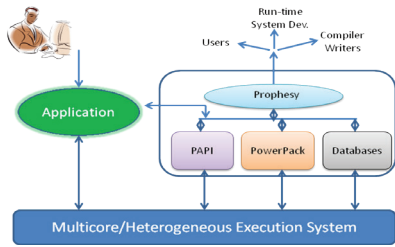


MUMMI

MULTIPLE METRICS MODELING INFRASTRUCTURE



MuMMI, or Multiple Metrics Modeling Infrastructure, is a project that is developing a framework to facilitate systematic measurement, modeling, and prediction of performance, power consumption, and performance-power tradeoffs for applications running on multicore and heterogeneous systems. MuMMI offers an integrated and extensible performance modeling and prediction framework for use by application developers, system designers, and scientific application users, helping them to make key choices for configuring and selecting appropriate multicore and hybrid systems and to evaluate and optimize power/performance on these systems.

POWER MEASUREMENT ON MULTICORE SYSTEMS

Recent Intel SandyBridge chips include the "Running Average Power Limit" (RAPL) interface. The internal circuitry can estimate current energy usage based on a model driven by hardware counters, temperature, and leakage models. The power model has been validated by Intel to closely follow actual energy being used. RAPL reports total combined energy used by all the cores (referred to as Power-Plane 0 (PPO)), including all of the processor caches, and energy reading for the DRAM interface (Power-Plane 1 (PP1)). PAPI provides access to the values returned by the power model.

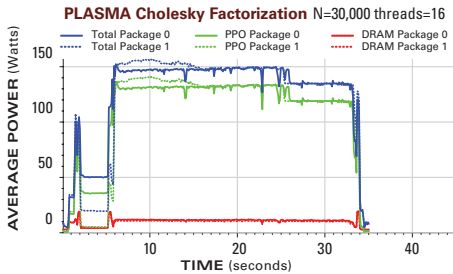


Fig. 1. PLASMA Cholesky power usage measured with RAPL on SandyBridge EP. Power Plane 0 (PPO) is total usage for all 8 cores in a package.

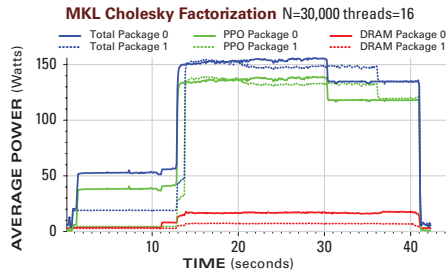


Fig. 2. Intel MKL Cholesky power usage measured with RAPL on SandyBridge. Power Plane 0 (PPO) is total usage for all 8 cores in a package.

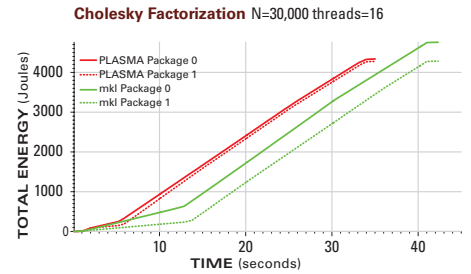


Fig. 3. Energy usage of two different implementations (PLASMA and MKL) of Cholesky on SandyBridge EP measured with RAPL.

Figure 1 shows average power measurements gathered while doing Cholesky factorization using the PLASMA library. Notice that the energy usage by each package varies, despite all of the cores doing similar work. Part of this is likely due to variations in the cores at the silicon level. Figure 2 shows the same measurements for the Intel MKL library.

Figure 3 shows energy measurements comparing the same Cholesky factorization using the PLASMA and Intel MKL libraries on the same hardware. The PAPI results show that for this case, PLASMA uses energy more quickly, but finishes faster and uses less total energy for the calculation.

POWER MEASUREMENT ON GPUS

Recent NVIDIA GPUs can report power usage via the NVIDIA Management Library (NVML). The power reported represents usage for the entire board, including GPU and memory. We have constructed an NVML component for PAPI and have validated the results using a "Kill-A-Watt" power meter. Figure 4 shows data gathered on an NVIDIA Fermi C2075 card running a MAGMA kernel using the LU algorithm with a matrix size of 10k. Our experiments have shown that the implementation of MAGMA GEMM operations on a GPU completely utilize it, maximizing the power consumption. This explains why the hybrid CPU+GPU LU factorization also maximizes the GPU power consumption and reduces time taken so that overall energy consumption is minimized.

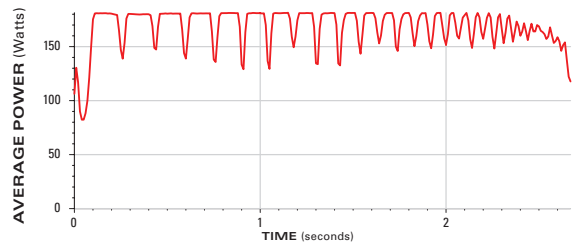


Fig. 3. MAGMA LU with size 10,000 power measurement on an NVIDIA Fermi C2075, gathered with NVML.

IN COLLABORATION WITH



FIND OUT MORE AT



www.mummi.org

SPONSORED BY



WITH SUPPORT FROM

