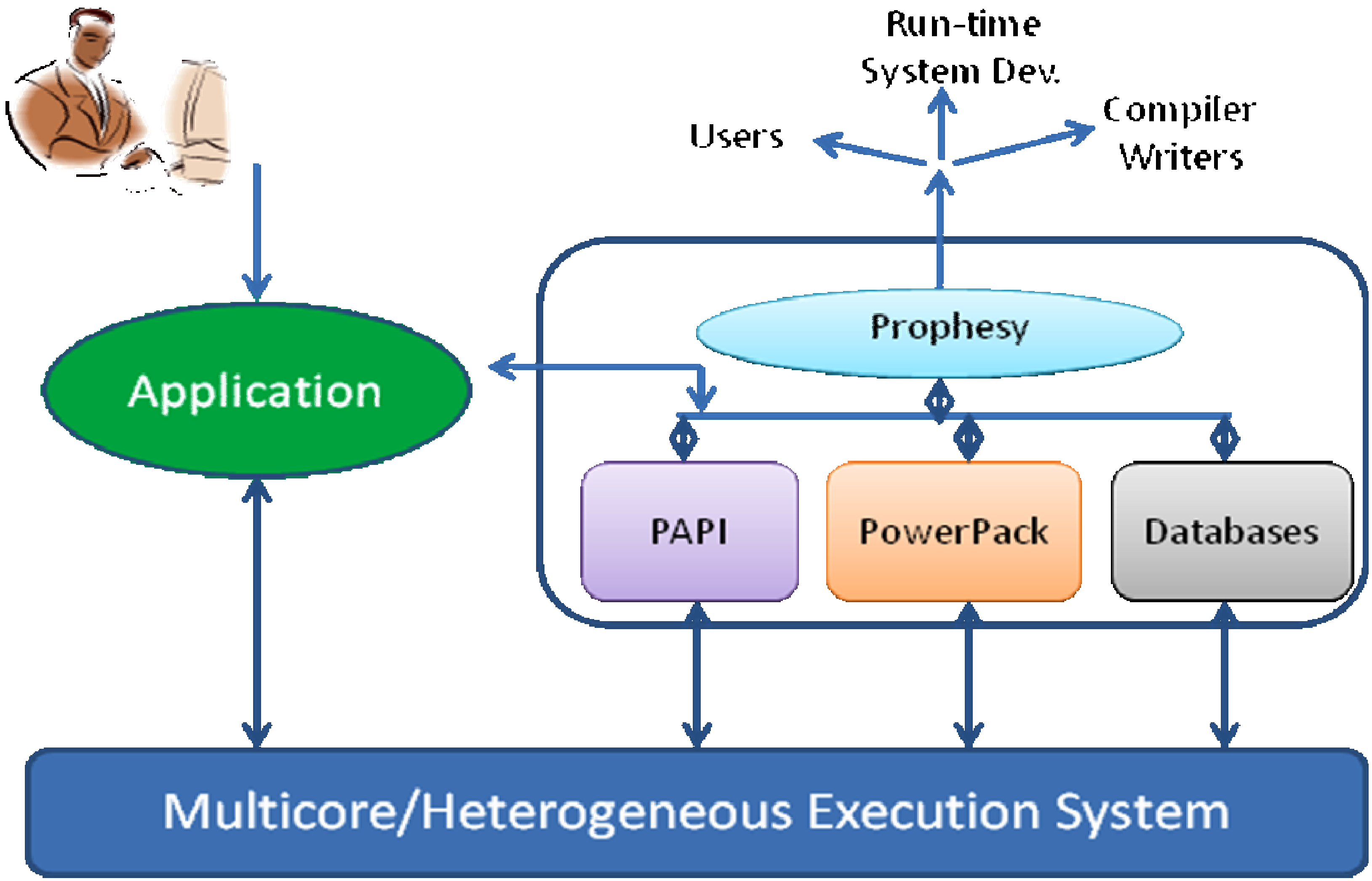
Combined Performance and Power Consumption Modeling and Optimization with MuMMI

Valerie Taylor, Xingfu Wu, Chee Wai Lee (TAMU)
Kirk Cameron, Hung-Ching Chang (Virginia Tech)
Dan Terpstra (UTK)
Shirley Moore (UTEP) (presenter)
svmoore@utep.edu

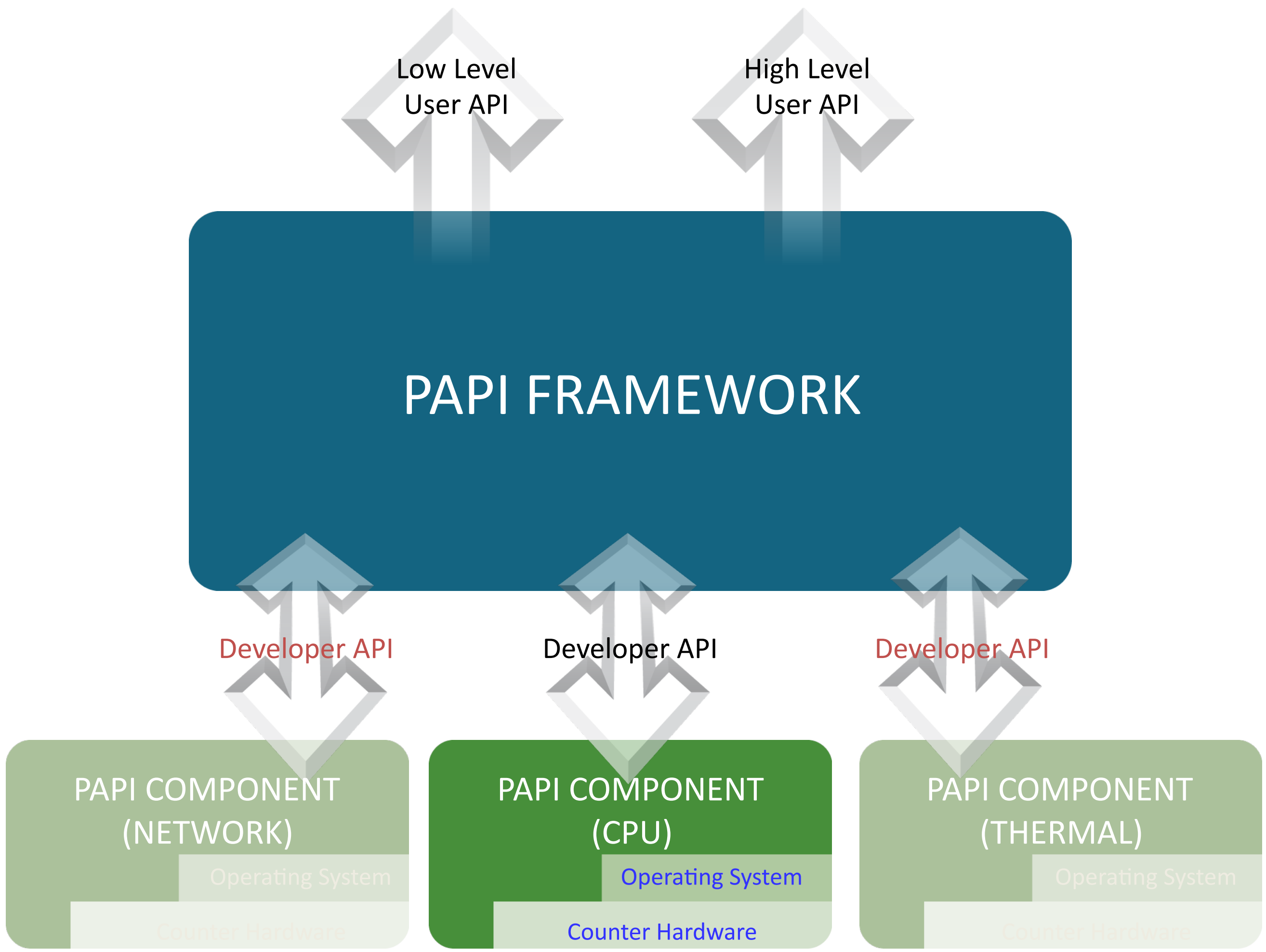
SC'12 Extreme Scale Performance Tools Workshop
November 16, 2012

MuMMI (Multiple Metrics Modeling Infrastructure) Project

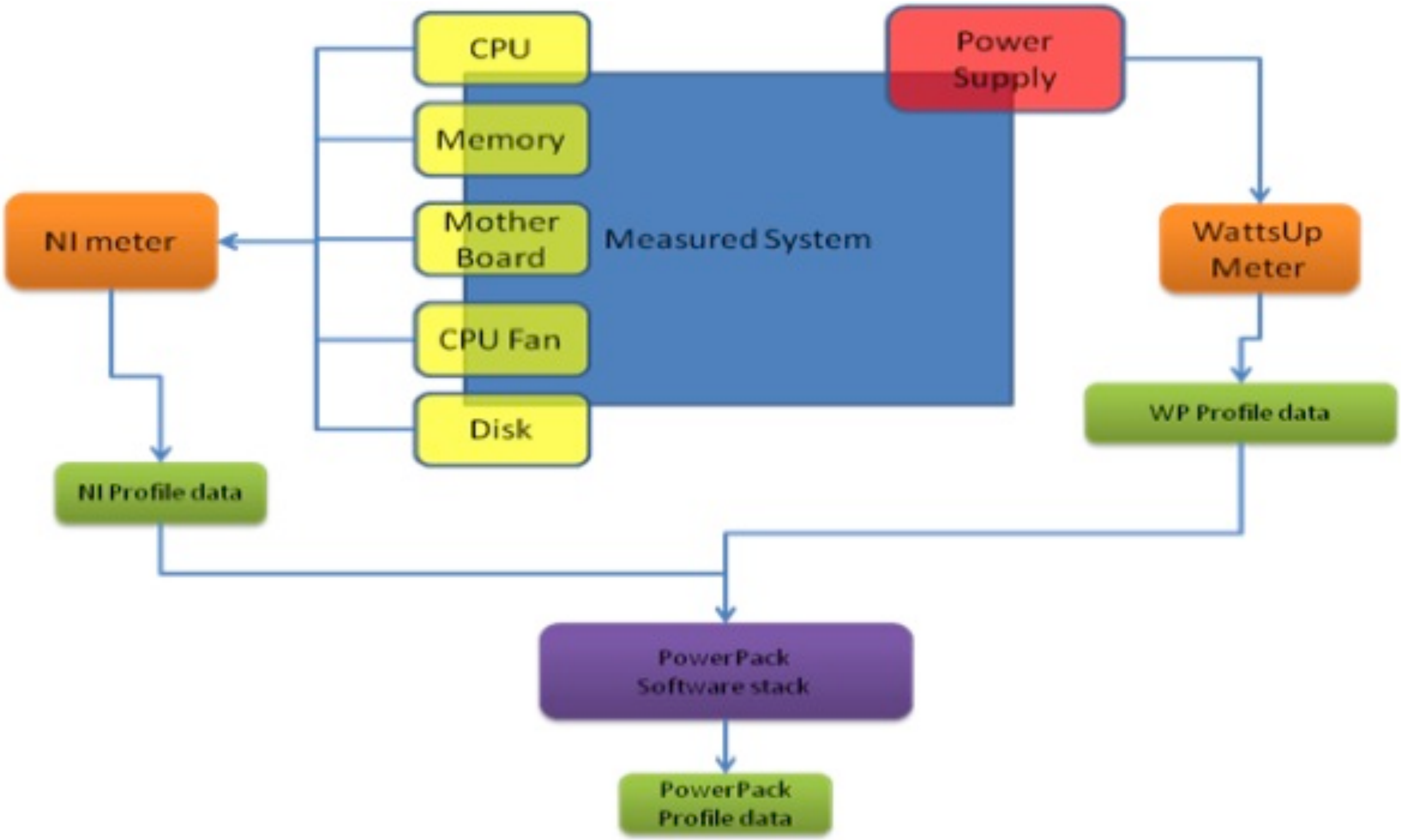


Project Website: http://www.mummi.org/

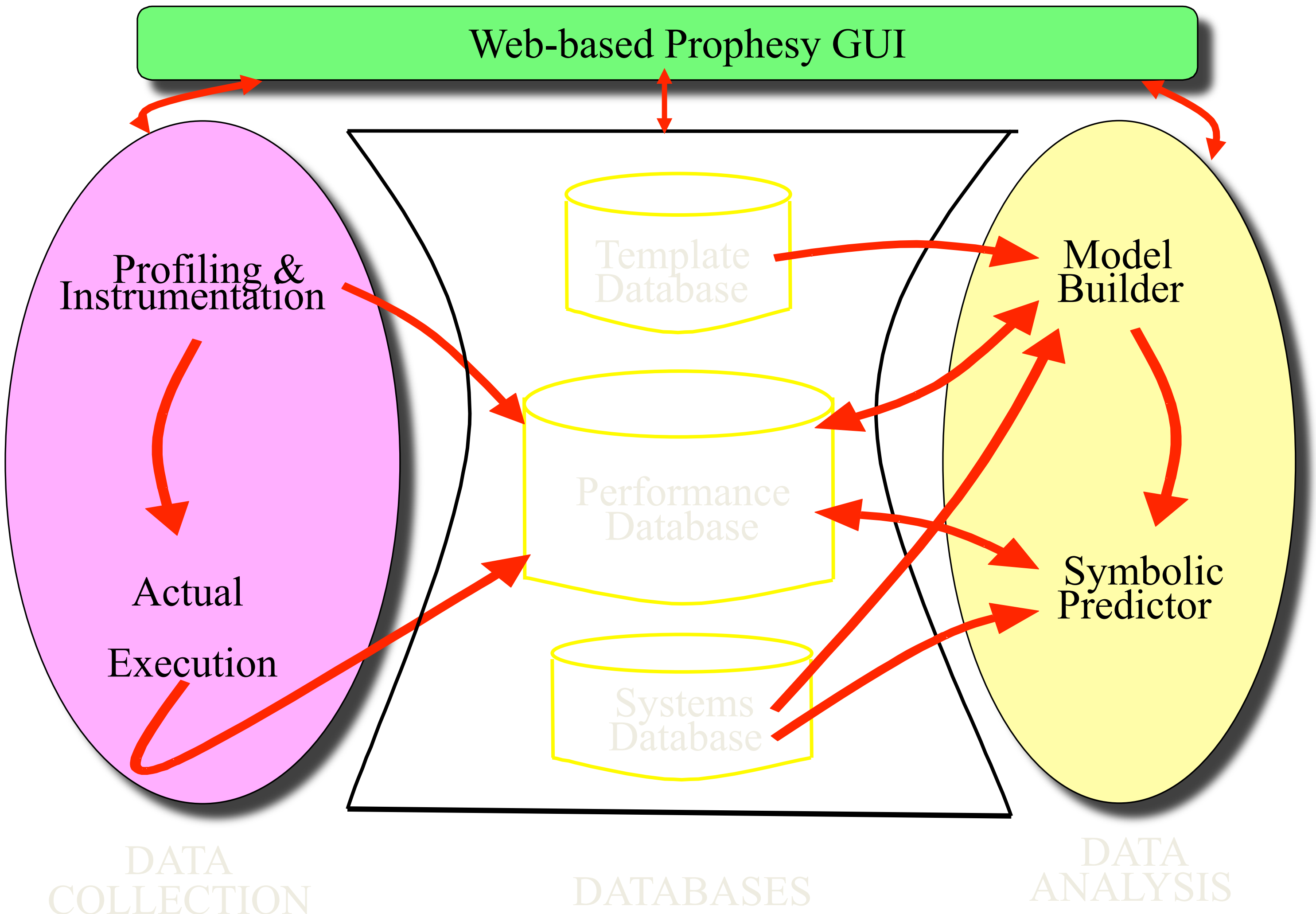
Component PAPI (UTK)



PowerPack (Virginia Tech)



Prophesy System



Power and Energy – Why do We Care?

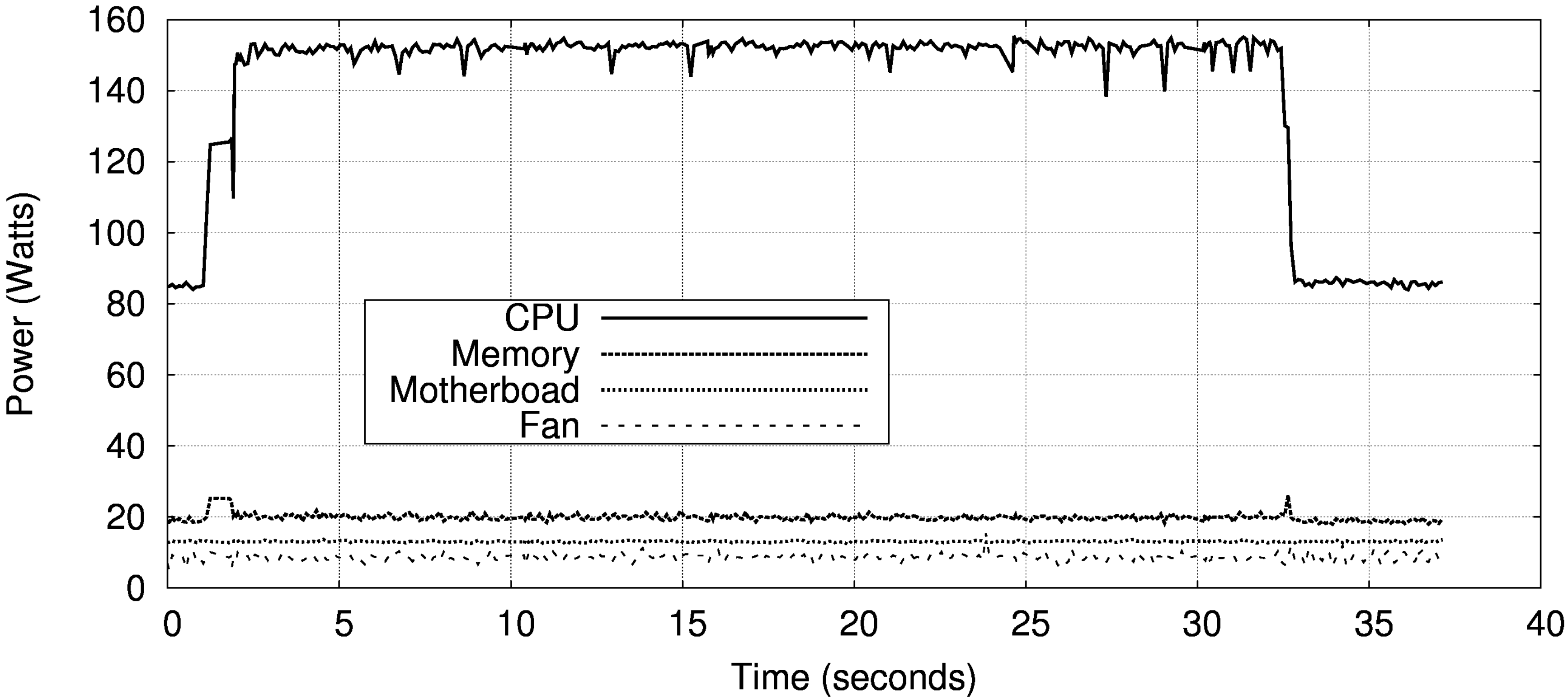
- New, massive HPC machines use impressive amounts of power

- When you have 100k+ cores, saving a few Joules per core quickly adds up

- To improve power/energy draw, you need some way of measuring it

PowerPack Measurement

Plasma/dposv results with Virginia Tech's PowerPack



Power Measurements with PAPI

- PAPI (Performance API) is a platform-independent library for gathering performance-related data
- PAPI-C interface makes adding new power measuring components straightforward
- PAPI can provide power/energy results in-line to running programs
- One interface for all power measurement devices
- Existing PAPI code and instrumentation can easily be extended to measure power
- Existing high-level tools (Tau, VAMPIR, etc.) can be used with no changes
- Easy to measure other performance metrics at the same time
- Example PAPI power measurement components: Intel RAPL and NVIDIA NVML

RAPL

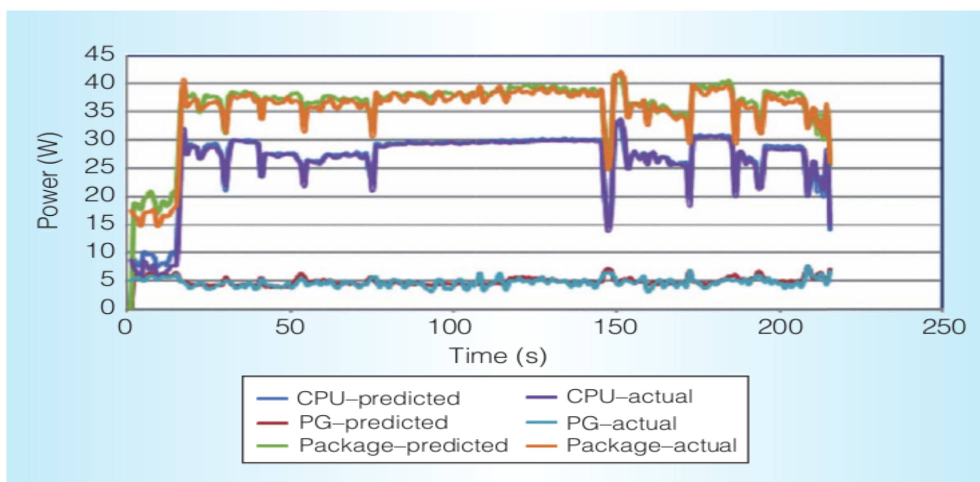
- **R**unning **A**verage **P**ower **L**imit
- Part of an infrastructure to allow setting custom per-package hardware enforced power limits
- User Accessible Energy/Power readings are a bonus feature of the interface.
- RAPL is *not* an analog power meter.
- RAPL uses a software power model, running on a helper controller on the main chip package.
- Energy is estimated using various hardware performance counters, temperature, leakage models and I/O models.
- The model is used for CPU throttling and turbo-boost, but the values are also exposed to users via a model-specific register (MSR).

Available RAPL Readings

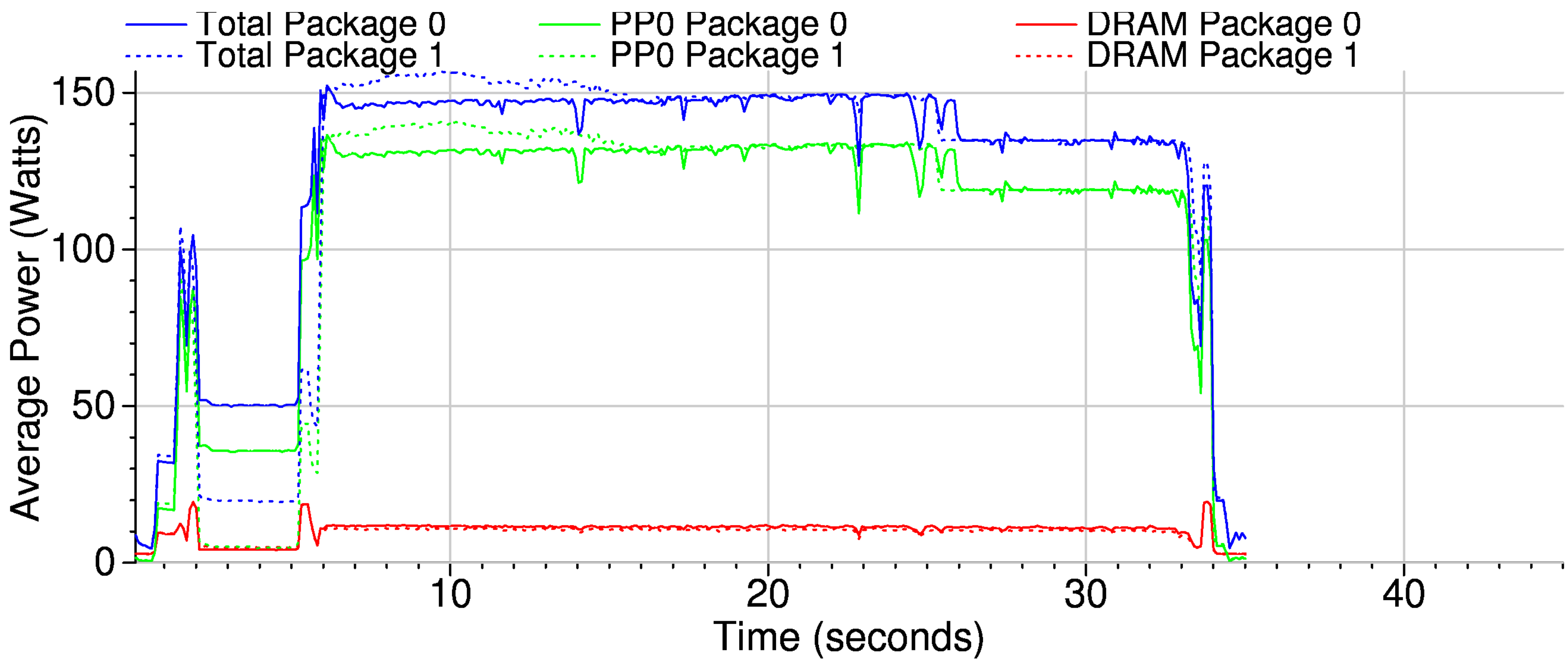
- PACKAGE_ENERGY: total energy used by entire package
- PP0_ENERGY: energy used by "power plane 0" which includes all cores and caches
- PP1_ENERGY: on original Sandybridge this includes the on-chip Intel GPU
- DRAM_ENERGY: on Sandybridge EP this measures DRAM energy usage. It is unclear whether this is just the interface or if it includes power used by all the DIMMs too.

RAPL Measurement Accuracy

- Intel Documentation indicates Energy readings are updated roughly every millisecond (1kHz)
- Rotem et al. in "Power-Management Architecture of the Intel Microarchitecture Code-Named Sandy Bridge" (IEEE Micro, March/April 2012) claim measurements closely match real power measurements:



RAPL Power Plot



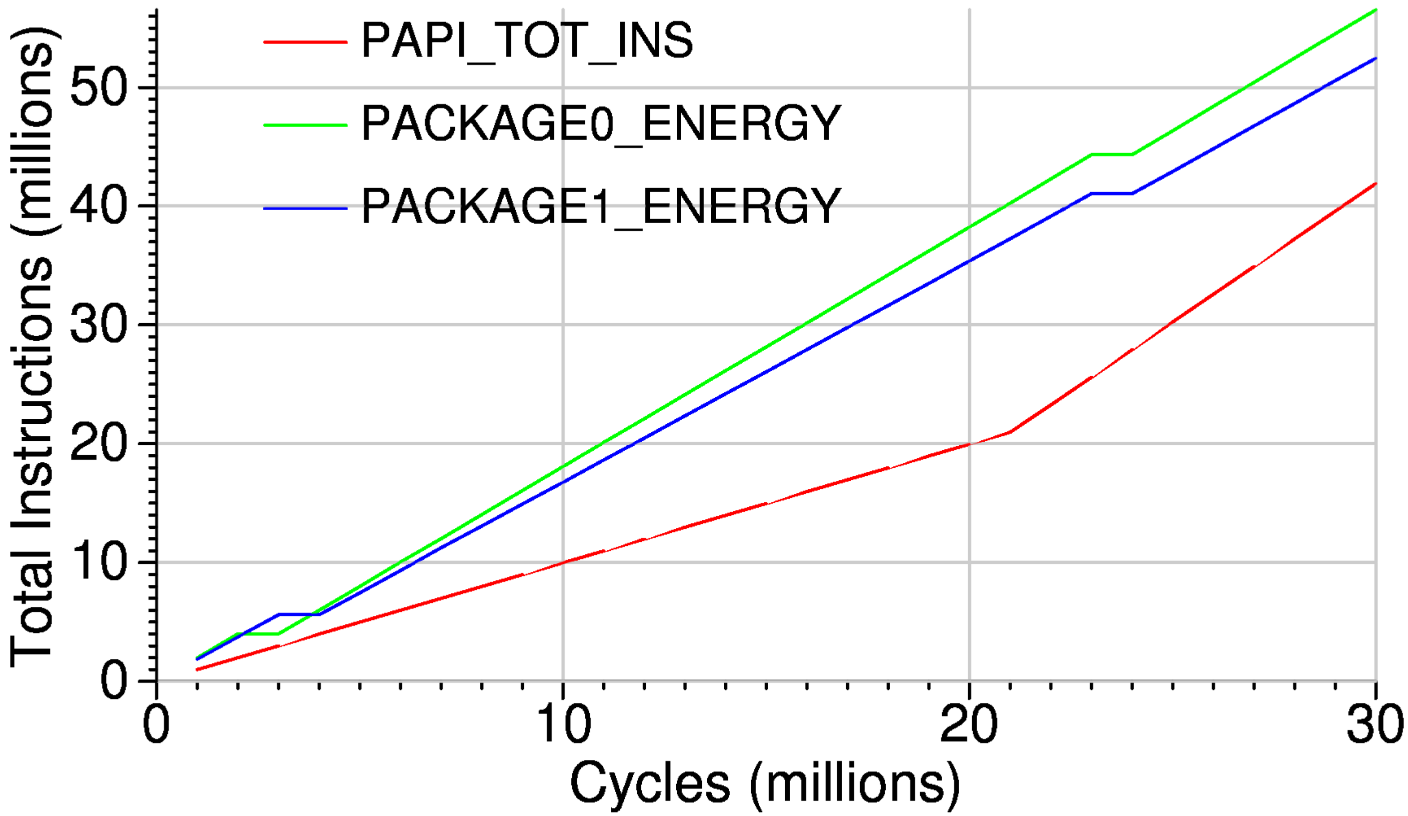
PLASMA Cholesky Factorization N=30,000 threads=16

Measured on SandyBridge EP

Listing Events

```
> papi_native_avail
==================================
 Events in Component: linux-rapl
==================================
| PACKAGE_ENERGY:PACKAGE0
|    Energy used by chip package 0
--------------------------------------
| PACKAGE_ENERGY:PACKAGE1
|   Energy used by chip package 1
--------------------------------------
```

Measuring Multiple Sources



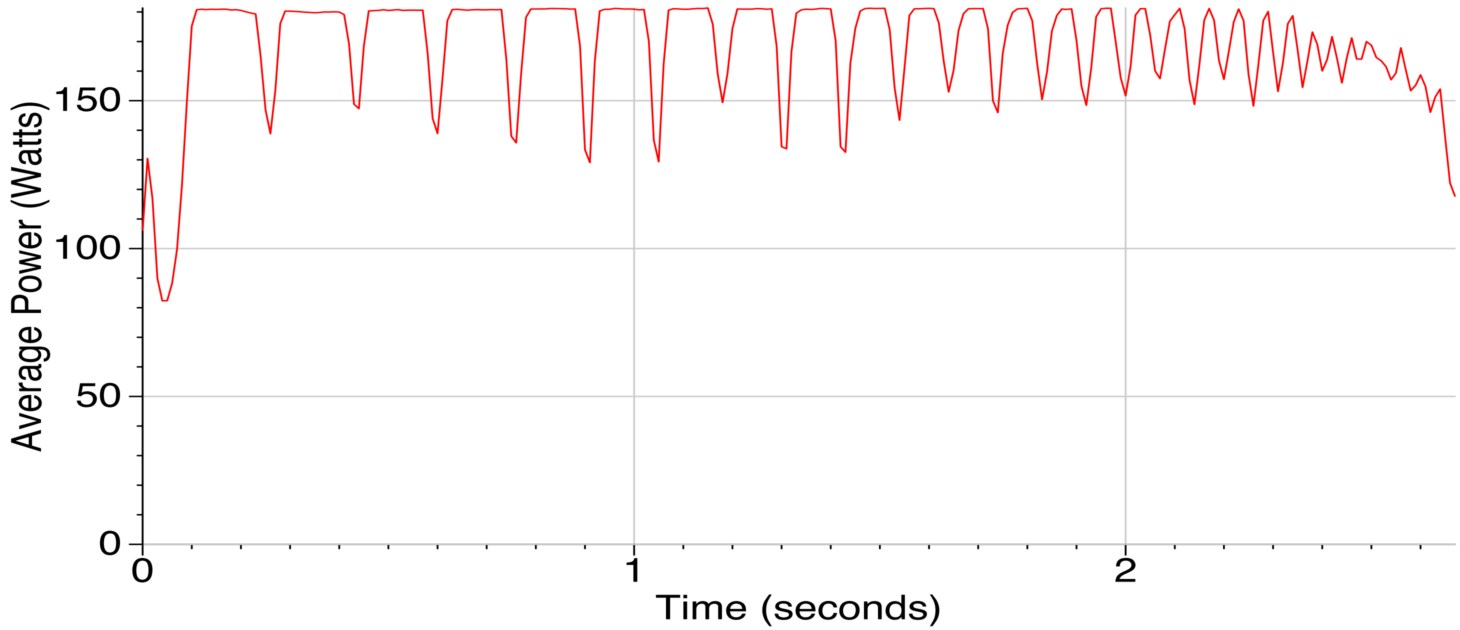
INT/FP RAPL Test

Measured on SandyBridge EP

NVIDIA NVML

- Recent NVIDIA GPUs can report power usage via the NVIDIA Management Library (NVML).
- Power reported is 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.

NVML Power Measurement of MAGMA Kernel



- Data gathered on an NVIDIA Fermi C2075 card running a MAGMA kernel using the LU algorithm with a matrix size of 10k.
- implementation of MAGMA GEMM operations on GPU completely utilize it, maximizing the power consumption.
- Hybrid CPU+GPU LU factorization also maximizes the GPU power consumption and reduces time taken so that overall energy consumption is minimized.

Hardware Counter Based Power Models

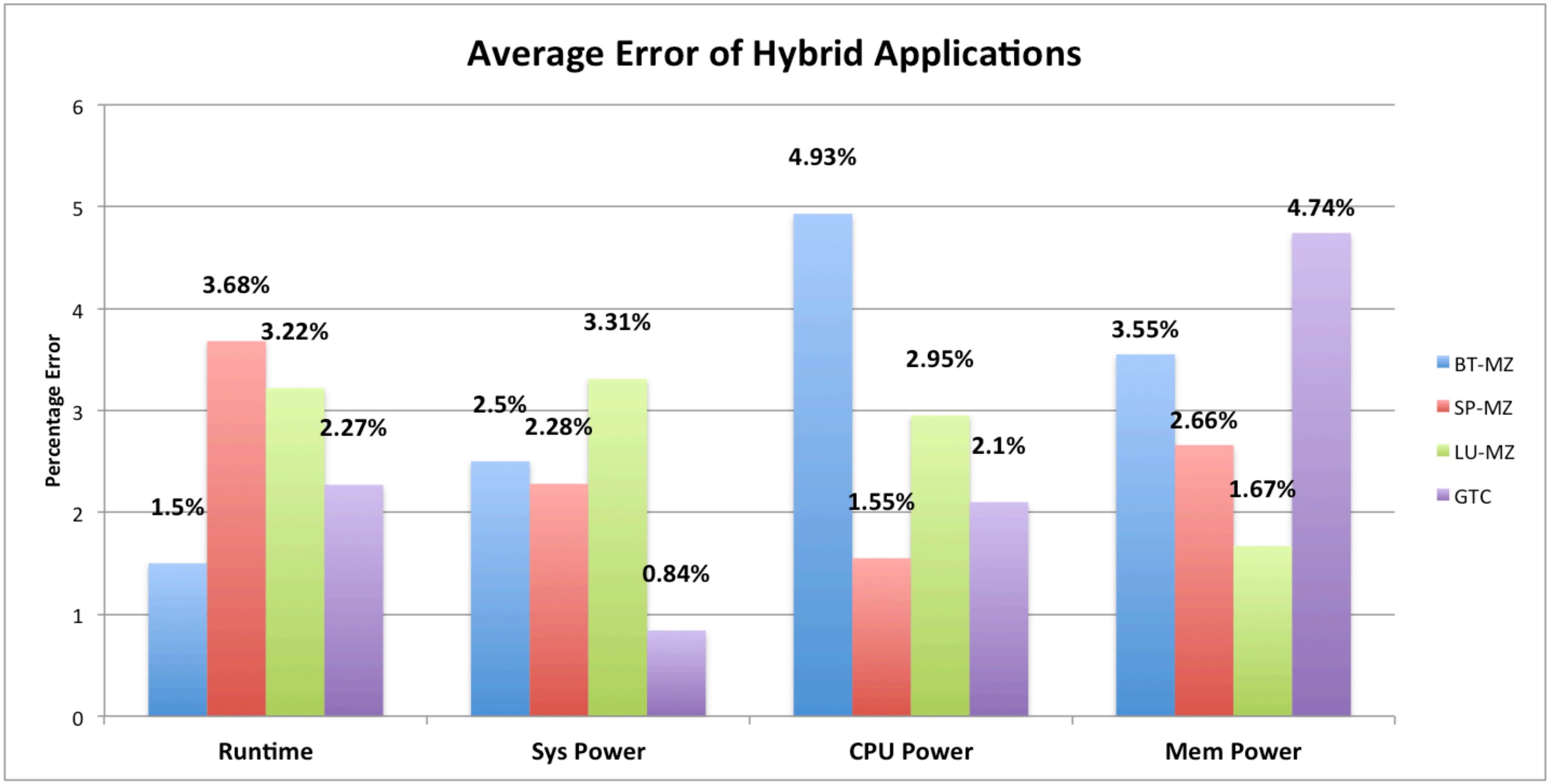
- Much related work on estimating energy/power using performance counters
- PAPI user-defined event infrastructure can be used to create power models using existing events
- Previous work (McKee et al.) shows accuracy to within 10%
- We have developed application-specific models with accuracy within 5%

Application-specific Power-Performance Models

Lively, Wu, Taylor, Moore, Chang, Su and Cameron, Power-Aware Predictive Models of Hybrid (MPI/OpenMP) Scientific Applications on Multicore Systems, *EnA-HPC2011*, Hamburg, Germany, Sept 2011.

	Time		**System Power**		**CPU Power**		**Memory Power**	
BT-MZ	Cache_FLD	-1.611	PAPI_L2_TCH	-1.6769	PAPI_L1_TCM	3.5432	PAPI_L1_TCA	0.0763
	PAPI_TOT_INS	0.0967	PAPI_L2_TCA	1.5967	PAPI_L2_TCH	-3.9389	PAPI_L1_DCM	4.0496
	PAPI_L2_TCH	0.2992	PAPI_RES_STL	0.0803	PAPI_RES_STL	0.3967	PAPI_L2_TCH	-1.9443
	PAPI_L2_TCA	1.2152					PAPI_L2_TCA	2.1806
SP-MZ	PAPI_TOT_INS	0.1818	PAPI_L1_ICA	0.355	LD_ST_stall	0.1917	Cache_FLD	0.4563
	PAPI_L1_TCA	0.0744	PAPI_L2_TCH	-1.3452	PAPI_L1_TCM	1.5008	LD_ST_stall	0.0192
	PAPI_L2_TCH	-1.2834	PAPI_L1_TCM	0.9911	PAPI_L2_TCH	-1.6914	PAPI_L2_TCH	-3.5895
	PAPI_L1_TCM	1.1761					PAPI_L2_TCA	3.1151
LU-MZ	Cache_FLD	-0.0006	LD_ST_stall	0.0166	LD_ST_stall	0.0869	PAPI_L1_TCA	0.27923
	PAPI_TOT_INS	0.0011	PAPI_L2_TCH	-0.9886	PAPI_L2_TCH	-8.0003	PAPI_L2_TCH	-3.9574
	PAPI_TLB_DM	3.9085	PAPI_L2_TCA	1.0411	PAPI_L2_TCA	7.9137	PAPI_RES_STL	-0.29141
	PAPI_L2_TCH	-0.0591	PAPI_RES_STL	0.025				
GTC	PAPI_TOT_INS	0.0006	PAPI_RES_STL	1.5689	PAPI_RES_STL	0.9261	PAPI_TOT_IN	0.169617
	PAPI_L2_TCH	-1.8976	PAPI_L2_TCH	-3.2505	PAPI_TOT_IN	0.2663	PAPI_L2_TCH	-2.881
	PAPI_L2_TCA	1.9351	PAPI_L1_TCA	1.6916	PAPI_L1_TCA	0.0816	PAPI_L2_ICM	2.7119
	PAPI_BR_INS	-0.0381			PAPI_L2_TCH	-1.2640		

Prediction Accuracy



Integration with Score-P

- Score-P and PAPI
  - Already integrated
  - Some refinements may be needed to handle power and energy measurements
- Score-P and MuMMI
  - Score-P produces OTF2 trace output.
  - OTF2 traces can be merged with Power Pack traces and output as MuMMI's output format. The SOAP scripts can then be used to upload this data into the MuMMI database.
  - We are currently able to simultaneously generate an OTF2 trace and a Power Pack trace. Work on a merging script is underway.
- Score-P and Power Pack
  - Difference between how Score-P currently expects data and how Power Pack expects to deliver data
  - Score-P expects to acquire performance data when an instrumentation point is encountered during execution.
  - Power Pack expects to be told about the encountered instrumentation point at the collector.
  - As such, Score-P needs to handle the acquisition of Power Pack information as a two-phase process. At the instrumentation point, Score-P needs to note a zero value and send the request package to the Power Pack data collector. At the end of the run, Score-P needs to pull in all the recorded Power Pack information in order for the data to bemerged before output.

Recent Publications

- Charles Lively, Xingfu Wu, Valerie Taylor, Shirley Moore, Hung-Ching Chang, and Kirk Cameron, Energy and Performance Characteristics of Different Parallel Implementations of Scientific Applications on Multicore Systems, *International Journal of High Performance Computing Applications (IJHPCA)*, Volume 25 Issue 3, August 2011, pp. 342 - 350.
- Charles Lively, Xingfu Wu, Valerie Taylor, Shirley Moore, Hung-Ching Chang, Chun-Yi Su and Kirk Cameron, Power-Aware Predictive Models of Hybrid (MPI/OpenMP) Scientific Applications on Multicore Systems, *International Conference on Energy-Aware High Performance Computing(EnA-HPC2011)*, Hamburg, Germany, September 7-9, 2011.
- Kiran Kasichayanula, Daniel Terpstra, Piotr Luszczek, Stan Tomov, Shirley Moore, and Greg Peterson. Power aware computing on GPUs. *Symposium on Application Accelerators in High Performance Computing* (*SAAHPC 2012)*, Argonne National Laboratory, July 10-11, 2012.
- Vince Weaver, Matthew Johnson, Kiran Kasichayanula, James Ralph, Piotr Luszczek, Daniel Terpstra, and Shirley Moore. Measuring energy and power with PAPI. *International Workshop on Power-Aware Systems and Architectures* (*PASA 2012)*, Pittsburgh, PA, September 10, 2012.
- Shirley Moore and James Ralph. User-defined events for hardware performance monitoring. *International Conference on Computational Science (ICCS)*, Singapore, June 2011.

Acknowledgments

- The work is partially supported by
  - NSF Grants CNS-0911023, CNS-0910899, CNS-0910784, CNS-0905187
  - DOE SciDAC Grant DE-FC02-06ER25761
  - NVIDIA

Questions?