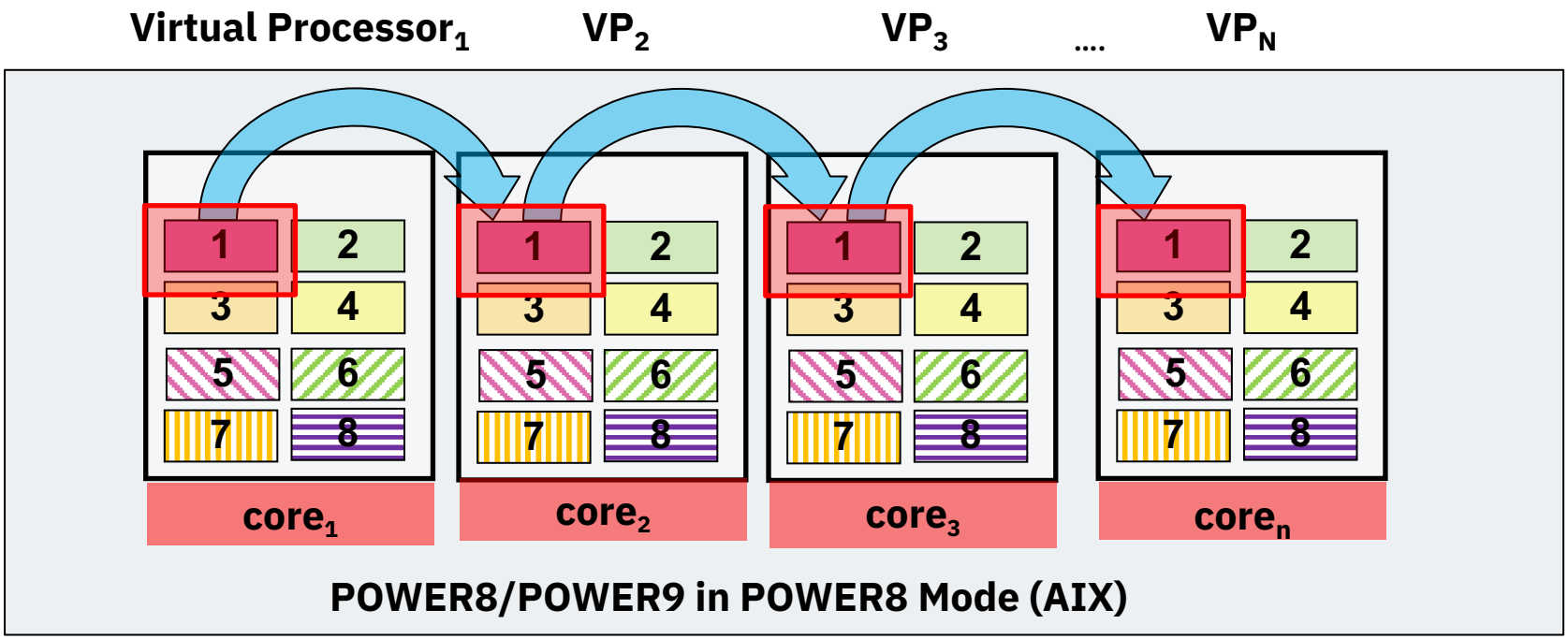


POWER9 PowerVM AIX/Linux
All POWER8 Environments

POWER9 OpenPOWER Linux

Dispatch Behavior in AIX

For POWER7 & POWER8, the default dispatch algorithm is known as Raw Throughput Mode. AIX will dispatch the first Virtual Processor SMT thread to a threshold of 50% utilization. Once all available first threads of all VPs are executing, it would then wrap around and use the second SMT thread for each VP.



Dispatch Behavior in AIX / POWER9

POWER9/AIX in POWER9 Mode behaves a little differently. The VP code is aware of the core architecture and will place/collapse smaller workloads slightly more aggressively when workloads are present:

- This optimization has the additional impact of reducing physical consumption
- Single thread utilization is calibrated to ~32% in SMT8 (~44% in SMT4), so below the default VP dispatch threshold of ~50% per core
- Because single-threads are calibrated lower and equivalent workloads will overall generate a lower utilization, they are more likely to fall below dispatch threshold, thus lowering physical consumed
- Linux not running on PowerVM dispatches to SMT4 cores and does not use hardware calibration

Customers using Scaled Throughput

Scaled Throughput Mode is an alternative AIX dispatch algorithm (vpm_throughput_mode), where SMT threads on the same Virtual Processor are executed more aggressively. In general, this mode:

- Reduces physical consumption by activating more SMT threads
- Was adopted by customers wanting to reduce physical consumption without the effort of reducing Virtual Processor counts in a migration
- Trades some performance/latency compared to Raw Mode

The common settings are 2, 4 & 8 and map to how many SMT threads are used before the next Virtual Processor is activated. Settings of 4 & 8 can be expected to have noticeable performance impacts for single-thread/latency sensitive workloads.

Guidance:

- Customers using Scaled Throughput on POWER8 and migrating to POWER9 should not expect significant per-thread performance increases
- There is nothing wrong with continuing to use Scaled Throughput in lieu of reducing VP counts, but customers in general should pursue one strategy over the other and not do both in a migration without testing

Best tool to review true VP and SMT activity

To view Virtual Processor and SMT activity on an existing workload, the best tool is mpstat with the –v option.

This option displays the actual Virtual Time Base (VTB) – the dispatch time for each Virtual Processor at the physical layer, physical consumption (pc) and the activity of the SMT threads.

#mpstat –v 2 5 (two samples, 5 second interval)

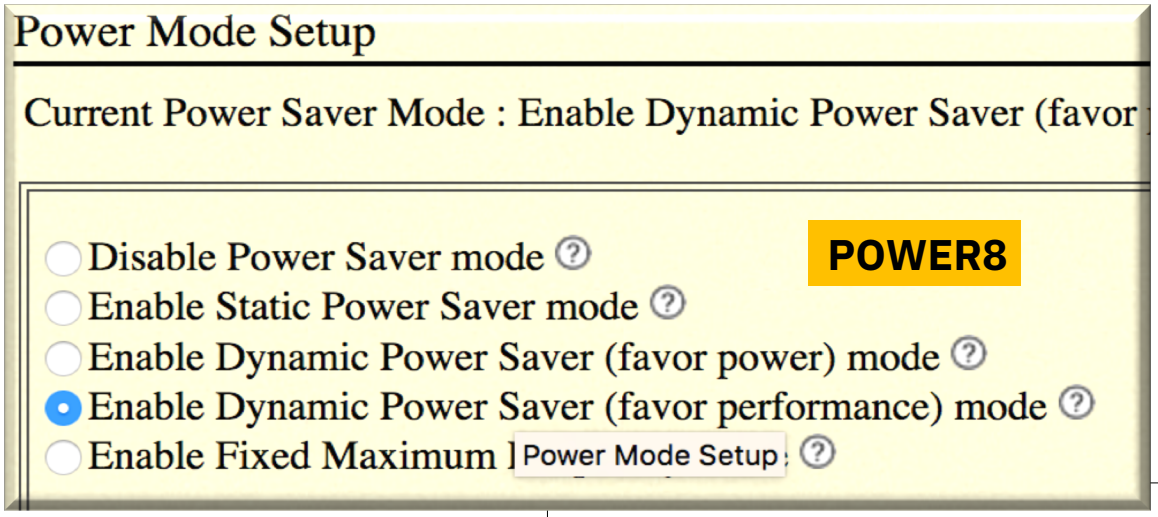
vcpu	lcpu	us	sy	wa	id	pbusy	pc	VTB(ms)
0		2.68	18.80	0.00	78.52	0.00[21.5%]	0.00[0.3%]	19
	0	2.68	16.28	0.00	20.92	0.00[19.0%]	0.00[39.9%]	-
	...							
1		58.97	0.02	0.00	41.01	0.59[59.0%]	1.00[99.9%]	4995
	4	58.97	0.01	0.00	0.00	0.59[59.0%]	0.59[59.0%]	-
	5	0.00	0.00	0.00	13.67	0.00[0.0%]	0.14[13.7%]	-
	6	0.00	0.00	0.00	13.67	0.00[0.0%]	0.14[13.7%]	-
	7	0.00	0.00	0.00	13.67	0.00[0.0%]	0.14[13.7%]	-

How many VPs are actually dispatching

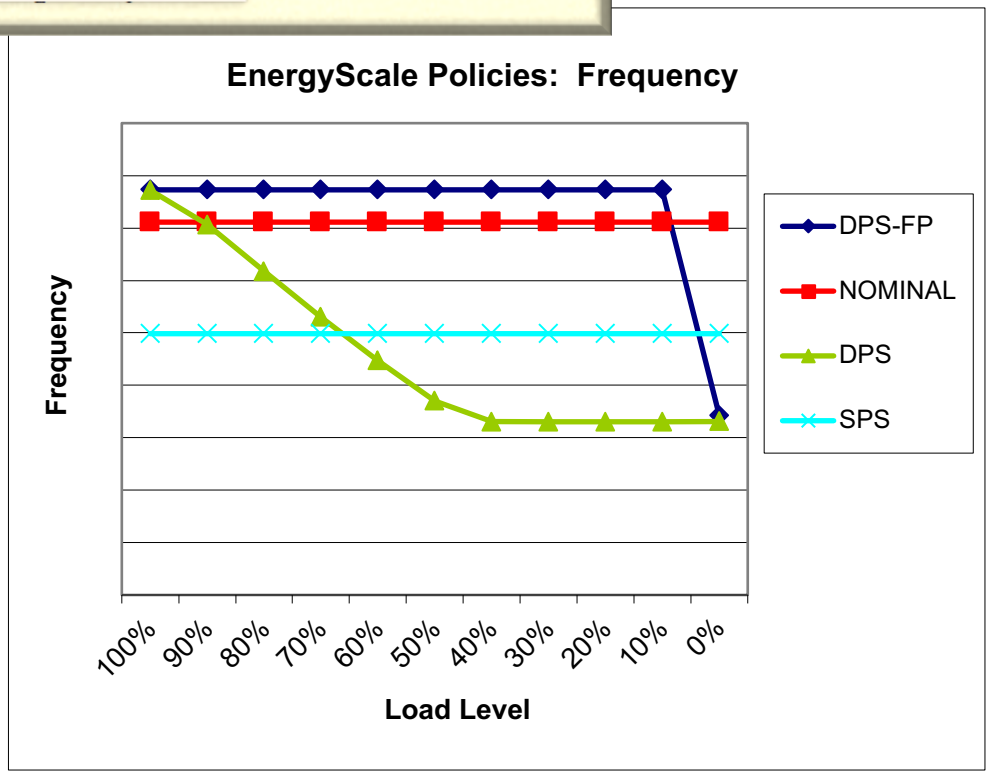
Dispatch time in milliseconds

AIX 7.1 TL3 SP2 or above required

POWER8 EnergyScale Overview



- System ships with fixed nominal frequency
- Maximum frequency (Turbo) achieved when system operating in nominal environment
- Workload will behave the same on same system configuration



POWER9 EnergyScale

Two new modes replace the POWER8 dynamic frequency modes

- POWER9 systems will ship with one of these two modes on by default
  - Dynamic Performance Mode – Enables dynamic frequency with some restrained power/thermal envelope. Default on POWER9 S914
  - Maximum Performance Mode – Enables dynamic frequency with highest performance operation. Default on S922 and S924 systems.
- Both modes dynamically adjust processor frequency to maximize performance
- Enable much higher CPU frequency range compared to POWER8
- For PowerVM systems, these are system wide modes but each CPU socket frequency is optimized separately

Factors used to determine the maximum CPU frequency

- CPU Utilization – Lighter workloads will run at higher frequencies
- Number of Active Cores – Less number of active cores will run at higher frequencies
- Environmental Conditions – Lower ambient temperatures will run at higher frequencies

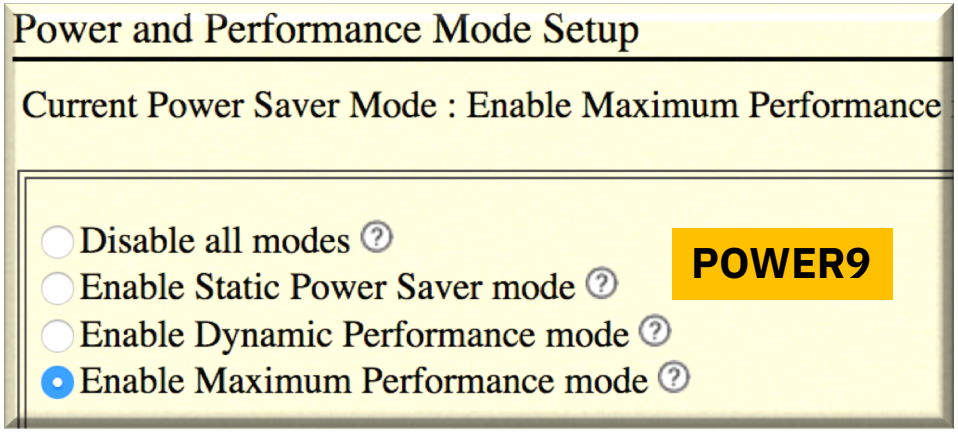
POWER9 Modes

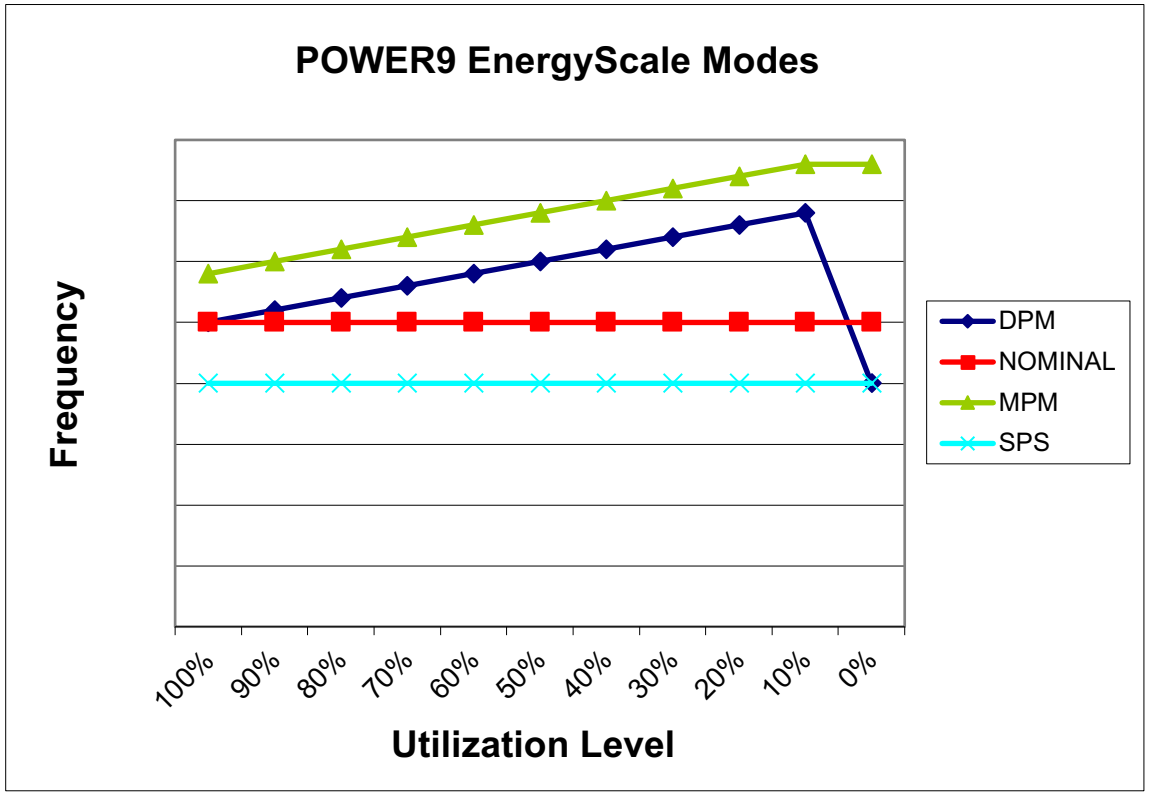
Dynamic Performance Mode

- Increased performance for typical workloads over Static Nominal
- Less active workloads can use higher frequencies
- Lower active core counts also increase top frequency potential

Maximum Performance Mode

- Increased performance over DPM for nominal environmental conditions
- Takes advantage of nominal environmental conditions by allowing increased CPU frequency and power draw
- Lighter workloads can exploit higher frequencies
- Idle state remains at high frequency





POWER9 Frequency ranges

Range of practical frequency variation experienced in maximum performance mode in an enterprise environment will be less than the minimum to maximum frequency range.

- ➢Minimum frequency defined by all cores running a computationally intensive high-performance computing loop
- ➢Maximum frequency only obtained for higher core count parts when a significant fraction of cores are turned off due to extended idle periods.
  - o rPerf ratings assume all-cores enabled so any additional variation from disabling cores is only upside
  - o Function delivered via future firmware update

Variation from a commercial workload running with all cores active to all-core maximum frequency is

- ➢<10% for 12 core offerings and for s924
- ➢<20% for lower core count offerings on s922 and s914

System	Cores	Power Management Mode	Minimum Frequency	Typical Enterprise Workload	All-core Maximum Frequency	Absolute Maximum Frequency
s924	12	Maximum	3.4	3.65	3.65	3.9
s924	10	Maximum	3.5	3.68	3.85	3.9
s924	8	Maximum	3.8	3.93	4.0	4.0
s922	12	Maximum	2.7	3.0	3.45	3.8
s922	10	Maximum	2.9	3.2	3.8	3.8
s922	8	Maximum	3.4	3.65	3.8	3.9
s922	4	Maximum	2.8	3.25	3.8	3.8
s914	8	Maximum	3.15	3.5	3.8	3.8
s914	6	Maximum	2.8	3.2	3.8	3.8
s914	4	Maximum	2.8	3.3	3.8	3.8

Monitoring Frequency

AIX

Currently, the AIX tooling only shows legacy value for Dynamic AND Maximum Performance Modes on POWER9 / AIX 7.2 (this is a bug)

```
lparstat -i | grep Saving
Power Saving Mode   : Dynamic Power Savings (Favor Performance)
```

Average of processors on LPAR:

```
lparstat -E 1 10
  Physical Processor Utilisation:
 --------Actual--------                 ------Normalised------
 user   sys  wait  idle      freq     user   sys  wait  idle
 ----  ----  ----  ----   ---------  ----  ----  ----  ----
0.352 0.003 0.000 0.645 2.3GHz[ 79%] 0.279 0.003 0.000 0.718
0.350 0.003 0.000 0.647 2.3GHz[ 79%] 0.277 0.002 0.000 0.721
0.614 0.005 0.000 0.381 3.7GHz[128%] 0.786 0.006 0.000 0.207
0.614 0.005 0.000 0.382 3.7GHz[128%] 0.785 0.006 0.000 0.209
```

POWER Processor counter interface:

```
pmcycles -M
This machine runs at 3475 MHz
```

Monitoring Frequency

Linux

List power management modes

```
dmesg | grep freq
[    0.000000] time_init: decrementer frequency = 512.000000 MHz
[    0.000000] time_init: processor frequency   = 2900.000000 MHz
```

Linux (PowerVM)

```
ppc64_cpu --frequency
```

Linux (Non-PowerVM)

List frequency of all cores

```
cat /sys/devices/system/cpu/cpu*/cpufreq/cpuinfo_cur_freq
```

Display nominal frequency range

```
cat /sys/devices/system/cpu/cpu*/cpufreq/scaling_available_frequences
```

Display frequency range

```
cat /sys/devices/system/cpu/cpu*/cpufreq/scaling_boost_frequencies
```

Proving Frequency

No tools in AIX support an indication of the range of frequencies possible on a system. Commands like prtconf or pmcycles will show lower frequency. You must apply a workload to see frequency changes.

Two simple examples would be to create a looping script or use Nigel's nstress package to generate a workload on a single cpu:

Create a script called: cpu_freq_test.sh

```
#!/usr/bin/ksh
while true
do
:
done
```

Set execute permissions and run:

```
chmod 755 cpu_freq_test.sh
./cpu_freq_test.sh
```

Download nstress and execute:

```
./ncpu -p 1 -s 120 &
```

Then execute pmcycles or lparstat –E as before:

```
pmcycles -M        (Do not use -m)
This machine runs at 3658 MHz
```

nstress available at:
https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power+Systems/page/nstress

Idle Power Saver

Idle Power Saver uses custom below-nominal thresholds for frequency adjustments. Can be combined with Disable All Modes, Dynamic Performance Mode and Maximum Performance Mode.

Do not experiment w/o support guidance. If you had previous guidance to modify these at older architecture levels, open a PMR and ask whether that should be continued for POWER9.

POWER8

Idle Power Saver

Idle Power Saver Enable
Current value: Enabled
New value: Enabled

Delay Time to Enter Idle Power
Current value: 240seconds
New value: 240 Range: MinVal-10seconds MaxVal-600sec

Utilization Threshold to Enter Idle Power
Current value: 15%
New value: 15 Range: MinVal-1% MaxVal-95%

Delay Time to Exit Idle Power
Current value: 10seconds
New value: 10 Range: MinVal-10seconds MaxVal-600sec

Utilization Threshold to Exit Idle Power
Current value: 25%
New value: 25 Range: MinVal-5% MaxVal-95%

POWER9

Idle Power Saver

Idle Power Saver Enable
Current value: Enabled
New value: Enabled

Delay Time to Enter Idle Power
Current value: 240Seconds
New value: 240 Range: MinVal-10Seconds MaxVal-600Seconds

Utilization Threshold to Enter Idle Power
Current value: 8%
New value: 8 Range: MinVal-1% MaxVal-95%

Delay Time to Exit Idle Power
Current value: 10Seconds
New value: 10 Range: MinVal-10Seconds MaxVal-600Seconds

Utilization Threshold to Exit Idle Power
Current value: 12%
New value: 12 Range: MinVal-5% MaxVal-95%

Note: Selecting a utilization threshold to enter idle power that is higher tha
unexpected behavior. Please see the EnergyScale™ white paper for more in

Other

Pre-reqs for Best Optimzation

- IBM Java JDK8 SR5
- Open JDK 1.8
- Compilers
  - AIX/xlc 2H2018
  - Linux/xlc v13.1.5, v15.1.6
  - Linux/gcc v7, -mtune=power9

Next generation SR-IOV

https://www.ibm.com/support/knowledgecenter/POWER9/p9hcd/fcec3l.htm
https://www.ibm.com/support/knowledgecenter/POWER9/p9hcd/fcec2r.htm
https://www.ibm.com/support/knowledgecenter/9009-22A/p9hcd/fcec2t.htm

Thank you!

Notices and disclaimers

© 2018 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those

customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers continued

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

.