POWER9 EnergyScale – Configuration & Management

Note: This document was originally copied from https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/POWER9%20EnergyScale%20-%20Configuration%20%26%20Management

Author: Hari Ganesh Muralidharan

POWER9 EnergyScale™

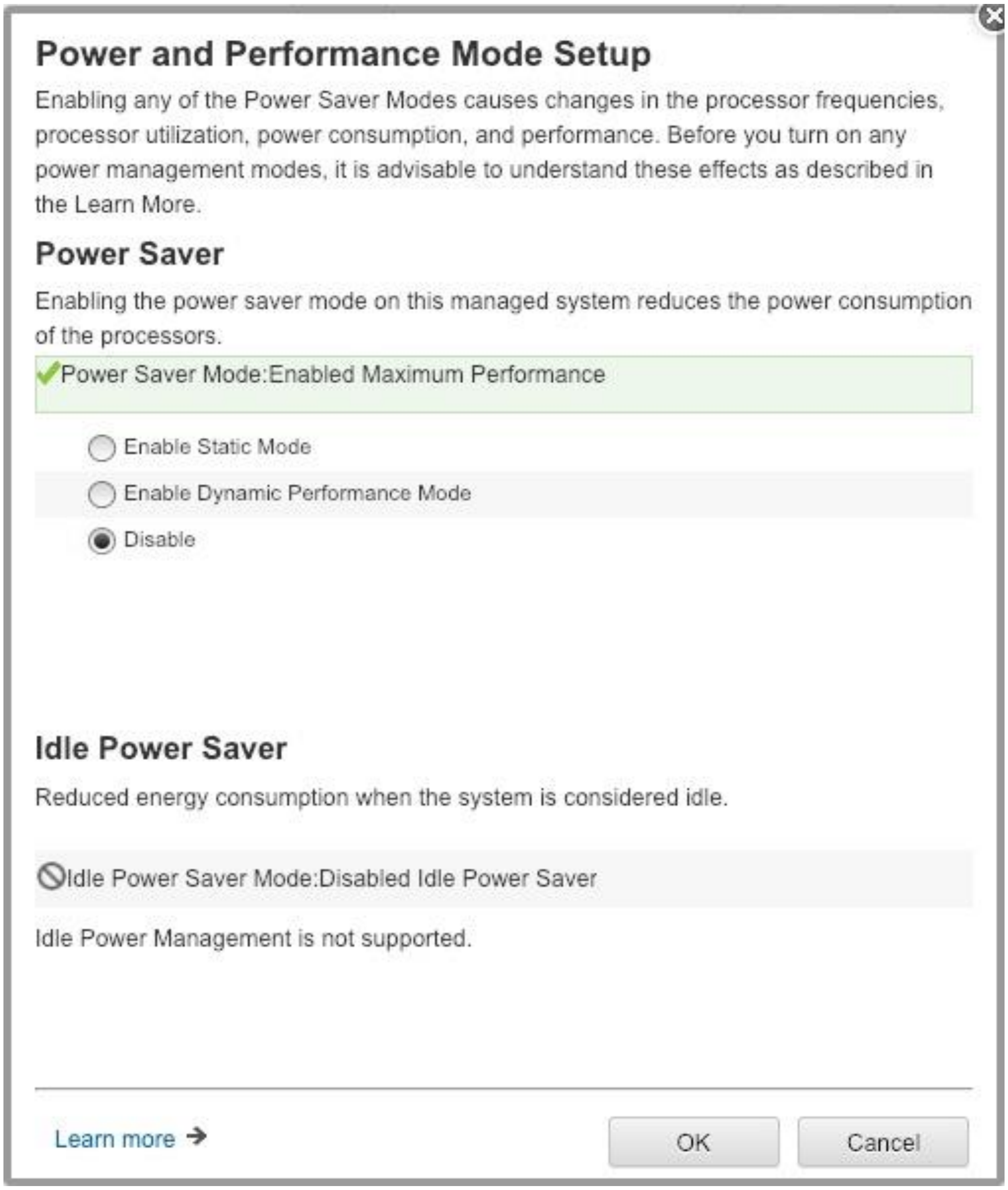
We previously published a blog that gave an overview of POWER9 EnergyScale. This is the second blog that covers how to configure and manage energy management on the server.

Changing the Energy Management Mode

The Energy Management mode can be changed in multiple ways; using the HMC GUI, using the HMC command line, or by the ASMI menus. All three options provide the same level of functionality. Note that changing the modes is a dynamic operation, it does not require a reboot, and takes effect immediately.

HMC GUI

The Power Mode Setup panel can be display by selecting the server, select "Actions", select "View All Actions" and select "Power Management". This will display the following panel:



Power and Performance Mode Setup

Enabling any of the Power Saver Modes causes changes in the processor frequencies, processor utilization, power consumption, and performance. Before you turn on any power management modes, it is advisable to understand these effects as described in the Learn More.

Power Saver

Enabling the power saver mode on this managed system reduces the power consumption of the processors.

✓Power Saver Mode:Enabled Maximum Performance

○ Enable Static Mode
○ Enable Dynamic Performance Mode
● Disable

Idle Power Saver

Reduced energy consumption when the system is considered idle.

⃠Idle Power Saver Mode:Disabled Idle Power Saver

Idle Power Management is not supported.

HMC Command Line

The HMC commands are listed below for those interested in scripting.

# ***chpwrmgmt -m <managed system name> -r sys -o enable -t <mode>***

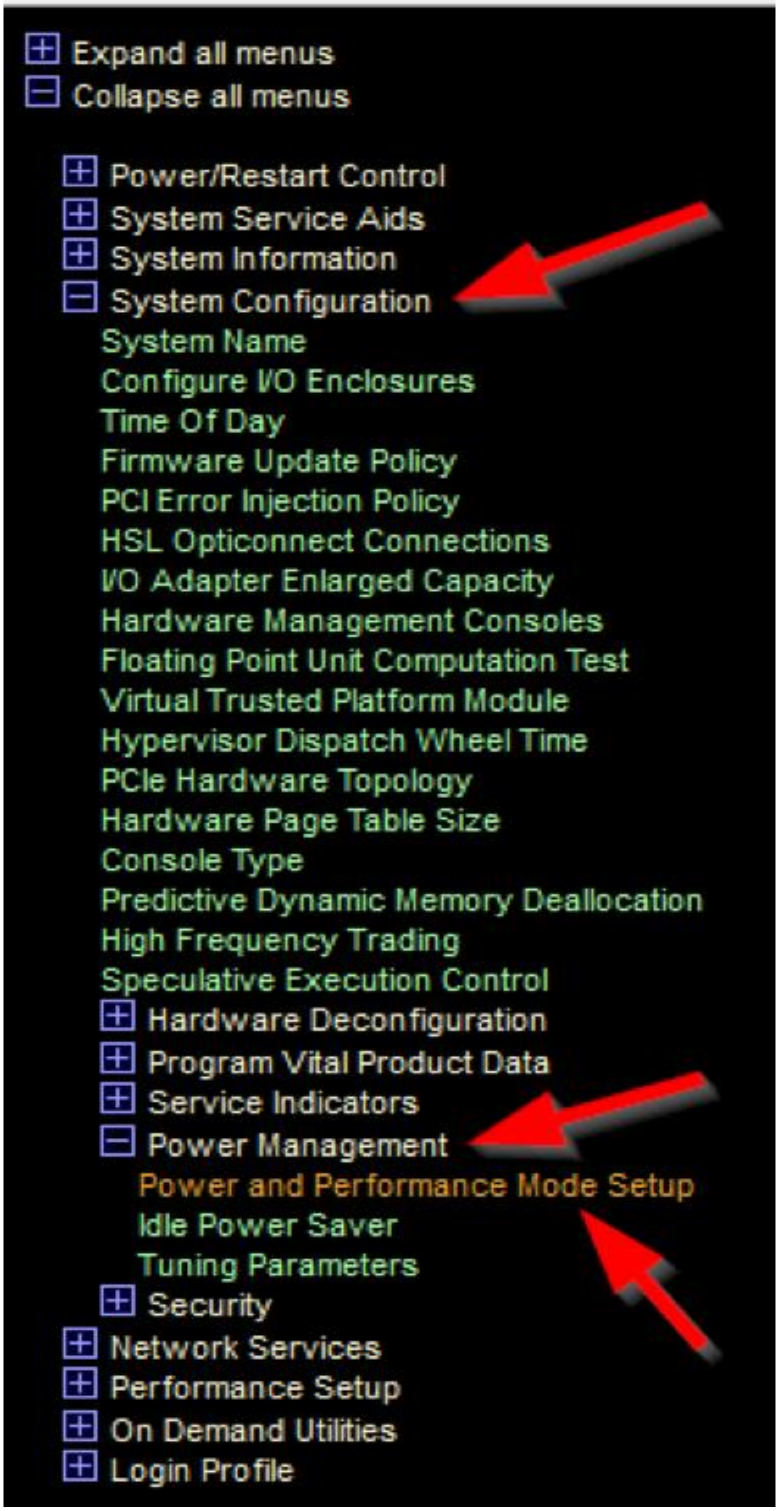
Where <mode> can be one of the supported modes from lspwrmgmt -m <managed system name> -r sys -F supported_power_saver_mode_types

Or to disable them and get nominal static:

# ***chpwrmgmt -m <managed system name> -r sys -o disable***

Advanced System Manager Interface (ASMI)

The figure below shows the ASMI menu options with the arrows pointing to the key selections.

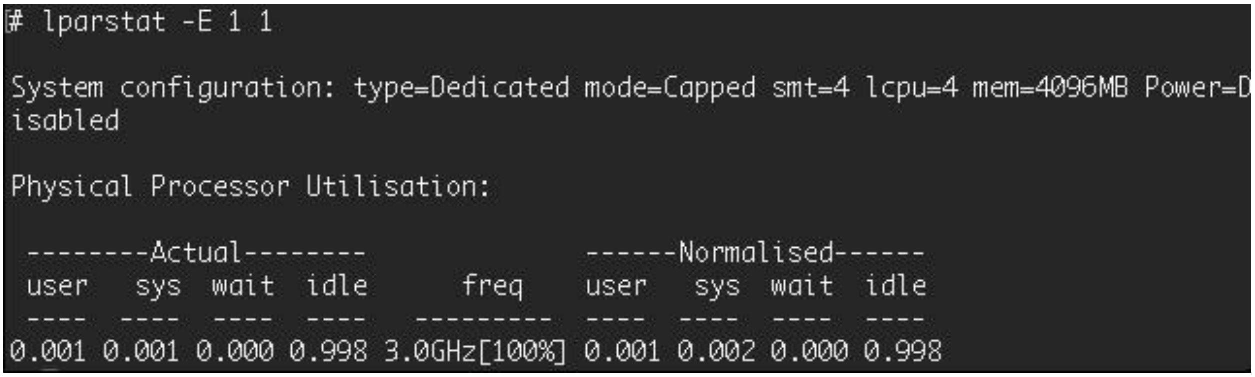


In addition to the four modes mentioned, there is the option of turning on Idle Power Saver. With Idle Power Saver enabled, the frequency will be reduced after long periods of complete system idleness (minutes). There are settings for idle power delay time and idle usage thresholds. Whether Idle Power Saver is on or off, the Dynamic Performance mode will still drop to the minimum frequency when there is little or no workload for milliseconds. When in Maximum Performance mode or the nominal mode, the frequency will only drop if Idle Power Saver is turned on. Further information on Idle Power Saver can be found at this web site: https://www.ibm.com/support/knowledgecenter/5148-22L/p8hby/ideal_power.htm.

Measuring the frequency from AIX

The recommended method for measuring core frequency in AIX is to use ***mpstat -E 1 1*** or ***lparstat -E 1 1***. The command ***mpstat -E 1 1*** provides individual core frequencies, while ***lparstat -E 1 1*** averages frequencies across all the cores in the lpar.

The ***lparstat –E 1 1*** command will show the current power saver mode and the current frequency:



```
# lparstat -E 1 1

System configuration: type=Dedicated mode=Capped smt=4 lcpu=4 mem=4096MB Power=D
isabled

Physical Processor Utilisation:

 --------Actual--------              ------Normalised------
 user   sys  wait  idle      freq    user   sys  wait  idle
 ----  ----  ----  ----   ---------  ----  ----  ----  ----
0.001 0.001 0.000 0.998 3.0GHz[100%] 0.001 0.002 0.000 0.998
```

Documentation for the lparstat and mpstat commands is available here:
https://www.ibm.com/support/knowledgecenter/en/ssw_aix_72/com.ibm.aix.cmds3/lparstat.htm
https://www.ibm.com/support/knowledgecenter/en/ssw_aix_72/com.ibm.aix.cmds3/mpstat.htm

Note: **The pmcycles command in AIX should NOT be used for reading the current processor frequency**. The recommended approach is to use lparstat - E 1 1 and mpstat -E 1 1 commands as discussed above.

Measuring the frequency from Linux

For Linux there are several ways to look at CPU frequencies depending on the flavor of Linux being used. The command, ***lscpu*** will show the running frequency on some flavors of Linux, but not all. The command ***dmesg*** (***dmesg | grep freq***) will also provide frequency information. The file /proc/cpuinfo contains frequency information

On some flavors of Linux, the commands below can be used.

Nominal frequency range:

cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_available_frequencies

Energy Scale Frequency range:

cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_boost_frequencies

Current running frequency of any core:

cat /sys/devices/system/cpu/cpu0/cpufreq/cpuinfo_cur_freq

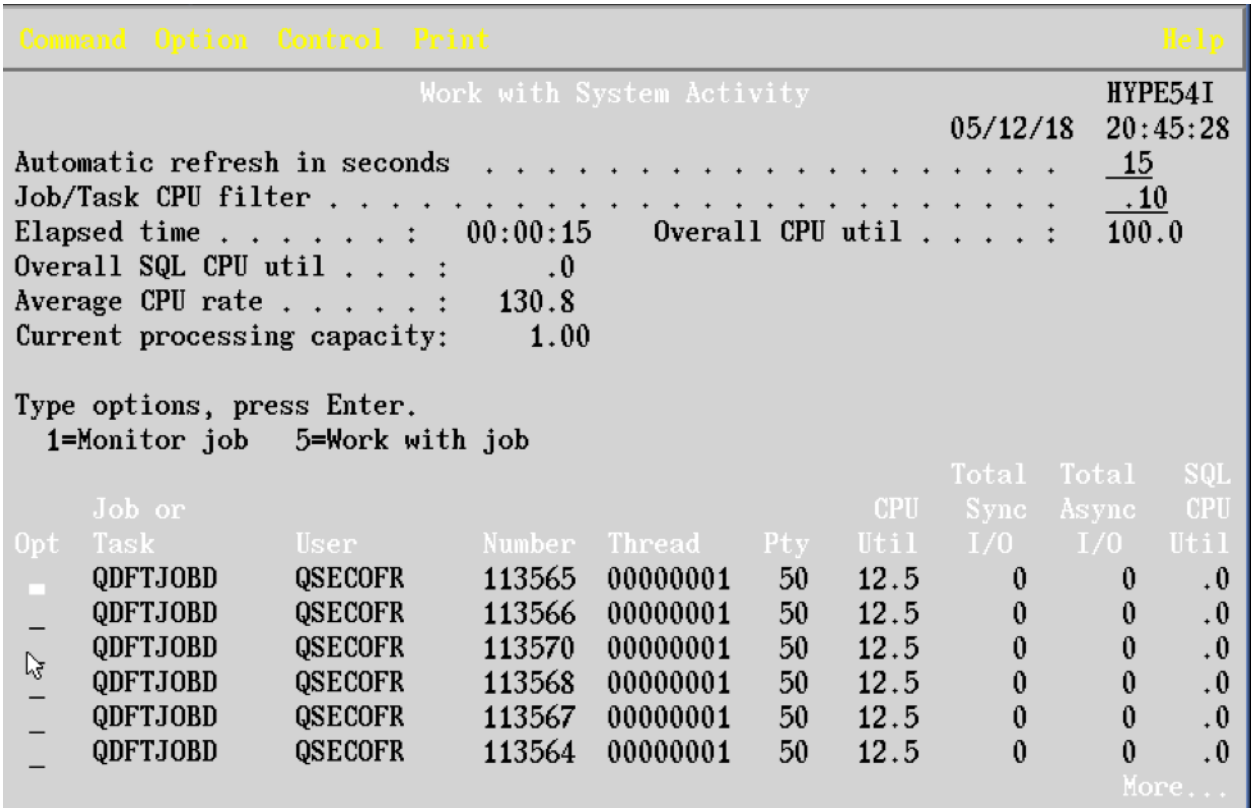
Note that in a purely Linux environment (with OPAL), the Linux operating system sets the frequency on a per core basis. In this case, the frequency is capped by the POWER On Chip Controller (OCC).

Measuring the frequency from IBM i

For the IBM i operating system, the following methods can be used to collect performance information.

IBM iDoctor for IBM i displays the CPU rate for the IBM i partition over time on the Collection Overview graph. The CPU rate for the partition is the ratio of scaled to unscaled processor utilized time, expressed as a percentage. There are two hardware registers that provide energy scaling information; the PURR and the SPURR. The Processor Utilization Register (PURR) is incremented monotonously as work is performed on a processor. The Scaled Processor Utilization register (PURR) is scaled in relation to the current frequency. If the processor is running at nominal frequency, the PURR and the SPURR will accumulate cycles at the same frequency. If the frequency is running higher than the nominal frequency, the ratio of the SPURR to the PURR will correspond to the increase in frequency. The processor utilized time reported by IBM i is the accumulation of non-idle virtual processor SPURR and PURR over each time interval.

The Work with System Activity (WRKSYSACT) command displays the Average CPU rate. The Average CPU rate for the partition is the ratio of scaled to unscaled processor utilized time, expressed as a percentage. The processor utilized time is the accumulation of non-idle virtual processor SPURR and PURR for the interval since the last refresh.



```
                           Work with System Activity                    HYPE54I
                                                             05/12/18  20:45:28
Automatic refresh in seconds  . . . . . . . . . . . . . . . . . . .     15
Job/Task CPU filter . . . . . . . . . . . . . . . . . . . . . . . .     .10
Elapsed time . . . . . . :   00:00:15    Overall CPU util . . . . :    100.0
Overall SQL CPU util . . . :       .0
Average CPU rate . . . . . :    130.8
Current processing capacity:      1.00

Type options, press Enter.
  1=Monitor job   5=Work with job
                                                                  Total  Total   SQL
     Job or                                               CPU     Sync   Async   CPU
Opt  Task          User         Number  Thread     Pty    Util     I/O    I/O   Util
     QDFTJOBD      QSECOFR      113565  00000001    50    12.5       0      0     .0
 _   QDFTJOBD      QSECOFR      113566  00000001    50    12.5       0      0     .0
     QDFTJOBD      QSECOFR      113570  00000001    50    12.5       0      0     .0
 _   QDFTJOBD      QSECOFR      113568  00000001    50    12.5       0      0     .0
 _   QDFTJOBD      QSECOFR      113567  00000001    50    12.5       0      0     .0
 _   QDFTJOBD      QSECOFR      113564  00000001    50    12.5       0      0     .0
                                                                             More...
```

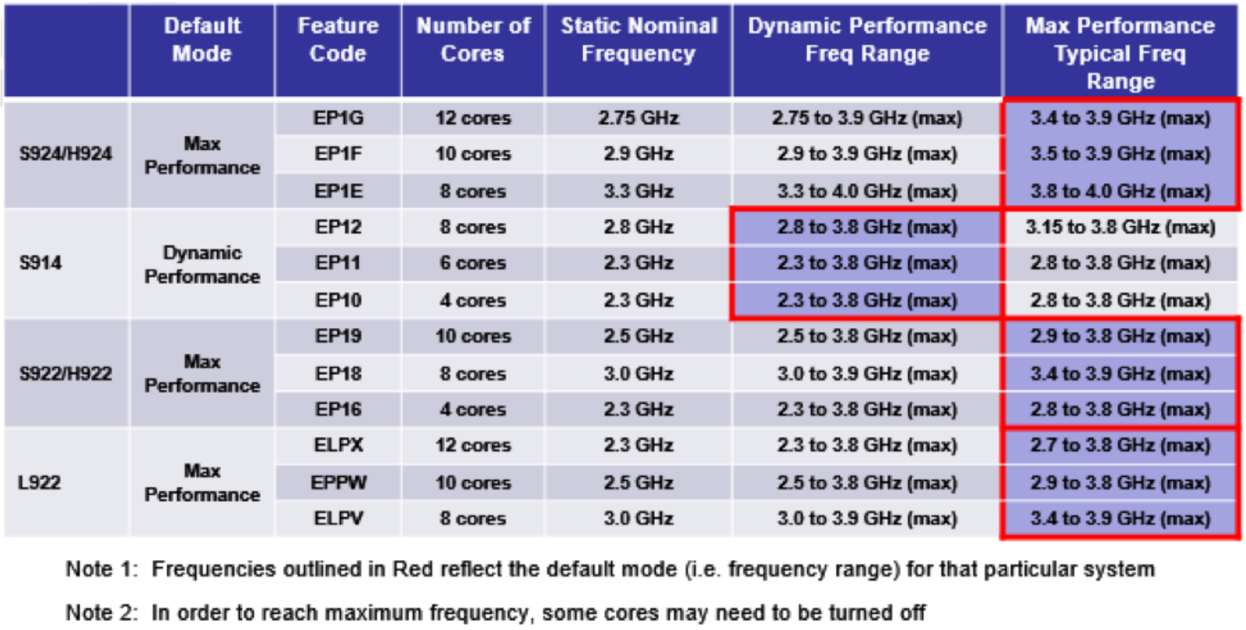
IBM i Collection Services Database file QAPMJOBMI contains time series data by task, primary thread, and secondary thread. Scaled and unscaled CPU times are available to calculate average CPU rate for processing activity of tasks and threads.

Database file QAPMSYSTEM contains time series system-wide (i.e. partition) accumulations of performance data. Scaled and unscaled CPU times are accumulated for various categories of processor usage. The ratio of scaled to unscaled time is the average CPU rate for the category of time accumulation. The processor utilized time is the accumulation of non-idle virtual processor SPURR and PURR for the time interval.

As of IBM i 7.3, the QAPMCONF database file key "NF" contains the processor nominal frequency in MHz. The processor nominal frequency can be used to convert average CPU rate to average processor frequency.

Frequency Range for POWER9 scale-out servers

The chart below lists the generally availability frequencies for the newly announced POWER9 scale-out systems. Note that these frequency values may change



	Default Mode	Feature Code	Number of Cores	Static Nominal Frequency	Dynamic Performance Freq Range	Max Performance Typical Freq Range
S924/H924	Max Performance	EP1G	12 cores	2.75 GHz	2.75 to 3.9 GHz (max)	3.4 to 3.9 GHz (max)
		EP1F	10 cores	2.9 GHz	2.9 to 3.9 GHz (max)	3.5 to 3.9 GHz (max)
		EP1E	8 cores	3.3 GHz	3.3 to 4.0 GHz (max)	3.8 to 4.0 GHz (max)
S914	Dynamic Performance	EP12	8 cores	2.8 GHz	2.8 to 3.8 GHz (max)	3.15 to 3.8 GHz (max)
		EP11	6 cores	2.3 GHz	2.3 to 3.8 GHz (max)	2.8 to 3.8 GHz (max)
		EP10	4 cores	2.3 GHz	2.3 to 3.8 GHz (max)	2.8 to 3.8 GHz (max)
S922/H922	Max Performance	EP19	10 cores	2.5 GHz	2.5 to 3.8 GHz (max)	2.9 to 3.8 GHz (max)
		EP18	8 cores	3.0 GHz	3.0 to 3.9 GHz (max)	3.4 to 3.9 GHz (max)
		EP16	4 cores	2.3 GHz	2.3 to 3.8 GHz (max)	2.8 to 3.8 GHz (max)
L922	Max Performance	ELPX	12 cores	2.3 GHz	2.3 to 3.8 GHz (max)	2.7 to 3.8 GHz (max)
		EPPW	10 cores	2.5 GHz	2.5 to 3.8 GHz (max)	2.9 to 3.8 GHz (max)
		ELPV	8 cores	3.0 GHz	3.0 to 3.9 GHz (max)	3.4 to 3.9 GHz (max)

Note 1: Frequencies outlined in Red reflect the default mode (i.e. frequency range) for that particular system

Note 2: In order to reach maximum frequency, some cores may need to be turned off

Varying frequencies is not new to Power systems, but new options are available, and the default mode that the system runs in has changed. What drove the change for the default mode? Why leave all that performance on the table, when the system is capable of so much more.

Summary

POWER9 EnergyScale gives you the flexibility to optimize your server to match the performance or energy requirements of your data center.

Contacting the PowerVM Team

Have questions for the PowerVM team or want to learn more? Follow our discussion group on LinkedIn IBM PowerVM