Performance Prediction Models for Large-scale Interconnection Networks in HPC System

Presented by: **Kishwar Ahmed**

FIU | School of Computing & Information Sciences

Content

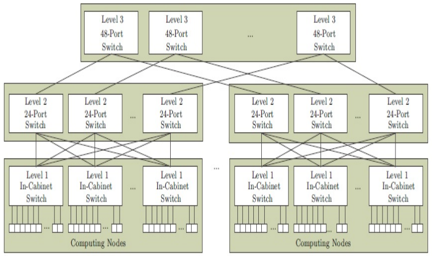
- Introduction
  - Motivation
  - Problem Statement and Our Contributions
- Performance Prediction Models in HPC System
- Conclusions

Why Performance Prediction in High Performance Computing (HPC) System?

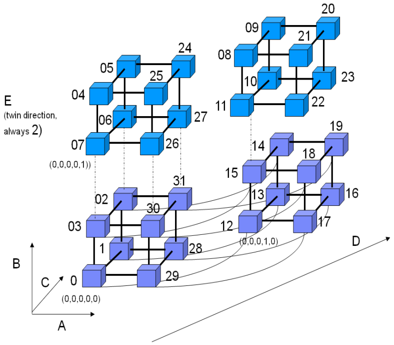
- Rapid changes in HPC architecture
  - E.g., introduction of many-core and multi-core architecture
- We are rapidly approaching towards exascale computing
  - New and advanced interconnect architecture to support high computation capacity
- Performance prediction facilitates
  - Evaluating design alternatives
  - Identifying performance issues

Interconnection Network Topology

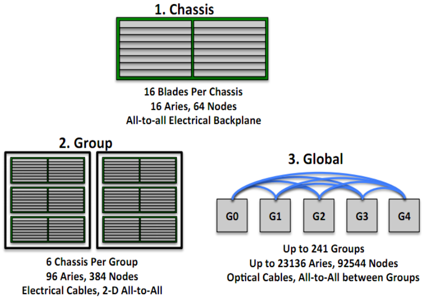
- **Interconnection network** specifies how to route data from
  - Processors to memory
  - One node (processor + memory) to another
- **Interconnect network topology**
  - Arrangement of nodes, switches
  - Affects routing, throughput, latency



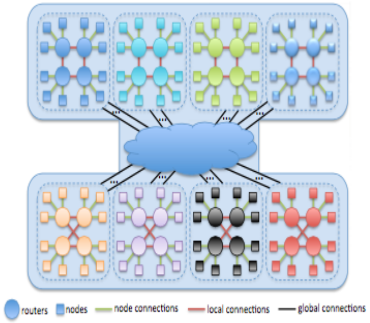
Fat-tree



Torus


1. Chassis
16 Blades Per Chassis
16 Aries, 64 Nodes
All-to-all Electrical Backplane
2. Group
6 Chassis Per Group
96 Aries, 384 Nodes
Electrical Cables, 2-D All-to-All
3. Global
G0
G1
G2
G3
G4
Up to 241 Groups
Up to 23136 Aries, 92544 Nodes
Optical Cables, All-to-All between Groups

Dragonfly

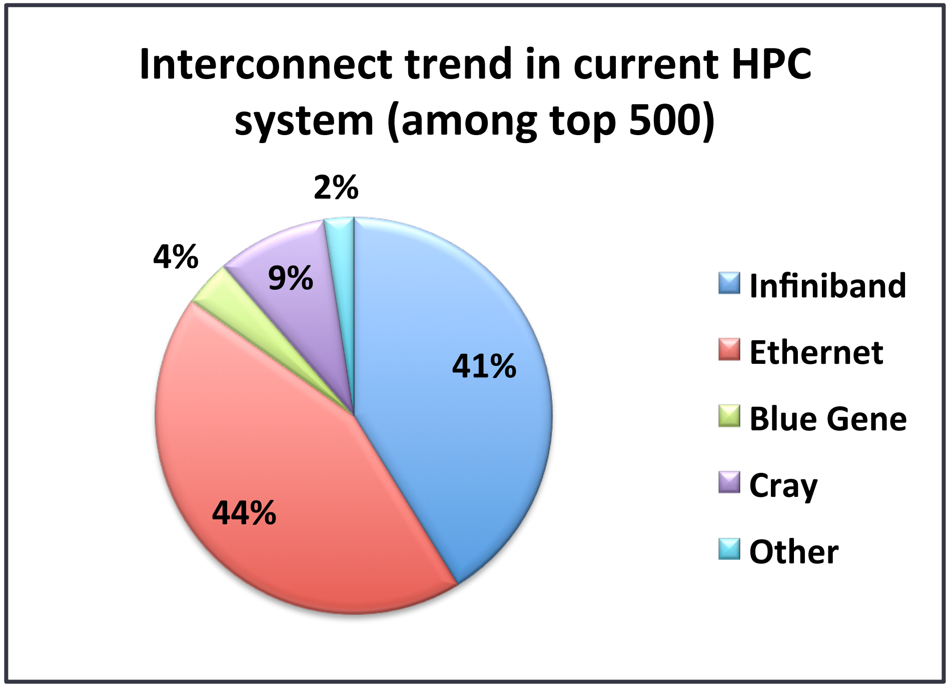


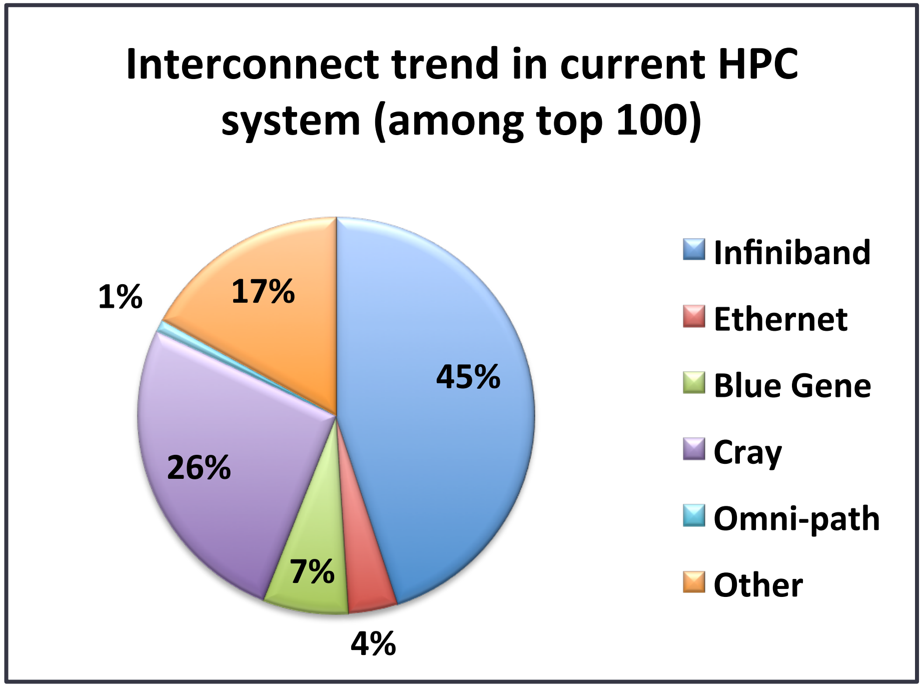
Slim Fly

And more…

Which Interconnect Topologies We Model?

- Dominant interconnection network topologies in current and future HPC systems: **Dragonfly**, **Fat-tree**, and **Torus**





Three topologies account for **54%** in **top 500**

Three topologies account for **82%** in **top 100**

Problem Statement and Solution

- Performance prediction of large-scale HPC system with following properties
  - **Accurate**: The prediction should produce accurate estimation of performance parameters (e.g., latency, bandwidth)
  - **Realistic**: The models should represent real-life implementation of the architecture
    - For example, Blue Gene/Q and Gemini for torus, Aries for dragonfly, Infiniband for fat-tree
  - **Applicable**: The models must be applicable for real-life HPC applications
- Our performance prediction models ensure all the three properties

We are here

- Introduction
- Performance Prediction Models in HPC System
  - Related Works
  - Background
  - Interconnection Network Models
  - Validation of Models and Results
- Conclusions

Related Works

- Performance prediction in large-scale interconnect
  - BigSim [Geng04]: *early efforts* for large-scale performance prediction
  - Structural Simulation Toolkit (SST) [Arun11]: a *comprehensive framework* for modeling large-scale HPC system
  - Co-Design of Exascale Storage System (CODES): torus [Ning11], dragonfly [Misbah14], fat-tree [Ning15])

- How our work differs:
  - Our interconnection network models reflect **accurate** and **actual** implementations of interconnect topologies (e.g., Aries, Blue Gene/Q, Infiniband)
    - We can study various interconnection networks of real (either existing or planned) HPC system.
  - We can model **real-life** scientific applications
    - e.g., SNAPSim using Edison supercomputer interconnect

A PDES Engine: Simian

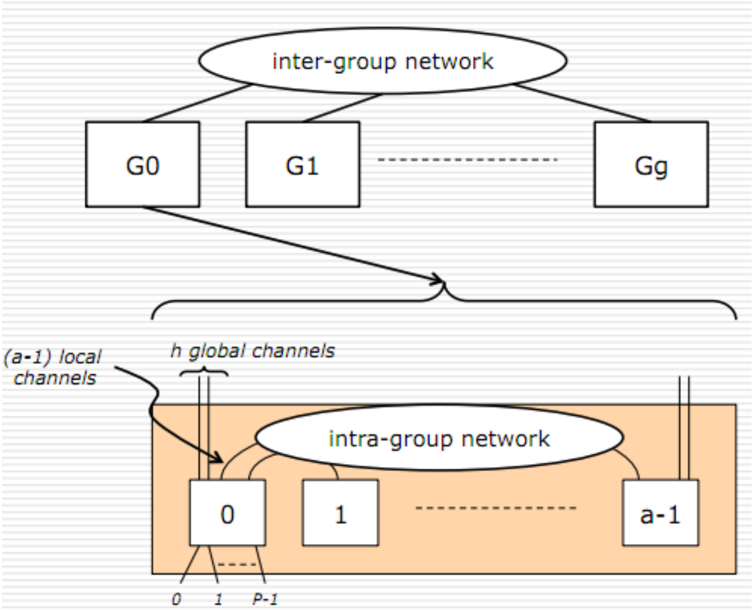
- An open-source, process-oriented parallel discrete-event (PDES) engine
- Distinct features
  - A minimalistic design (only around 500 lines of code)
  - Minimal dependency to third-party libraries
  - A very simplistic application programming interface (API)

MPI Models

- Message Passing Interface (MPI)
  - One of the most popular parallel programming tools on HPC platform
- We used different MPI functions to perform communication among nodes
  - Point-to-point (e.g., MPI_Send, MPI_Recv)
  - Collective (e.g., MPI_Bcast, MPI_Reduce)
  - Group and collective operations (e.g., MPI_Comm_dup, MPI_Group_size)

Dragonfly Topology

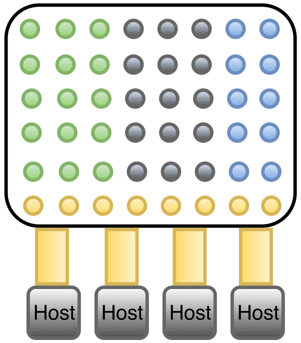
- A cost-efficient topology
  - Exploits the economical, optical signaling technologies for long distance communication
  - High-radix (virtual) router

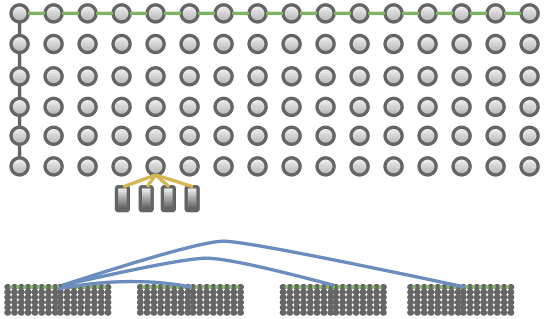


Kim, John, et al. "Technology-driven, highly-scalable dragonfly topology." *ACM SIGARCH Computer Architecture News*. Vol. 36. No. 3. IEEE Computer Society, 2008.

Cray’s Aries Interconnect

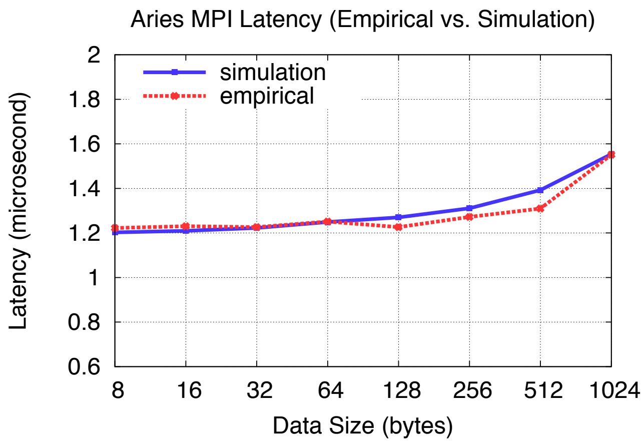
- Used by many supercomputers as interconnect architecture
- Uses dragonfly topology
- Consists of cabinets
  - Two cabinets per group
  - Three chassis per cabinets
  - Six chassis per group
  - Sixteen Aries blades per chassis





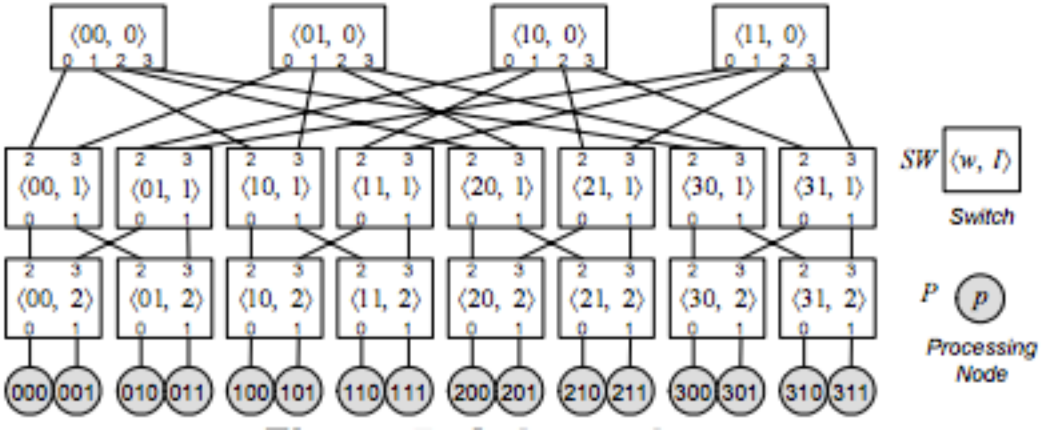
Aries Interconnect (Validation#1)

- Trinity@LANL
  - Ranked 7th in Top500 list
  - Consists of 9436 nodes and 301,952 cores
  - Uses a Cray XC40 system
    - Nodes connected via Aries interconnect
- We measured
  - Average end-to-end latency
- Compared
  - With empirical results
  - In general, close resemblance



Fat-tree Model

- Fat-tree widely-used in HPC clusters and data center networks
- Many popular variations of fat-tree topologies
  - m-port n-tree, k-ary n-tree
  - We used the **m-port n-tree** in our work

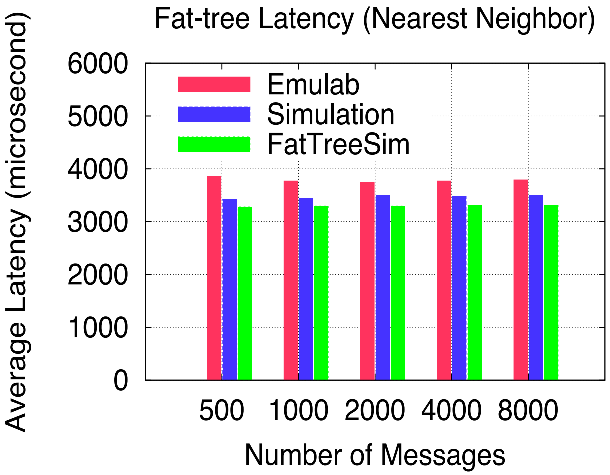


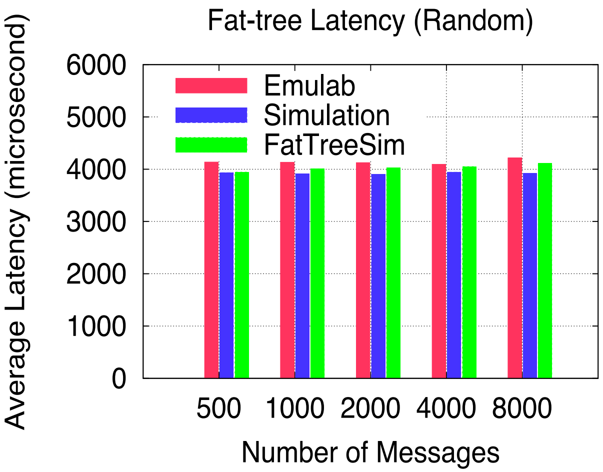
A 4-port 3-tree

A Multiple LID Routing Scheme for Fat-Tree-Based InfiniBand Networks, Xuan-Yi Lin, Yeh-Ching Chung, and Tai-Yi Huang

Validation: Fat-tree Model

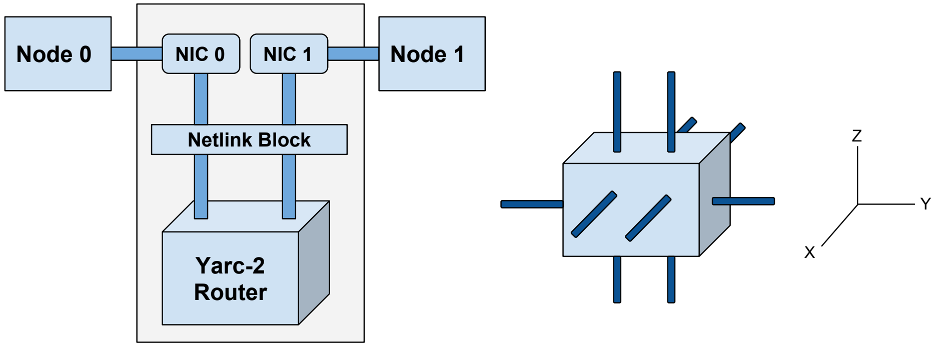
- Stampede@TACC
  - Ranked 12th in top500 list
  - Consists of 6,400 nodes connected via fat-tree-based Infiniband FDR network
- Validated our model with a recently-proposed fat-tree simulator: FatTreeSim
- Similar setup used as in FatTreeSim and Emulab





3-D Torus (Gemini) Model

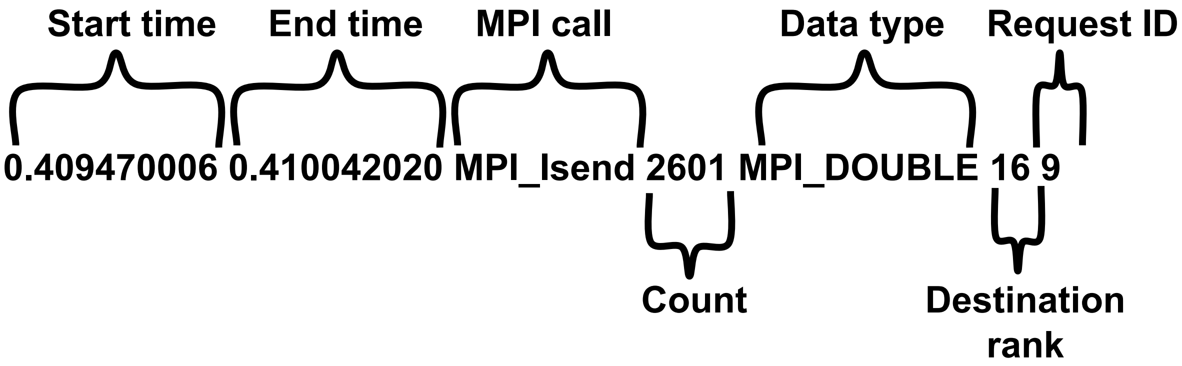
- Cray’s XE6 system uses 3-D torus-based Gemini architecture
- In Gemini, each Application-Specific Integrated Circuit (ASIC) contains
  - Two AMD Opteron nodes
  - 48-port YARC router
  - Each router gives
    - Ten torus connections
    - Two connections per direction in the “X” and “Z” dimension
    - One connection per direction in the “Y direction”



- We validated Gemini
  - Using Hopper@NERSC

DOE Communication Traces

- Application communication traces provided by the NERSC
- We use the open-source SST DUMPI toolkit to process the traces



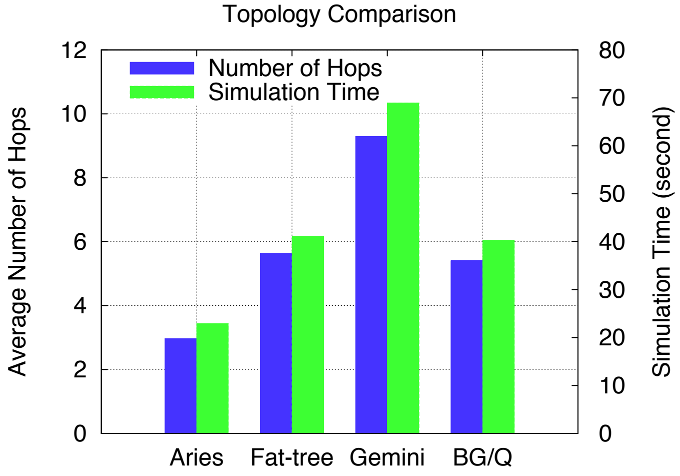
0.409470006 0.410042020 MPI_Isend 2601 MPI_DOUBLE 16 9

Format of MPI calls in the processed trace file
(there is one trace file for each MPI rank)

A Comparative Study (Contd.)

- Configurations:
  - Aries: Trinity@LANL
  - Fat-tree: Stampede@TACC
  - Gemini: Hopper@NERSC
    - 6,834 nodes connected via Gemini interconnect at 17X8X24
  - Blue Gene/Q: Mira@ANL
    - 49,152 nodes connected via Blue Gene/Q at 8X12X16X16X2
- Results:
  - Aries has the minimum # of hops
  - Gemini has the maximum # of hops
  - Simulation time is consistent

  with # of hops traversal



Conclusions

- We presented performance prediction models for all the major interconnection network topologies in HPC system
- Designed real-life interconnection architectures based on the interconnect topologies
- Validated the accuracy of the models with existing and planned HPC architectures
- Our prediction models are capable of running real-life communication and scientific applications

References

- [Arun11] Arun F Rodrigues, K Scott Hemmert, Brian W Barrett, Chad Kersey, Ron Oldfield, Marlo Weston, R Risen, Jeanine Cook, Paul Rosenfeld, E CooperBalls, et al. The structural simulation toolkit. ACM SIGMETRICS 2011
- [Geng04] Gengbin Zheng, Gunavardhan Kakulapati, and Laxmikant V Kal ́e. Bigsim: A parallel sim- ulator for performance prediction of extremely large parallel machines. IPDPS 2004
- [Ming12] Ming-yu Hsieh, Rolf Riesen, Kevin Thompson, William Song, and Arun Rodrigues. Sst: A scalable parallel framework for architecture-level performance, power, area and thermal simulation. The Computer Journal, 2012
- [Misbah14] Misbah Mubarak, Christopher D Carothers, Robert B Ross, and Philip Carns. Using massively parallel simulation for MPI collective communication modeling in extreme-scale networks, WSC 2014
- [Ning11] Ning Liu and Christopher D Carothers. Modeling billion-node torus networks using massively parallel discrete-event simulation. IEEE PADS 2011
- [Ning15] Ning Liu, Adnan Haider, Xian-He Sun, and Dong Jin. FatTreeSim: Modeling large-scale fat-tree networks for HPC systems and data centers using parallel and discrete event simulation, ACM SIGSIM PADS 2015

Thank you! Questions?