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

Presented by: Kishwar Ahmed
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

And more…

Fat-tree

Torus

Dragonfly

Slim Fly
Which Interconnect Topologies We Model?

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

Interconnect trend in current HPC system (among top 500)

- Infiniband: 41%
- Ethernet: 44%
- Blue Gene: 9%
- Cray: 2%
- Other: 4%

Interconnect trend in current HPC system (among top 100)

- Infiniband: 45%
- Ethernet: 26%
- Blue Gene: 17%
- Cray: 7%
- Omni-path: 4%
- Other: 1%

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

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 7\textsuperscript{th} 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

![Aries MPI Latency (Empirical vs. Simulation)](chart.png)
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 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 12\textsuperscript{th} 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

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


- [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?