Power Shifting in Thrifty Interconnect Networks

Jian Li [1] (Email: jianli@us.ibm.com)

with Wei Huang[1], Charles Lefurgy[1], Wolfgang E. Denzel[2], Richard Treumann[3], Kun Wang[4], Lixin Zhang[5]

[1] IBM Resarch – Austin
[2] IBM Research - Zurich
[3] IBM System and Technology Group, Poughkeepsie
[4] IBM Research - China
[5] Institute of Computing Technologies, Chinese Academy of Science (formerly IBM Research - Austin)

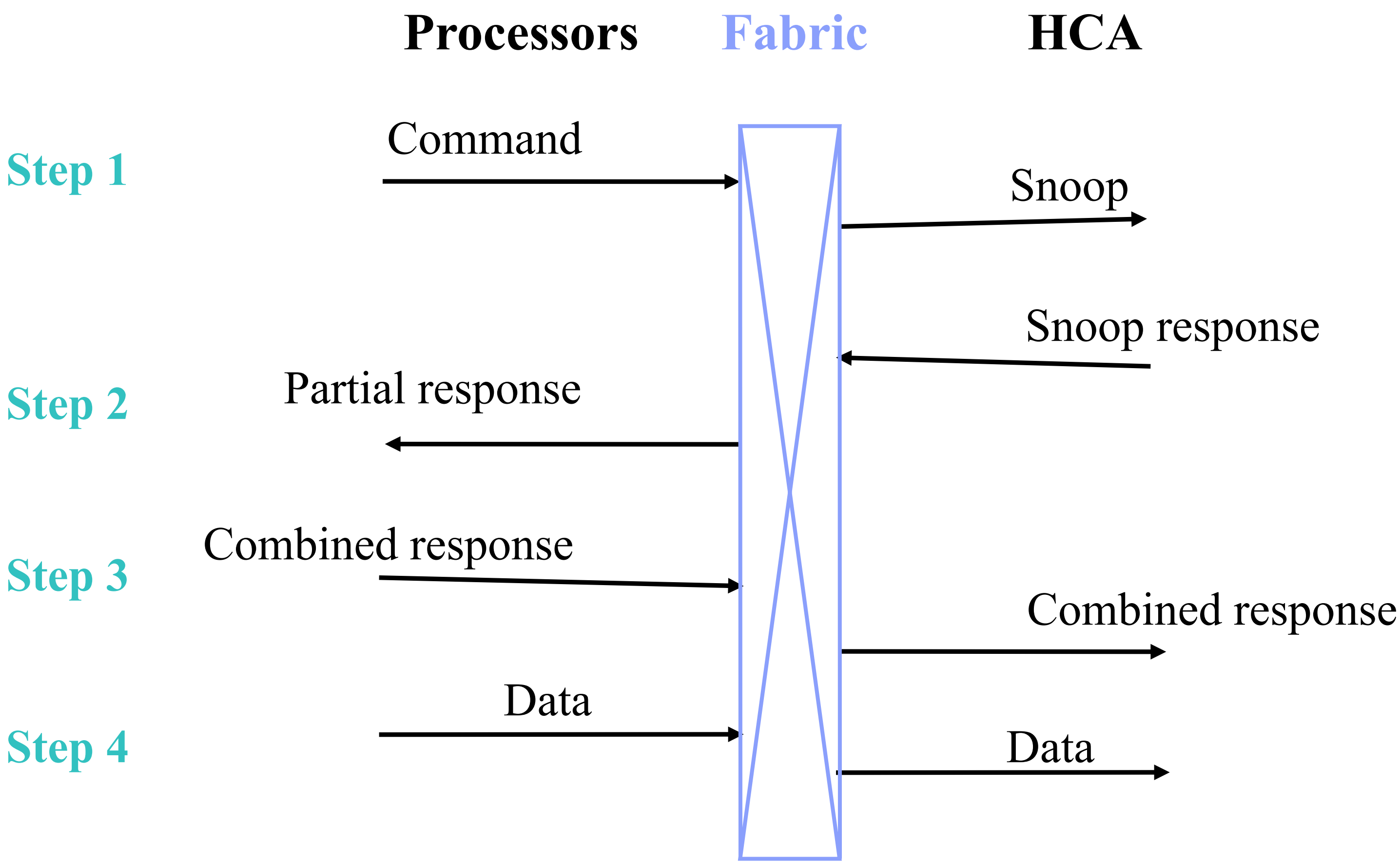
This Paper

- **Thrifty interconnection network**
  - Save power/energy and guarantee performance w/o prediction by utilizing inherent system events
- **Power shifting**
  - Between compute nodes and network components
- **Issues for further investigation and collaboration**

Motivation

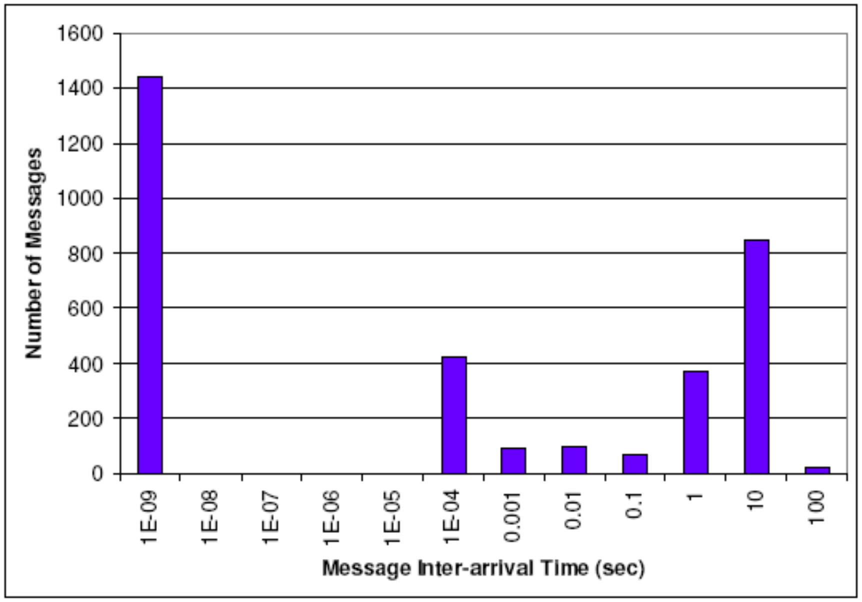
- **Power is a critical problem in modern supercomputing systems**
  - E.g., PERCS (Productive, Easy-to-use, Reliable Computing System), IBM's response to DARPA's HPCS Program
- **Link power can be significant portion of total system power**
  - 64% of the power budget of an IBM 8-port Infiniband 12X switch
  - 10% – 30% of the total system power in many HPC systems
    - 50% in ISCA'10 paper by Google
- **Link power characterization**
  - Constant training between transmitter and receiver
  - Average power almost identical to worst power
- **Protocol operations, between first command to initiate network message and data ready to go, can take time**

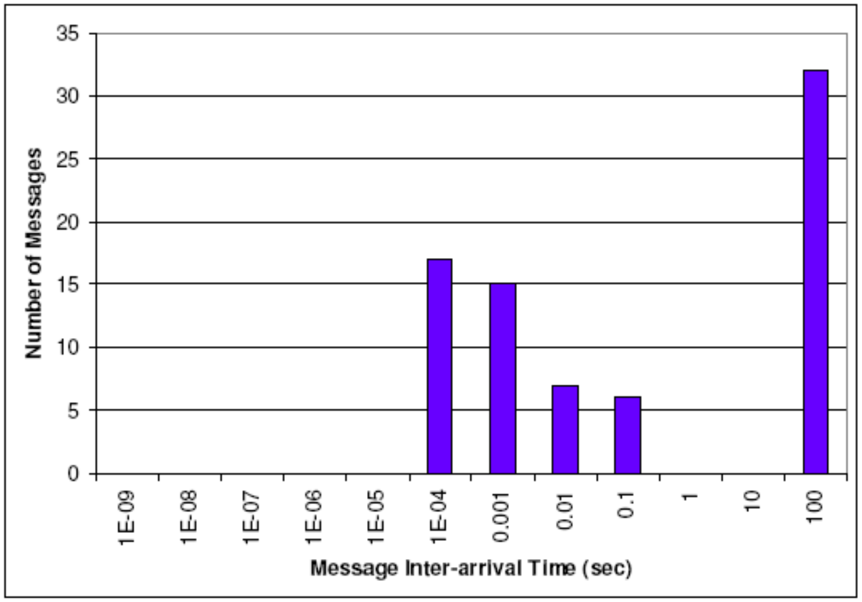
Protocol Overhead in an SMP Bus Transaction



Bursty Traffic Pattern

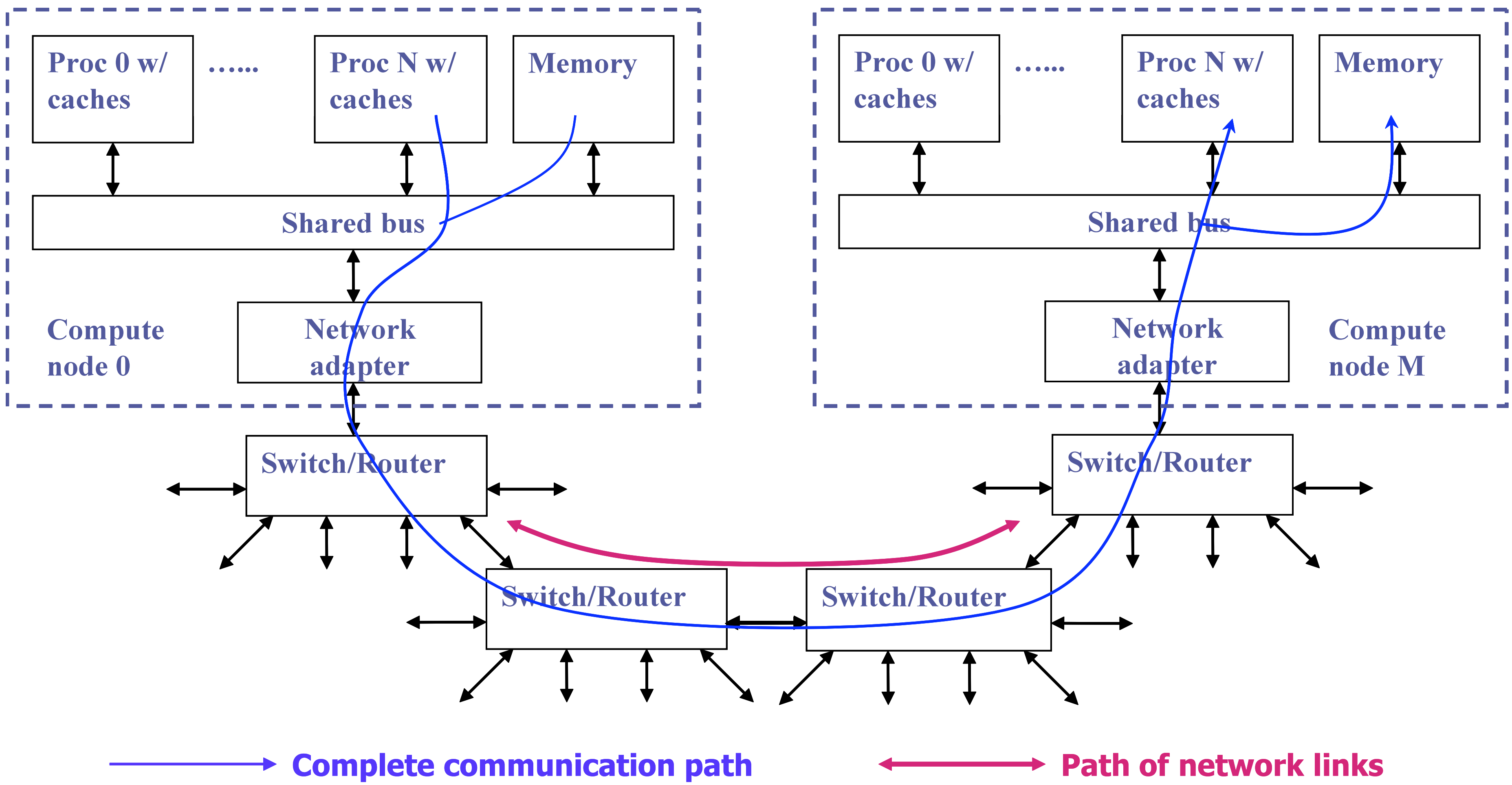
Distribution of MPI message inter-arrival time of SPPM (left) and WRF (right)



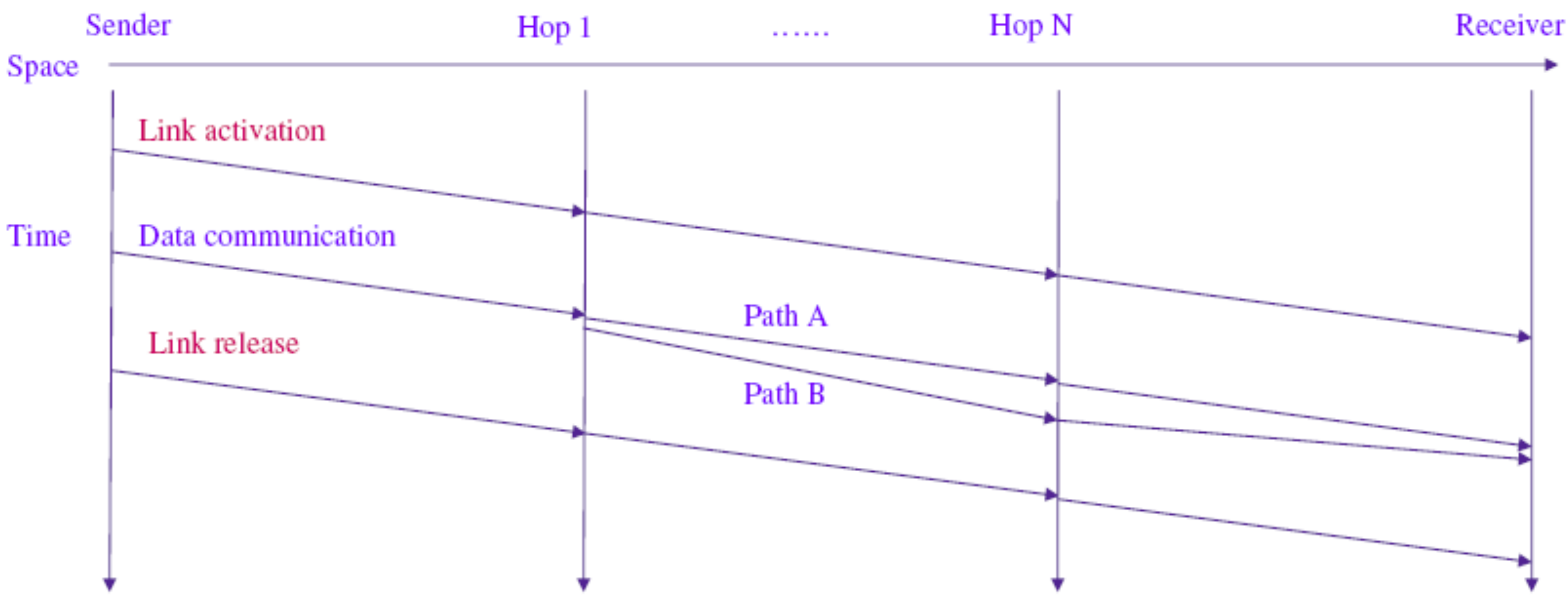


Overlapping computation and communication can be hard for many workloads

A Data Transmission Path in an Interconnected System

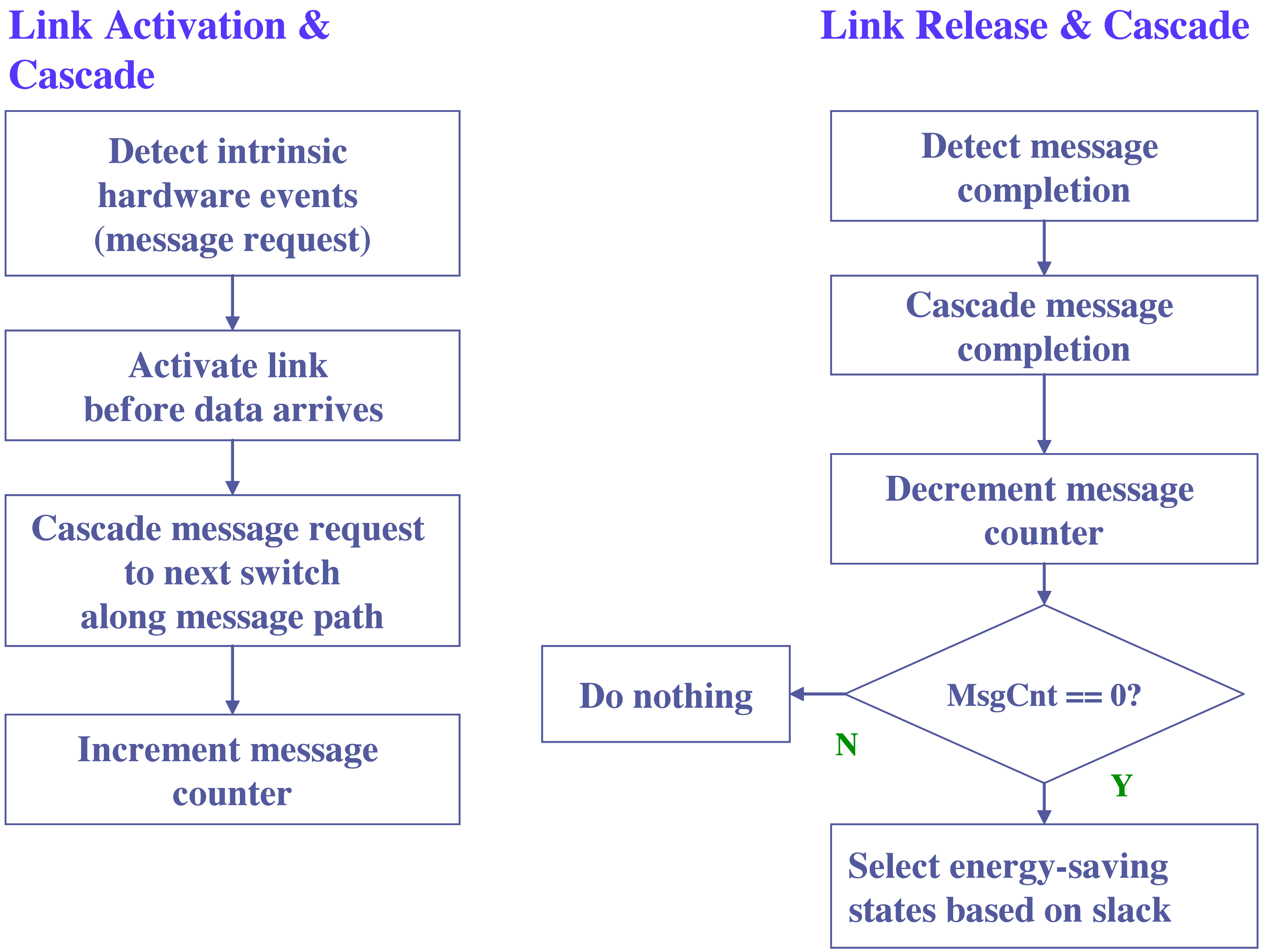


Communication in Thrifty Interconnection Network

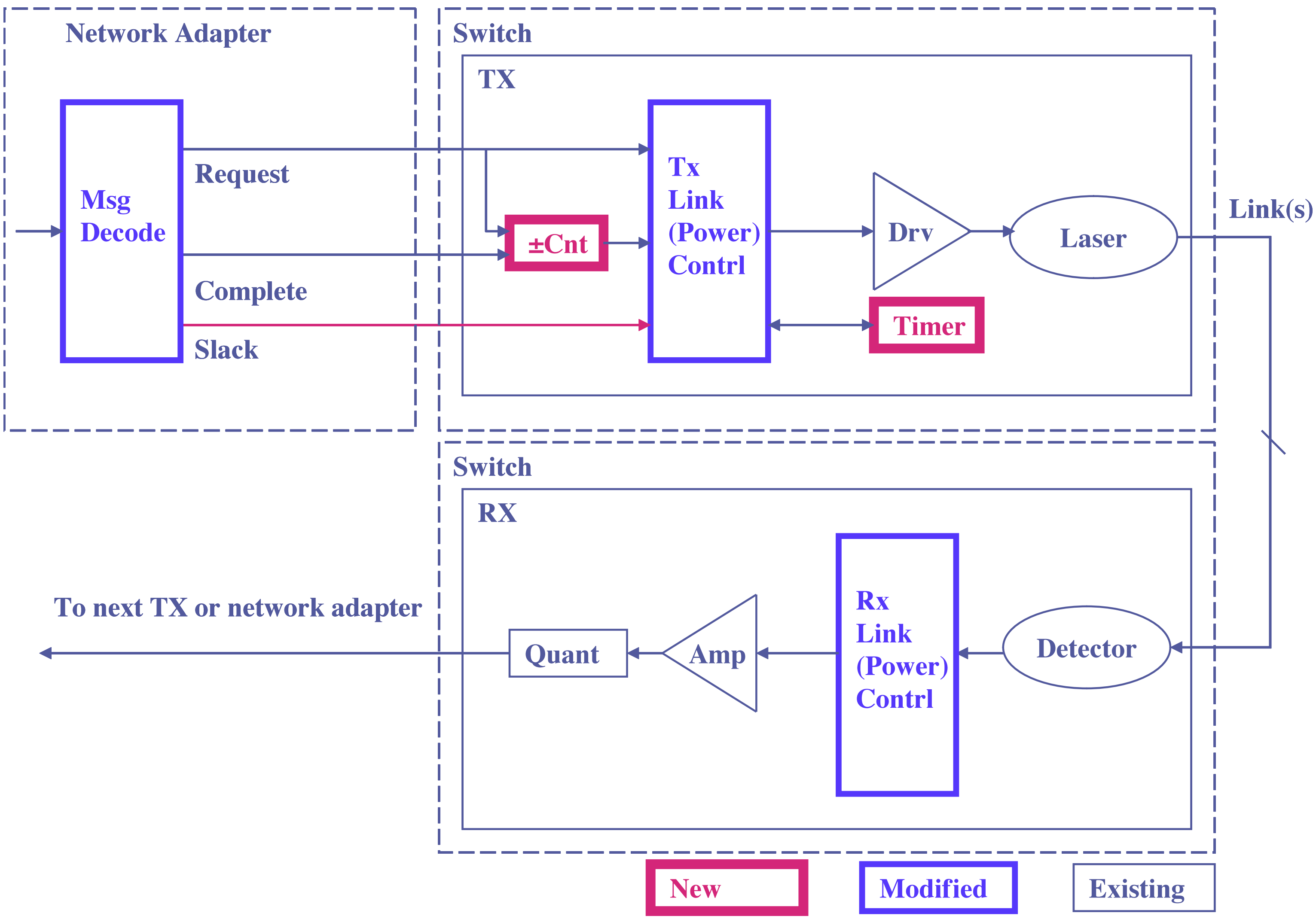


- **Activation and release via <u>separate control network</u>**
- **Packets in a network message can be routed through different paths**

Link Policy for Activation (left) and Release (right)



Hardware Support for Link Management

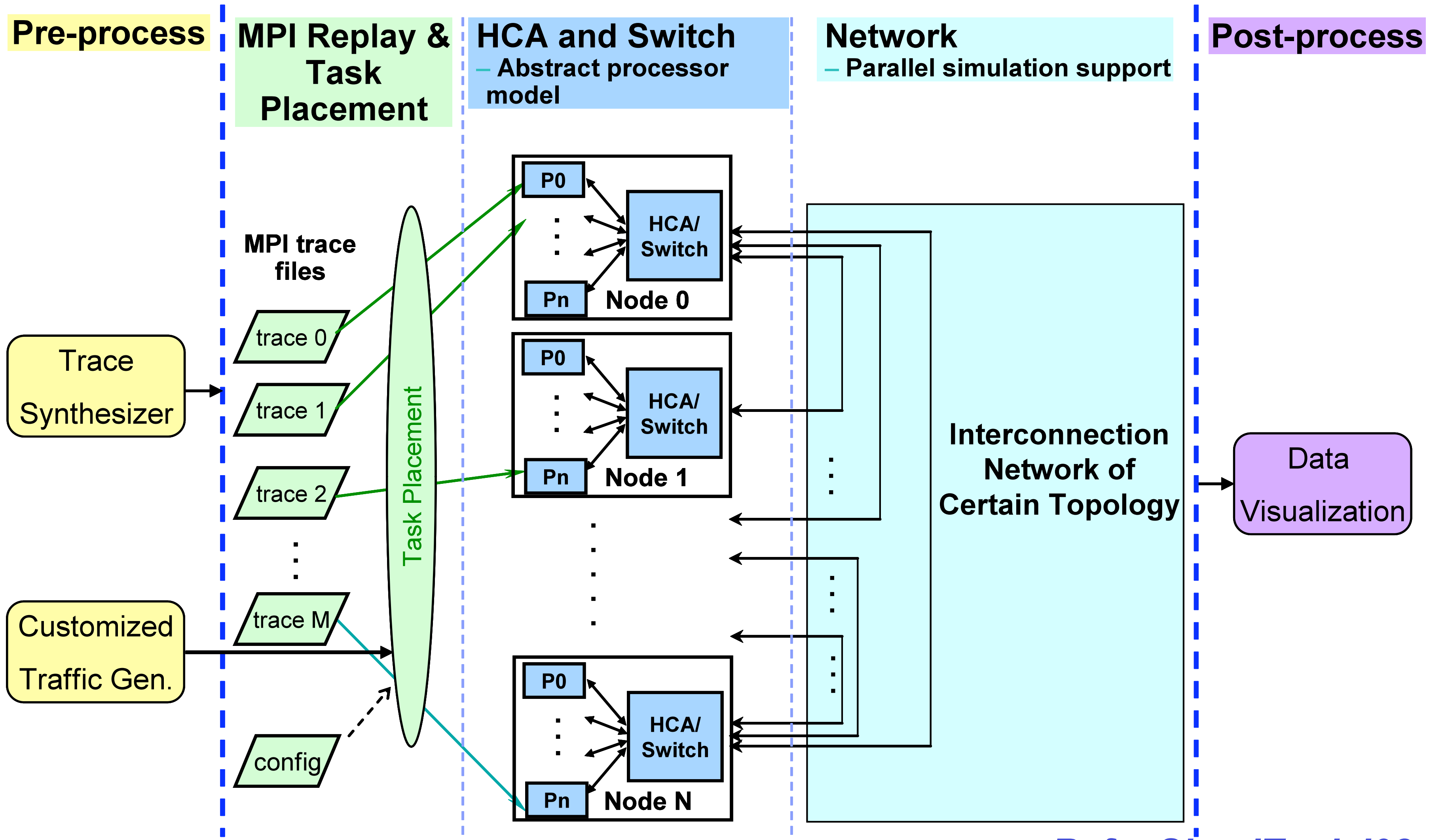


Example of Software Support

```
……
LINK_ACTIVATION(L) /* Command to inform network adapter to activate links that MPI_CALL_A will use. */
……
MPI_CALL_A( , , , )
LINK_RELEASE(L, SHUTDOWN) /* Command to hint network adapter to shutdown corresponding links. */
……
LINK_ACTIVATION(M) /* To activate a different set of links. */
……
MPI_CALL_B( , , , )
LINK_RELEASE(M, KEEP_ON) /* If too close together, this command and the next, LINK_ACTIVATION(M), can be
eliminated. */
……
LINK_ACTIVATION(M)
……
MPI_CALL_C( , , , )
LINK_RELEASE(M, SHUT_DOWN)
LINK_ACTIVATION(N) /* To activate a different set of links. */
……
MPI_CALL_D( , , , )
LINK_RELEASE(N, SHUT_DOWN)
……
```

- **Link shutdown can only be triggered by the message counter in switch**
- **LINK_RELEASE() is more a hint for optimization than a command for link shutdown**

Simulation Framework



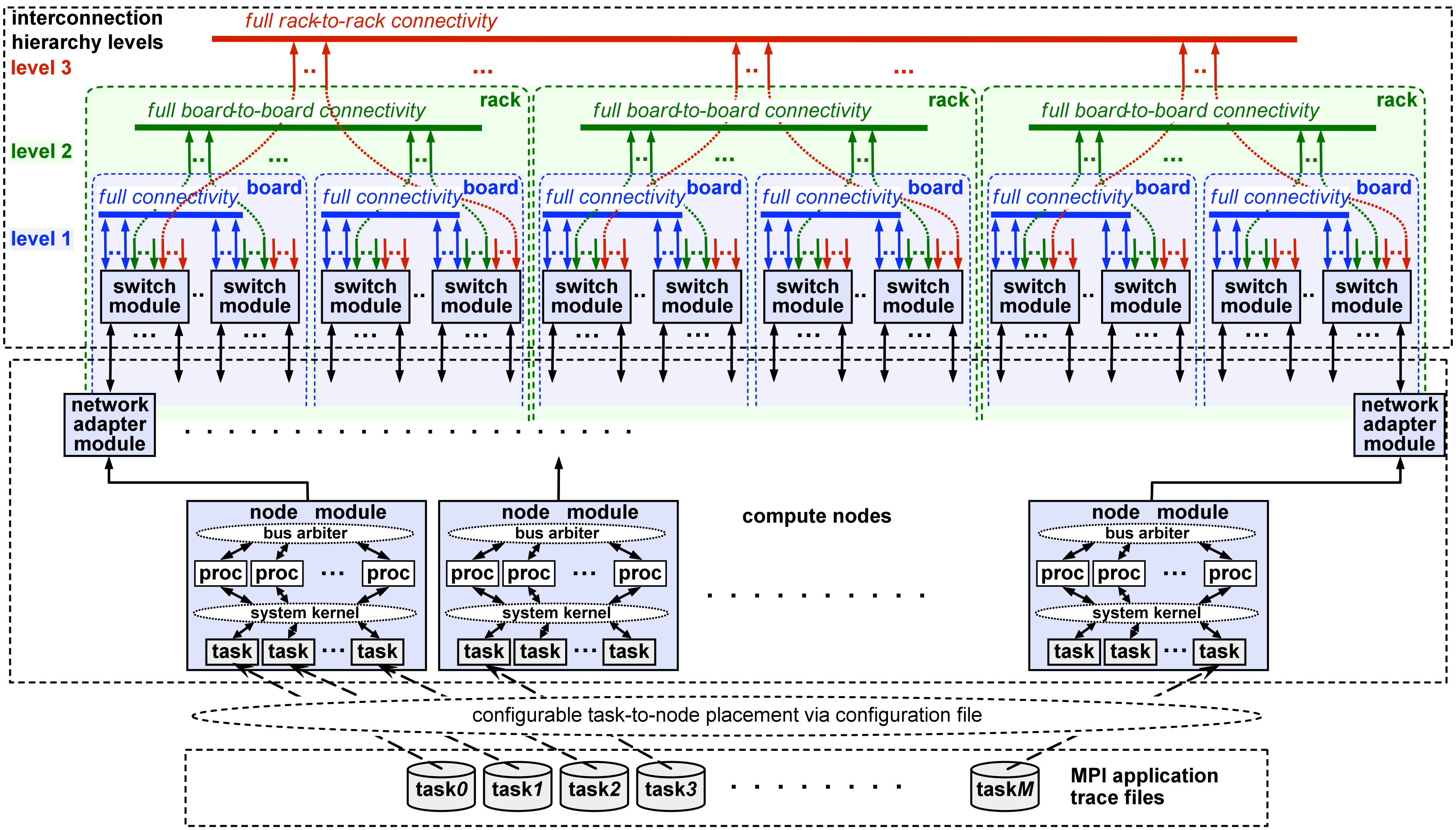
Refs: SimulTools'08

MPI Trace Format for Simulation

 # MPI Task Trace Sample (similar to MyMPI format)

```
1996807979380294 20 0 0 Init X "j25" 32 1996807979.380295
1996807979380294 20 0 0 Comp E
1996807979380367 20 -1 0 Comp X 1709 361 118 148 18 0 0.211235
1996807979380367 20 -1 0 Disable E
1996807984051219 20 -1 0 Enable X
1996807984051219 20 -1 0 Comp E
1996807984051303 20 -1 0 Comp X 91458 85647 26241 10095 21973 8941 0.936463
1996807984051303 20 -1 0 Allreduce E 3 14 2 0
1996807984051430 20 -1 0 Allreduce X 3 14 2 0
1996807984051430 20 -1 0 Comp E
1996807984053218 20 -1 0 Comp X 3025399 1378017 499996 233540 581190 144938 0.455483
1996807984053218 20 -1 0 Allreduce E 1 8 0 0
1996807984056071 20 -1 0 Allreduce X 1 8 0 0
1996807984056071 20 -1 0 Comp E
1996807984056103 20 -1 0 Comp X 2150 365 111 135 18 0 0.169767
1996807984056103 20 -1 0 Mark E 2
1996807984056103 20 -1 0 Comp E
1996807984056129 20 -1 0 Comp X 17749 7483 2476 1579 1362 1 0.421601
1996807984056129 20 -1 0 Irecv E 1317 14 4 0 0 804394832
1996807984056147 20 -1 0 Irecv X 1317 14 4 0 0 0
. . . . . . .
```

A Hierarchical Direct Interconnect Architecture



Refs: IBM PERCS, SimulTools'08, ISCA'08, HotInterconnet'10

System Configuration

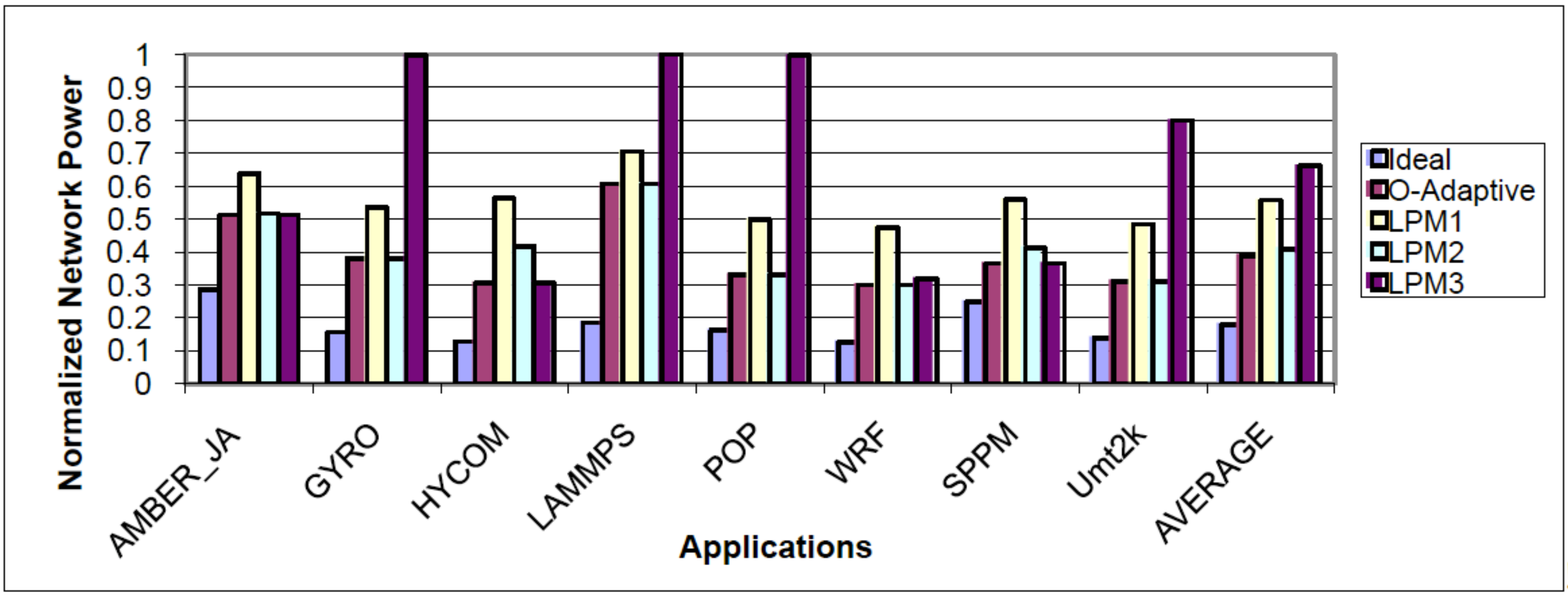
Processor	POWER5-like @ 2 GHz with caches, 45 nm, 120 W peak power
Switch	3-stage with buffers @ 2 GHz, 45 nm, 50 W peak power including link power
Level-1 link	4 @ 2 GB/s bidirectional, 3 ns
Level-2 link	2 @ 2 GB/s bidirectional, 6 ns
Level-3 link	4 @ 1 GB/s bidirectional, 30 ns
Control link	1/8 BW of its companion data link
Optical transceivers	3 W peak power
Memory/storage/etc	80 W peak power

Workloads

AMBER	The collective name for a suite of programs that carry out molecular dynamics simulations, particularly on biomolecules
GYRO	5D Eulerian gyro-kinetic-Maxwell (GKM) solver that computes the turbulent radial transport of particles and energy in tokamak plasma
HYCOM	The Hybrid Coordinate Ocean Model that implements a general circulation model of open ocean to shoreline regions
LAMMPS	Classical molecular dynamics code LAMMPS, which stands for Large-scale Atomic/Molecular Massively Parallel Simulator
POP	Parallel Ocean Program that solves the three-dimensional primitive equations for fluid motions in ocean circulation
SPPM	3D gas dynamics solver for a uniform Cartesian mesh, using a simplified version of the Piecewise Parabolic Method code
UMT2K	Benchmark of a class of computationally intensive application codes at Lawrence Livermore National Laboratory (LLNL)
WRF	Weather Research and Forecasting (WRF) modeling system

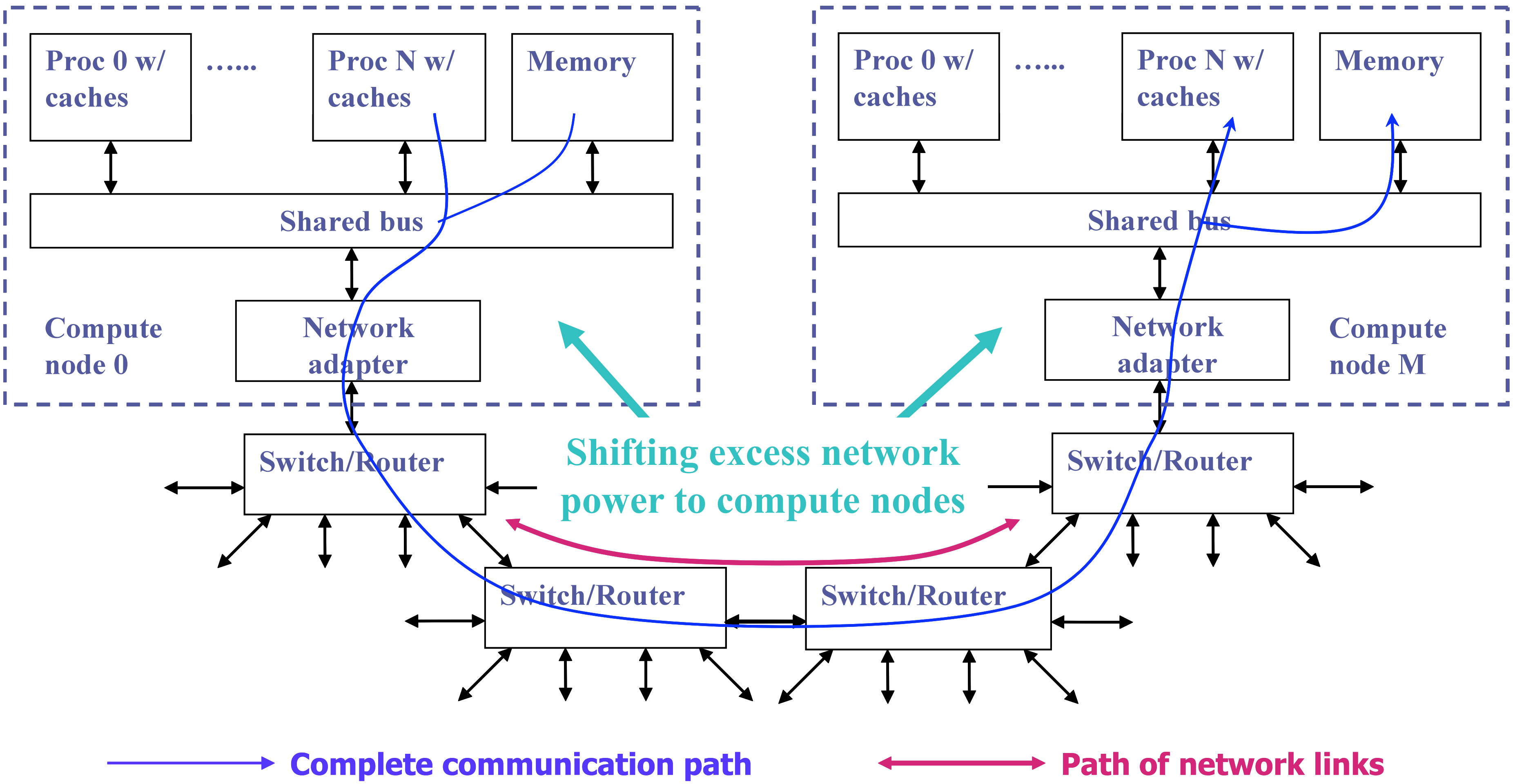
Average Power of Thrifty Interconnection Network

The lower the better



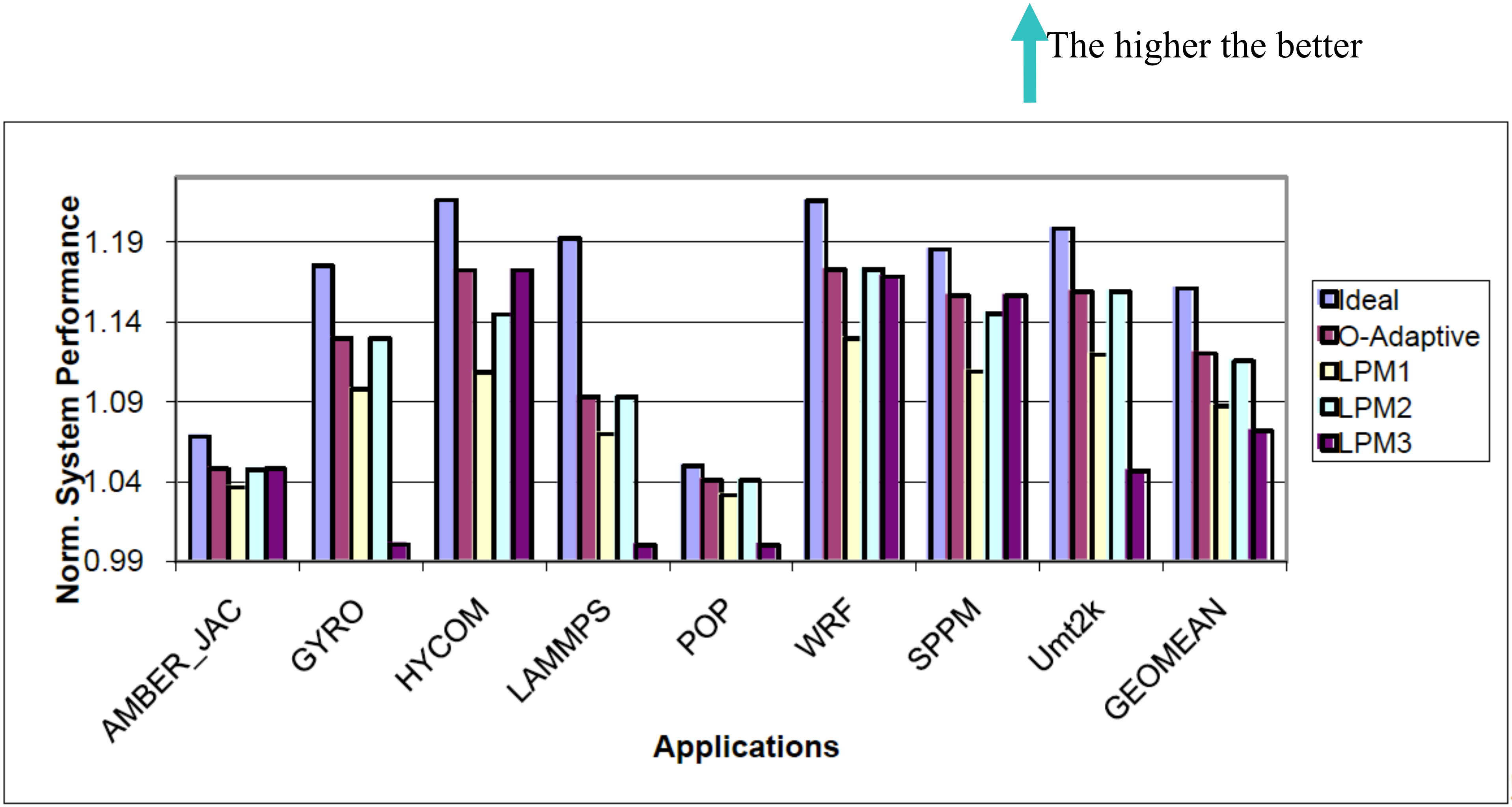
- 32 SMP nodes

Power Shifting



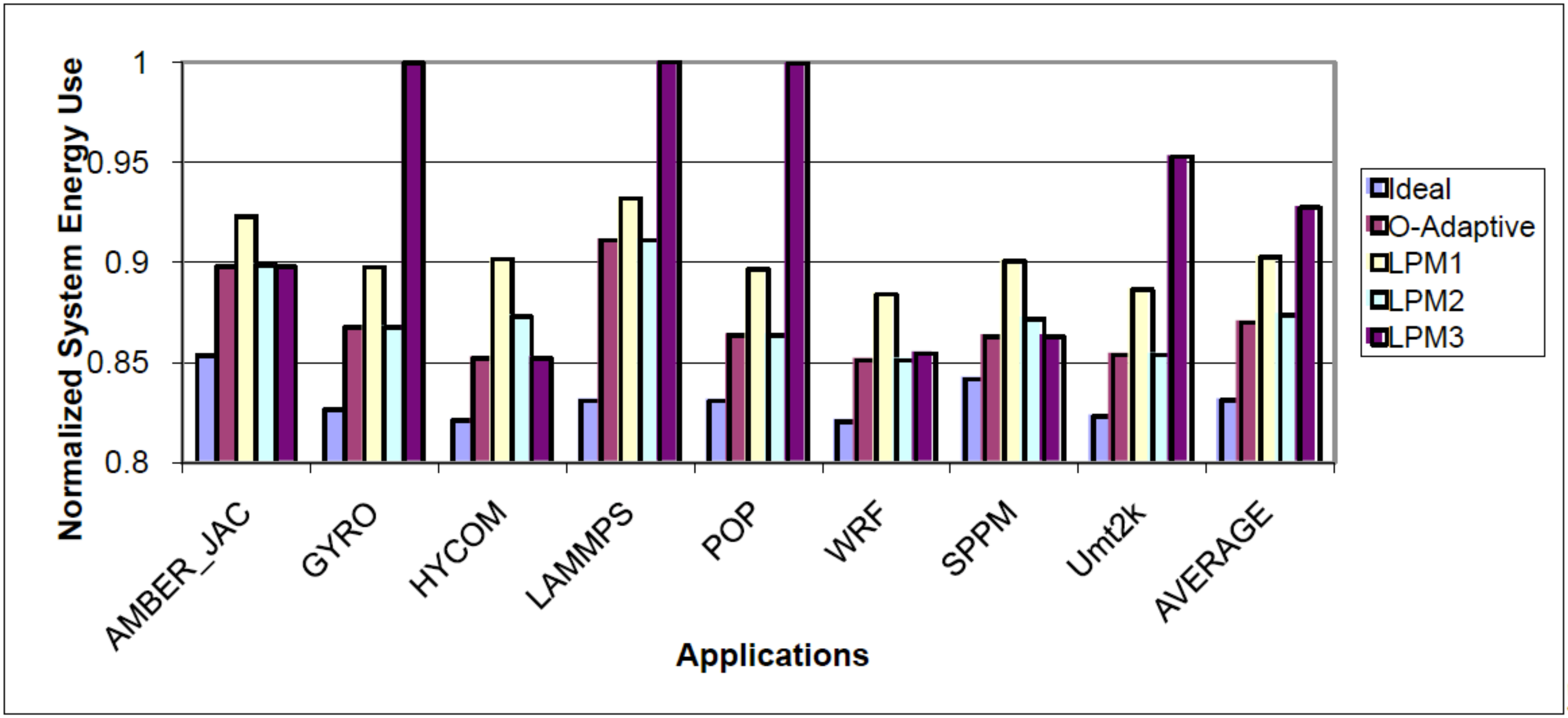
- Power shifting between compute nodes and their switch/Router and links
- 20% cap for frequency boost at compute nodes due to thermal concerns

System Performance Improvement

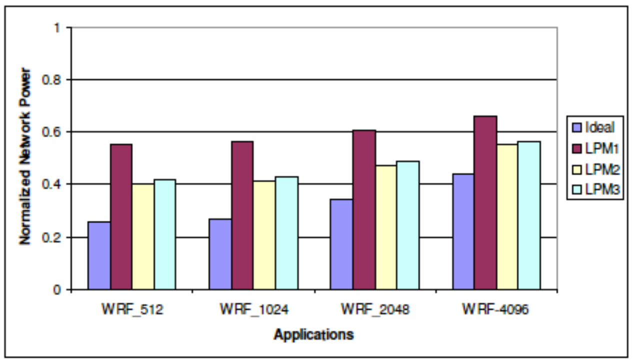


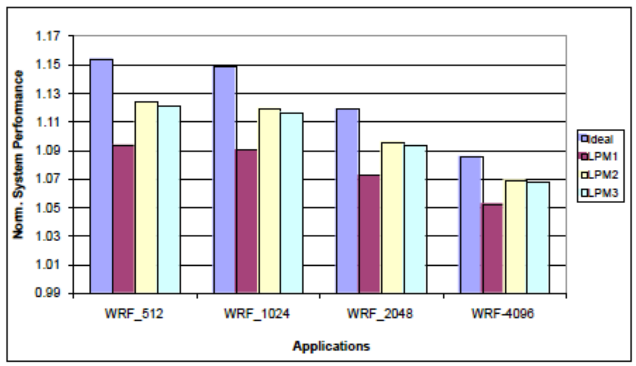
System Energy Improvement

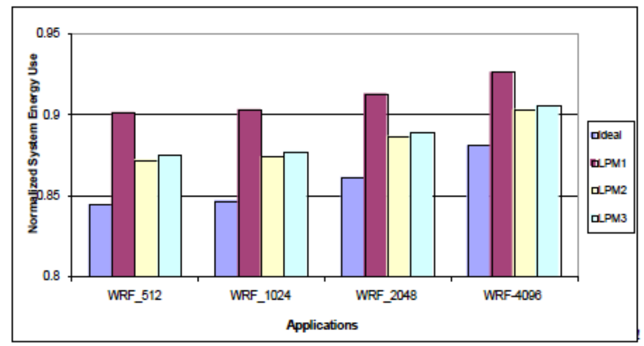
The lower the better



Scalability







- Lin-log scale
- Strong scaling of WRF with 512, 1024, 2048 and 4096 MPI tasks
    - Weak scaling would perform better (not shown)

Summary & Future work

- **Thrifty interconnection network**
  - Deterministic link activation and release service without prediction overhead
  - 60% average network power reduction, 12% performance speedup and 13% energy reduction for the studied workloads
- **Power shifting**
  - Dynamically shifts the total power budget between the compute nodes and the interconnection network that connect them
- **Future work**
  - Software-hardware co-design: <u>Robust</u> interaction with compiler/run-time, MPI, etc
  - <u>True</u> Power shifting in the network (ref: PowerRouting ASPLOS'10)
  - Leverage power and reliability in large-scale systems

Power Shifting in Thrifty Interconnect Networks

Jian Li
IBM Research - Austin
Email: jianli@us.ibm.com