Multistage Switches are not Crossbars: Effects of Static Routing in High-Performance Networks

- A Case Study with InfiniBand -

Torsten Hoefler, Timo Schneider, and Andrew Lumsdaine

Open Systems Lab
Indiana University
Bloomington, USA

IEEE Cluster 2008

Tsukuba, Japan

September, 29th 2008

Introduction

- large-scale networks are common on HPC
- huge variety of different technologies (IB, QSNet, Myrinet)
- offering: offload, onload, OS bypass
- we focus on topologies and routing!

Topologies

- flat: Ring, Kautz, k-ary n-cubes (Torus, Hypercube)
- MIN: Omega, Banyan, Clos, k-ary n-tree (Fat Tree)

Routing

- oblivious: fully random, random paths, online, ...
- adaptive: simple adaptive, probing adaptive, ...

$\Rightarrow$ focus on Fat Tree Topologies with oblivious routing!

Why Fat Trees?

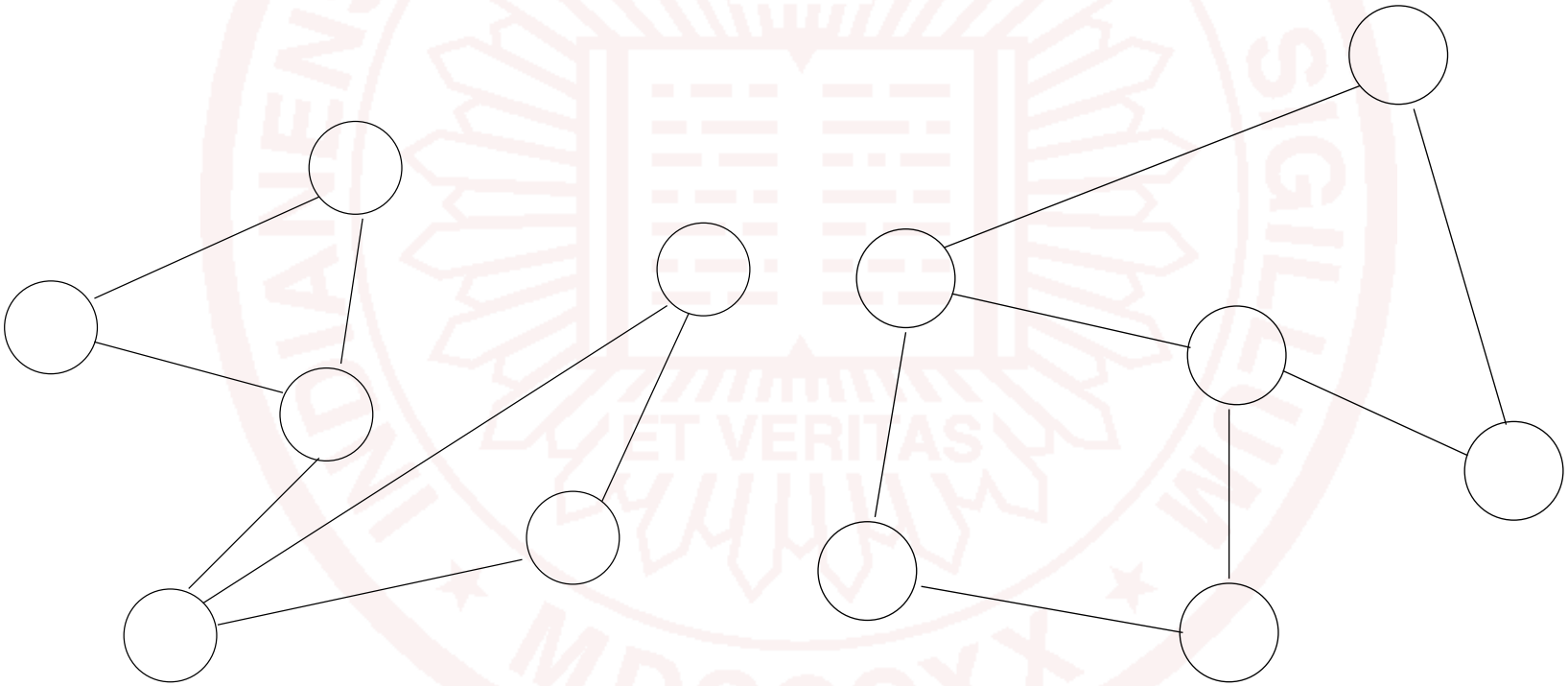
Fat Tree networks seem to have several advantages:

- simple construction rule
- Clos networks are a special case
- high bandwidth at large scale
- well understood since the 60s (Telephone)
- used by many switch vendors
- can be built with full bisection bandwidth (FBB)
- maps many (all?) patterns optimally
- simple deadlock-free routing

... so it seems ...

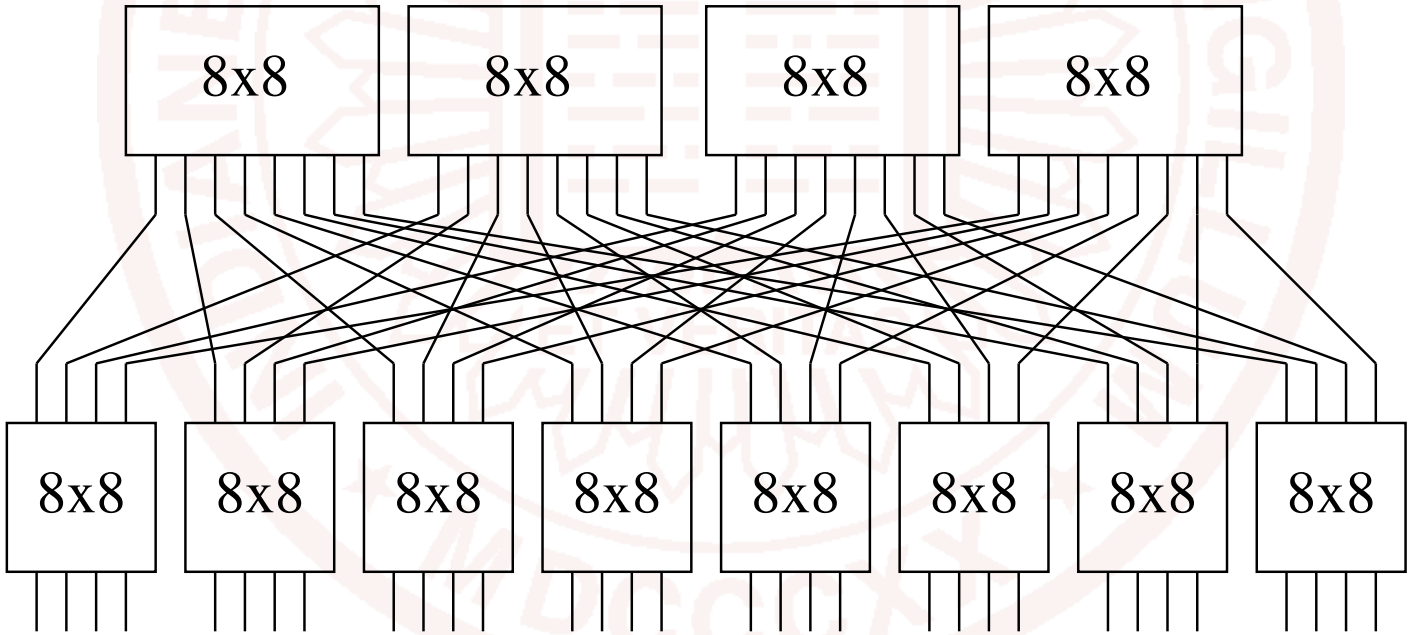
What is Bisection Bandwidth?

Definition: For a general network with N endpoints, represented as a graph with a bandwidth of one on every edge, BB is defined as the minimum number of edges that have to be removed in order to split the graphs into two equallysized unconnected parts.



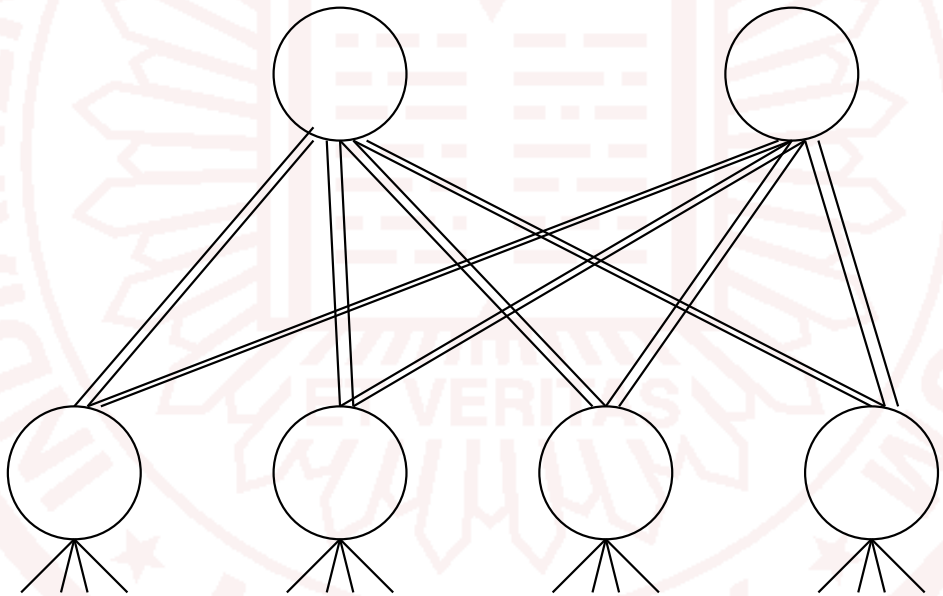
Clos Networks

- see [Clos'53] for details
- can be built blocking, rearrangable non-blocking, strictly non-blocking
- rearrangable non-blocking is most widely used
- $\frac{N}{2} + N$ $N \times N$ crossbars
- $\frac{N}{2} \cdot N$ endpoints and connections ("cables")



k-ary n-trees (Fat Trees)

- see [Leiserson'90] for details
- "generalisation" of Clos networks
- much more flexible in size and bandwidth
- similar principles



Oblivious Routing and InfiniBand

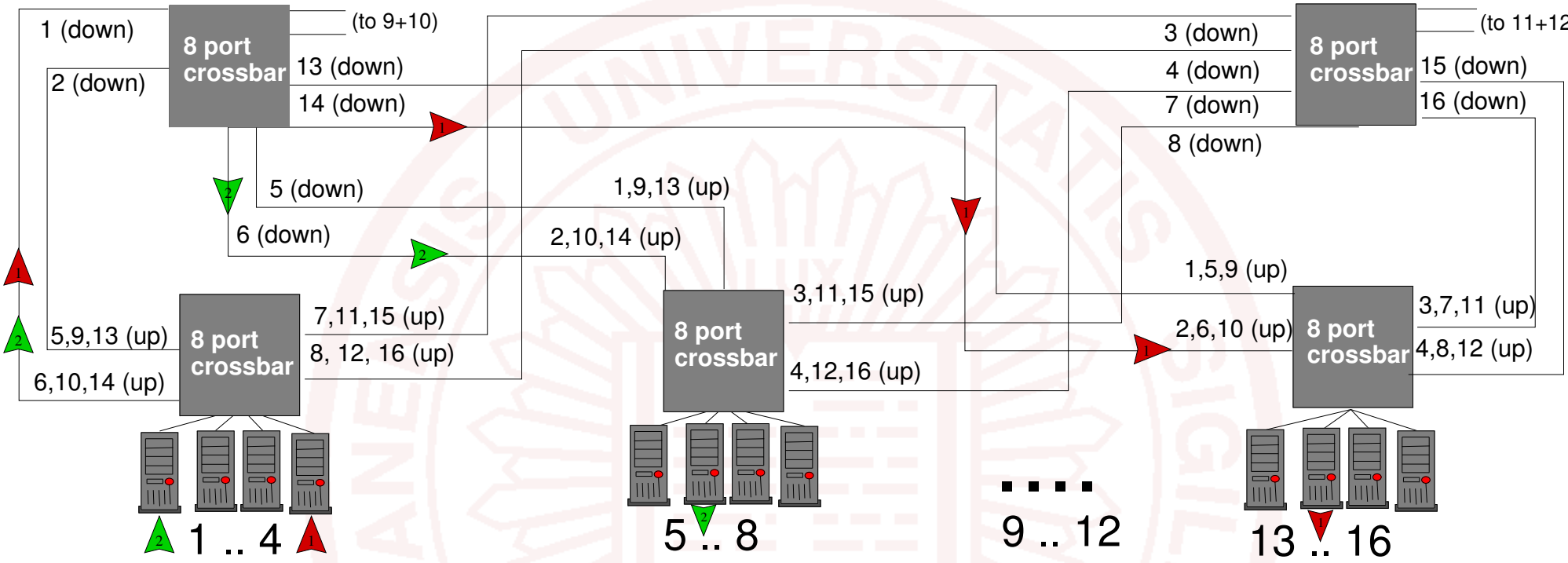
Oblivious Routing

- static routing without considering the traffic demands
- e.g., Ethernet, InfiniBand, IP, ...
- adaptive routing has limits (fast changing patterns with small packets)

InfiniBand

- Subnet Manager (SM) discovers topology and computes routes
- crossbars have destination-based forwarding tables
- 24-port crossbars -> Clos network has 288 ports
- recursive Clos up to 41472 ports
- biggest chassis has 3456 ports (Fat Tree)

An FBB Network Pattern



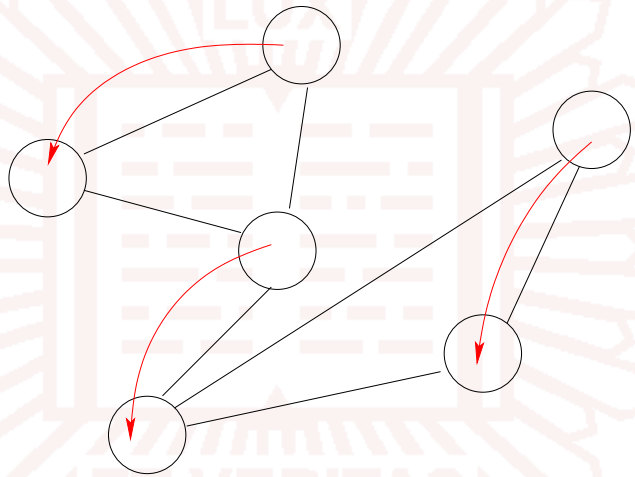
two distinct communications (1 to 6 and 4 to 14) in an FBB network

$\Rightarrow$ no full bandwidth!

Bandwidth depends on traffic patterns, routing and topology!

What is the essence of Bisection Bandwidth?

- Is it an upper bound to real bandwidth?
  - no, see example on last slide!
- Is it a lower bound to real bandwidth?
  - no, see:



- Is it the expected (average) bandwidth?
  - not easy to assess
  - simulate different traffic patterns!

The effective Bisection Bandwidth (eBB)

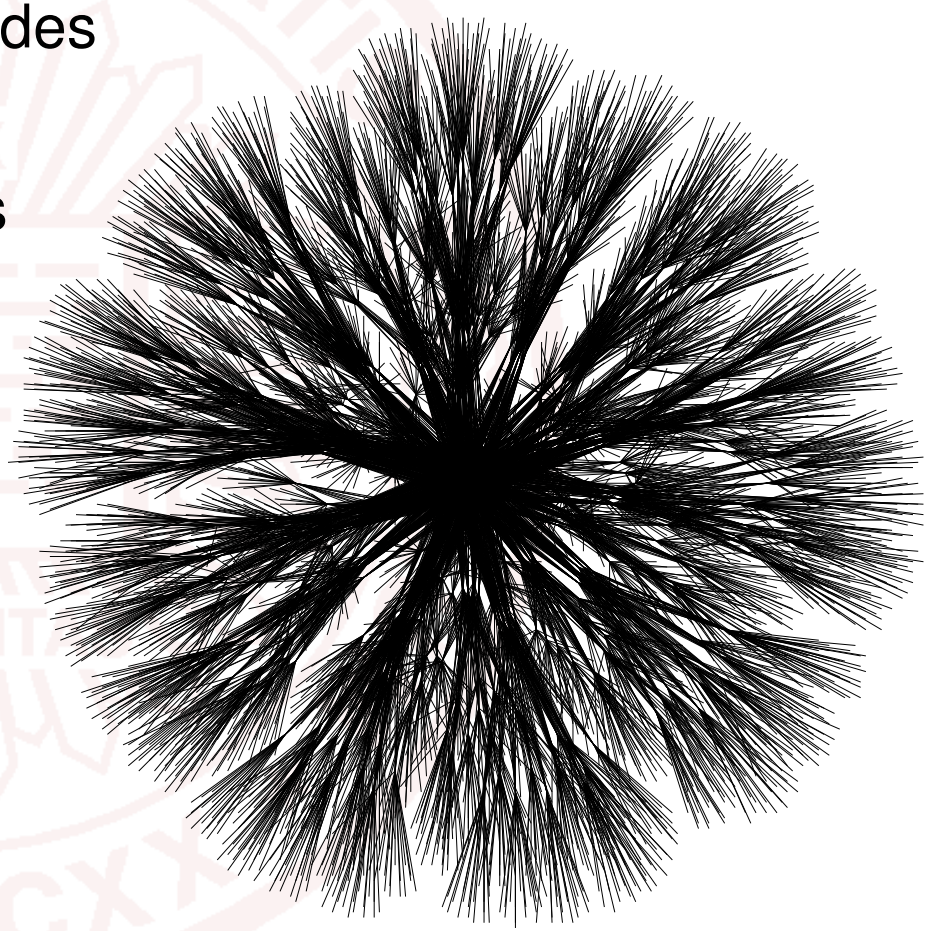
- models real bandwidth as the average bandwidth of all *bisect* patterns
- constructing a *bisect* pattern:
  - divide network in two equal partitions A and B
  - find exactly one peer in B for each node in A
- $\binom{P}{\frac{P}{2}}$ ways to partition P nodes
- $\frac{P}{2}!$ ways to pair $\frac{P}{2}$ nodes
- $\longrightarrow$ huge number of patterns
  - many of them have full bandwidth
  - no closed form yet, thus simulate

The Network Simulator

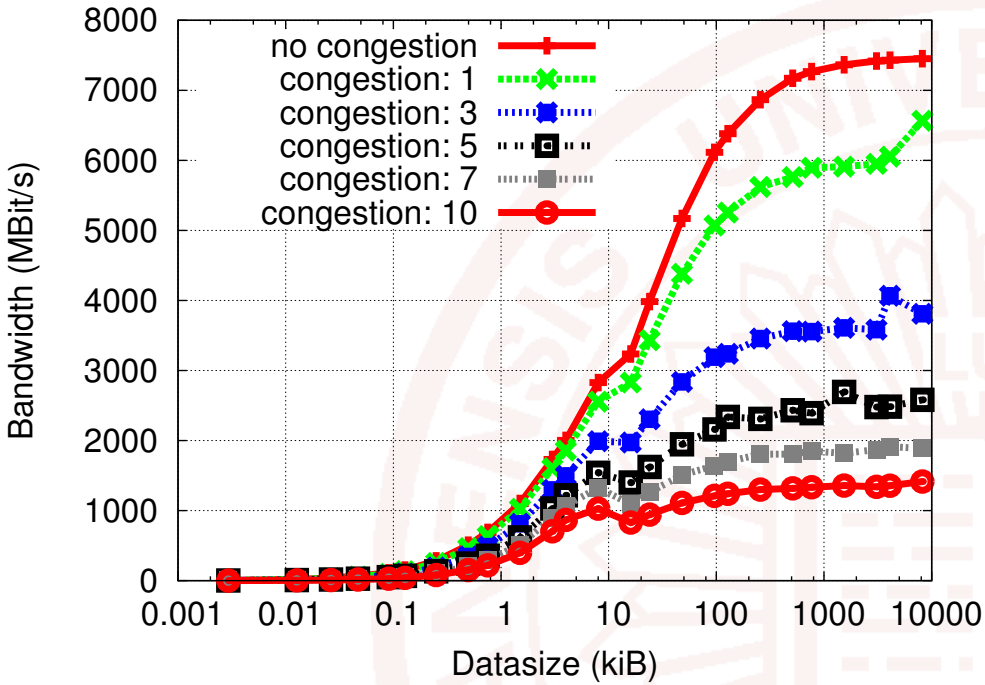
- model physical network as a graph
- construct random bisect patterns
  - simulate packet routing and record edge congestion
  - find maximum congestion c along each path
- compute average bandwidth per path $b = \frac{1}{c}$
- repeat simulation with many patterns

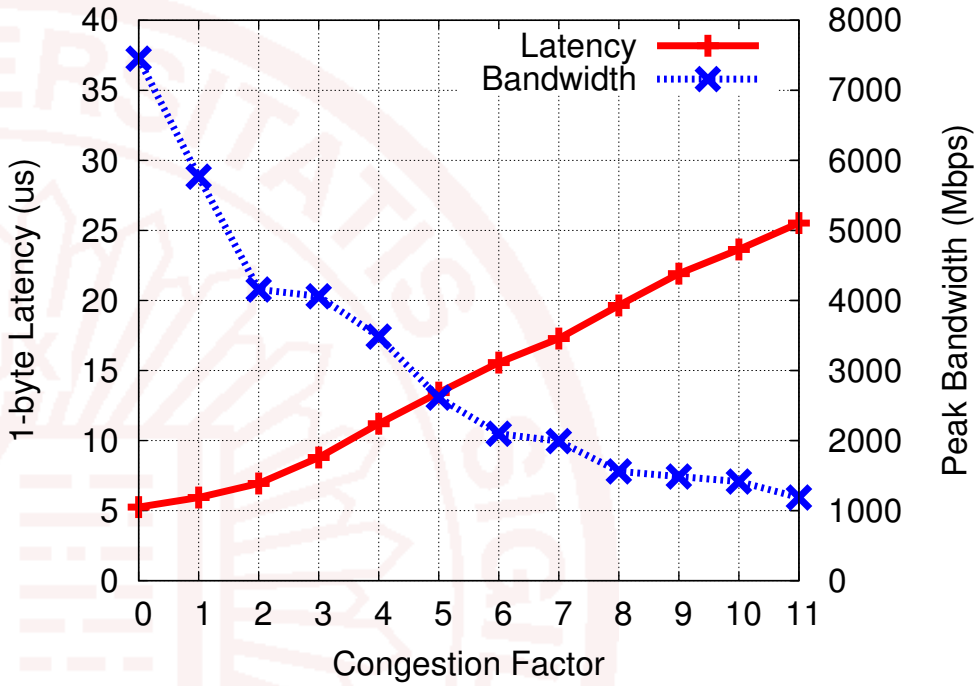
Simulating Real-World Systems

- retrieved topology via `ibnetdiscover` and `ibdiagnet`
- four large-scale InfiniBand systems queried:
- Thunderbird at SNL - 4390 nodes
- Atlas at LLNL - 1142 nodes
- Ranger at TACC - 3908 nodes
- CHiC at TUC - 566 nodes



Influence of Head of Line Blocking

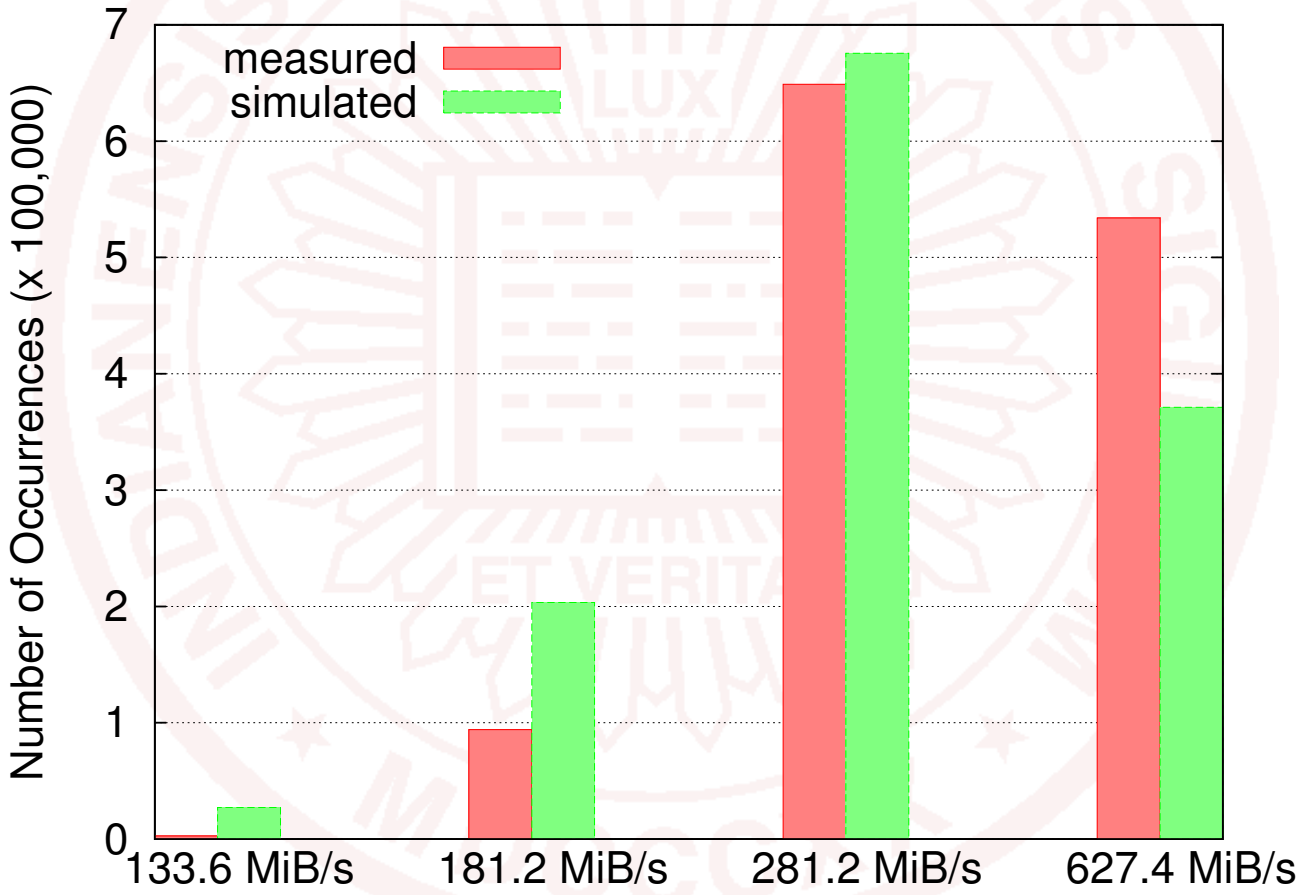




- communication between different pairs (*bisect*)
- laid out to cause congestion
- 24 ports → max. congestion is 11

Simulation and Reality

- compare 512 node CHiC run with 566 node simulation
- random bisect patterns, bins of size 50 MiB/s
- measured and simulated $>$ 99.9% into only 4 bins!



Results on other Systems

Effective Bisection Bandwidth

- Ranger (3908 nodes, FBB): 57.5%
- Atlas (1142 nodes, FBB): 55.6%
- Thunderbird (4390 nodes, 1/2 FBB): 40.6%

Other Effects of Congestion?

- bandwidth varies with comm. pattern
- not easy to predict/model
- effects on latency are not trivial (buffering etc.)
- leads to network skew (problem at large scale)

Influence on Applications?

Application Analysis

- MPQC - MPI_Reduce, MPI_Bcast, MPI_Allreduce
- MILC - Neighbor Exchange, MPI_Allreduce
- POP - Neighbor Exchange, MPI_Allreduce
- Octopus - Neighbor Exchange, MPI_Allreduce

Conclusions?

- many applications use collective communication
- nearest neighbor exchange is also collective
- patterns can be scaled up!
- simulate collective patterns: tree, dissemination, six neighbor

Results on other Systems

Six Neighbor Bandwidth

- Ranger: 62.4%
- Atlas: 60.7%
- Thunderbird: 37.4%

Tree Bandwidth

- Ranger: 69.9%
- Atlas: 71.3%
- Thunderbird: 57.4%

Dissemination Bandwidth

- Ranger: 41.9%
- Atlas: 40.2%
- Thunderbird: 27.4%

Conclusions and Future Work

Conclusions

- bisection bandwidth does reflect practice well
- effective bisection bandwidth is harder to assess but more realistic
- applications bandwidths suffer, even on FBB networks

Future Work

- develop better oblivious routing for IB
- analyze more systems and applications
- look into adaptive routing options? or LMC??
- look at other interconnects and topologies

Conclusions and Future Work

Conclusions

- bisection bandwidth does reflect practice well
- effective bisection bandwidth is harder to assess but more realistic
- applications bandwidths suffer, even on FBB networks

Future Work

- develop better oblivious routing for IB
- analyze more systems and applications
- look into adaptive routing options? or LMC??
- look at other interconnects and topologies