PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

MuMI

Energy and Performance Characteristics of MPI and Hybrid Scientific Applications on Multicore Systems

Xingfu Wu, Charles Lively, and Valerie Taylor

Department of Computer Science & Engineering
Texas A&M University

CCGSC2010, Sep. 8, 2010
Highland Lake Inn, Flat Rock, North Carolina

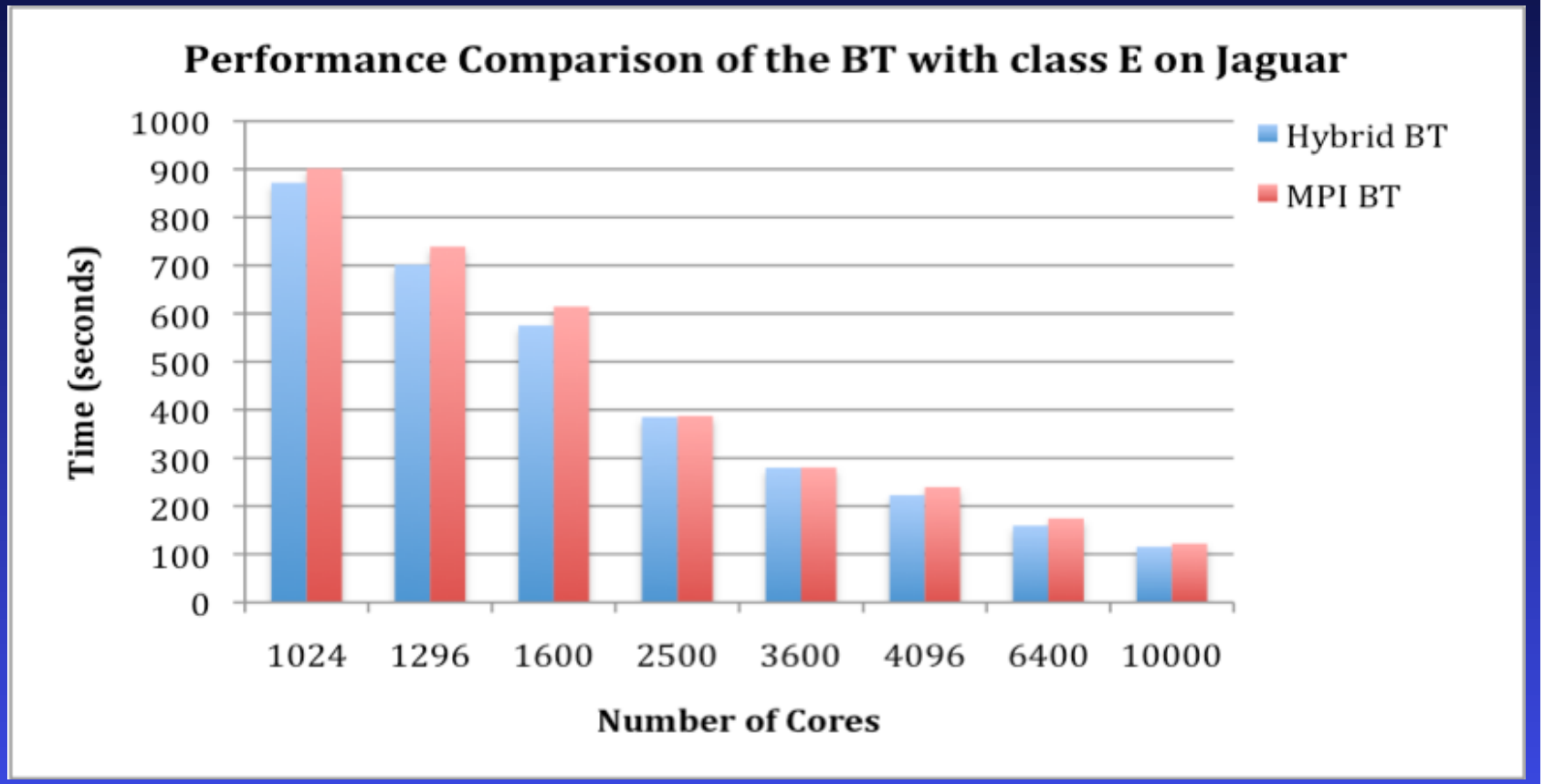
Outline

- Introduction
- Scientific Applications
- Methodology and Experiment Platform
- Experimental Results
- NSF-funded MuMI Project

Introduction

- Energy consumption becomes a major challenge for using multicore to build peta- or exa-flops systems
- Multicore systems appear to be a natural match for hybrid MPI/OpenMP applications
- Conducted initial studies related to performance implications for MPI versus hybrid for multicore systems

Performance Comparison of BT
(strong scaling)



Scientific Applications

- NAS Parallel Benchmark Suite (version 3.3): Block Tridiagonal (BT)
  - Class B
  - Strong scaling
- Gyrokinetic Toroidal Code (GTC)
  - 100 particles per cell
  - Weak scaling
- Lattice Boltzmann Application (LBM)
  - 3D Mesh, 64x64x64
  - Strong scaling

Experimental 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

Methodology

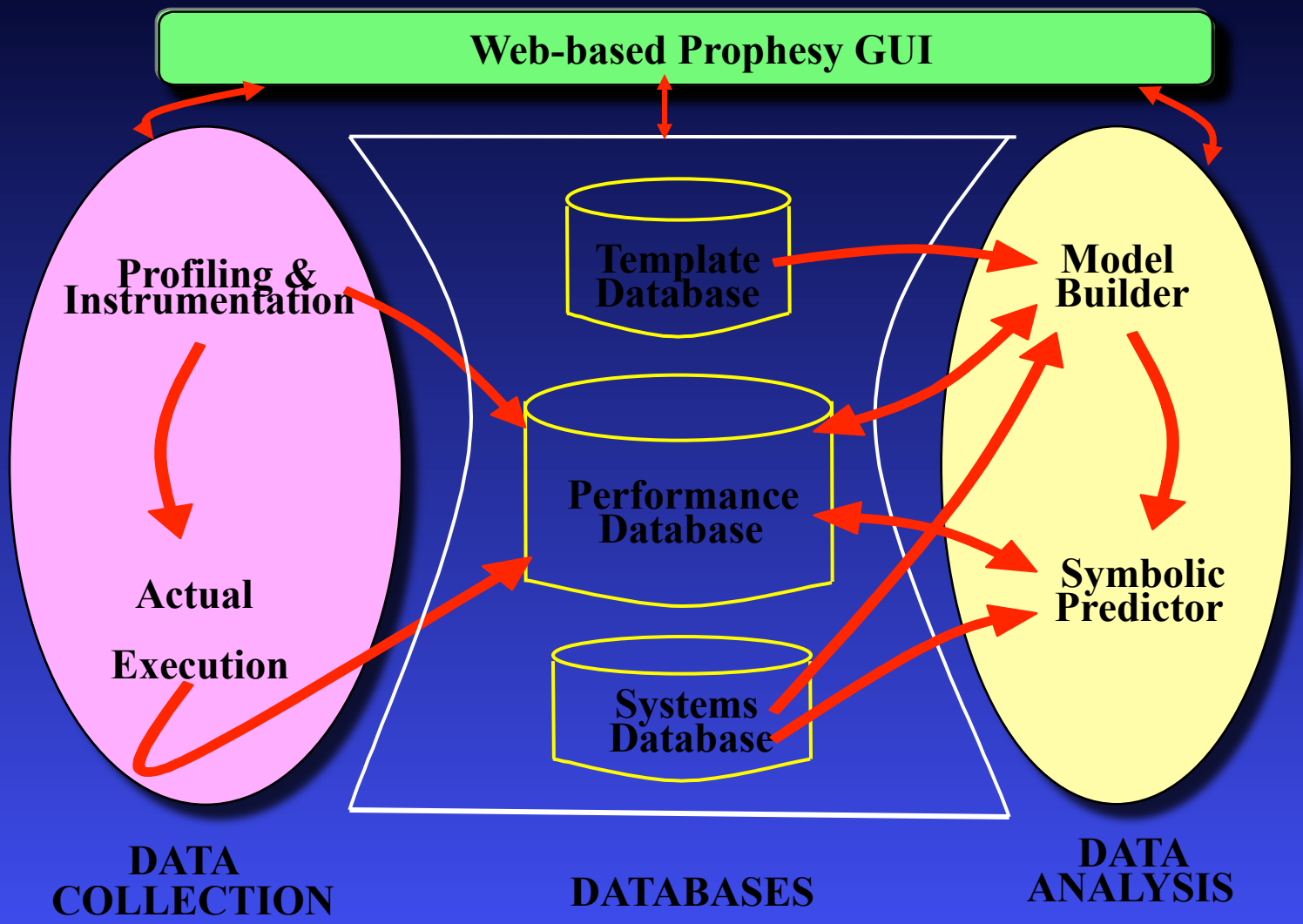
- Use Prophesy to collect application performance data
- Use PowerPack to obtain power profiling data for the application execution

PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

Prophesy System



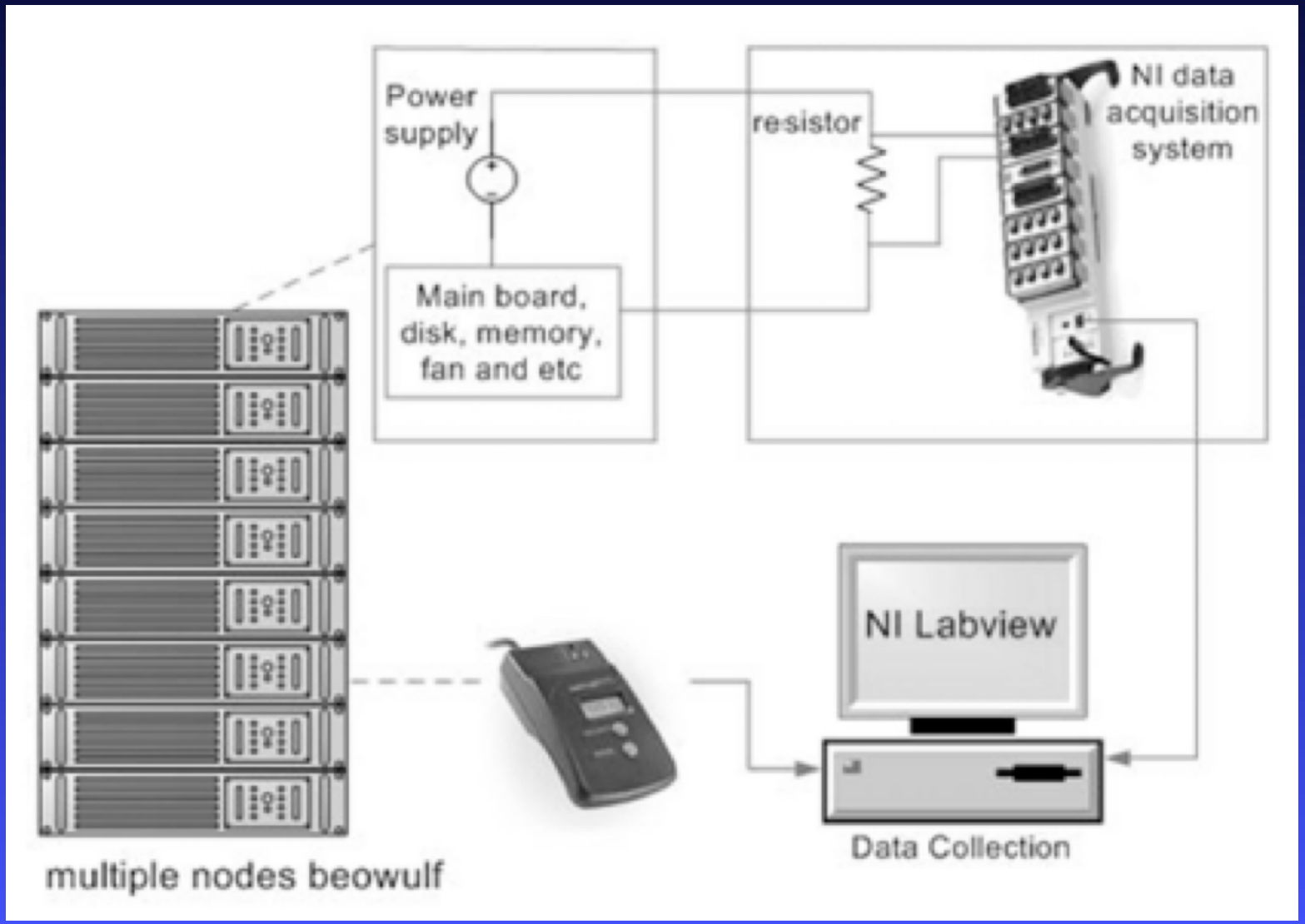
MuMI

PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

PowerPack (Virginia Tech)



MuMI

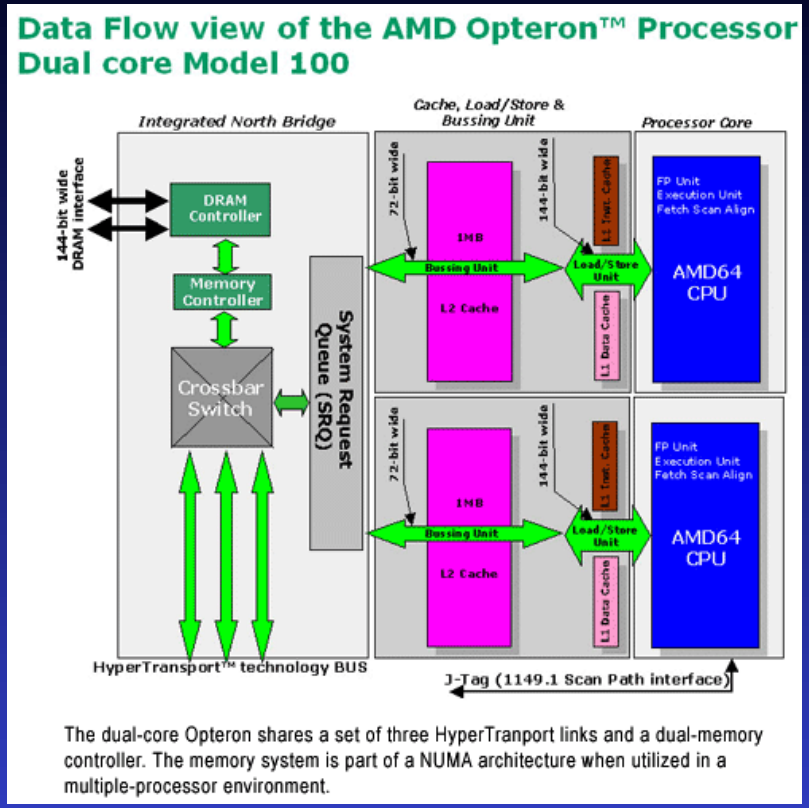
PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

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**



The dual-core Opteron shares a set of three HyperTranport 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: 1.8Ghz, 1.6Ghz, 1.4Ghz, 1.2Ghz, 1.0Ghz

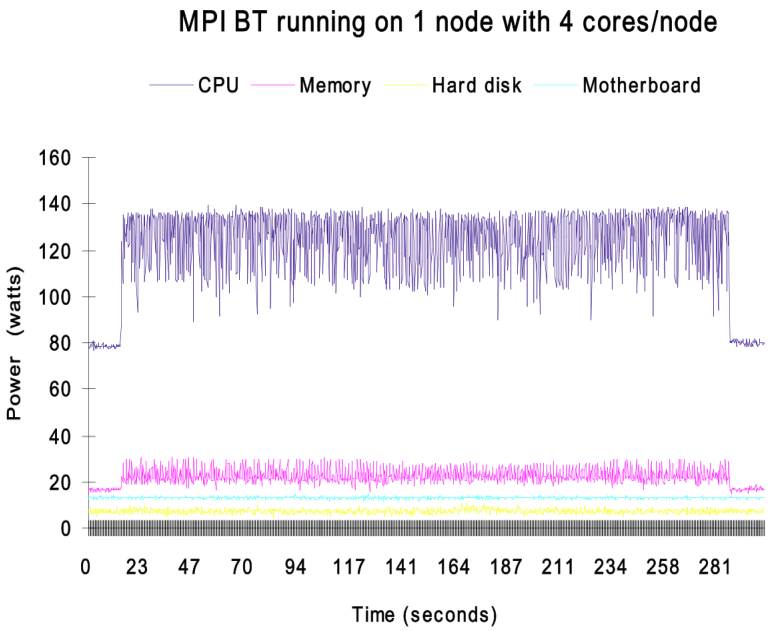
MuMI

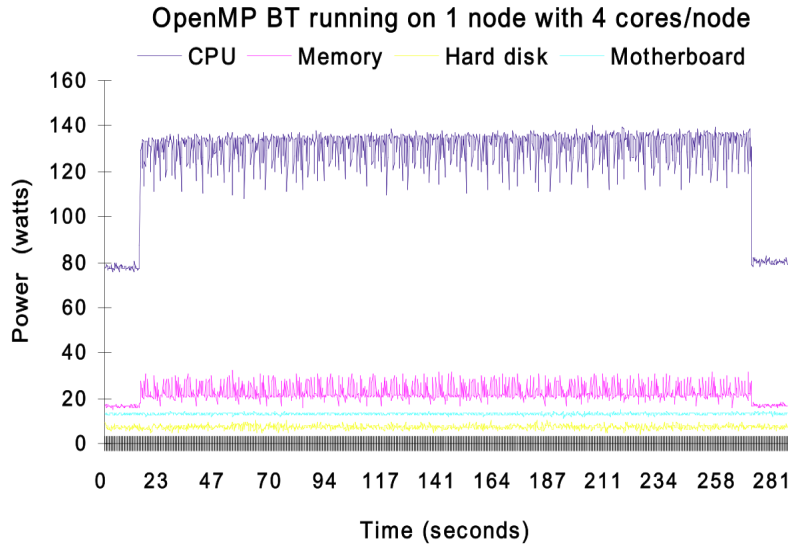
PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

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%





MuMI

PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

MuMI

Performance and Energy for BT with Class B on 16 cores

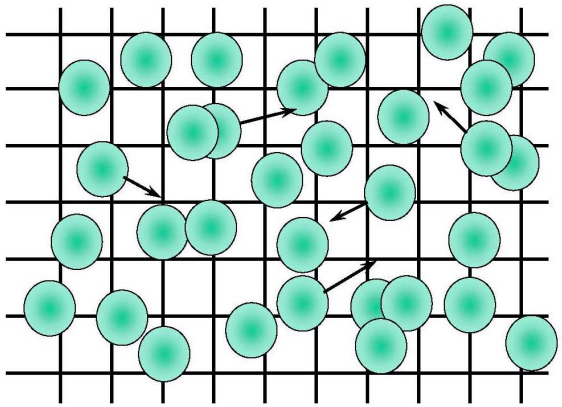
On 4 nodes	Performance	Total Energy
MPI BT	76.174 s	16,702.200 J
Hybrid BT	71.723 s	15,941.091 J
% improvement	6.2 %	4.56 %

CPU Speed	BT Type	Runtime(s)	System Energy(J)	CPU Energy (J)	Memory Energy (J)	Disk Energy (J)	Motherboard Energy (J)
1.8Ghz	Hybrid	71.723 (-25.31%)	15941.091 (10.36%)	9453.668 (22.88%)	1580.718 (-26.76%)	508.679 (-30.83%)	919.018
	MPI Only	76.174 (-27.82%)	16702.200 (15.63%)	9986.521	1765.706	554.347	997.488
1.6Ghz	Hybrid	76.139 (-21.80%)	15058.230 (4.25%)	8737.304 (13.57%)	1713.132 (-20.62%)	566.728 (-22.94%)	1000.655
	MPI Only	81.841 (-15.94%)	15903.052 (10.1%)	9088.220	1858.386	598.208	1062.258
1.4Ghz	Hybrid	84.849 (-12.86%)	14732.076 (1.99%)	8186.828 (6.41%)	1852.877 (-14.15%)	601.683 (-18.19%)	1091.145
	MPI Only	90.530 (-7.02%)	15624.080 (8.17%)	8577.551	1992.754	639.778	1163.661
1.2Ghz	**Hybrid (BASELINE)**	**97.366**	***14444.036***	**7693.547**	**2158.369**	**735.495**	**1288.510**
	MPI Only	101.990 (4.74%)	15088.793 (4.46%)	8101.081	2330.107	803.493	1372.160
1.0Ghz	Hybrid	111.947 (14.97%)	17041.246 (17.98%)	9325.778 (21.22%)	2480.800 (14.93%)	873.530 (18.77%)	1503.207
	MPI Only	117.394 (20.56%)	17774.750 (23.06%)	9630.939	2606.256	898.152	1559.354

GTC Code

- Gyrokinetic Toroidal code (GTC)
  - A 3D particle-in-cell application developed at the Princeton Plasma Physics Laboratory to study turbulent transport in magnetic fusion
- A flagship DOE SciDAC fusion microturbulence code
- 100 particles per cell and 100 time steps
- Weak scaling

Stephane Ethier's Talk in 2005 BlueGene Applications Workshop



The PIC Steps

- "SCATTER", or deposit, charges on the grid (nearest neighbors)
- Solve Poisson equation
- "GATHER" forces on each particle from potential
- Move particles (PUSH)
- Repeat…

PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

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

#Cores	GTC Type	Runtime(s)	System Energy(KJ)	CPU Energy (KJ)	Memory Energy (KJ)	Disk Energy (KJ)	Motherboard Energy (KJ)
1x4	Hybrid	1302.773 (**-59.3%**)	270.223 (**-60.8%**)	162.969 (-62.65%)	27.086 (-50.31%)	9.699 (-59.24%)	17.119 (-59.01%)
	MPI (baseline)	2075.376	434.524	265.071	40.714	15.445	27.221
2x4	Hybrid	1395.322 (**-59.9%**)	576.674 (**-60.47%**)	353.826 (-62.23%)	61.887 (-52.27%)	23.801 (-61.55%)	38.753 (-60.84%)
	MPI (baseline)	2231.652	925.401	574.003	94.238	384.501	62.333
4x4	Hybrid	1434.491 (**-62.05%**)	1182.959 (**-62.35%**)	711.065 (-64.76%)	118.186 (-52.99%)	41.824 (-64.64%)	74.670 (-62.81%)
	MPI (baseline)	2324.707	1920.578	1171.572	180.825	68.858	121.571
8x4	Hybrid	1463.457 (**-72%**)	2419.985 (**-72.03%**)	1457.945 (-73.59%)	244.013 (-60.58%)	86.806 (-71.54%)	153.596 (-71.82%)
	MPI (baseline)	2528.556	4162.998	2530.861	391.842	148.906	263.909

MuMI

PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

Performance and Energy for GTC on 16 Cores

CPU Speed	GTC Type	Runtime(s)	Total Energy (KJ)	CPU Energy (KJ)	Memory Energy (KJ)	Disk Energy (KJ)	Motherboard Energy (KJ)
1.8Ghz	Hybrid	1434.491 (-8.62%)	1182.959 (3.72%)	711.065 (7.1%)	118.186 (-8.09%)	41.824 (-10.81%)	74.669 (-9.26%)
	MPI	2324.707 (48.1%)	1920.578 (68.50%)	1171.572	180.825	68.858	121.571
1.6Ghz	**Hybrid (Baseline)**	**1569.960**	***1139.831***	**664.098**	**128.594**	**46.894**	**82.292**
	MPI	2511.532 (59.97%)	2057.516 (80.51%)	1253.041	196.440	76.902	133.030
1.4Ghz	Hybrid	1773.444 (12.96%)	1143.615 (0.03%)	661.161 (0.04%)	153.450 (19.39%)	59.649 (27.19%)	98.017 (19.10%)
	MPI	2791.607 (77.81%)	1778.682 (8.00%)	1040.457	230.353	93.187	153.778
1.2Ghz	Hybrid	2094.598 (33.40%)	1162.393 (1.97%)	628.386 (-5.37%)	176.897 (37.56%)	68.966 (47.1%)	114.914 (39.64%)
	MPI	3126.446 (99.1%)	1724.057 (51.26%)	940.227	254.275	103.819	171.746
1.0Ghz	Hybrid	2445.155 (37.87%)	1393.650 (22.26%)	769.366 (15.85%)	204.96 (4.34%)	81.417 (73.61%)	134.758 (63.76%)
	MPI	3553.982 (127.37%)	2015.483 (76.82%)	1112.277	285.326	115.870	193.778

MuMI

Parallel Lattice Boltzmann Method

- Lattice Boltzmann method (LBM) is widely used in simulating fluid dynamics.
- LBM is based on the kinetic theory, which entails a more fundamental level in studying the fluid than Navier-Stokes equations.
- This code (MPI) was developed by our Aerospace Engineering department.
- 3D Mesh partition is based on number of MPI processes

PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

Energy and Performance for LBM with the problem size of 64x64x64

#Cores	LBM Type	Runtime(s)	System Energy(KJ)	CPU Energy (KJ)	Memory Energy (KJ)	Disk Energy (KJ)	Motherboard Energy (KJ)
1x4	Hybrid	30.022 (25.33%)	6.337 (41.45%)	3.682	0.818	0.243	0.411
	MPI (baseline)	22.418	3.710	2.224	0.421	0.190	0.310
2x4	Hybrid	21.045 (15.78%)	8.629 (28.28%)	5.246	0.916	0.354	0.584
	MPI (baseline)	17.724	6.189	3.731	0.667	0.296	0.489
4x4	Hybrid	13.248 (5.46%)	10.534 (9.54%)	6.276	1.229	0.455	0.738
	MPI (baseline)	12.524	9.529	5.595	1.177	0.412	0.693
8x4	Hybrid	11.929 (-21.32%)	17.903 (-17.26%)	10.723	2.088	0.822	1.327
	MPI (baseline)	15.161	21.637	12.784	2.526	1.039	1.683

MuMI

What We've Learned

- The hybrid BT outperforms the MPI BT on 16 cores by 6.2% in performance and 4.56% in energy saving
- The hybrid GTC outperforms MPI GTC on 32 cores by 72% in performance and 72% in energy saving
- For LBM, the MPI outperforms the hybrid
  - 3D mesh partition based on number of MPI processes
  - Does not adequately take advantage of OpenMP
    - Requires a significant revision
- CPU power consumption is dominated (more than 55%)
  - DVFS is a good method to reduce power at the expense of an increase in execution time
- Further work is needed to explore larger number of cores and additional scientific applications

PROPHESY

http://prophesy.cs.tamu.edu

http://www.mumi-tool.org

MuMI (Multicore application Modeling Infrastructure) Project

Run-time System Dev.

Users

Compiler Writers

Application

Prophesy

PAPI

PowerPack

Databases

Multicore/Heterogeneous Execution System

Project Website: http://www.mumi-tool.org

MuMI

Acknowledgement

- Kirk Cameron, Hung-ching Chang from Virginia Tech for the use of PowerPack and Dori
- Shirley Moore from University of Tennessee (UTK) for providing the GTC code
- Bah Rabiou Ashraf, our REU student this summer, for his work on the hybrid LBM