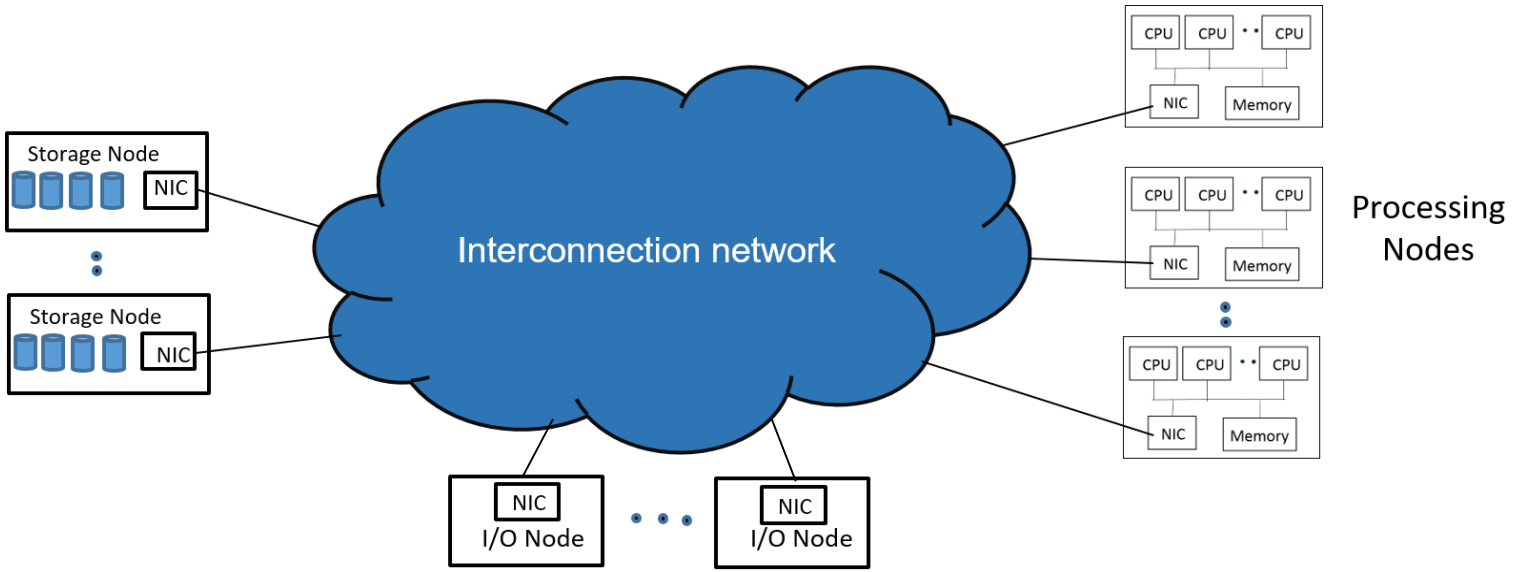
Modeling and Comparison of Large-Scale Interconnect Designs

Md Atiqul Mollah(mollah@cs.fsu.edu)

Advisor: Xin Yuan

Florida State University

The Big Picture



- For communication intensive applications, Interconnect is a major performance bottleneck of parallel HPC systems.
- Future generation HPC systems need highly scalable high-performance interconnects

Dissertation Goal

- Devise and evaluate new designs for HPC interconnects
- Measure performance of large-scale interconnects through modeling and simulation
- Investigate the key performance components of interconnect design
  - topology, routing, network control

Topologies Investigated

- Current-generation Interconnects:
  - Fat-tree
  - Dragonfly
- Up-and-coming:
  - Slimfly
  - De Bruijn Graph
  - Random Regular Graph

Top500 Rank (June 2017)	System	Topology
1	Sunway TaihuLight	Fat-tree
2	Tianhe-2 (MilkyWay-2)	Fat-tree
3	Piz Daint	Dragonfly
4	Titan	3D Torus
5	Sequoia	5D Torus
6	Cori	Dragonfly
7	Oakforest-PACS	Fat-tree
8	K computer	6D torus
9	Mira	5D Torus
10	Trinity	Dragonfly

Selected Contributions

1. Rapid Calculation of Max-Min Fair Rates for Multi-commodity Flows in Fat-Tree Networks (Cluster 2015, TPDS preprint)
2. Random Regular Graph and Generalized De Bruijn Graph with k-Shortest Path Routing(IPDPS 2016, TPDS preprint)
3. A Comparative Study of SDN and Adaptive Routing on Dragonfly Networks (SC 17)
4. Modeling UGAL on the Dragonfly Topology(Submitted work)
5. Study of Interconnect design approaches(Ongoing work)

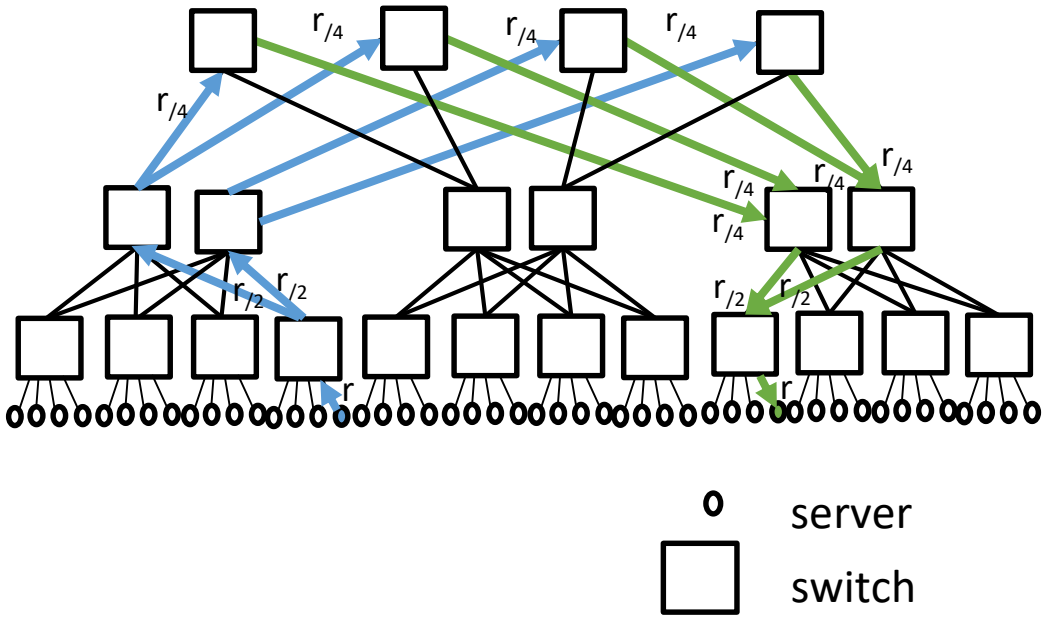
Fast Calculation of Max-Min Fair rates on Fat trees

Max-Min Fair (MMF) Rate Allocation

- Max-Min fairness(MMF) is a well known fairness metric.
- MMF rate allocation problem is to optimize the *fair* bandwidth rate for all traffic flows in a given network
  - No Routing scheme specified a priori
- Desirable to achieve, but computationally expensive
  - Requires solving Linear Programming problems iteratively
- On arbitrary network with *E* links and *D* flows, complexity is: O(min(E,D) * LP(2D+E,D*E))

Fat-tree Network

- Widely used in HPC and data centers
- Hierarchical, well-defined tree structure
- Servers on leaf node, routers on non-leaf nodes
- Uniform All-shortest Path Routing(UAPR):Flow bandwidth gets uniformly shared to all available shortest paths



MMF-MCF on fat trees

- In a fat tree, If a rate allocation is feasible for any routing scheme, it is also feasible using UAPR
- Therefore, solving MMF-MCF for fat trees is equivalent to solving MMF rate with known optimal routing UAPR
- Based on the above property, 3 new Algorithms are developed to calculate optimal MMF on fat-trees.

1. Simplified LP formulation(LP)	**O(min(E,F) * LP(E, 1))**
2. Progressive filling approach(PF)	**$O(E * D^2)$**
3. Optimized progressive filling(OPT)	**O((E+D)*D)**

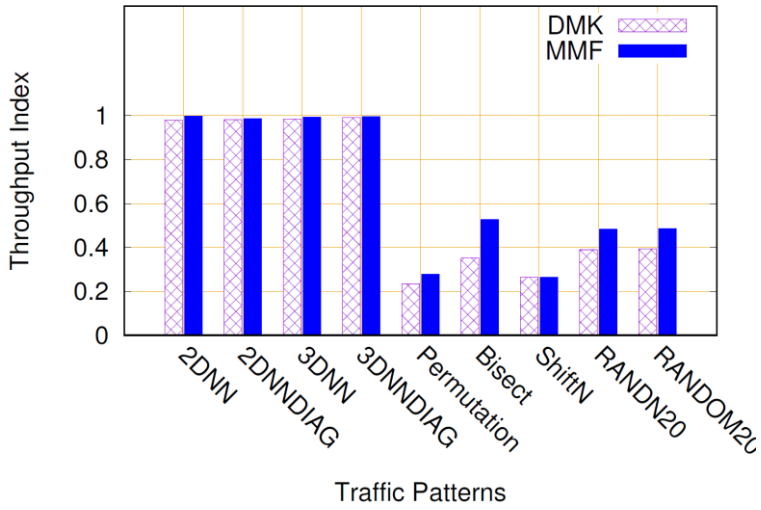
Performance Evaluation

Topology	Pattern	Average Execution Time				
		GEN	*LP*	*PF*	*OPT*	*DMK*
$XGFT(2;12,24;1,12)$ (288 proc. nodes)	*Perm.*	>30h	0.593s	0.150s	0.003s	0.096s
	2DNN	>30h	1.228s	0.145s	0.002s	0.002s
	RANDN(20)	>30h	127.618s	0.592s	0.033s	0.093s
$XGFT(2;18,36;1,18)$ (648 proc nodes)	*Perm.*	>30h	1.016s	0.348s	0.007s	0.513s
	2DNN	>30h	3.606s	0.324s	0.006s	0.005s
	RANDN(20)	>30h	323.986s	3.672s	0.007s	0.532s
$XGFT(3;12,12,24;1,12,12)$ (3,456 proc. nodes)	*Perm.*	>30h	27.222s	4.172s	0.031s	0.326s
	2DNN	>30h	7.783s	2.278s	0.024s	0.027s
	RANDN(20)	>30h	30,080.077s	3,717.002s	10.189s	29.63s
$XGFT(3;18,18,36;1,1818)$ (11,664 proc. nodes)	*Perm.*	>30h	230.870s	26.151s	0.150s	1.083s
	2DNN	>30h	38.447s	13.532s	0.089s	0.134s
	RANDN(20)	>30h	>30h	96,730.001s	310.650s	782.923s

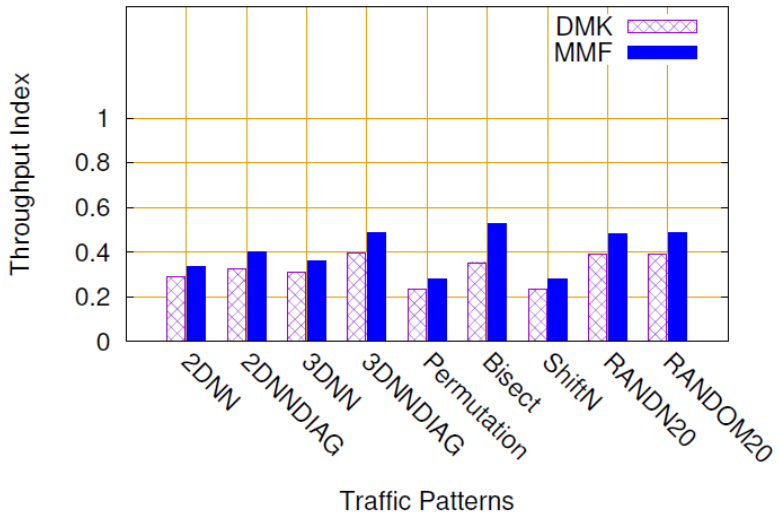
GEN: General Approach
DMK: D-mod-k Routing(single path, not optimal)

Perm. : Random Permutation Traffic Pattern
2DNN: 2D Nearest Neighbor communication pattern

Application: Routing scheme Evaluation



a) Direct Node mapping



b) Random Node mapping

MMF throughput of d-mod-k routing(DMK) and Optimal routing(MMF) on a 20,736 node fat-tree

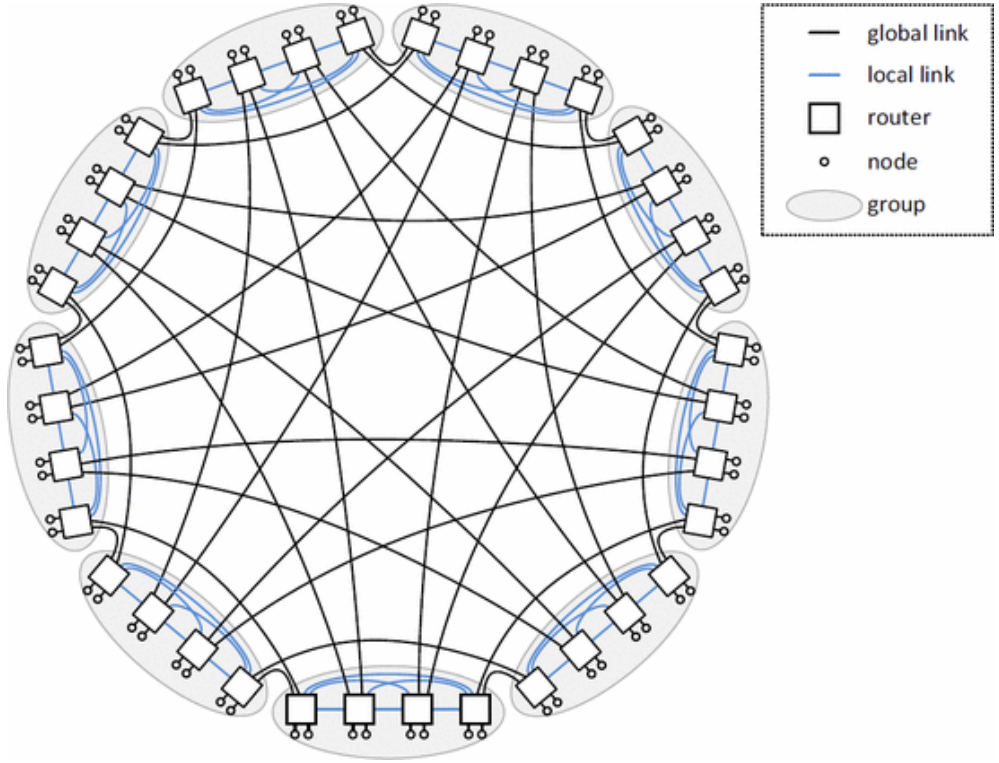
Summary

- In a fat-tree network, Uniform All-shortest Path Routing(UAPR) is the optimal routing scheme that yields max-min fair(MMF) throughput.
- Using newly developed algorithms, MMF rates of a fat-tree network can be rapidly calculated.
- Useful to measure performances of current fat-tree routing schemes.

Modeling UGAL Routing on the Dragonfly Topology

The Dragonfly Network

- Scalable, low-cost topology for connecting thousands of HPC nodes
- Dragonfly uses Universal Globally Adaptive Load-balancing (UGAL) Routing
- I analyze the performance characteristics of UGAL by constructing throughput model based on UGAL features



72 node dragonfly with fully connected inter- and intra-group
By M. García *et al.*, "On-the-Fly Adaptive Routing in High-Radix Hierarchical Networks," ICPP*2012*

UGAL Features

- **Feature 1(limited paths):** UGAL considers only 2 types of paths for routing:
  1) Minimal(MIN) hop count paths
  2) Valiant Load-balancing(VLB) paths where traffic is detoured to a random node first, and then forwarded to destination.

- **Feature 2(random path control):** UGAL uniform-randomly selects a small number of MIN and VLB paths as candidate paths for each packet.

- **Feature 3(path length bias):** Among the randomly chosen paths, UGAL implicitly differentiates paths of different lengths.
  - Biased towards picking shorter paths

Modeling UGAL

- Find the Maximum Concurrent Flow(MCF) of the Dragonfly network
  - -The bandwidth at which ALL flows can inject traffic
  - -First-order approximation of MMF
- Express as an LP optimization problem
- Model the three UGAL features(limited path, random path control, path length bias) in different extents

Model No. 0

Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

- Apply feature 1 to Both MIN and VLB paths

- Each MIN and VLB path of a flow have individual rate allocation

- Given:

  F =traffic pattern/set of flows

  E = set of links

  x_d = bandwidth used by d, $d \in F$

  P_d= set of all MIN and VLB paths, $d \in F$

  $C\ (e)$= Link capacity function

- Maximize α
- Subject to:

$$\alpha - x_d \leq 0\ d \in F \quad \forall d \in F \tag{1}$$

$$x_d = x_d^1 + x_d^2 + \ \ldots + x_d^{|P_d|} \ \forall d \in F \tag{2}$$

$$\sum_{,p \in P_d, d \in F, p\ uses\ link\ e} x_d^p \leq C(e)\ \forall e \in E \tag{3}$$

Model No. 1

Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

- Apply feature 1 to MIN, feature 2 to VLB paths
- Each MIN have individual rate allocation
- All VLB paths of same length of a flow have the same rate allocation

- Given:

F =traffic pattern/set of flows

E = set of links

x_d = bandwidth used by d, $d \in F$

$P_{d,MIN}$= set of all MIN paths, $d \in F$

$P_{d,MIN}(e)$= set of all MIN paths using link e, $d \in F, e \in E$

$P^L_{d,VLB}(e)$= set of all L-hop VLB paths using link e, $d \in F, e \in E$

C (e)= Link capacity function

H=maximum path hop length

Maximize α

Subject to:

$$\alpha - x_d \leq 0 \; d \in F \quad \forall d \in F \tag{1}$$

$$x_d = x_d^1 + x_d^2 + \ldots + x_d^{|P_{d,MIN}|} + x_d^{VLB,1} + x_d^{VLB,2} + \ldots + x_d^{VLB,H}, \forall d \in F \tag{2}$$

$$\sum_{p \in P_{d,MIN}(e), d \in F} x_d^p + \sum_{d \in F, 0 < L \leq H, P^L_{d,VLB}(e) \neq \emptyset} |P^L_{d,VLB}(e)| \times x_d^{VLB,L} \leq C(e) \; \forall e \in E \tag{3}$$

Models Summary

- Model No. 0
  - All MIN and VLB paths have individual rates
- Model No. 1
  - All VLB paths of same length of a flow have the same rate
- Model No. 2
  - All VLB paths of a flow have the same rate
- Model No. 3:
  - All MIN paths of same length of a flow have the same rate
  - All VLB paths of same length of a flow have the same rate
- Model No. 4:
  - All MIN paths of same length of a flow have the same rate
  - All VLB paths of a flow have the same rate
- Model No. 5:
  - All MIN paths of a flow have the same rate
  - All VLB paths of a flow have the same rate

Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

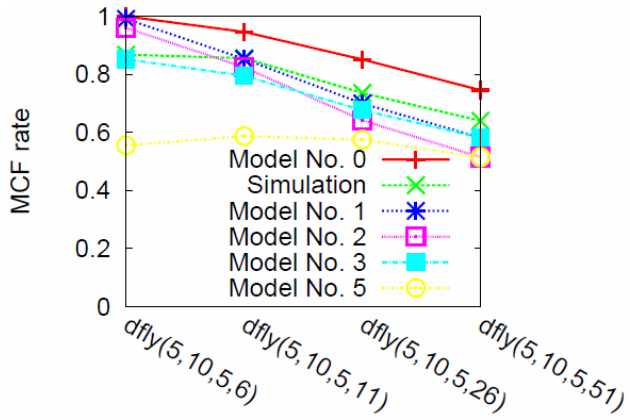
Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

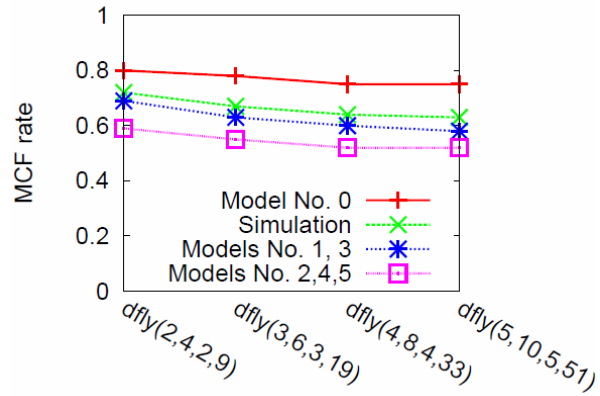
Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

Feature 1(limited paths)
Feature 2(random path control)
Feature 3(length bias)

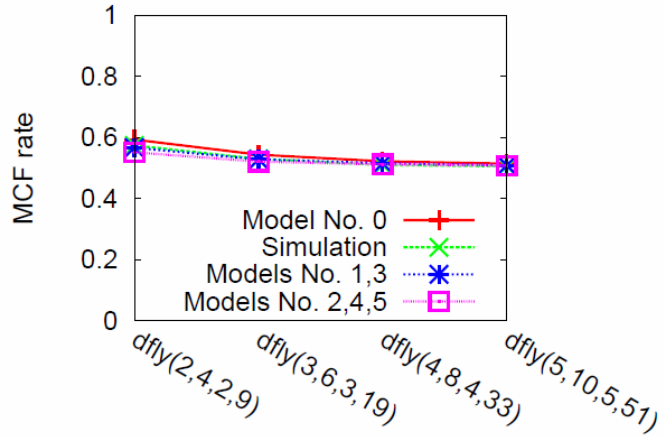
Model validation

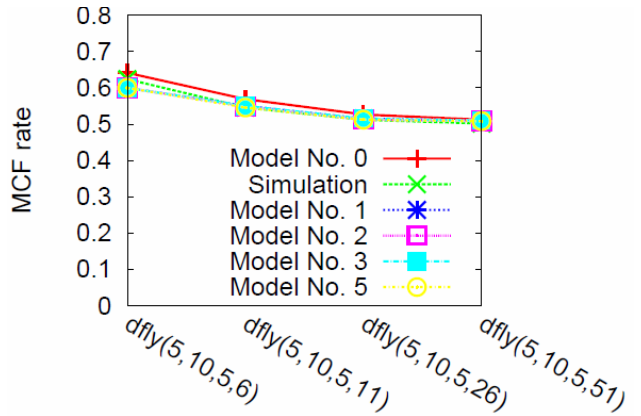
- Formulated 6 LP models for different Dragonfly topologies
- Simulated UGAL-G on same topologies in Booksim* packet-level Simulator
- UGAL-G best approximated by Model No. 3
- Model No. 0 gives a performance upper bound of UGAL





Random Permutation Traffic





Adversarial Traffic

Thank You!