NsimPower: Interconnect Simulator for Power and Performance Prediction

Koji Inoue

Kyushu University, Japan

Power/Performance Issues on Interconnection Network (IN)

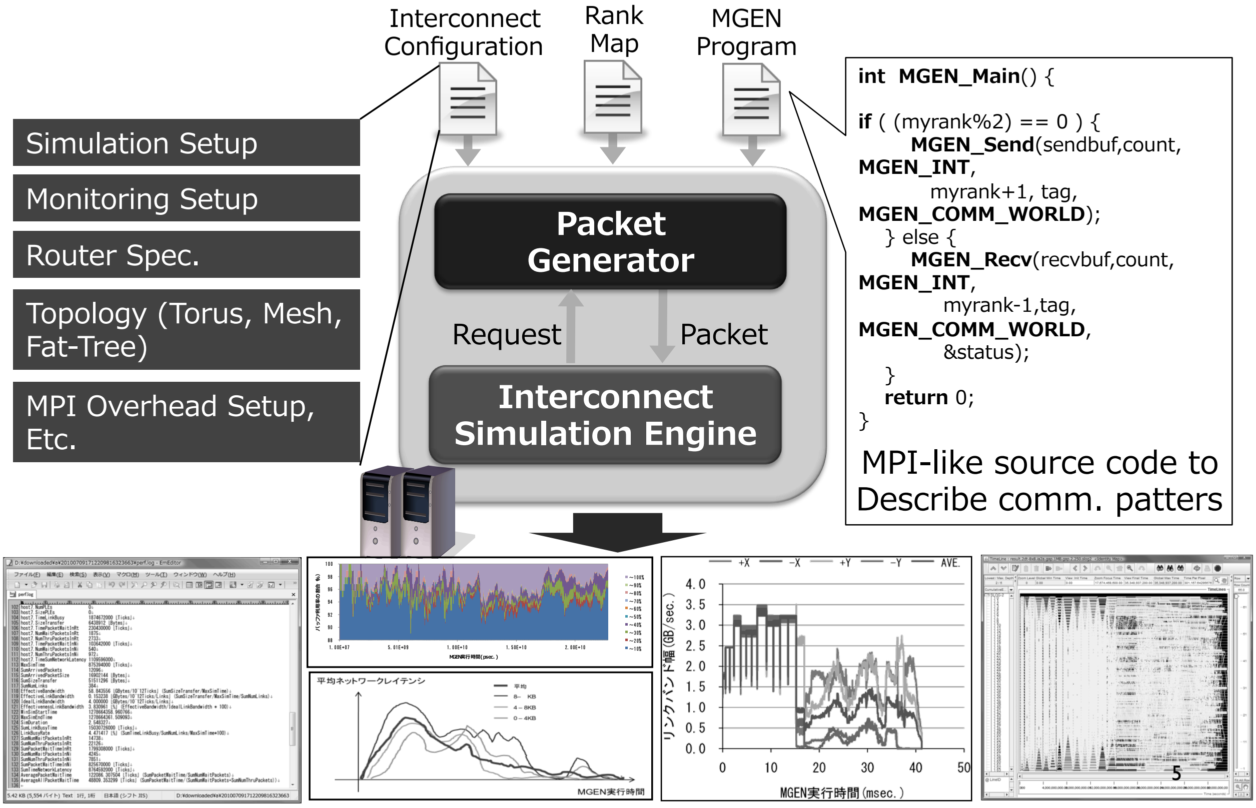
- Is interconnect power problem?
  - Roughly 10 to 30 % of total power
  - Increase in the number of computing nodes
  - High-bandwidth/Low-latency requirements for strong scaling
- Toward to power/energy efficient supercomputing
  - Need to consider computing node, memory, and interconnection network at the same time!
  - Bandwidth, latency and energy efficiency optimization from the view point of interconnects

Why We Need Interconnection Network Simulators?

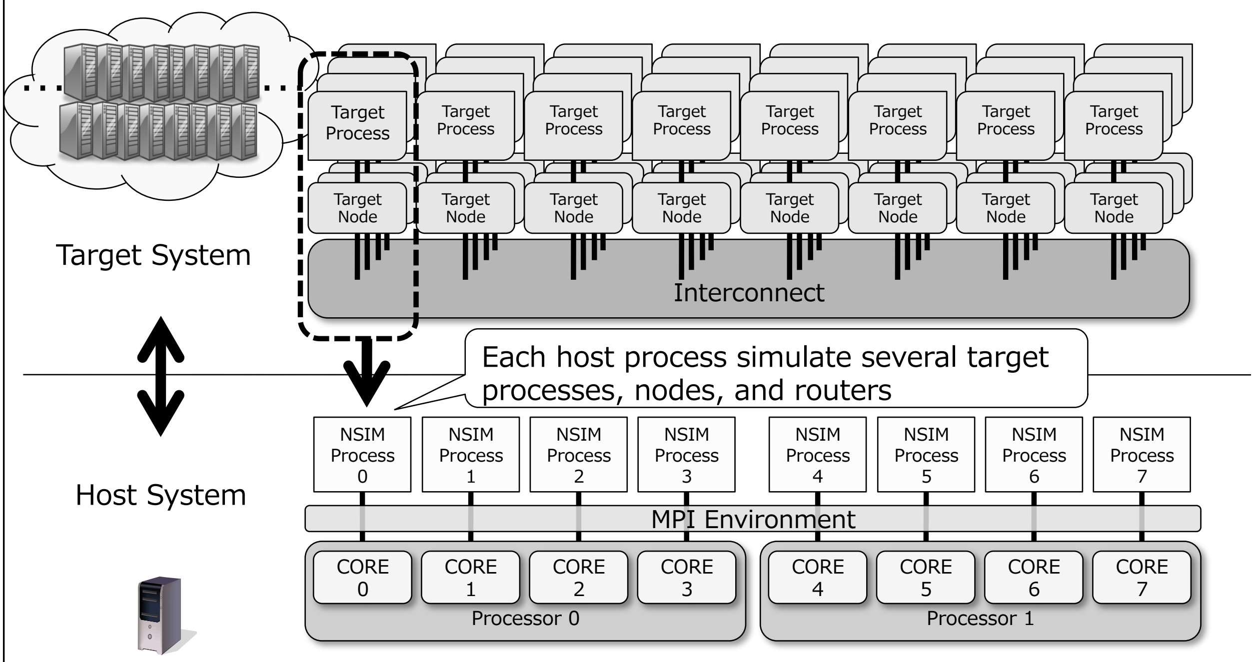
- For system designers
  - Design space exploration for high-performance, power-efficient large scale supercomputers
  - Detailed analysis for hardware (e.g. buffer size) and software (e.g. all-to-all algorithm) design parameters
- For application users
  - Understand execution behavior of own programs
  - Can be exploited for program optimizations

WHAT IS NSIM?

NSIM: Execution Driven Interconnection Network Simulator



NSIM Execution Image



Comparison with Other Simulators

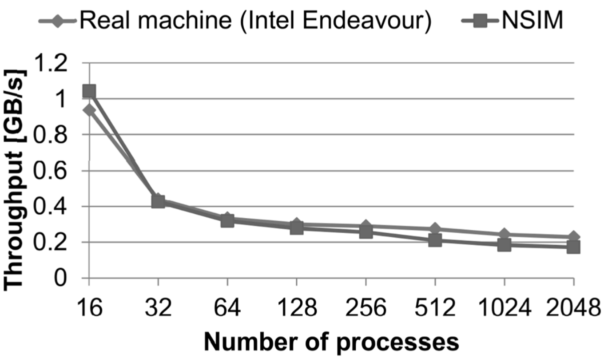
Simulator	BookSim [1]	BigNetSim [2]	BlueGene/L Interconnect Simulator[3]	NSIM
Developer	Stanford	UIUC	IBM	Kyushu Univ. ISIT, Fujitsu
Simulation Method	Execution Driven	Trace Driven	Trace Driven	Execution Driven
Parallel Simulation	---	Discrete Event Driven (Optimistic)	Discrete Event Driven (Optimistic)	Discrete Event Driven (Conservative)
Granularity	Flit Level	Packet Level		
Exe. Platform	Sequential Exe.	Parallel Machine (Distributed Memory)	Parallel Machine (Shared Memory)	Parallel Machine (Distributed Memory)
Simulation Target Size	Small (e.g. ~10 Nodes)	Large (e.g. 32K~64K Nodes)		

[1] W. Dally and B. Towles, "Principles and Practices of Interconnection Networks," Morgan Kaufmann Publishers Inc. 2003. http://cva.stanford.edu/books/ppin/
[2] N. Choudhury, T. Mehta, T. L. Wilmarth, E. J. Bohm, and L. V. Kale, "Scaling an optimistic parallel simulation of large-scale interconnection networks," Proc. of the Winter Simulation Conference, pp. 4-7, Dec 2005.
[3] N. R. Adiga, M. A. Blumrich, D. Chen, P. Coteus, A. Gara, M. E. Giampapa, P. Heidelberger, S. Singh, B. D. Steinmacher-Burow, T. Takken, M. Tsao, and P. Vranas, "Blue Gene/L torus interconnection network," IBM Journal of Research & Development, Vol. 49, No. 2/3, pp.265–276, 2005.

Accuracy

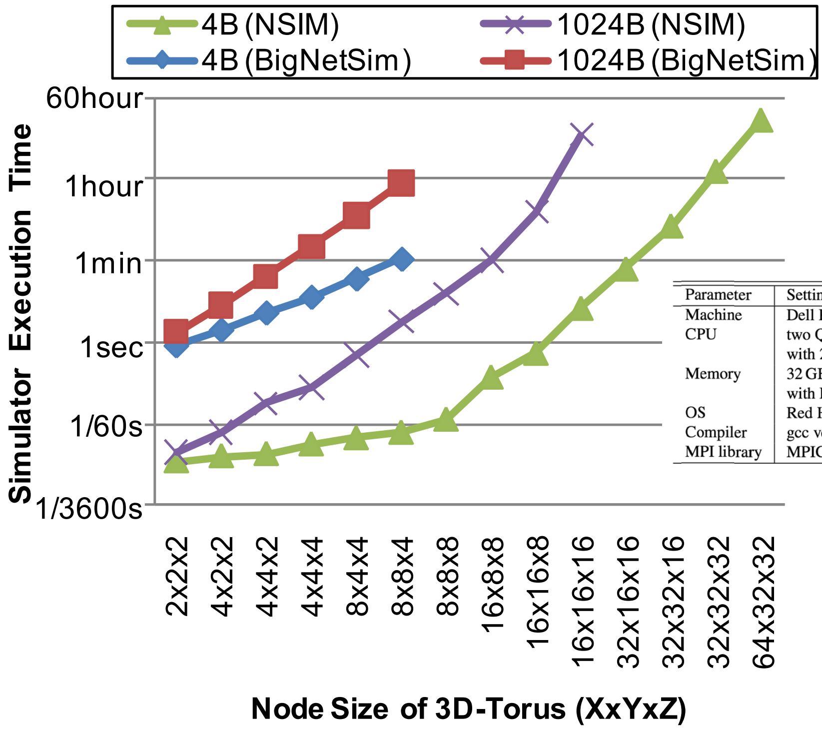
InfiniBand Fat-Tree, Random Ring, 2MB Messages

Type	Parameter	Value
Router	Unidirectional link throughput	4.0 GB/s
	Switch throughput	4.0 GB/s
	Routing computation time	4.0 ns
	Virtual-channel allocation time	4.0 ns
	Switch allocation time	4.0 ns
	Switch latency	90 ns
	Cable latency	10 ns
Node	DMA transfer rate	6 GB/s
	Memory bandwidth	6 GB/s
	MPI overhead	1.6 us
	Number of process	1 process/node



- Other evaluation
  - BlueGene/L (IBM)
  - Kei-Supercomputer (RIKEN/Fujitsu)
  - FX10 (Fujitsu)

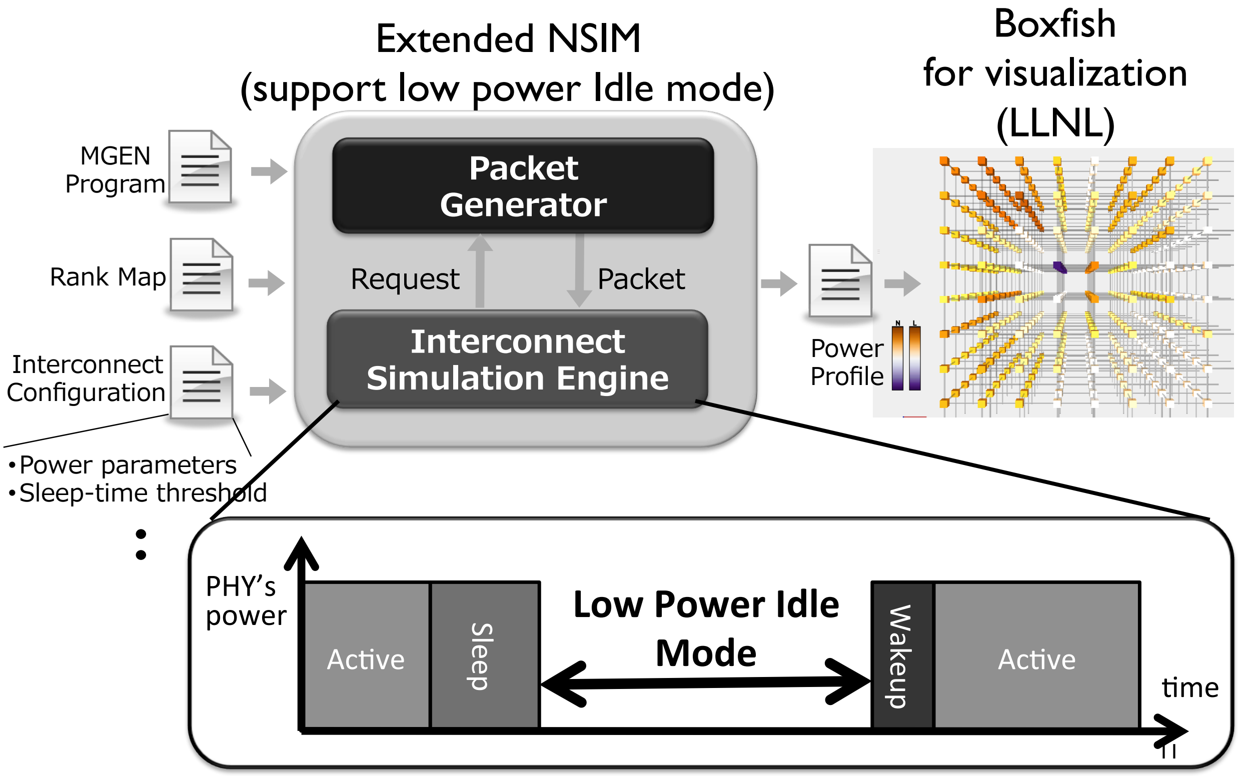
Simulation Performance
~The Case for Bruck's All-to-All~



Parameter	Setting
Machine	Dell Precision T7400 workstation
CPU	two Quad-Core Intel Xeon (R) processors E5440 2.8 GHz with 2x6 MB L2 cache
Memory	32 GB quad-channel DDR2-SDRAM memory 667 MHz with ECC
OS	Red Hat Enterprise Linux 5
Compiler	gcc version 4.1.2
MPI library	MPICH2 1.0.8 [16]

EXTENSION FOR POWER-PERFORMANCE ANALYSIS

Overview of NsimPower



Chunk based Power Modeling

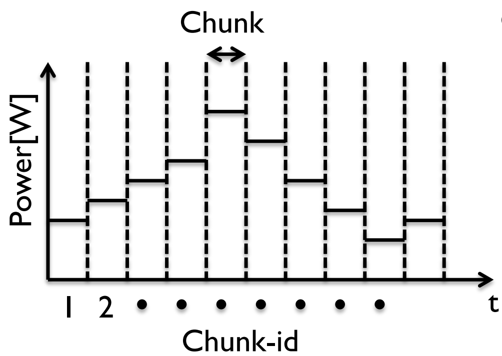
Power of router-*j* in chunk-*i*

#of links connected to router-*j*

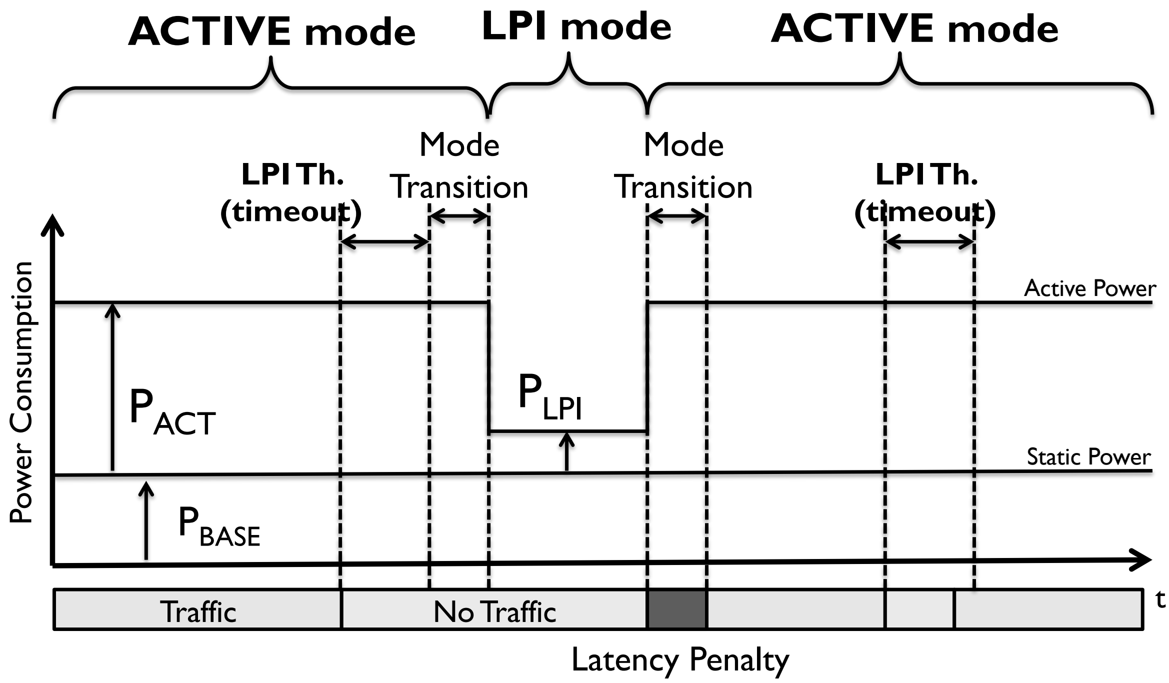
$$P_{ij} = \sum_{k=1}^{Nlink} \{P_{ACT}\} + P_{BASE}$$

Ave. active link power of router-*j*

Ave. static power of router-*j*



Supporting Low-Power Idle (LPI) Technology



Power Model Supporting Low-Power Idel Operations

Power of router-*j* in chunk-*i*

#of links connected to router-*j*

LPI rate of link-*k* in chunk-*i*

$$P_{ij} = \sum_{k=1}^{Nlink} \left\{ P_{ACT} \times (1 - R_{LPI-k}) + P_{LPI} \times R_{LPI-k} \right\} + P_{BASE}$$

Ave. link active power of router-*j*

Ave. link idle power of router-*j*

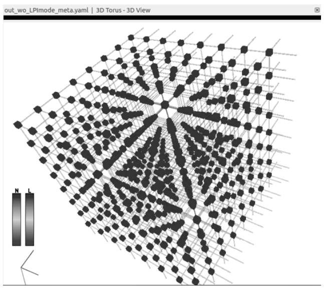
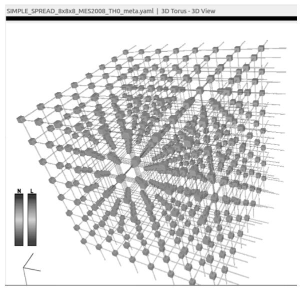
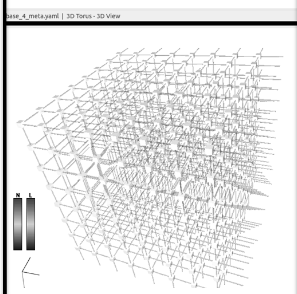
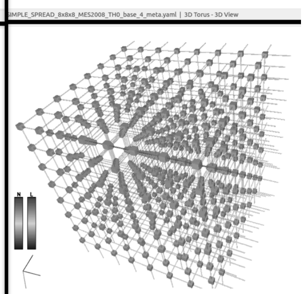
Ave. static power of router-*j*

CASE STUDY

Case Study

Ave. Base Power per Router	17.80 W (**1.0x, 0.25x**)
Ave. Power on ACTIVE mode per link	1.02 W
Ave. Power on LPI mode power link	0.10 W
Wake-up Transition Time	0 ns -> **Ideal case**
Sleep Transition Time	0 ns -> **Ideal case**
LPI Threshold	0 µs -> **Ideal case**
Chunk Length	50,000 ns
Topology	3D Torus (8x8x8)
Link Bandwidth	5GB/s
Packet Size	2,048 B
Communication	All-to-All (simple spread)

Potential of LPI Optimization

P_{BASE}=17.8W		P_{BASE}=4.5W (0.25x)	
w/o LPI	w/ LPI	w/o LPI	w/ LPI
			
12.25 KW	10.36 KW (-15.4%)	5.41 KW	3.58 KW (-33.8%)

Concusions

- Summary
  - NsimPower: Large Scale Interconnection Network Simulator for Power/Performance Analysis
  - Japan/US collaborated work
- On-going Work
  - Verify the accuracy of power estimation
  - Apply to large-scale power-performance analysis
  - Extend to system-wide power-performance prediction