Modeling UGAL routing on the Dragonfly topology

Md Atiqul Mollah, Peyman Faizian,
Md Shafayat Rahman, Xin Yuan

Florida State University

Scott Pakin, Michael Lang
Los Alamos National Laboratory

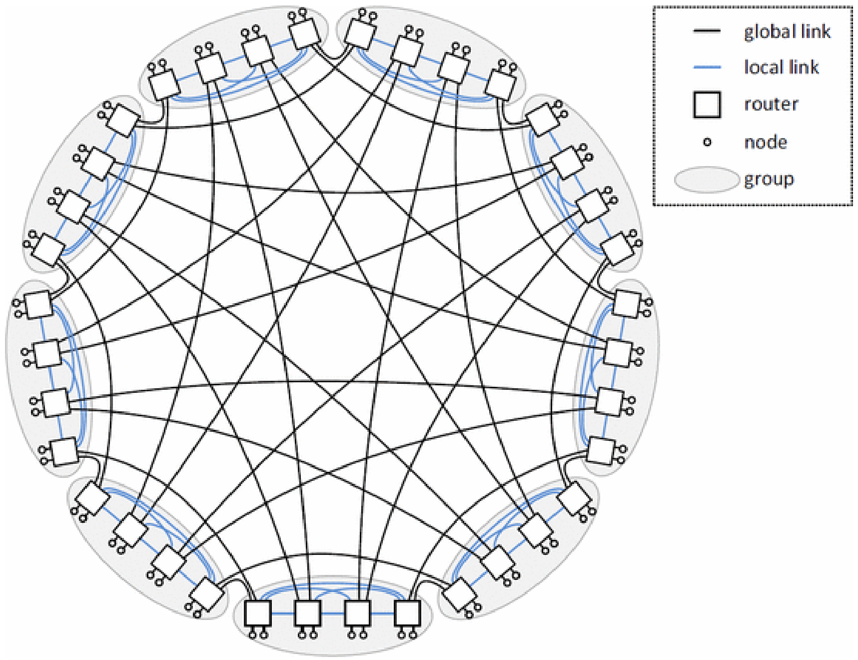
Motivation

- Interconnect is a vital part of modern day HPC systems
  - potential performance bottleneck of the entire system
  - especially on exascale(or even near-exascale)

- Important to measure interconnect performance while designing
  - through modeling and simulation
  - Modeling provides a holistic view
  - Simulation provides a more component-level view
  - In this work, we focus on modeling

Dragonfly Topology

- Used in current generation interconnects
- Scalable, cost-efficient design
- Used in Cray® Cascade system/XC® series
- In TOP500*:
  - Piz Daint(#3), Cori(#6), Trinity(#10)
  - 28 in the top 100

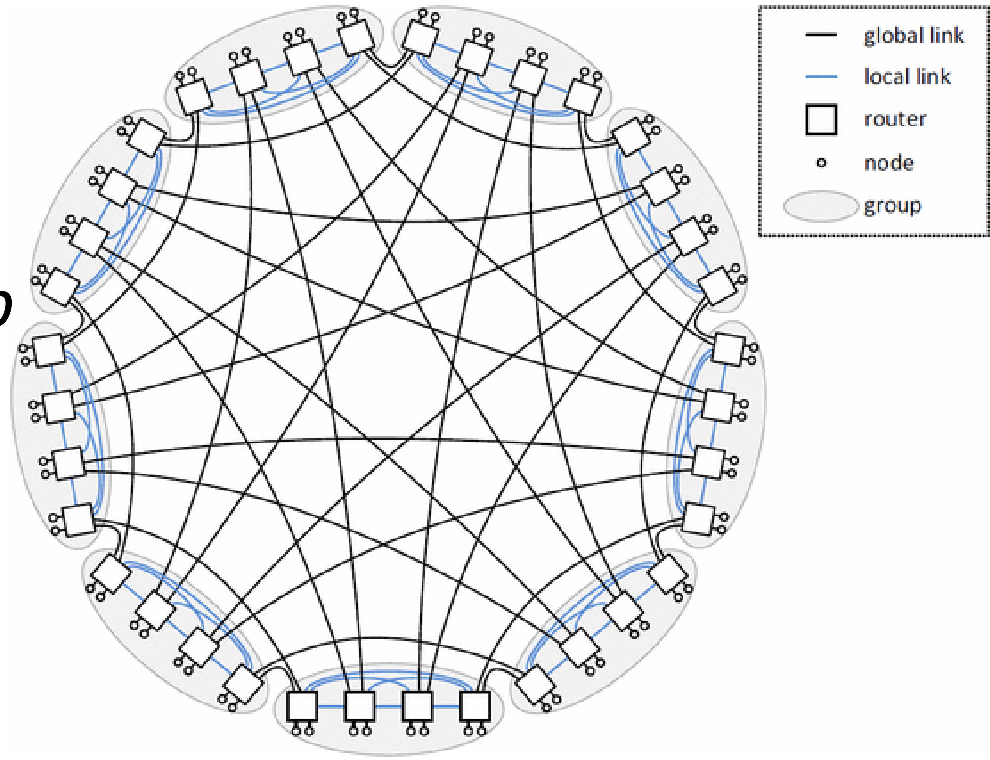


72 node dragonfly with fully connected inter- and intra-group
By M. García *et al.*, "On-the-Fly Adaptive Routing in High-Radix Hierarchical Networks," ICPP*2012*

* https://www.top500.org/list/2017/06/

Dragonfly Construction

- 2-level hierarchical design
- Local interconnection of links forms a *group*
  - topology of choice
  - each group imitates a high-radix router
- Fully connected inter-group topology
  - using long *global* links

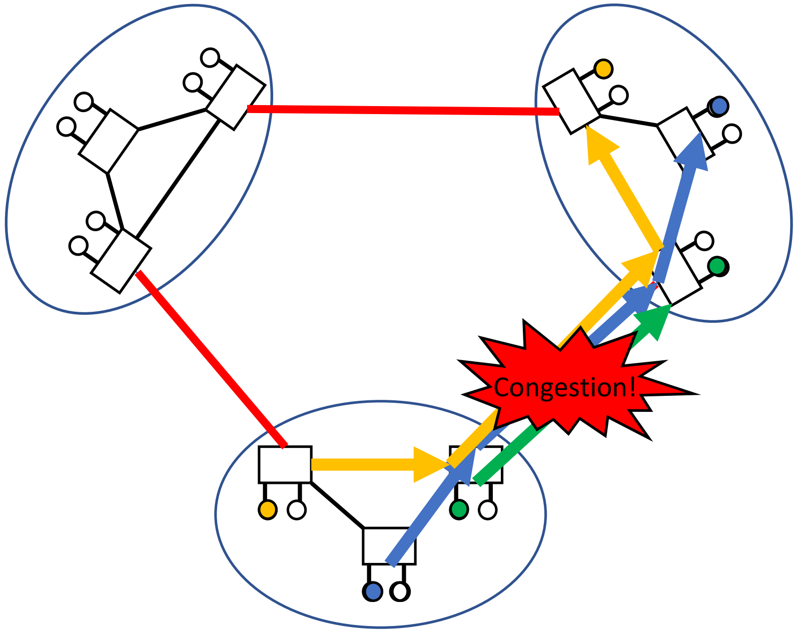


72 node dragonfly with fully connected inter- and intra-group
By M. García *et al.*, "On-the-Fly Adaptive Routing in High-Radix Hierarchical Networks," ICPP*2012*

* https://www.top500.org/list/2017/06/

Dragonfly Routing

- Minimal Path Routing (MIN)
  - Local routing in source group
  - Global link hop
  - Local routing in destination group

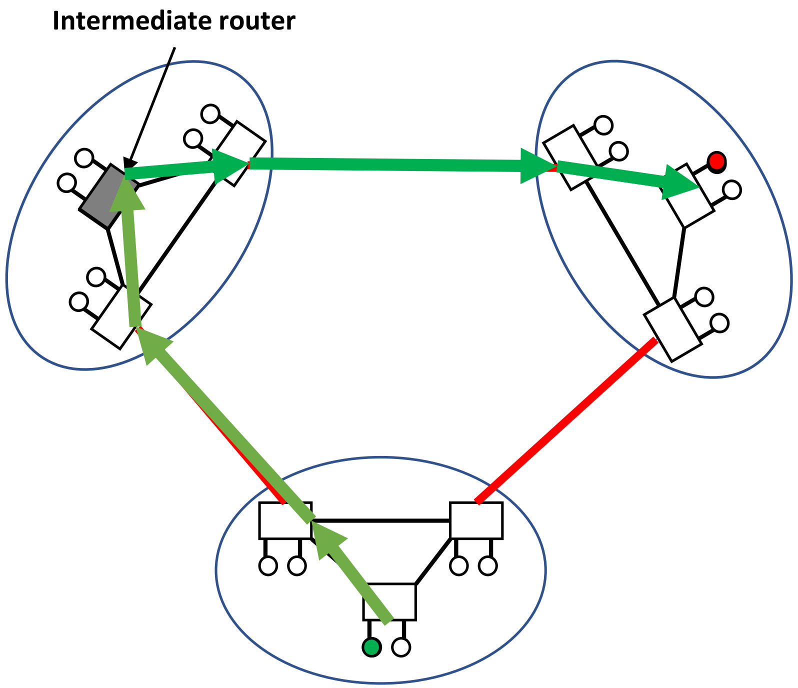


- Performs well for *benign*(e.g. uniform) traffic

- Dragonfly has limited MIN path diversity
  - canonical design has only 1 MIN path per node pair
  - leads to bottleneck on certain *adversarial* traffic patterns

Dragonfly Routing

- Valiant's Load-balancing (VLB)
  - Choose intermediate(intm.) router randomly
  - MIN route from source to intm.
  - MIN route from intm. to dest.



- VLB diffuses any bottleneck traffic
  - High path diversity
  - Also high end-to-end latency(2 times that of MIN)

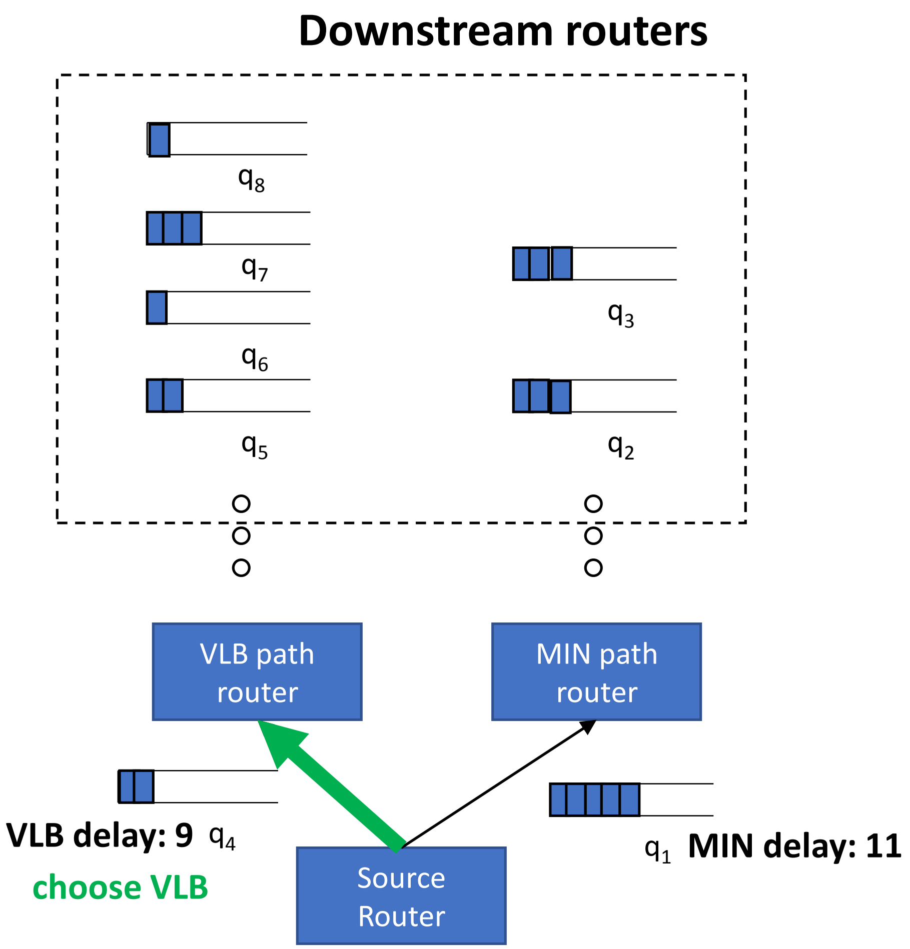
Dragonfly Adaptive Routing

- A single routing scheme(MIN/VLB) does not suit all traffic patterns

- Adaptive routing combines the benefit of both schemes.
  - Choose between a MIN and a VLB path based on traffic condition
  - Routing decision taken on-the-fly for each packet

Universal Globally Adaptive Load-balanced(UGAL) Routing

- For each packet, Pick one MIN and one VLB randomly
- From the two, choose path with minimum estimated latency
- Obtained from router queue length information
- Performs well for both benign and adversarial traffic patterns



Characterizing UGAL

- Why does UGAL perform so well?
  - just a greedy heuristic to maximize network throughput
  - Only a small subset of dragonfly designs have been studied
  - Lacking formal analysis

- How close is UGAL performance to its upper bound?

- Can other routing schemes perform better than UGAL?

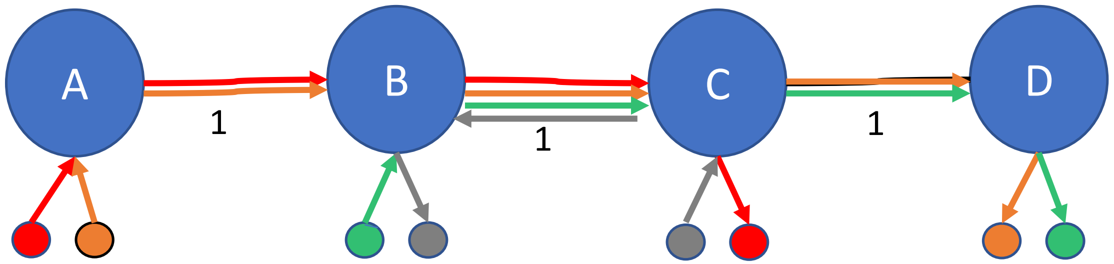
Modeling UGAL for dragonfly topology

1. Model the throughput optimization problem
2. Identify distinguishing features of UGAL
3. Modify throughput optimization model based on UGAL characteristics
4. Verify model by comparing to simulation results

1. Modeling Throughput optimization

- Given a traffic pattern, we find the Maximum Concurrent Flow(MCF) rate
  - MCF is the bandwidth at which **ALL** communications can inject traffic
  - In other words, it is the ***guaranteed throughput*** for any communication

- Example: For traffic pattern A->C, A->D, B->D, C->B
  - MCF = max$\{\frac{capacity(l)}{load(l)} : l \in E\}$
  - = 0.3333



¥ Farhad Shahrokhi and D. W. Matula. 1990. The maximum concurrent flow problem. J. ACM 37, 2 (April 1990), 318-334. DOI=http://dx.doi.org/10.1145/77600.77620

LP formulation for MCF ¥ [Shahrokhi et al. '90]

Given

F =traffic pattern/set of flows

E = set of links

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

P_d= *set of all paths,* $d \in F$

P_d*(e) = paths using link e,* $d \in F, e \in E$

C *(e)= Link capacity function,* $e \in E$

Maximize

Concurrent flow rate α

Subject to

$$\alpha - x_d = 0 \qquad \forall \text{ d} \in F$$

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

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

¥ Farhad Shahrokhi and D. W. Matula. 1990. The maximum concurrent flow problem. J. ACM 37, 2 (April 1990), 318-334. DOI=http://dx.doi.org/10.1145/77600.77620

2. UGAL Features

- Feature 1: UGAL considers all MIN and VLB paths
  - Instead of all possible paths
- Feature 2: UGAL randomly selects a small number of MIN and VLB paths as candidate paths for each packet.
  - All paths equally likely to be selected
- Feature 3: UGAL implicitly differentiates paths of different lengths.
  - Biased towards picking shorter paths

3. Modify MCF model based on UGAL Features

- Feature 1: UGAL considers all MIN and VLB paths
- For each flow $d \in F$, consider
  - P_d^{MIN} = all available MIN paths
  - P_d^{VLB}; = all available VLB paths
- Modify MCF model

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

$$x_d = \sum_{p \in P_d^{MIN}} x_d^p + \sum_{q \in P_d^{VLB}} x_d^q$$

3. Modify MCF model based on UGAL Features

- Feature 2: UGAL randomly selects candidate MIN and VLB paths for each packet
  - For large enough sample space all MIN paths of a flow could be used equally

$$x_d^p = x_d^{MIN} \qquad \forall p \in P_d^{MIN}$$

  - All VLB paths of a flow could be used equally

$$x_d^q = x_d^{VLB} \qquad \forall q \in P_d^{VLB}$$

3. Modify MCF model based on UGAL Features

- Feature 3: UGAL differentiates paths of different lengths
  - all **same-length** MIN paths of a flow could be used equally

$$x_d^p = x_d^{MIN,L} \qquad \forall p \in P_d^{MIN}, |p| = L$$

  - All **same-length** VLB paths of a flow could be used equally

$$x_d^q = x_d^{VLB,L} \qquad \forall q \in P_d^{VLB}, |q| = L$$

Step 3: Modify MCF model based on UGAL

- Three level of control for MIN and VLB paths:
  - **Individual:** all paths may have unique, optimized bandwidth (least restricted)
  - **Path-length-based random:** all paths of the same length treated equally, have same bandwidth
  - **All-random:** all paths treated equally, have same bandwidth (most restricted)

- We do not know which feature dictates overall performance
- Therefore, we introduce these features in different extents
- Model for MIN and VLB separately

Model 1

- Individual control over MIN paths
- Path length-based control over VLB paths

Maximize α

Subject to:

$$\alpha - \left(\sum_{p \in P_d^{MIN}} x_d^p + \sum_{P_d^{VLB,L} \neq \emptyset} |P_d^{VLB,L}| \times x_d^{VLB,L}\right) \leq 0, \ \forall d \in F$$

$$\sum_{p \in P_d^{MIN}(e), d \in F} x_d^p + \sum_{P_d^{VLB,L}(e) \neq \emptyset, d \in F} |P_d^{VLB,L}(e)| \times x_d^{VLB,L} \leq C_e, \forall e \in E$$

Model 2

- Individual control over MIN paths
- All-Random control over VLB paths

Maximize α

Subject to:

$$\alpha - \left(\sum_{p \in P_d^{MIN}} x_d^p + |P_d^{VLB}| \times x_d^{VLB}\right) \leq 0, \ \forall d \in F$$

$$\sum_{p \in P_d^{MIN}(e), d \in F} x_d^p + \sum_{P_d^{VLB}(e) \neq \emptyset, d \in F} |P_d^{VLB}(e)| \times x_d^{VLB} \leq C_e, \forall e \in E$$

Model 3

- Path length-based control over both MIN and VLB paths

Maximize α

Subject to:

$$\alpha - \left(\sum_{P_d^{MIN,L} \neq \emptyset} |P_d^{MIN,L}| \times x_d^{MIN,L} + \sum_{P_d^{VLB,L} \neq \emptyset} |P_d^{VLB,L}| \times x_d^{VLB,L}\right) \leq 0, \forall d \in F$$

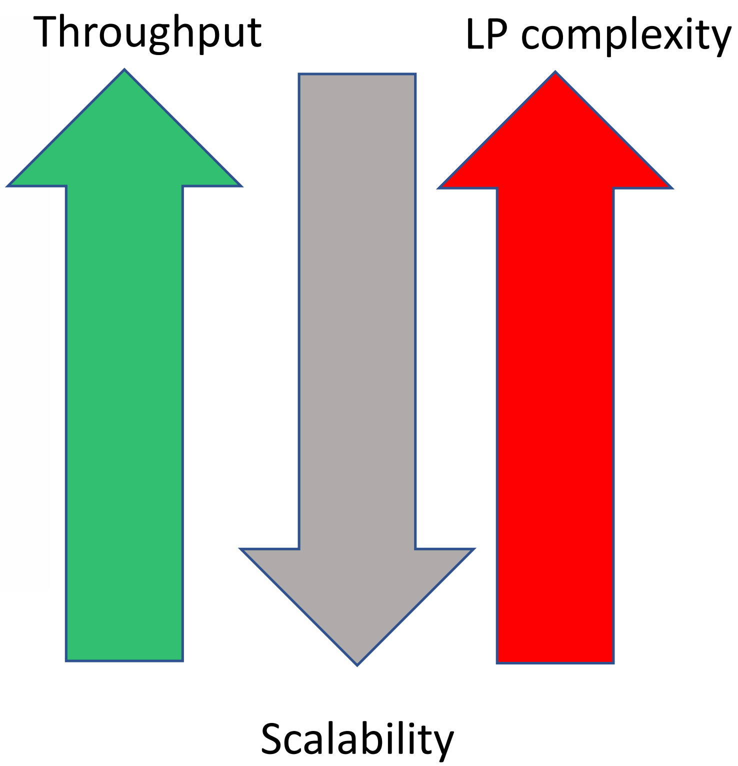
$$\sum_{P_d^{MIN,L}(e) \neq \emptyset, d \in F} |P_d^{MIN,L}(e)| \times x_d^{MIN,L} + \sum_{P_d^{VLB,L}(e) \neq \emptyset, d \in F} |P_d^{VLB,L}(e)| \times x_d^{VLB,L} \leq C_e, \forall e \in E$$

More models!

- Model 4:
  - Path length-based MIN path rates
  - All-random VLB path rates
- Model 5:
  - All-random MIN and VLB path rates

Models Summary

Model	MIN	VLB
No. 0	individual	individual
No. 1	individual	path-length-based random
No. 2	individual	all random
No. 3	path-length-based random	path-length-based random
No. 4	path-length-based random	all random
No. 5	all random	all random



Step 4. Validation and Analysis

- Used LP models to calculate the Max. Concurrent Flow rate

- Topologies:
  - Dragonfly *dfly(p,a,h,g)*:
    - fully connected groups of *a* routers, *p* nodes and *h* global links per router, *g* groups
  - Cascade
    - 96-router group in 16 x 6 HyperX, 4 nodes per router, 6 groups
    - global connections taken from NERSC's Edison topology dump**

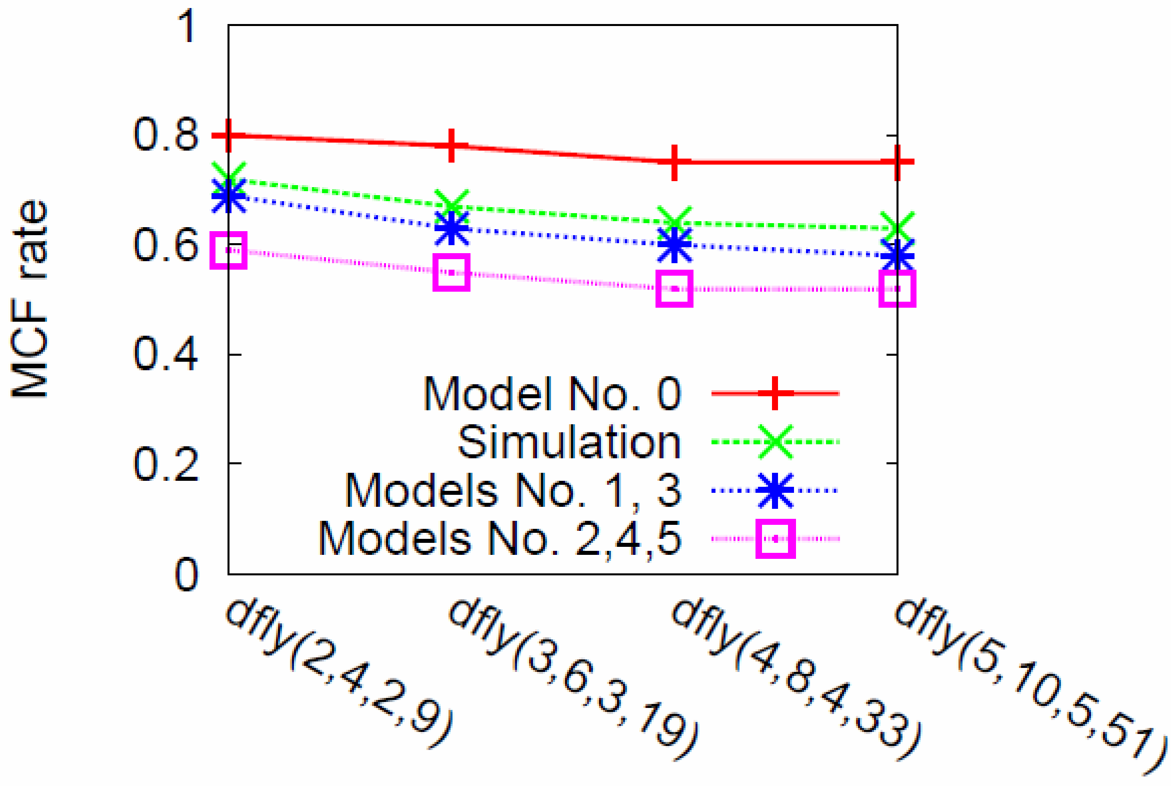
- Simulated UGAL on same topologies in Booksim* Interconnect Simulator

http://www.nersc.gov/users/computational-systems/edison/

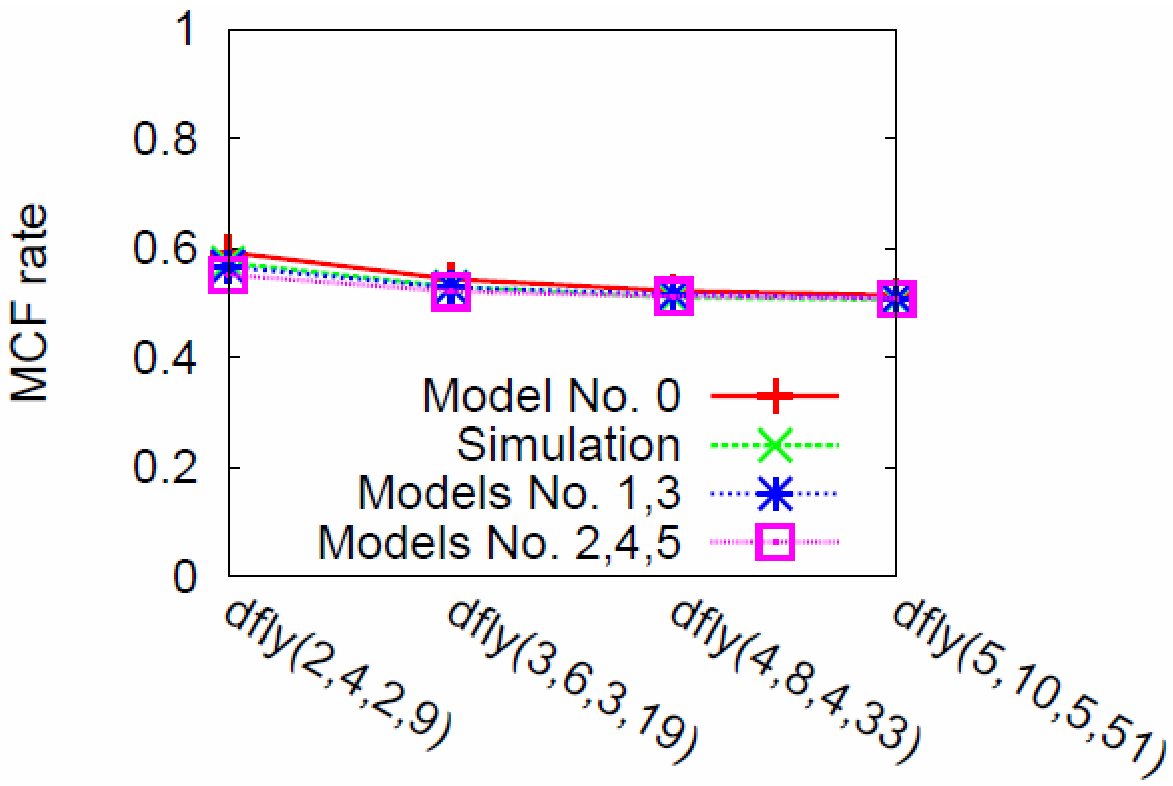
*N. Jiang, J. Balfour, D. U. Becker, B. Towles, W. J. Dally, G. Michelogiannakis, A detailed and flexible cycle-accurate network-on-chip simulator, ISPASS' 2013

Model validation: Canonical Dragonfly

Random Permutation Traffic

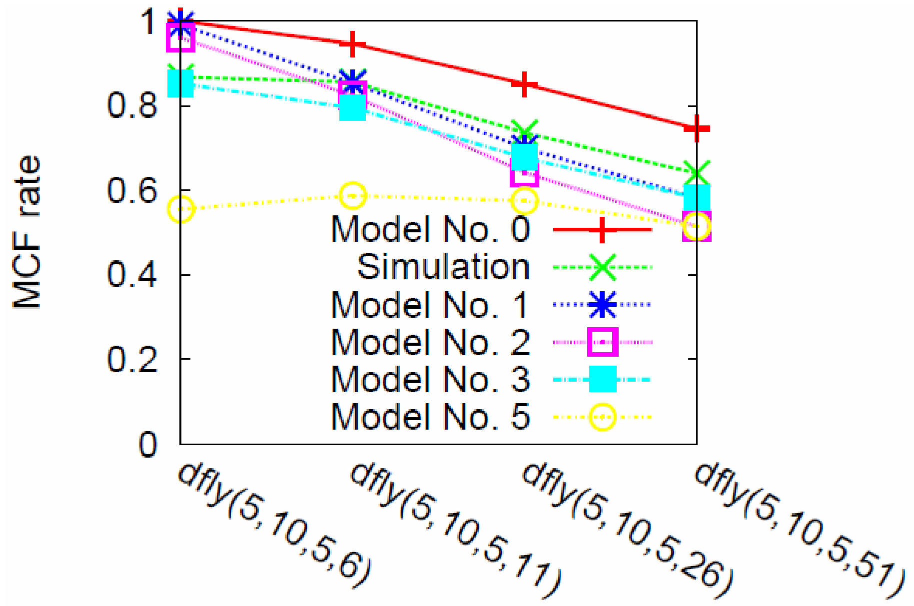


Random Shift Traffic(adversarial)



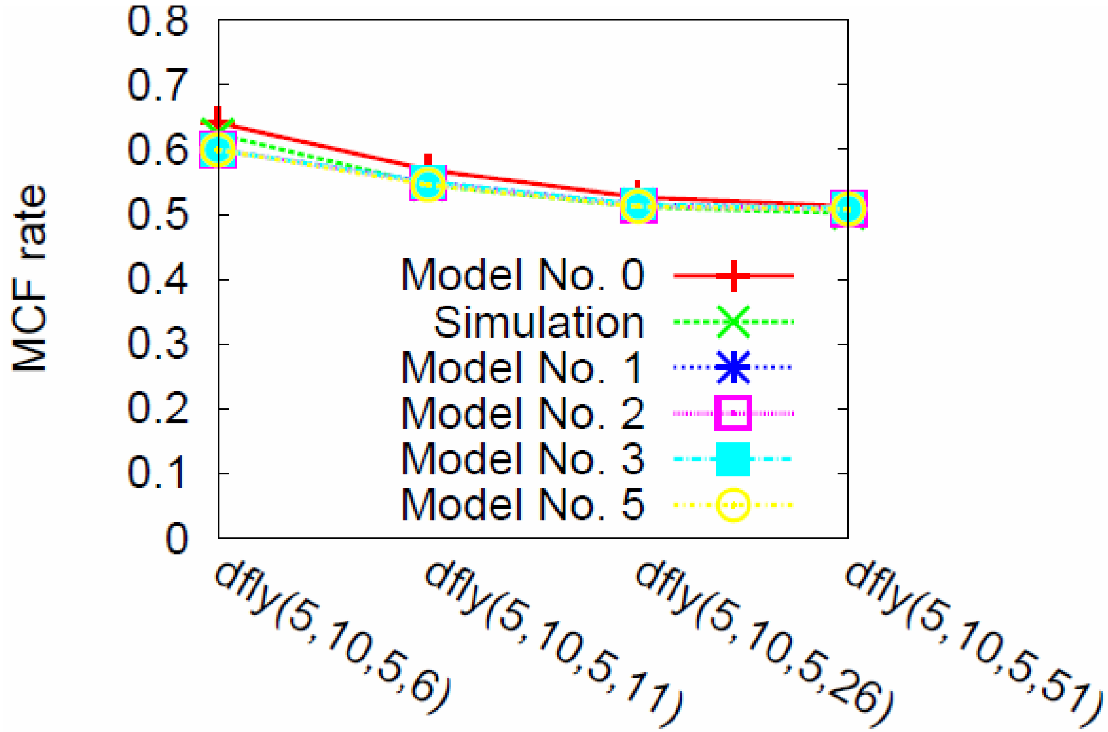
Model validation: Varying # of groups

Random Permutation Traffic



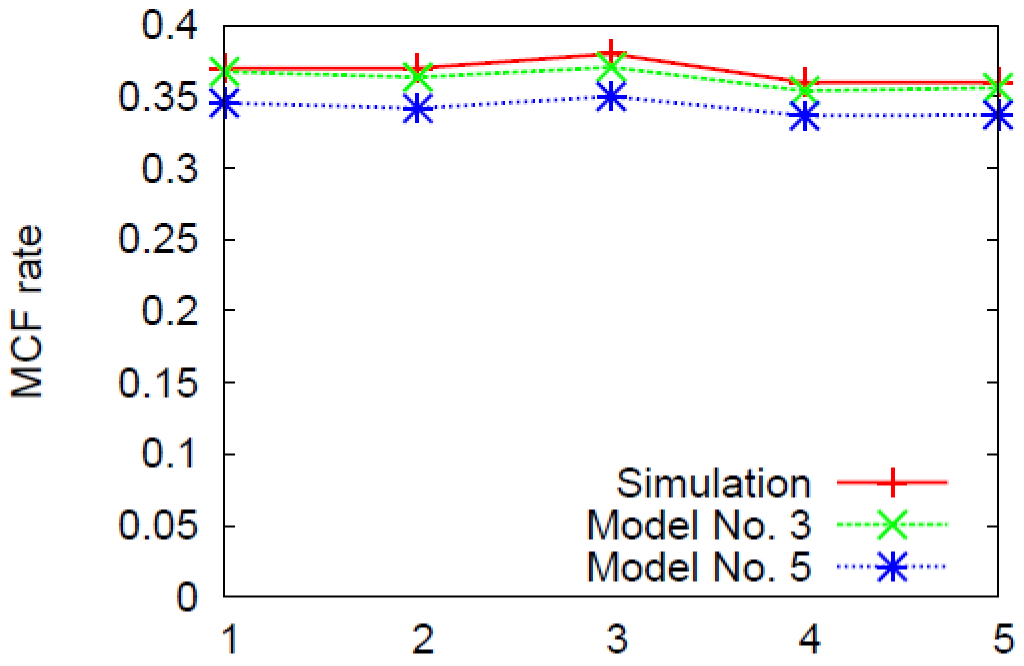
Increasing network size ->
<-higher number of MIN paths

Random Shift Traffic

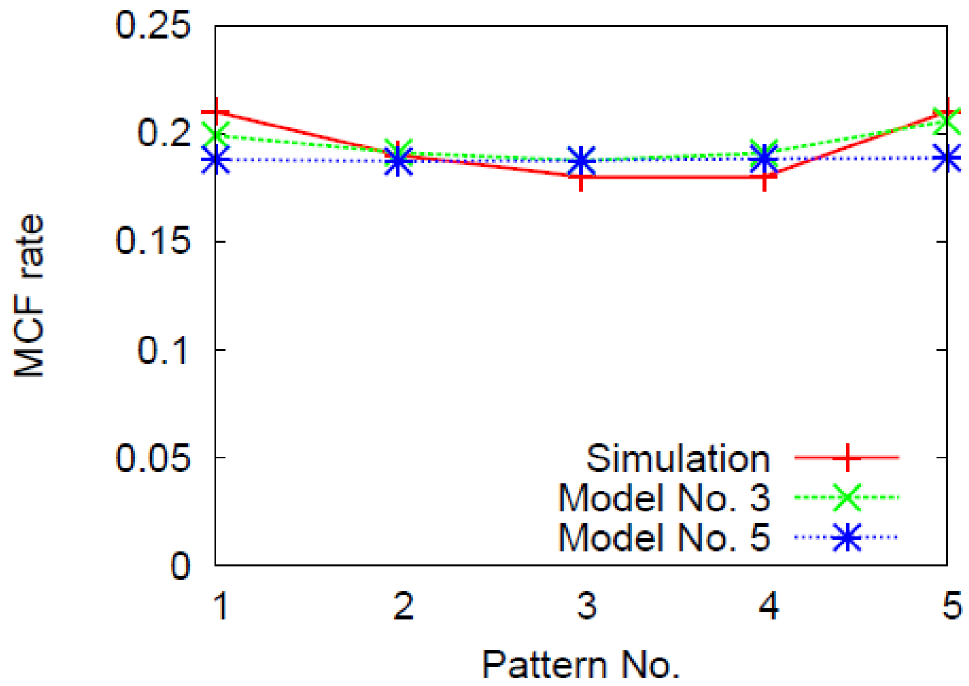


Model validation: Cascade Topology

5 random permutations



5 random shift patterns



Summary

- We develop a set of throughput models for UGAL on Dragonfly topology
- We identify an efficient model that accurately characterize UGAL on various Dragonfly designs
- We learn that UGAL on dragonfly optimizes throughput performance partially, based on path length

Thank you!
Questions?

Md Atiqul Mollah

mollah@cs.fsu.edu