

April 19, 2006



# **BG/L Status and Update**

The 3<sup>rd</sup> BG/L Systems Software and Applications Workshop

Nobuyuki Koizumi IBM Engineering Technology Services

© 2006 IBM Corporation

		- AL		
				-
-	1000			_
	_			
			1000	
			-	
				_

# What is driving the need for more compute cycles?

Science Area	Breakthrough Target Requiring 100's of Teraflops
Fusion	Simulate hydrogen plasma torus (ITER reactor) operating at over 100 million to produce 500 MW of fusion power
Climate	Perform full ocean / atmosphere climate model with 0.125 degree spacing and an ensemble of 8-10 runs
Nanoscience	Simulate nanostructures with 100's to 1000's of atoms
Combustion	Simulate laboratory scale flames with high fidelity
Astrophysics	Simulate explosion of a supernova with a full 3D model

Source: "Towards Petascale Computing for Science," Horst Simon, Lawrence Berkeley National Laboratory, ICCSE 2005 Istanbul, June 2005



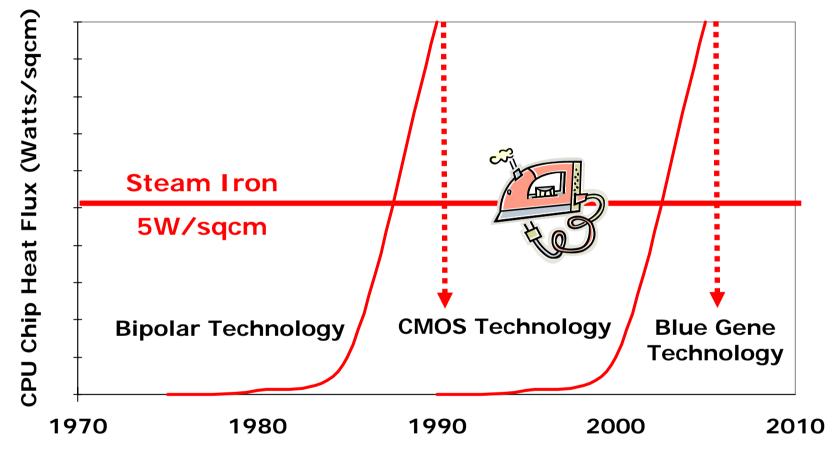
# What is driving the need for more compute cycles?

Commercial Area	Breakthrough Target Requiring 100's of TeraFLOP/S
Petroleum	Run PSTM and PSDM routines for large-scale seismic surveys in < <sup>1</sup> / <sub>2</sub> day
Finance	Perform Monte Carlo simulations using > 500M variables for risk analysis, options pricing, derivatives hedging
Life Sciences	Conduct real-time simulations of the brain's neocortical column using a cellular-level model
Engineering	Perform CFD simulation of full airframe using <b>100's more</b> mesh points than required to simulate an airplane wing
Supply Chain Management	Model optimized deployment of 1000's of part numbers across 100's of parts depots with turnaround time < 1 hr

#### April 19, 2006

		- 10 A		
-	1000			_
	_			
			Including Street	
			-	
				_

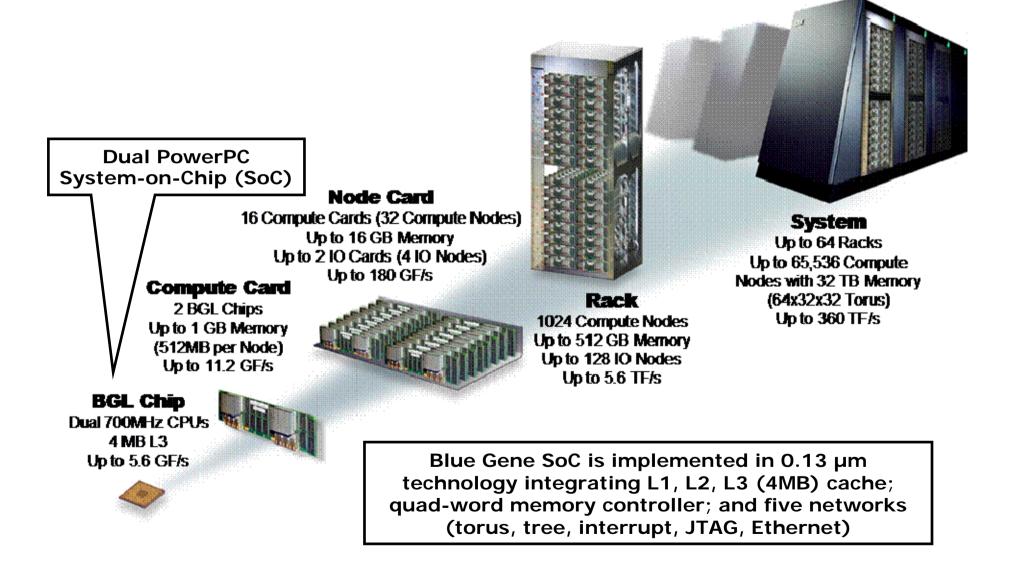
# What is the roadblock to more compute cycles?



Blue Gene is an evolutionary, innovative technology which reduces "time to solution" for many computational science problems through ultrascalability and modularity with the lowest power consumption, smallest footprint, highest reliability and easiest manageability in the industry.

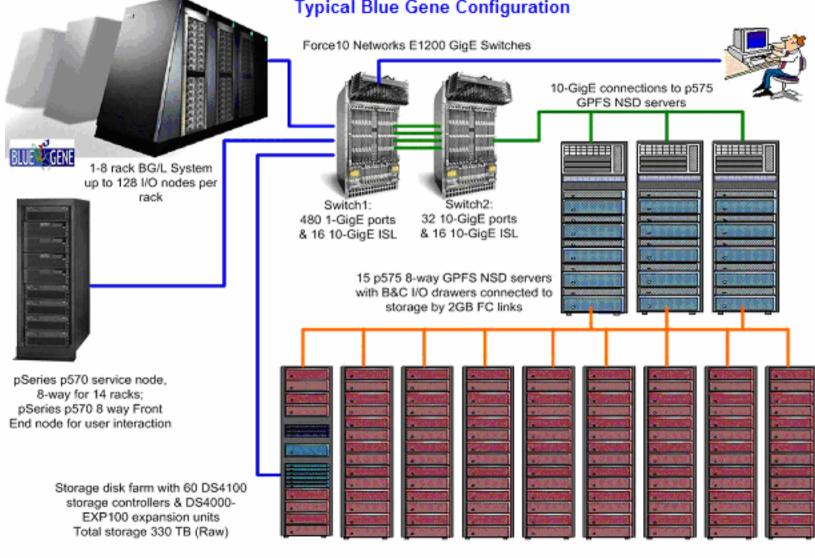
	- AL		
-		Internal Volume 1	
	100		

#### What is Blue Gene?



April 19, 2006





Typical Blue Gene Configuration

10GigE connections 2GB FC connections 1GB Ethernet (Single / Multiple connections)



Blue Gene balances massive scale-out capacity while preserving familiar user/administrator environments



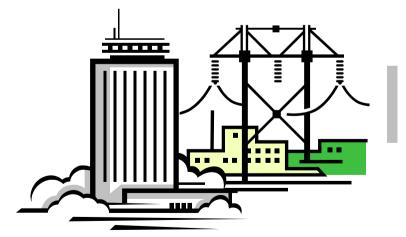
- > Scalability to > 65K processors
- Transparent high-speed, low latency networking
- Standards-based Message Passing Interface (MPI)
- > Familiar, standard programming environment
  - Linux based development environment (SuSE SLES9)
  - Blue Gene's Compute Node Kernel (CNK) provides  $\checkmark$ **POSIX** system calls with restrictions to ensure scaling
  - ✓ Automatic SIMD (Single-Instruction Multiple-Data) FPU exploitation enabled by Fortran, C, and C++ compilers



#### **Energy and space savings with Blue Gene**

"Commodity microprocessors and Linux cluster architectures offer the potential for unprecedented levels of hardware price/performance. They also, however, dramatically increase data center space, power and cooling requirements. Costs for these may more than offset the advantages of inexpensive hardware, and may create reliability and manageability problems that impair system effectiveness. It is from this perspective that the potential role of the IBM Blue Gene system should be viewed."

Business Case for IBM eServer Blue Gene Systems: Reducing the Costs of Commercial Supercomputing, International Technology Group, Nov. 2005



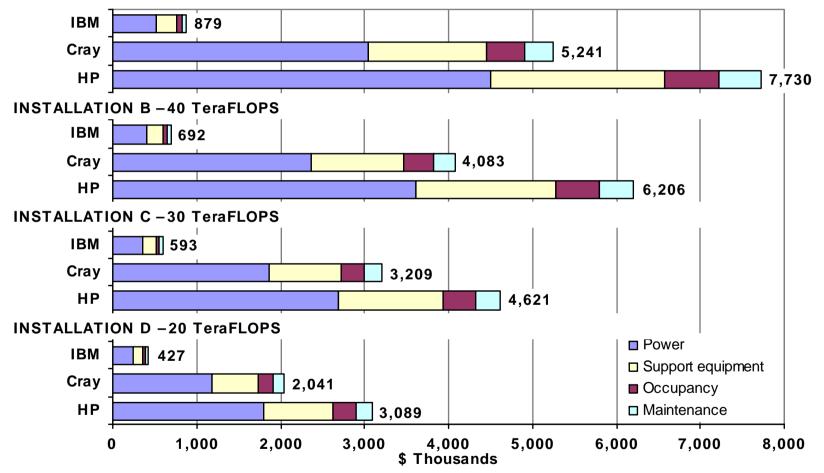
Blue Gene requires 75-80% less power and space than COTS cluster





#### **Comparative 3-year infrastructure costs** (existing facilities)

#### **INSTALLATION A – 50 TeraFLOPS**

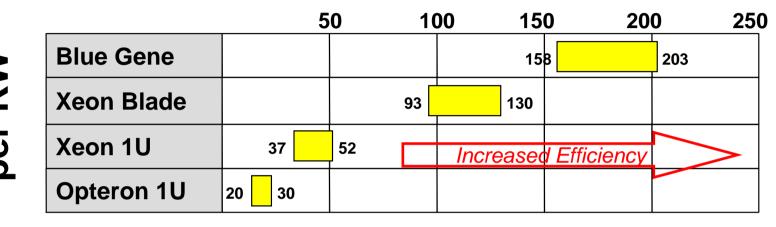


Business Case for IBM eServer Blue Gene Systems: Reducing the Costs of Commercial Supercomputing, International Technology Group, November 2005

	 	£
_		
	1000	
	-	
_	 	_

# **Blue Gene power and space efficiency**

Linpack and Peak GF per KW



Linpack and Peak GF per Rack

	1	<u>250</u>	25	<u>500 37</u>	<u>750 50</u>	000 6	<u>5250</u>
Blue Gene					4420	570	D
Xeon Blade	410 10	75					
Xeon 1U	330 514			Increase	d Efficiency		
Opteron 1U	265 369						
	Xeon Blade Xeon 1U	Blue GeneXeon Blade41010Xeon 1U330	Blue GeneXeon Blade4101075Xeon 1U330	Blue Gene   Xeon Blade   410   1075   Xeon 1U   330   514	Blue Gene   Xeon Blade   410   1075     Xeon 1U   330   514   Increased	Blue Gene 4420   Xeon Blade 410   1075 1075   Xeon 1U 330 514 Increased Efficiency	Blue Gene 4420 5700   Xeon Blade 410 1075   Xeon 1U 330 514   Increased Efficiency

Linpack GF

Peak GF

			_	
100		-		
			1000	
_	_	-	-	
_		_		_

# **Blue Gene award-winning performance**

- 3 in Top10 (#1 and #2)
- 7 in Top50 (11-50)
- 9 in Top100 (51-100)
- 19 in Top500

www.top500.org

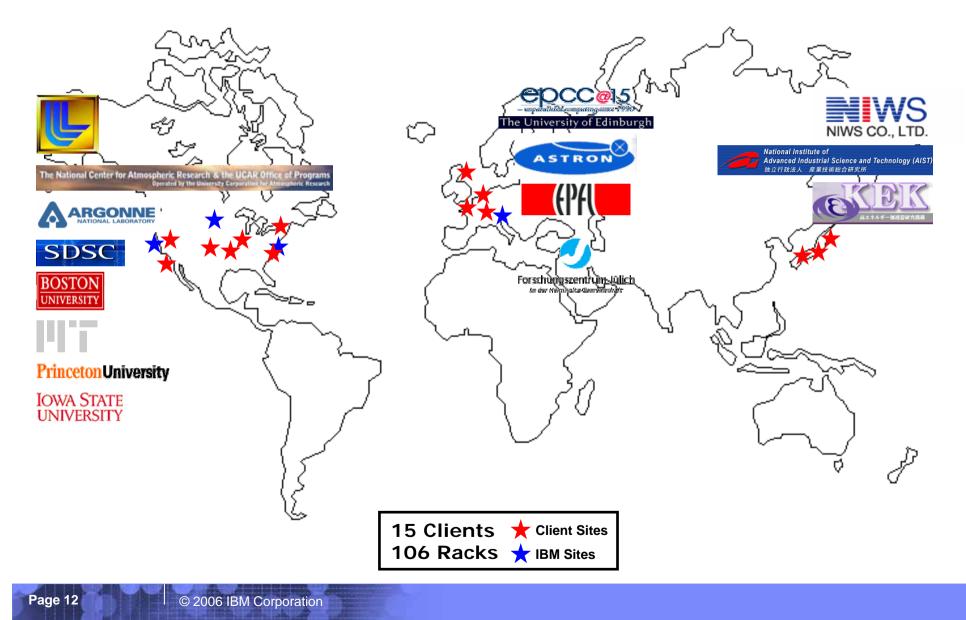
SUPERCOMPUTER S

TOP

HPC Challenge Benchmarks	IBM Blue Gene 65,536 nodes 64 racks	Cray XT3 (Opteron) 5200 nodes 56 racks
HPL (TFLOP/s) – Linpack TPP benchmark which measures floating point rate of execution for solving linear system of equations	259.21	20.53
<u>RANDOMACCESS</u> (GUP/s) – measures rate of integer random updates of memory	35.46	0.69 (7.69 on Cray X1E)
FETE (GFLOP/s) – measures floating point rate of execution of double precision complex one-dimensional Discrete Fourier Transform (DFT)	2311.09	905.57
<u>STREAM</u> (GB/s) – simple synthetic that measures sustainable memory bandwidth and corresponding computation rate for simple vector kernel	160,064	26,021



#### Where in the world is Blue Gene?





# How are they using Blue Gene?

#### Weapons Research

Lawrence Livermore National Laboratory in California has the cream of the Blue Gene crop: a 64-rack machine ... used for nuclear-weapons research.

#### **Radioastronomy**

The Netherlands' Low Frequency Array radio telescope uses a six-rack Blue Gene machine based at the University of Groningen ... to analyse data from 15,000 antennas spanning an area 350 km in diameter.

#### **Protein Folding**

Researchers at the University of Edinburgh, UK, are using a single-rack Blue Gene machine, capable of 6 trillion operations per second, to simulate protein folding and fluid mixing.

#### **Climate Research**

A single-rack Blue Gene is being shared by the National Center for Atmospheric Research and the University of Colorado, which runs simulations of ocean, weather and climate behaviour.

#### **Cosmology**

A team at the San Diego Supercomputing Center is using a single-rack Blue Gene to run Enzo — the centre's software that simulates how galaxies evolved from the Big Bang. The simulation, which has already predicted how the first stars formed, produced around 30,000 gigabytes of data on a typical recent run.

#### **Drug Development**

Japan's National Institute of Advanced Industrial Science and Technology is using its four-rack device to boost drug development. Researchers there hope to shed light on how drugs interact with their targets in the body.

Source: "Virtual Big Bangs and Digital Mushroom Clouds," Nature Magazine, July 2005



# "Killer Apps" on Blue Gene

Discipline	Application
Condensed Matter	CPMD, ddCMD, LSMS, ParaDis
Molecular Mechanics (ab initio)	Amber, CPMD, Qbox, abinit, CASTEP
Classical Molecular Dynamics	ddcMD, NAMD, GRASP, LAMMPS, MDCASK, NAMD, SPaSM, DL_POLY
BioInformatics/Life Science	Smith-Waterman, BLAST, mpiBLAST, pNeo
Neurosciences	Blue Brain (EPFL), SPLIT (KTH)
Astrophysics	FLASH, Enzo, Capreole
High Energy Physics	QCD
Weather / Climate	WRF, HOMME, POP
Plasma Physics	GTC, NIMROD
Computational Fluid Dynamics	AVBP, Miranda, NEK5000, Overflow, FUN3D, Raptor, LUDWIG
Nuclear Physics	QMC



# **Blue Gene Consortium**

#### > Purpose

A community of Blue Gene users who share experiences, help understand and develop the application space, assist in application porting and tuning and provide feedback on functional requirements for next generation

# > Membership Criteria

Membership is open to may be from any geography and any industry or organization, committed to provide a skill or undertaking such as performance measurements, application porting, OS/compiler testing

#### Consortium Facilitator

Argonne National Lab (ANL) http://www-fp.mcs.anl.gov/bgconsortium/default.htm



#### **BG/L Consortium Members**

#### LABORATORIES:

Ames National Lab/Iowa State University Argonne National Laboratory Brookhaven National Laboratory Fermi National Laboratory Jefferson Laboratory Lawrence Berkeley National Laboratory Lawrence Livermore National Lab Oak Ridge National Laboratory Pacific Northwest National Laboratory Princeton Plasma Physics Laboratory

#### UNIVERSITIES:

Boston University California Institute of Technology Columbia University Cornell University DePaul University Harvard University Illinois Institute of Technology Indiana University Iowa State University Louisiana State University Massachusetts Institute of Technology

#### UNIVERSITIES (cont'd):

National Cntr for Atmospheric Research New York University - Courant Institute Northern Illinois University Northwestern University Ohio State University Pennsylvania State University Pittsburgh Super Computing Center Princeton University Purdue University **Rutgers University** Stony Brook University **Texas A&M University** University of California – Irvine University of California - San Francisco University of California - SDSC University of Chicago University of Colorado - JILA University of Delaware University of Hawaii University of Illinois Urbana-Champaign University of Minnesota University of North Carolina

#### UNIVERSITIES (cont'd):

University of S California - ISI University of Texas at Austin - TACC University of Utah University of Wisconsin INDUSTRY: Engineered Intelligence Corporation Gene Network Sciences IBM **INTERNATIONAL:** Allied Engineering Corporation AIST ASTRON/LOFAR CERT EPFL John von Neumann Inst. for Computing National University of Ireland/ICHEC NIWS Co., Ltd. Trinity College, Trinity Centre for HPCC University of Edinburgh, EPCC University of Tokyo



# **Blue Gene Collateral**

#### **Brochure Specifications Sheet** IBM Redbooks 1037 IBM @server' Blue Gene Solution IBM @server\* Blue Gene Solution IBM The IBM9 atturning Blue Genetit Solution is the result of an IBM Bike Gene sustomers note that time-to-solution for many Unfolding the IBM project began over five years ago, deducted to building a new family of supercomputers optimized for bandwidth, scalability and the the long sets at he antellizations has been substant and number two supercent the TOMOS INF alongwill artists in the top 10, there helded supercomputer is in the Tomostane of Execution by an order of magnitude. Scientists can make a new runs mure often allowing them to *eserver* Blue Gene ability to handle large amounts of signore alternative models and approaches to problems. Blue Gene is making a demonstrately the Department of Broggy Nuclear Security Administ data while consuming a fraction of the power and floor space required by today's fastest systems. Solution Notice Society Administration Landow Disense Hadmin Mitmany (UNI), The My on the Server years of UNA, the administration administration attage, The OOI anterior on a shange in the way science can for done Today Blue Gene ranks as the Understand the Blue Gene architecture number one fastest supercomputer on the TOPSco list, and holds 5 of the top 10 positions. The world's And Blue Game delivers more than ultrascalable performance Because of unique design input anguility industry mpoper Tertifull-gherverill file lated Tokin is bereat of an BM spe-Highlights Open Distances Select suitable applications for Candening performance in a grane carring power advance package for the rescriber and by notig performance comparing applications development of the maximum order to explore designs for the effective computers. Solution to be required in a million region-comparing project legan meer free pairs ago diadocatorito balding a sale landy of region responses optimized free toroenethy, socializity and the acting to handle large semantic of data write fastest supercomputer is installed solves that allow decay · Leadership performance in a Leadership performance in a space warms, power-efficient package for the most demanding HPC Administration's Lawrence applications Laboratory. packaging of processors, memory and interconnects, Blue Gene offers leadership Learn about our es in porting Building Specific Adjusts experiences in parallel applica bit has been that prime has been obtained breaks. It is also adverted breaks of origin beings pro-ative breaksabling of proefficiency in the areas of pow and foor space companytion converting a fraction of the preser and Darge Involutions for advance-Nor space required by today's right Besign innovations for advancement. Many university and powerment INVESTIGATION AND AND AN ADDRESS OF ADDRESS AND ADDRES Available in configurations of acience across an important set of research labs have also deployed vital workloads Bibe Gene for computational ranging from one to 64 racks, Bias Gene III the innovative re solution from BM that delivers an ubreacelable solution witho fedwaring accounting to a with other, had only who may studies in measure renearch, radio anga of soletion, no worker and developers attronomy, protein folding, climate research, cosmology, and drug development. Technology accessible to a wide Antoine Tabary Pascal Vezolle Hari Reddy ii Garlos Sosa Sheeba Prakash Octavian Lascu range of acientists, researchers and sauficing efficiency. ooks ITG STATUS REPORT **Application Notes** Blue Gene's Teraflop Attack BUSINESS CASE FOR IBM eSERVER BLUE GENE SYSTEMS: REDUCING THE COSTS OF COMMERCIAL SUPERCOMPUTING When the last residuant of the TOPTWI regressing starting many patholes have been V and water a semantical publication of PM Marci Gare equations. We addly that the Charles of the semantical transmission that the discussion y 2002 and half the exp are, the mature Harci Gare diployment half half yield of the regress partners, and PM and the Care diployment half half yield of the regress. **Research Papers** Outrik Note: Challenges At the and al 2004, there were freer than 10 appercemptions worksheet exploite of districting users than 20 Treat/2004 at L35-AdX, Safera performance. The current rate of growth occurines, by the end of the detotals a least 10-2004 for the tau an. Technology admices, accompanies by detatable detailses to the cost of processing power, are estimate an admiced and the safera admices and the safera admices admices and the safera admices and safera admices. The safera admices admices admices admices admices admices admices admices models transmission is administed and events of processing admices a exercise systems in soil of task solar Har Gaugaraily. The TOPOR •Porting Applications to the IBM nemer promotion, in stat of multi-molecular there can uncertain. The TOPIOS of the second-gradient is a schedule difference magnitude of only them more obleg intelligation. But it along great maniformality integration of the stream or discontinue are listed. For example, to obly with dominant/ly developed operations -down its one of them or another-spart are maintenen-dementered operations -down its one of them or another-spart are maintened dominant interestingly and <sup>1</sup> eServer Blue Gene/L System for many companies, conventional strategies for leadership through R&D will erode as large-scale spercomputing becomes the norm not only among Portuge 1000 corporations, but also among simalle players seeking to challenge them through aggressive innovation. Competitive pressures will force exclusing supercomputer investments. TOPUT new he designed by designs has the designed in an Solution The dynamics of dis process will vary between industries. Transformations, however, may be rapid, and comparise need to plan for them now. A great deal of work will be required in application conception and development. For this should not be the only frozen. Citized infrastructure challenges are also ported. that Game and its cannot be one. Not Mills Property representation des processing present date ins-representation des processing present des ins-representations des processions and seatures present by public gebeurauch of seature presents togradues in deven, Mar Gauer sales it aufhand segres samparing Massive Parallel Quantum entropyments on an annuments of an only once, contain an announce to annumpt of an opport-"containing" and annumpt of an opported and an opported and an opported and an opported and here is of and ready properties and a line and any annumpt of an opported and and and and and only once minimizing and annumpt in the opported and annumpt of an anguarder materials and may once minimizing and annumpt problem that imput of prime shorteress. In it from this preparation that the postation of the of an IMIM Bind contrast of the strend. Computer Simulator • Early Experience with Scientific Har-Connorder Market Barry Bar Infrastructure Costs This report summarizes preliminary results of research to project infrastructure costs - including costs o data cester floor space, power and support equipment - for commercial supercomputing systems. The focus is on systems in the 20 to 50 TernPLOPG performance range. Applications on the Blue Gene/L namparat WH transfer prostnets in War Catal War a constraint, and y forward in a comparational for large - quoticiting for an activity, due folding interactions of formus proteins. Is provided this as a comparing "general large" in the same with a site 10 Gene War option and the same multi-form ingrin Genery Koopers in 1997. How Gene insuft has an day support compares are served in its down of a shifted or and multi-form and the same days. Three platforms are compared: IBM Blue Gene systems, Hewlett-Packard (HD) Ranium 2-based Cluster Blatform 6000 systems, and Gray Opteron-based XT3 systems. Supercomputer Hentisting data center facilities are employed, three-year information costs for Blae Gene systems range from 5.5.7 million to 5.1.0 million less than for the Context Flatform 6000 configurations, and from 35.5.7 million to 31.0.7 million less than for Cray XT3 configurations delivering comparable performance. Derivate, Inc. Connected & 1000 Composition for well-positing. To not reproduce Scaling Physics and Material at a faile (1997 head-out). Life If new data center facilities must be comparable (infrastructure costs for Blue Gene systems range from 36.56 to 33.61 million less than for comparable HD Claster Flatform 6000, and from 34.60 to 32.57 million less than for comparable Cray XT3 configurations. ind in last 2003 when 1904 years Science Applications on a MILLION lines cannot for completeness on sy or a stream and fairer treats is information million dy by the international Technology Orage. where he hered to Massively Parallel Blue Gene/L System TCO Study

 _		_	
		_	-
_		_	-
-		_	-
 	_	_	_
	_	_	

# What applications are currently running on Blue Gene and what is their performance?



# **Applications Performance on Blue Gene**

- BG is first HPC system to break barrier of 100+ TeraFlop/s sustained performance on real applications (Molecular Dynamics)
  - ddcMD 101.5 TeraFlop/s (7 hrs of Uranium atoms on 64 racks)
  - CPMD 110.4 TeraFlop/s
- Several other applications have achieved two orders of magnitude or more higher performance than previously possible – successful scaling achieved from 1K to 100K processors
- Gordon Bell Prize competition at SC 2005
  - 4 of 6 finalists based on Blue Gene
  - LLNL/IBM team won for "100+ TFlop Solidification Simulations on Blue Gene/L"

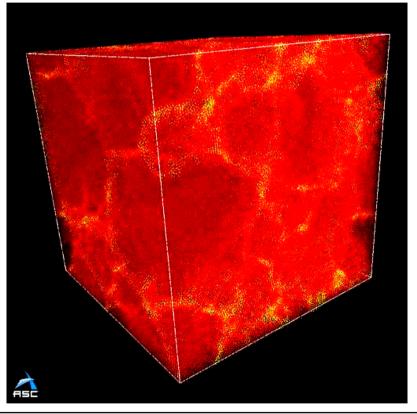


#### ddcMD *Classical MD* 2005 Gordon Bell Prize Winner

 Scalable, general purpose code for performing classical molecular dynamics (MD) simulations using highly accurate MGPT potentials

 MGPT semi-empirical potentials, based on a rigorous expansion of many body terms in the total energy, are needed in to quantitatively investigate dynamic behavior of transitions metals and actinides

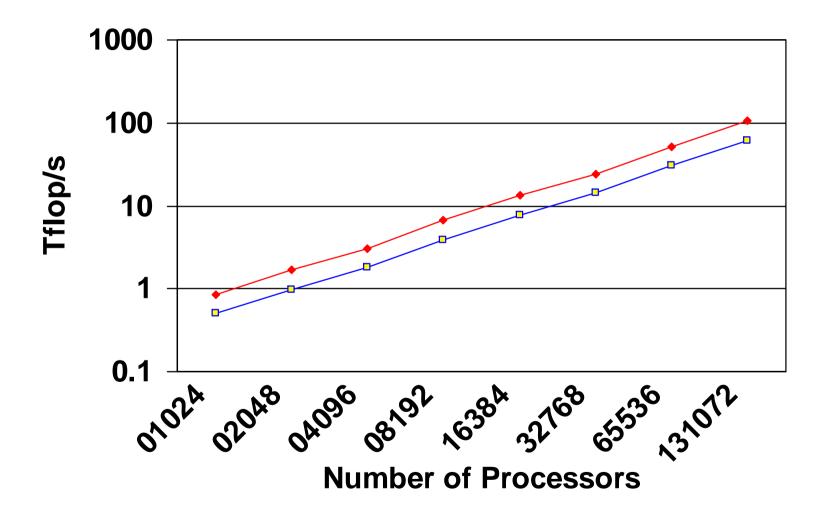
•Visualization of important scientific findings already achieved on BG/L: Molten Ta at 5000K demonstrates solidification during isothermal compression to 250 GPa



524 million atom simulations on 64K nodes are orders of magnitude larger than any previously attempted runs; *superb strong and weak scaling* expected for full machine - ("very impressive machine" says PI)



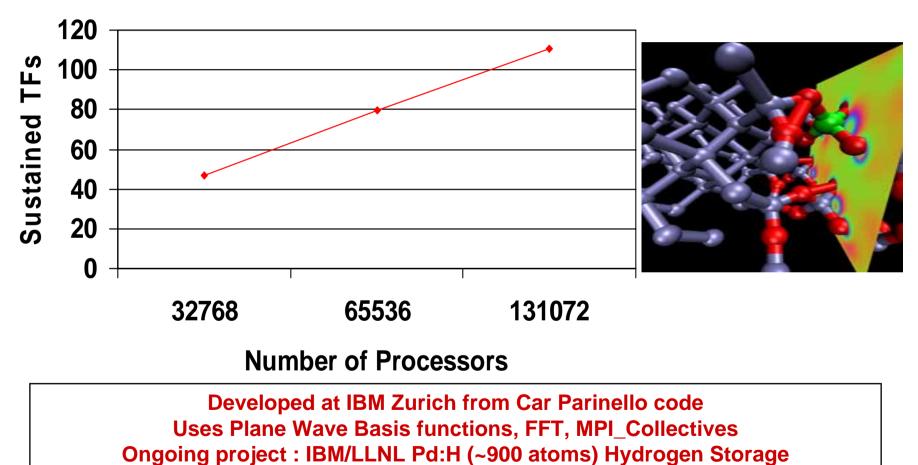
#### Performance of ddcMD on Blue Gene Weak scaling: MGPT Uranium and Tantalum





#### **CPMD**

Alessandro Curioni, Salomon Billeter, Wanda Andreoni CPMD Performance on BG/L



Achieved 110.4 Teraflop/s sustained on 64 racks BG/L (excellent strong scaling)



# Conclusions

Blue Gene represents an innovative way to scale to multi-TF/s

- Leadership performance and price/performance
- ✓ Massive scalability
- Efficient packaging drives low power, cooling and floor space requirements
- Blue Gene balances massive scale-out capacity and preservation of familiar programming environments
- Blue Gene is applicable to a variety of computationally intensive workloads
  - ✓ High-energy physics
  - ✓ Astrophysics
  - Molecular dynamics and modeling
  - Climate modeling
  - Risk management and portfolio optimization
- Blue Gene technology is easily accessible for porting and optimization, including availability in IBM's On Demand Center

#### IBM's ongoing R&D commitment ensures the viability and vitality of the Blue Gene solution



# Read more about it...

#### http://www.ibm.com/servers/deepcomputing/bluegene.html

#### http://www-fp.mcs.anl.gov/bgconsortium/default.htm

#### http://www.research.ibm.com/journal/rd49-23.html

http://www.top500.org