IBM Linux Technology Center

Linux on POWER for Green Data Center

POWER7™
Built on Power™

Vaidyanathan Srinivasan <svaidy@linux.vnet.ibm.com>
IBM India Systems and Technology Labs

Agenda

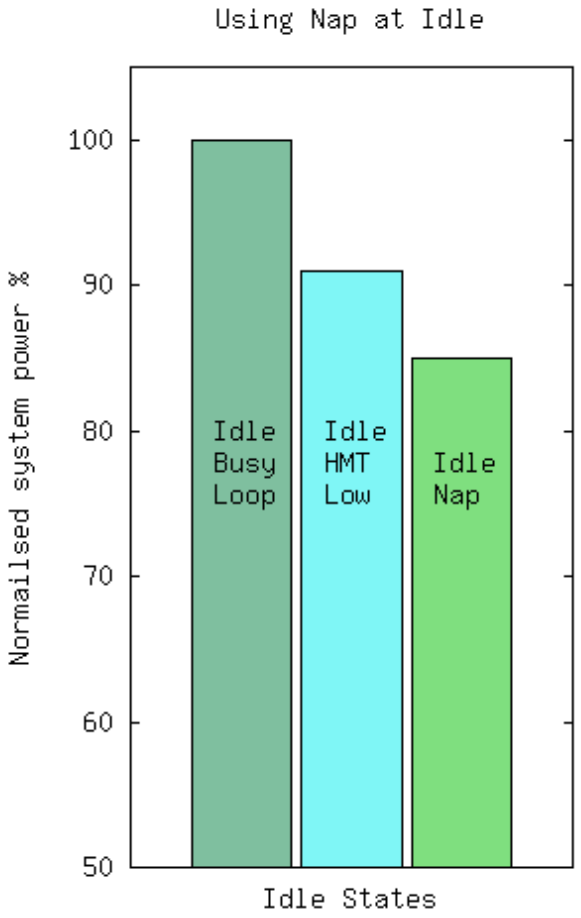
- POWER platform energy management features
- Linux OS energy management at idle
  - Short idle cycles – nap at idle
  - Long idle cycles – cpu offline
- Linux scheduler power saving tunables
- Platform power modes
- Experimental Results
- Reference
  - Linux patches
  - POWER platform reference

POWER platform energy management features

- Dynamic and voltage and frequency scaling (DVFS)
  - Thermal Power Management Device (TPMD) handles utilization based frequency and voltage control
  - System Power mode policies like *Static Power Saver* and *Dynamic Power Savers* are implemented in the platform
- Low power idle states
  - Processor Nap at idle using hypervisor calls
  - Hints in H_CEDE() hypervisor call for long term idle (POWER7)
- Power capping
  - Limit server power consumption below a preset value

Linux Energy Management features on POWER: Nap at idle

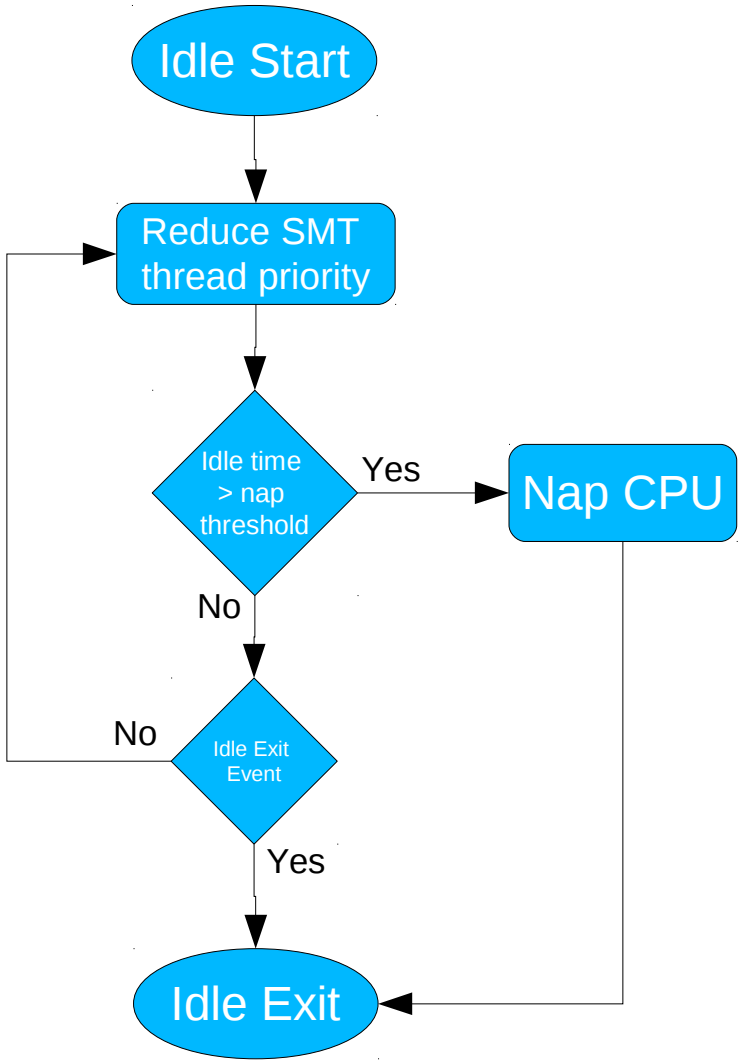
- Idle task in Linux can yield CPU cycles to hypervisor
- Hypervisor can transition CPU to low power nap state
- CPUs in nap can be woken-up on external interrupt or timer event



Measurement on 16 core system in Nominal power mode

Linux idle loop will reduce thread priority (HMT Low) first and then transition to Nap for longer idle cycles

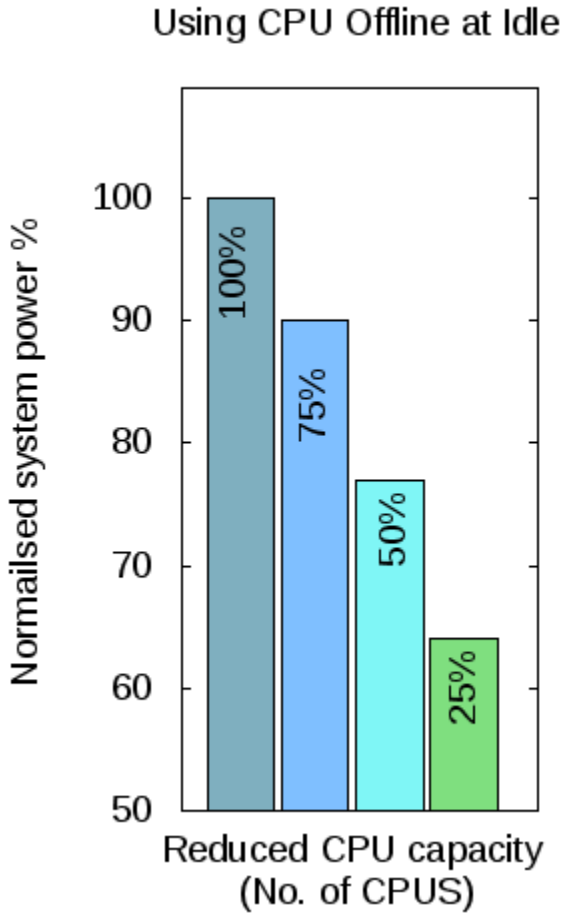
Nap at idle in Linux



Exploiting long idle cycles using cpu offline states

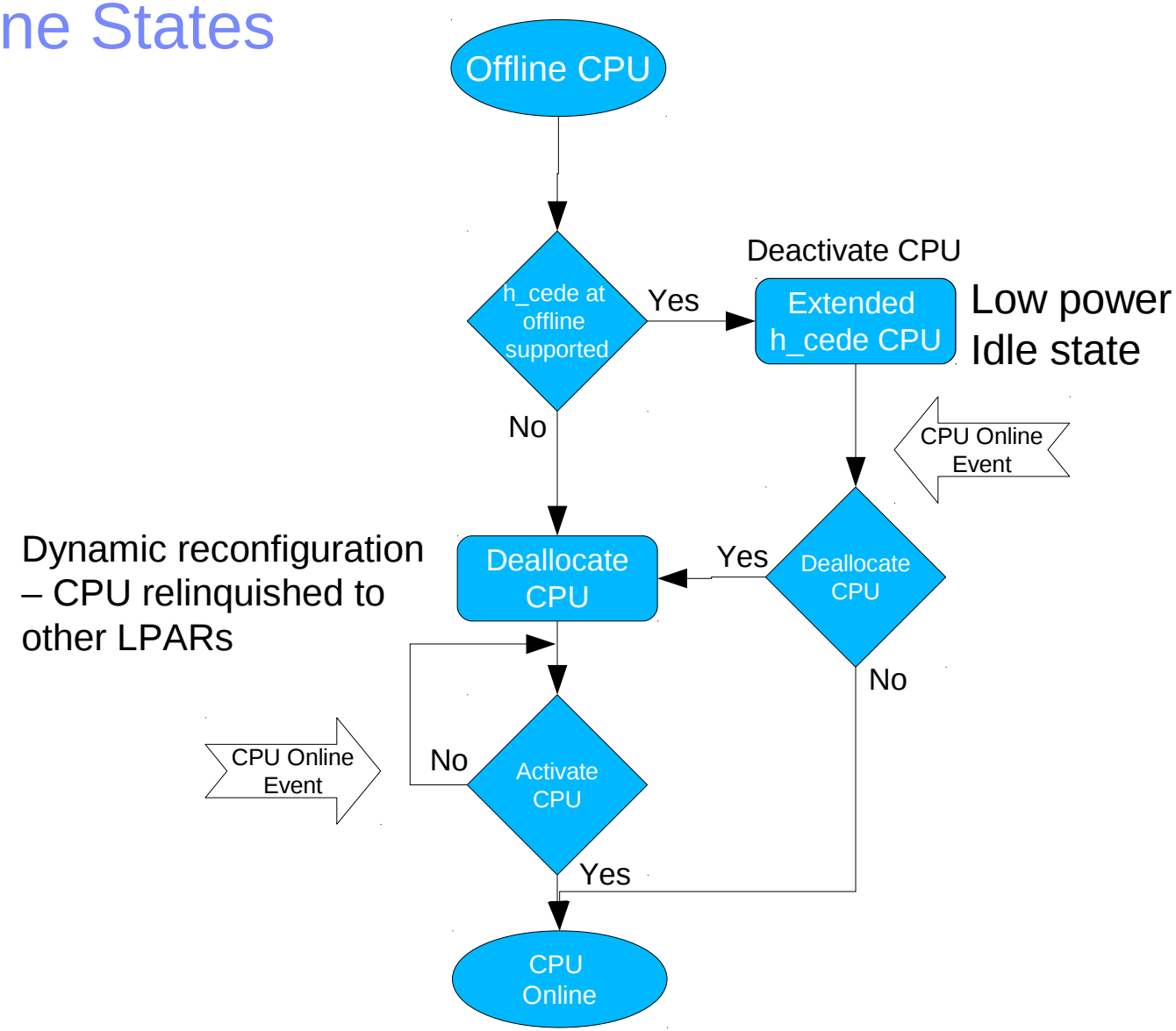
- Linux cpuoffline framework allows runtime deallocation of logical CPUs
- Multiple offline states allow exploitation of the framework for power savings (deactivate) and deallocations

```
echo 0 > /sys/devices/system/cpu/cpu1/online
```



Measurement on 32 core system in Nominal power mode

CPU Offline States



Hints for cpuoffline: H_BEST_ENERGY

- Hypervisor provided hints for list of CPUs to activate or deactivate
- Hints will provide optimal configuration based on overall system utilization across all partitions

```
# cat /sys/devices/system/cpu/activate_hint_list
16,20,24,28,56,60
# cat /sys/devices/system/cpu/deactivate_hint_list
0,4,8,12,32,36,40,44,48,52
```

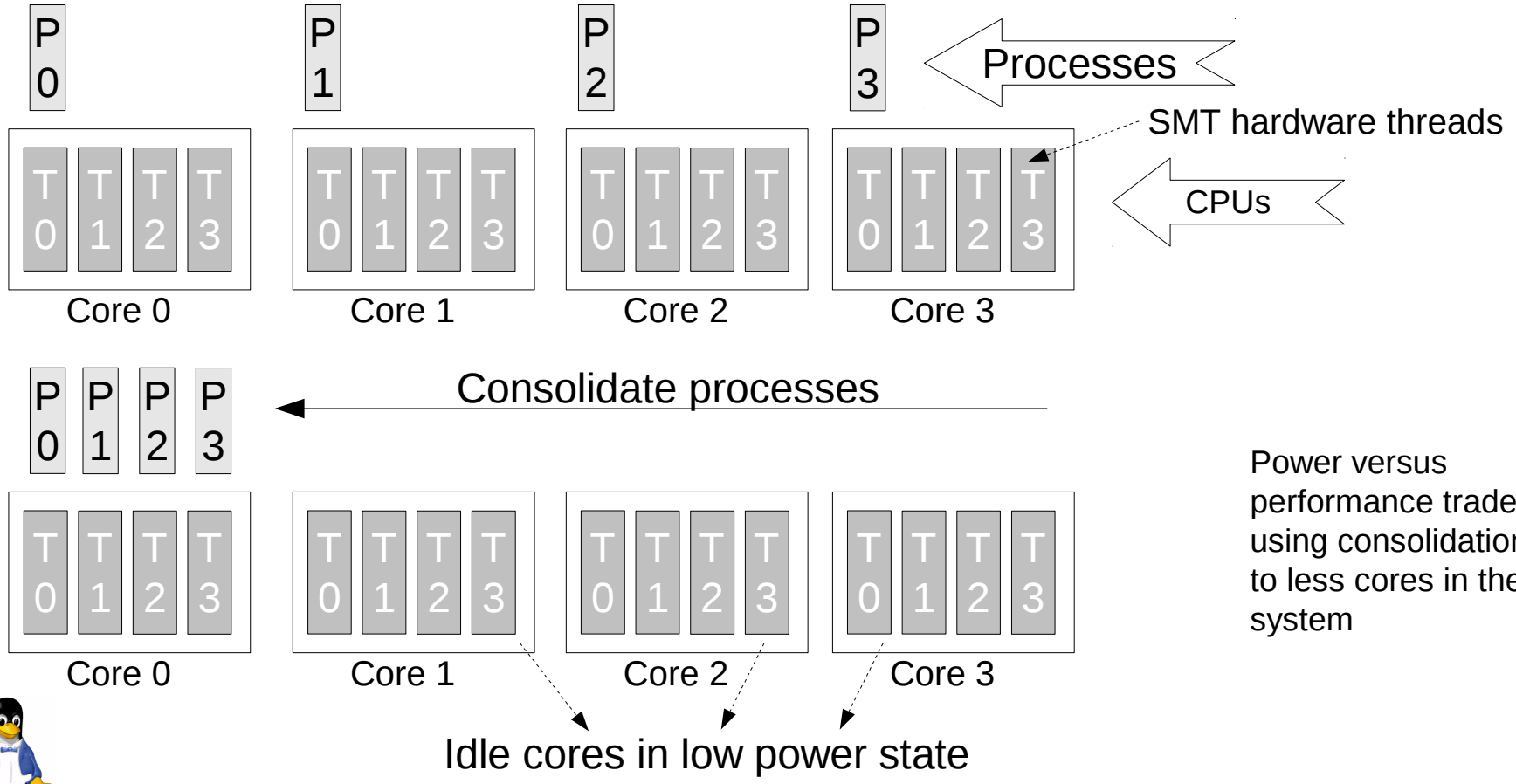
Patch under development and discussion in linuxppc-dev mailing list

CPU offline and DLPAR operations

- Dynamic Logical Partition configuration framework allows CPU and other resources to be added or removed to an OS instance at runtime on POWER platform
- CPU Offline for power savings cooperates with DLPAR operations
- Dynamic removal of a CPU in low power (deactivated) state will initiate a deallocate operation

Scheduler power saving tunables

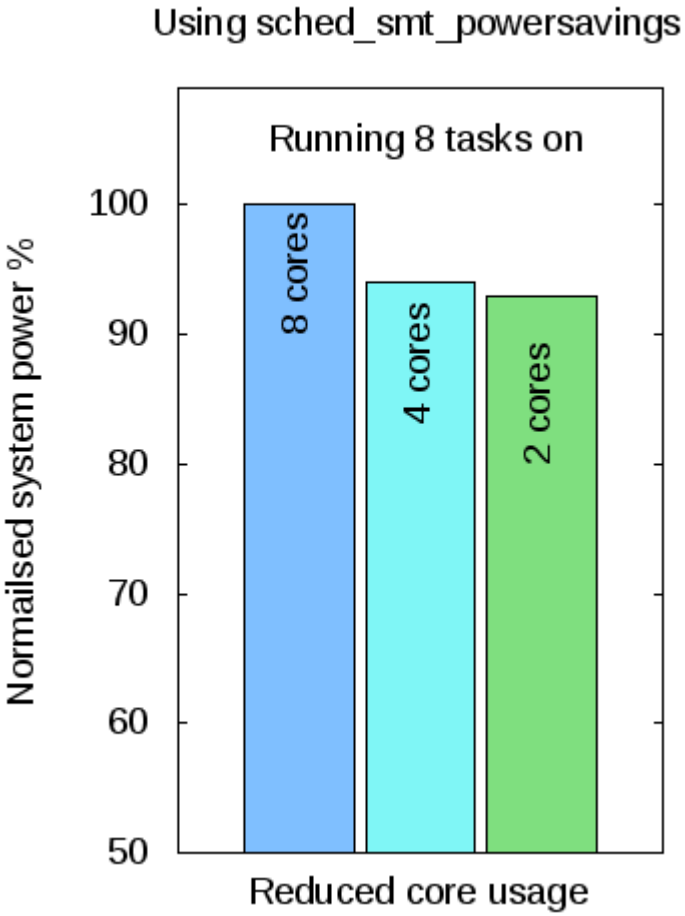
- Task consolidation to cores using sched_smt_powersavings



Power versus performance tradeoff using consolidation to less cores in the system

Linux scheduler sched_smt_powersavings

- Running 8 tasks on reduced number of cores by exploiting SMT4 modes
- Increases core nap residency and saves power
- Degradation in performance is workload dependent 10-30% for two tasks per core and higher for 4 tasks per core
- Ideally used in conjunction with static power saver mode where the policy favours power savings

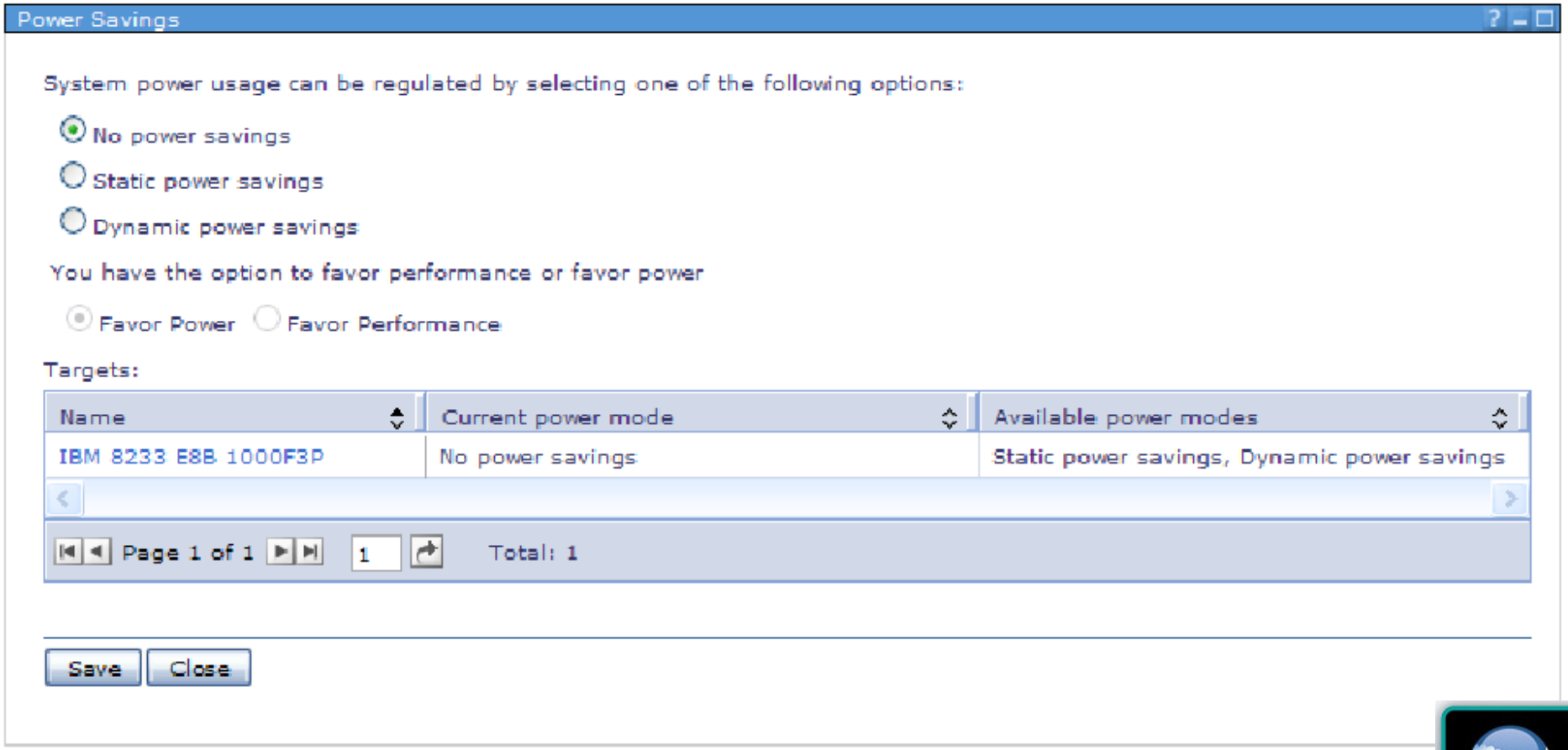


Measurement on 16 core system in SPS mode

Platform power modes

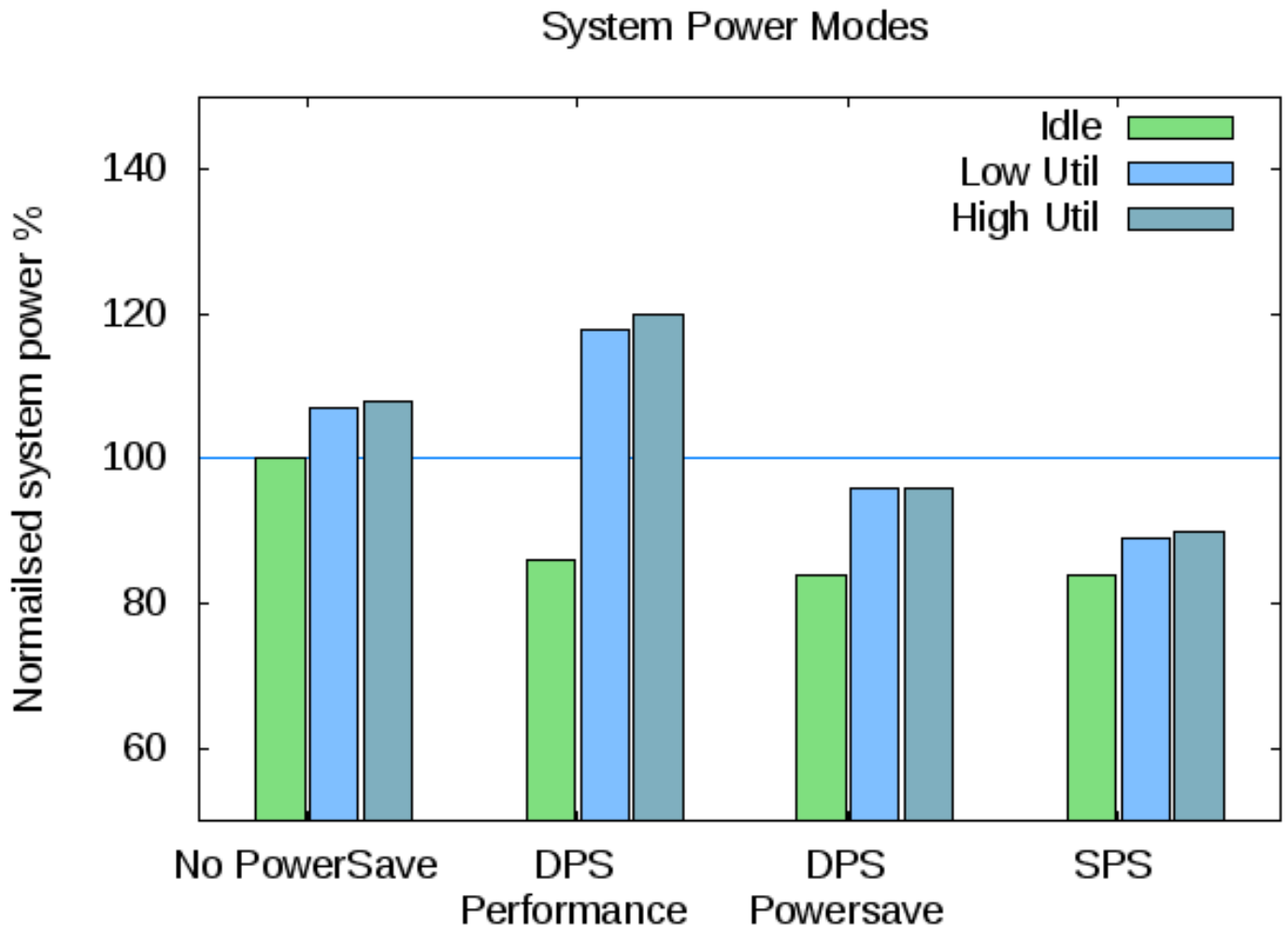
- Platform power modes are set using IBM Systems director with Active Energy manager (AEM) plugin
- Static Power Saver
  - Run CPUs at a fixed lower frequency and voltage to reduce power consumption
- Dynamic Power Saver
  - Change CPU frequency and voltage based on runtime CPU utilization
  - SubModes
    - Favour power savings
    - Favour performance

IBM Systems Director – Active Energy Manager



Source: IBM POWER7 EnergyScale WhitePaper

System Power Modes - Illustration

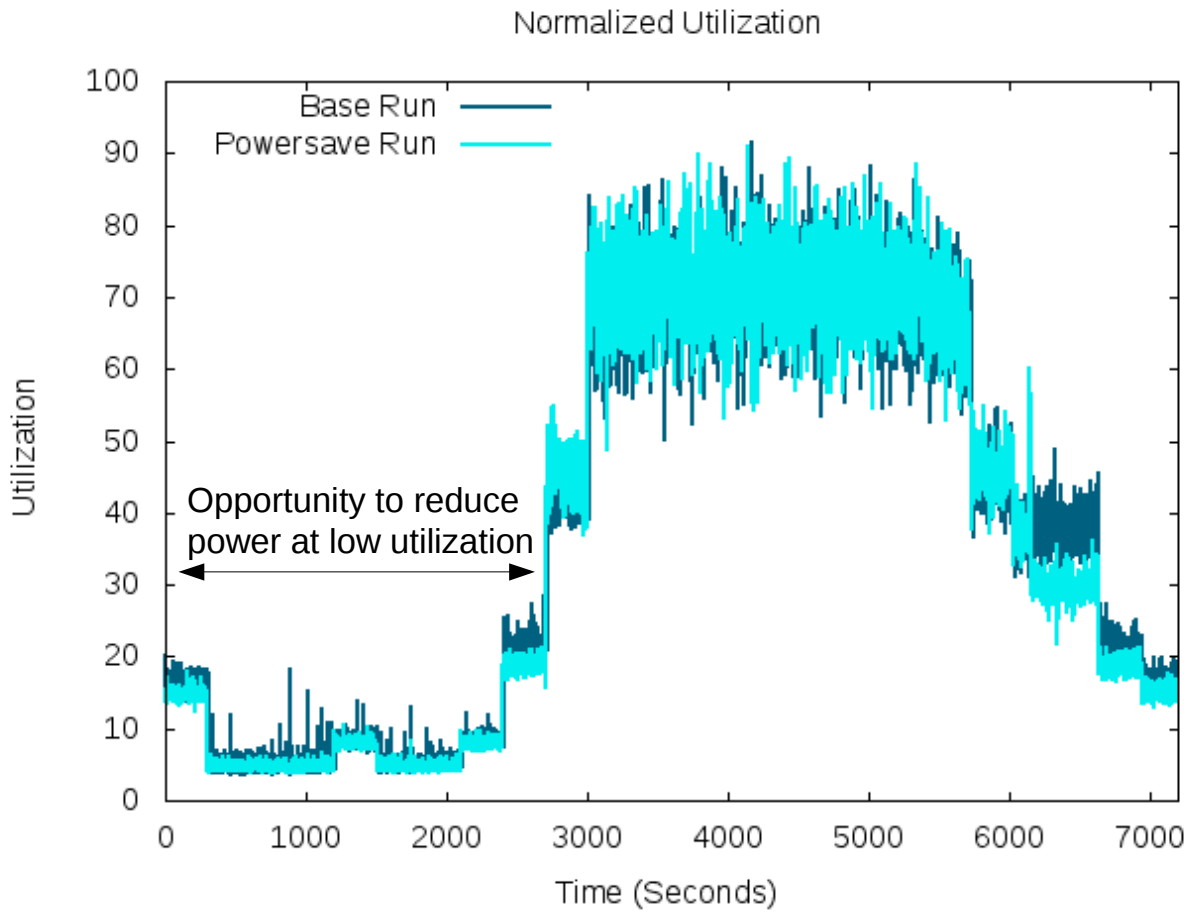


Linux interface to system power modes

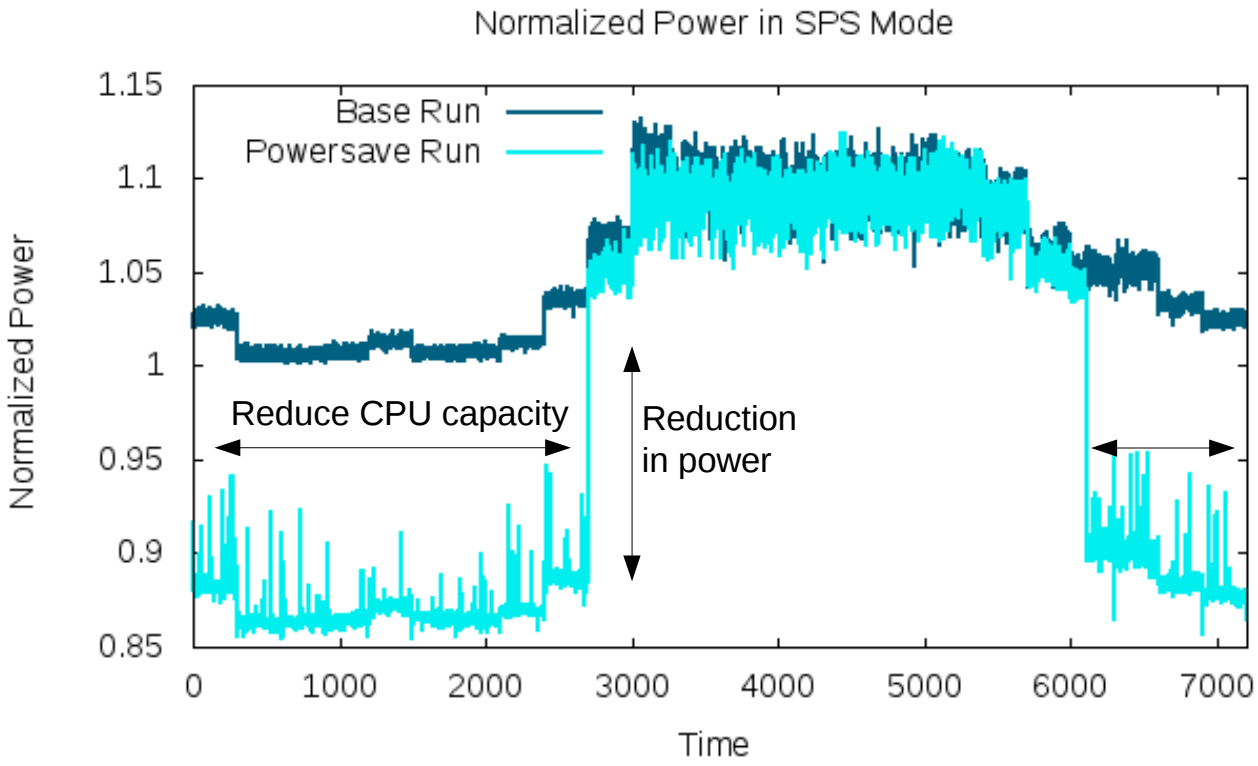
- Power mode information can be obtained in Linux OS using hypervisor call H_EM_GET_PARMS
- Linux OS specific optimizations can be done based on power modes

```
#cat /proc/powerpc/lparcfg
…
<lpar info>
…
power_mode_data=xxxx
```

Experimental Results: Workload pattern (Trade6)



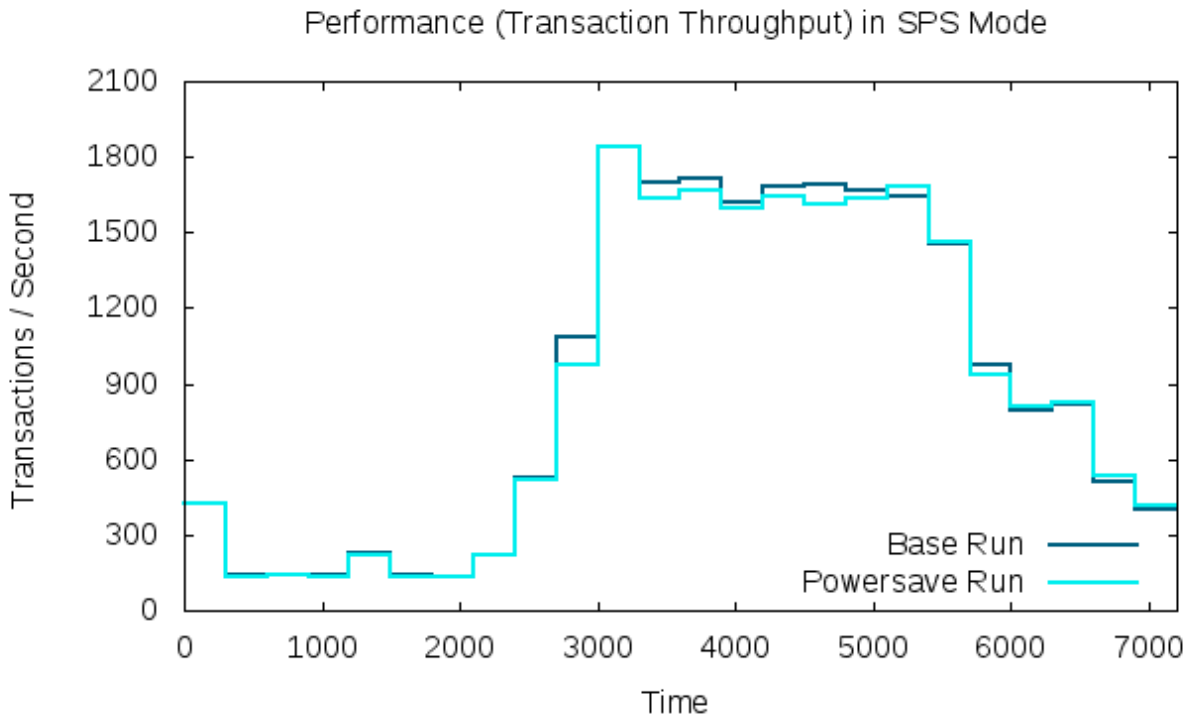
Experimental results: Static Power Saver (power)



Typical workload pattern emulated by Trade6 benchmark
System capacity statically reduced to 30% during off peak hours

Measurement on 16 core system in SPS mode

Experimental results: Static Power Saver (performance)



Negligible performance impact since only very low utilization periods are exploited for power savings

Reference: Linux patches

- Cpuoffline state framework
  http://git.kernel.org/?p=linux/kernel/git/torvalds/linux-2.6.git;a=commit;h=3aa565f53c396914a9406388efaa238e9c937fc6
- sched_smt_powersavings http://lwn.net/Articles/287924/
- POWER h_em_get_parms http://patchwork.ozlabs.org/patch/49190/
- POWER h_best_energy http://patchwork.ozlabs.org/patch/46851/
- POWER cpuidle framework http://lkml.org/lkml/2010/2/9/33

Reference: Linux publications

- OLS2008:Energy-aware task and interrupt management in Linux
  http://ols.fedoraproject.org/OLS/Reprints-2008/srinivasan1-reprint.pdf
- LFCollabSummit 2009: Tweaking Linux For Green Data Center
  http://events.linuxfoundation.org/slides/lfcs09_srinivasan.pdf

Reference: POWER related resources

- POWER7 http://www-03.ibm.com/press/us/en/pressrelease/29315.wss
- POWER7 EnergyScale WhitePaper
  http://www-03.ibm.com/systems/power/hardware/whitepapers/energyscale7.html
- POWER7 and EnergyStar http://www-03.ibm.com/press/us/en/pressrelease/29796.wss
- ActiveEnergyManager
  http://www.ibm.com/systems/management/director/plugins/actengmgr.html
- Linux on POWER7 https://www.ibm.com/developerworks/wikis/display/LinuxP/Home
- IBM Linux Technology Center http://www-03.ibm.com/linux/ltc/

Acknowledgments

- Arun Bharadwaj
- Christoph Arenz
- Dipankar Sarma
- Gautham R Shenoy
- Jeffery J Heroux
- Naren Devaiah
- Nathan Fontenot
- Premalatha M Nair
- Susanne Libischer
- Trinabh Gupta

Questions/Discussions?

Thank You

Legal Statements

- Copyright International Business Machines Corporation 2010.
- Permission to redistribute in accordance with Linux Foundation Collaboration Summit submission guidelines is granted; all other rights reserved.
- This work represents the view of the authors and does not necessarily represent the view of IBM.
- IBM, IBM logo, ibm.com are trademarks of International Business Machines Corporation in the United States, other countries, or both.
- Intel is a trademark or registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.
- Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
- Other company, product, and service names may be trademarks or service marks of others.
- References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.

Legal Statements

- INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.