Scalability of GeoFEM on BG/L prototype

Kengo Nakajima
The 21st Century Earth Science COE Program
Department of Earth & Planetary Science
The University of Tokyo. and
Japan Science & Technology Agency (JST)

The 3rd BG/L Systems Software & Applications Workshop
April-19-2006, CBRC/AIST, Tokyo Japan.

This study …

- Parallel FEM benchmarks using IBM BG/L prototype (500MHz) up to 512 PE's in March 2004.

Acknowledgements

- IBM/T.J.Watson Research Center.
  - Dr. David Klepacki
  - Dr. Robert Walkup
- IBM, Japan.
  - Mr. Yasuo Kurita
  - Dr. Hiroki Nakano
- NERSC/Lawrence Berkeley National Laboratory.
  - Dr. Esmond G. Ng
- Earth Simulator Center, Japan.
- GeoFEM Project
- CREST/JST
- CBRC/AIST

- Introduction
  - Features of FEM
  - Ideal Scalable System
- Benchmarks
  - Hardware Environment('s)
  - Software
  - Results

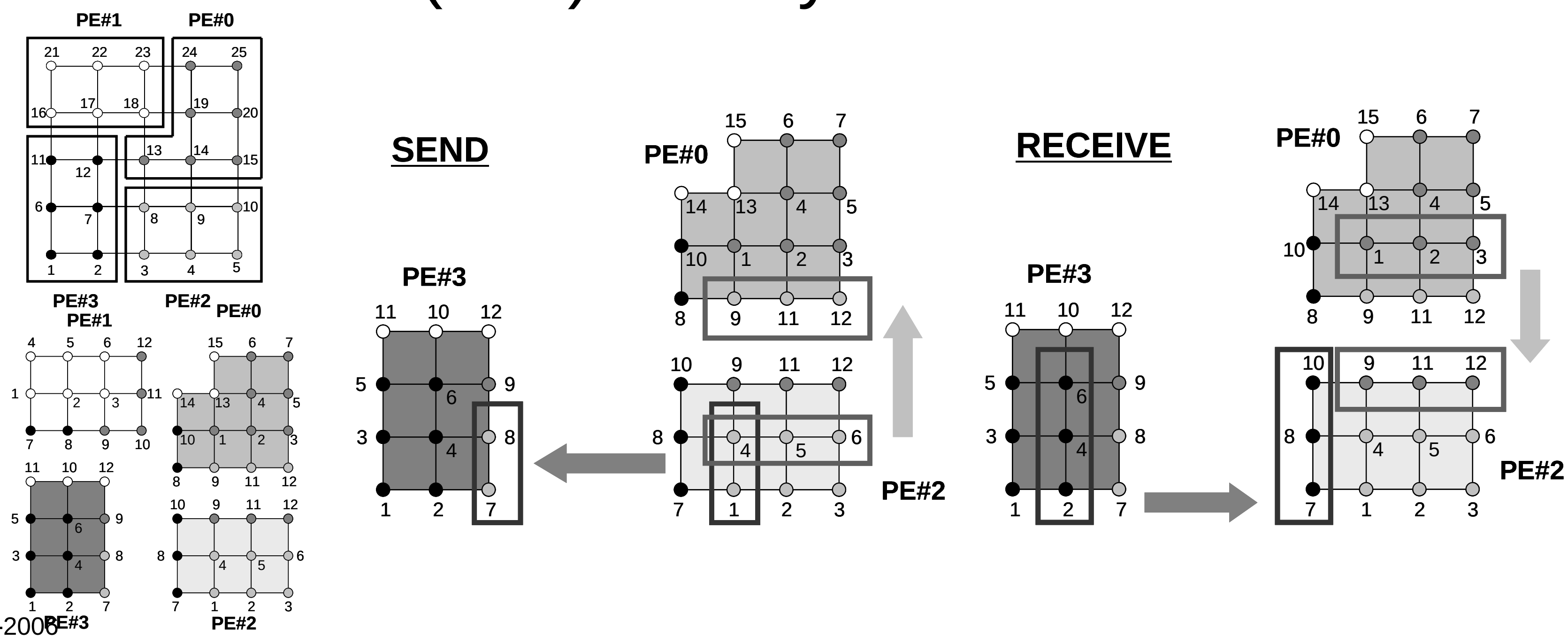
Features of FEM applications

- **HUGE** “indirect” accesses
  - memory intensive
    - element-by-element
    - vertex-by-vertex (node-by-node)

- Local “element-by-element” operations
  - sparse coefficient matrices
  - suitable for parallel computing

Features of FEM applications in parallel computation

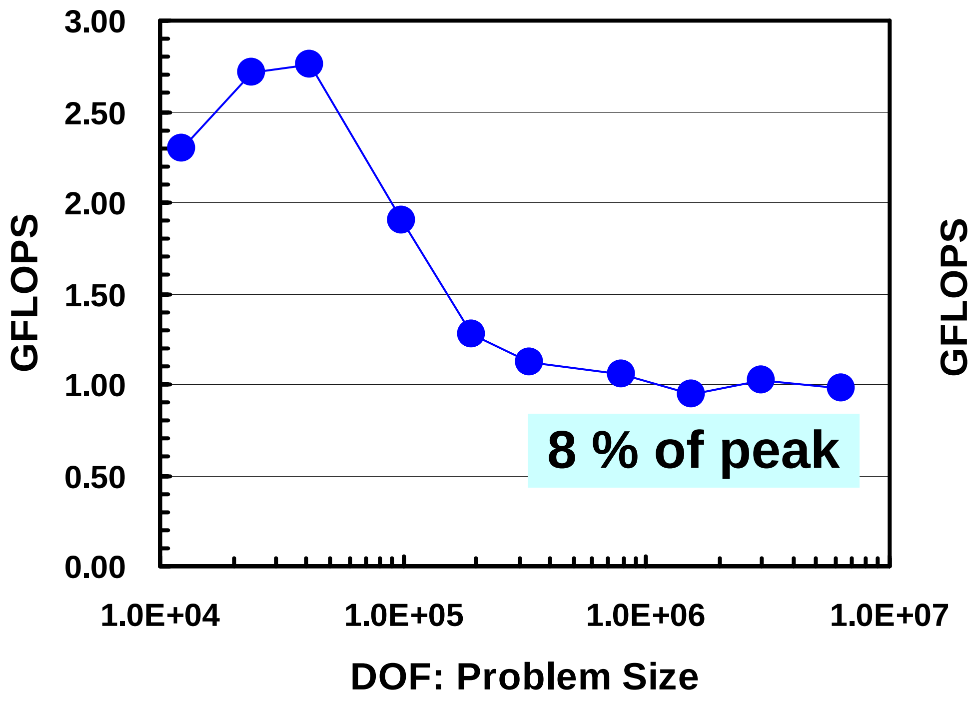
- communications with ONLY neighbors (except "dot products" etc.)
- amount of messages are relatively small because only values on domain-boundary are exchanged.
- communication (MPI) latency is critical

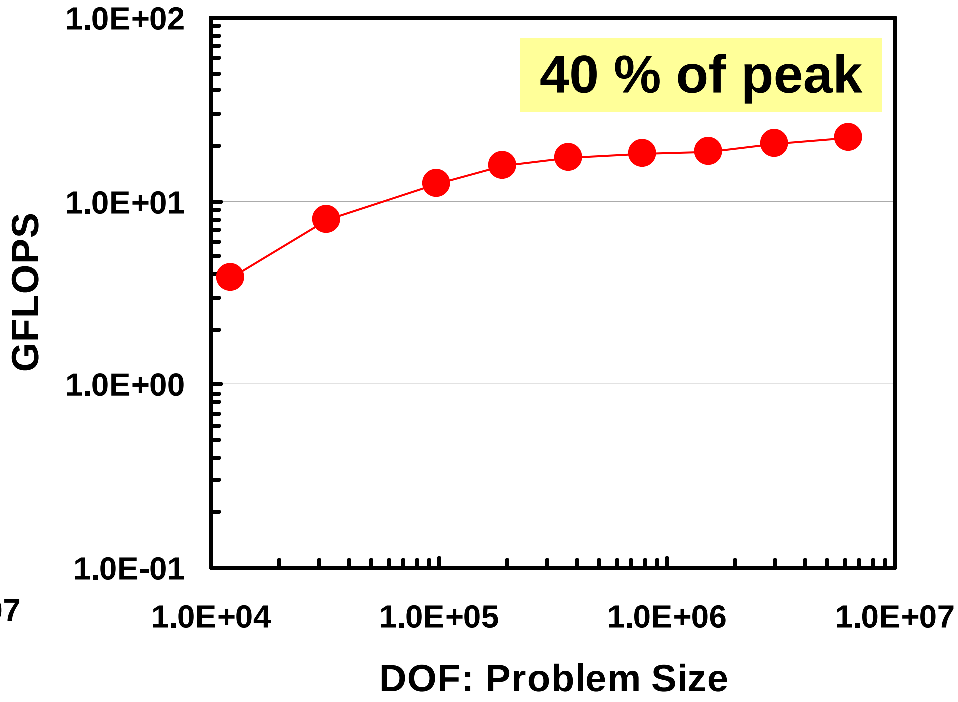


Performance of FEM Applications

- Scalar Processor
  - Big gap between clock rate and memory bandwidth.
  - Very low sustained/peak performance ratio
    - e.g. 5-8 % on IBM Power-3/Power-5
- Vector Processor
  - Very high sustained/peak performance ratio with special tuning
    - e.g. 35-40 % on the Earth Simulator
    - Sufficiently long loops (= large-scale problem size) are required for certain performance

Performance vs. Problem Size

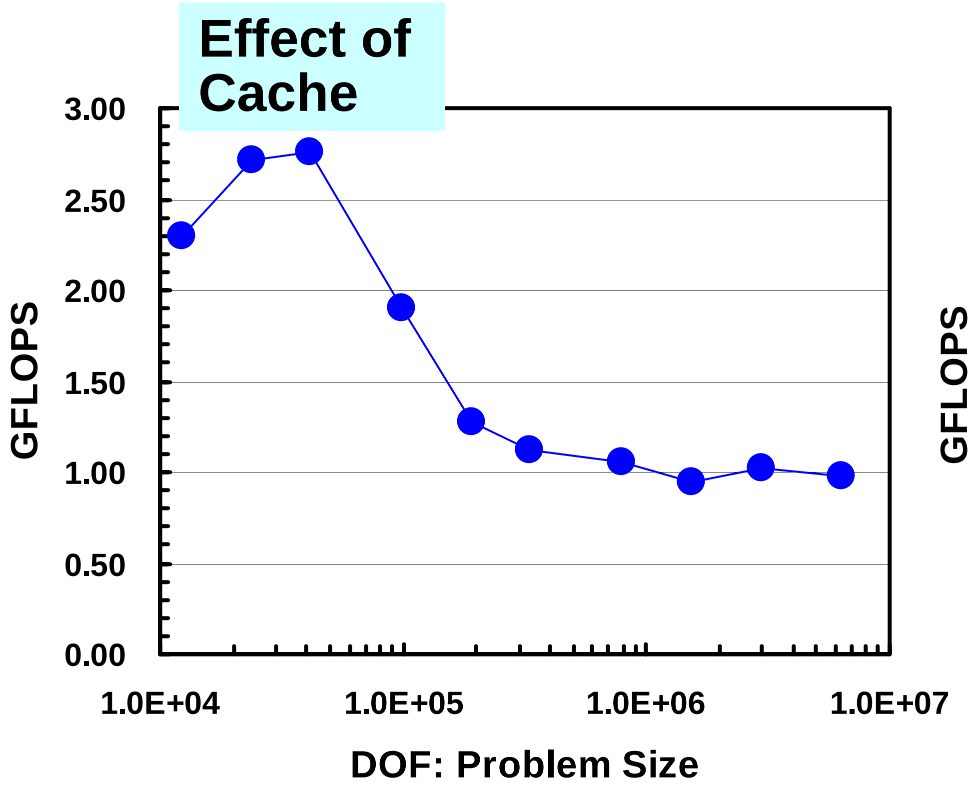


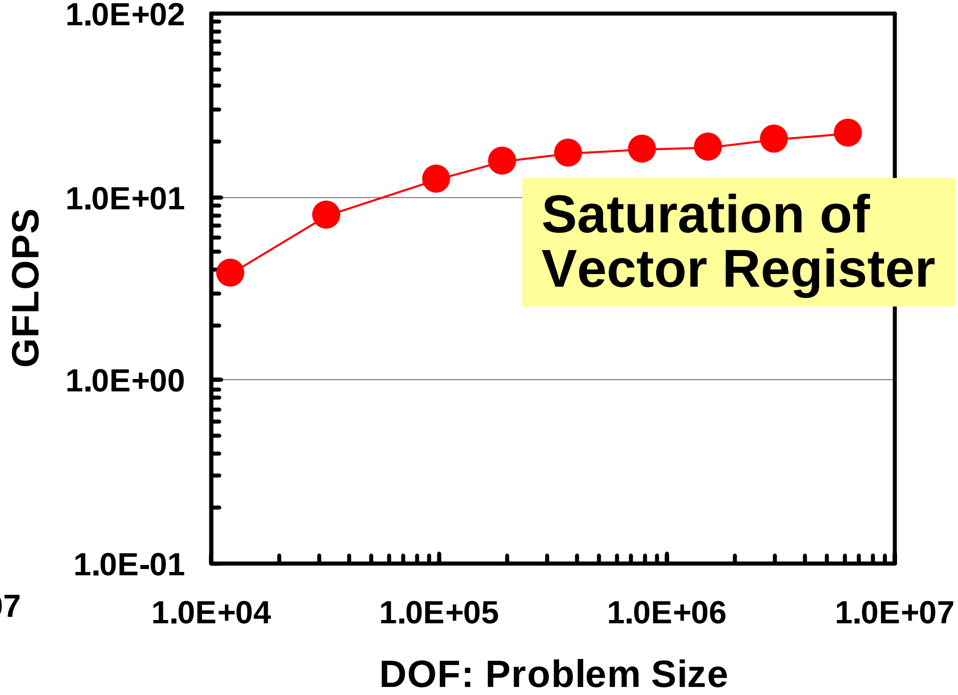


IBM-SP3:
Performance is good for small problems due to cache effect.

Earth Simulator:
Performance is good for large scale problems due to long vector length.

Performance vs. Problem Size

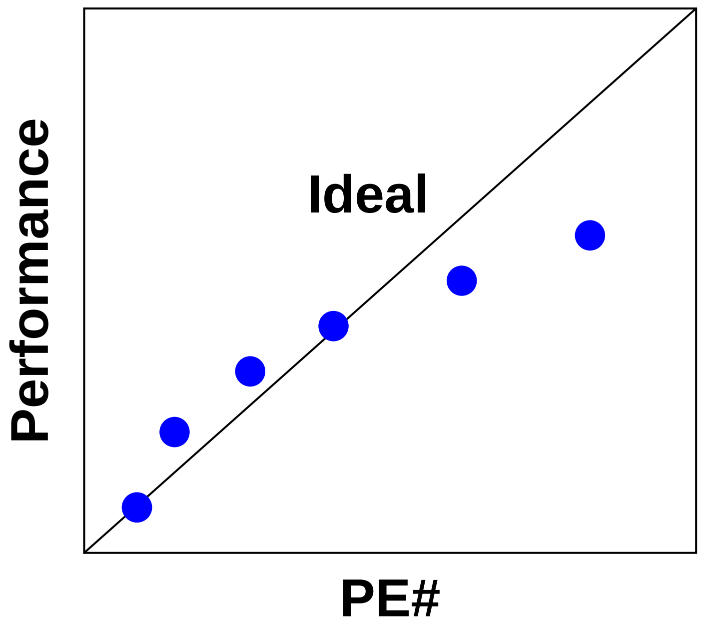


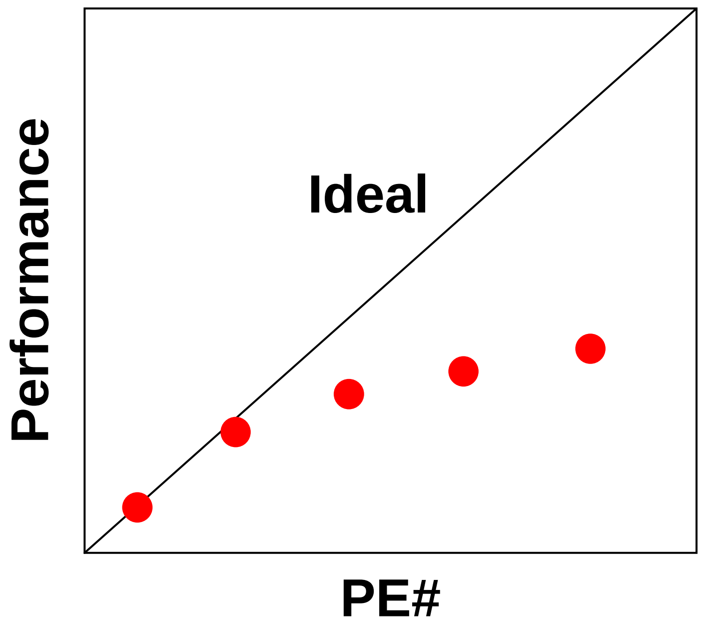


IBM-SP3:
Performance is good for small problems due to cache effect.

Earth Simulator:
Performance is good for large scale problems due to long vector length.

Parallel Computing
Strong Scaling (Fixed Prob. Size)

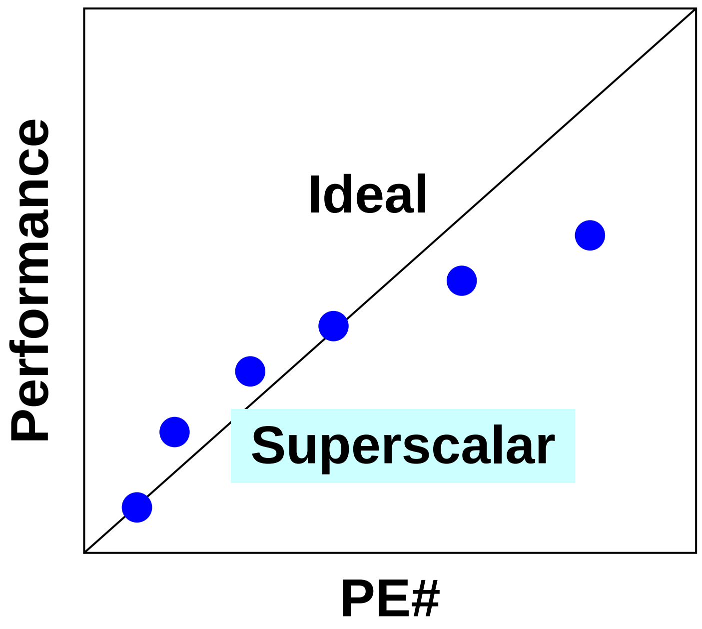


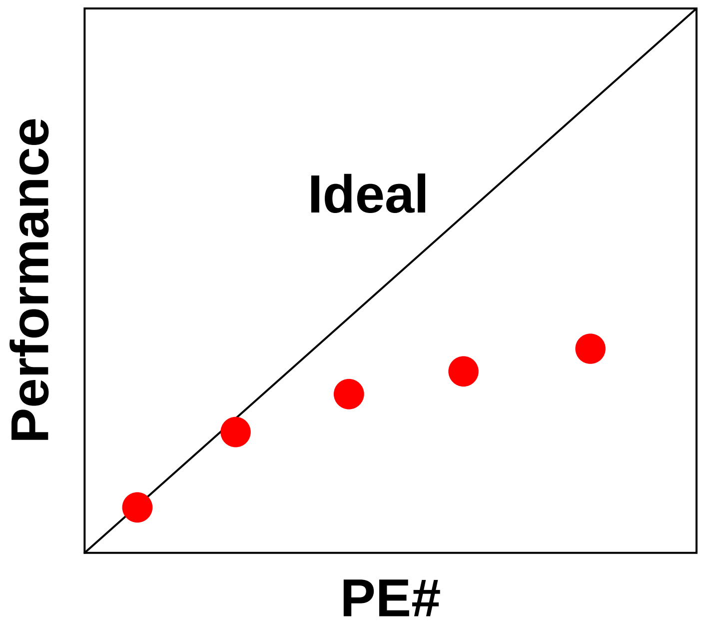


<u>IBM-SP3:</u>
Super-scalar effect for small number of PE's. Performance decreases for many PE's due to comm. overhead.

<u>Earth Simulator:</u>
Performance decreases for many PE's due to comm. overhead and **<u>small vector length</u>**.

Parallel Computing
Strong Scaling (Fixed Prob. Size)

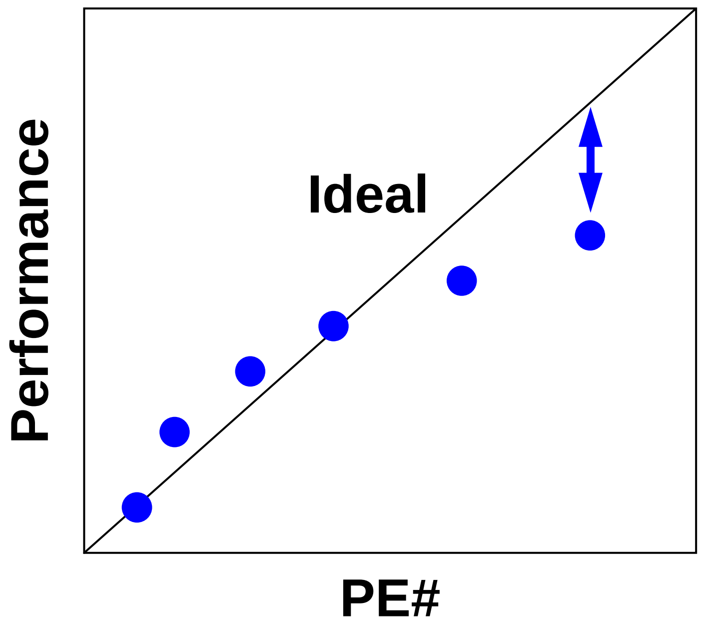


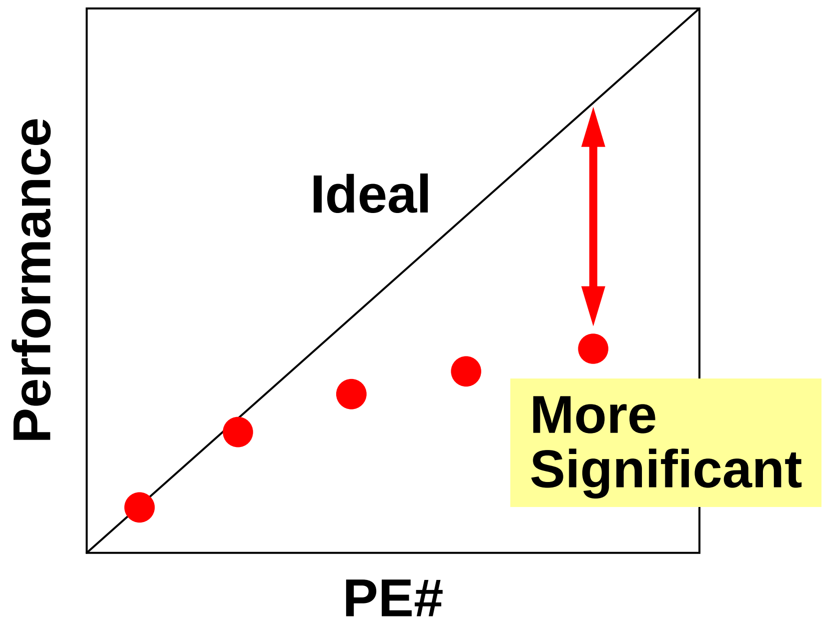


IBM-SP3:
Super-scalar effect for small number of PE's. Performance decreases for many PE's due to comm. overhead.

Earth Simulator:
Performance decreases for many PE's due to comm. overhead and **small vector length**.

Parallel Computing
Strong Scaling (Fixed Prob. Size)

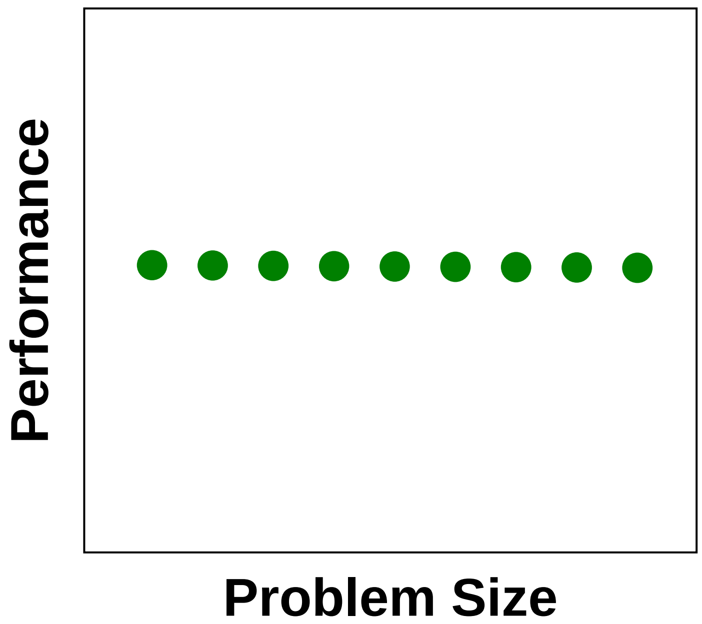


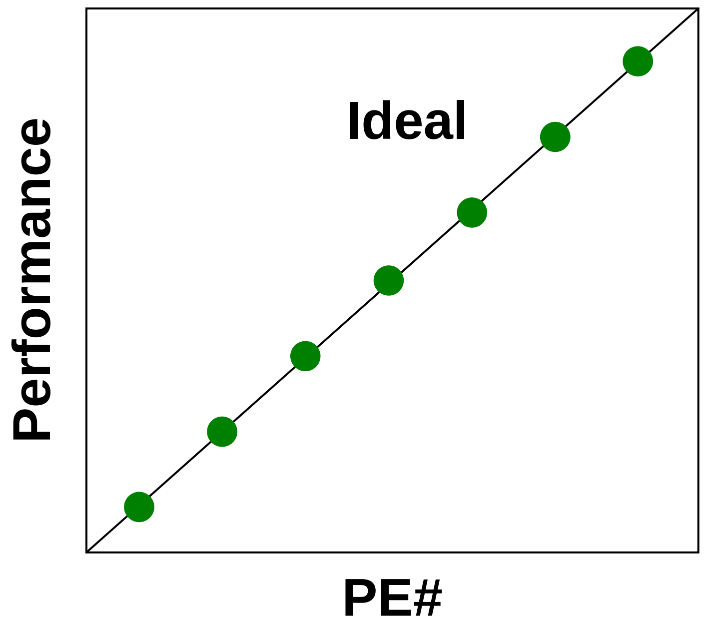


IBM-SP3:
Super-scalar effect for small number of PE's. Performance decreases for many PE's due to comm. overhead.

Earth Simulator:
Performance decreases for many PE's due to comm. overhead and **small vector length**.

What is the Ideal Scalable System ?





Constant (**and high**) sustained performance for a wide range of problem size, and **<u>for a wide range of applications.</u>**

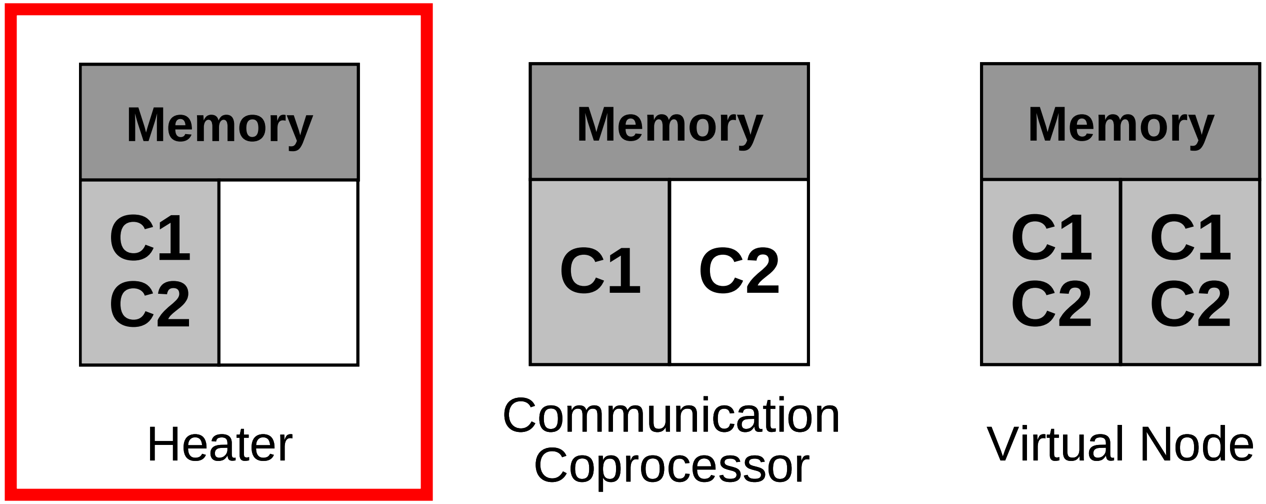
Low communication overhead.

- Introduction
  - Features of FEM
  - Ideal Scalable System
- **Benchmarks**
  - **Hardware Environment(‘s)**
  - **Software**
  - **Results**

Hardware Environment('s)

- 512-node Prototype@500MHz, 256MB/node
  - IBM Rochester, Minnesota
  - with early version of OS & compiler (March 2004)
- 1 PE/node (HEATER mode)
- -qarch=440 (single FPU/PE)
  - 1 GFLOPS for peak performance/PE

Processor Configurations on Each Node.



Heater mode:
one processor run, the other is idle.
Communication coprocessor mode:
one processor is dedicated to communication and the other for general processing.
Virtual node mode:
both CPUs process and communicate.
C1: Computation, **C2**: Communication

Hardware Environment(‘s)

- 512-node Prototype@500MHz, 256MB/node
  - IBM Rochester, Minnesota
  - with early version of OS & compiler (March 2004)
- 1 PE/node (HEATER mode)
- -qarch=440 (single FPU/PE)
  - 1 GFLOPS for peak performance/PE

- Comparison with
  - IBM SP-3
    - “Seaborg” at Lawrence Berkeley National Laboratory, USA.
    - 375MHz, peak performance= 1.5 GLOPS/PE
    - 8 of 16 PE’s on each SMP node used.
  - Earth Simulator

Comparison with IBM SP3 at NERSC/LBNL, and ES

	BG/L	Seaborg at NERSC/LBNL	Earth Simulator
Architecture	IBM Power PC 440	IBM POWER3 Nighthawk 2	NEC SX-6 based
PE#/node	2	16	8
Clock rate	700 MHz	375 MHz	500 MHz
Peak performance/PE	1.40 GFLOPS (primary FPU only)	1.50 GFLOPS	8.00 GFLOPS
Memory/node	512 MB ~ 2 GB	16GB ~ 64 GB	16 GB
L1 Cache/PE (data/instruct)	32/32 KB	64/32 KB	-
L2 Cache	2 KB/PE	8 MB/PE	-
L3 Cache	4 MB/node	-	-
Memory-PE Bandwidth	5.5 GB/sec/node	16 GB/sec/node	256 GB/sec/node
Bidirectional Communication Bandwidth/node	2.1 GB/sec	2.1 GB/sec	12.3 GB/sec
MPI Latency	5.5-8.5μs	16.3μs	5.0-5.6μs

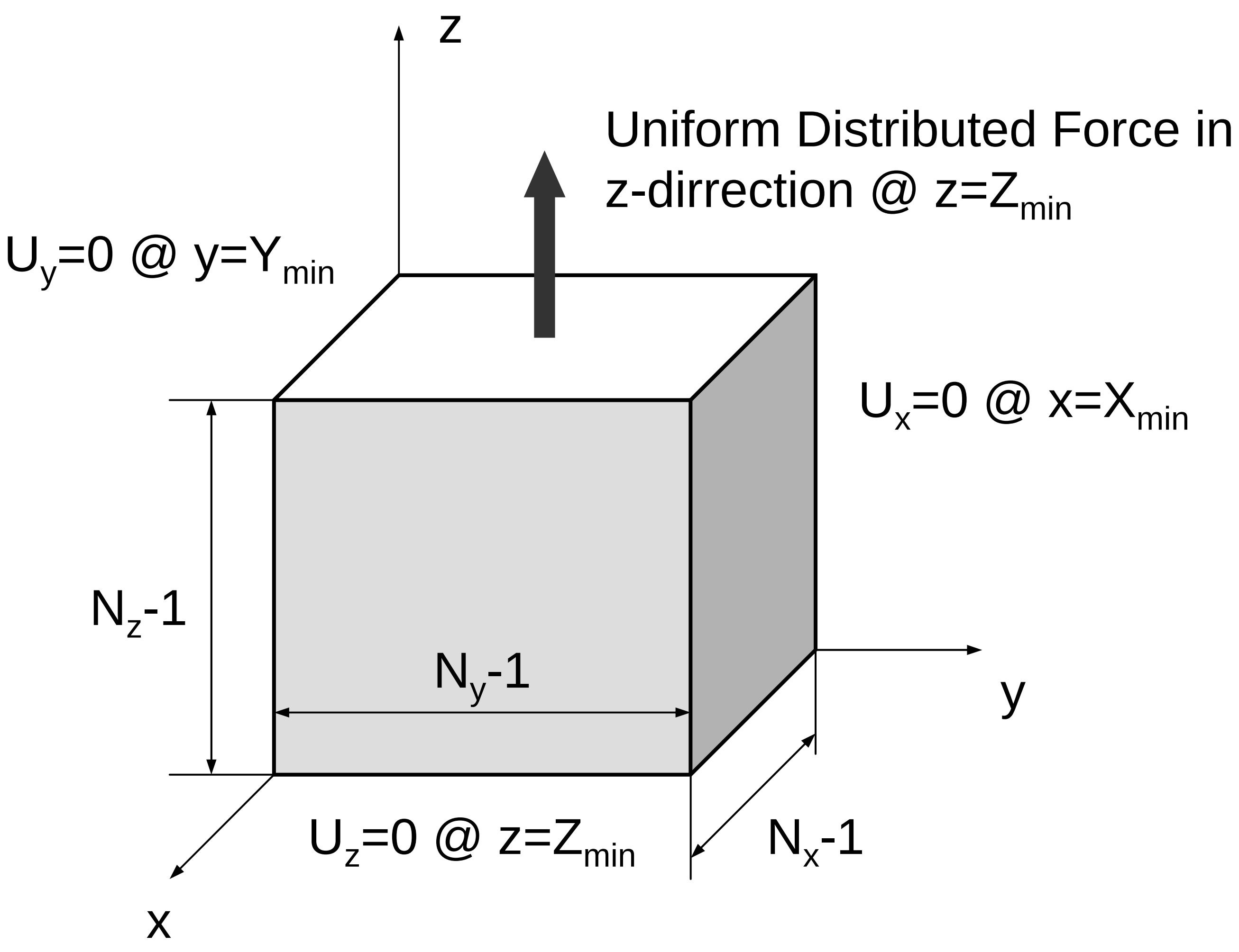
Software

- Parallel preconditioned iterative solvers for unstructured grids.
  - Elastic-linear problems
    - Block ILU(0)
  - Contact problems
    - Selective-blocking (SB) for contact problems
- Codes are originally optimized for the Earth Simulator
- Distributed data structure for parallel FEM procedures developed in ***GeoFEM***
  - **http://geofem.tokyo.rist.or.jp/**
- NO special attention to network topology

Example I: Block ILU(0)

- 3D Linear Elastic Problems
- Parallel Iterative Linear Solver
  - Node-based Local Data Structure
  - Localized Block ILU(0) Preconditioning (Block Jacobi)
  - Additive Schwartz Domain Decomposition（ASDD）
  - Reordering for Vector/Parallel Performance
  - http://geofem.tokyo.rist.or.jp/

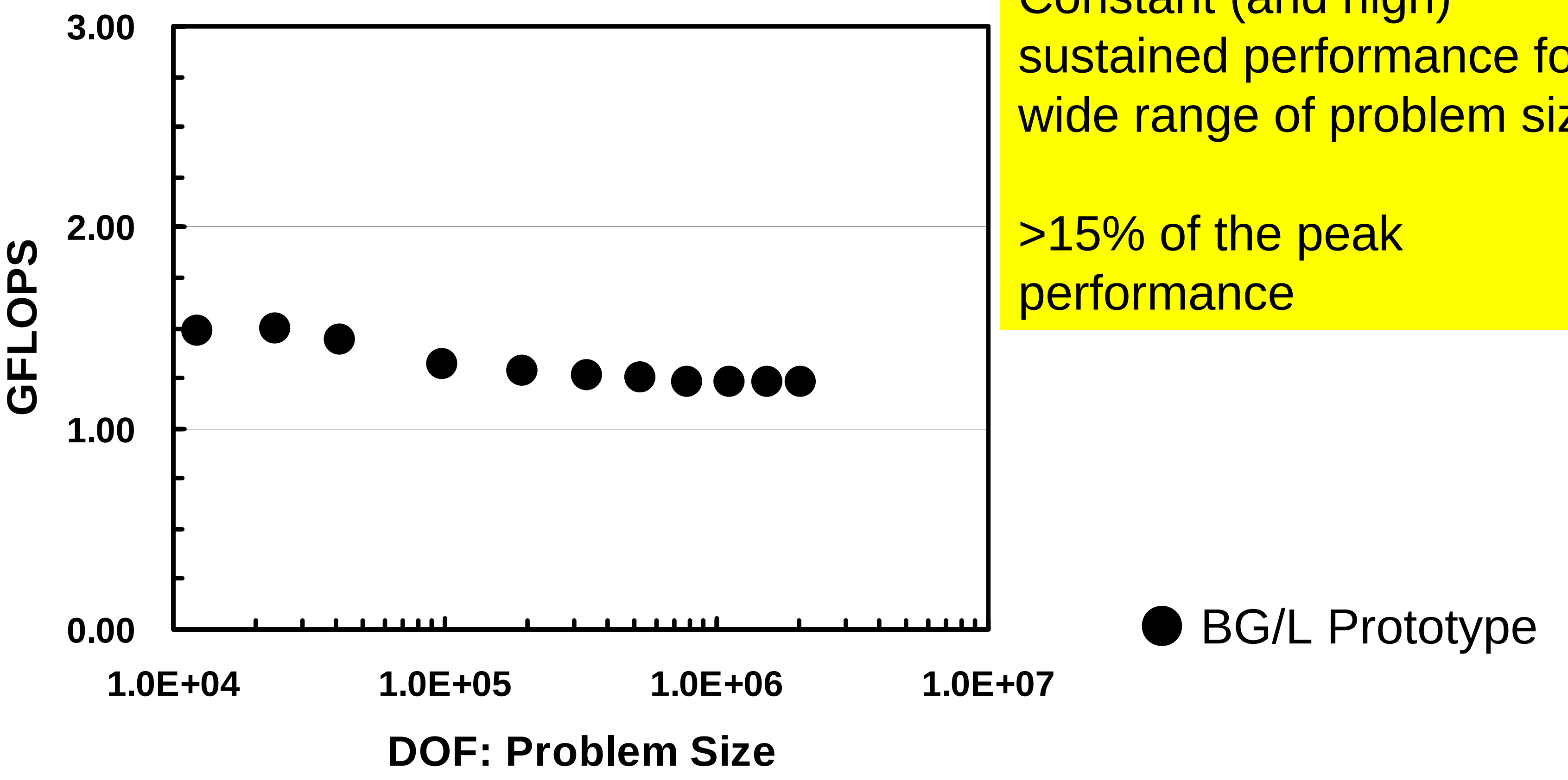
Simple 3D Cubic Model



Effect of Problem Size

Results on 8 nodes (1PE/node)

BG/L 512-n prototype@500MHz

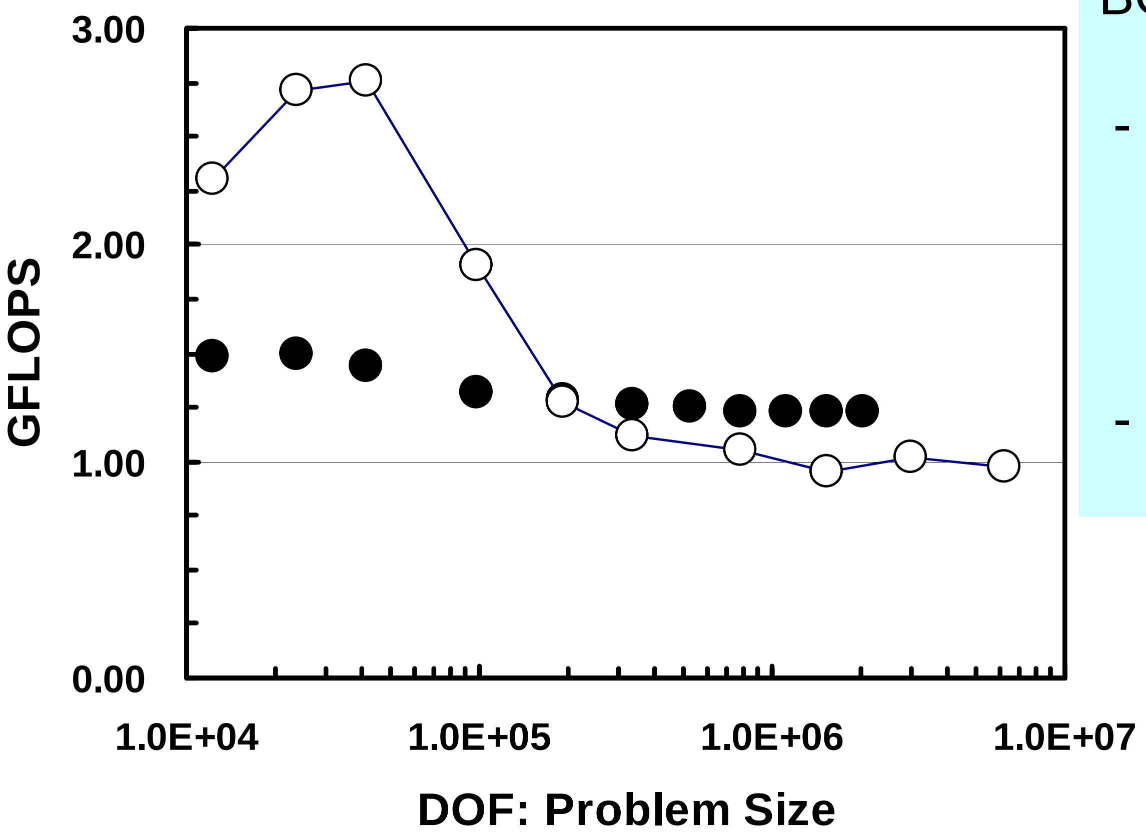


Constant (and high) sustained performance for a wide range of problem size

>15% of the peak performance

Effect of Problem Size

Results on 8 nodes (1PE/node)



BG/L Prototype

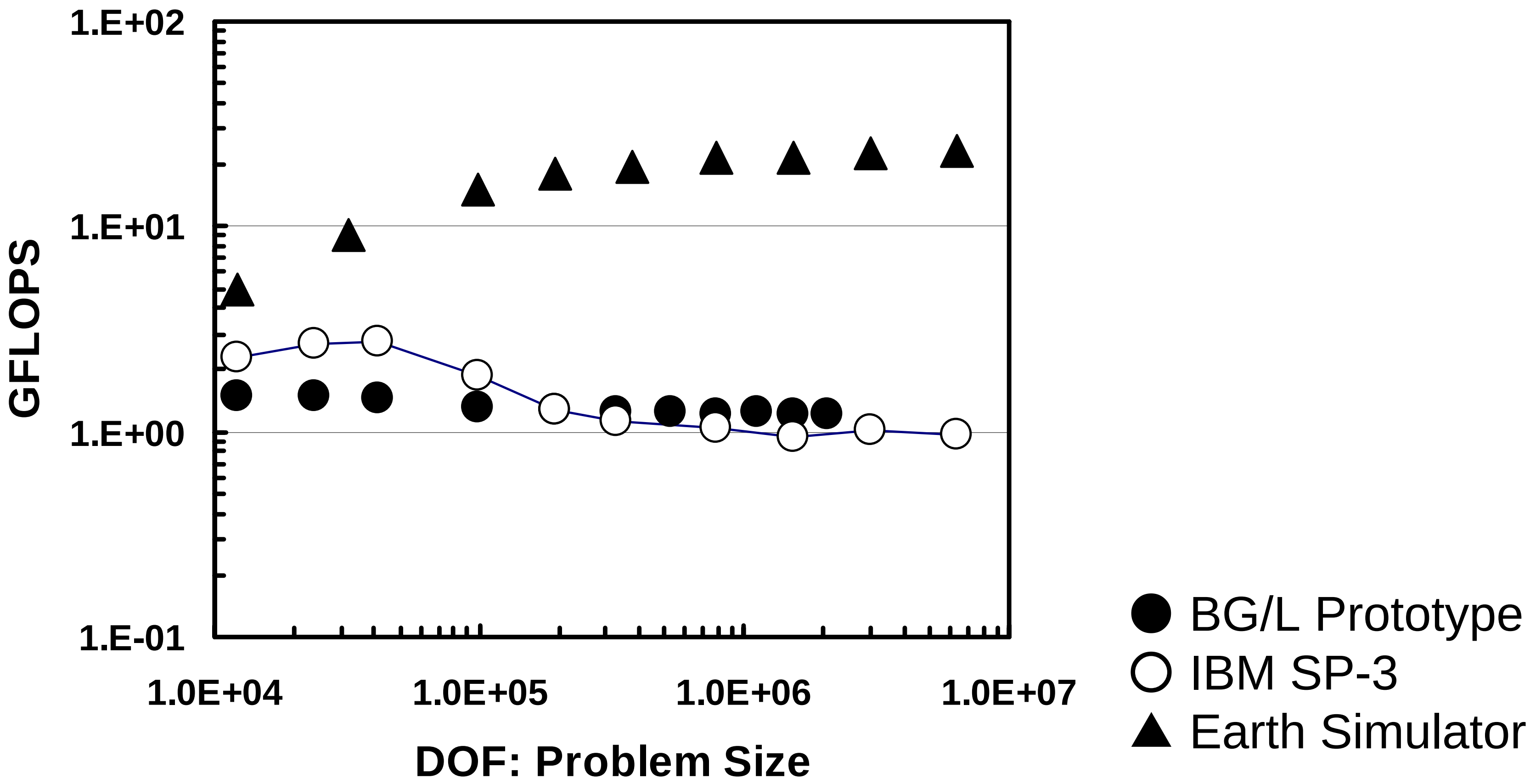
- relatively smaller
  peak performance (1.5:1.0)
  cache size (4:1)
  memory latency (2:1)

- relatively larger
  memory bandwidth (1:2)

● BG/L Prototype
○ IBM SP-3

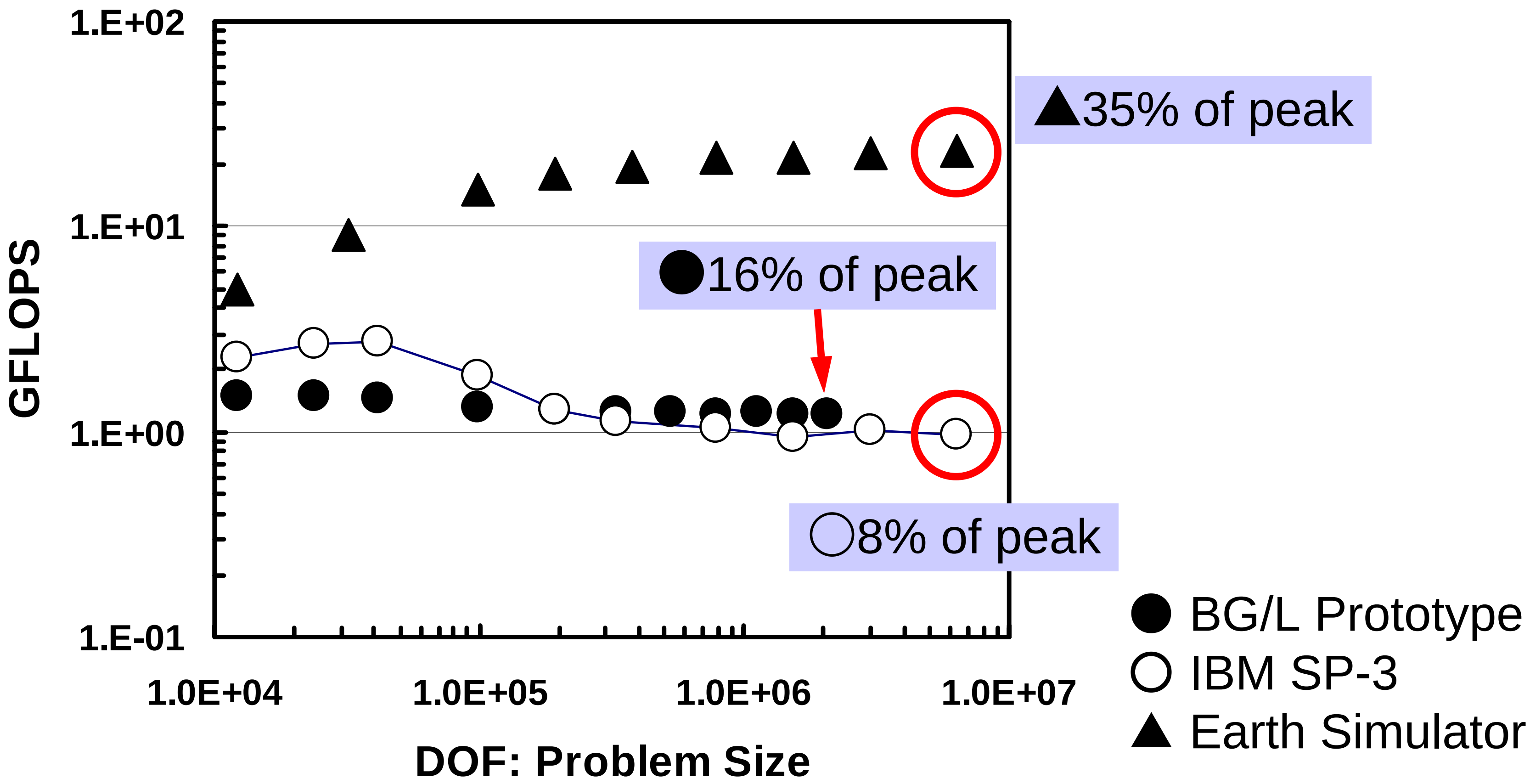
Effect of Problem Size

Results on 8 nodes (1PE/node)



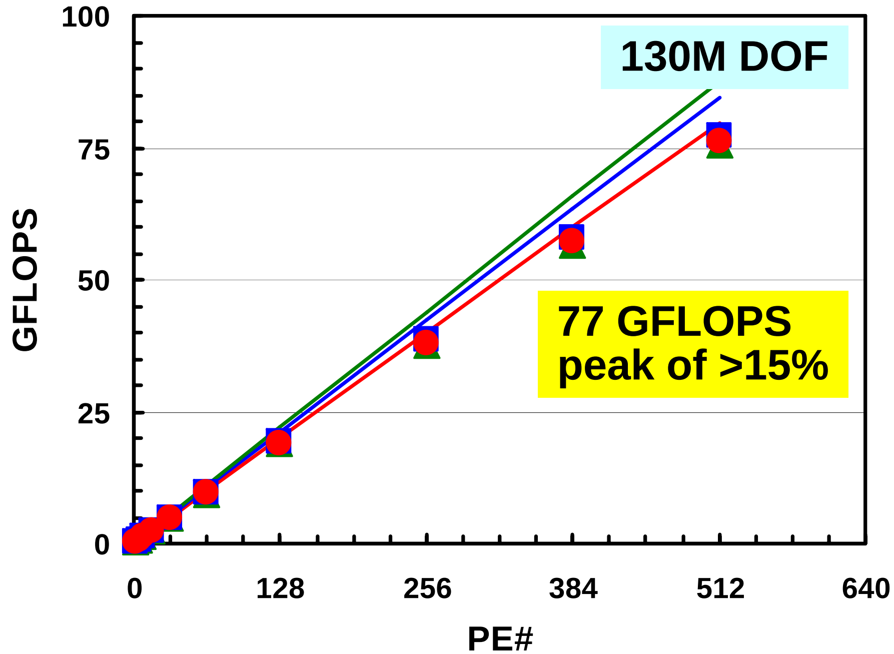
Effect of Problem Size

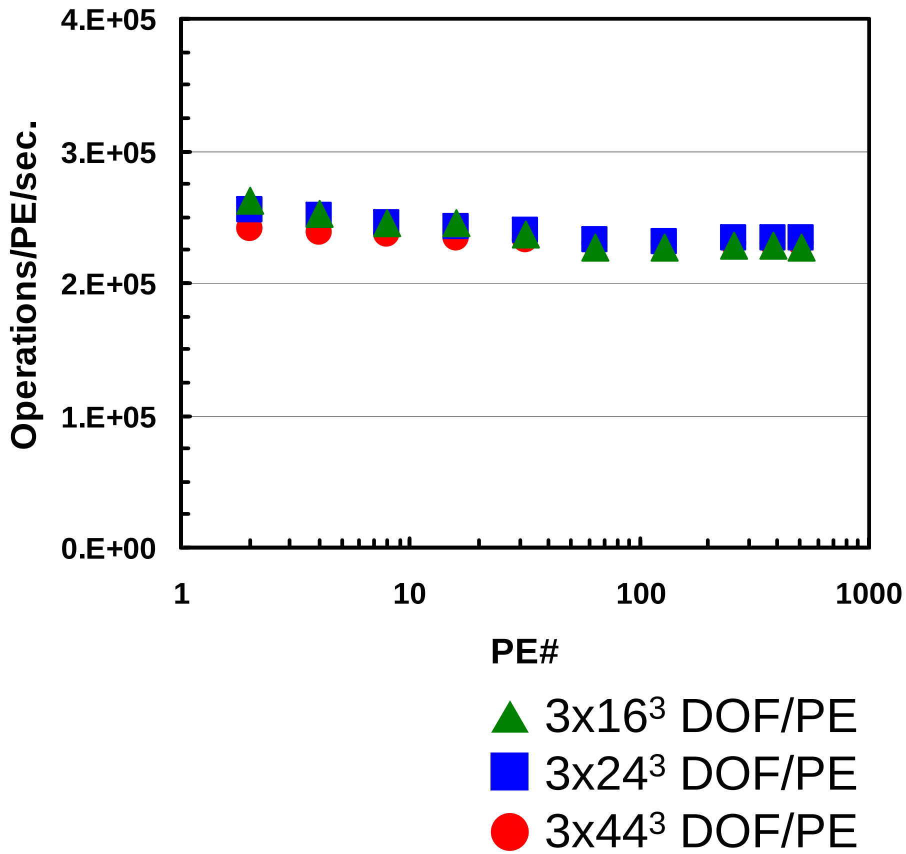
Results on 8 nodes (1PE/node)



Weak Scaling Test
Max. 130M DOF for 512 nodes

BG/L 512-node prototype@500MHz





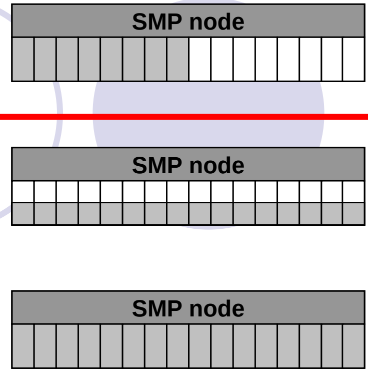
Weak Scaling Test

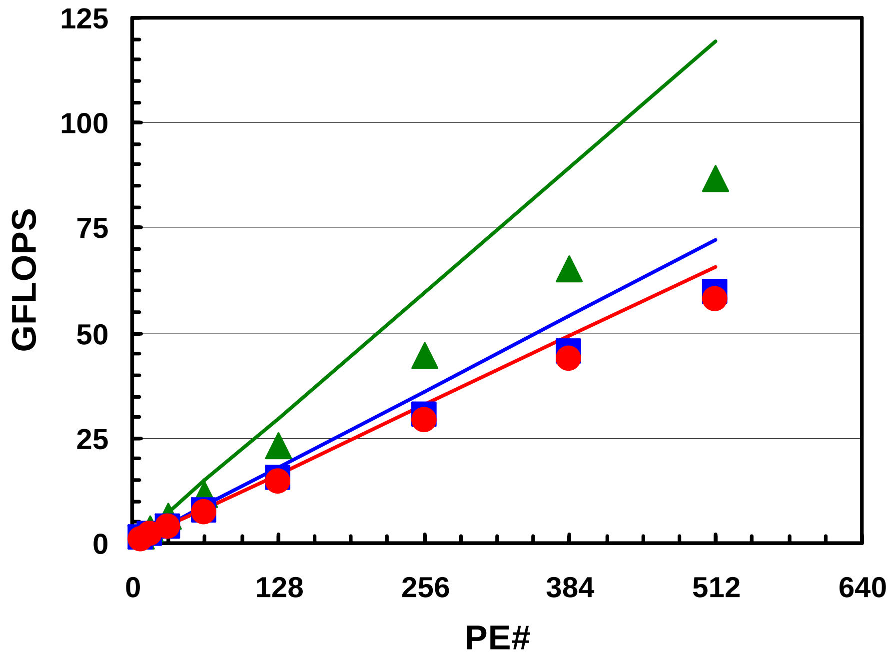
IBM-SP3/Seaborg type-2A

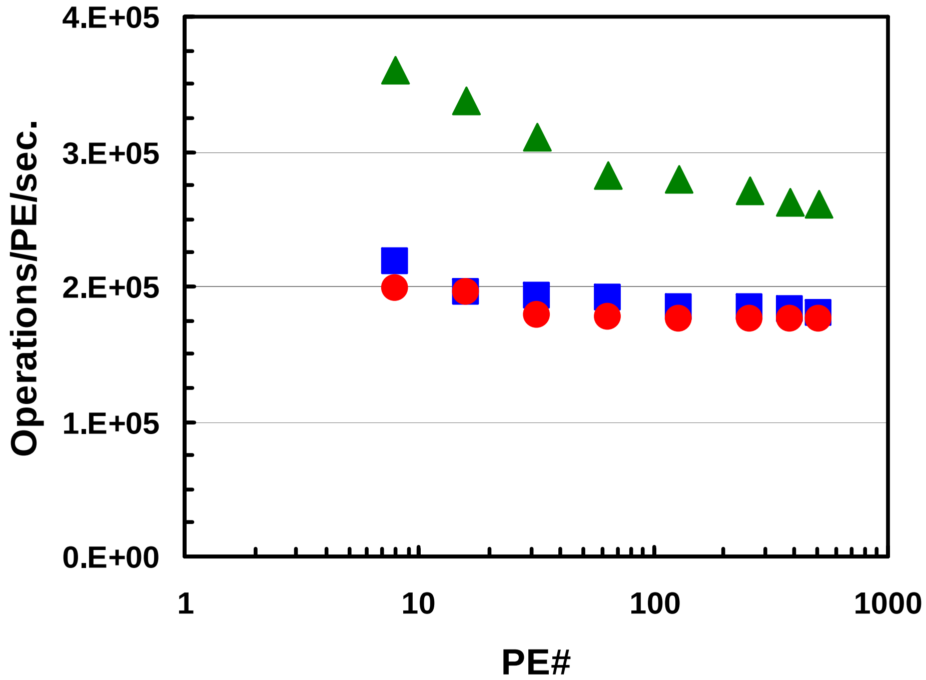
Type-2A
8 of 16 PE's are used

Type-2B
16 of 16 PE's are used
Problem size/SMP node is
as same as that of Type-2A.

Type-2C
16 of 16 PE's are used
Problem size/PE is
as same as that of Type-2A.





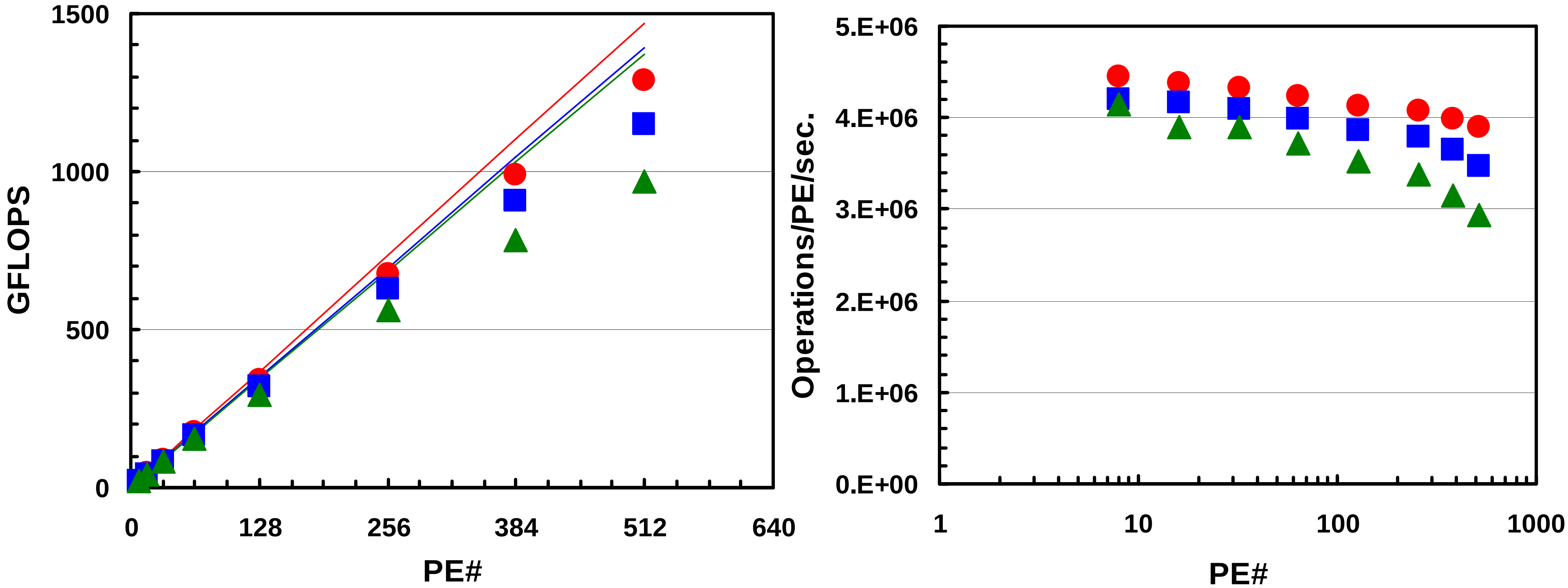


▲ 3×16^3 DOF/PE
■ 3×24^3 DOF/PE
● 3×44^3 DOF/PE

Scalability of BG/L prototype is much better

Weak Scaling Test

Earth Simulator

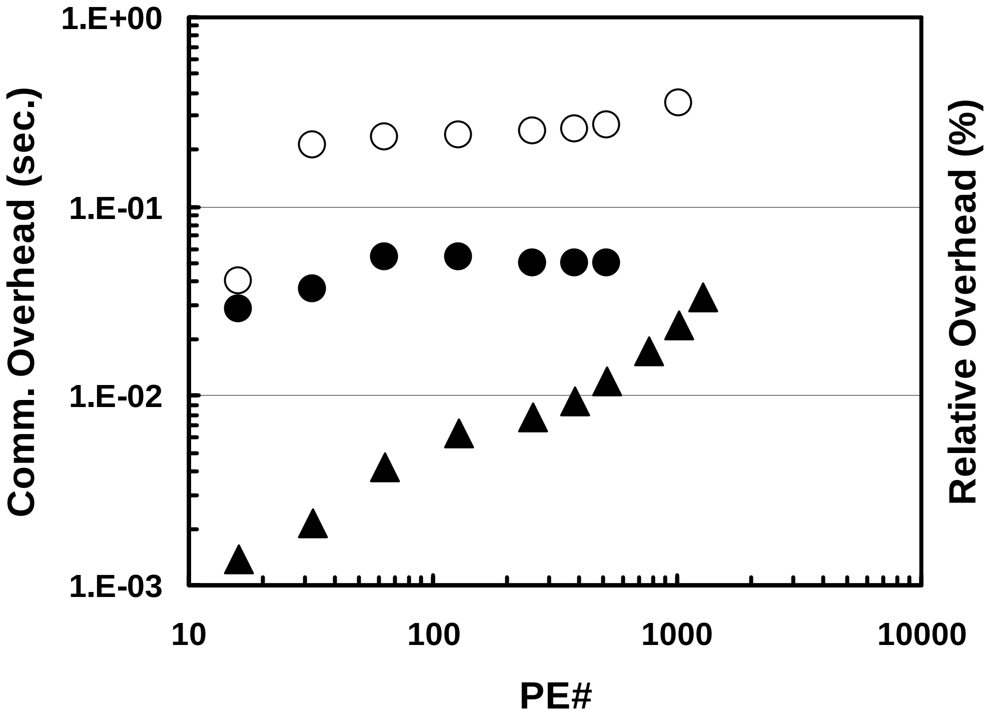


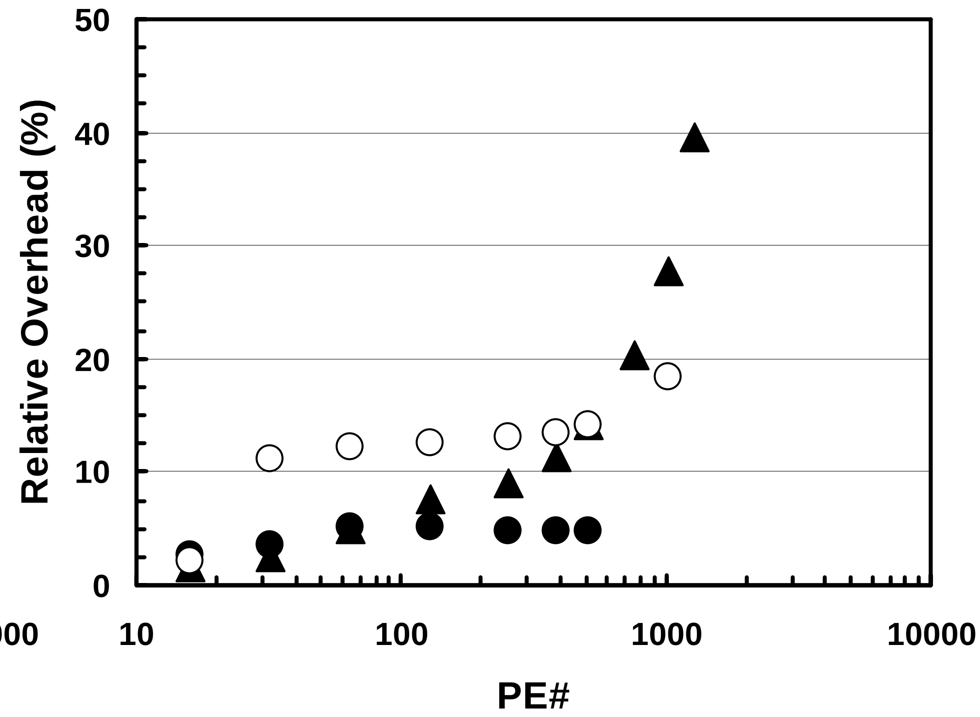
Scalability of BG/L prototype is much better

▲ $3x32^3$ DOF/PE
■ $3x40^3$ DOF/PE
● $3x50^3$ DOF/PE

Communication Overhead
Weak Scaling Test

Effect of MPI latency is significant if PE# is large.





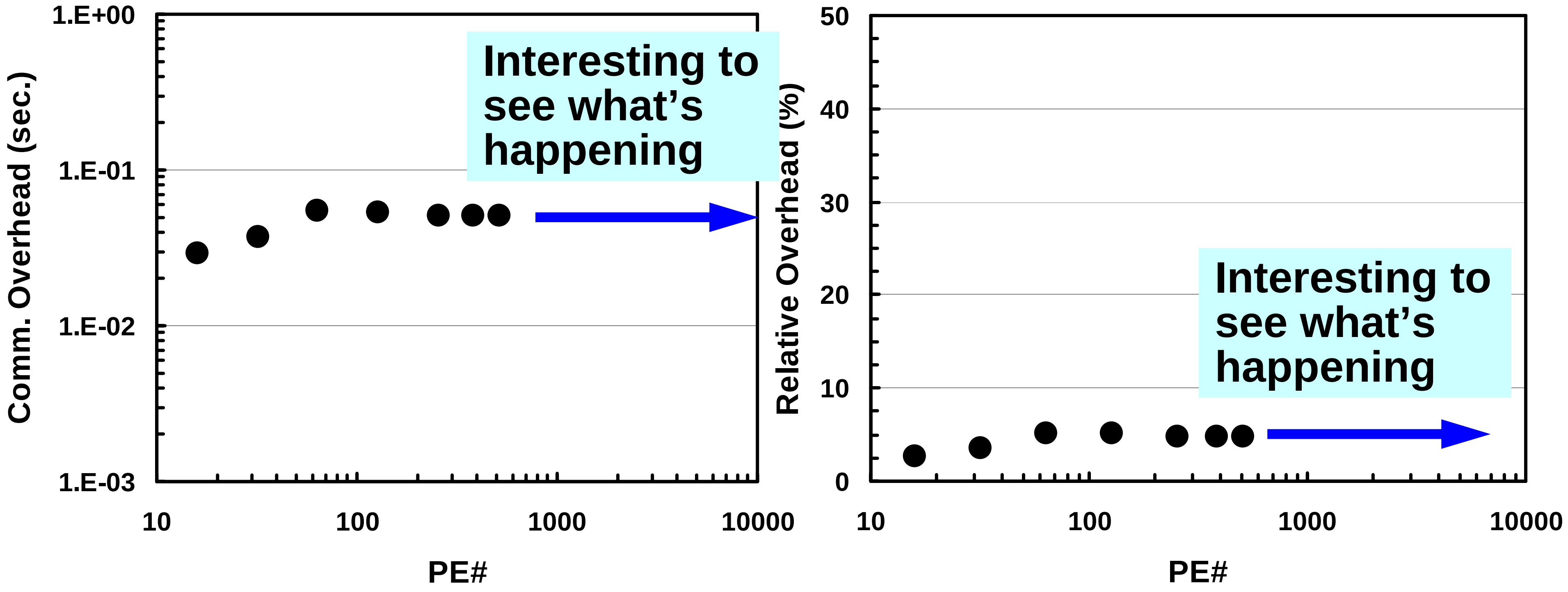
● BG/L Prototype 3x44^3 DOF/PE

○ SP-3 3x44^3 DOF/PE

▲ ES 3x50^3 DOF/PE

Communication Overhead
Weak Scaling Test

Effect of MPI latency is significant if PE# is large.

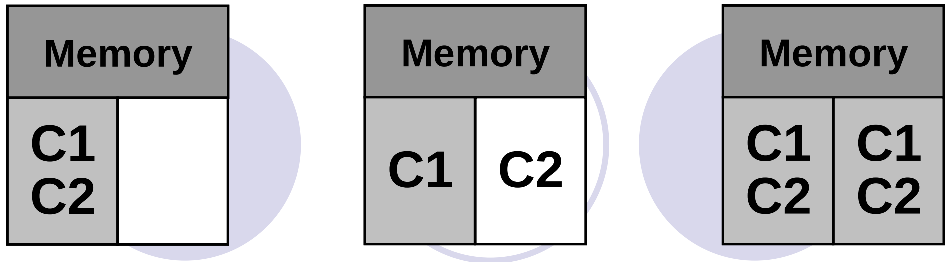


● BG/L Prototype $3x44^3$ DOF/PE

Comparison with IBM SP3 at NERSC/LBNL, and ES

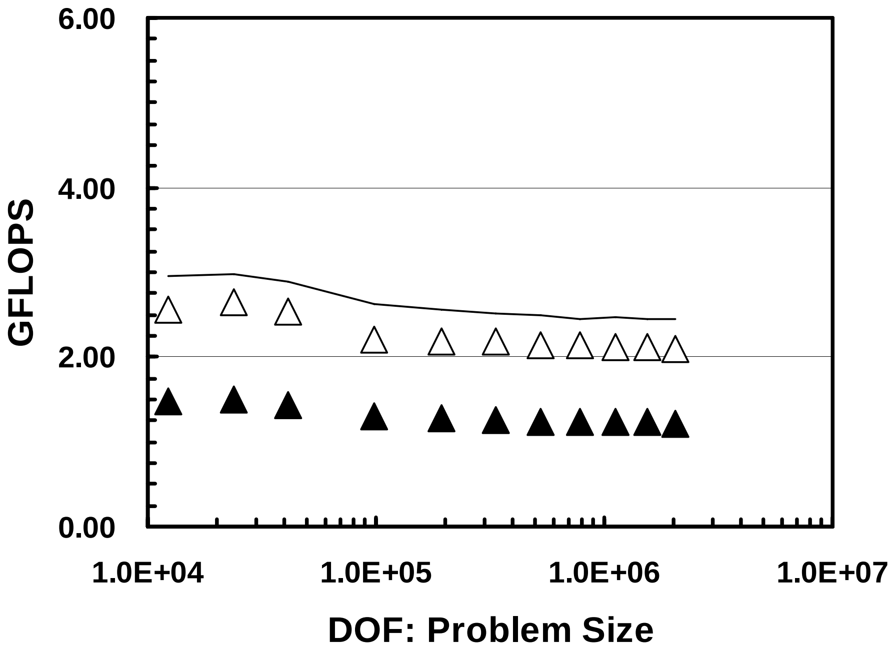
	BG/L	Seaborg at NERSC/LBNL	Earth Simulator
Architecture	IBM Power PC 440	IBM POWER3 Nighthawk 2	NEC SX-6 based
PE#/node	2	16	8
Clock rate	700 MHz	375 MHz	500 MHz
Peak performance/PE	1.40 GFLOPS (primary FPU only)	1.50 GFLOPS	8.00 GFLOPS
Memory/node	512 MB ~ 2 GB	16GB ~ 64 GB	16 GB
L1 Cache/PE (data/instruct)	32/32 KB	64/32 KB	-
L2 Cache	2 KB/PE	8 MB/PE	-
L3 Cache	4 MB/node	-	-
Memory-PE Bandwidth	5.5 GB/sec/node	16 GB/sec/node	256 GB/sec/node
Bidirectional Communication Bandwidth/node	2.1 GB/sec	2.1 GB/sec	12.3 GB/sec
MPI Latency	5.5-8.5μs	16.3μs	5.0-5.6μs

Effect of PE# on Each Node



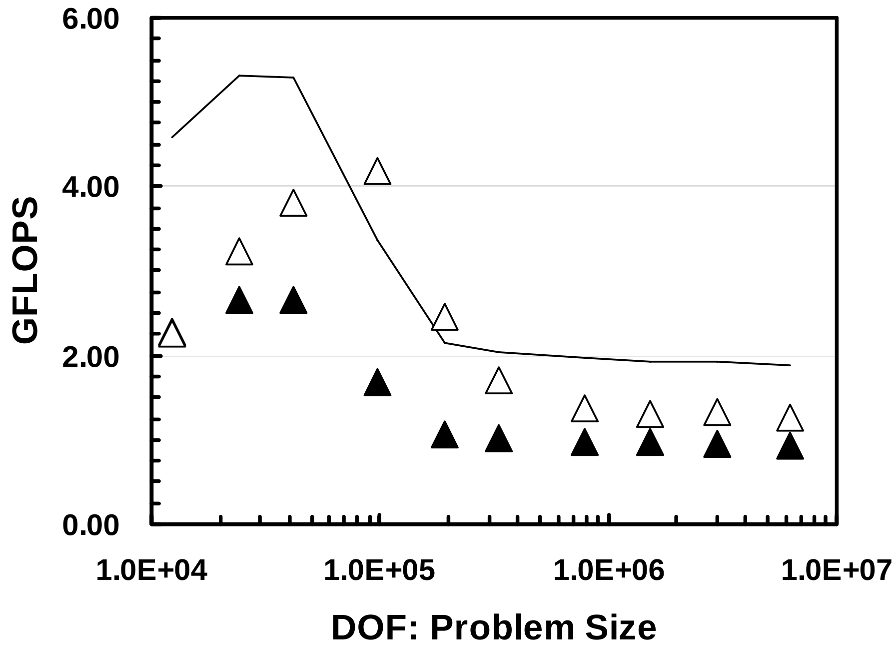
Heater | Communication Coprocessor | Virtual Node

IBM BG/L Prototype



▲ 8 nodes x 1PE/node= 8 PE's : **Heater**
△ 8 nodes x 2PE/node= 16 PE's : **Virtual Node**
— Estimated Ideal Performance of 16PE case extrapolated from results of 8 PE case

IBM SP-3/Seaborg



▲ 1 node x 8PE/node= 8 PE's
△ 1 node x 16PE/node= 16 PE's
— Estimated Ideal Performance of 16PE case extrapolated from results of 8 PE case

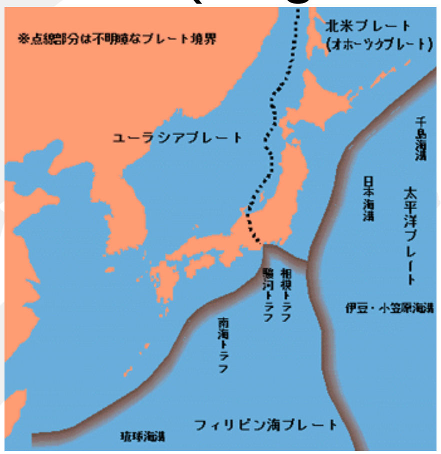
Summary on Elastic Linear Problem

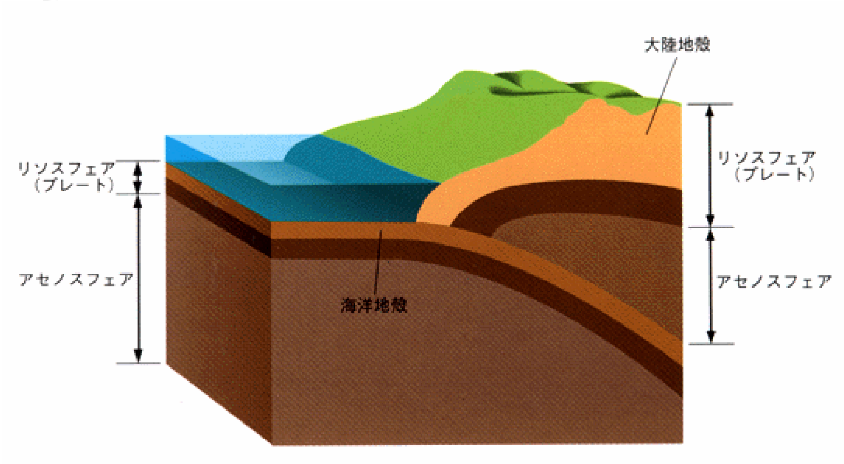
- Constant performance of BG/L for wide range of problem size
  - Single PE/parallel performance
  - 1 or 2 PE’s on each node
  - This point is very peculiar to other H/W, such as IBM-SP3, and the Earth Simulator.
    - large memory bandwidth/FLOPS ratio
    - small memory latency
  - >15% of peak performance (1GFLOPS/PE base)

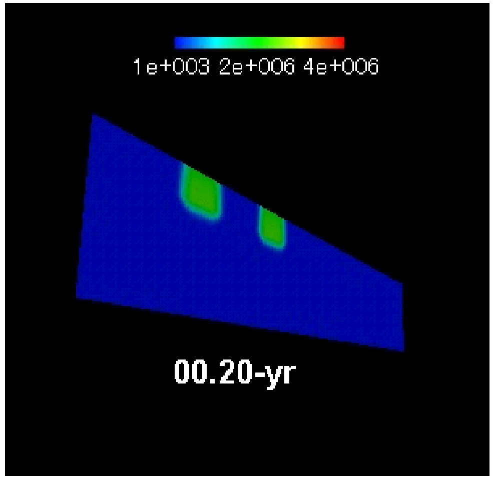
- Performance is better than IBM-SP3.

Preconditioning for Contact Problems

- Contact Problems in Simulations for Earthquake Generation Cycle by GeoFEM.
  - Quasi-Static Stress Accumulation Process at Plate Boundary.
  - Non-linear Contact Problems with Newton-Raphson Iter's
  - Ill-conditioned problem due to penalty constraint by ALM (Augmented Lagrangian).

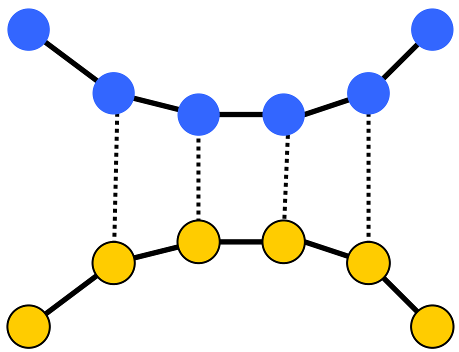






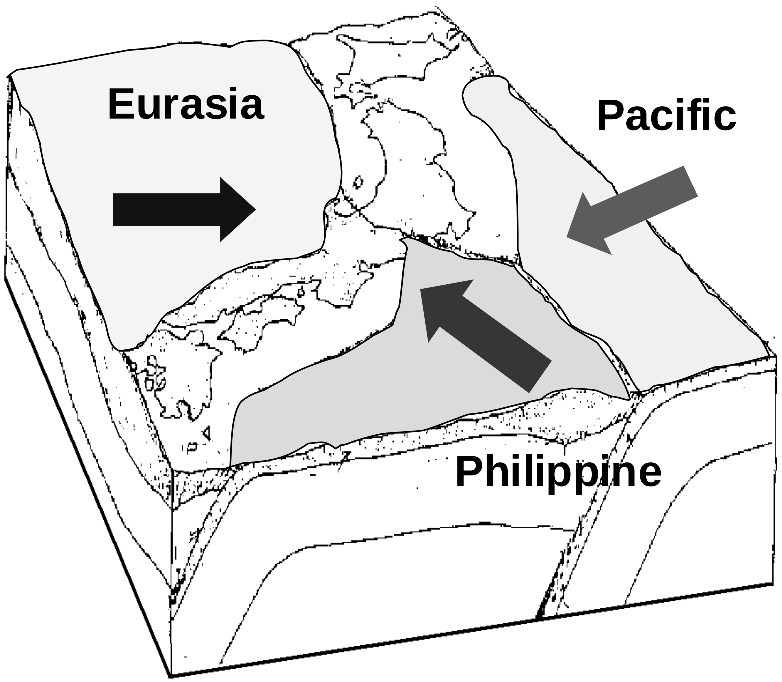
Preconditioning for Contact Problems (cont.)

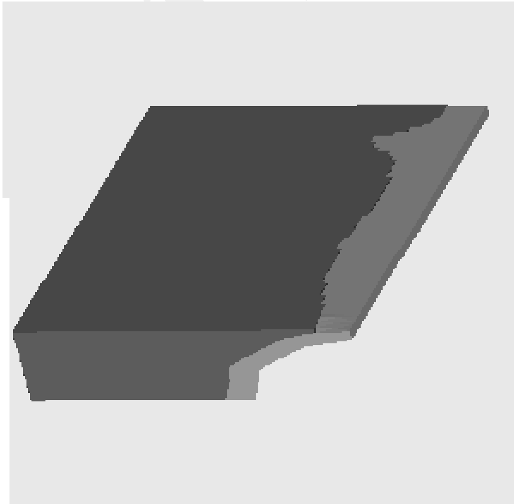
- Assumptions
  - Infinitesimal deformation, static contact relationship.
    - Location of nodes is in each "contact pair" is identical.
  - No friction : Symmetric coefficient matrix
- Special preconditioning : ***Selective Blocking.***
  - provides robust and smooth convergence in 3D solid mechanics simulations for geophysics with contact.

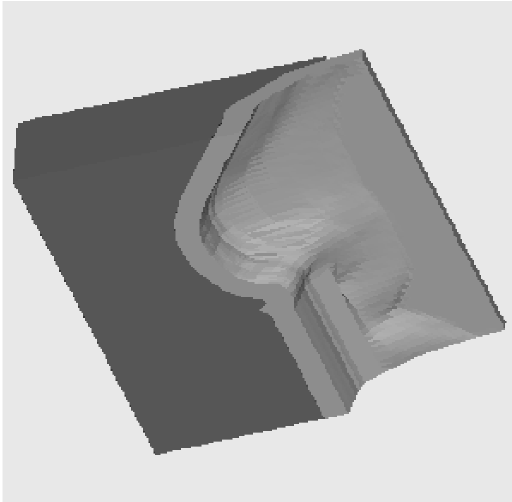


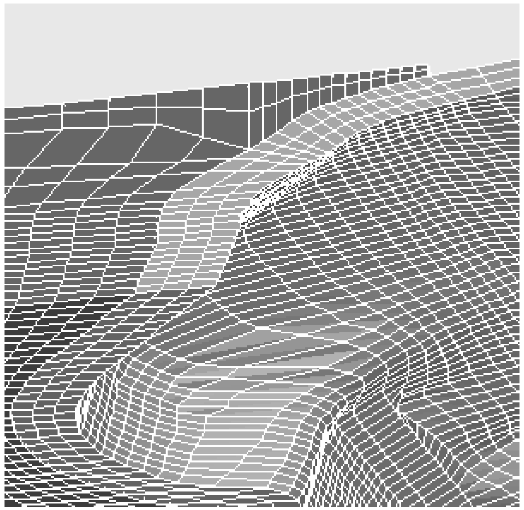
South-West Japan (SWJ)

fixed at $z=z_{min}$ + body force



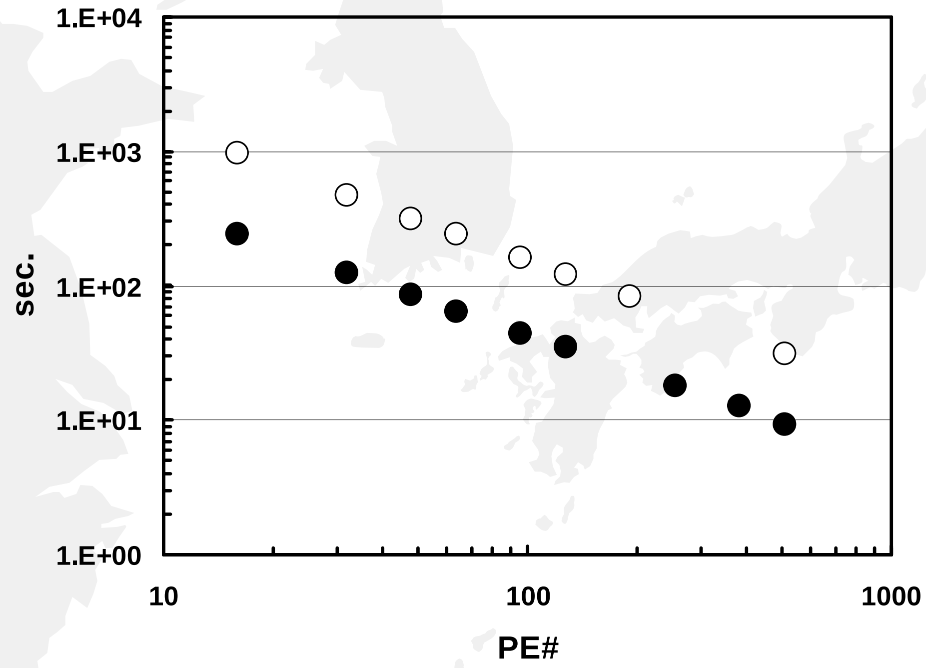


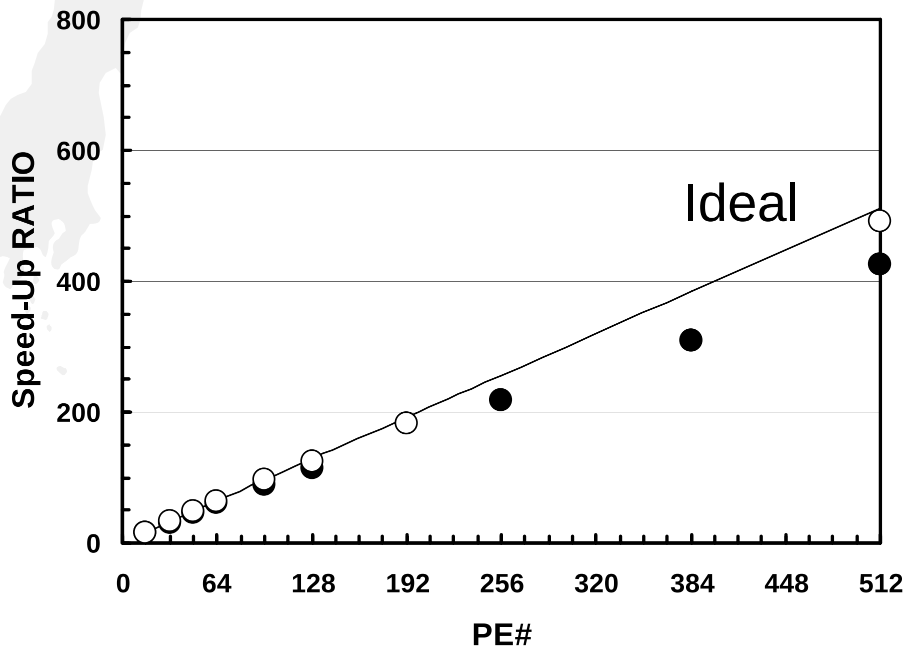




Strong Scaling Test of SB-BIC-CG

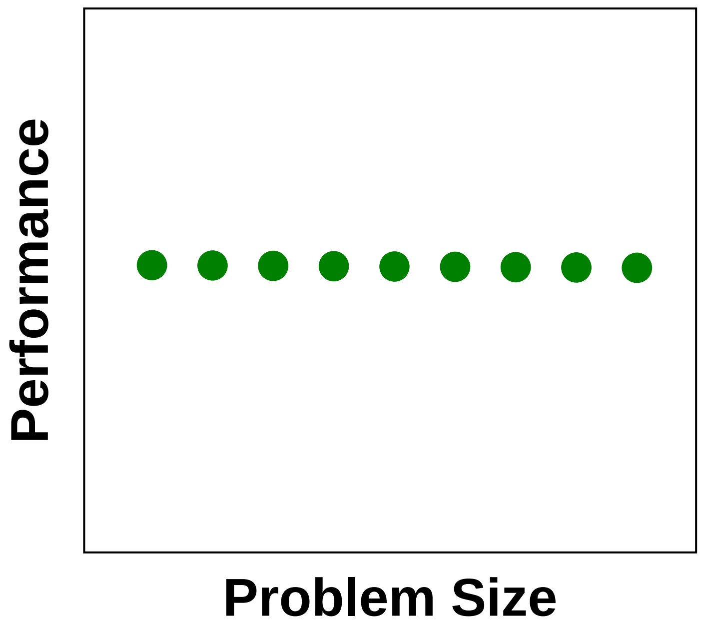
$\lambda/E=10^6$, 16-512 PE's of IBM BG/L Prototype Entire Prob. Size Fixed.

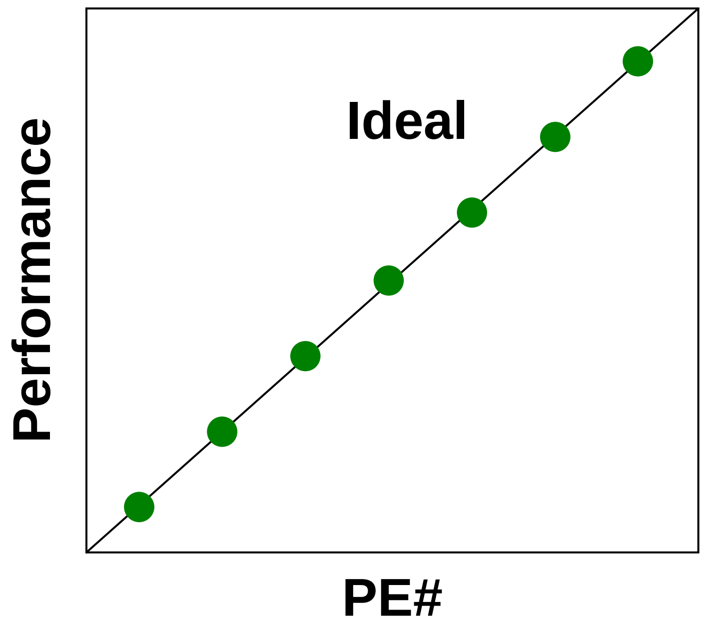




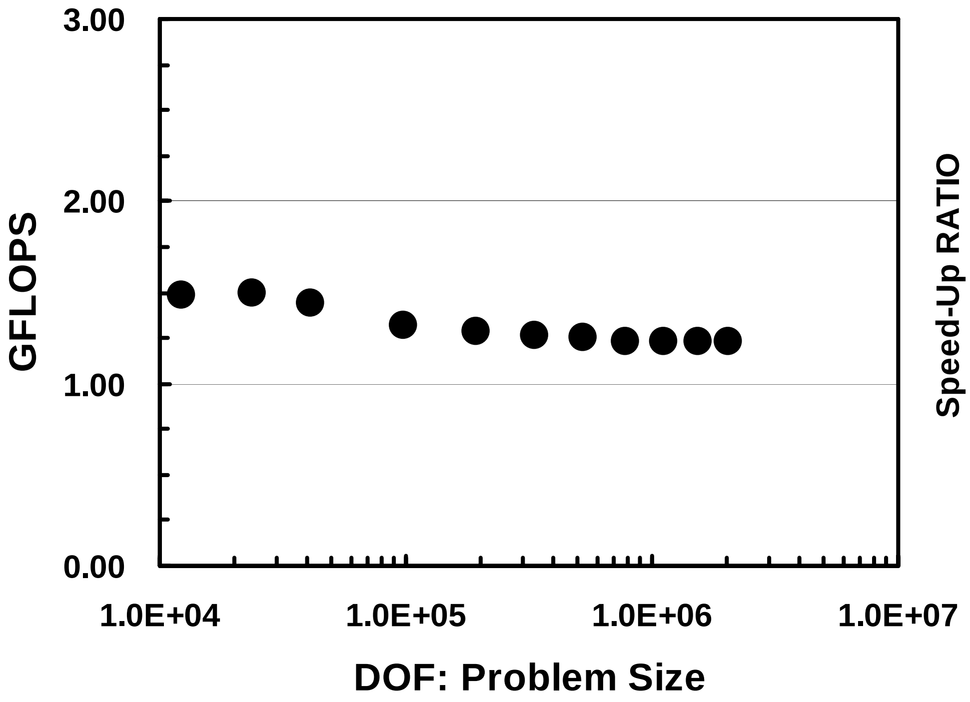
● Simple Block(2,471,439 DOF)
○ SWJ(2,992,266 DOF)

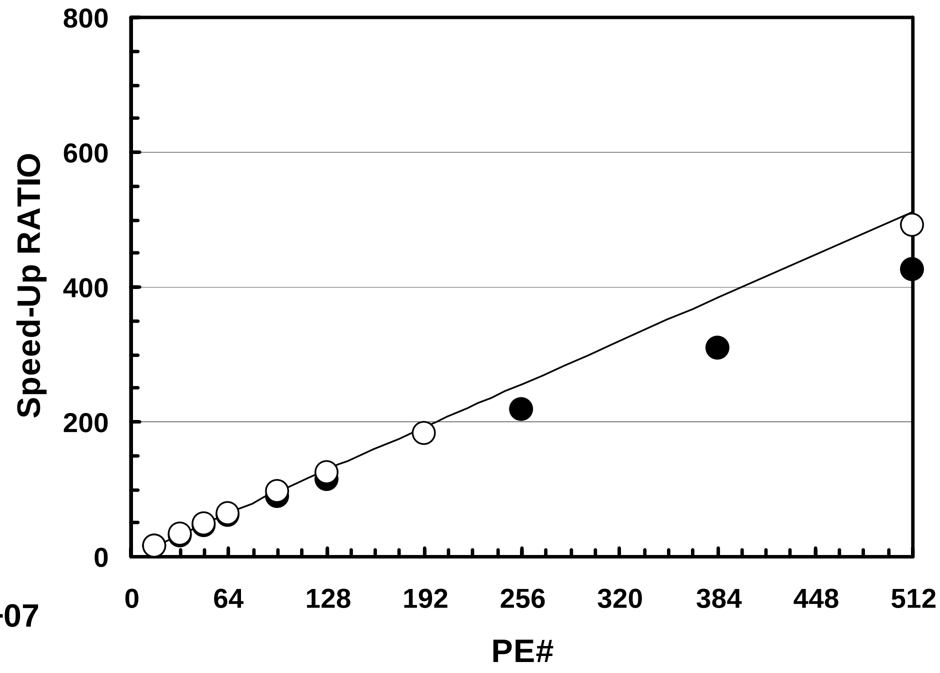
What is the Ideal Scalable System ?





What is the Ideal Scalable System ?





This is that !