# Design and Evaluation of Hierarchical Rings with Deflection Routing

**Rachata Ausavarungnirun**, Chris Fallin, Xiangyao Yu,
Kevin Chang, Greg Nazario, Reetuparna Das,
Gabriel H. Loh, Onur Mutlu

Carnegie Mellon

SAFARI

AMD

MICHIGAN

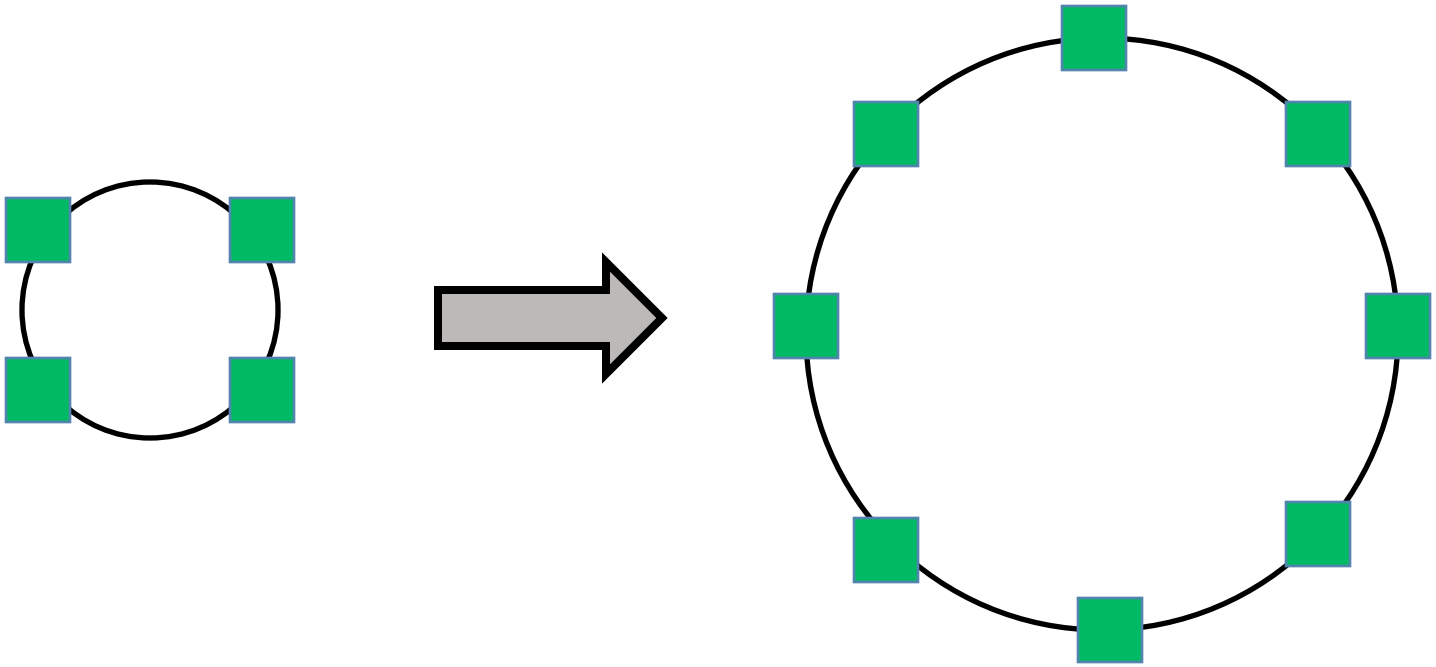
# Executive Summary

- Rings do not scale well as core count increases
- Traditional hierarchical ring designs are complex and energy inefficient
  - Complicated buffering and flow control
- **Solution:** Hierarchical Rings with Deflection (HiRD)
  - Guarantees livelock freedom and delivery
  - Eliminates all buffers at local routers and most buffers at bridge routers
- HiRD provides higher performance and energy-efficiency than hierarchical rings
- HiRD is simpler than hierarchical rings

# Outline

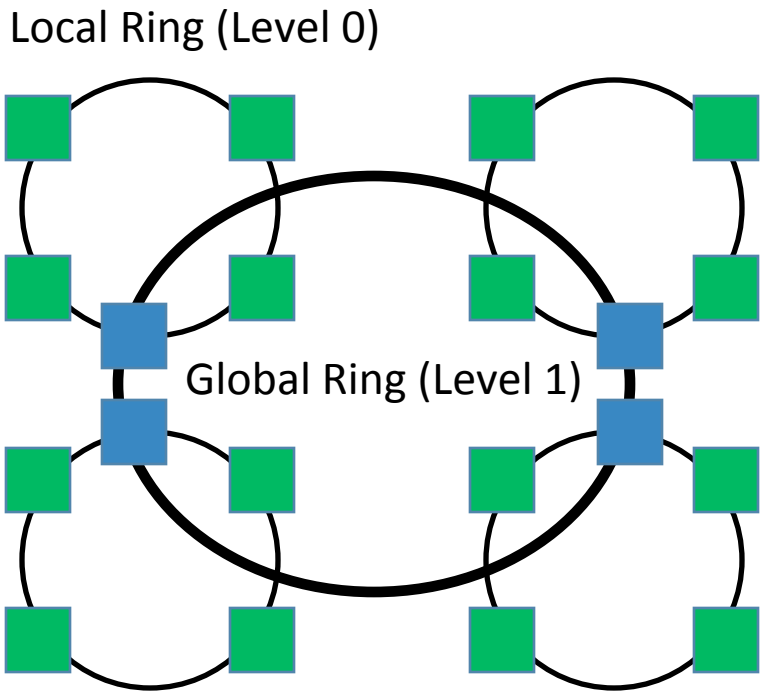
- **Background and Motivation**
- Key Idea: Deflection Routing
- End-to-end Delivery Guarantees
- Our Solution: HiRD
- Results
- Conclusion

# Scaling Problems in a Ring NoC



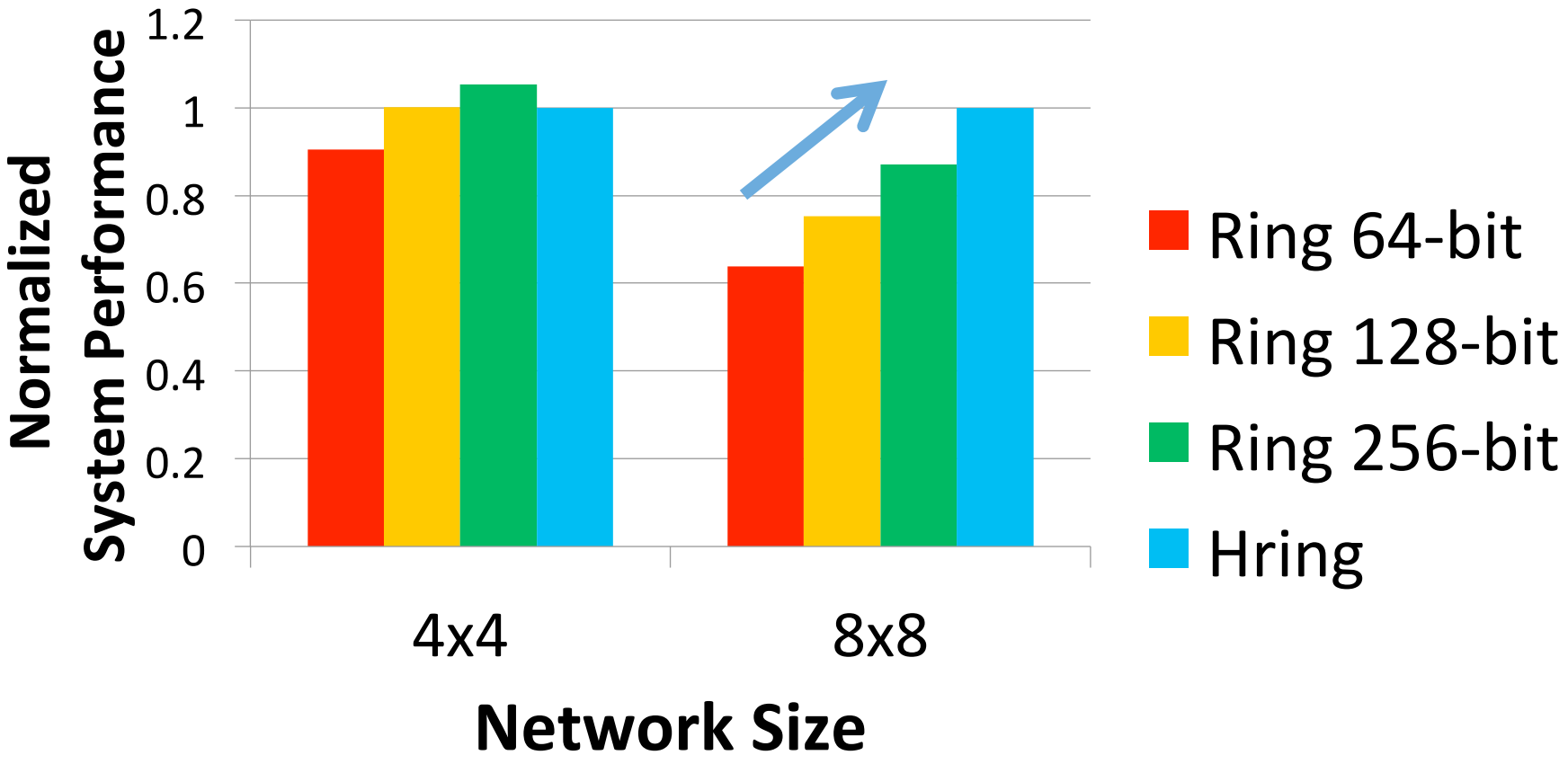
- As the number of cores grows:
  - Lower performance
  - More power

# Alternative: Hierarchical Designs



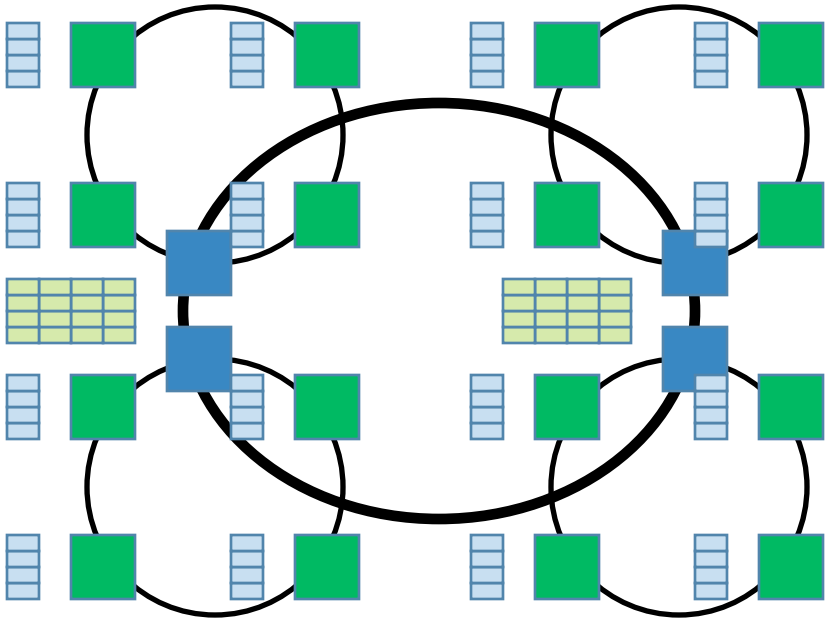

Packets can reach far destination in fewer hops

# Single Ring vs. Hierarchical Rings



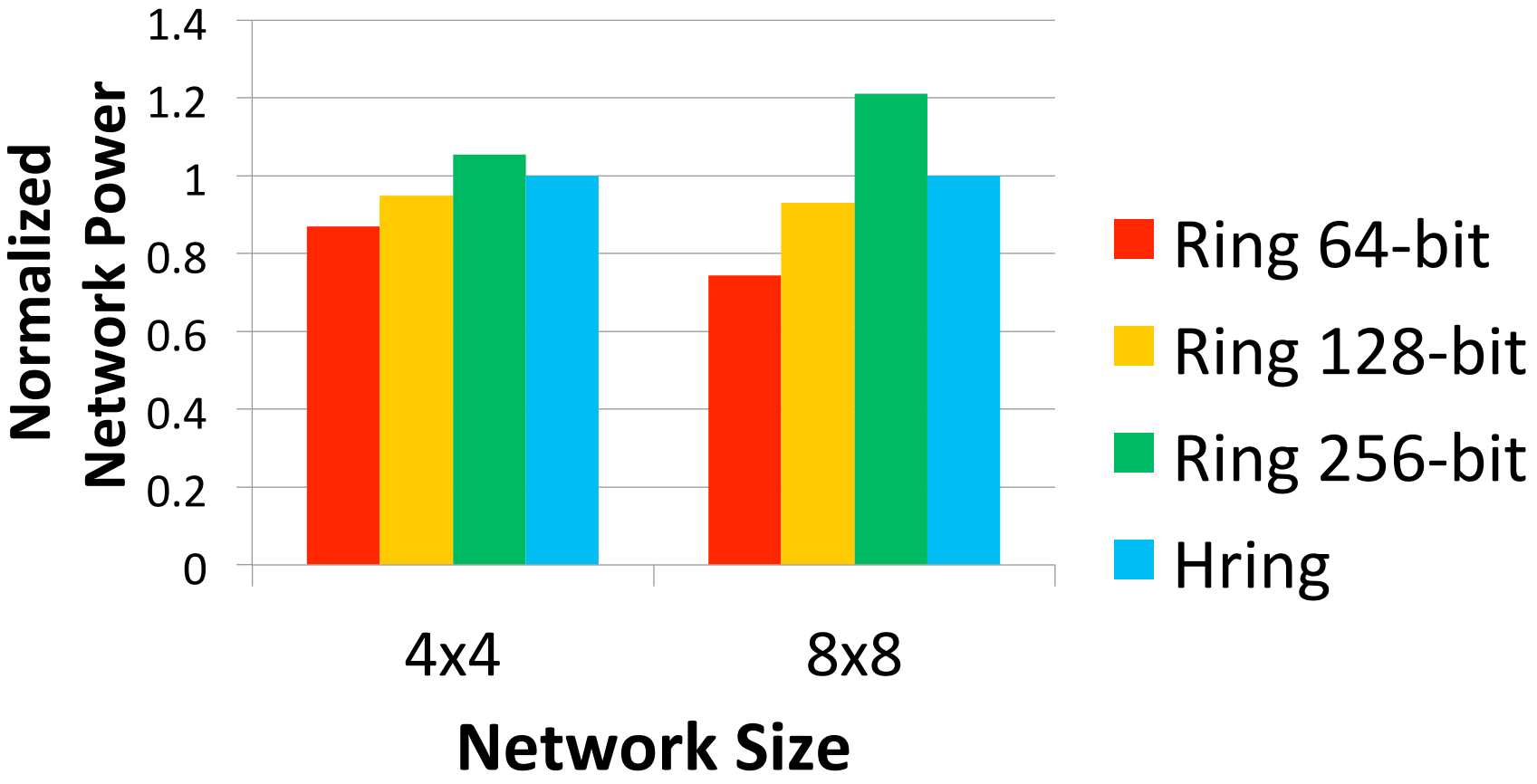

A hierarchical design provides better performance as the network scales

# Complexity in Hierarchical Designs



Complex buffering and flow control

# Single Ring vs. Hierarchical Rings




Design complexity increases power consumption

# Our Goal

- Design a hierarchical ring that
  has lower complexity
  without sacrificing performance

# Outline

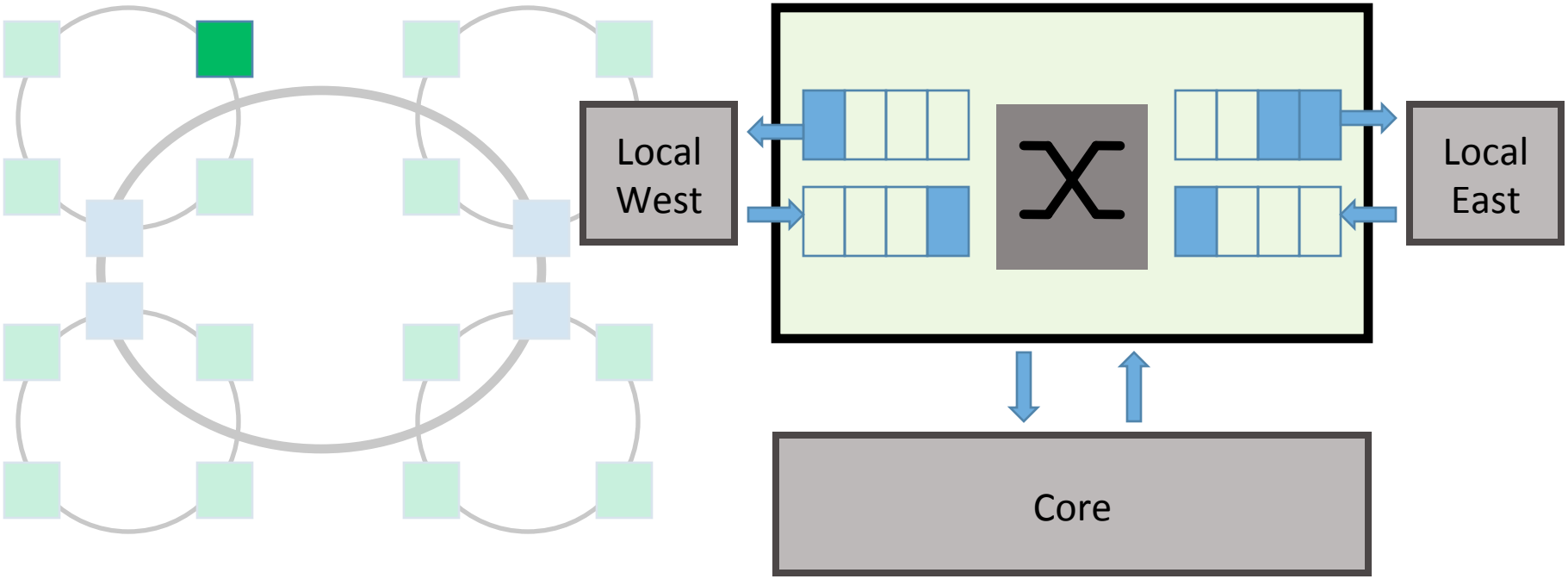
- Background and Motivation
- **Key Idea: Deflection Routing**
- End-to-end Delivery Guarantees
- Our Solution: HiRD
- Results
- Conclusion

# Key Idea

- Eliminate buffers

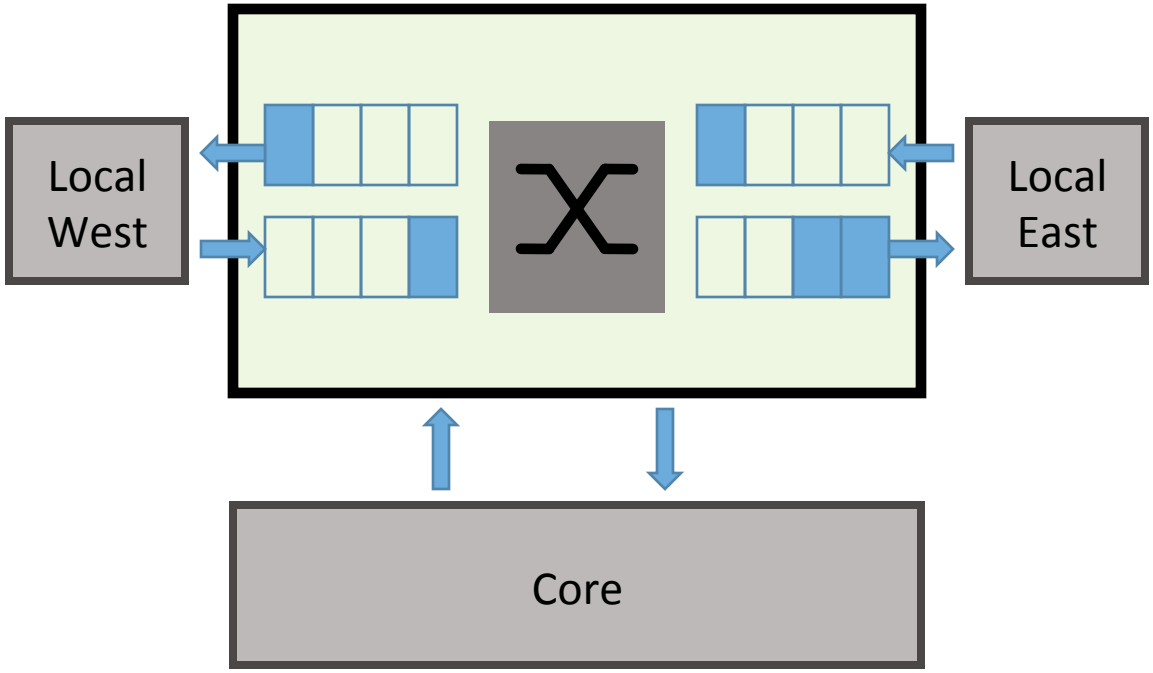
- Use deflection routing
  - Simpler flow control

# Local Router



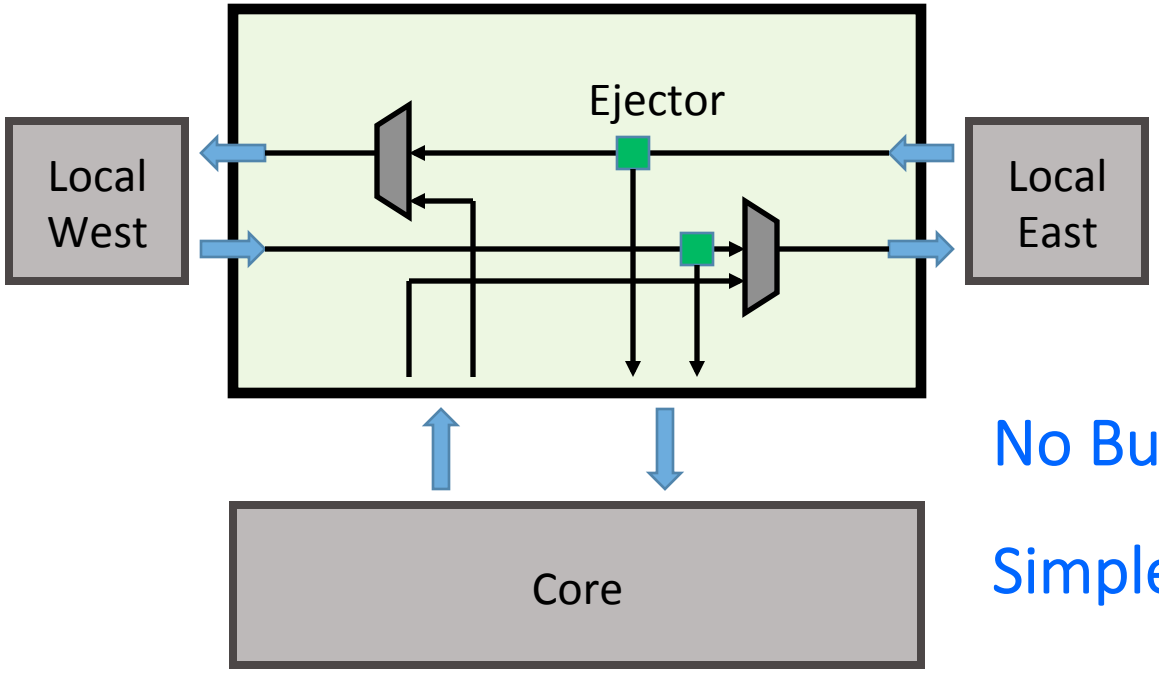

- Key functionality:
  - Accept new flits
  - Pass flits around the ring

# Eliminating Buffers in Local Routers



# Eliminating Buffers in Local Routers

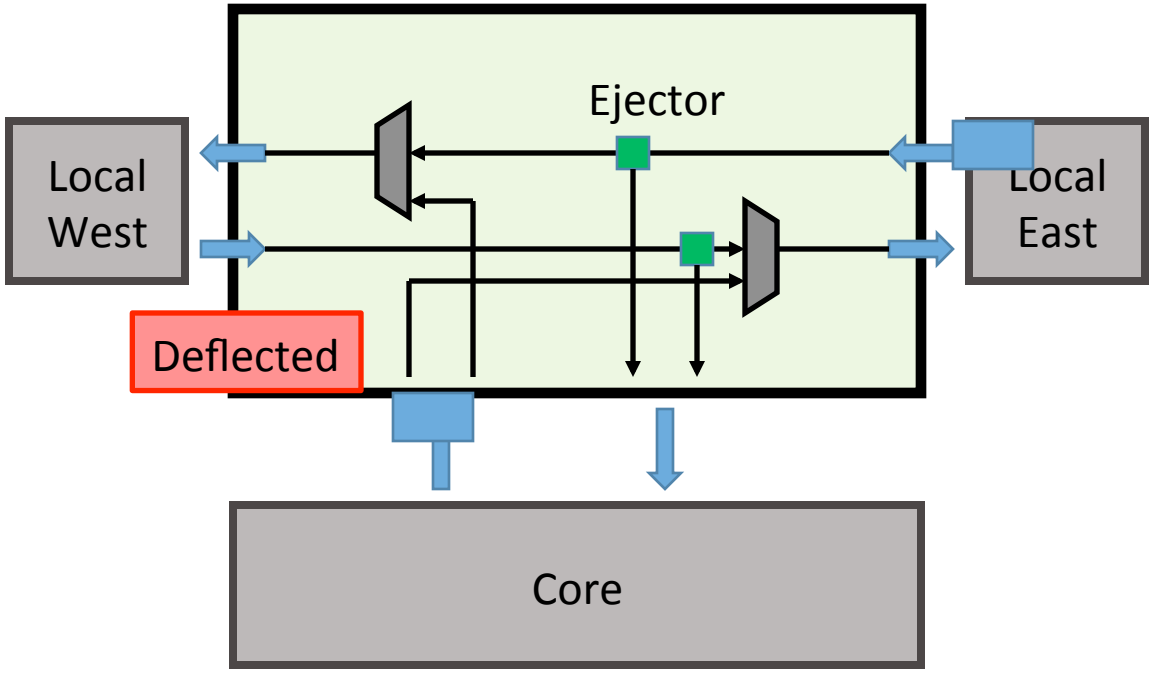
- Flits can enter the ring if the output is available



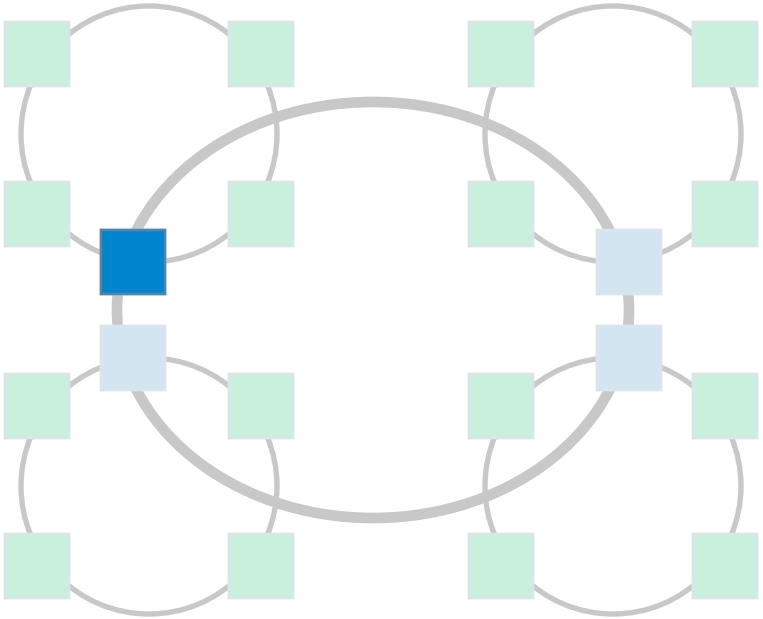

No Buffer

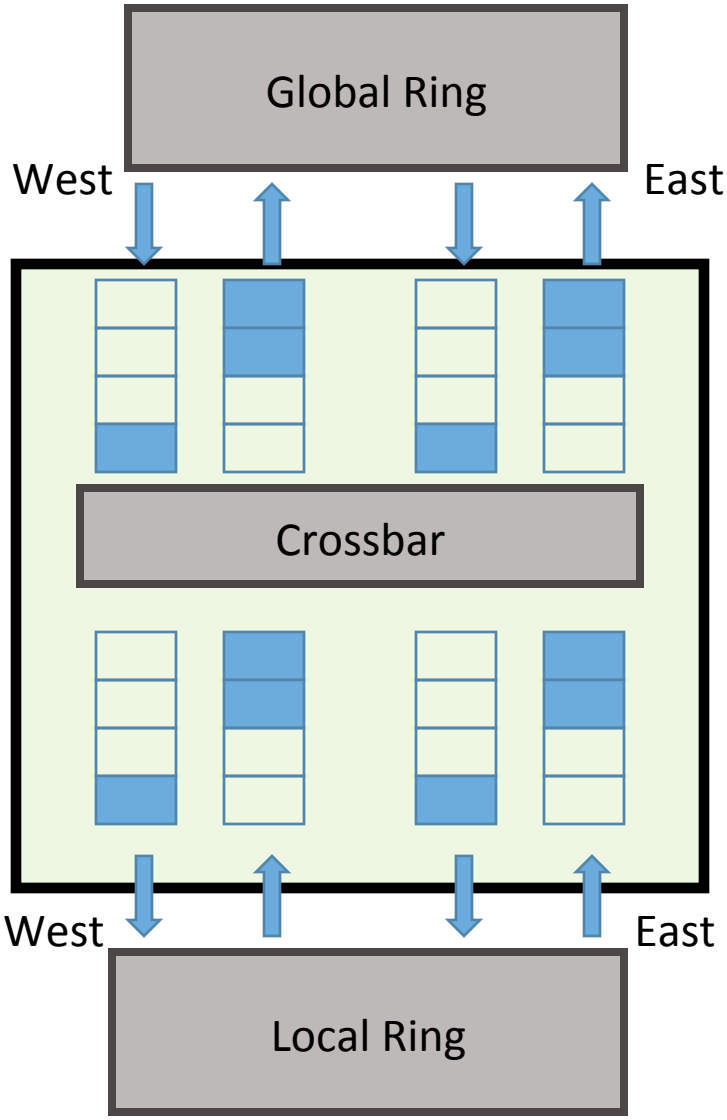
Simpler Crossbar

# Deflection Routing

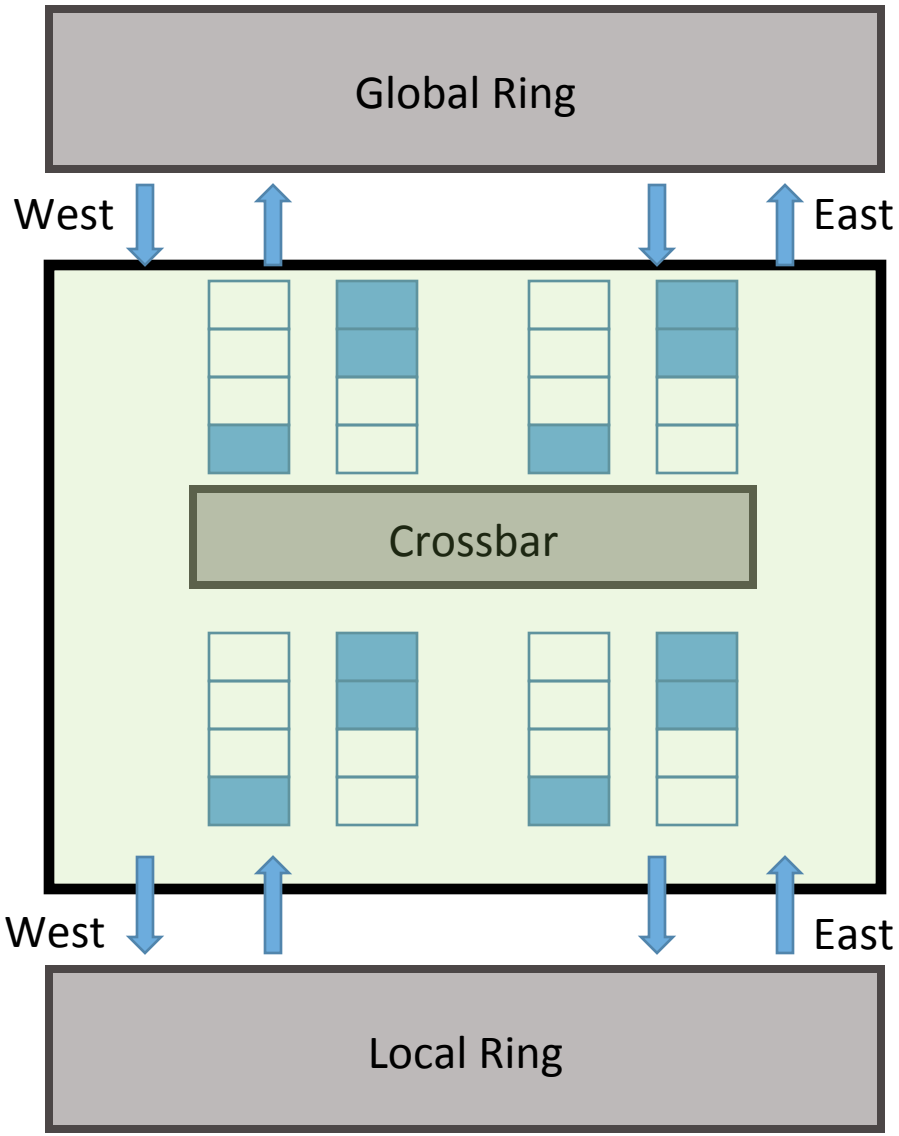


# Bridge Router

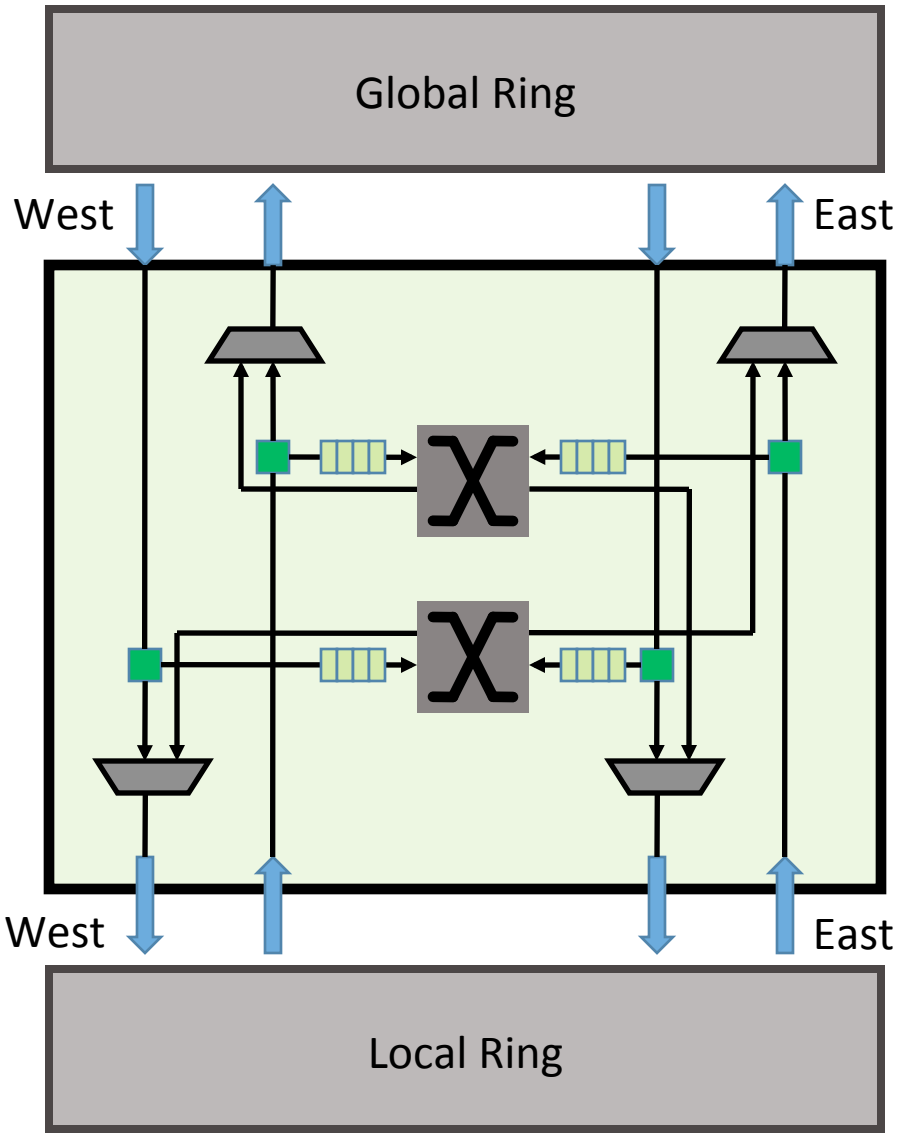




# Eliminating Buffers in Bridge Routers



# Eliminating Buffers in Bridge Routers




Simpler Buffering

Fewer Buffers

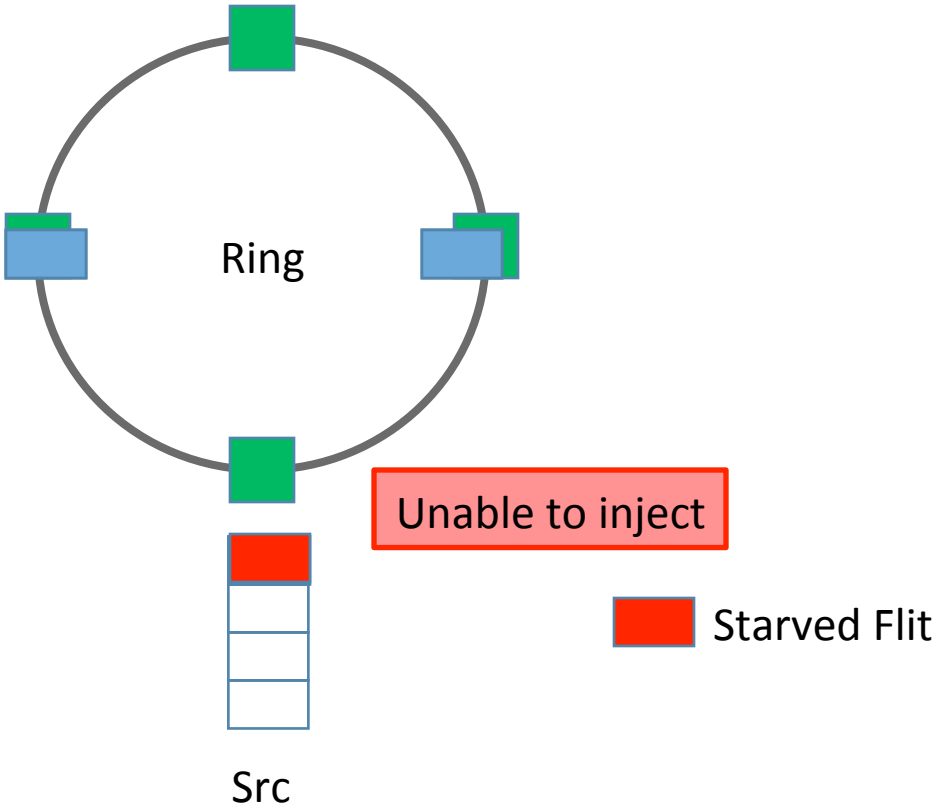
Simpler Crossbar

# Outline

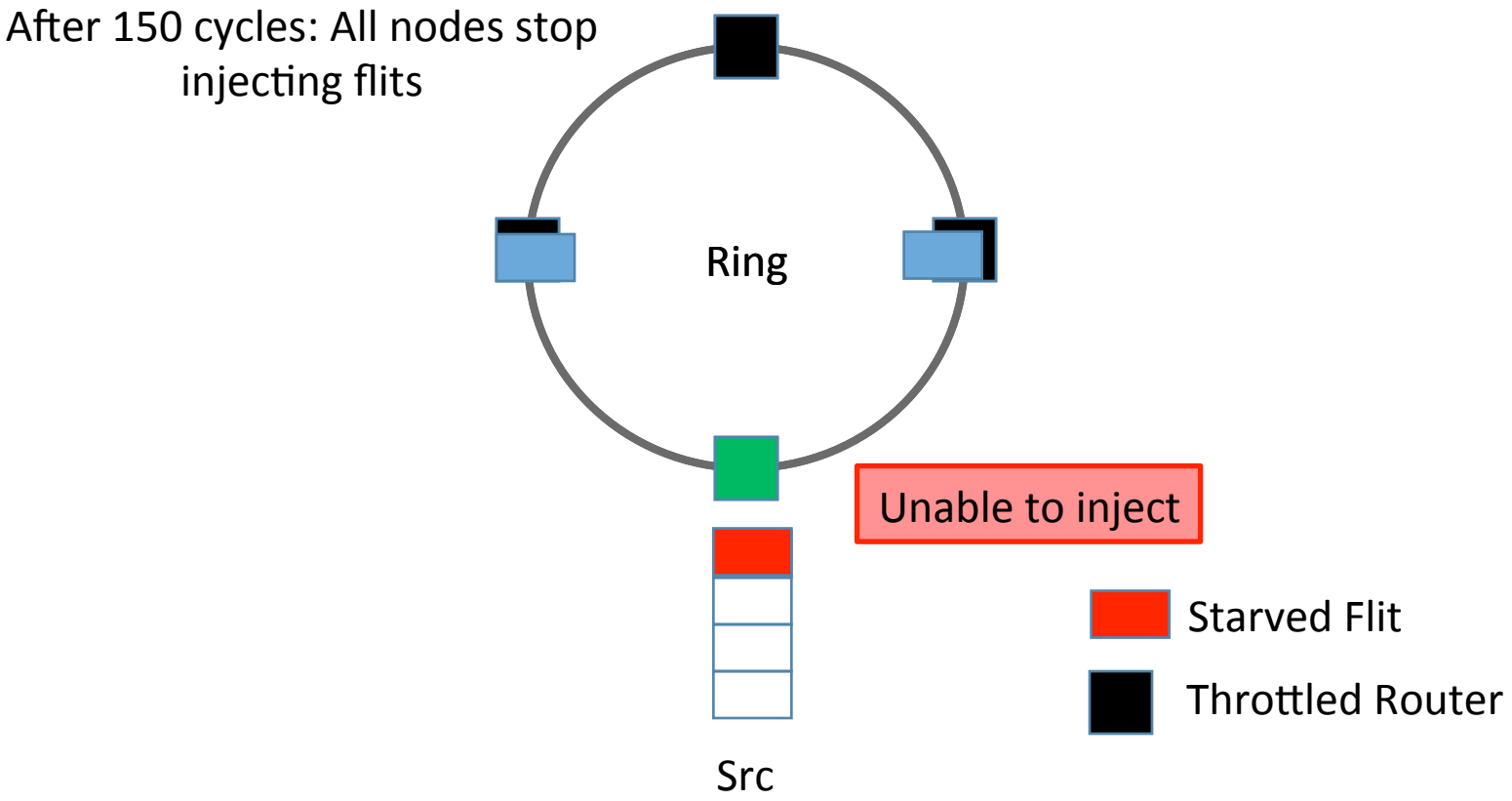
- Background and Motivation
- Key Idea: Deflection Routing
- **End-to-end Delivery Guarantees**
- Our Solution: HiRD
- Results
- Conclusion

# Livelock in Deflection Routing

- Injection starvation



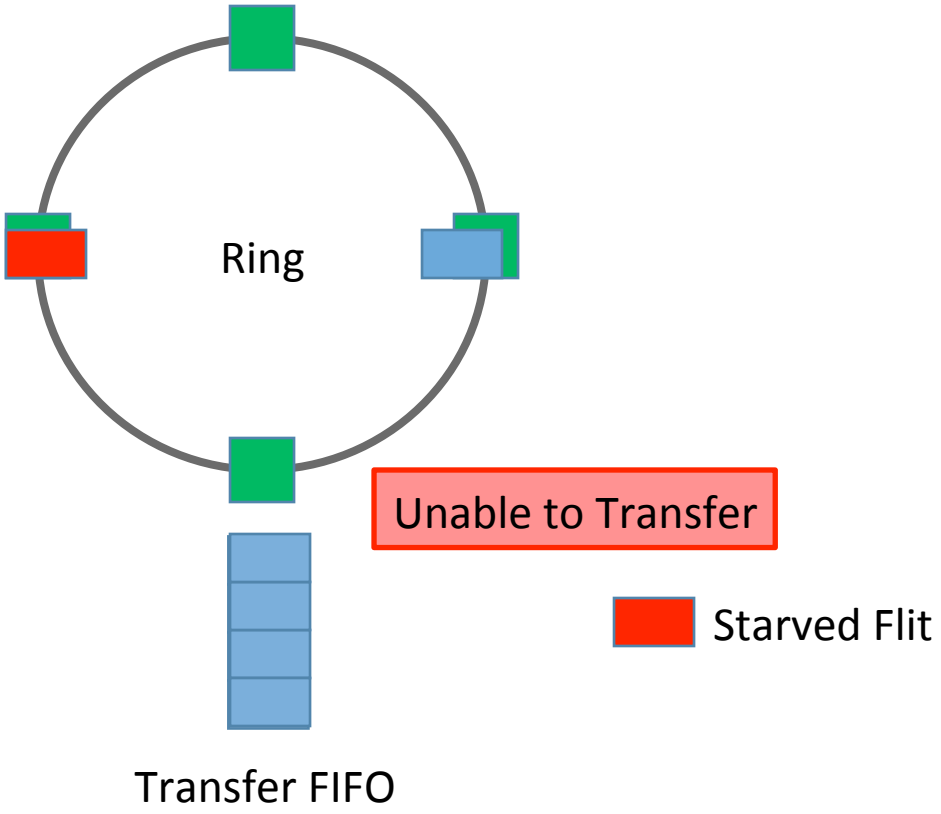
# HiRD: Injection Guarantee



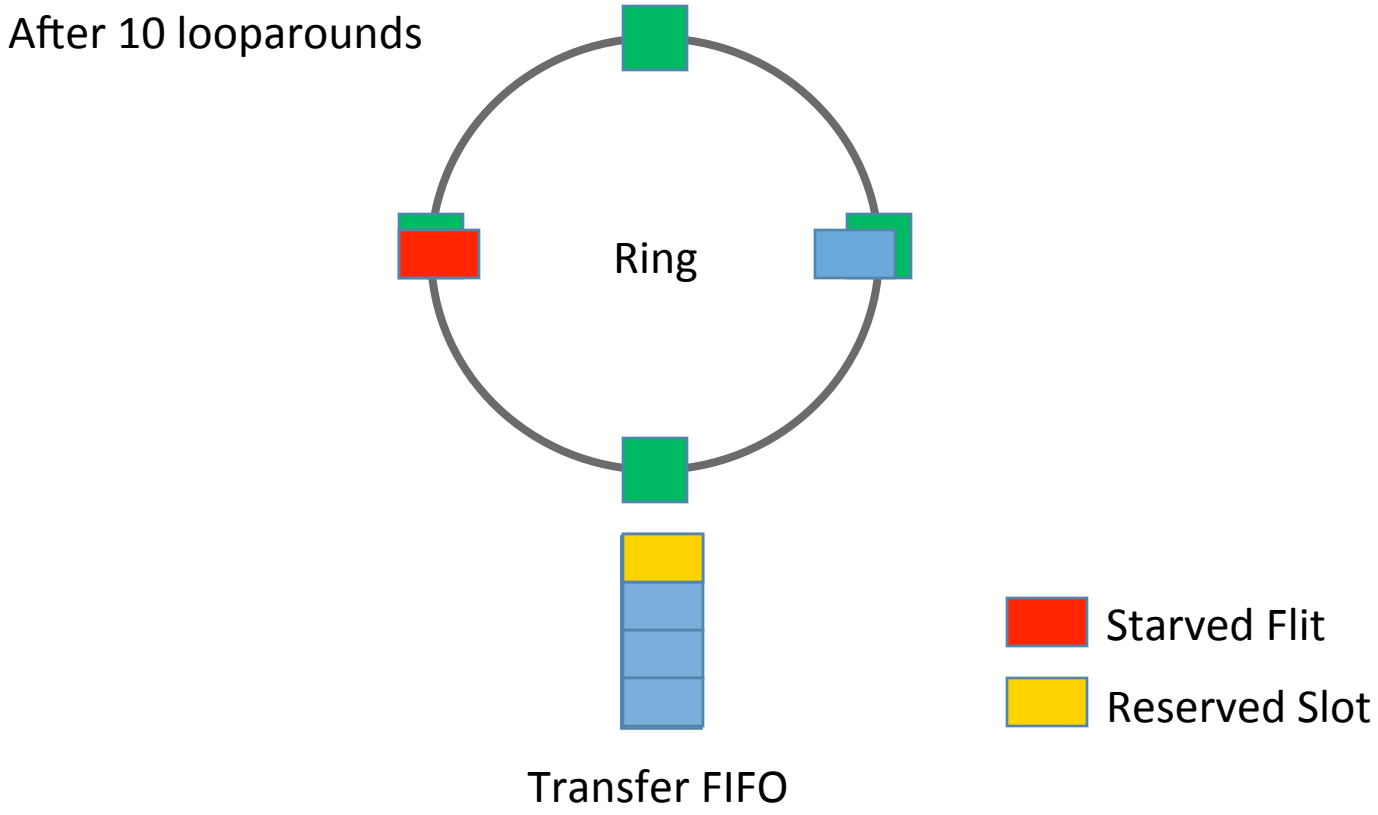

- Throttling provides injection guarantee

# Livelock in Deflection Routing

- Transfer starvation



# HiRD: Transfer Guarantee



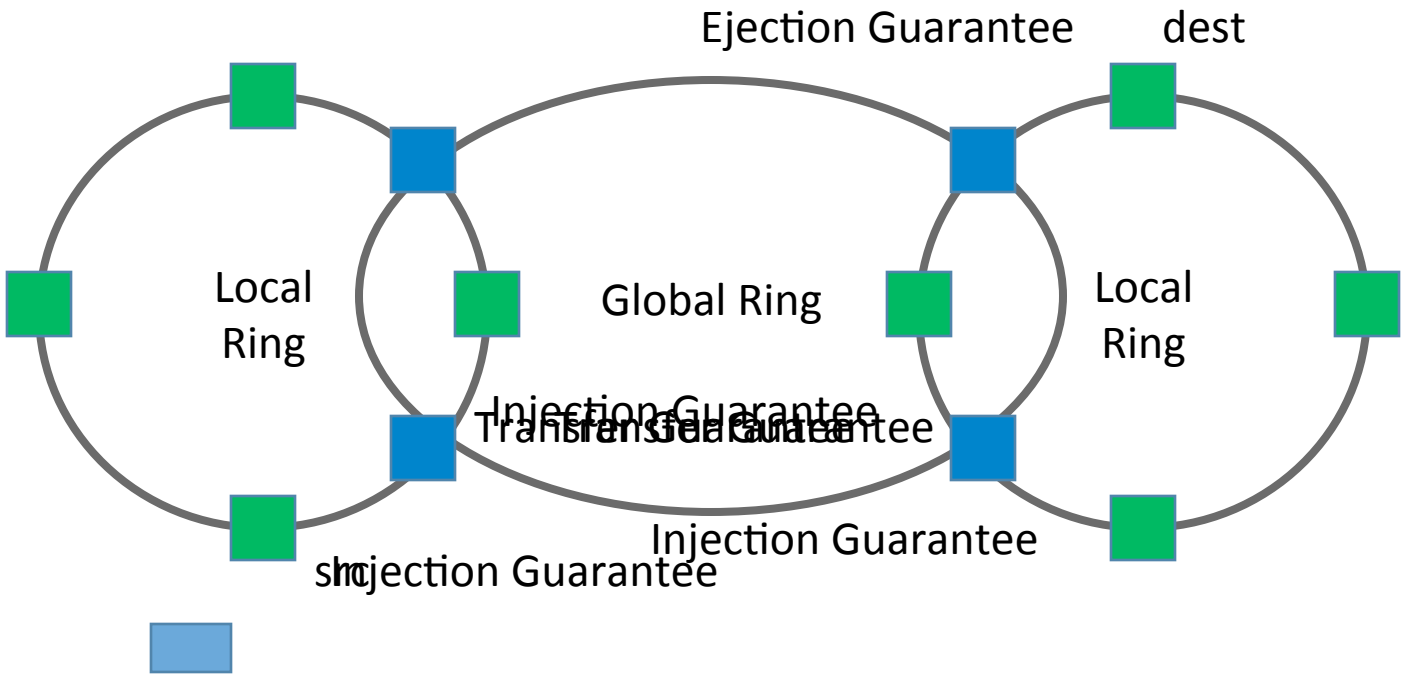

- Reservation provides transfer guarantee

# Ejection Guarantee

- Provided by a prior work

- Re-transmit once [Fallin et al., HPCA'11]
  - Drop a flit if there is no available slot
  - Reserve a buffer slot at the destination if a flit was dropped

# End-to-end Delivery Guarantees



# Outline

- Background and Motivation
- Key Idea: Deflection Routing
- End-to-end Delivery Guarantees
- **Our Solution: HiRD**
- Results
- Conclusion

# An Overview of HiRD

- Deflection routing
- No buffers in the local rings
- Simpler bridge routers
- Provides end-to-end delivery guarantees
  - Injection guarantee by throttling
  - Transfer guarantee by reservation

# Putting It All Together

- Deflection routing
  - Simpler flow control
  - Simpler crossbars and control logic
- No buffers in the local rings
  - Simpler and faster local routers
- Simpler bridge routers
  - Lower power, less area and simpler to design
- Provides end-to-end delivery guarantees
  - Injection guarantee by throttling
  - Transfer guarantee by reservation

# Outline

- Background and Motivation
- Key Idea: Deflection Routing
- End-to-end Delivery Guarantees
- Our Solution: HiRD
- **Results**
- Conclusion

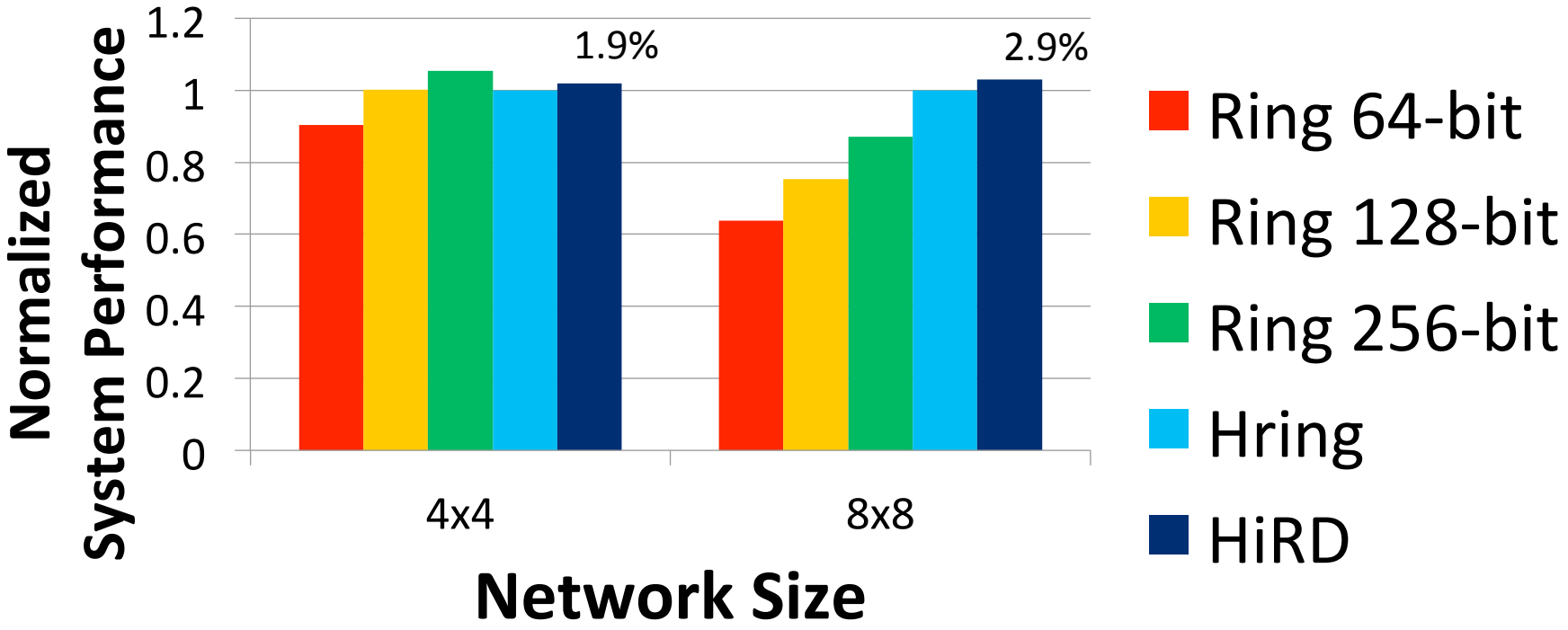
# Methodology

- Cores
  - 16 and 64 OoO CPU cores
  - 64 KB 4-way private L1
  - Distributed L2
- Network
  - 1 flit local-to-global buffer
  - 4 flits global-to-local buffers
  - 2-cycle per hop latency for local routers
  - 3-cycle per hop latency for global routers
- 60 workloads consisting of SPEC2006 apps

# Comparison to Previous Designs

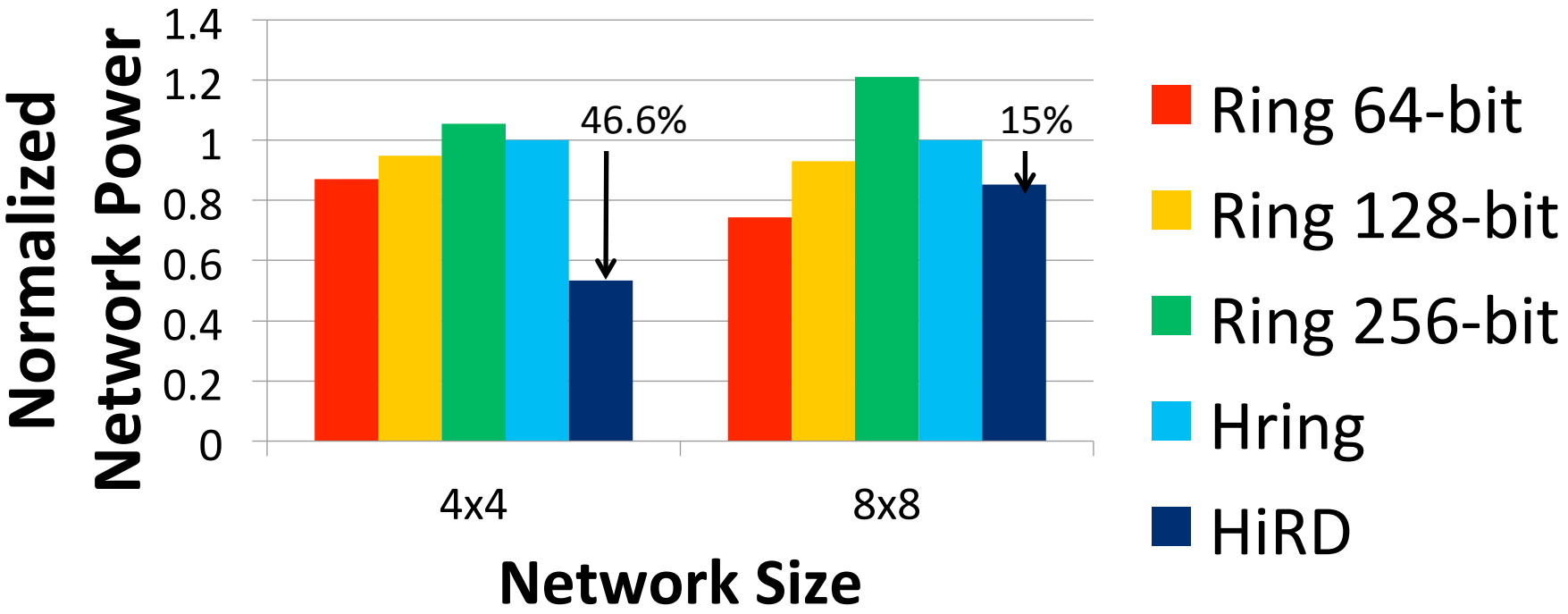
- Single ring design
  - Kim and Kim, NoCArc'09
  - 64-bit links
  - 128-bit links
  - 256-bit links
- Buffered hierarchical ring design
  - Ravindran and Stumm, HPCA'97
  - Identical topology
  - Identical bisection bandwidth
  - 4-flit buffers in both local and global routers

# Results: System Performance




1) Hierarchical designs provide better performance than a single ring on a larger network
2) HiRD performs better compared to buffered hierarchical rings due to lower latency in local routers and throttling

# Results: Network Power




1) Hierarchical designs consume much less power than the highest-performance single ring
2) Routers and flow control in HiRD are simpler than routers in buffered hierarchical rings

# Router Area and Critical Path

- 16-node network with 8 bridge routers
- Verilog RTL design using 45nm Technology

- HiRD reduces NoC area by 50.3% compared to a buffered hierarchical ring design
- HiRD reduces local router critical path by 29.9% compared to a buffered hierarchical ring design

# Additional Results

- Detailed power breakdown
- Synthetic evaluations
- Energy efficiency results
- Worst case analysis
- Techical Report:
  - Multithreaded evaluation
  - Average, 90$^{th}$ percentile and max latency
  - Comparison against other topologies
  - Sensitivity analysis on different link bandwidths and number of buffers

# Outline

- Background and Motivation
- Key Idea: Deflection Routing
- End-to-end Delivery Guarantees
- Our Solution: HiRD
- Results
- Conclusion

# Conclusion

- Rings do not scale well as core count increases
- Traditional hierarchical ring designs are complex and energy inefficient
  - Complicated buffering and flow control
- **Solution:** Hierarchical Rings with Deflection (HiRD)
  - Guarantees livelock freedom and delivery
  - Eliminates all buffers at local routers and most buffers at bridge routers
- HiRD provides higher performance and energy-efficiency than hierarchical rings
- HiRD is simpler than hierarchical rings

# Design and Evaluation of Hierarchical Rings with Deflection Routing

**Rachata Ausavarungnirun**, Chris Fallin, Xiangyao Yu,
Kevin Chang, Greg Nazario, Reetuparna Das,
Gabriel H. Loh, Onur Mutlu

Carnegie Mellon

SAFARI

AMD

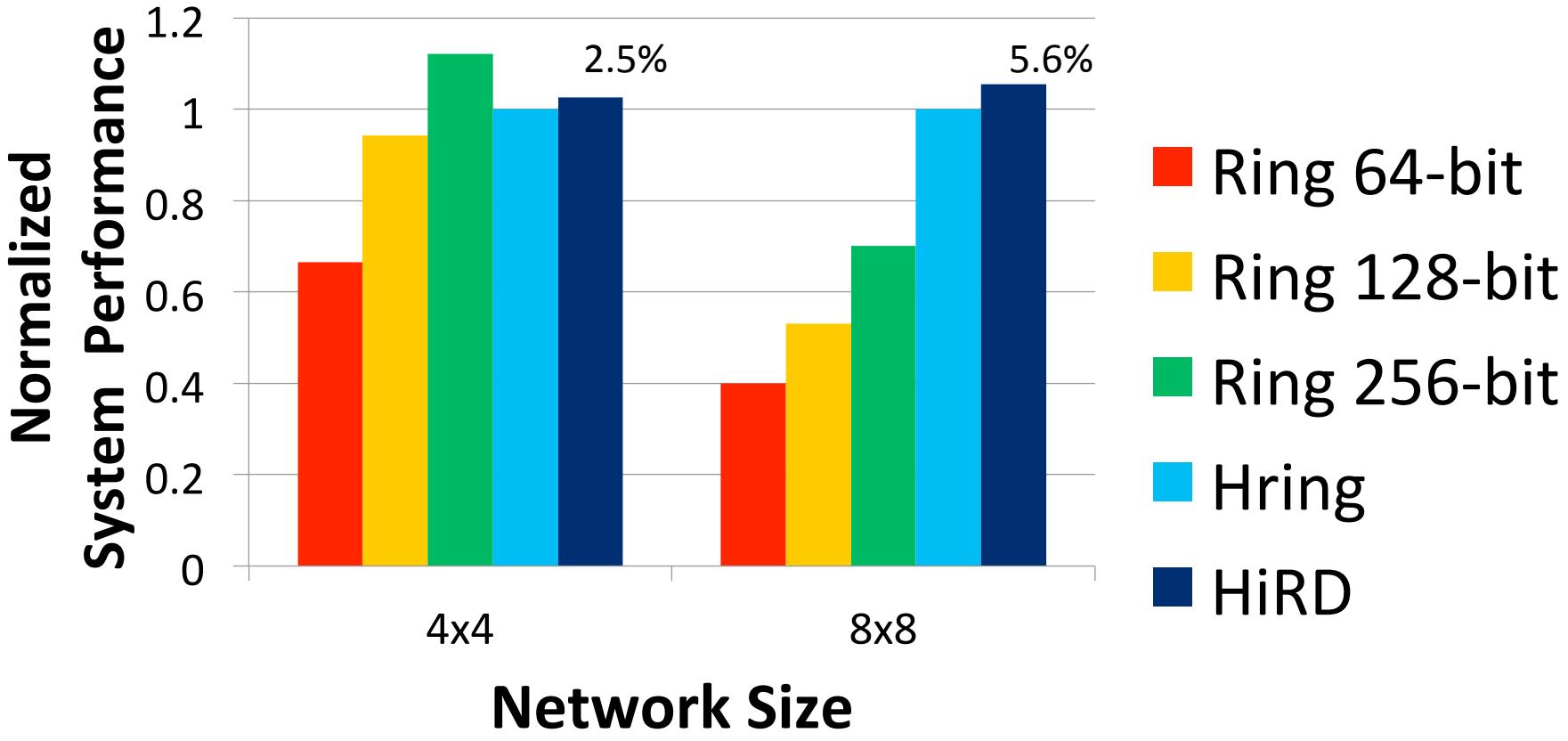
MICHIGAN

# Backup Slides

# Network Intensive Workloads

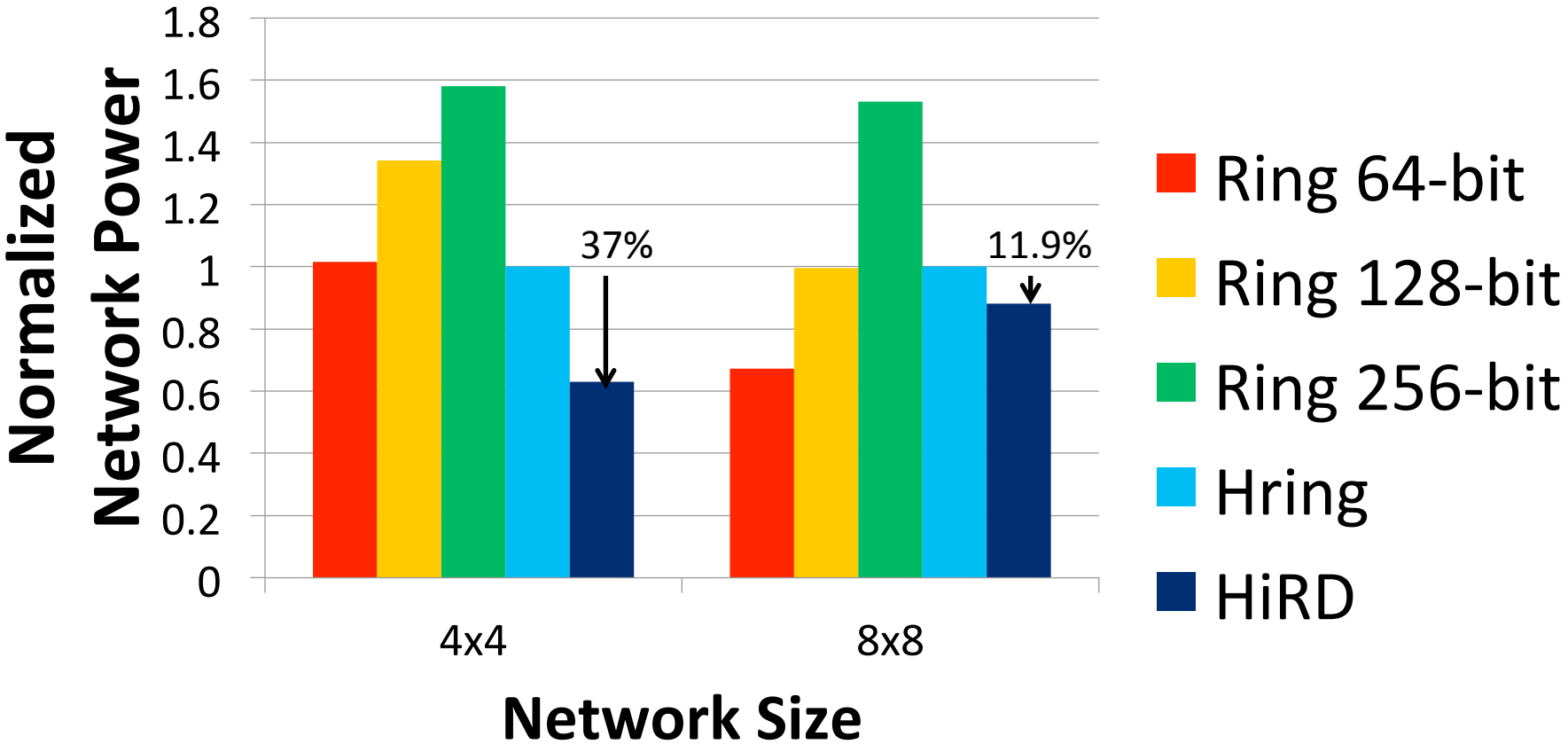
- 15 network intensive workloads

# System Performance



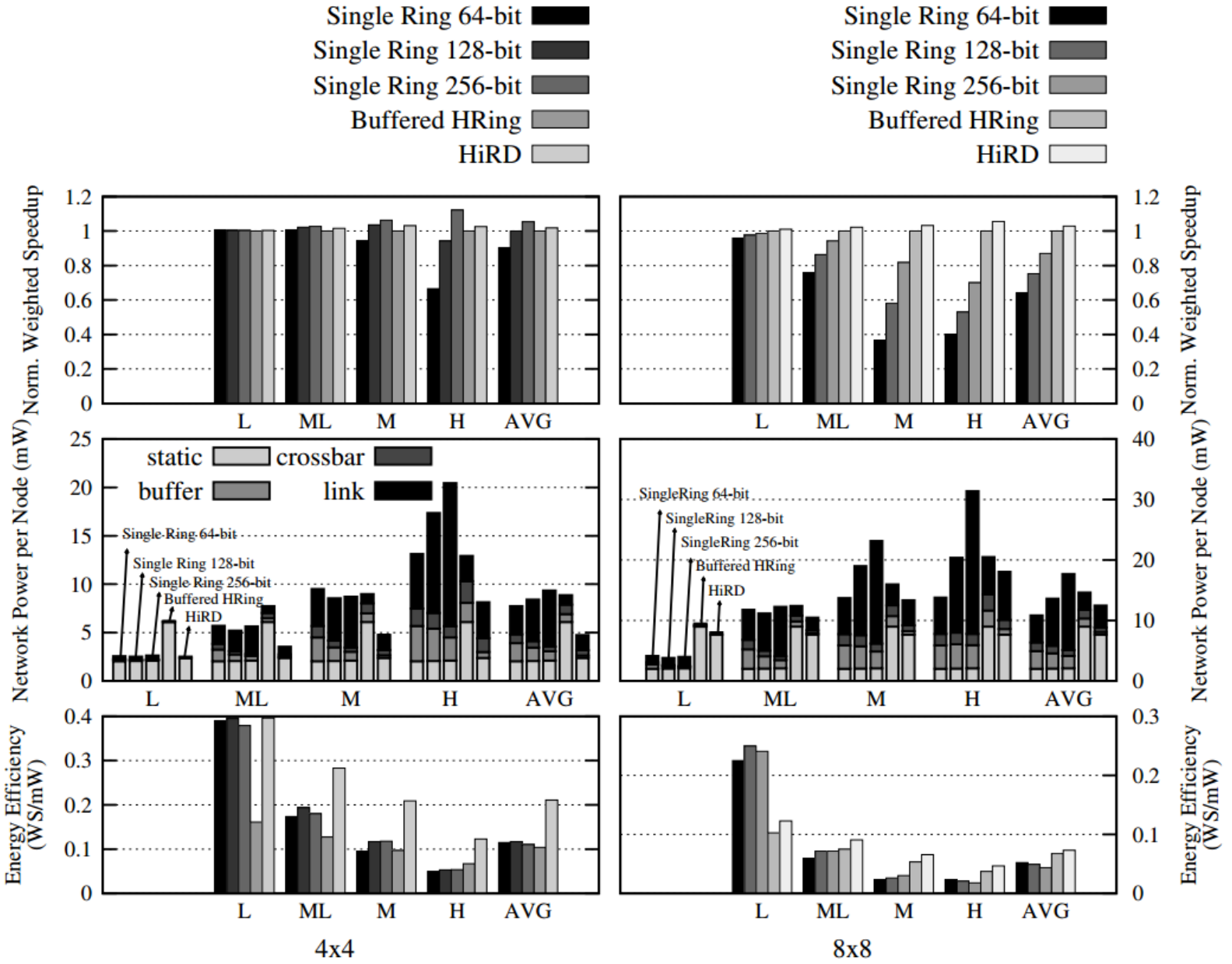

Deflections balance out the network load
Thorttling reduces congestion

# Network Power

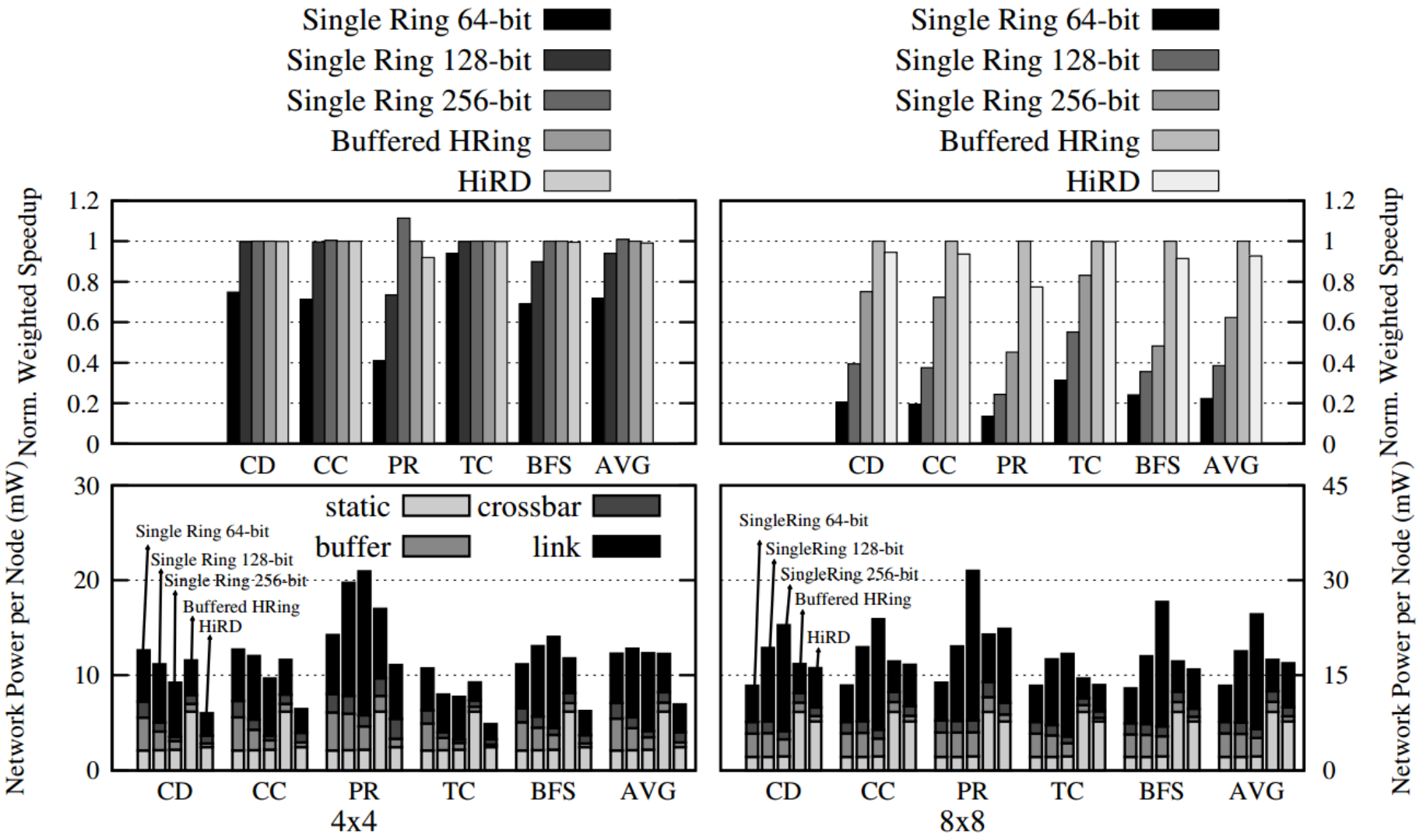



More deflections happen when the network is congested

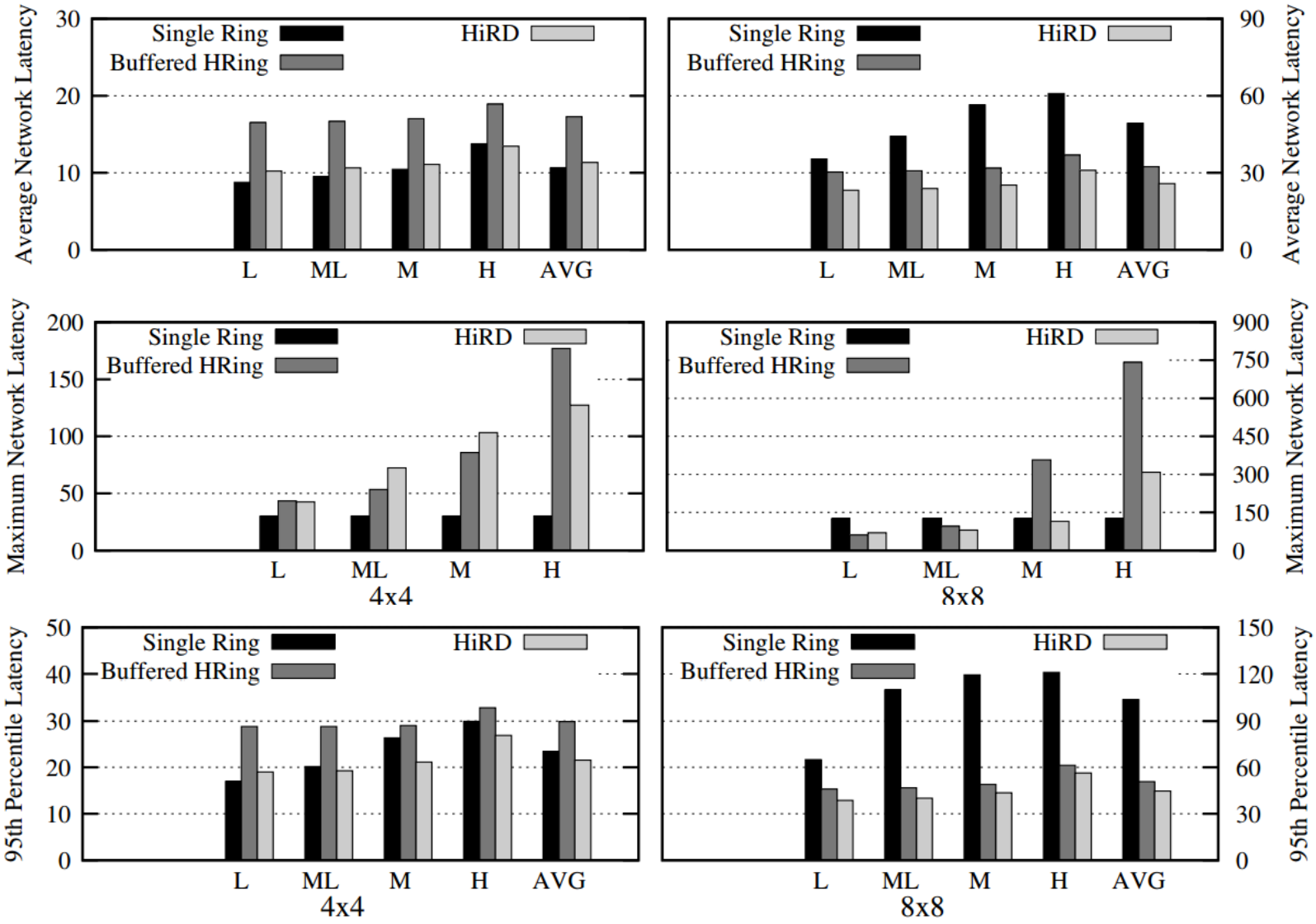
# Detailed Results



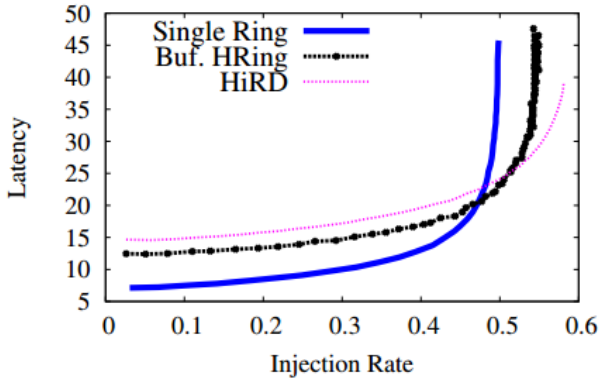
# Multithreaded Applications



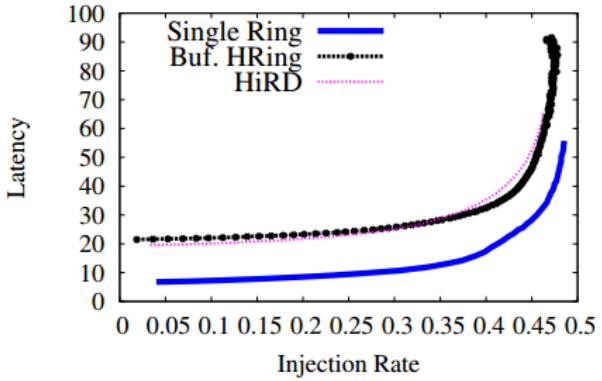
# Network Latency



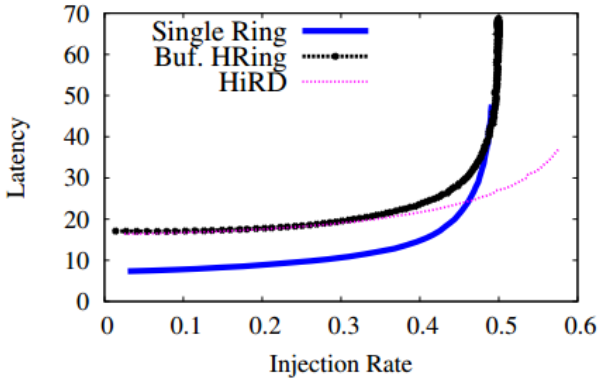
# Synthetic Traffic Evaluations



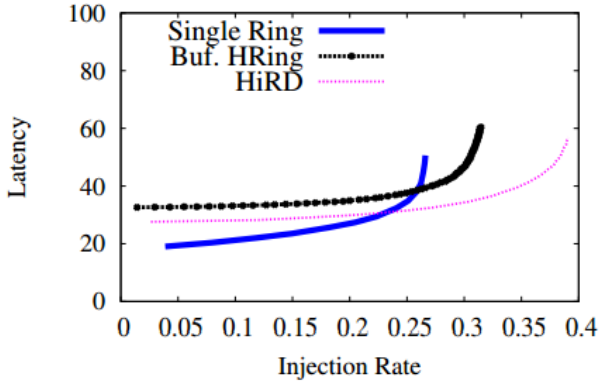

(a) Uniform Random, 4x4



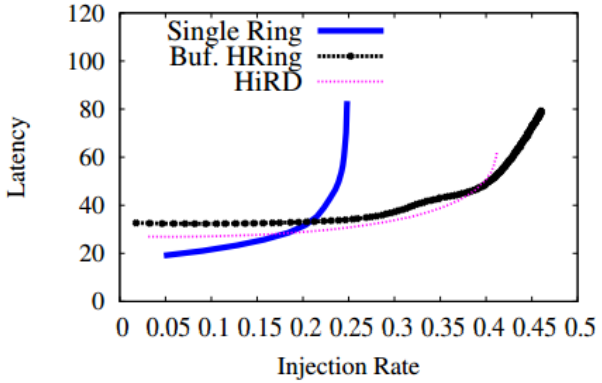

(b) Bit Complement, 4x4



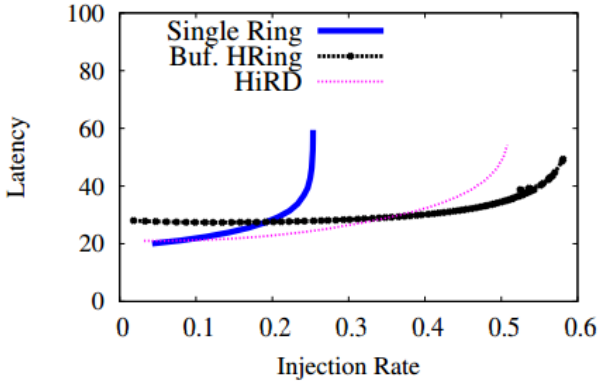

(c) Transpose, 4x4




(d) Uniform Random, 8x8




(e) Bit Complement, 8x8

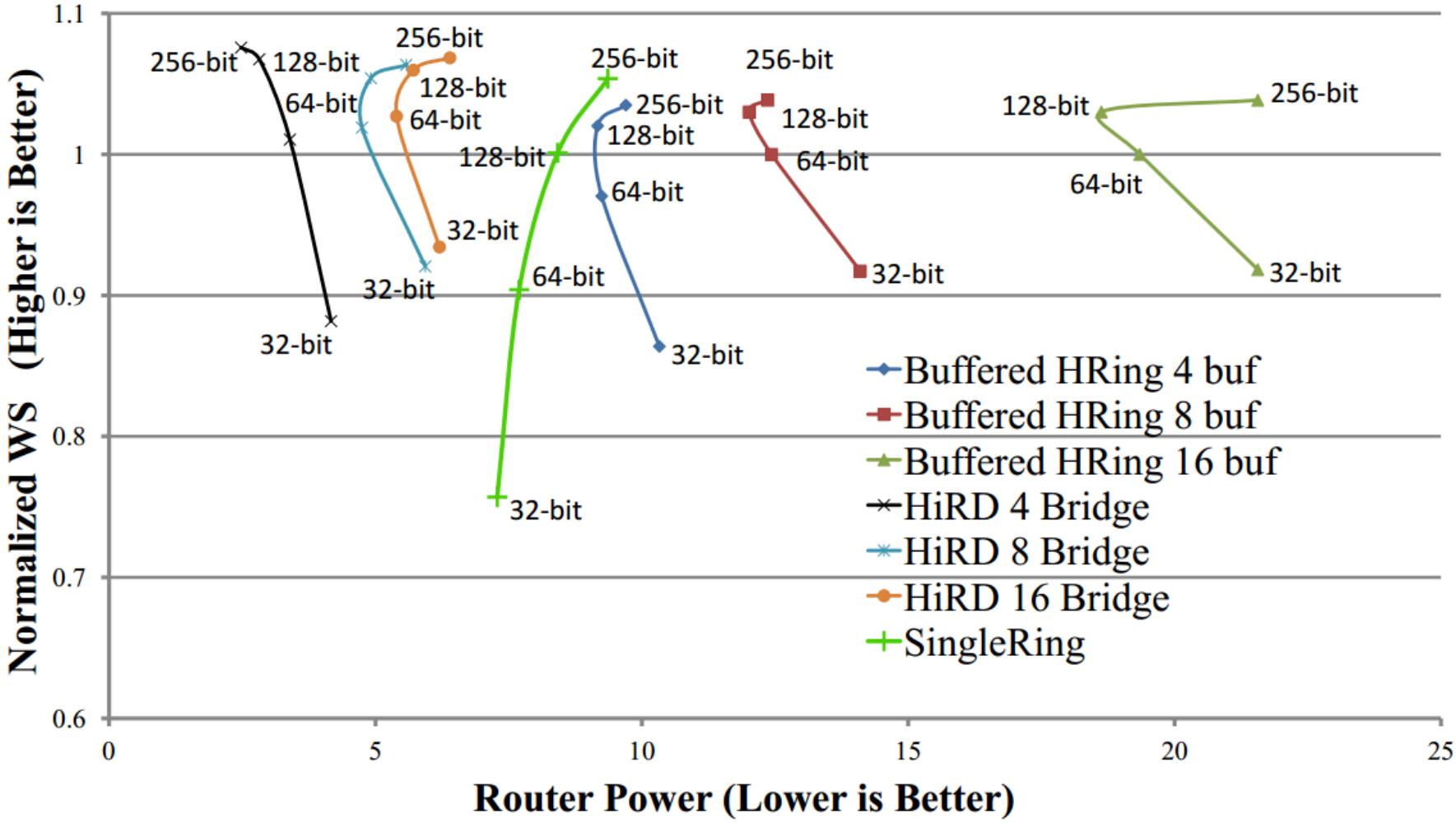



(f) Transpose, 8x8

# Topology Comparison

| Topologies | 4x4 | | 8x8 | |
|---|---|---|---|---|
| | Norm. WS | Power (mWatts) | Norm. WS | Power (mWatts) |
| Single Ring | 0.904 | 7.696 | 0.782 | 13.603 |
| Buffered HRing | 1 | 12.433 | 1 | 16.188 |
| Buffered Mesh | 1.025 | 11.947 | 1.091 | 13.454 |
| CHIPPER | 0.986 | 4.631 | 1.013 | 7.275 |
| Flattened Butterfly | 1.037 | 10.760 | 1.211 | 30.434 |
| HiRD | 1.020 | 4.746 | 1.066 | 12.480 |

# Sweep over Different Bandwidth



# Packet Reassembly

- Borrowed from CHIPPER [Fallin et al. HPCA'10]
  - Retransmit-Once → Destination node reserves a buffer slot for a dropped packet
  - Provides ejection guarantee

# Other Optimizations

- Map cores that communicate with each other a lot on the same local ring
  - Takes advantage of the faster local ring routers

# Related Concurrent Works

- Clumsy Flow Control
  [Kim et al., IEEE CAL'13]
  - Requires coordination between cores and memory controllers

- Transportation inspired NoCs
  [Kim et al., HPCA'14]
  - tNoCs require an additional credit network
  - tNoCs have more complex flow control
  - HiRD is more lightweight

# Some Related Previous Works

- Hierarchical Bus [Udipi et al., HPCA'10]
  - HiRD provides more scalability
- Concentrated Meshe [Das et al., HPCA'09]
  - Several nodes share one router
  - Used on meshed network
  - Less power efficient than HiRD
- Low-cost Mesh Router [J. Kim, MICRO'09]
  - Specifically designed for meshes
  - Does not solve issues in deflection-based flow control (HiRD does)