

Energy and Performance Characteristics of MPI and Hybrid (MPI/OpenMP) Scientific Applications on Multicore Systems

Charles Lively III*, Xingfu Wu*, Valerie Taylor*, Shirley Moore+,
Hung-Ching Chang^, Chun-Yi Su^, and Kirk Cameron^

*Department of Computer Science & Engineering, Texas A&M University

+Electrical Engineering and Computer Science, University of Tennessee-Knoxville

^Department of Computer Science, Virginia Tech

Outline

- Introduction
- Background
- Two Approaches
 - Trade-Off Approach
 - Modeling Approach
- Conclusions & Future Work

Introduction

- Energy consumption becomes a major challenge for large-scale systems, e.g., peta- or exa-flops systems.
- Large-scale systems are hierarchical and consist of heterogeneous components.
- Models are needed to handle the complexity of large-scale systems and scientific apps
- **Goal: Efficient use of large-scale systems in terms of runtime and power consumption for scientific apps**
 - Initial focus on multicore systems

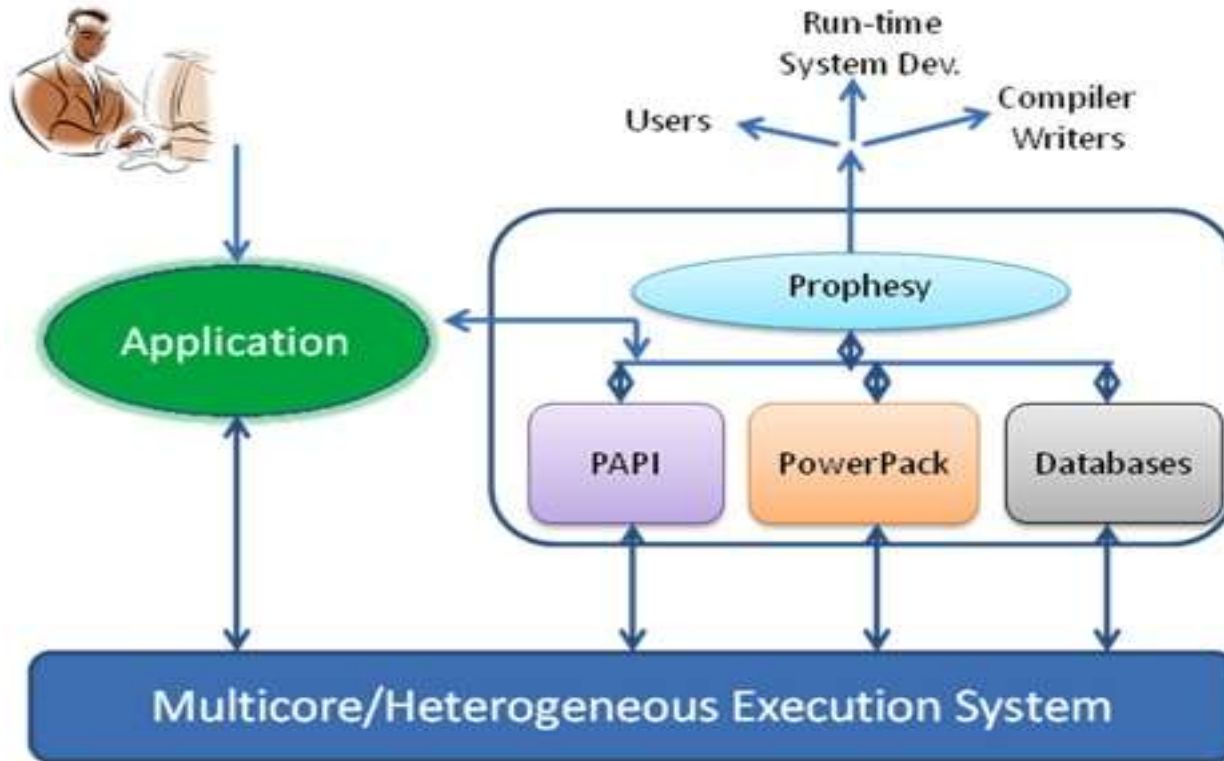
Background – Top 500 List

Rank	Site	Number of Cores	Power (KW)
1	RIKEN Advanced Institute for Computational Science (AICS), Japan	548352	9898.56
2	National Supercomputing Center in Tianjin, China	186368	4040.00
3	DOE/SC/Oak Ridge National Laboratory, USA	224162	6950.60
4	National Supercomputing Centre in Shenzhen (NSCS), China	120640	2580.00
5	GSIC Center, Tokyo Institute of Technology, Japan	73278	1398.61

Predicted Exa-System Configurations

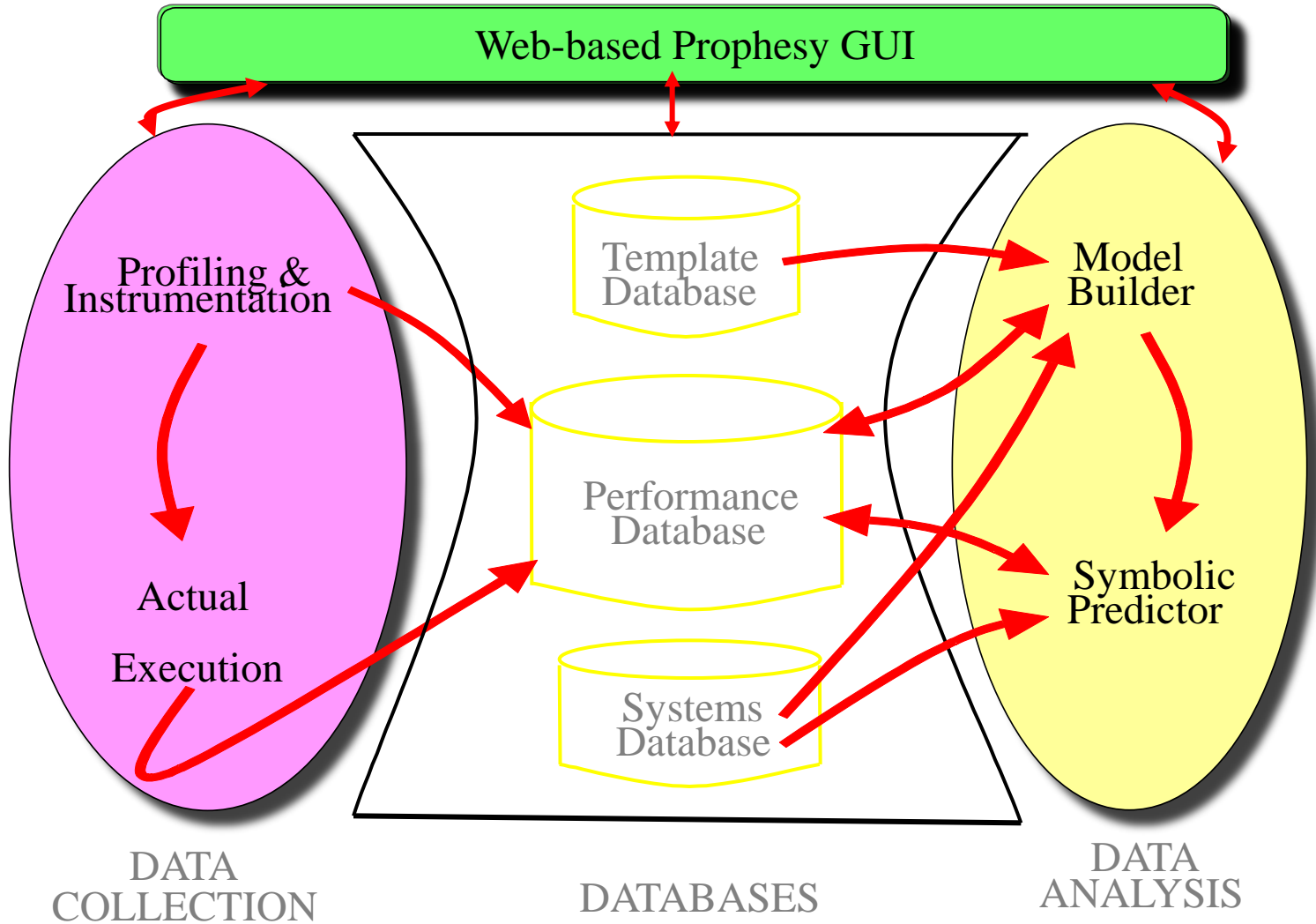
Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	O(Million)
Total Concurrency	225,000	3 Million	50 Million	O(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	O(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

MuMMI Framework

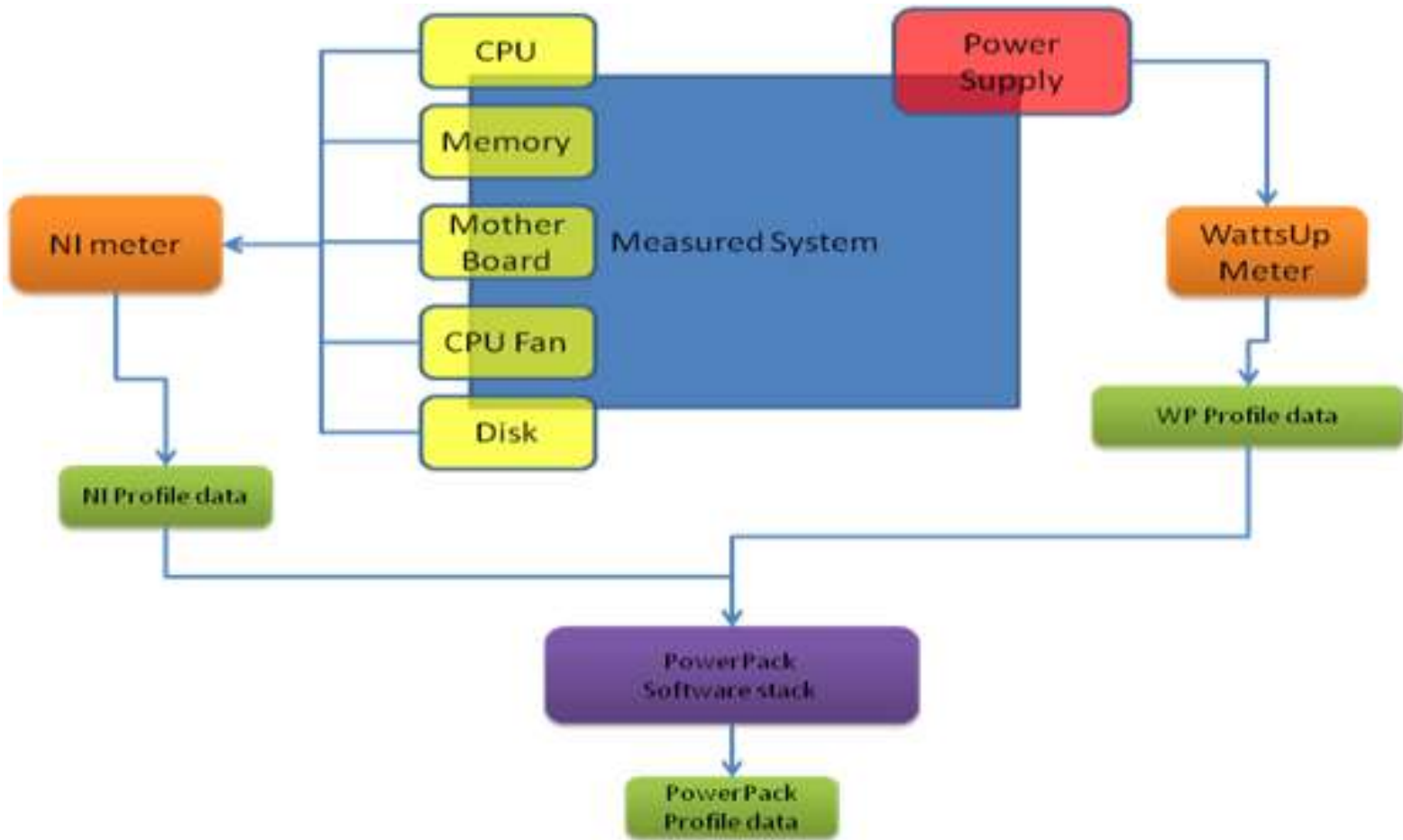


Multiple Metrics Modeling Infrastructure (MuMMI)

Prophesy Framework



PowerPack Schema

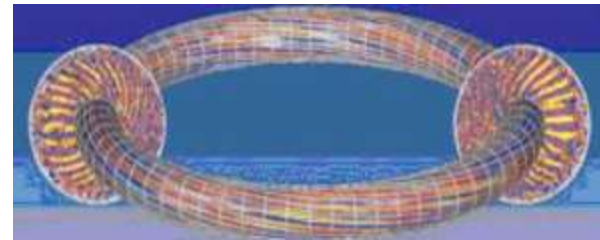


PowerPack 3.0 Framework

Outline

- Introduction
- Background
- Two Approaches
 - Trade-Off Approach
 - Modeling Approach
- Conclusions & Future Work

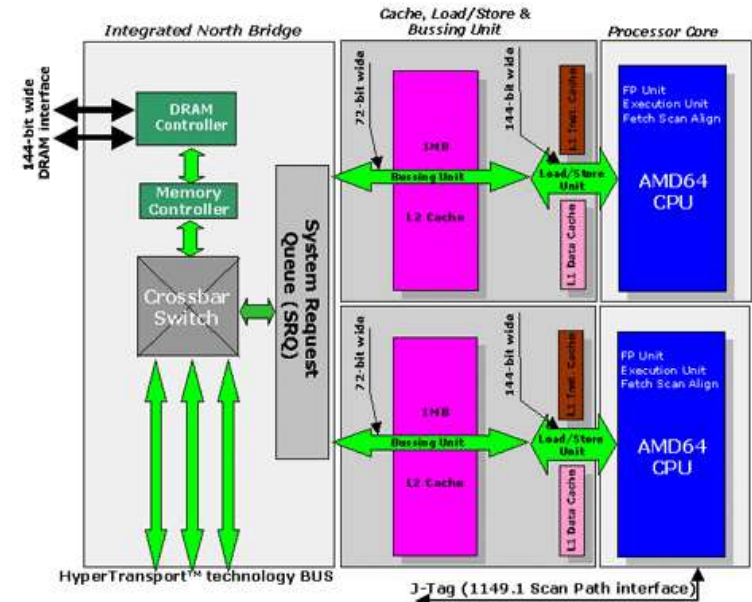
- NAS Multizone Benchmark Suite
 - Consists of BT-MZ, SP-MZ, and LU-MZ
 - written in Fortran
 - Uses MPI and OpenMP for communication
 - **Block Tri-diagonal algorithm (BT-MZ)**
 - represents realistic performance case for exploring the discretization meshes in parallel computing
- Large-Scale Scientific Application
 - **Gyrokinetic Toroidal code (GTC)**
 - 3D particle- in-cell application
 - Flagship SciDAC fusion microturbulence code
 - written in Fortran90
 - Uses MPI and OpenMP for communication



Execution Environment

Specification of Dori	
Configurations	Dori Virginia Tech
Number of Compute Nodes	8
CPUs per Node	4
CPU type	1.8GHz Opteron (dual-core)
Memory per Node	6GB

Data Flow view of the AMD Opteron™ Processor Dual core Model 100



The dual-core Opteron shares a set of three HyperTransport links and a dual-memory controller. The memory system is part of a NUMA architecture when utilized in a multiple-processor environment.

PowerPack is available on Dori for Power Profiling

Five frequency settings: 1.0Ghz - 1.8Ghz

Trade-Off Approach

- Multicore systems appear to be a natural match for hybrid MPI/OpenMP applications.
- Conducted detailed studies related to performance implications for MPI versus hybrid for multicore systems.
 - Comparison of performance of application implementations.
 - Analysis of application trends with frequency scaling.

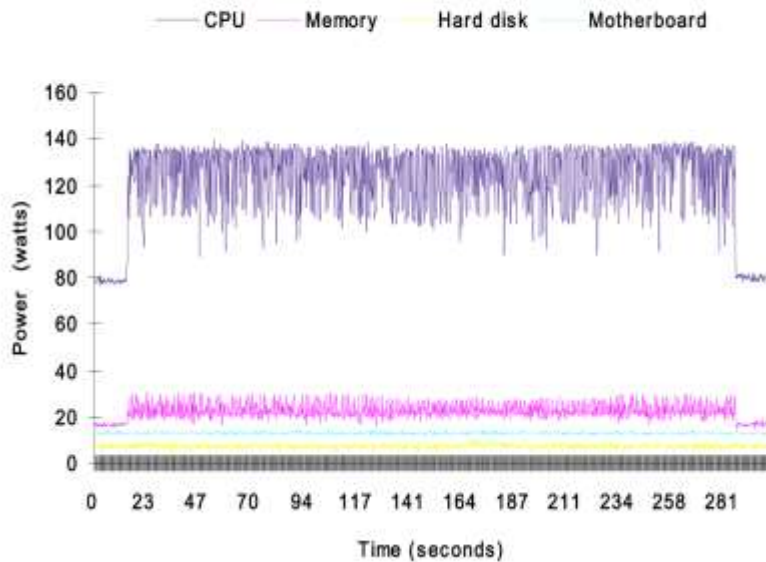
Execution Setup

- Constant Factors
 - Application Algorithm
 - Compiler
 - Programming languages and libraries
- Parallel programming paradigms
 - MPI used for inter-node communication
 - OpenMP used for intra-node communication
 - Different communication patterns

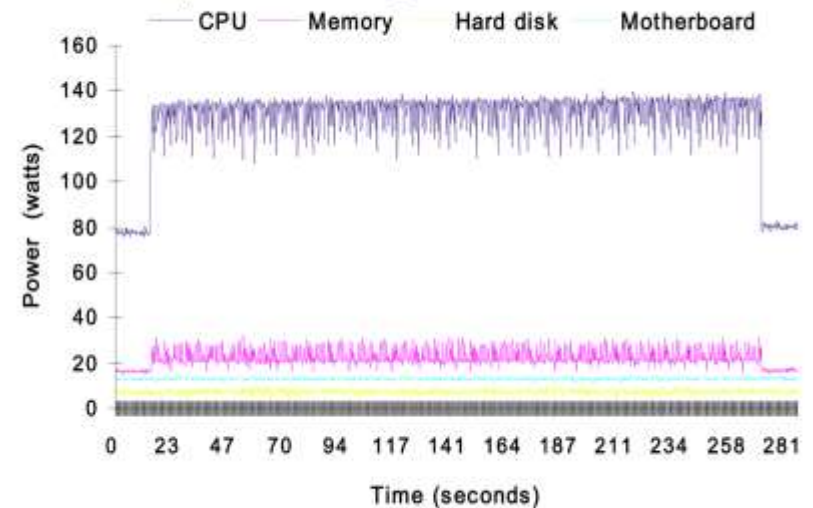
Performance and Energy for BT with Class B on 4 cores

On One Node	Performance	Total Energy
MPI BT	269 s	58,643 J
Hybrid BT	257 s	57,779 J
% improvement	4.46%	1.47%

MPI BT running on 1 node with 4 cores/node



OpenMP BT running on 1 node with 4 cores/node



Performance and Energy for BT with Class B on 16 cores

On 4 nodes	Performance	Total Energy
MPI BT	76.174 s	16,702.200 J
Hybrid BT	71.723 s	15,941.091 J
% improvement	6.2 %	4.56 %

CPU Speed	BT Type	Runtime(s)	System Energy(J)
1.8Ghz	Hybrid	71.723 (-25.31%)	15941.091 (10.36%)
	MPI Only	76.174 (-27.82%)	16702.200 (15.63%)
1.6Ghz	Hybrid	76.139 (-21.80%)	15058.230 (4.25%)
	MPI Only	81.841 (-15.94%)	15903.052 (10.1%)
1.4Ghz	Hybrid	84.849 (-12.86%)	14732.076 (1.99%)
	MPI Only	90.530 (-7.02%)	15624.080 (8.17%)
1.2Ghz	Hybrid (BASELINE)	97.366	14444.036
	MPI Only	101.990 (4.74%)	15088.793 (4.46%)
1.0Ghz	Hybrid	111.947 (14.97%)	17041.246 (17.98%)
	MPI Only	117.394 (20.56%)	17774.750 (23.06%)

Performance and Energy (default CPU Frequency of 1.8GHz) for GTC

#Cores	GTC Type	Runtime(s)	System Energy(KJ)
1x4	Hybrid	1302.773 (-59.3%)	270.223 (-60.8%)
	MPI (baseline)	2075.376	434.524
2x4	Hybrid	1395.322 (-59.9%)	576.674 (-60.47%)
	MPI (baseline)	2231.652	925.401
4x4	Hybrid	1434.491 (-62.05%)	1182.959 (-62.35%)
	MPI (baseline)	2324.707	1920.578
8x4	Hybrid	1463.457 (-72%)	2419.985 (-72.03%)
	MPI (baseline)	2528.556	4162.998

Performance and Energy for GTC on 16 Cores

CPU Speed	GTC Type	Runtime(s)	Total Energy (KJ)
1.8Ghz	Hybrid	1434.491 <i>(-8.62%)</i>	1182.959 <i>(3.72%)</i>
	MPI	2324.707 <i>(48.1%)</i>	1920.578 <i>(68.50%)</i>
1.6Ghz	Hybrid (Baseline)	1569.960	1139.831
	MPI	2511.532 <i>(59.97%)</i>	2057.516 <i>(80.51%)</i>
1.4Ghz	Hybrid	1773.444 <i>(12.96%)</i>	1143.615 <i>(0.03%)</i>
	MPI	2791.607 <i>(77.81%)</i>	1778.682 <i>(8.00%)</i>
1.2Ghz	Hybrid	2094.598 <i>(33.40%)</i>	1162.393 <i>(1.97%)</i>
	MPI	3126.446 <i>(99.1%)</i>	1724.057 <i>(51.26%)</i>
1.0Ghz	Hybrid	2445.155 <i>(37.87%)</i>	1393.650 <i>(22.26%)</i>
	MPI	3553.982 <i>(127.37%)</i>	2015.483 <i>(76.82%)</i>

Modeling Approach

- **Application-centric models** are used to explore common and different characteristics of Hybrid (MPI/OpenMP) Scientific Applications.
 1. Which combination of performance counters should be used to model each performance components?
 - Runtime and Power Consumption of System, CPU, and Memory.
 2. What application characteristics affect runtime and power consumption in Hybrid scientific applications?
 3. What characteristics of Hybrid applications need to be optimized to improve performance on multicore systems?

SystemG (Virginia Tech)



- Largest Power-Aware Compute System in the World.
- Over 30 power and thermal sensors per node.
- 324 nodes; 2,592 cores

- Training Set: 5 training execution points.
 - 1x1, 1x2, 1x3, 1x8, and 2x8
- 16 Larger execution points were predicted.
 - 1x4, 1x5,...3x8, 4x8, 5x8,16x8
- 40 performance counter events are captured.
 - Using PAPI and Perfmon Library
- Performance counter events are normalized per cycle.
- Performance-Tuned Supervised Principal Component Analysis Method utilized to select combination of performance counters for each application.

Performance-Tuned Supervised PCA

1. Compute Spearman's rank correlation for each application and performance component.
1. Eliminate counters with low correlation.
2. Compute regression model based upon performance counter event rates.
3. Eliminate performance counters with negligible regression coefficients.
4. Compute principal components of reduced performance counter event rates.
5. Use the performance counters with highest PCA vectors to build multivariate linear regression model.

Performance Counter Events

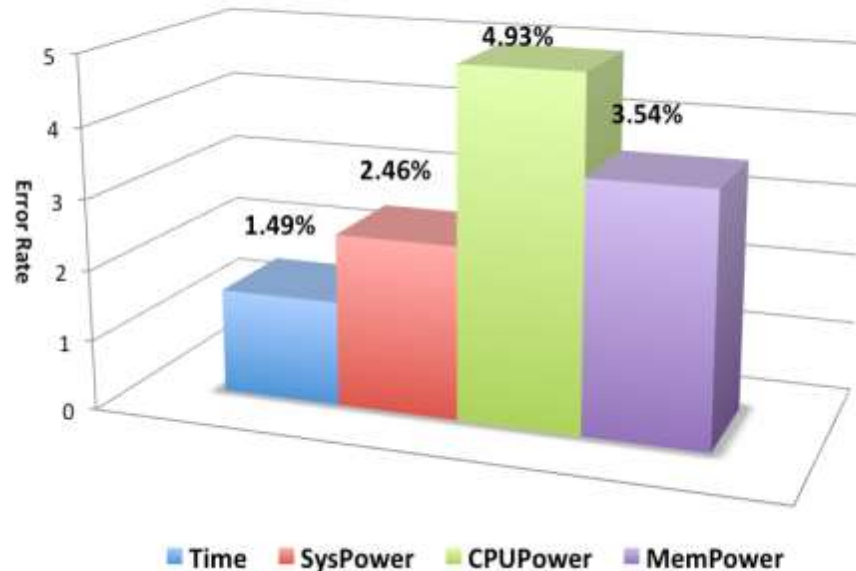
- 15 Performance Counters used in this Work

Hardware Counter	Description
PAPI_TOT_INS	Total instructions completed
PAPI_TLB_DM	TLB misses
PAPI_L1_TCA	L1 cache total accesses
PAPI_L1_ICA	L1 instruction cache accesses
PAPI_L1_TCM	L1 total cache misses
PAPI_L1_DCM	L1 data cache misses
PAPI_L2_TCH	L2 total cache hits
PAPI_L2_TCA	L2 total cache accesses
PAPI_L2_ICM	L2 instruction cache misses
PAPI_BR_INS	Branch instructions completed
PAPI_RES_STL	System stalls on any resource
Cache_FLD_per_instruction	L1 writes/reads/hits/misses
LD_ST_stall_per_cycle	Load/stores stalls per cycle

BT-MZ Results

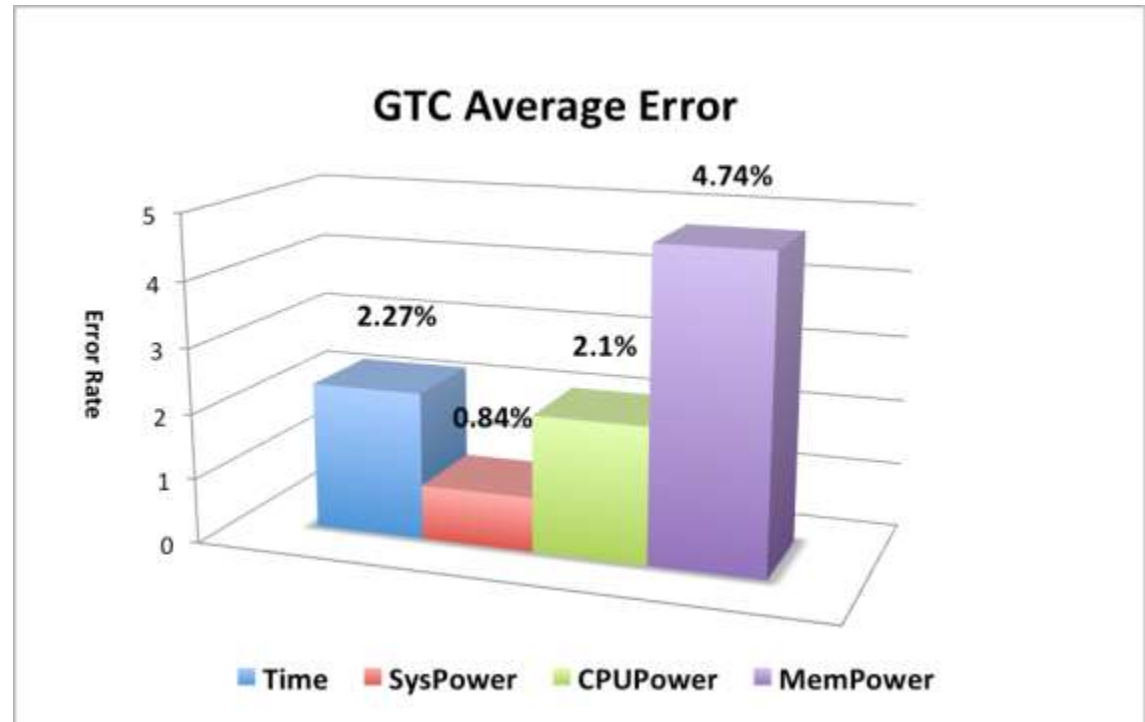
	Time		System Power		CPU Power		Memory Power	
BT-MZ	Cache_FLD	-1.611	PAPI_L2_TCH	-1.6769	PAPI_L1_TCM	3.5432	PAPI_L1_TCA	0.0763
	PAPI_TOT_INS	0.0967	PAPI_L2_TCA	1.5967	PAPI_L2_TCH	-3.9389	PAPI_L1_DCM	4.0496
	PAPI_L2_TCH	0.2992	PAPI_RES_STL	0.0803	PAPI_RES_STL	0.3967	PAPI_L2_TCH	-1.9443
	PAPI_L2_TCA	1.2152					PAPI_L2_TCA	2.1806

BT-MZ Average Error



GTC Results

	Time		System Power		CPU Power		Memory Power	
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		



Conclusions

- CPU power consumption is dominated (more than 55%)
- Predictive performance models analyze the performance characteristics of Scientific Applications.
 - Execution time
 - System Power Consumption
 - CPU Power Consumption
 - Memory Power Consumption
- 97+% Accuracy across four Hybrid (MPI/OpenMP) Scientific Applications.
- Determines application characteristics that affect component performance.

Future Work

- Incorporation of additional Large-Scale Hybrid Scientific Applications.
 - Parallel Lattice Boltzman Method
 - Parallel Ocean Program (POP)
 - Etc....
- Compare performance and modeling of Hybrid (MPI/OpenMP) versus MPI-only implementation.
- Modeling and Prediction across different application input sizes and frequency settings.
- Exploration of nonlinear prediction models.

Questions?

