A Low Latency Scalable 3D NoC Using BFT Topology with Table Based Uniform Routing

Avik Bose†, Prasun Ghosal†‡, Saraju P. Mohanty‡

† Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, WB, India

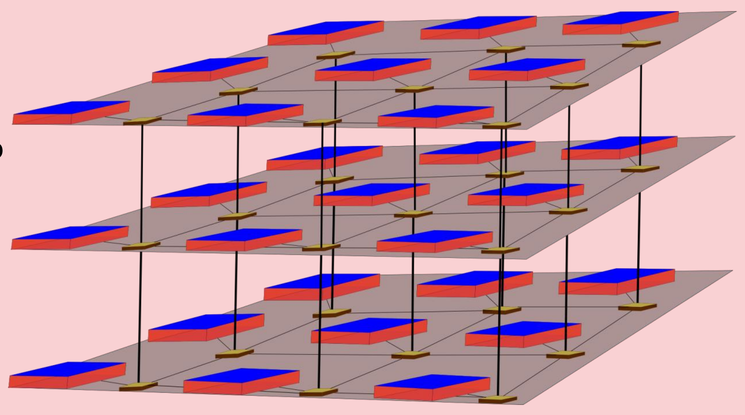
‡ University of North Texas, Denton, TX 76203, USA

Email: avikbose1145@gmail, prasun@ieee.org, saraju.mohanty@unt.edu

PROBLEM DESCRIPTION

- Due to the close proximity of layers in 3D NoC structure the signals travelling in the vertical (inter-layer) direction is much faster than the horizontal (intralayer) in their 2D counterpart.
- An inter-layer connection requires the addition of two more links (up and down) to each router that leads to an increase in complexity as well as the blocking probability inside the router.
- Being a multi-hop communication fabric, the traditional NoC routers can not be placed on the vertical path in a NoC as the multi-hop delay and the router delay would overshadow the ultra fast propagation time.

- Thus it is desirable to have single hop communication among the layers because of the short distance between them.
- Also, the number of vertical pillars should be kept low to reduce the manufacturing cost of a 3D NoC.
- It induces a new problem for the IP blocks with close vicinity to the pillar nodes on a layer giving more advantage in case of inter layer communication than those that are at relatively distant position from the pillar nodes.

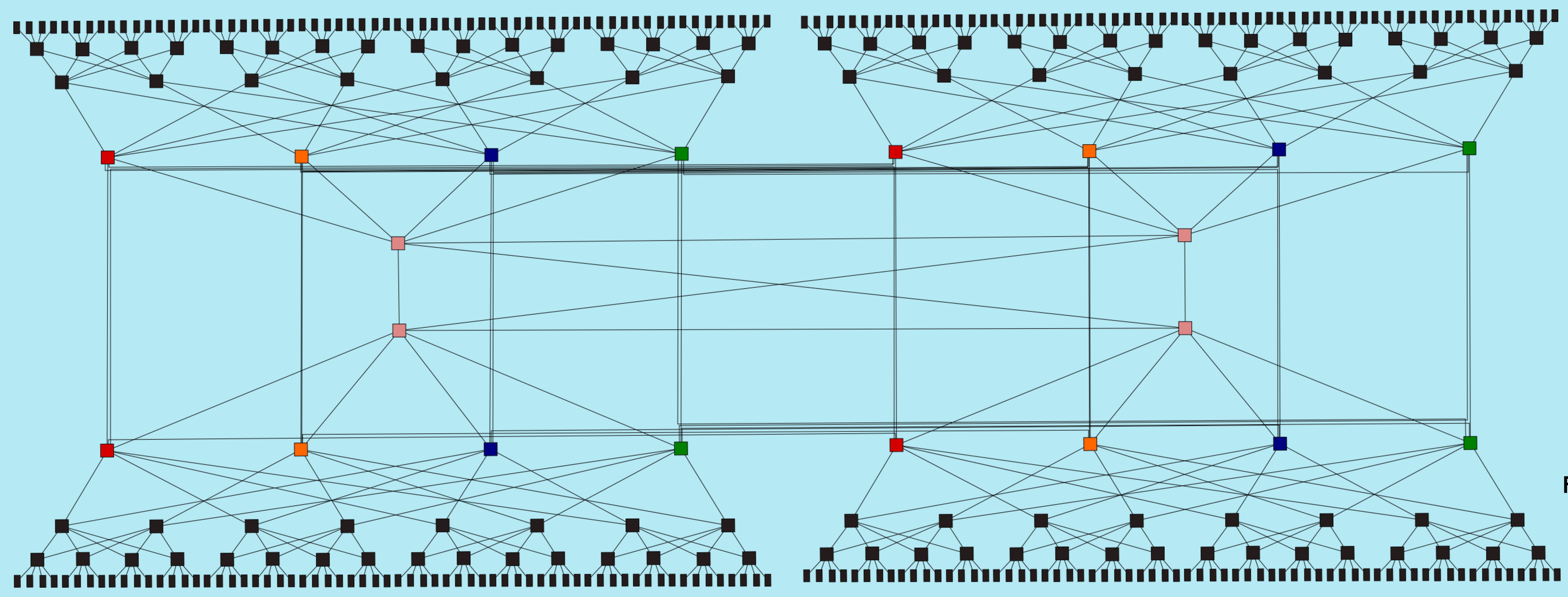


Connection Structure in a typical 3D Mesh topology.

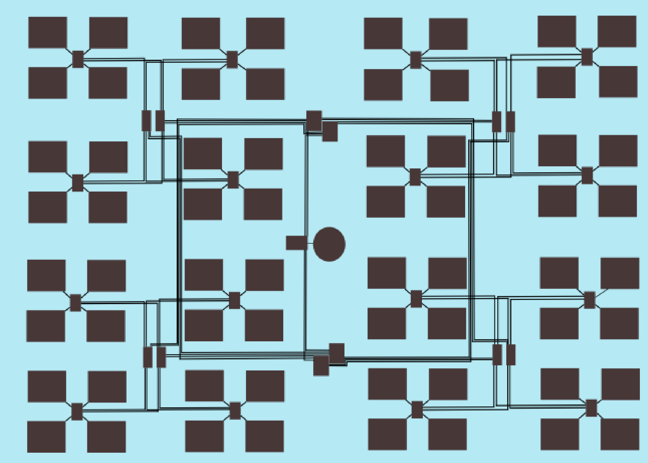
Objective

To propose a novel 3D NoC topology with proper routing method that can address all the above issues.

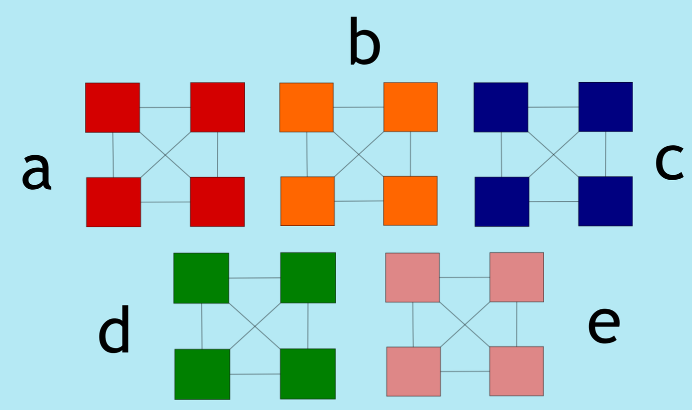
PROPOSED SOLUTION



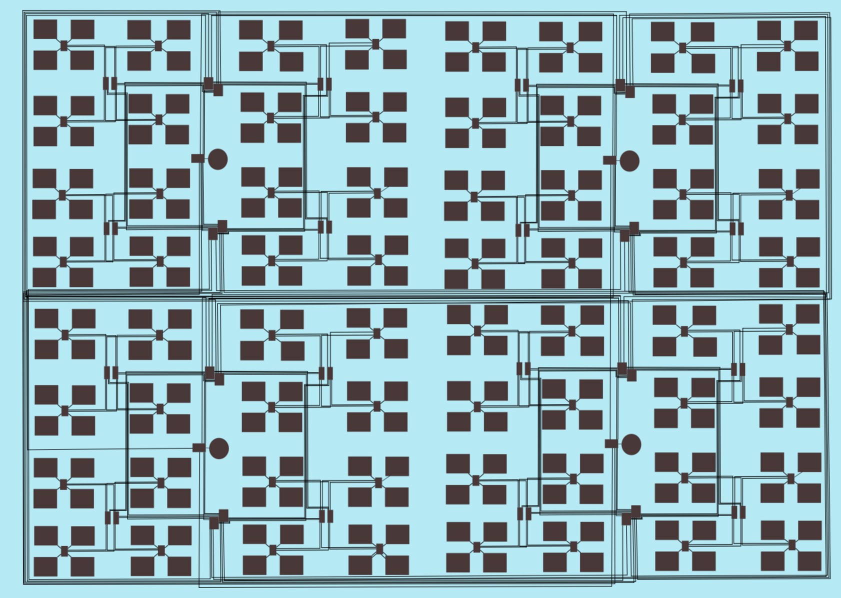
Logical diagram of four connected Butterfly Fat Trees.



Floor plan of a single BFT as a locality of a pillar node.



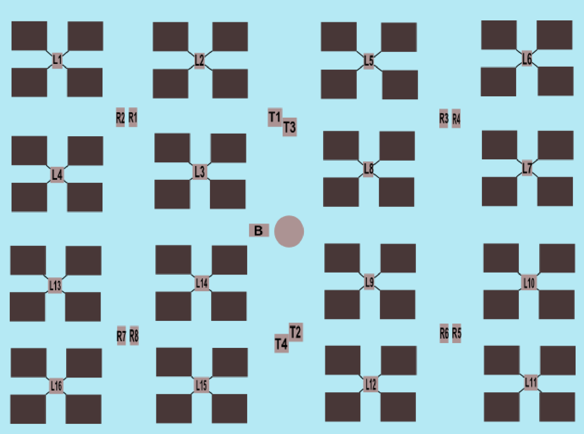
Connectivity between border and root routers. Each of (a),(b),(c), and (d) is a four root level router, each pertaining to a particular BFT. (e) Border routers of a floor, each of which is dedicated to a particular BFT.



Floor plan of a complete NoC layer comprising four BFTs as the localities of their respective pillar nodes.

- Overall architecture of a layer in the proposed design of 3D NoC based on Butterfly Fat Tree topology. Four BFTs are connected together having four root nodes each (colored red, orange, blue, and green).
- Root nodes with same color are connected together to form complete graph. Reason behind connecting the root nodes in the above manner is to reduce the network latency in terms of hop count.
- For inter layer communication DTDMA pillars are used that eliminate transactional character commonly associated with buses [8], [9] by employing a dynamic bus arbitration (thus close to 100 % bandwidth efficient). Single-hop communication and transaction-less arbitrations allow for low and predictable latencies. Furthermore, hybridization of NoC router with bus architectures requires only one additional link (in the place of two) on NoC router.
- The circular DTDMA pillar node is shown in the center of the floor plan picture of a single BFT. It is connected to a special router called as border router (responsible for regulating traffic across different layers of the chip). This router is the gateway for inter layer communication.

DIFFERENT ROUTING SCOPES



Distribution of routers across different scopes for a single BFT.

Layer Number	Tree Number	Region Number	Locality Number	Node Number
m bits	2 bits	2 bits	2 bits	2 bits

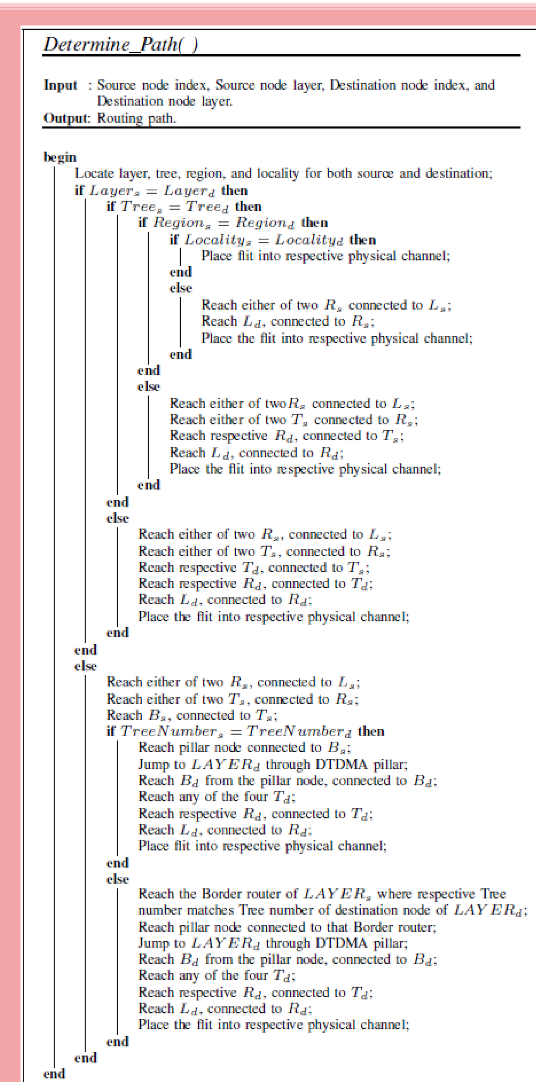
Address format of an IP block.

- First m bits specify the layer number.
- Next 2 bits specify the tree number out of four BFTs in a layer.
- Next 2 bits contain the region number out of four regions of a BFT.
- Next 2 bits denotes the locality number out of four localities of a region.
- Last 2 bits field contains the node number in that locality out of four IP blocks in that locality.

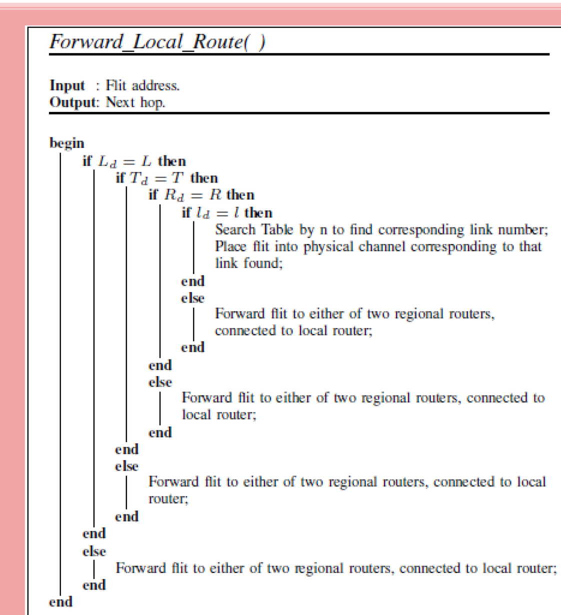
- ✓ *Layer Scope:* A 3D NoC chip consists of several layers.
- ✓ *Tree Scope:* Every layer is made up of four BFTs. (denoted by T)
- ✓ *Region Scope:* In every BFT there are four regions. Each region is made up of two regional Routers (denoted by R) and their siblings.
- ✓ *Local Scope:* Each region is made up of four localities, where each locality comprises of one local router (denoted by L) and four IP blocks connected to it

DIFFERENT ROUTING TABLES ACCORDING TO SCOPES	
Local Routing Table	Table (Node Number, Link Number)
Regional Routing Table	Table (Locality Number, Link Number)
Root Routing Table	Table1 (Region Number, Link Number) Table2 (Tree Number, Link Number)
Border Routing Table	Table (Router Number, Link Number)

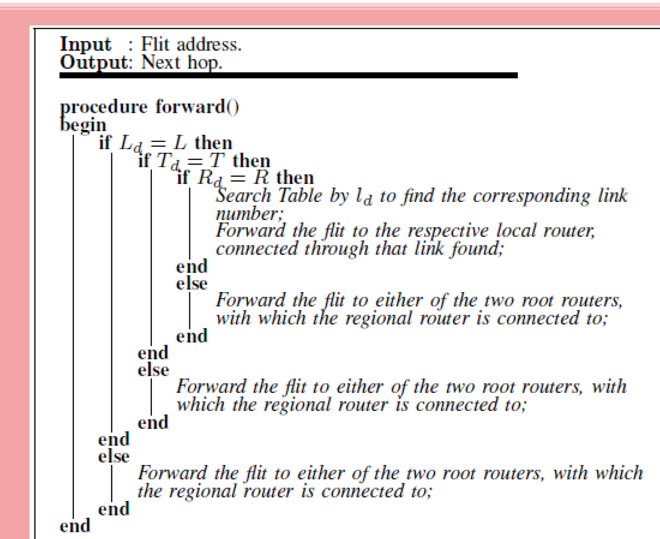
ROUTING ALGORITHMS FOR DIFFERENT SCOPES



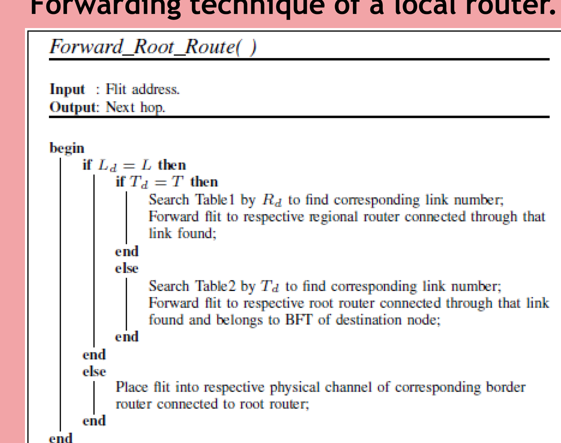
The main routing methodology.



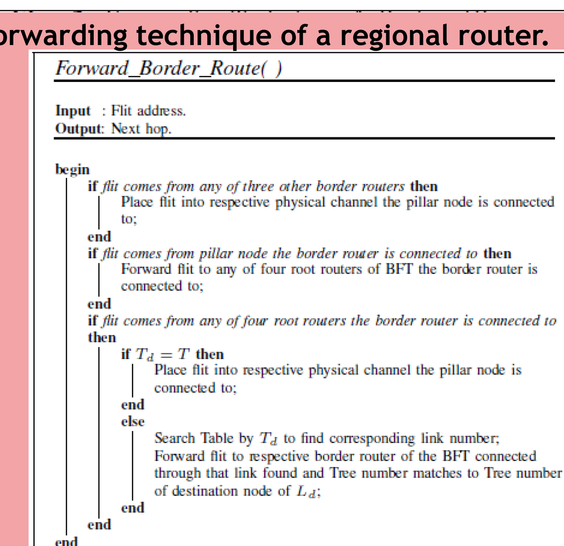
Forwarding technique of a local router.



Forwarding technique of a regional router.

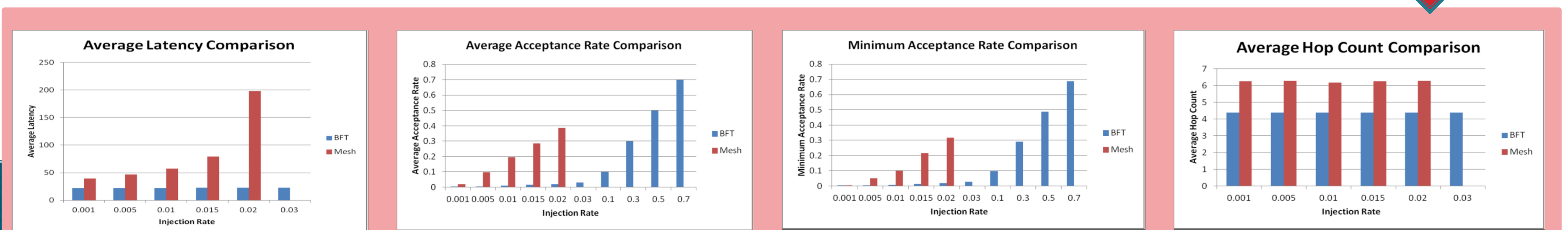


Forwarding technique of a root router.

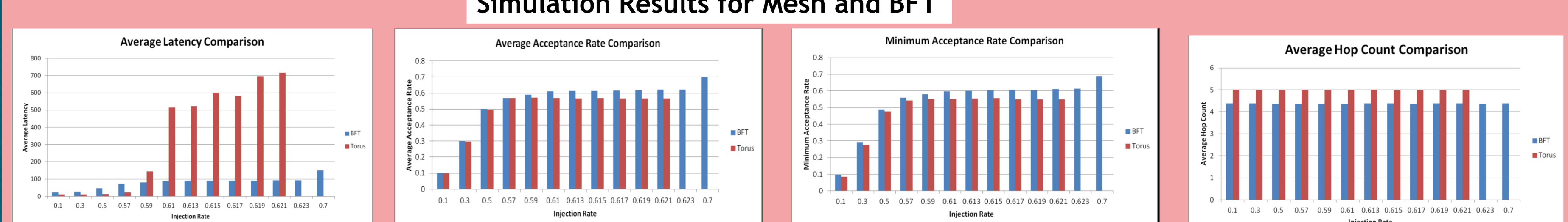


Forwarding technique of a border router.

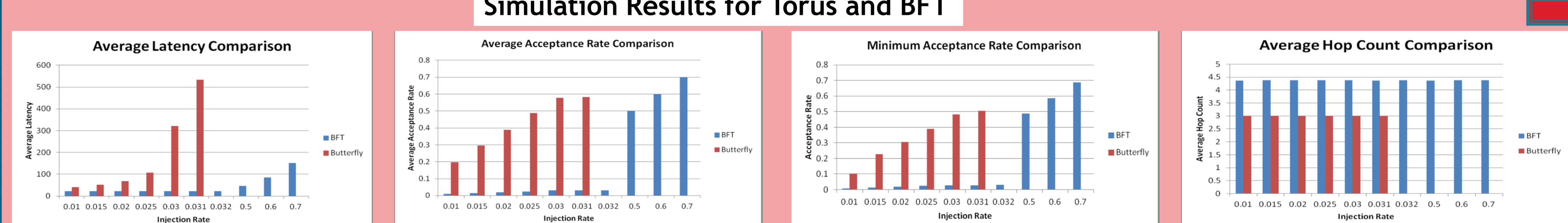
EXPERIMENTAL RESULTS



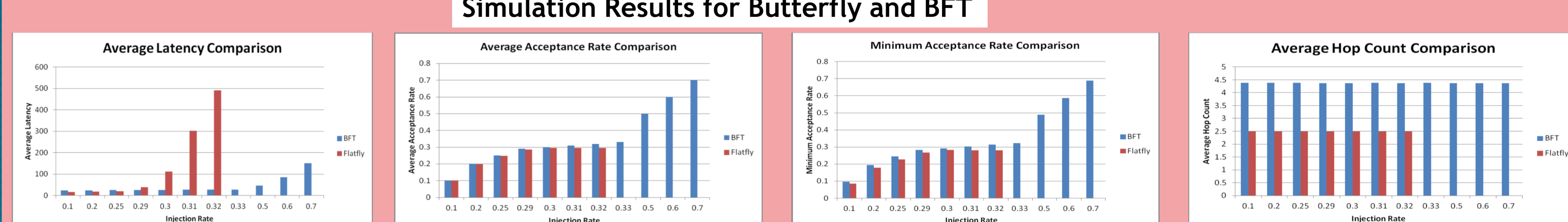
Simulation Results for Mesh and BFT



Simulation Results for Torus and BFT



Simulation Results for Butterfly and BFT



Simulation Results for Flattened Butterfly and BFT

SUMMARY OF COMPARATIVE IMPROVEMENTS (%) FOR DIFFERENT PERFORMANCE METRICS

Performance metrics / Topologies	Avg Hop Count	Avg Acceptance Rate	Avg Latency	Min Acceptance Rate
Mesh	30	NIL	43-89	NIL
Torus	13	6-9	83-88	6-13
Butterfly	NIL	NIL	46-96	NIL
Flattened Butterfly	NIL	1-8	31-95	5-14

CONCLUSION AND FUTURE WORK

- Proposed BFT topology can withstand heavy workload while still maintaining low latency, and the acceptance rate also increases with increasing injection rate. This is because of the uniform and load balancing connectivity of BFT where we have more than one path between a pair of source and destination but with same hop count. On the other hand, all other topological designs have failed to balance the load and sometimes crash.
- If routers can be designed in such a way that they can have the capability to balance load and control congestion efficiently, then with this design we can achieve a really effective NoC system for interactive applications with threading capability.
- Future works may be in the direction of investigating the thermal effects and optimizing it accordingly with a suitable core placement strategy, investigating and improving performance using real time application mapping and so on.

REFERENCES

[1] "International Technology Roadmap for Semiconductors." Available: http://www.itrs.net/.

[2] R. Mullins, A. West, and S. Moore, "Low-latency virtual channel routers for on-chip networks," in 31st Annual International Symposium on Computer Architecture, 2004.

[3] L. Benini and G. D. Micheli, "Networks on Chips: A New SoC Paradigm," IEEE Ccomputer, vol. 35, no. 1, pp. 70-78, 2002.

[4] K. Lahiri, A. Raghunathan, and S. Dey, "Evaluation of the Traffic- Performance Characteristics of System-on-Chip Communication Architectures," in Proceedings of the Fourteenth International Conference on VLSI Design, pp. 29-35, 2001.

[5] A. Agarwal, C. Iskander, and R. Shankar, "Survey of Network on Chip (NoC) Architectures & Contributions," Journal of Engineering, Computing and Architecture, vol. 3, pp. 21-27, 2009.

[6] F. Gebali, H. Elmiligi, and H. W. El-Kharashi, Networks-on-Chips - Theory and Practice. CRC Press, Taylor & Francis Croup, 2009.

[7] T. C. Xu, P. Liljeberg, and H. Tenhunen, "A Study of Through Silicon Via Impact to 3D Network-on-Chip Design," in International Conference on Electronics and Information Engineering, 2010.

[8] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, and M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," in 33rd International Symp. Computer Architecture (ISCA 2006), pp. 130–141, June 2005.

[9] C. Nicopoulos, V. Narayanan, and C. R. Das, Network-on-Chip Archi-tectures - A Holistic Design Exploration. Springer, 2009.