SUPER's Green HPC: Improving the Energy-efficiency of DoE HPC

SUPER INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

Laura Carrington (lead), Ananta Tiwari
UCSD/SDSC

Rob Fowler
RENCI

Barry Roundtree, Bronis de Supinski
LLNL

Dan Terpstra
UTK

Abstract: SUPER's Energy thrust is charged with understanding how computation and communication patterns affect the overall energy requirements of HPC applications. We then leverage this understanding to design software- and hardware-aware optimization techniques that reduce the DOE's HPC energy footprint. Three sub-thrusts have emerged within this thrust: research on power measurement devices, software solutions that provide fine-grained access to the power measurements, and energy efficiency research that utilizes the first two to develop green optimization strategies. We highlight recent accomplishments in each of the three research themes and present empirical results that illustrate SUPER's contributions in minimizing DOE's HPC energy requirements.

Developing component level power measurement

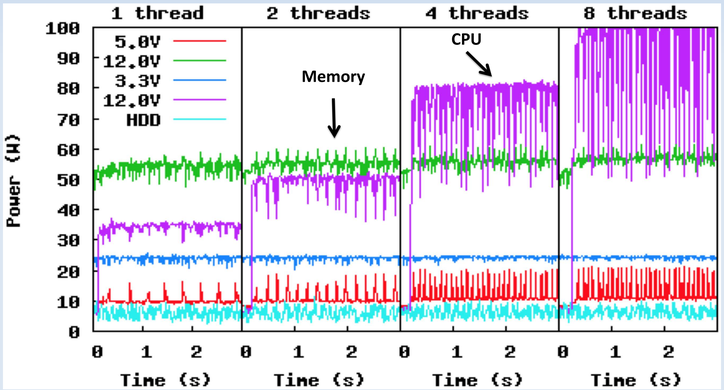
Component level power measurements helps us understand, exploit and limit hardware policies for transparent and adaptive power control. Such understanding allows the implementation of more effective hardware-software co-managed energy optimizations.

DC measurement device – PowerMon2

PowerMon2 facilitates fine grain DC power monitoring and evaluation of DVFS for energy reduction on per application phase level. It allows evaluation of power supply efficiency and derivation and validation of energy models based on application characteristics.

PowerMon2 device

- Fits into a 3.5'' internal hard drive bay
- Cable harness for mounting between the power supply, motherboard and other components
- Independent monitoring of voltage and current on up to 8 power rails
- Device control and data over USB port
- Approximate 1K sample sets/sec



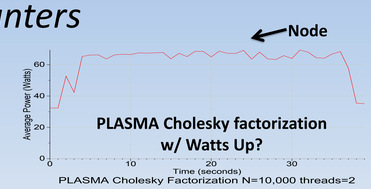
Power draw plot for 5 rails during the execution of NAS OpenMP BT run using up to 8 threads on 4 core hyper-threaded Nehalem. Iteration structure is clearly visible on CPU and memory power draws.

PAPI interface for power measurements: devices & counters

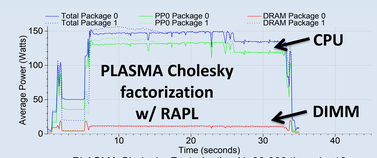
PAPI provides access to hardware performance counters, enabling software engineers to measure relationships between software performance and processor or other hardware events.

Component PAPI can be extended to allow energy and power measurements without any changes to core infrastructure or measurement tools. Current users of PAPI on HPC systems can now analyze power and energy with little additional effort. Recent and ongoing component development supports measuring energy and power usage via a variety of sources including:

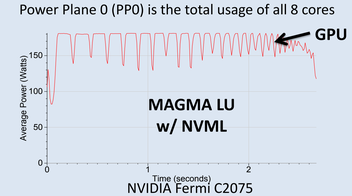
- PowerMon2 : RENCI's real-time power monitoring card
- RAPL: Intel's Running Average Power Limit model
- NVML: NVIDIA's Management Library for Tesla cards
- Watts Up?: AC line voltage and current monitor



PLASMA Cholesky Factorization N=10,000 threads=2
Intel Core2 Laptop measurement via Watt's Up Pro with one-second sampling frequency

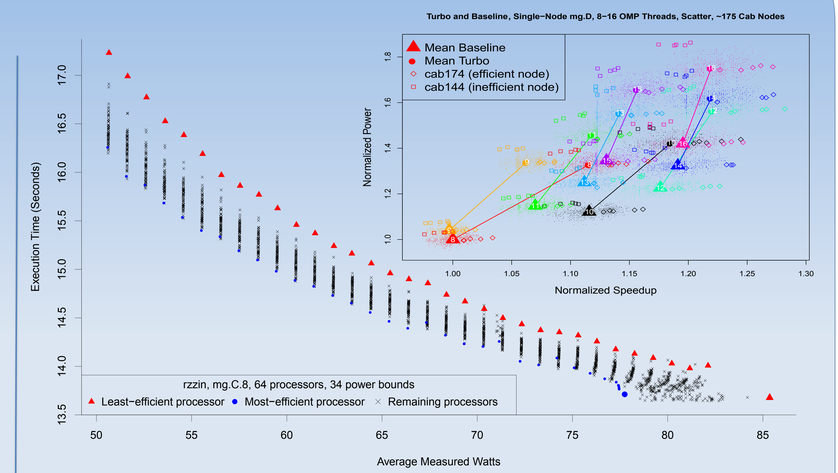


PLASMA Cholesky Factorization N=30,000 threads=16
Sandybridge EP measured with RAPL.
Power Plane 0 (PP0) is the total usage of all 8 cores



NVIDIA Fermi C2075
MAGMA LU size 10,000

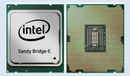
Intel's RAPL on SandyBridge based clusters



Use of Intel's Runtime Power Average Limit technology on Sandy Bridge processors has, for the first time, allowed per-processor power measurement and power capping to occur at scale. The work above shows unexpected variation in power draw and performance across processors: variation that must be accounted for in any proposed exascale design.

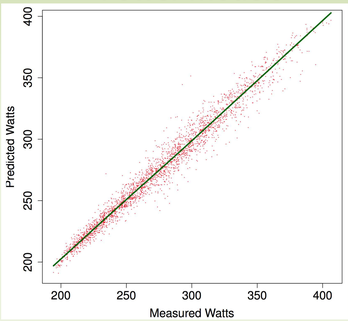
Table 14-2. RAPL MSR Interfaces and RAPL Domains

Domain	Power Limit (Offset 0)	Energy Status (Offset 1)	Policy (Offset 2)	Perf Status (Offset 3)	Power Info (Offset 4)
PKG	MSR_PKG_POWER_LIMIT	MSR_PKG_ENERGY_STATUS	RESERVED	MSR_PKG_RAPL_PERF_STATUS	MSR_PKG_POWER_INFO
DRAM	MSR_DRAM_POWER_LIMIT	MSR_DRAM_ENERGY_STATUS	RESERVED	MSR_DRAM_RAPL_PERF_STATUS	MSR_DRAM_POWER_INFO
PP0	MSR_PP0_POWER_LIMIT	MSR_PP0_ENERGY_STATUS	MSR_PP0_POLICY	RESERVED	RESERVED
PP1	MSR_PP1_POWER_LIMIT	MSR_PP1_ENERGY_STATUS	MSR_PP1_POLICY	RESERVED	RESERVED



Utilizing component level power measurements to advance energy-efficiency HPC research

Component Level Power Models



PMaC's Green Queue is an automated framework that collects characterization data for large scale applications (e.g. memory & computational behavior); then use this data to construct power draw models and to develop customized application-aware *inter-node* and *intra-node* DVFS settings.

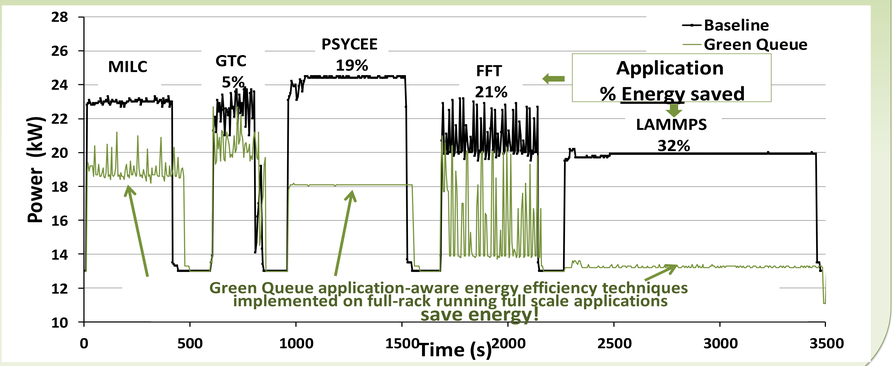
Power models relate application characterizations to component level power draw.

Power models are used to customize DVFS settings for application phases in Green Queue.

Inter-Node technique takes advantage of load imbalances within HPC applications to save energy. The CPU clock can be lowered on the hardware, which has less work to do with minimal to no performance impact while saving energy.

Intra-Node technique exploits opportunities to scale down CPU frequency, and therefore power draw, for application phases, which lack computational work (e.g., phases where CPU spends significant time in waiting for the data motion from memory to complete).

Green Queue on SDSC-Gordon's 1024 cores



Support for this work was provided through the Scientific Discovery through Advanced Computing (SciDAC) program funded by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (and Basic Energy Sciences/Biological and Environmental Research/High Energy Physics/Fusion Energy Sciences/Nuclear Physics).