

Performance Analysis and Optimization of Large-scale Scientific Applications on IBM POWER and BlueGene Supercomputers

Xingfu Wu, Valerie Taylor, Charles Lively, and Sameh Sharkawi

Department of Computer Science
Texas A&M University

IBM ScicomP14, May 20, 2008, Hudson Valley, NY

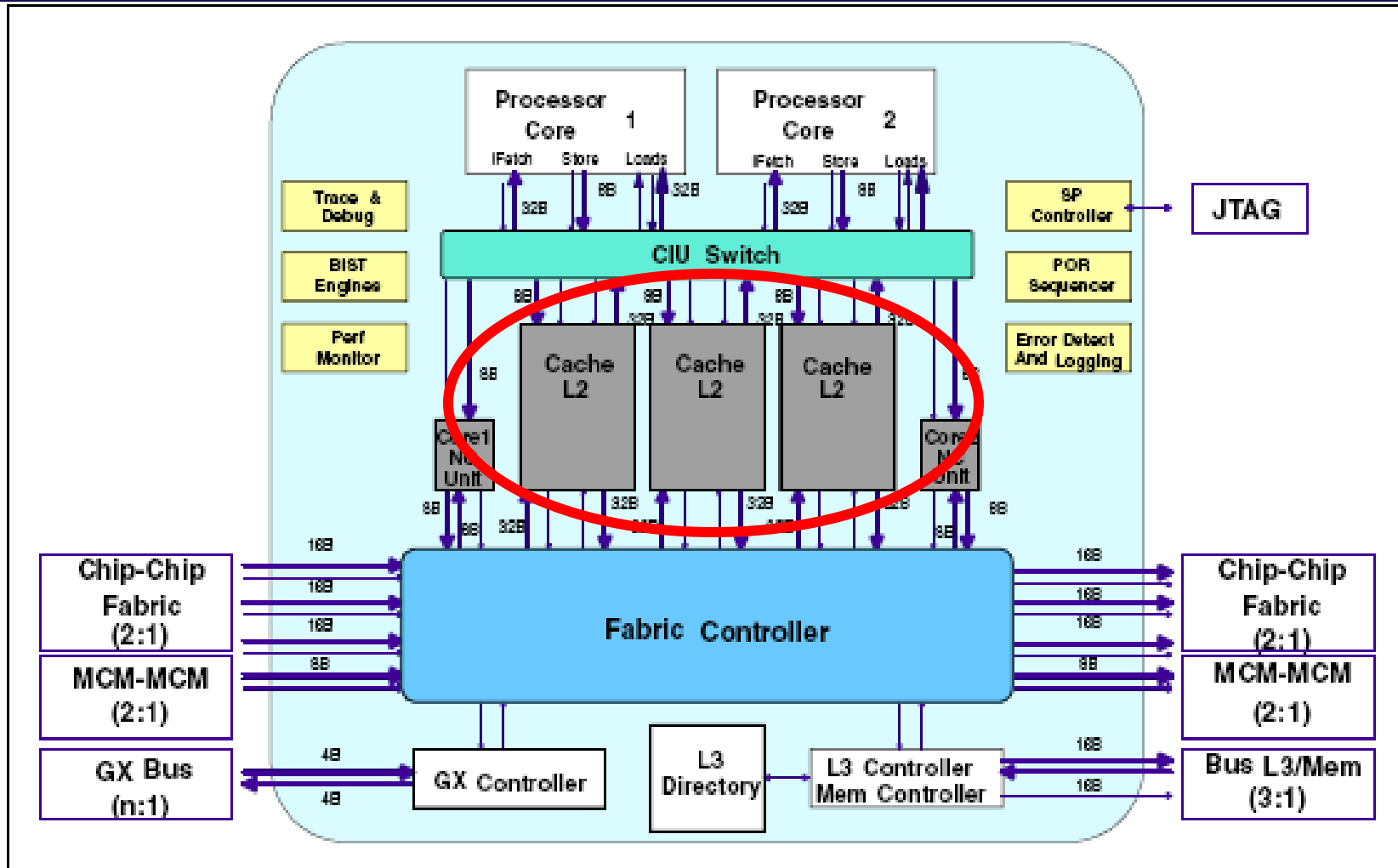
Outline

- **IBM POWER and BlueGene Systems**
- **Performance Analysis and Optimization Methods**
- **Case Studies: GTC, LBM, and GYRO**
- **Summary**

CMP (Chip MultiProcessors)

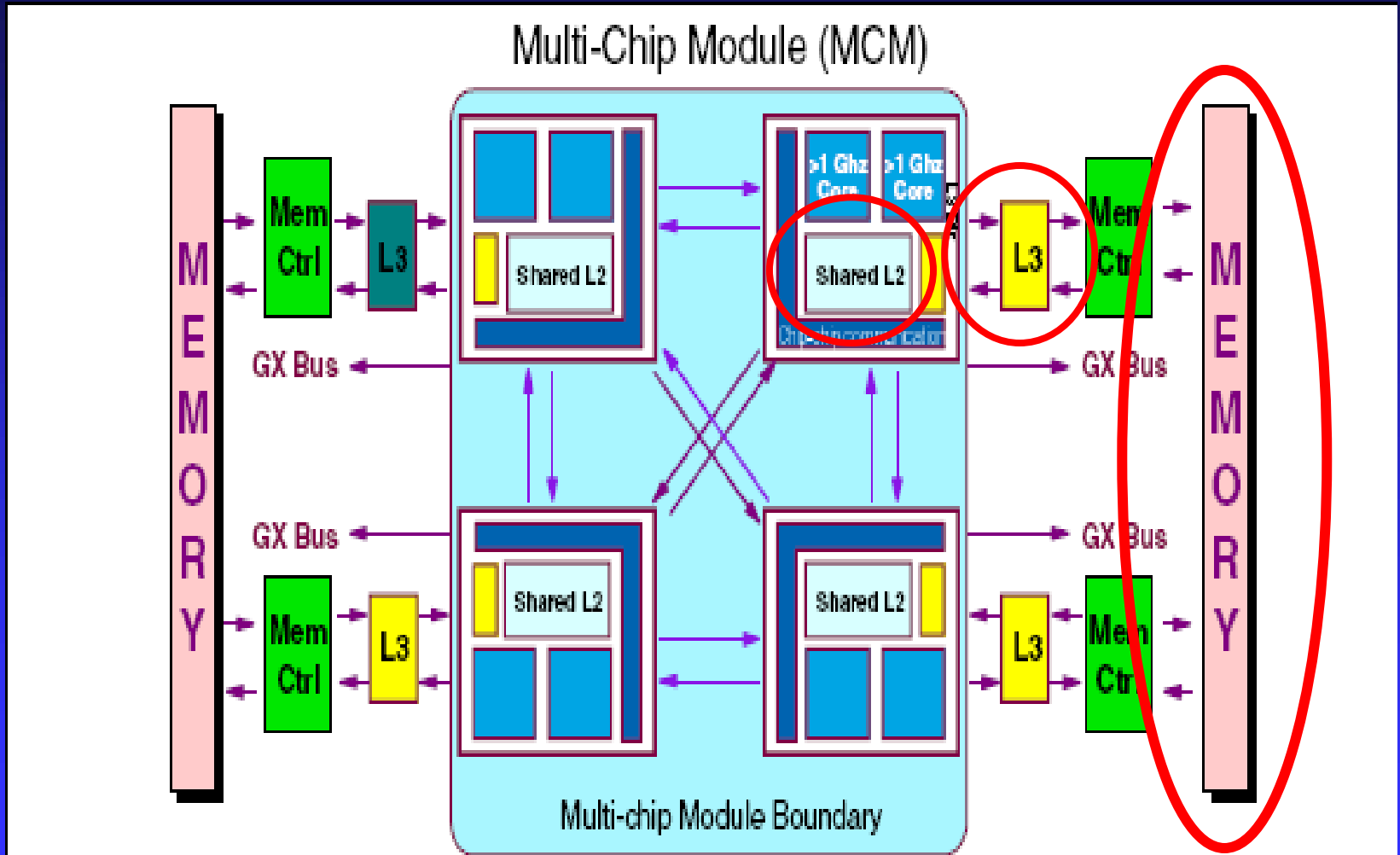
- Today, chip multiprocessors (CMPs) with multiple cores integrated on a single chip are made available for general purpose high performance computing.
- Trends toward multiple cores will continue.
 - ◆ IBM, AMD and Intel manufactured quad-core.
- Modern high performance computing systems have been shifting towards cluster systems with CMPs, for which some current systems have 2 to 32 processors per node.

POWER4 chip

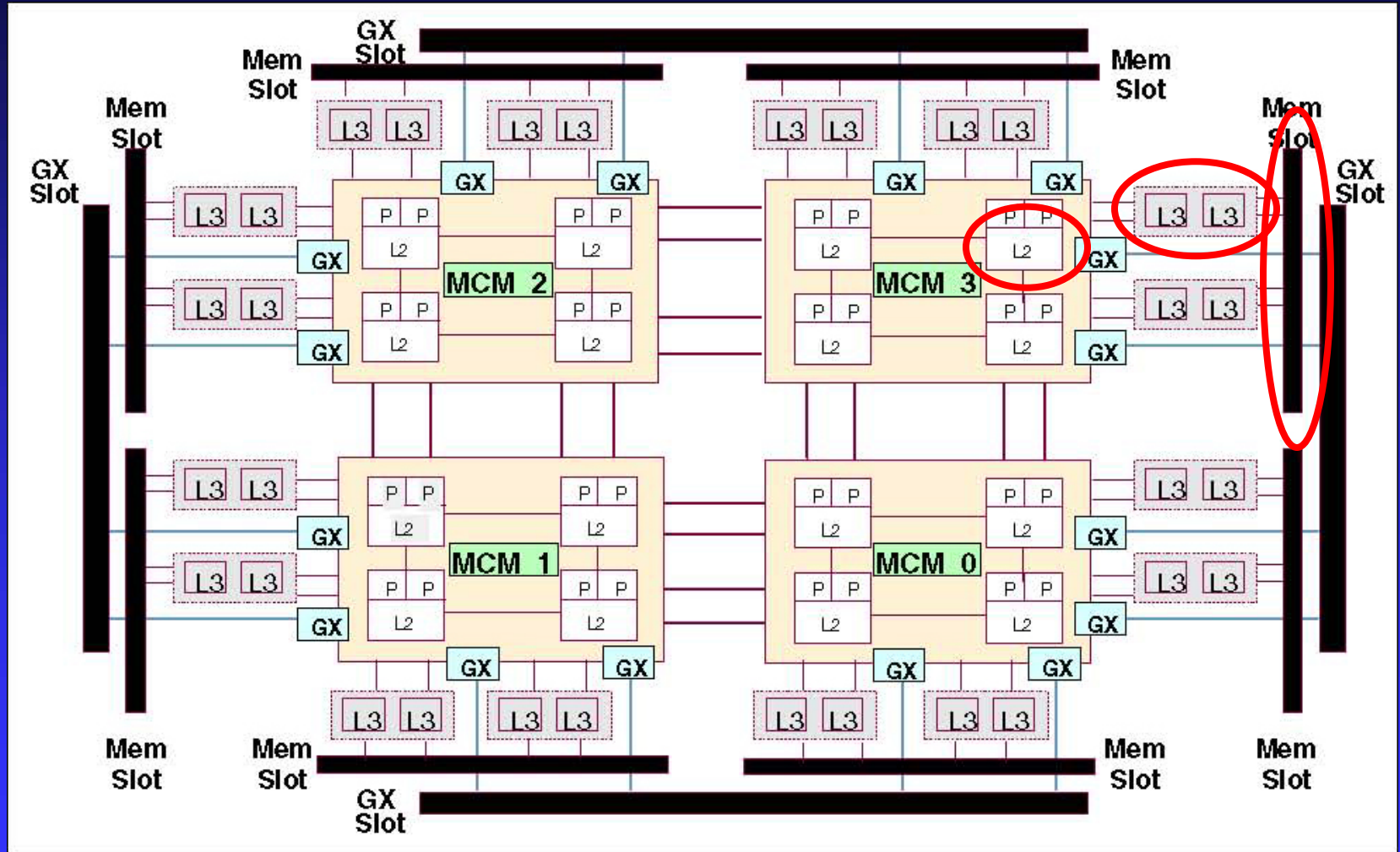


S. Behling, R. Bell, et al., *The POWER4 Processor Introduction and Tuning Guide*, IBM Redbooks, Nov. 2001.

POWER4 MCM (Multi-Chip-Module)



Multiple MCM Interconnection



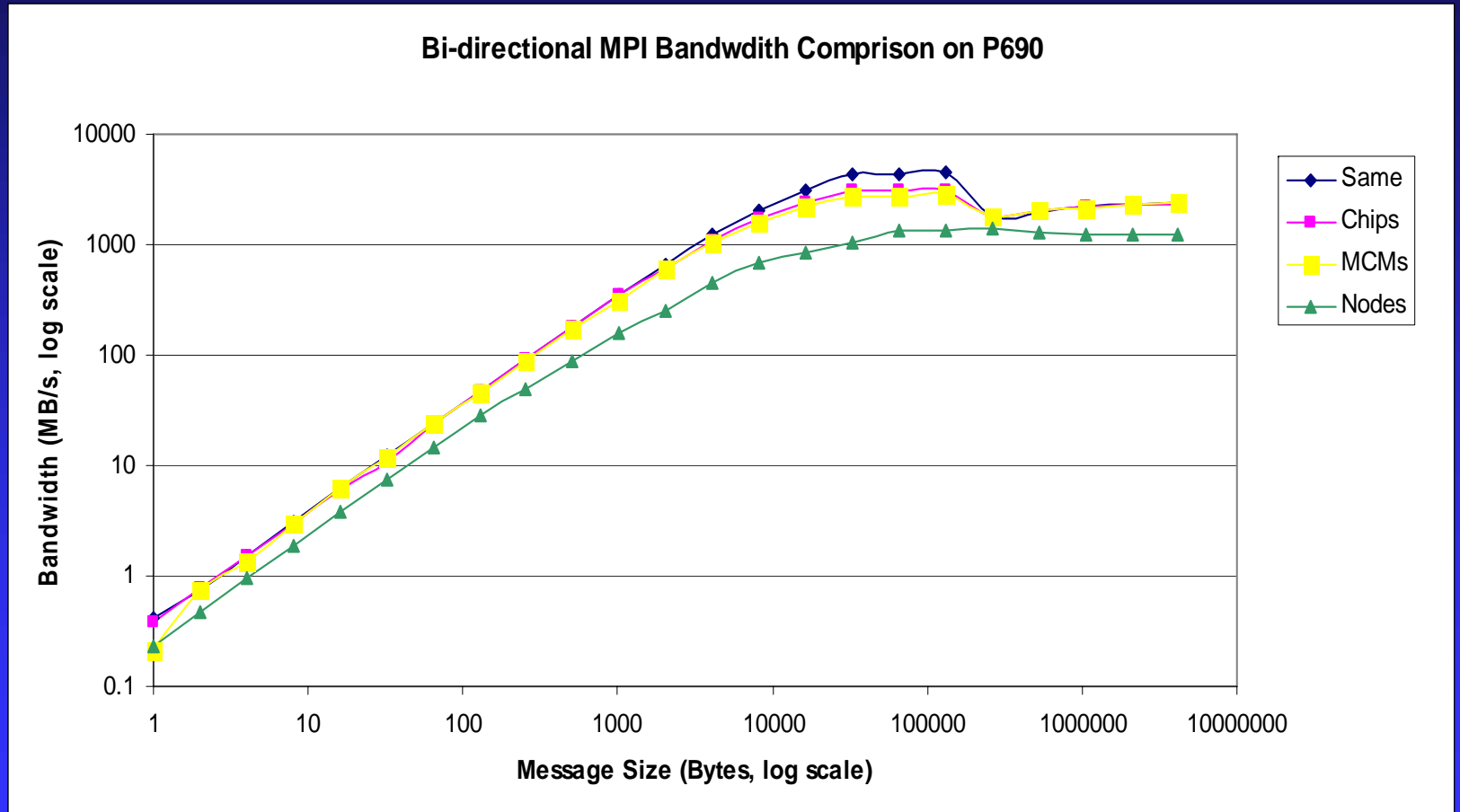
Execution Platforms

Configurations	UNC BlueGene/L	SDSC P655	SDSC P690
# Processors	2048	2176	227
# Nodes	1024	272	7
MCMs/Node	NA	1	4
chips/MCM	NA	4	4
Cores/chip	2	2	2
CPUs / Node	2	8	32
CPU type	700 MHz PowerPC	1.5,1.7GHz POWER4	1.7GHz POWER4
Memory/Node	1GB	16,32GB	128GB
L1 Cache/CPU	32KB	64/32 KB	64/32 KB
L2 Cache/chip	16 128-byte lines	1.41MB	1.41MB
L3 Cache/chip	4MB	32MB	32MB
Network	3D Torus	Federation	Federation

Processor Affinity Policy for POWER

- Process to processor assignment
- Order of assignment:
 - ◆ Chip
 - ◆ MCM
 - ◆ Node

MPI Performance on P690



Outline

- IBM POWER and BlueGene Systems
- **Performance Analysis and Optimization Methods**
- Case Studies: GTC, LBM, and GYRO
- Summary

Performance Analysis

- One issue is how many processors per node to use for efficient execution.
- It is expected that the best number of processors per node is dependent upon the application characteristics and system configurations.
- Quantify the performance gap resulting from using different number of processors per node for application execution (for which we use the term processor partitioning) .

Processor Partitioning

- Understand how processor partitioning impacts system & application performance
- Investigate how an application is sensitive to communication and memory access patterns
- Explore how we use the information to optimize applications in order to efficiently utilize all processors per node

Performance Optimization

- Take advantage of the hierarchical organization of processors
- Reduce conflicts from resource sharing
- Optimizations considered
 - ◆ Computation Optimization
 - ◆ Loop blocking
 - ◆ Loop unrolling/fusion
 - ◆ Communication Optimization
 - ◆ MPI_Allreduce, MPI_Alltoall ...

Communication Optimization

- **MPI_Allreduce: GTC and GYRO**
- **MPI_Allreduce Hybrid Method**
 - ◆ Measure performance of the original MPI_Allreduce with different message sizes on each system,
 - ◆ Measure performance of Rabenseifner's algorithm for allreduce in [TR05],
 - ◆ Compare both performance to decide message size ranges for the best performance,
 - ◆ Implement a hybrid method for MPI_Allreduce based on the message size ranges (basically using the original algorithm at small message sizes (1KB for P655, 0.25KB for BGL) and the Rabenseifner's algorithm otherwise).

Note: Rabenseifner's algorithm: reduce-scatter followed by an allgather

Outline

- IBM POWER and BlueGene Systems
- Performance Analysis and Optimization Methods
- **Case Studies: GTC, LBM, and GYRO**
- Summary

Large-scale Scientific Applications

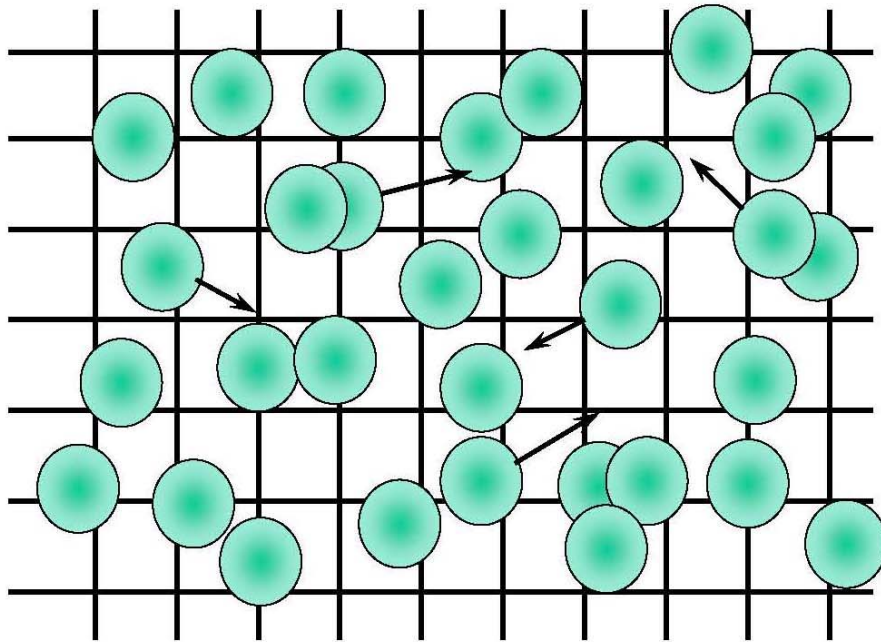
Application	Discipline	Problem Sizes	Programming Environments
GTC	Magnetic Fusion	100 particles per cell 100 time steps	Fortran90, MPI, OpenMP
LBM	Fluid Dynamics	128x128x128 512x512x512	C, MPI
GYRO	Plasma Physics	B1-std: 6x140x4x4x8x6 B2-cy: 6x128x4x4x8x6	Fortran90, MPI

Cast Study I: GTC (Barrier: memory contentions)

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.
- The flagship SciDAC fusion microturbulence code

PIC Steps of GTC Code

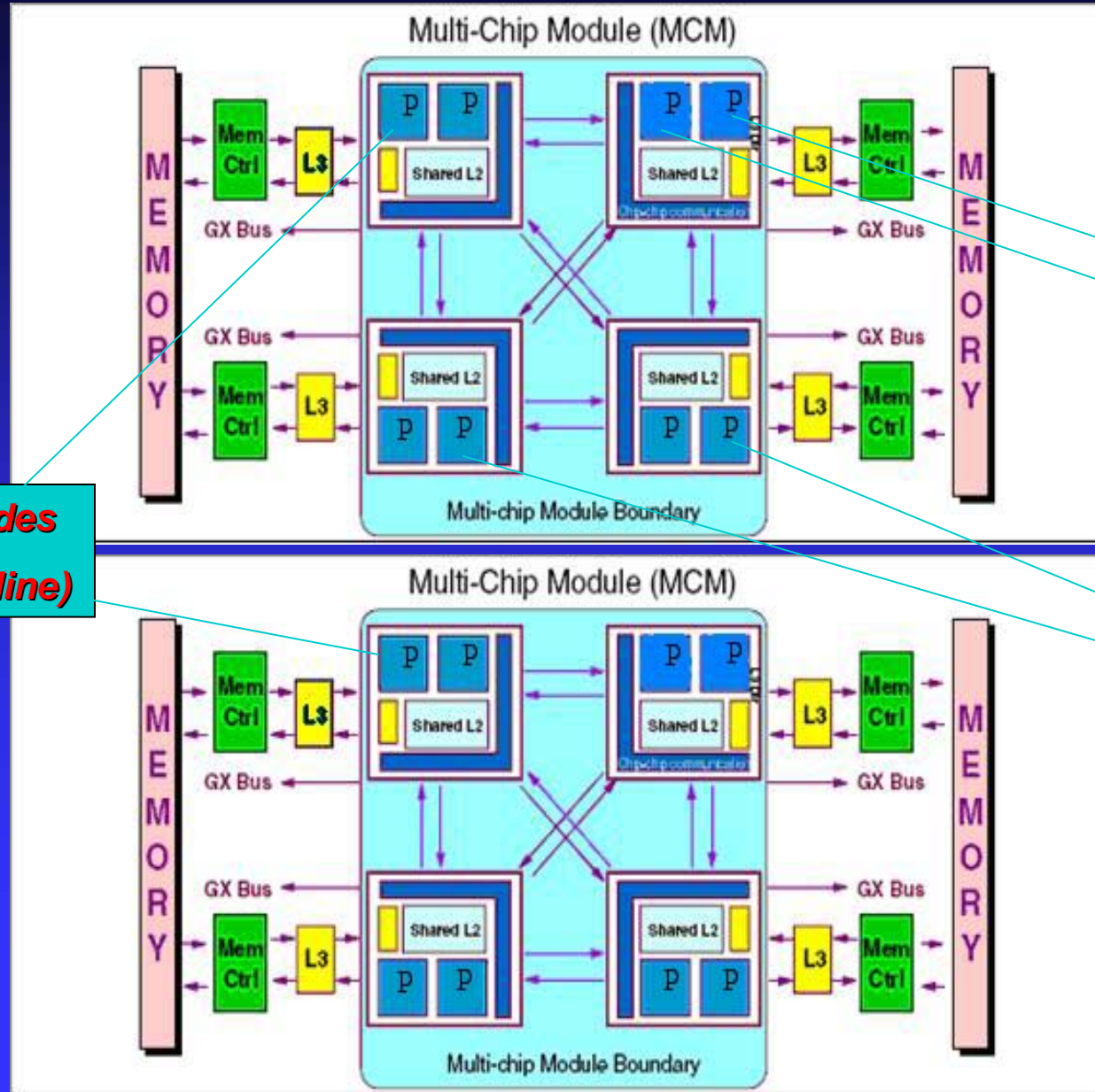


The PIC Steps

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

Stephane Ethier's Talk in 2005 BlueGene Applications Workshop

Performance on 2 Processors on P655



**2 Nodes
(baseline)**

**1 Chip
13.21%**

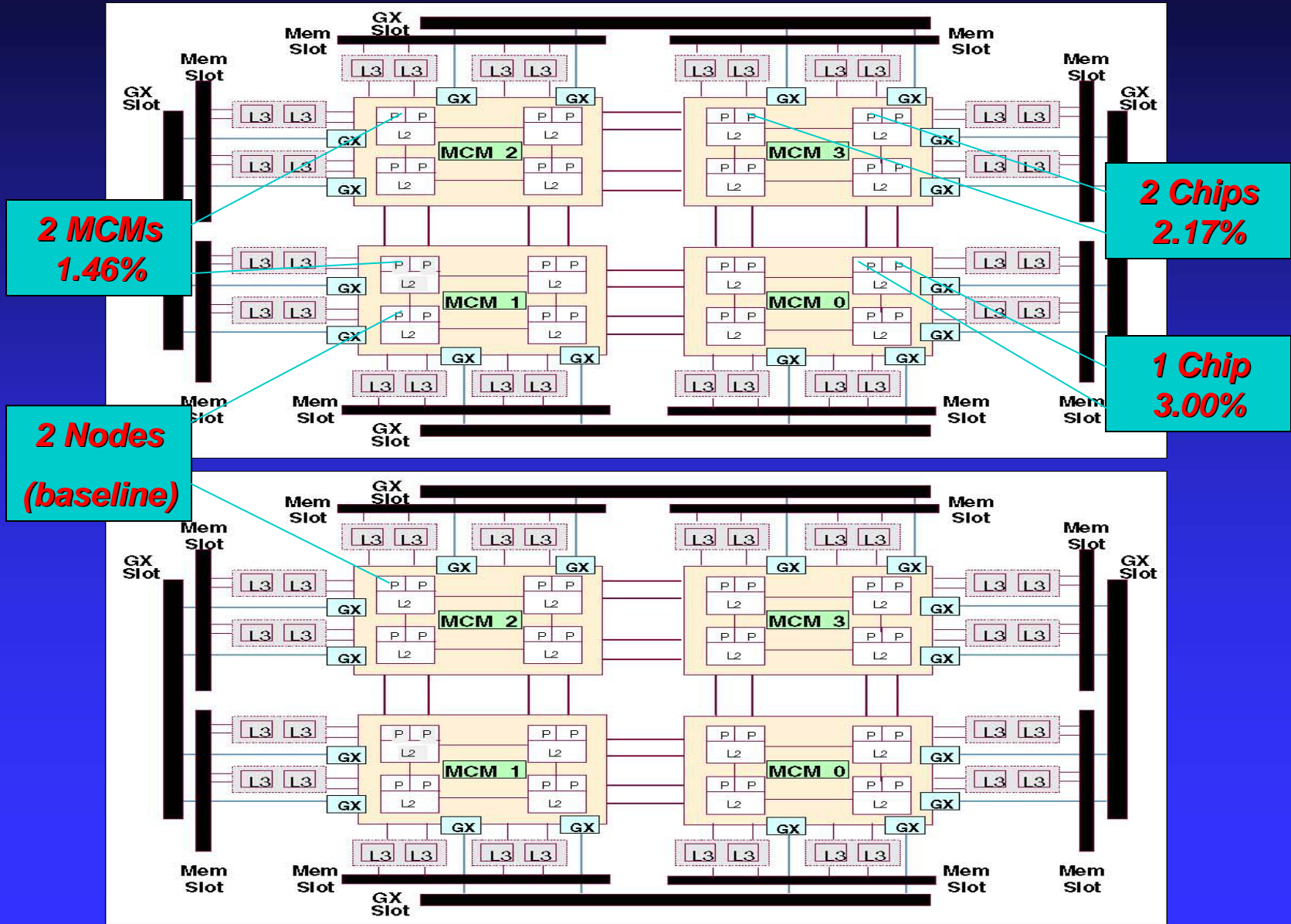
**2 Chips
0.59%**

Performance of GTC on P655

	Across 2 nodes	Across 2 chips	Within a chip
Metrics	2x1/1PPC	1x2/1PPC	1x2/2PPC
Runtime (% difference)	984.26 (baseline)	990.02 (0.59%)	1114.32 (13.21%)
L1 hit rate	92.356%	92.375%	92.386%
TLB miss rate	0.005%	0.005%	0.005%
L2 bandwidth/processor	4618.906MB/s	4582.185MB/s	4071.585MB/s
% accesses from L2	2.228%	2.214%	2.235%

PPC: Processors per Chip

Performance on 2 Processors on P690



Performance of GTC on P690

	Across 2nodes	Across 2MCMs	Across 2 chips	Within a chip
Metrics	2x1/1PPM- 1PPC	1x2/1PPM- 1PPC	1x2/2PPM- 1PPC	1x2/2PPM- 2PPC
Runtime (% difference)	980.23 (baseline)	994.58 (1.46%)	1001.49 (2.17%)	1009.66 (3.00%)
L1 hit rate	92.515%	92.48%	92.457%	92.475%
TLB miss rate	0.005%	0.005%	0.005%	0.005%
L2 bandwidth per processor	4578.505 MB/s	4499.834 MB/s	4491.888 MB/s	4457.119 MB/s
% accesses from L2	2.211%	2.181%	2.177%	2.16%

PPC: Processors per Chip

PPM: Processors per MCM

Performance on 64 Processors

System	64x1	32x2	16x4	8x8	4x16	2x32
P655	1203.32 (baseline)	1253.83 (4.2%)	1266.49 (5.25%)	1305.74 (8.51%)	--	--
P690	--	--	--	--	1210.26 (baseline)	1297.12 (7.18%)

Using Processor Binding on P655

32x2			16x4		
2PPC	1PPC	% difference	2PPC	1PPC	% difference
1253.83	1218.31	2.92	1266.49	1230.47	2.93

PPC: Processors per Chip

Optimization Results of GTC on BlueGene/L

Processors (MxN)	original	optimized	% improvement
8 (8x1)	3804.03	3082.54	18.97%
16 (16x1)	3834.98	3124.80	18.52%
32 (32x1)	3869.60	3166.56	18.17%
64 (64x1)	3937.46	3221.31	18.19%
128 (128x1)	3919.06	3202.81	18.28%
256 (256x1)	3913.07	3208.95	17.99%
512 (512x1)	3820.28	3120.65	18.31%
1024 (1024x1)	3788.49	3096.36	18.27%
2048 (1024x2)	3808.03	3108.40	18.37%

Optimization Results of GTC on P655

Processors (MxN)	original	optimized	% improvement
8 (1x8)	1207.07	1144.73	5.16%
16 (2x8)	1242.56	1174.80	5.45%
32 (4x8)	1273.75	1203.32	5.53%
64 (8x8)	1305.73	1240.52	4.99%
128 (16x8)	1263.71	1201.99	4.88%
256 (32x8)	1237.22	1177.27	4.85%
512 (64x8)	1228.28	1172.25	4.56%

Cast Study II: LBM (Barrier: particular kernel)

Lattice Boltzmann Method

- Widely used in simulating fluid flows
- A non-conventional method that approximates the Navier-Stokes equation with a cellular automaton
- The code is developed by Aerospace Engineering Department at TAMU

Execution times (seconds) for different system configuration on 32 processors

System	Problem size	32x1	16x2	8x4	4x8	2x16	1x32
P655	128x128 x128	33.09 (baseline)	33.38 (0.88%)	34.57 (4.47%)	36.41 (10.03%)	--	--
	512x512 x512	2154.94 (baseline)	2224.76 (3.24%)	2361.67 (9.59%)	2601.66 (20.73%)	--	--
P690	128x128 x128	--	--	--	29.15 (baseline)	30.35 (4.12%)	36.01 (23.53%)
	512x512 x512	--	--	--	2198.90 (baseline)	2248.62 (2.26%)	2590.18 (17.79%)
Blue Gene/L	128x128 x128	33.40	33.42	--	--	--	--
	512x512 x512	2252.99	2253.27	--	--	--	--

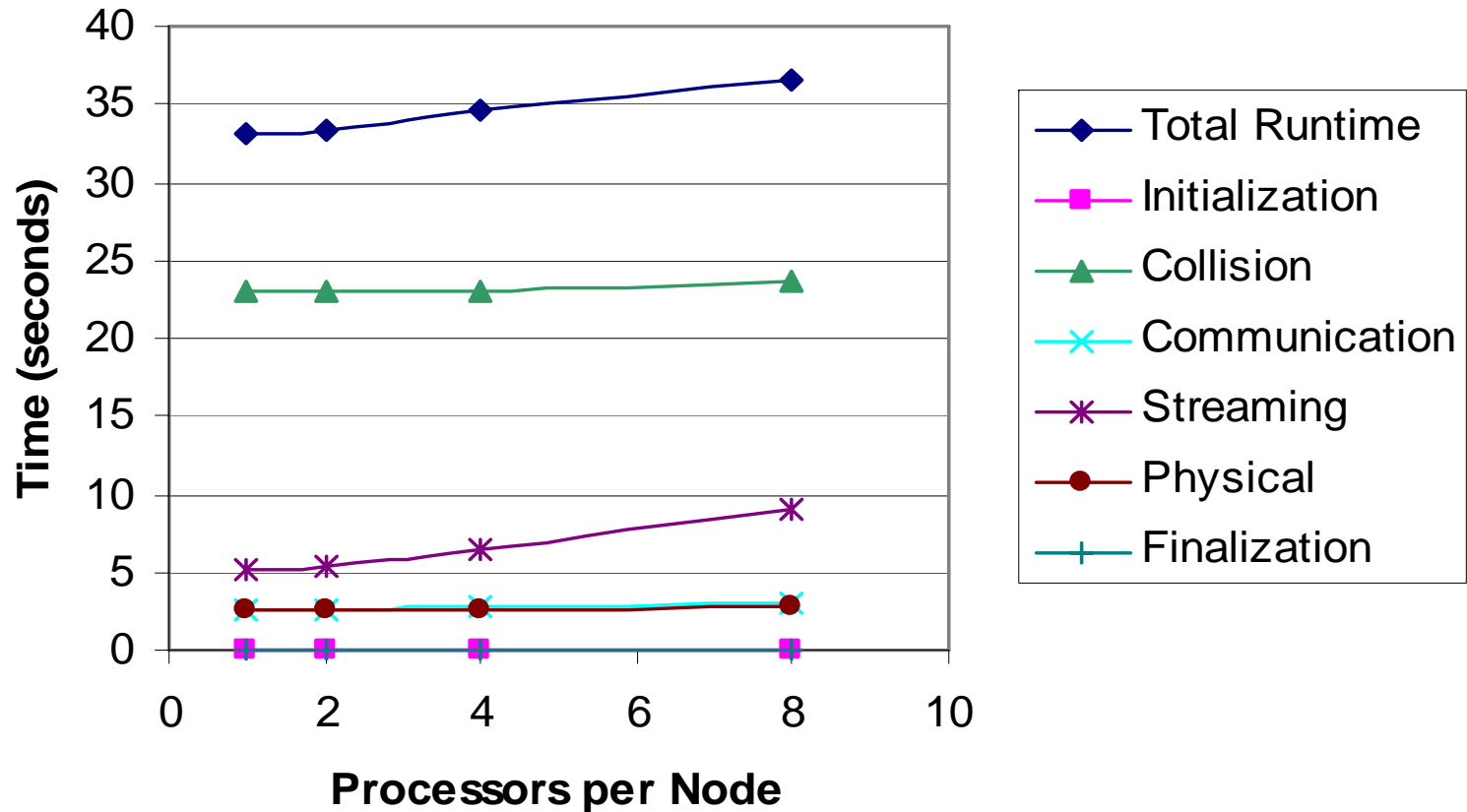
Performance of LBM with 512x512x512 on P655 using processor binding

16x2			8x4		
2PPC	1PPC	% difference	2PPC	1PPC	% difference
2224.76	2124.66	4.71	2361.67	2212.28	6.75

PPC: Processors per Chip

Performance of LBM with 128x128x128 on 32 Processors on P655

Processors Partition on SDSC DataStar p655



Performance Characteristics

- *collision* dominates most of the application runtime
 - ◆ But, its runtime remains almost unchanged.
- *streaming* has the similar performance trend as the application runtime.
- Although optimizing *collision* can improve performance, it doesn't eliminate performance impact caused by processor partitioning.
- In order to efficiently utilize all processors per node, *streaming* should be optimized.

Optimization Results for LBM with 512x512x512 on P655

Processors (MxN)	original	optimized	% improvement
32 (4x8)	2601.66	2191.28	15.77%
64 (8x8)	1306.55	1110.71	14.99%
128 (16x8)	653.31	561.56	14.04%
256 (32x8)	319.71	287.71	10.01%
512 (64x8)	163.44	157.53	3.62%
1024 (128x8)	83.90	79.37	5.40%
2048 (256x8)	48.56	45.30	6.71%

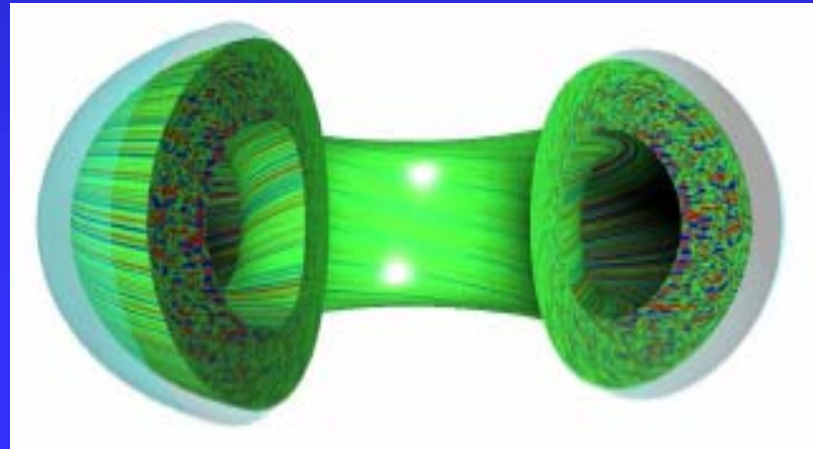
Optimization Results for LBM with 512x512x512 on BlueGene/L

Processors(MxN)	original	optimized	% improvement
32 (16x2)	2253.27	1954.95	13.24%
64 (32x2)	1280.72	1092.11	14.73%
128 (64x2)	699.71	606.83	13.27%
256 (128x2)	356.09	311.35	12.56%
512 (256x2)	153.95	150.73	2.09%

Cast Study III: GYRO (Barrier: global communication)

GYRO

- Physics simulation application
 - ◆ Developed by Fusion Group at General Atomics
 - ◆ Ability to solve gyro-kinetic Maxwell equations
 - ◆ Linear and nonlinear simulations



Execution Time for 32 processors

System Name	Problem size	32x1	16x2	8x4	4x8	2x16	1x32
P655	B1-std	390.14 (baseline)	449.75 (15.28%)	484.13 (24.09%)	554.24 (42.06%)	--	--
	B2-cy	910.09 (baseline)	972.17 (6.82%)	1046.47 (14.99%)	1158.01 (27.24%)	--	--
Blue Gene/L	B1-std	1378.41	1379.20	--	--	--	--
	B2-cy	2983.37	2988.42	--	--	--	--
P690	B1-std	--	--	--	443.64 (baseline)	501.01 (12.93%)	691.15 (55.79%)
	B2-cy	--	--	--	987.61 (baseline)	1146.11 (16.05%)	1489.69 (50.84%)

Performance on P655 using processor binding

16x2			8x4		
2PPC	1PPC	% difference	2PPC	1PPC	% difference
449.75	423.36	5.89	484.13	449.48	7.16

PPC: Processors per Chip

Optimization Results with B1 on P655

Processors (MxN)	original	optimized	% improvement
16 (2x8)	1130.29	1024.75	9.33%
32 (4x8)	554.24	510.93	7.81%
64 (8x8)	276.78	260.58	6.21%
128 (16x8)	146.91	136.79	6.88%
256 (32x8)	84.81	81.05	4.43%
512 (64x8)	61.46	58.939	4.10%

Optimization for B1 on BlueGene/L

Processors (MxN)	original	optimized	% improvement
16 (8x2)	2746.85	2409.21	12.29%
32 (16x2)	1379.20	1210.66	12.22%
64 (32x2)	761.45	697.09	8.45%
128 (64x2)	410.70	372.69	9.25%
256 (128x2)	235.09	214.00	8.97%
512 (256x2)	161.02	150.01	6.83%
1024 (512x2)	121.38	114.37	5.77%

Summary

- **Performance Analysis and Optimization Methods**
- **Case Studies: GTC, LBM, and GYRO**
 - ◆ **Performance improvements:**
Up to 18.97% for GTC, 15.77% for LBM,
and 12.29% for GYRO
- **This work will be published in ICPP2008 Proceedings this September.**