



# **IBM EnergyScale for Power10 Processor-Based Systems**

*July 2022*

*Martha Broyles*

*Christopher J. Cain*

*Chris Francois*

*Stuart Jacobs*

*Todd Rosedahl*

*Vaidyanathan Srinivasan*

*Brian Veale*

|  |    |
|--|----|
| Executive Overview .....                             | 4  |
| EnergyScale Features .....                           | 5  |
| Power and Thermal Trending .....                     | 5  |
| Performance Boost .....                              | 5  |
| Power Capping .....                                  | 5  |
| “Soft” Power Capping .....                           | 5  |
| Energy-Optimized Fan Control .....                   | 6  |
| Processor Stop States .....                          | 6  |
| Processor Folding .....                              | 6  |
| Processor Folding in AIX .....                       | 6  |
| Processor Folding in IBM i .....                     | 7  |
| Processor Folding in Linux .....                     | 8  |
| Safe Mode .....                                      | 9  |
| Reasons for Temporary Safe Mode .....                | 9  |
| Permanent Safe Mode .....                            | 9  |
| System Power and Performance Mode .....              | 10 |
| Query System Power and Performance Mode in AIX ..... | 12 |
| Static Mode .....                                    | 12 |
| Power Saving Mode .....                              | 12 |
| Maximum Performance Mode (MPM) .....                 | 14 |
| Idle Power Saver .....                               | 14 |
| EnergyScale for I/O .....                            | 14 |
| Frequency and Performance Measurement .....          | 14 |
| AIX .....  | 14 |
| IBM i .....  | 15 |
| User Interfaces .....                                | 16 |
| Overview .....                                       | 16 |
| Open BMC .....                                       | 18 |
| Setting System Power and Performance Mode .....      | 18 |
| Setting Power Cap .....                              | 19 |
| Idle Power Saver .....                               | 19 |
| ASMI .....   | 20 |
| Setting System Power and Performance Mode .....      | 20 |
| HMC .....  | 21 |
| Setting System Power and Performance Mode .....      | 21 |

|   |    |
|---|----|
| Redfish .....                                     | 23 |
| Appendix I: System Requirements .....             | 24 |
| Release Level FW1010 .....                        | 25 |
| Feature Support .....                             | 25 |
| Frequency .....                                   | 26 |
| Release Level FW1020 .....                        | 28 |
| Feature Support .....                             | 28 |
| Frequency .....                                   | 29 |
| Appendix II: Processor Usage and Accounting ..... | 30 |
| Appendix III: Resources .....                     | 32 |

## Executive Overview

The energy required to power and cool computers can be a significant cost to a business – reducing profit margins and consuming resources. In addition, the cost of creating power and cooling infrastructure can be prohibitive to business growth. In response to these challenges, IBM developed EnergyScale™ Technology for IBM® Power®. EnergyScale provides functions that help the user to understand and control IBM server power and cooling usage. This enables better facility planning, provides energy and cost savings, enables peak energy usage control, and increases system availability. Administrators may leverage EnergyScale capabilities to control the power consumption and performance of Power processor-based systems to meet their particular data center needs.

In this paper, the functions provided by EnergyScale are described along with usage examples, and hardware and software requirements. Support and awareness of EnergyScale extends throughout the system software stack, and is included in the AIX, IBM i, and Linux operating systems. This paper focuses on features and functions found in systems based on Power10 processors. For previous-generation IBM Power, please refer to companion paper, “IBM EnergyScale for Power9 Processor-based Systems”.

## EnergyScale Features

EnergyScale provides features to control the power consumption, energy efficiency, and performance of Power10 Systems.

### ***Power and Thermal Trending***

EnergyScale provides continuous collection of real-time server power consumption. Administrators may use the power information to predict data center power consumption at various times of the day, week, or month. In addition, data may be aggregated to identify anomalies, manage electrical loads, and enforce system-level power budgets.

A measured ambient temperature can help identify data center “hot-spots” that need attention. Please note that the ambient temperature reported may vary from system model to system model due to variations in the placement of the ambient temperature sensor in the system.

See [User Interfaces](#) for supported interfaces to collect power and thermal data.

### ***Performance Boost***

EnergyScale provides additional performance. The system is designed to provide enough power and cooling for the processors to run most workloads at a specified base frequency. However, most workloads do not consume all available power at the base frequency. In these cases, the frequency can be increased above that base frequency to take advantage of the additional power and thermal headroom in the system.

### ***Power Capping***

Power Capping enforces a user-specified limit on power consumption. See [User Interfaces](#) for supported interfaces to set a power cap. In most data centers and other installations, when a server is installed, a certain amount of power is allocated to it. Generally, the amount is what is considered to be a “safe” value, and it typically has a large margin of reserved, extra power that is never used. This is called the *margin power*. The main purpose of the power cap is not to save power but rather to give a data center operator the capability to reallocate power from existing systems to new systems by reducing the margin assigned to the existing servers. That is, power capping gives an operator the capability to add extra servers to a data center which previously had all available power allocated to its existing systems. It does this by guaranteeing that a system will not use more power than assigned to it by the operator. This is also called a “hard power cap”.

Previously, the data center administrator had to plan for the power consumption of the data center based on the Underwriters' Laboratories (UL) rating on the back of the servers being installed. The UL rating (commonly referred to as “label power”) on today's servers indicates the most power that a system could ever draw and is based on the capacity of the power supplies. It has to take into account a fully-configured system with the highest power-usage parts installed at the highest possible utilization.

#### **“Soft” Power Capping**

There are two power ranges into which the power cap may be set. When a power cap is set in the guaranteed range (“hard power cap” described above), the system is guaranteed to use less power than the cap setting. In order to meet this guarantee, extreme system configuration and environmental conditions must be accounted for. Setting a power cap in this region allows for the recovery of the margin power, but in many cases cannot be used to save power. Soft power capping extends the allowed power capping range further, beyond a region that can be guaranteed in all configurations and

conditions. By setting a power cap in this soft region, the system can be set to save power by running at a lower power/performance point. If the power management goal is to meet a particular power consumption limit, then soft power capping is the mechanism to use. Note that the failure to enforce a “soft” power cap below the minimum guaranteed range is not an error, and will not result in any error. The system will continue to operate normally with all EnergyScale features at the minimum-supported frequency until the soft power cap is disabled or raised.

## ***Energy-Optimized Fan Control***

Cooling fans contribute significantly to the overall power consumption of a given computer. In order to minimize energy expended on cooling and to minimize the energy wasted “over-cooling” a system, firmware on all Power10 systems will adjust fan speed in response to real-time temperatures of the system components.

## ***Processor Stop States***

The IBM Power processors uses a combination of clock and power gating to achieve different levels of power saving and associated latency to enter and exit these levels known as stop states. For more information see

<https://community.ibm.com/community/user/power/blogs/amit-tendolkar1/2020/09/20/an-overview-of-idle-states-in-the-power9-processor>

## ***Processor Folding***

While Processor Core Nap and sleep provide energy savings and performance boost when processors become idle, additional savings/boost can be realized if processors remain idle by intent rather than by happenstance. Processor Folding is a consolidation technique that dynamically adjusts, over the short-term, the number of processors available for dispatch to match the number of processors demanded by the workload. As the workload increases, the number of processors made available increases; as the workload decreases, the number of processors made available decreases. Processor Folding increases energy savings/boost during periods of low to moderate workload because unavailable processors remain in low-power idle states longer than they otherwise would. Processor Folding achieves power savings similar to those that could be achieved by intelligent, utilization-based logical partition (LPAR) configuration changes, but it does so with much greater efficiency and fidelity, and without impacting the configuration or processor utilization of the LPAR.

### **Processor Folding in AIX**

In AIX the processor folding policy can be configured from the command line via the `schedo` command. The `vpm_fold_policy` tunable is a 4-bit value where each bit indicates the configuration of a different setting. The following table shows the various settings that are controlled.

| <i>Bit</i> | <i>Setting</i>   |
|------------|--|
| <b>0</b>   | =1 processor folding is enabled when the partition is using shared processors    |
| <b>1</b>   | =1 processor folding is enabled when the partition is using dedicated processors |

| <i>Bit</i> | <i>Setting</i>  |
|------------|---|
| <b>2</b>   | =1 disables the automatic setting of processor folding when the partition is in Power Saving mode |
| <b>3</b>   | =1 processor affinity will be ignored when making folding decisions                               |

Table 1: `vpm_fold_policy` is a 4-bit value, in which each bit controls an aspect of folding.

The following command displays the current setting of `vpm_fold_policy`:

```
# schedo -L vpm_fold_policy
NAME                CUR    DEF    BOOT    MIN    MAX    UNIT  TYPE
DEPENDENCIES
-----
vpm_fold_policy     1      1      1       0     15     D
-----
```

To enable processor folding on a partition using dedicated partitions when the current value of `vpm_fold_policy` is set to 1, the following command would be issued to set the value to 3:

```
# schedo -o vpm_fold_policy=3
```

To disable processor folding, the value of `vpm_fold_policy` can be set to the value 4 using the following command:

```
# schedo -o vpm_fold_policy=4
```

By default, AIX will attempt to consider processor affinity (or topology) information when making processor folding decisions. This allows for the workload to remain spread across the processor nodes (e.g., chips depending on the system) and benefit from improved performance. Bit 3 of the `vpm_fold_policy` tunable allows this default behavior to be disabled. For example, if `vpm_fold_policy` is currently set to 6, indicating that processor folding is enabled when the partition is using dedicated partitions and that the partition will automatically enable processor folding when in Power Saving mode, the following command would change the setting to indicate that the operating system should no longer consider processor affinity when making folding decisions:

```
# schedo -o vpm_fold_policy=14
```

For more information see the help information via the `-h` option of the `schedo` command:

```
# schedo -h vpm_fold_policy
```

## Processor Folding in IBM i

In an IBM i partition, processor folding is configured and controlled by the operating system by default. On Power10 servers, the operating system enables processor folding by default in shared processor LPARs or when Power Saving mode is enabled. Operating system control of processor folding may be overridden via the **QWCCTLSW** limited availability API, which provides a key-based control language programming interface to a limited set of IBM i tunable parameters. Processor folding control is accessed via **QWCCTLSW** key 1060. The following sequence of calls cycles through the various options. Changes to key 1060 take effect immediately but do not persist across IPLs.

Get current status:

```
> CALL QWCCTLSW PARM('1060' '1')
KEY 1060 IS *SYSCTL.
KEY 1060 IS SUPPORTED ON THE CURRENT IPL.
KEY 1060 IS CURRENTLY ENABLED.
```

Explicitly disable processor folding:

```
> CALL QWCCTLSW PARM('1060' '3' )
KEY 1060 SET TO *OFF.
```

Explicitly enable processor folding:

```
> CALL QWCCTLSW PARM('1060' '2' 1)
KEY 1060 SET TO *ON.
```

Re-establish system control of processor folding:

```
> CALL QWCCTLSW PARM('1060' '2' 2)
KEY 1060 SET TO *SYSCTL.
```

## Processor Folding in Linux

It is essential to install a daemon package based on the host OS to enable utilization-based processor folding for Power Saving:

```
pseries-energy-1.4.0-1.el7.ppc64.rpm
pseries-energy-1.4.0-1.el6.ppc64.rpm
pseries-energy-1.4.0-1.sles11.ppc64.rpm
```

Version 5.4 has the necessary user space tools required to enable CPU Folding.<sup>1</sup>

Once this package is installed, the `energyd` daemon will monitor the system power mode and activate processor folding when system power mode is set to "Power Saving" and deactivate processor folding in all other modes. The utilization-based CPU folding daemon will deactivate unused cores and transition them to low power idle states until the CPU utilization increases and those cores are activated to run a workload.

Utilization-based processor folding can be manually disabled using the following commands:

```
/etc/init.d/energyd stop #Stop daemon now, activate all cores
chkconfig energyd off    #Do not restart daemon on startup
```

-or-

```
rpm -e pseries-energy    #un-install the package completely
```

Alternatively, CPU cores can be folded or set to low power idle state in any power mode manually using the following command line:

---

<sup>1</sup>Refer to <http://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/installtools/home.html> for more details.



```
echo 0 > /sys/devices/system/cpu/cpuN/online #Where N is the
logical CPU number
```

Please note that all active hardware threads of a core need to be taken off-line using the above command in order to move the core to a low power idle state.

The cores can be activated again with the following command:

```
echo 1 > /sys/devices/system/cpu/cpuN/online #Where N is the
logical CPU number
```

## Safe Mode

All systems and firmware releases support safe mode. “Safe Mode” is a system mode where the firmware will automatically drop to a fixed lower frequency to keep the system thermal and power safe

### Reasons for Temporary Safe Mode

The following *may* cause the system to temporarily enter safe mode, no user action is required to exit safe mode from these conditions and no user notification is provided when it occurs.

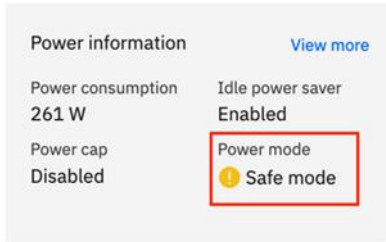
- Concurrent code update
- Error recovery

### Permanent Safe Mode

Certain hardware and firmware failures that cannot be recovered from can cause the system firmware to stay in safe mode. There will always be an error log generated when safe mode is entered due to a hard failure. In addition to a posted error log safe mode will also be indicated in the ASMI Power and Performance Mode Setup menu:

The screenshot shows the IBM Advanced System Management (ASMI) interface. At the top, there is a header with the IBM logo and the text "Advanced System Management". To the right of the header, it says "Copyright © 2002, 2021 IBM Corporation. All rights reserved." Below the header, there is a "Log out" button. The main content area is divided into two columns. The left column contains a list of menu items, including "Time Of Day", "Firmware Update Policy", "PCI Error Injection Policy", "Monitoring", "HSL Opticonnect Connections", "I/O Adapter Enlarged Capacity", "Hardware Management Consoles", "Virtual I/O Connections", "Firmware License Agreement", "PCIe Hardware Topology", "Hardware Page Table Size", "Estimated Corrosion Rates", "Console Type", "Predictive Dynamic Memory", "Deallocation", "High Frequency Trading", "Hardware Deconfiguration", "Program Vital Product Data", "Service Indicators", "Power Management", and "Power and Performance Mode". The right column displays the "Power and Performance Mode Setup" menu. The "Current Power Saver Mode" is highlighted as "Safe mode". Below this, there is a message: "The system is currently operating with reduced performance due to Safe Mode. Please review Error/Event logs to determine the cause. The system will attempt to exit Safe Mode on the next power on attempt or Service Processor reset."

Or on the BMC GUI:

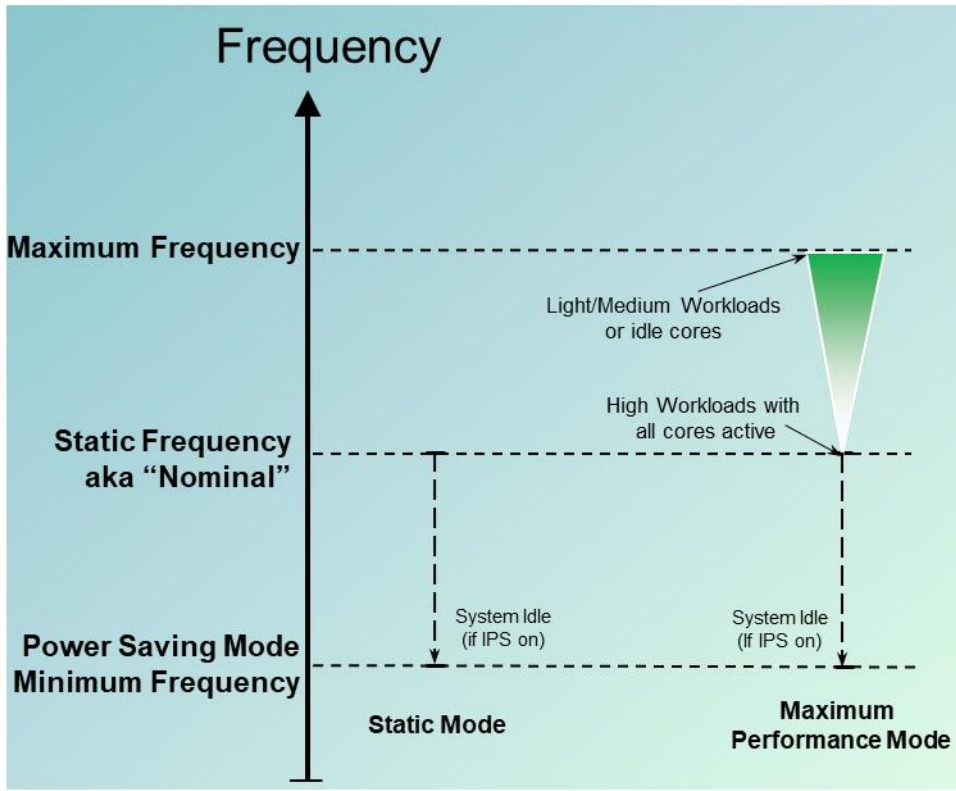


To re-enable full EnergyScale functionality, a reboot, firmware update, FRU replacement, or complete A/C power cycle of the system must be completed. If the problem which originally caused the system to enter safe mode is still present, the system will re-enter safe mode and generate an additional error log entry.

## ***System Power and Performance Mode***

The mode that a system is in can be queried from ASMI, BMC, HMC or the OS but it can only be changed from ASMI, BMC or the HMC.

| <b><i>Power9</i></b>                         | <b><i>Power10</i></b> | <b><i>Notes</i></b>   |
|--|-----------------------|---|
| Modes Disabled also referred to as “nominal” | Static                | Re-name only. Functionally the same between Power9 and Power10                        |
| Static Power Saver                           | Power Saving          | Re-name only. Functionally the same between Power9 and Power10                        |
| Dynamic Performance                          | Not supported         |   |
| Maximum Performance                          | Maximum Performance   | No change. All Power10 systems will be shipped in Maximum Performance mode by default |



The System Power and Performance mode setting persists across system boots, service processor resets and loss of ac power (unless the power outage is long enough to drain the Service Processor NVRAM battery).

## Query System Power and Performance Mode in AIX

Starting with AIX 7.2 TL 3 SP 5 and AIX 7.1 TL 5 SP 9, lparstat -N displays information about Power and Performance Modes, Processor Frequency, and Processor Folding.

### # lparstat -N

```
Power and Performance Mode      : Maximum Performance
Idle Power Saver Status         : Active ← System has met Idle Power Saver enter criteria, frequency is at minimum
Minimum Frequency (Mhz)        : 2000
Static Frequency (Mhz)         : 3200
Maximum Frequency (Mhz)        : 4000
Processor Folding Status        : Enabled
```

### # lparstat -N

```
Power and Performance Mode      : Maximum Performance
Idle Power Saver Status         : Enabled ← IPS is enabled but system is not actively idle, frequency is normal
Minimum Frequency (Mhz)        : 2000
Static Frequency (Mhz)         : 3200
Maximum Frequency (Mhz)        : 4000
Processor Folding Status        : Disabled
```

### # lparstat -N

```
Power and Performance Mode      : Maximum Performance
Idle Power Saver Status         : Not Supported
Minimum Frequency (Mhz)        : 3251
Static Frequency (Mhz)         : 3550
Maximum Frequency (Mhz)        : 3900
Processor Folding Status        : Disabled
```

## Static Mode

This is also known as "Nominal" mode. In this mode, the frequency is set to fixed value that can run all normal workloads under all normal environmental conditions. There is no frequency boost based on idle cores, lower workloads, nor favorable ambient conditions.

## Power Saving Mode

Power Saving mode lowers the processor frequency and voltage to the minimum, reducing the power consumption of the system while still delivering predictable performance. In addition, this mode automatically enables processor folding in dedicated processor partitions. See [Appendix I](#) for details on the actual Power Saving frequency by system and firmware release.

Power Saving could be enabled based on regular variations in workloads, such as predictable dips in utilization overnight, or over weekends. Power Saving can be used to reduce peak energy consumption, which can lower the cost of all power used. Please note that when Power Saving is enabled for certain workloads with low CPU utilization, workload performance will not be impacted, though CPU utilization may increase due to the reduced processor frequency.

The only time that a system does not support operating at the Power Saving voltage and frequency is during a system boot or re-boot. Power Saving may be enabled at any time; however, if Power Saving was enabled prior to a system boot the voltage and frequency will remain at the default boot values until the platform firmware reaches a standby or running state. Immediately before the platform

firmware starts executing on the system's processors, the voltage and frequency will drop to the Power Saving mode values. If a system re-boot occurs while in Power Saving mode, the voltage and frequency will return back to boot values and following a successful system re-boot the voltage and frequency will be dropped back to Power Saving mode.

## Maximum Performance Mode (MPM)

Maximum Performance mode will allow the system to reach the maximum frequency by taking advantage of the thermal and power headroom provided by idle cores, lower workloads, less I/O devices attached to the processor and favorable environmental conditions.

## Idle Power Saver

This mode, which is enabled/disabled independently from all other modes and functions, reduces the energy usage to a low level *when the entire system is determined to be idle*. When idle, the voltage/frequency of the processor is reduced to the minimum and static power saver mode is reported to the OS to enable processor folding. This can cause OS folding policy values to be updated at runtime. When not idle the system is managed in accordance with the configured power mode (i.e. Maximum Performance). Idle Power Saver can be enabled/disabled via BMC. Additionally, the utilization levels that determine idleness and the time delays for entry/exit can also be modified from the BMC. See [Appendix I](#) for system support details such as which systems support this mode and which systems have it enabled by default.

## EnergyScale for I/O

IBM Power automatically power off pluggable PCI adapter slots that are not being used to save approximately 14 watts per slot. A PCI adapter slot is considered not being used when the slot is empty, when the slot is not assigned to a partition, or when the partition to which the slot is assigned is not powered on. A PCI slot is powered off immediately by system firmware when it is dynamically removed from the partition to which it was assigned, and when the partition to which it is assigned is powered off. Furthermore, system firmware automatically scans all pluggable PCI slots at regular intervals looking for those that meet the criteria for being not in use and powers them off. This ensures among other things that slots left on after platform power-on are subsequently powered off if they are not in use. This is supported on all Power10 processor-based systems, and the expansion units that they support. Note that it applies to hot-pluggable PCI slots only. Power controls for other types of I/O features and built-in, or embedded, PCI adapters are not available and so they cannot be powered off independently from their enclosure power.

## Frequency and Performance Measurement

Methods to measure the frequency and performance vary by operating system.

### AIX

lparstat -E and mpstat -E are used to view the current processor frequency

- lparstat -E reports the frequency averaged across all of the Virtual Processors assigned to the LPAR
- mpstat -E reports the frequency per Virtual Processor

Usage: lparstat/mpstat -E [ *Interval* [ *Count* ] ]

```
# lparstat -E 1 1
System configuration: type=Dedicated mode=Capped smt=4 lcpu=4 mem=4096MB Power=Disabled
Physical Processor Utilization:
-----Actual-----          -----Normalised-----
user  sys  wait  idle  freq          user  sys  wait  idle
----  -
0.001 0.002 0.000 0.997  3.0GHz [100%] 0.001 0.002 0.000 0.997
```

```
# mpstat -E 1 1
System configuration: lcpu=4 mode=Capped
vcpu      pbusy      physc      freq      scaled physc
----      -
0         0.0022[0%]  1.0005[100%] 3.0GHz [100%] 1.0048[100%]
```

The *Interval* parameter must be supplied to lparstat/mpstat -E to read the current frequency. Without the interval parameter, the commands generate a single report containing statistics since the last boot of the LPAR by default.

**NOTE: For AIX version 7.2 and earlier, pmcycles command in AIX should NOT be used for reading the current processor frequency.** The recommended approach is to use lparstat -E and mpstat -E as discussed above.

References:

[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/lparstat.htm)

[https://www.ibm.com/support/knowledgecenter/en/ssw\\_aix\\_72/com.ibm.aix.cmds3/mpstat.htm](https://www.ibm.com/support/knowledgecenter/en/ssw_aix_72/com.ibm.aix.cmds3/mpstat.htm)

## IBM i

### **IBM iDoctor for IBM i**

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. The processor utilized time is the accumulation of non-idle virtual processor SPURR and PURR<sup>2</sup> over each time interval. The ratio of SPURR and PURR accumulated over an interval represents the processor frequency versus nominal over the interval.

### **WRKSYSACT**

The Work with System Activity (WRKSYSACT) command displays the Average CPU rate since last refresh for the partition in output shown on the display station. 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 accumulation of non-idle virtual processor SPURR and PURR for the interval since the last refresh.

### **IBM i Collection Services**

Database file QAPMJOBMI contains time series data by task, primary thread, and secondary thread. Scaled and unscaled CPU times, both charged and used, 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 accumulation of non-idle virtual processor SPURR and PURR for the time interval.

Note: 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.

---

<sup>2</sup> See [Appendix II](#) for more info.

## User Interfaces

### Overview

The table below summarizes the BMC, ASMI, HMC and Redfish support. Refer first to [Appendix I: System Requirements](#) to know if the specific interface is supported for a particular release and system.

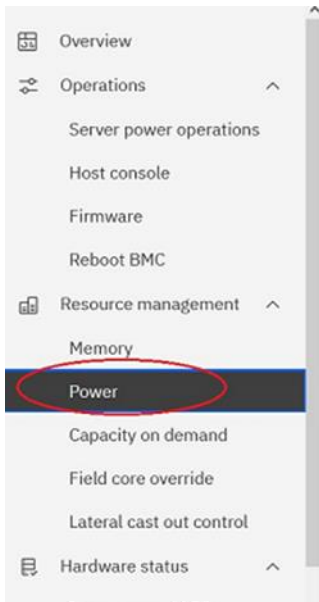
|                                    | <i>BMC</i> | <i>ASMI</i> | <i>HMC</i> | <i>Redfish</i> |
|------------------------------------|------------|-------------|------------|----------------|
| <b>Power and Thermal Reporting</b> | Y          | n/a         | n/a        | Y              |
| <b>Maximum Performance Mode</b>    | Y          | Y           | Y          | Y              |
| <b>Static Mode</b>                 | Y          | Y           | Y          | Y              |
| <b>Power Saving Mode</b>           | Y          | Y           | Y          | Y              |
| <b>Idle Power Saver</b>            | Y          | n/a         | n/a        | Y              |
| <b>Power Capping</b>               | Y          | n/a         | n/a        | n/a            |





## Open BMC

Supported EnergyScale features can be found on the BMC under the “Power” menu.



## Setting System Power and Performance Mode

Power and performance mode

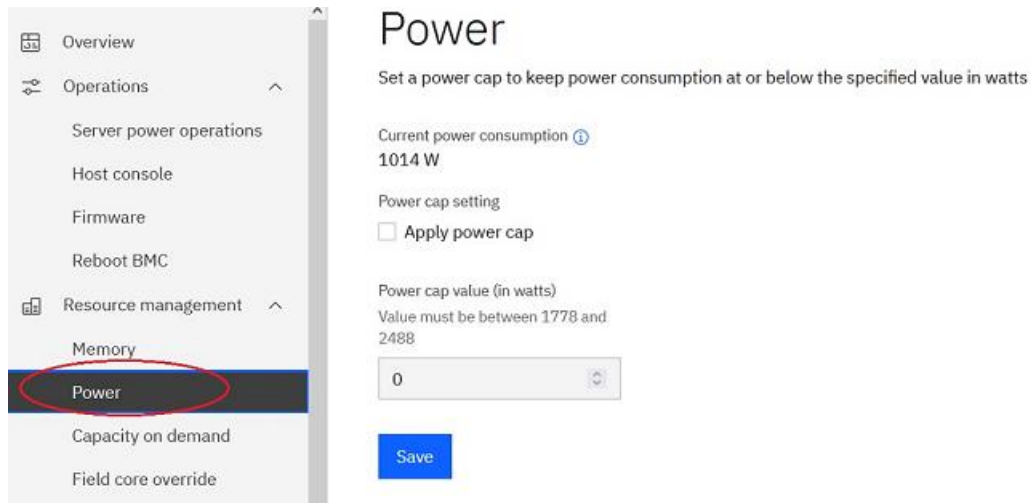
[^ View effects of each mode](#)

Select mode

- Static
- Power saving
- Maximum performance

[Update power saver mode](#)

## Setting Power Cap



The screenshot shows a management interface with a sidebar on the left and a main content area on the right. The sidebar has a menu with items: Overview, Operations (with a sub-menu: Server power operations, Host console, Firmware, Reboot BMC), Resource management (with a sub-menu: Memory, Power, Capacity on demand, Field core override). The 'Power' item is highlighted with a red oval. The main content area is titled 'Power' and contains the following text: 'Set a power cap to keep power consumption at or below the specified value in watts'. Below this, it shows 'Current power consumption 1014 W' with a help icon. Under 'Power cap setting', there is an unchecked checkbox for 'Apply power cap'. Below that, it says 'Power cap value (in watts)' and 'Value must be between 1778 and 2488'. A numeric input field contains the value '0'. At the bottom of the main content area is a blue 'Save' button.

## Idle Power Saver

### Idle power saver

Enable idle power saver

#### To enter

Delay time (in seconds)

240

Utilization threshold (in %)

8

#### To exit

Delay time (in seconds)

10

Utilization threshold (in %)

12

Update idle power saver

Reset to default

# ASMI

Supported EnergyScale features can be found on the Advanced System Management Interface (ASMI) under the “System Configuration” → “Power Management” menu.

## Setting System Power and Performance Mode

The screenshot shows the ASMI interface with the following elements:

- Header:** IBM logo, "Advanced System Management", and "Copyright © 2002, 2021 IBM Corporation. All rights reserved."
- Navigation:** "Log out" button.
- Left Sidebar (Menu):**
  - FISL Opticonnect Connections
  - I/O Adapter Enlarged Capacity
  - Hardware Management Consoles
  - Virtual I/O Connections
  - Firmware License Agreement
  - PCIe Hardware Topology
  - Hardware Page Table Size
  - Estimated Corrosion Rates
  - Console Type
  - Predictive Dynamic Memory
  - Deallocation
  - High Frequency Trading
  - Hardware Deconfiguration
  - Program Vital Product Data
  - Service Indicators
  - Power Management
  - Power and Performance Mode Setup** (circled in red)
  - Security
- Main Content Area:**
  - Power and Performance Mode Setup**
  - Current Power Saver Mode : Maximum Performance mode
  - Radio button options:
    - Static mode ?
    - Power Saving mode ?
    - Maximum Performance mode ?
  - Note: Enabling any of the Power Saver modes will cause changes in the processor frequencies, changes in processor utilization, changes in power consumption, and performance to vary. Other effects are possible as well. Please see the EnergyScale™ white paper for more information on power saving modes.
  - Continue ? button

## HMC

The HMC (Hardware Management Console) is a management console that controls managed systems, logical partitions, managed frames, other features provided through the managed objects, and the HMC itself. The HMC provides both graphical user interface (GUI) and command line interfaces. Users can use the user interfaces to configure or manage various features offered by the managed objects.

### ***Setting System Power and Performance Mode***

The user can list which power management modes are supported using the `lspwrmgmt` command on the command line:

```
lspwrmgmt -m <managed system name> -r sys -F supported_power_saver_mode_types
```

Example output:

```
"static,power_saving,max_perf"
```

A user can then enable one of the supported power management modes using the `chpwrmgmt` command on the command line as illustrated below enabling Maximum Performance mode:

```
chpwrmgmt -m <managed system name> -r sys -o enable -t max_perf
```

To query the current power management mode use `lspwrmgmt` command on the command line as illustrated below:

```
lspwrmgmt -m <managed system name> -r sys
```

Example output:

```
curr_power_saver_mode=Enabled,curr_power_saver_mode_type=max_perf,desired_power_saver_mode=Enabled,desired_power_saver_mode_type=max_perf,"supported_power_saver_mode_types=static,power_saving,max_perf",idle_power_saver_mode=unavailable
```

NOTE: The new mode may not take effect immediately. Normally, if the operation is performed before the system is powered on, the desired mode won't take effect until the system is up and running. If the mode is in transition, any changes will be blocked.

With the HMC GUI, users can reach this task by selecting the managed system -> Operations -> Power Management.



## **Redfish**

Redfish is an Open industry standard specification for hardware management. All resources linked from the Service root /redfish/v1/

Power and thermal sensors are available from /redfish/v1/Chassis/chassis/Sensors

For more information on sensor collection see

[https://www.dmtf.org/sites/default/files/standards/documents/DSP2051\\_1.0.0.pdf](https://www.dmtf.org/sites/default/files/standards/documents/DSP2051_1.0.0.pdf)

## **Appendix I: System Requirements**

Due to differences in each release, this appendix details the systems and EnergyScale features supported by release including the actual frequency limits for various power management modes.



# Release Level FW1010

## Feature Support

Refer to the EnergyScale Features chapter earlier in this document for a definition of each feature.

|                                    | Power E1080                           |
|------------------------------------|---------------------------------------|
| <b>Power and thermal reporting</b> | n/a                                   |
| <b>Power Saving Mode</b>           | ASM, HMC                              |
| <b>Maximum Performance Mode</b>    | <b>Enabled by Default</b><br>ASM, HMC |
| <b>Static Mode</b>                 | ASM, HMC                              |
| <b>Idle Power Saver</b>            | n/a                                   |
| <b>User Set Power Capping</b>      | n/a                                   |

## Frequency

Depending upon the Power savings setting selected, the maximum and minimum frequency may change. For a definition of each setting, please refer to the EnergyScale Features chapter earlier in this document.

### Support Notes

<sup>1</sup> Note that CPU frequencies greater than the system base are *not* guaranteed. The actual maximum frequency may vary based on environmental conditions, system configuration, firmware version, component tolerances, and workload.

<sup>2</sup> For frequency to be guaranteed to be in the normal operating frequency range the ambient must be below this temperature when in Maximum Performance mode

|                       | <i>Guaranteed Ambient Temperature<sup>2</sup></i> | <i>Maximum Performance Mode Typical Operating Frequency Range<sup>1</sup></i> | <i>Power Saving Mode / Minimum Frequency</i> |
|-----------------------|---|---|--|
| <b>E1080 @3.55GHz</b> | 27C   | 3.55 - 4.00 GHz   | 3.25 GHz                                     |
| <b>E1080 @3.60GHz</b> | 27C   | 3.60 – 4.15 GHz   | 3.4 GHz                                      |
| <b>E1080 @3.65GHz</b> | 27C   | 3.65 - 3.90 GHz   | 3.25 GHz                                     |



# Release Level FW1020

## Feature Support

Refer to the EnergyScale Features chapter earlier in this document for a definition of each feature.

|                                    | Power S1014                           | Power S1022                           | Power S1024                           | Power E1050                           |
|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
| <b>Power and thermal reporting</b> | Redfish, BMC                          | Redfish, BMC                          | Redfish, BMC                          | Redfish, BMC                          |
| <b>Power Saving Mode</b>           | BMC, HMC                              | BMC, HMC                              | BMC, HMC                              | BMC, HMC                              |
| <b>Maximum Performance Mode</b>    | <b>Enabled by Default</b><br>BMC, HMC | <b>Enabled by Default</b><br>BMC, HMC | <b>Enabled by Default</b><br>BMC, HMC | <b>Enabled by Default</b><br>BMC, HMC |
| <b>Static Mode</b>                 | BMC, HMC                              | BMC, HMC                              | BMC, HMC                              | BMC, HMC                              |
| <b>Idle Power Saver</b>            | <b>Enabled by Default</b><br>BMC, HMC | <b>Enabled by Default</b><br>BMC, HMC | <b>Enabled by Default</b><br>BMC, HMC | <b>Enabled by Default</b><br>BMC, HMC |
| <b>User Set Power Capping</b>      | BMC                                   | BMC                                   | BMC                                   | BMC                                   |

## Frequency

Depending upon the Power savings setting selected, the maximum and minimum frequency may change. For a definition of each setting, please refer to the EnergyScale Features chapter earlier in this document.

### Support Notes

<sup>1</sup> Note that CPU frequencies greater than the system base are *not* guaranteed. The actual maximum frequency may vary based on environmental conditions, system configuration, firmware version, component tolerances, and workload.

<sup>2</sup> For frequency to be guaranteed to be in the normal operating frequency range the ambient must be below this temperature when in Maximum Performance mode

|                               | <i>Guaranteed Ambient Temperature<sup>2</sup></i> | <i>Maximum Performance Mode Typical Operating Frequency Range<sup>1</sup></i> | <i>Power Saving Mode / Minimum Frequency</i> |
|-------------------------------|---|---|--|
| <b>S1022 @2.45GHz</b>         | 27C   | 2.45 - 3.90 GHz   | 2.0 GHz                                      |
| <b>S1024 @2.75GHz</b>         | 27C   | 2.75 – 3.90 GHz   | 2.0 GHz                                      |
| <b>S1022 @2.75GHz</b>         | 27C   | 2.75 – 4.00 GHz   | 2.0 GHz                                      |
| <b>S1022 @2.90GHz</b>         | 27C   | 2.90 – 4.00 GHz   | 2.0 GHz                                      |
| <b>S1014 / S1022 @3.00GHz</b> | 27C   | 3.00 - 3.90 GHz   | 2.0 GHz                                      |
| <b>S1024 @3.10GHz</b>         | 27C   | 3.10 - 4.00 GHz   | 2.0 GHz                                      |
| <b>S1024 @3.40GHz</b>         | 27C   | 3.40 - 4.00 GHz   | 2.0 GHz                                      |
| <b>E1050 @2.95GHz</b>         | 27C   | 2.95 – 3.90 GHz   | 2.0 GHz                                      |
| <b>E1050 @3.20GHz</b>         | 27C   | 3.20 - 4.00 GHz   | 2.0 GHz                                      |
| <b>E1050 @3.35GHz</b>         | 27C   | 3.35 - 4.00 GHz   | 2.0 GHz                                      |

## Appendix II: Processor Usage and Accounting

For historical reasons, process accounting charges and processor utilization are usually formulated in terms of time. The processors used in early time-sharing computer systems were fixed-frequency and single-threaded; processing capacity per unit of time was relatively constant. Consequently, it was convenient to express process accounting charges in terms of processor time and processor utilization as the percentage of time that the processor was not idle over an interval of interest.

With the introduction of multi-threaded processors (i.e. processors capable of executing multiple programs simultaneously), it was no longer desirable to report processor utilization based on the percentage of time that the processor was not idle. Consider a processor that supports 2-way Simultaneous Multi-Threading (SMT). When both threads are idle, utilization should be 0%. When both threads are not idle, utilization should be 100%. When one thread is idle and the other is not idle, there are several options:

1. Treat the processor as idle (0% utilization). This is obviously wrong, as the process is consuming more than 50% of the processing capacity.
2. Continue to treat the processor as not idle and charge the process with 100% of the time (100% utilization). This is a little better than the first option, but it causes utilization to be over-reported since it does not recognize the capacity available in the idle thread. The process is also significantly over-charged compared to when it shares the processor with another process.
3. Treat processor threads as individual processors, charging the process with 100% of the time and reporting processor-thread based utilization (50% utilization). This causes utilization to be under-reported because SMT efficiencies are much less than 100%, that is, the not idle thread typically represents only 15%-30% additional capacity. The process is also significantly over-charged compared to when it shares the processor with another process.

To address this problem, the Processor Utilization Resource Register (PURR) was introduced on Power5™ for the purpose of apportioning processor time among the processor's threads. The PURR is defined such that  $\sum (\Delta \text{PURR}) = \Delta \text{TIME}$  over any interval. For each SMT context, PURR ticks are apportioned among the idle and non-idle processor threads, so that non-idle PURR as a portion of available time for the processor reflects the relationship between throughput and processor utilization observed for a representative commercial processing workload. By expressing processor utilization as the ratio of not idle PURR ticks to available time and by expressing process accounting in terms of PURR ticks, the historical definition and relationship between processor utilization and process accounting was maintained.

IBM Power10 processor-based systems with EnergyScale employ variable processor speed technology to dynamically optimize the speed and energy usage of the processor to the demand of the workload. Because the PURR ticks at a constant rate independent of processor speed, PURR-based processor utilization remains a useful and accurate metric, but PURR-based process accounting charges can vary depending on processor speed. To address this problem, the IBM Power6 processor included a new per-thread processor timekeeping facility to normalize the relationship between processor time and processor speed. The new facility was named the Scaled Processor Utilization Resource Register (SPURR), and it represented processor time at nominal (i.e. 100%) speed.<sup>3</sup> The SPURR was primarily intended to improve process accounting consistency, but it can also be used in conjunction with the

---

<sup>3</sup>When the Power10 processor is operating at full speed, the PURR and SPURR tick in lockstep; when the Power10 processor is operating at reduced speed, the SPURR ticks slower than the PURR; when the Power10 processor is operating in excess of full speed, the SPURR ticks faster than the PURR.

PURR in processor speed and capacity calculations. For example, a SPURR to PURR percentage of 85% indicates that the processor operated at 85% nominal speed over a sample interval.

While the PURR and SPURR provide the information necessary to measure accurate processor utilization and to perform consistent process accounting in the variable processor speed environment, they do not address the ambiguity of CPU time in the historical context. Simply stated, whether a CPU time value in a legacy software interface should represent PURR-based CPU time or SPURR-based CPU time is open to some interpretation. There is no single best choice to handle all cases, and in fact, the issue has not been uniformly addressed by all operating systems or even among versions of the same system. There is general agreement that SPURR-based process accounting is preferable to PURR-based accounting since the results are more consistent across EnergyScale modes and within modes that vary the processor speed dynamically.

## **Appendix III: Resources**

### **AIX:**

<http://publib.boulder.ibm.com/infocenter/pseries/v6r1/index.jsp>

(Search for "spurr" to discover references regarding enablement in APIs and/or tools)

### **IBM i:**

[IBM Docs – IBM i Documentation](#)

(Search for “Energy management”, “Processor folding”, “Scaled processor time attribute”, “Processor time”, “Scaled processor time”, “Processor utilized time”, “Processor scaled utilized time”, “Processor interrupt time”, “Processor scaled interrupt time”, “Processor stolen time”, “Processor scaled stolen time”, “Processor donated time”, “Processor scaled donated time”, “Processor idle time”, “Processor scaled idle time”)

### **IBM EnergyScale for Power9 Processor-Based Systems:**

[https://www.ibm.com/downloads/cas/6GZMODN3?mhsrch\\_a&mhq=EnergyScale](https://www.ibm.com/downloads/cas/6GZMODN3?mhsrch_a&mhq=EnergyScale)

### **Power Systems:**

<https://www.ibm.com/it-infrastructure/power>

### **Redfish:**

[https://www.dmtf.org/sites/default/files/standards/documents/DSP2051\\_1.0.0.pdf](https://www.dmtf.org/sites/default/files/standards/documents/DSP2051_1.0.0.pdf)



The Power Architecture and Power.org wordmarks and the Power and Power.org logos and related marks are trademarks and service marks licensed by Power.org.

UNIX is a registered trademark of The Open Group in the United States, other countries or both.

Linux is a trademark of Linus Torvalds in the United States, other countries or both.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. In the United States and/or other countries.

All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, to evaluate the performance of a system they are considering buying.

Photographs show engineering and design models. Changes may be incorporated in production models.

Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM.



© IBM Corporation 2021  
IBM Corporation  
Systems and Technology Group  
Route 100  
Somers, New York 10589

Produced in the United States of America  
October 2021  
All Rights Reserved

This document was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this document in other countries.

The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.

IBM, the IBM logo, ibm.com, AlX, EnergyScale, i5/OS, Power, Power6, Power7, Power8, Power9, Power10, System i and System p are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml)

Other company, product, and service names may be trademarks or service marks of others.

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

This equipment is subject to FCC rules. It will comply with the appropriate FCC rules before final delivery to the buyer.

Information concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of the non-IBM products should be addressed with those suppliers.

When referring to storage capacity, 1 TB equals total GB divided by 1000; accessible capacity may be less.

The IBM home page on the Internet can be found at: <http://www.ibm.com>.

The IBM Power home page on the Internet can be found at: <http://www.ibm.com/systems/power/>