QoS Aware BiNoC Architecture

Shih-Hsin Lo, Ying-Cherng Lan, Hsin-Hsien Yeh, Wen-Chung Tsai, Yu-Hen Hu, and Sao-Jie Chen

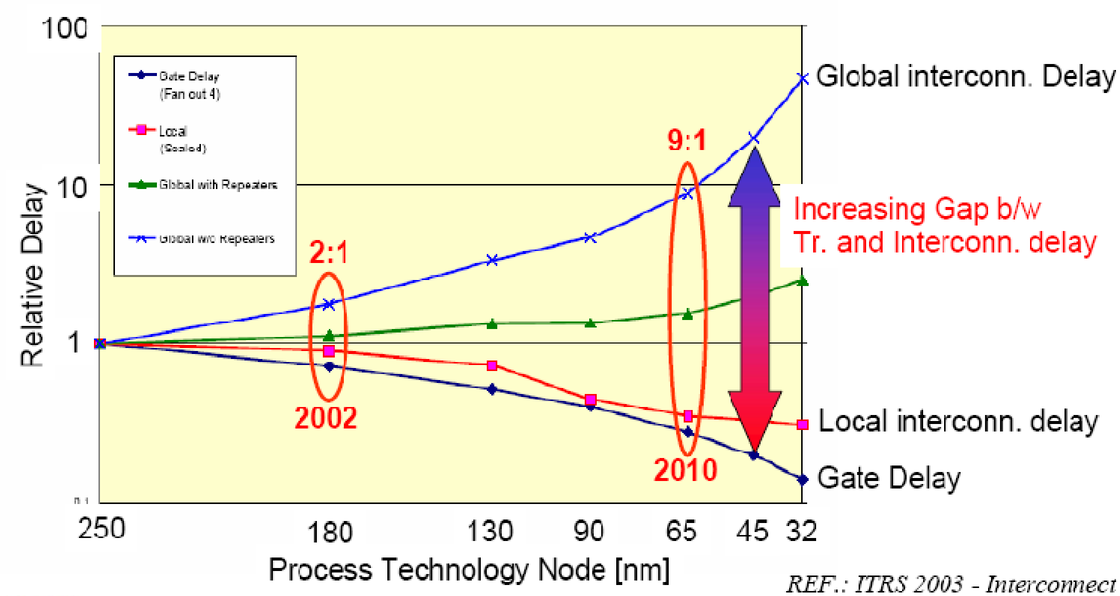
Ying-Cherng Lan

CAD System Lab
Graduate Institute of Electronics Engineering
National Taiwan University
Taipei, Taiwan, ROC

Introduction

- The trend toward many-core processing chips is now a well established one

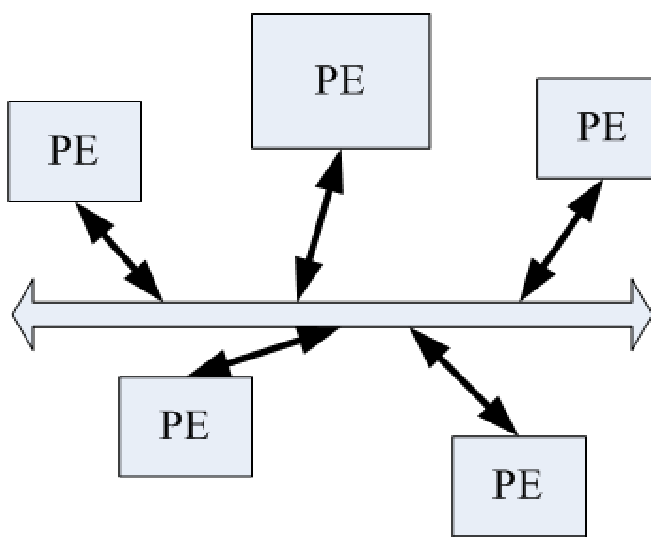
- Interconnect delay dominates gate delay
  - Global interconnect delay continuously increasing
  - Need multiple clock cycles to cross chip die
  - Limits the performance of microprocessors



Communication Centric Design

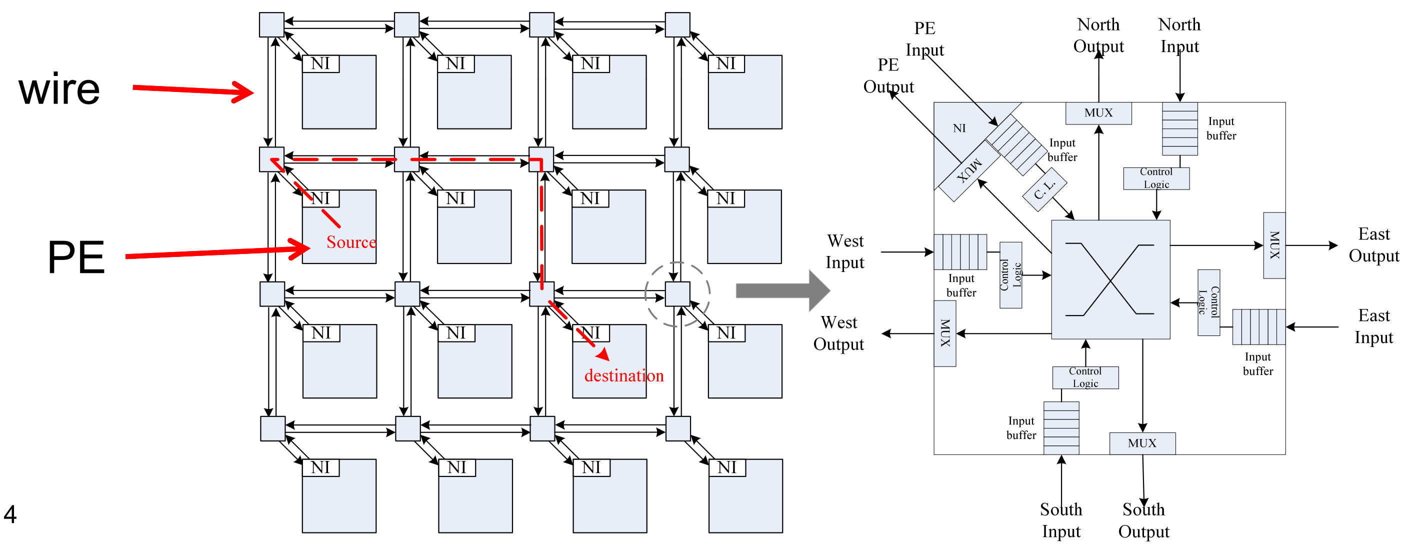
- The design concept of a system is moving gradually from computation-centric to communication centric.

- Conventional bus-based architecture becomes no longer a feasible communication scheme in terms of bandwidth, and scalability.



Network on Chip

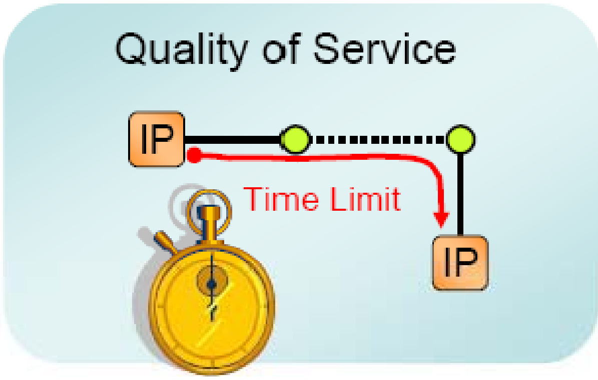
- Network on Chip (NoC) is a promising solution to mitigate the ever increasing communication complexity and provide better scalability.
  - W.J. Dally and B. Towles, “Route Packets, Not Wires: On-Chip Interconnection Networks,” in *Proceedings of DAC*, pp. 684-689, Jun. 2001.
  - L. Benini and G. DeMicheli, “Networks on Chips: a New SoC Paradigm,” *IEEE Computer,* vol. 35, no. 1, pp. 70-78, Jan. 2002.
  - A. Jantsch and H. Tenhunen (Eds.), *Networks on Chip*, Kluwer Academic Publishers, 2003.



Quality of Service for NoC

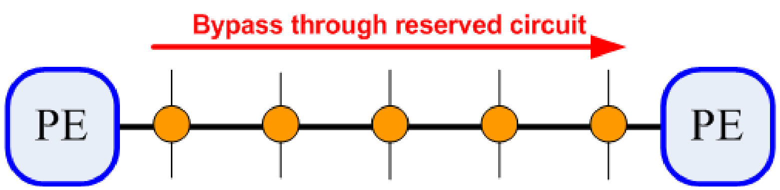
- Since many of the system applications have real-time requirements, the system and the network have to be predictable.

- To proceed a practical application, there are numerous type of packets in different importance need to be transmitted.
  - GS (guaranteed service) : guaranteed in latency. (e.g., real time stream)
  - BE (best effort) : guaranteed only in correctness

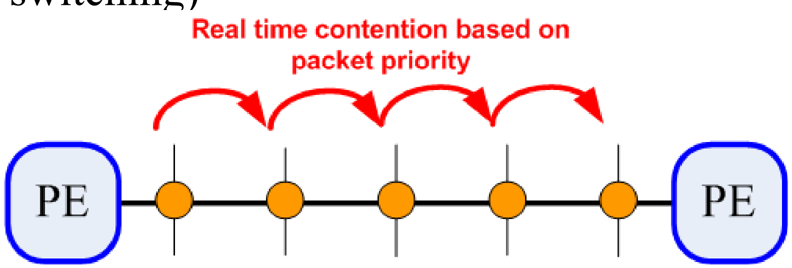


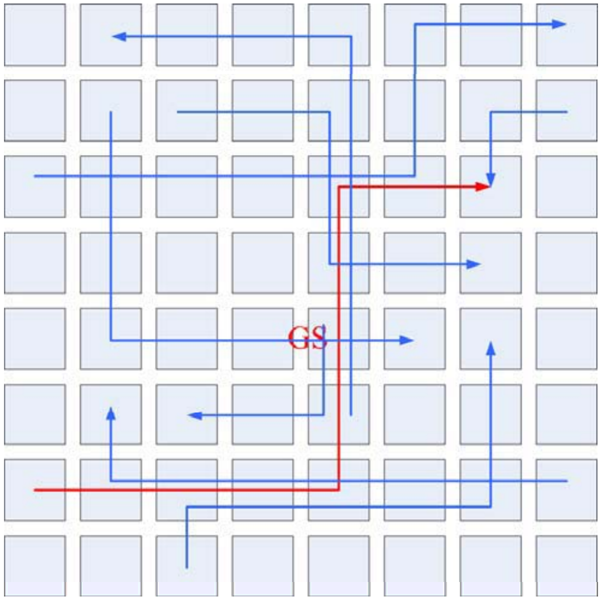
How to provide QoS for NoC

- To provide QoS in network on chip, two communication scheme have been proposed.
  - Connection-oriented mechanism (Circuit switching)



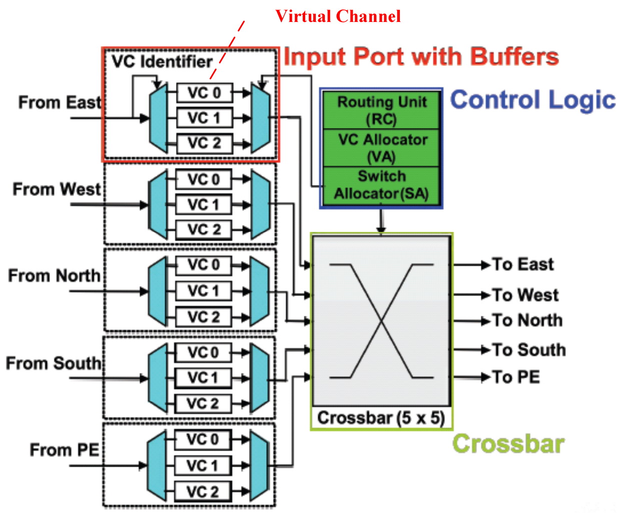
  - Connection-less mechanism (Packet-switching)

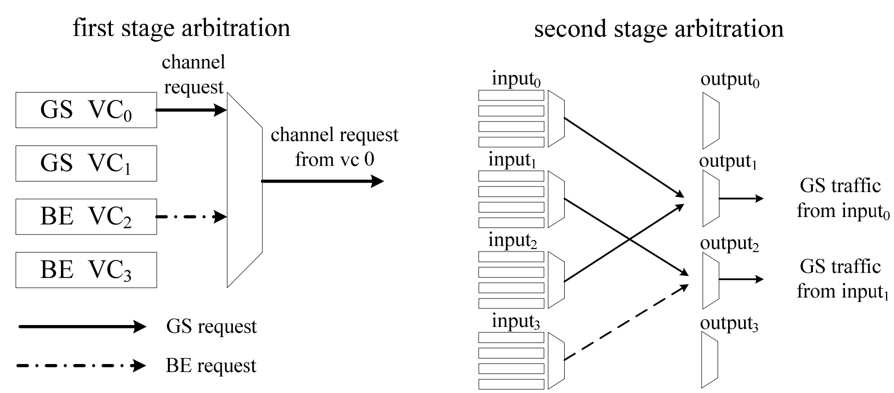




Related Works

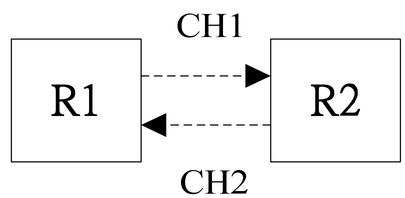
- It is proven that connection less scheme is better in a variable bit rate application.
  - M. D. Harmanci, **"Quantitative modelling and comparison of communication schemes to guarantee Quality-of-Service in Networks-on-Chip"** ISCAS '05
- In a typical connection-less QoS scheme, the packets with different priorities can be adapted to a virtual channel NoC router.
  - E. Bolotin, "**QNoC: QoS architecture and design process for network on chip**", *J. Syst. Architecture: EUROMICRO J '04*



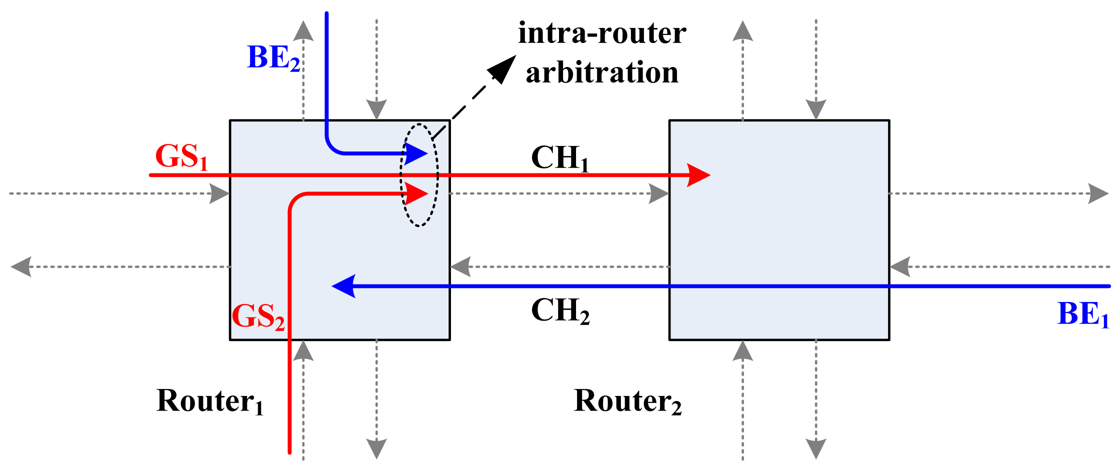


Motivational Example

- Conventional uni-directional inter-router communication channel

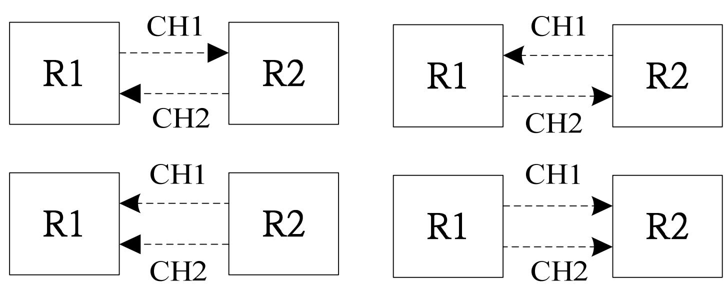


- Under the typical uni-directional NoC, only GS1 is granted while another channel with opposite direction is used by the BE1 flow with the lower QoS requirement.

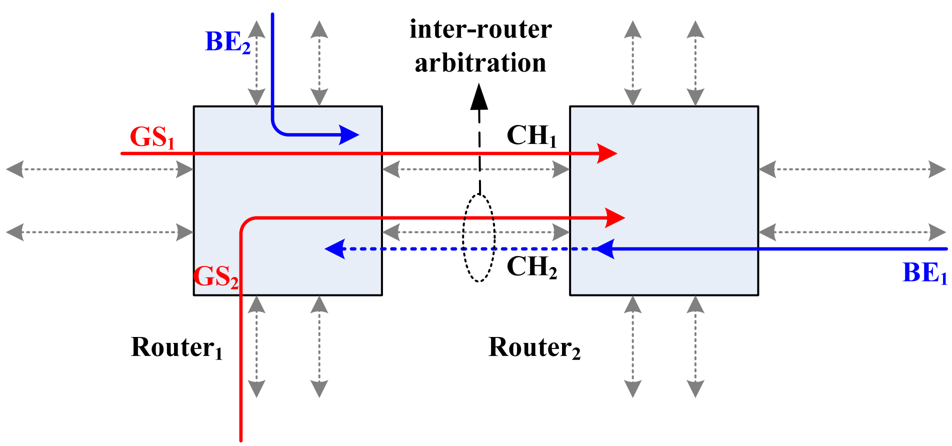


Motivational Example

- Enhance real-time traffic routing flexibility by dynamic reconfigurable bi-directional channel.

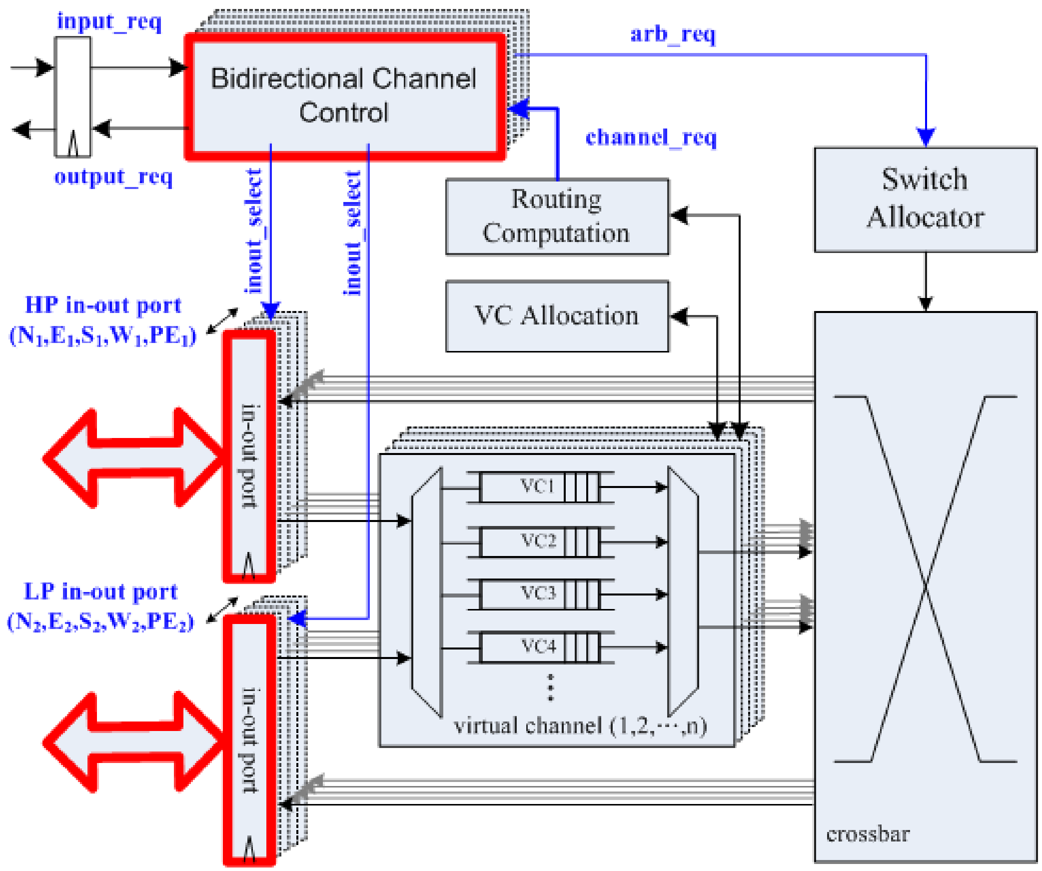


- The inter-router arbitration can be applied to further enhance the channel usage priority for the GS traffic.



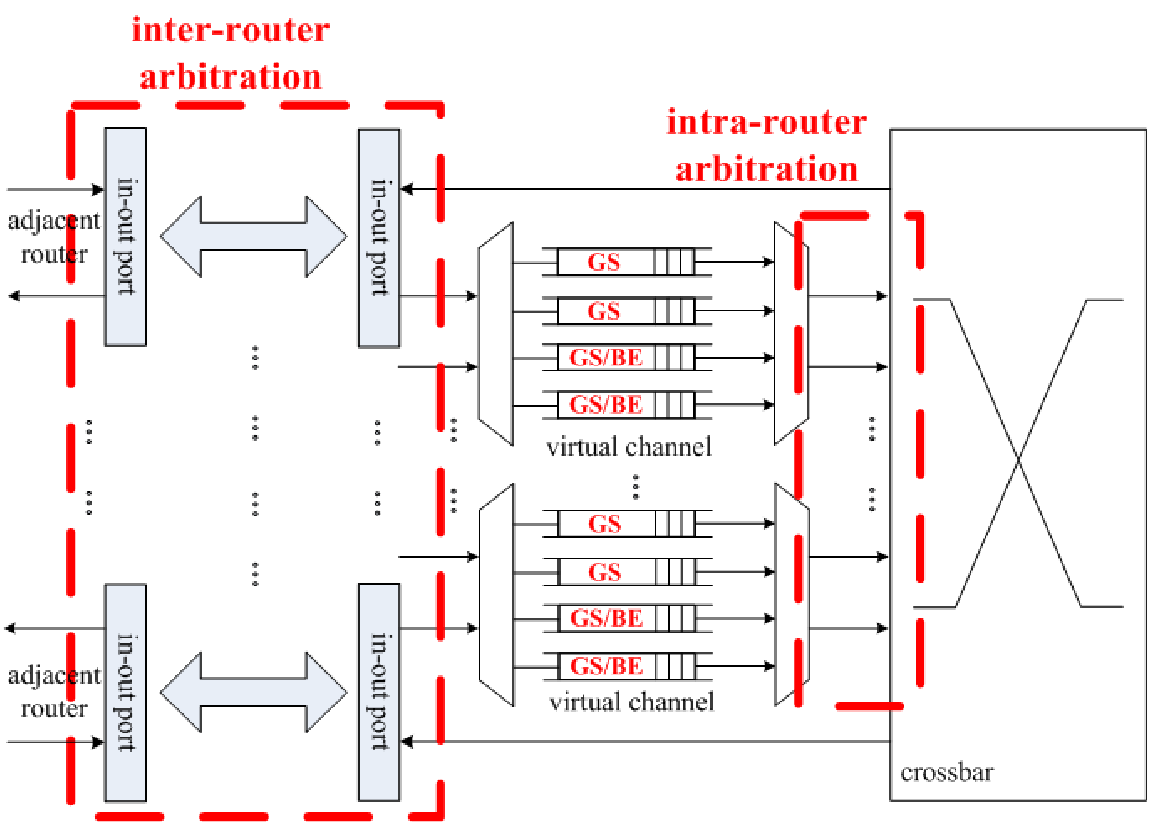
QoS Aware BiNoC Architecture

- Bidirectional channel direction control module are implemented for inter-router arbitration.



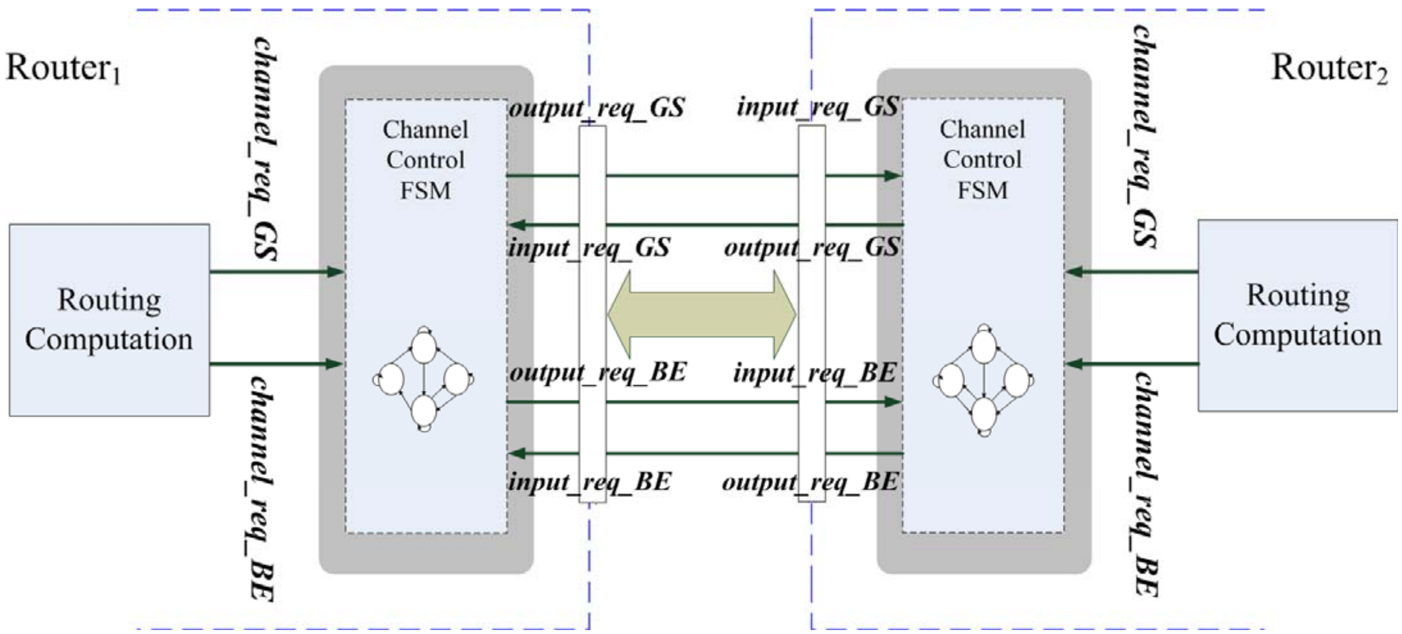
Prioritized Virtual Channel Management and Inter-router Arbitration

- GS packets always has the higher priority to get the output bandwidth during the intra-router arbitration.
- inter-router arbitration improve the channel utilization for GS packets.



Inter-Router Channel Direction Control Scheme

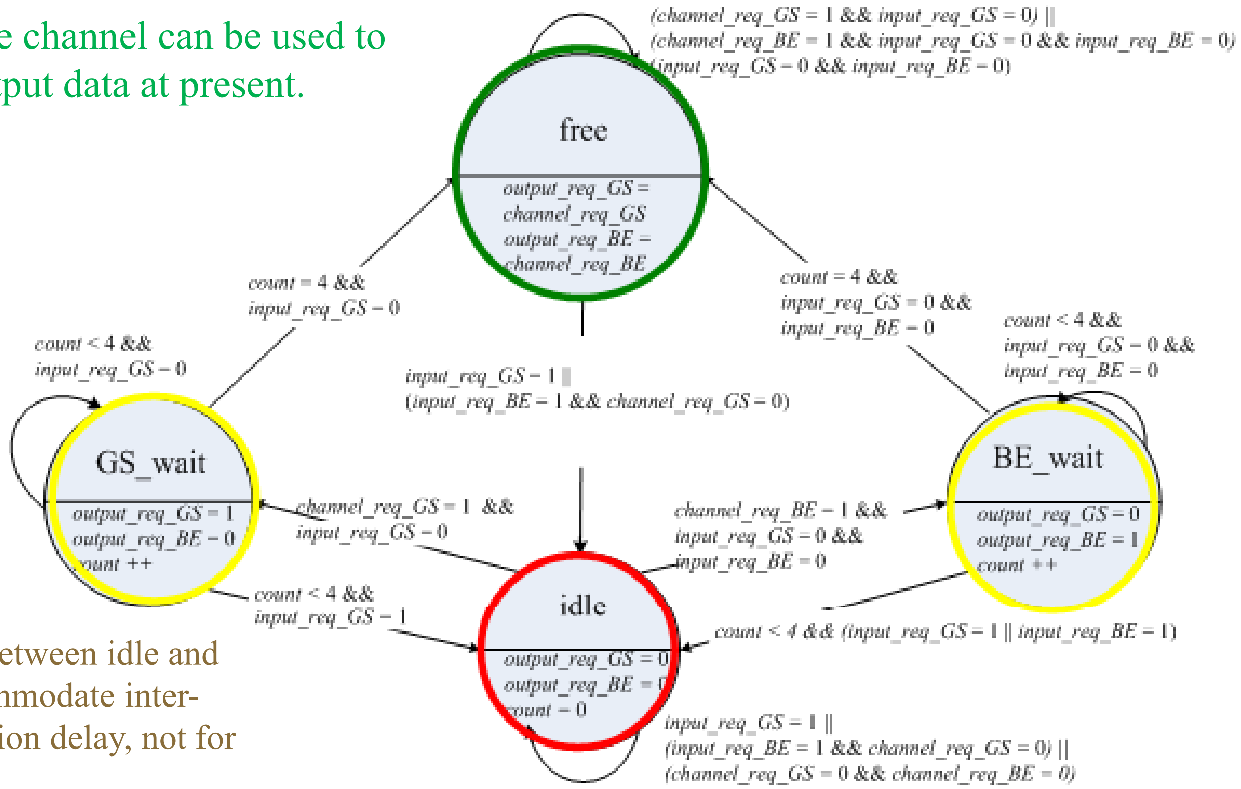
- Configuration of a bi-directional channel is managed by a finite state machine in the channel control modules.



Inter-Router Transmission Scheme

- The channel state reflects whether this port can be used to output data currently or not.

The channel can be used to output data at present.

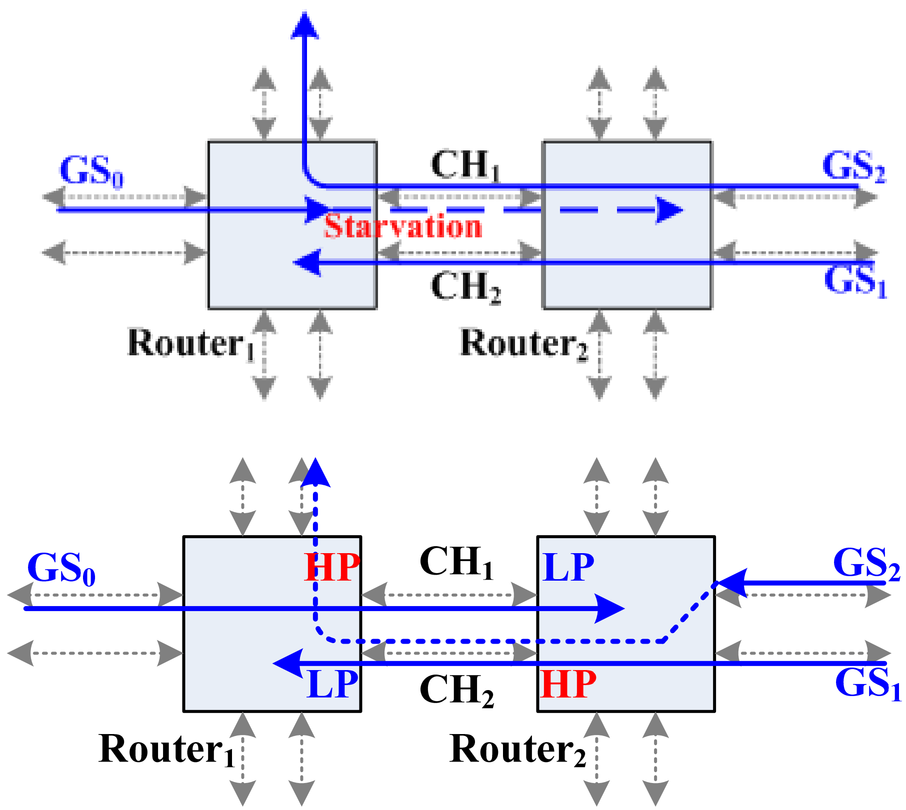


Transitional state between idle and free , used to accommodate inter-router communication delay, not for output currently.

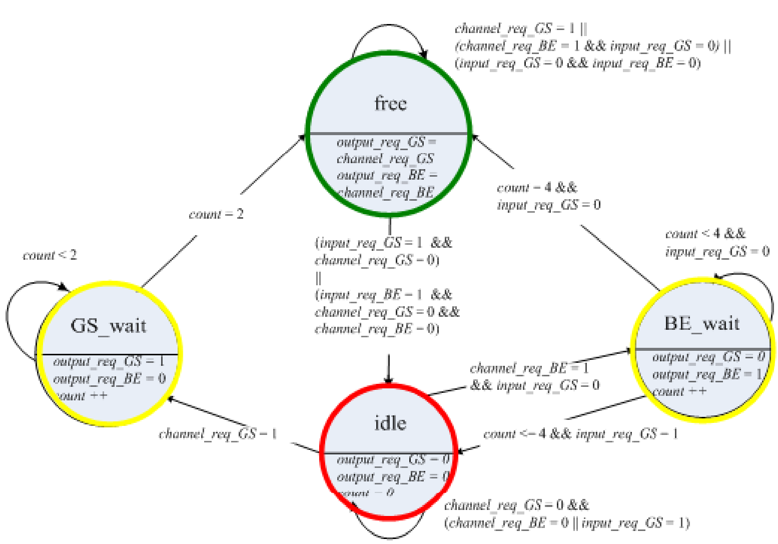
The channel direction is inward, not for output currently.

Starvation Avoidance

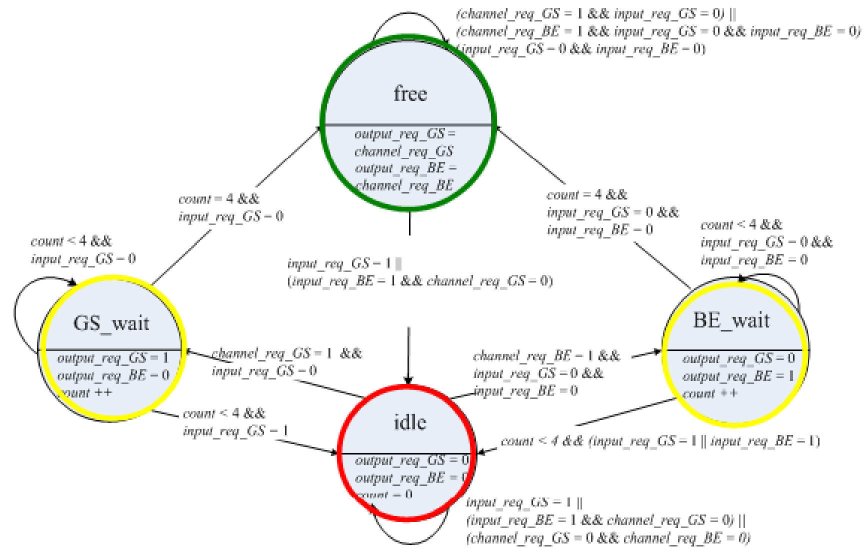
- To prevent the inter-router starvation problem, one of these two FSMs will be designated with a higher priority (HP) and the other with a lower priority (LP).



Prioritized Channel Control FSM



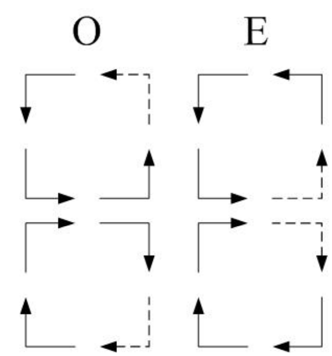
High priority FSM



Low priority FSM

Prioritized Routing Restriction

- A prioritized routing restriction is applied to leave more available communication bandwidth for GS traffic.
  - BE traffic : deterministic routing
  - GS traffic : adaptive routing
  - Odd-Even Turn model is applied to prevent deadlock



G. M. Chiu, "**The Odd-Even Turn Model for Adaptive Routing**," *IEEE Transactions on Parallel and Distributed Systems*, vol. 11, no. 7, pp. 729-738, Jul. 2000.

- The prioritized routing can help the GS traffic to exploit more channel resource for transmission.

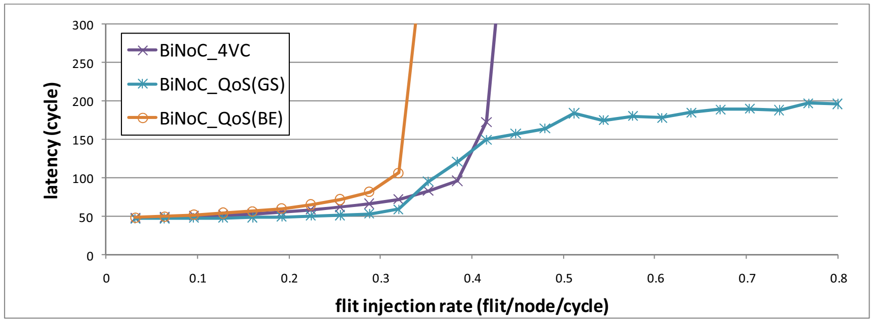
Experimental Setup

- Simulation setup
  - 8x8 2-D mesh
  - Cycle accurate HDL simulation
  - 32 flit buffer implemented in 4 virtual channels in each direction
  - Uniform, transpose and hotspot traffic

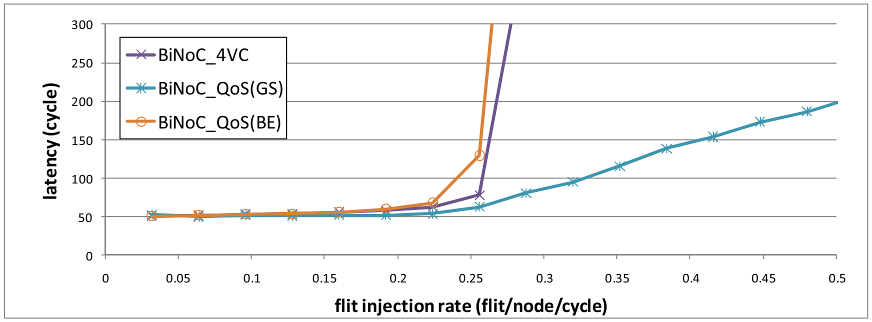
Architecture	NoC_QoS	BiNoC_4VC	BiNoC_QoS	BiNoC_QoS_OE
QoS Mechanism	GS/BE	None	GS/BE	GS/BE
Total Channels	5-in 5-out	10-inout	10-inout	10-inout
GS Routing	deterministic	deterministic	deterministic	adaptive

Experimental Results

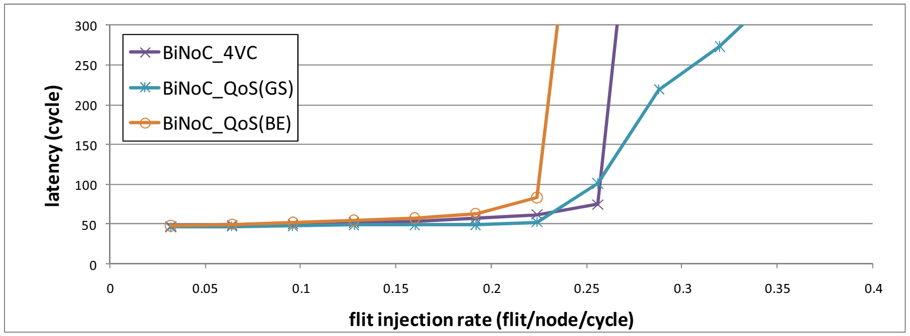
- latency results between BiNoC_4VC, and BiNoC_QoS.
- GS traffic occupies 20% of the total traffic



uniform



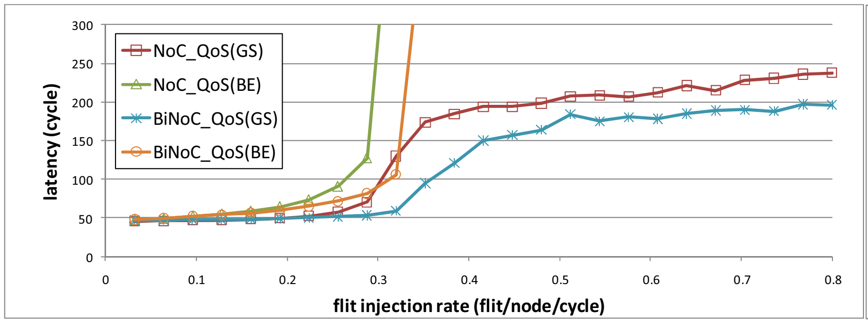
transpose



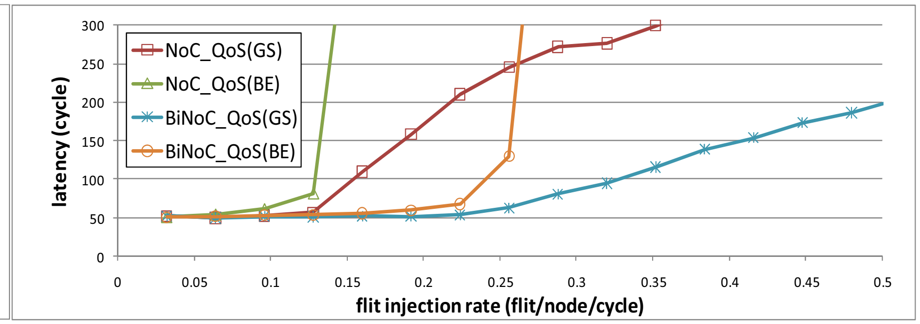
hotspot

Experimental Results

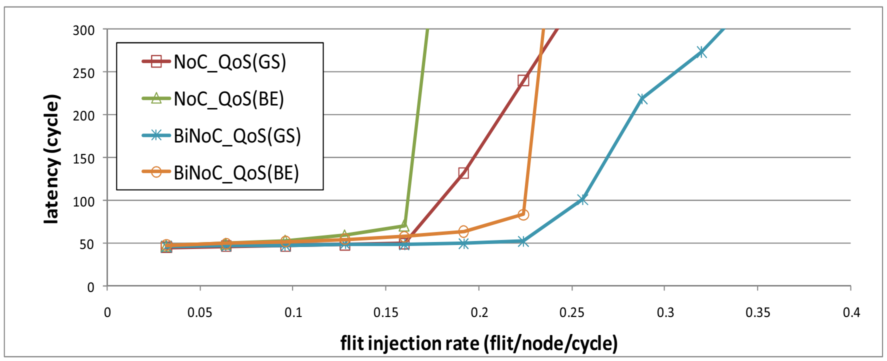
- latency results between NoC_QoS, and BiNoC_QoS.
- Inter-router arbitration can further reduce the latency of GS packets because of the doubled bandwidth utilization flexibility



uniform



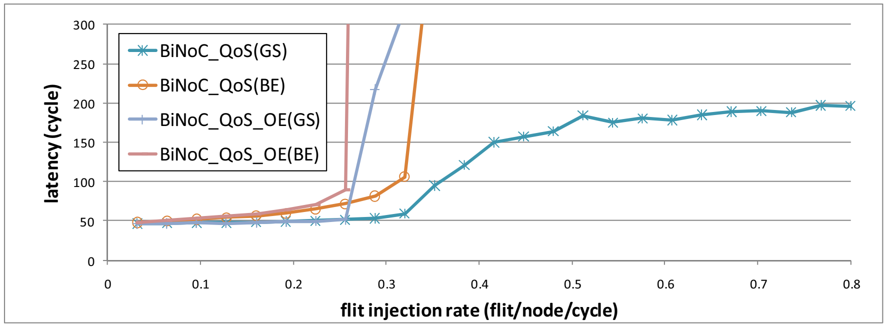
transpose



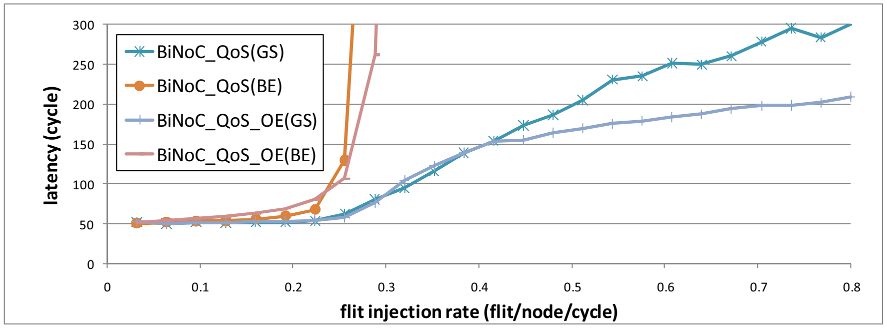
hotspot

Experimental Results

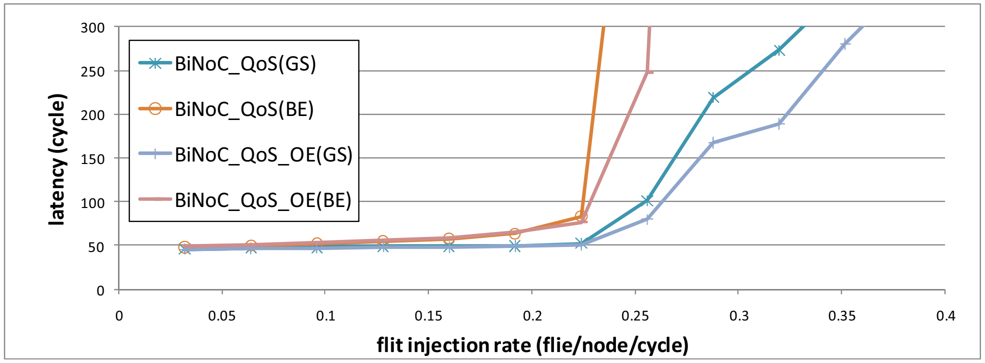
- latency results between BiNoC_QoS, and BiNoC_QoS_OE.
- The prioritized routing restriction can help the GS packets to avoid the blocking nodes thus reduces the latency.



uniform



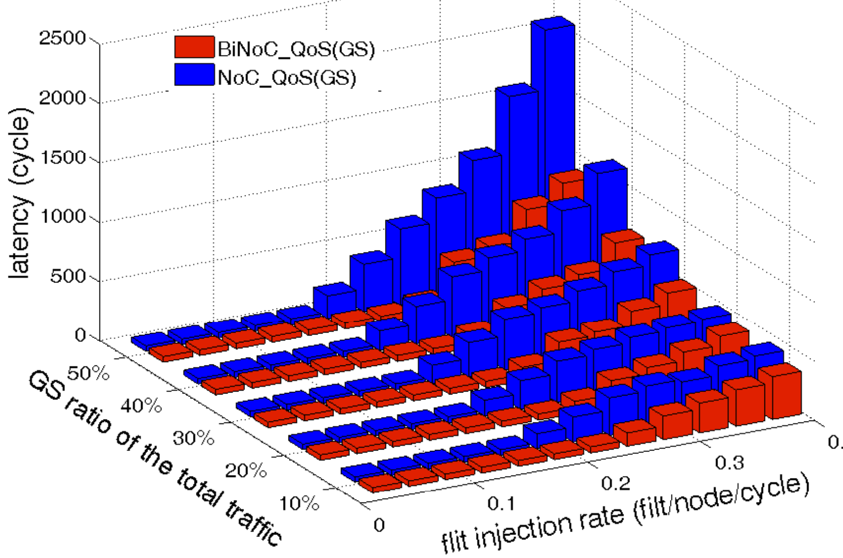
transpose

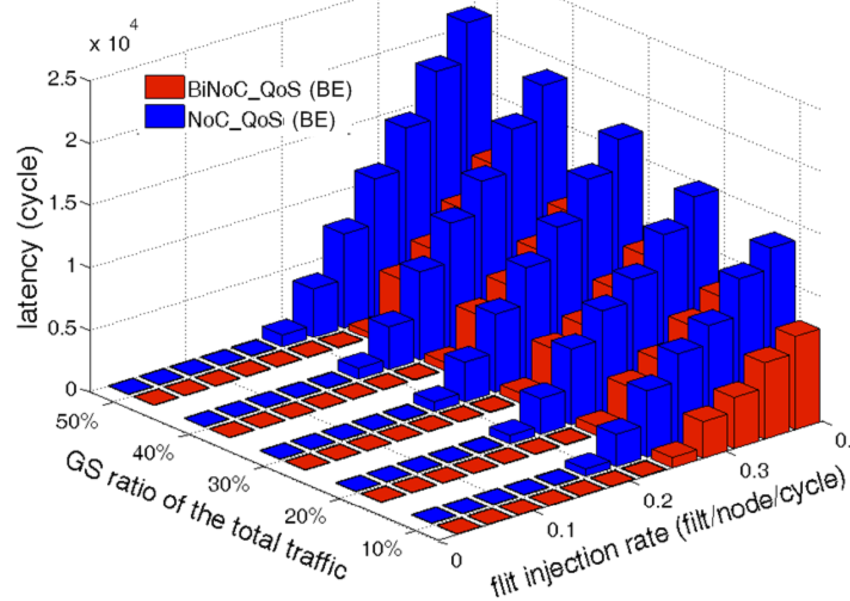


hotspot

Experimental Results

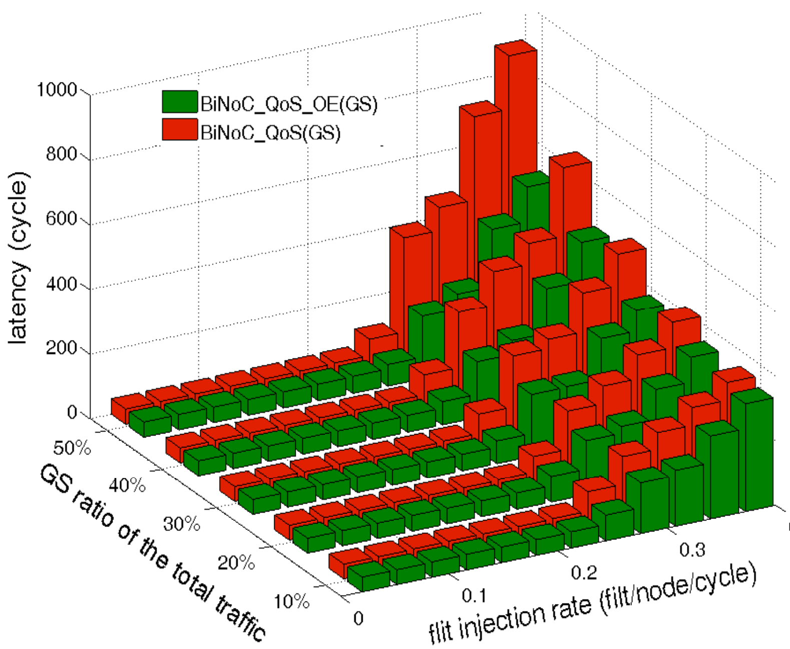
- latency results between NoC_QoS, and BiNoC_QoS in various GS ratios of the total traffic.
- Results are obtained by running hotspot traffic.

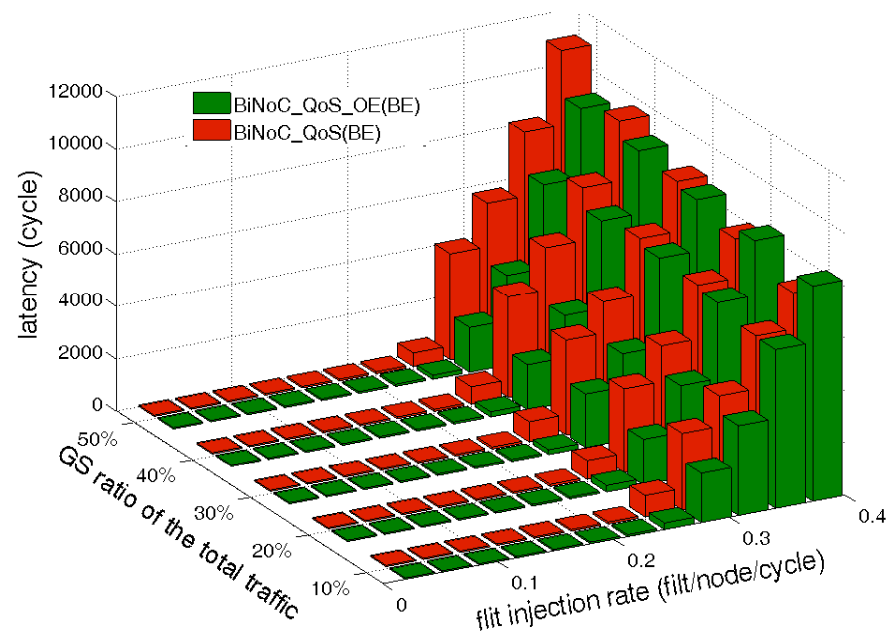




Experimental Results

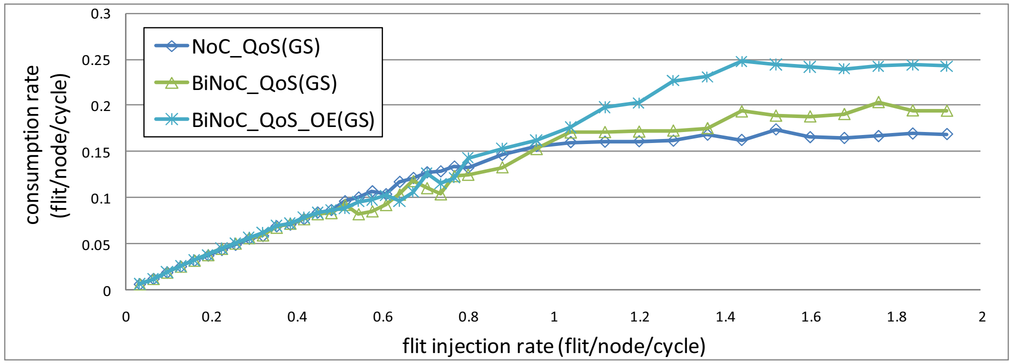
- latency results between BiNoC_QoS, and BiNoC_QoS_OE in various GS ratios of the total traffic.
- Results are obtained by running hotspot traffic.

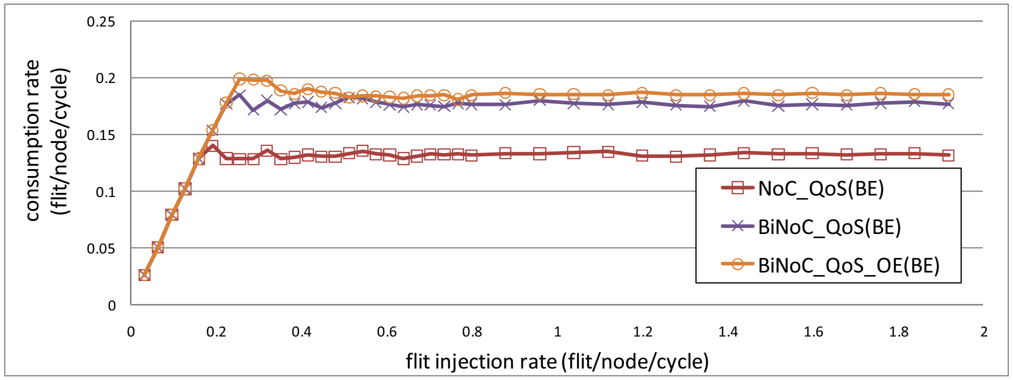




Experimental Results

- Flit consumption rate of GS and BE packets under hotspot traffic with 20% GS ratio to the total traffic.





Closing Remarks

- A connection-less QoS mechanism based on the bi-directional channel NoC (BiNoC) backbone is proposed.
- A flexible virtual channel management mechanism and a novel prioritized routing policy are integrated
- the proposed inter-router arbitration scheme can significantly improve the channel utilization for GS packets