Lightspeed Datacenter Network

RAM Cloud Meeting 03/16/12

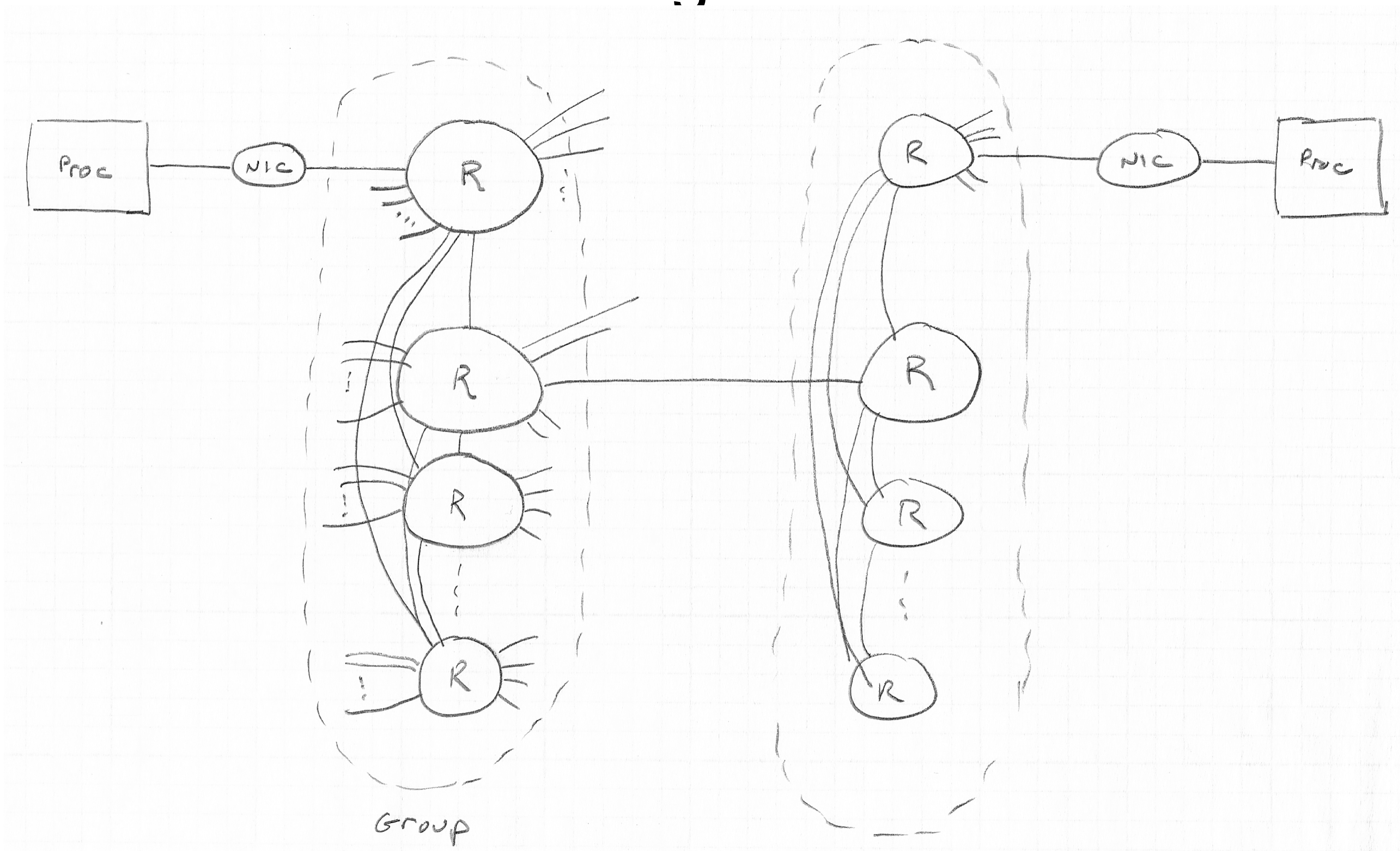
Prof. William Dally
Stanford University

STANFORD ENGINEERING

Assumptions/Requirements

- Need to connect up to 100,000 endpoints
- Each endpoint requires 10Gb/s bandwidth
  - Can be scaled up as needed (25Gb/s in near future)
  - Can gang – 40Gb/s (100Gb/s) if needed
- Flat bisection bandwidth
- Mostly benign (load balanced) traffic patterns
  - But must handle adversarial traffic, unbalanced, hot spots
- Must support both short packets (~ 64B) and long flows (64MB)
- Must provide congestion control for in-cast traffic
- Latency as close to time-of-flight as possible
- Standard QSFP active optical cables
- PCIe (today) interface to NICs
  - Integrated in the CPU in the future

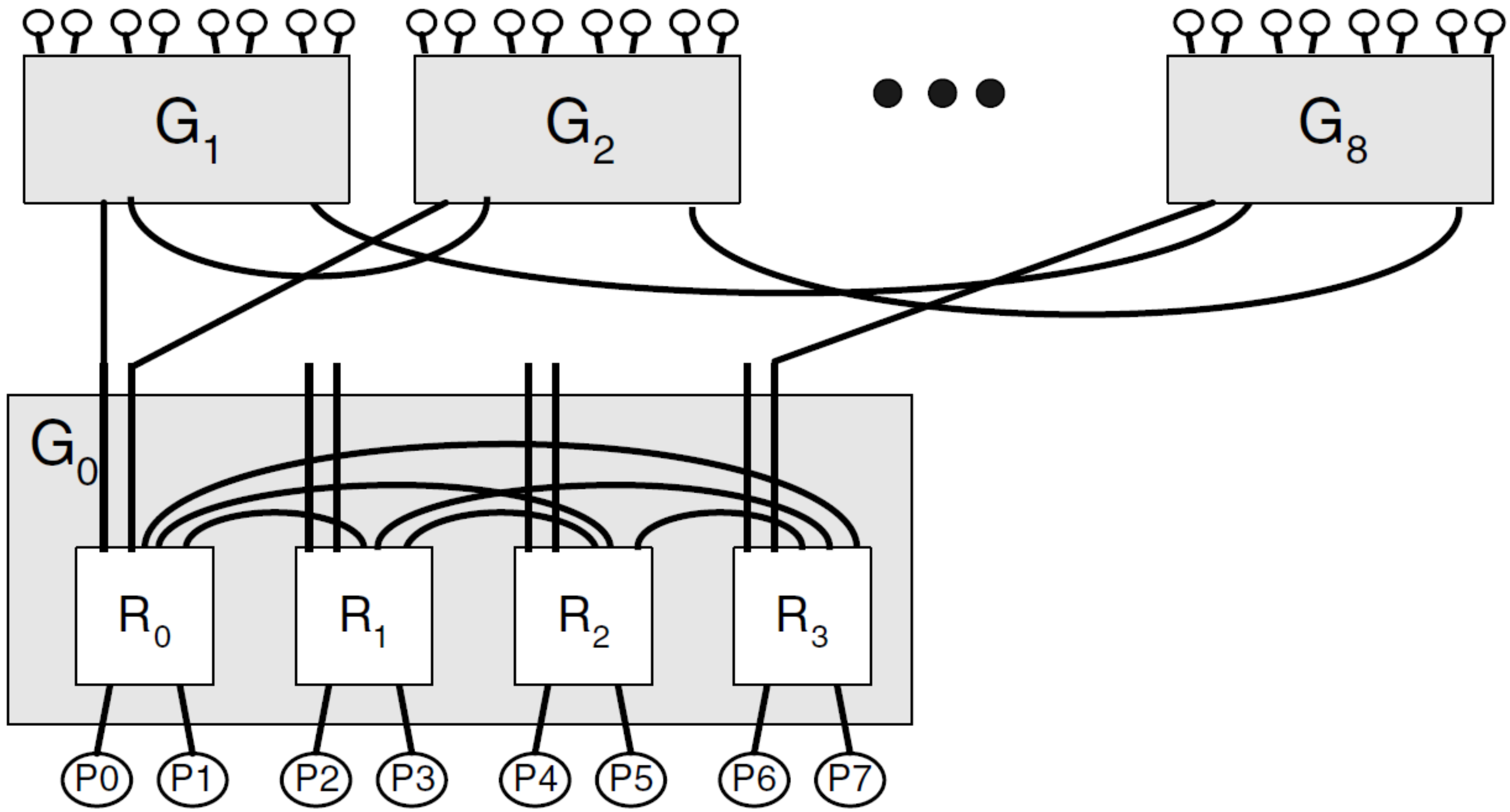
The Big Picture



Issues/Decisions

- Topology – Dragonfly (ISCA 2008, IEEE Micro 2009)
  - Only one long (expensive) hop per load-balanced route
    - Two for adversarial traffic
  - Four routers along a typical route
- Routing – Indirect Adaptive Routing (ISCA 2009)
  - Take minimal route if unloaded
  - Take non-minimal route (Valiant) if loaded
- Flow-Control – Virtual Channel with Speculative Reservation (HPCA 2012)
  - Lossless, low latency (minimal queueing), prevents tree saturation
- Routers – High-Radix (ISCA 2005, ISCA 2006)
  - Up to 256 10Gb/s (25Gb/s) ports per router
  - Monitoring/system management
- NICs
  - Support SRFC, per-job protection, monitoring
  - Fast launch of short messages
  - PCIe today, integrated tomorrow

Dragonfly Topology



Dragonfly by the Numbers

Endpoints per router	P	42
Side-links per router	S=3P	126
Global links per router	G=2P	84
Total router links	6P	252
Maximum routers/group	S+1	127
Endpoints per max group	P(S+1)	5,334
Global links per max group	G(S+1)	10,668
Max number of groups	G(S+1)+1	10,669
Max endpoints	P(S+1)(G(S+1)+1)	56,908,446

P:S:G = 1:3:2 provides 100% throughput on adversarial traffic vs (1:2:1).
More endpoints than needed – use redundant paths or lower radix routers.
Radix 64 routers can reach 100,000 endpoints.

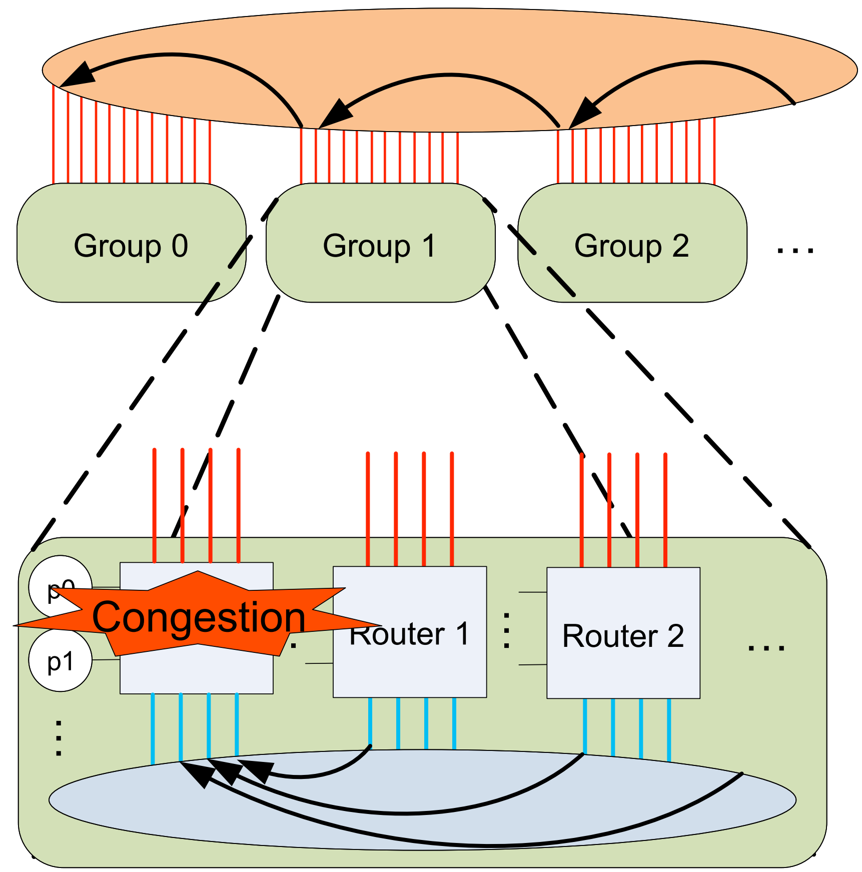
Dragonfly Packaged

- 84 endpoints + 2 routers per cabinet
  - Electrical connection from endpoint to router
- 32 cabinets in a group
  - 2,688 endpoints
  - 5,376 global channels
- 64 groups in a data center
  - 2,048 cabinets
  - 172,032 endpoints
  - 84 channels between every pair of groups

- 3-level Clos (fat tree) achieves similar numbers with same radix router

Routing on the Dragonfly

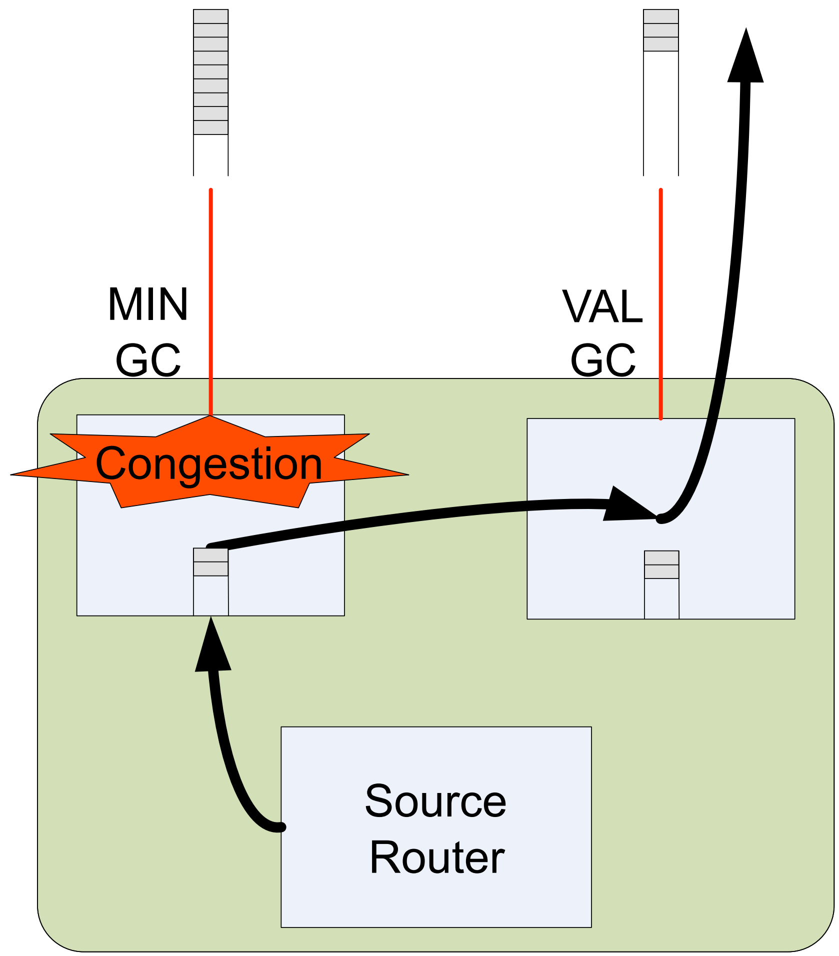
- Minimal Routing (MIN)
  1. Source local network
  2. Global network
  3. Destination local network
- Some Adversarial traffic congests the global channels
  - Each group *i* sends all packets to group *i+1*
- Oblivious solution: Valiant's Algorithm (VAL)
  - Poor performance on benign traffic



Progressive Adaptive Routing

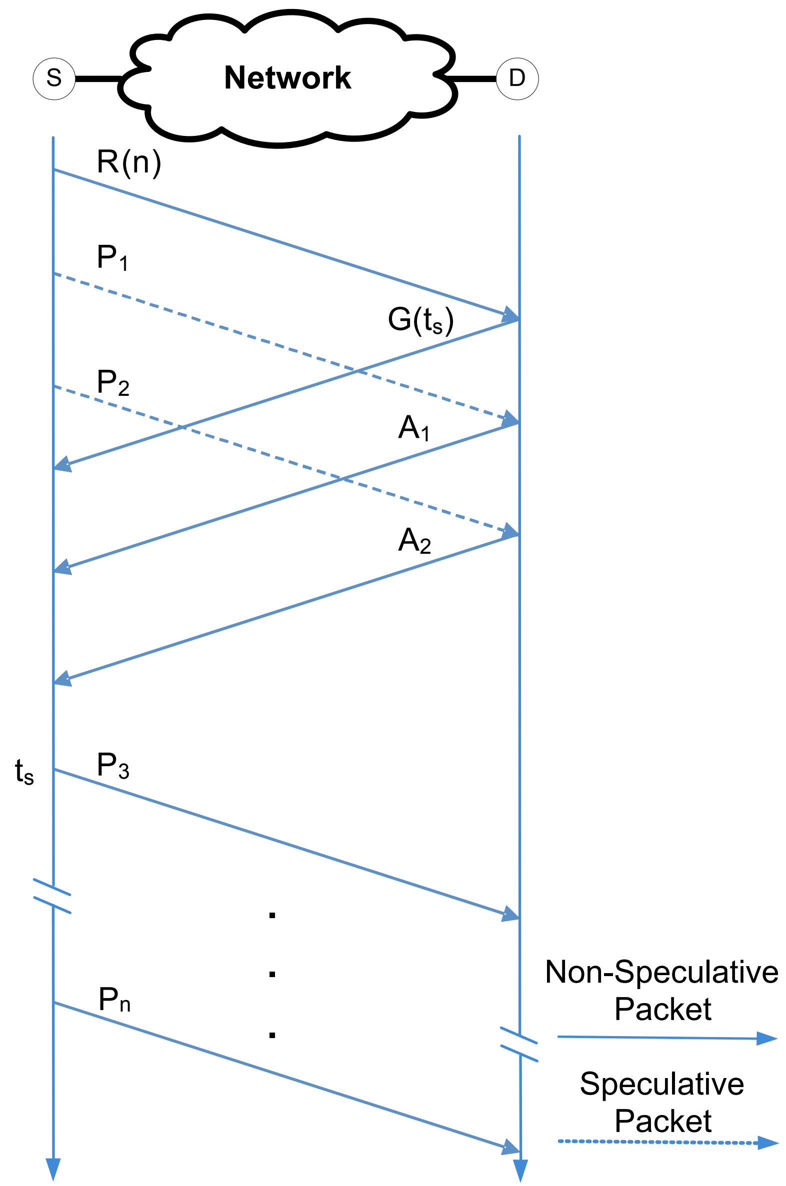
- MIN routing decisions at the source are not final
- VAL decisions are final
- Switch to VAL when encountering congestion

- Uses an additional virtual channel to avoid deadlock



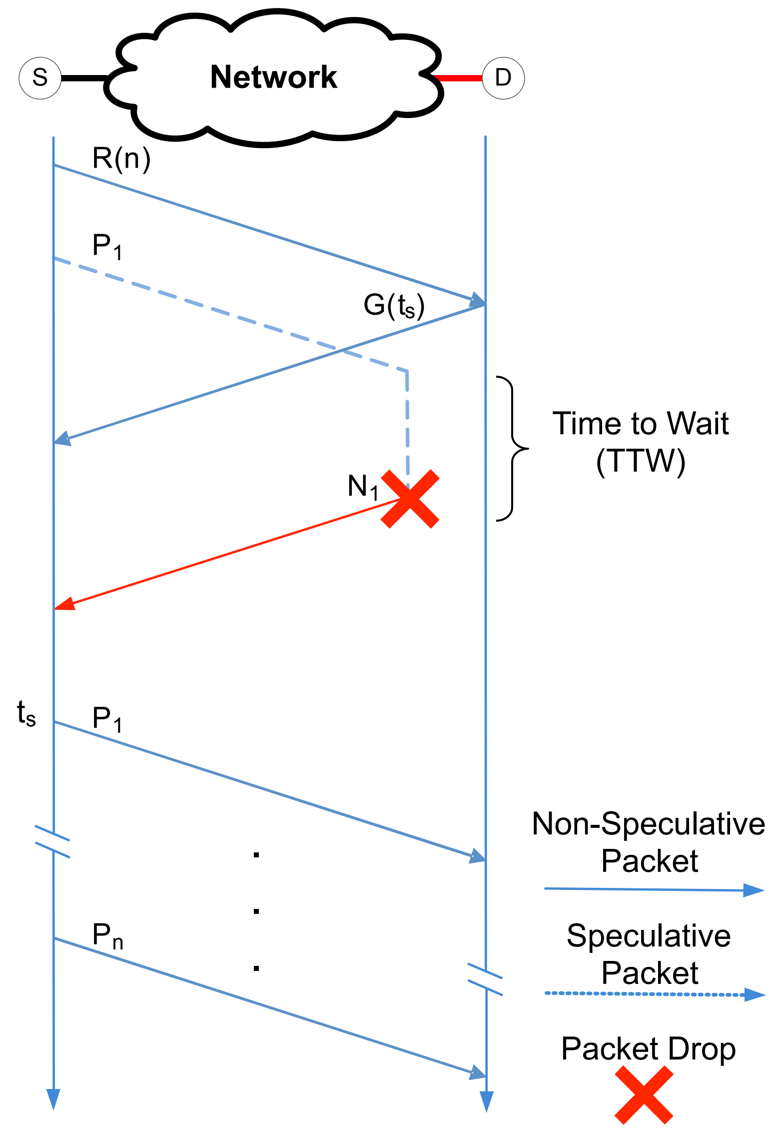
Speculative Reservation Protocol

- Network source issues a reservation indicating transmission size, R(n)
- Network destination replies with a grant indicting when the source can transmit, G (t_s)
- Waiting for reply, source can transmit packets speculatively, P1 and P2
- Speculative packets requires acknowledgements, A1 and A2
- After the granted time, t_s, the source can transmit non-speculatively, P3 to Pn

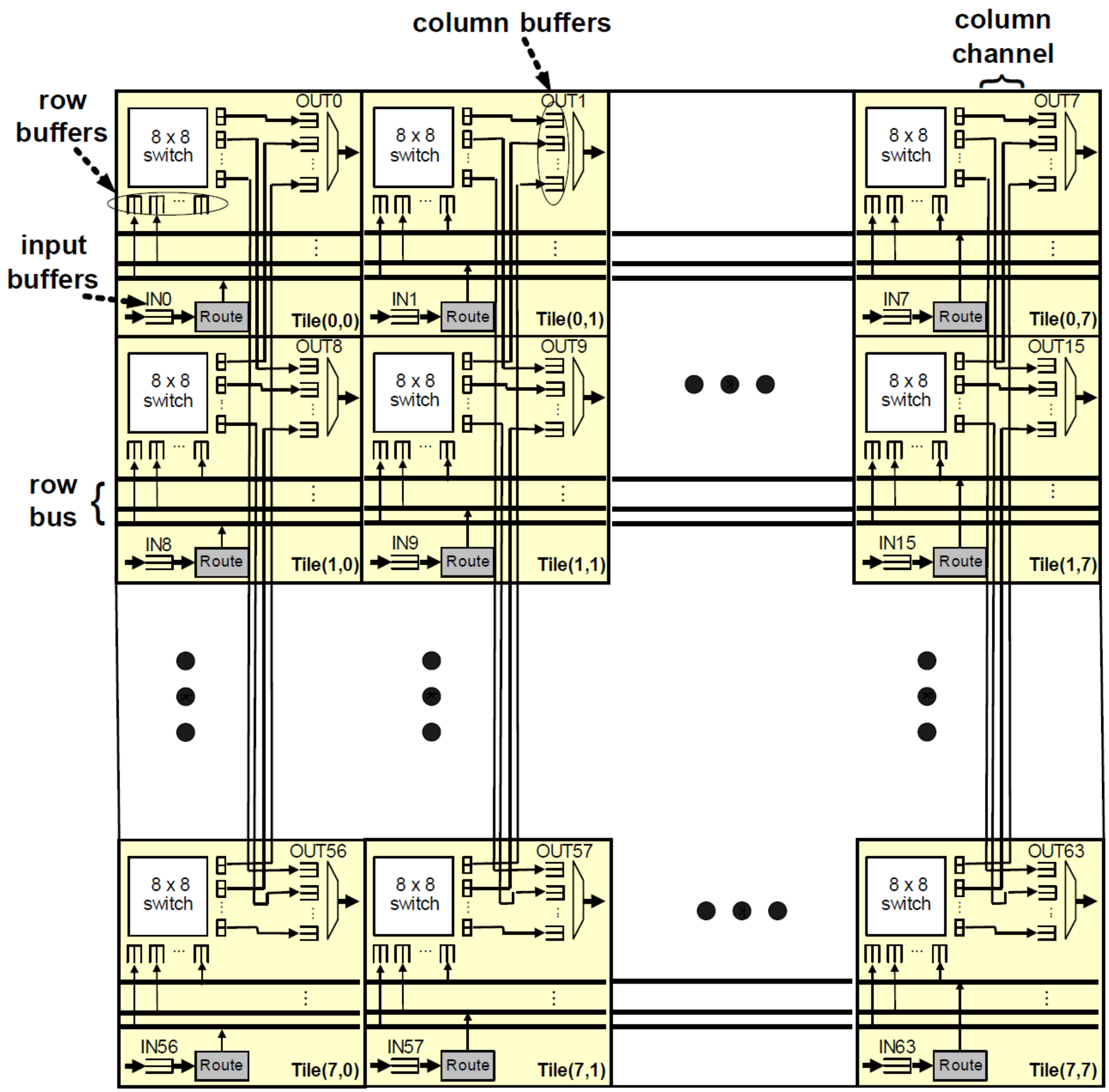


Network Congestion Example

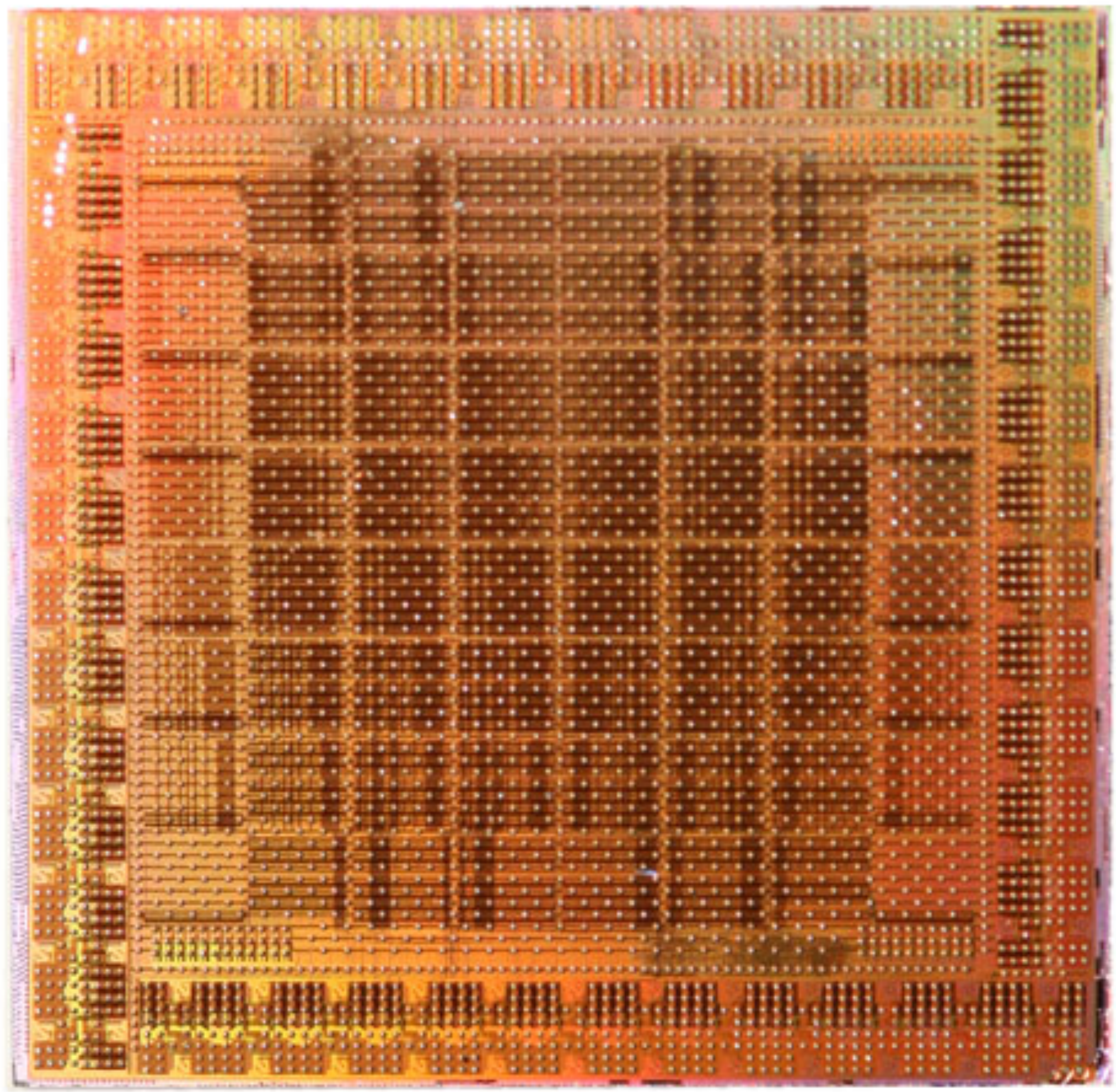
- Source starts normally by sending reservation R(n) and speculative P1
- R(n) has high network priority and quickly reaches the destination
- P1 encounters congestion and is buffered in the network
- P1 is dropped after a period of time and a negative acknowledgement is returned, N1
- Dropped speculative packets are retransmitted after the granted time t_s



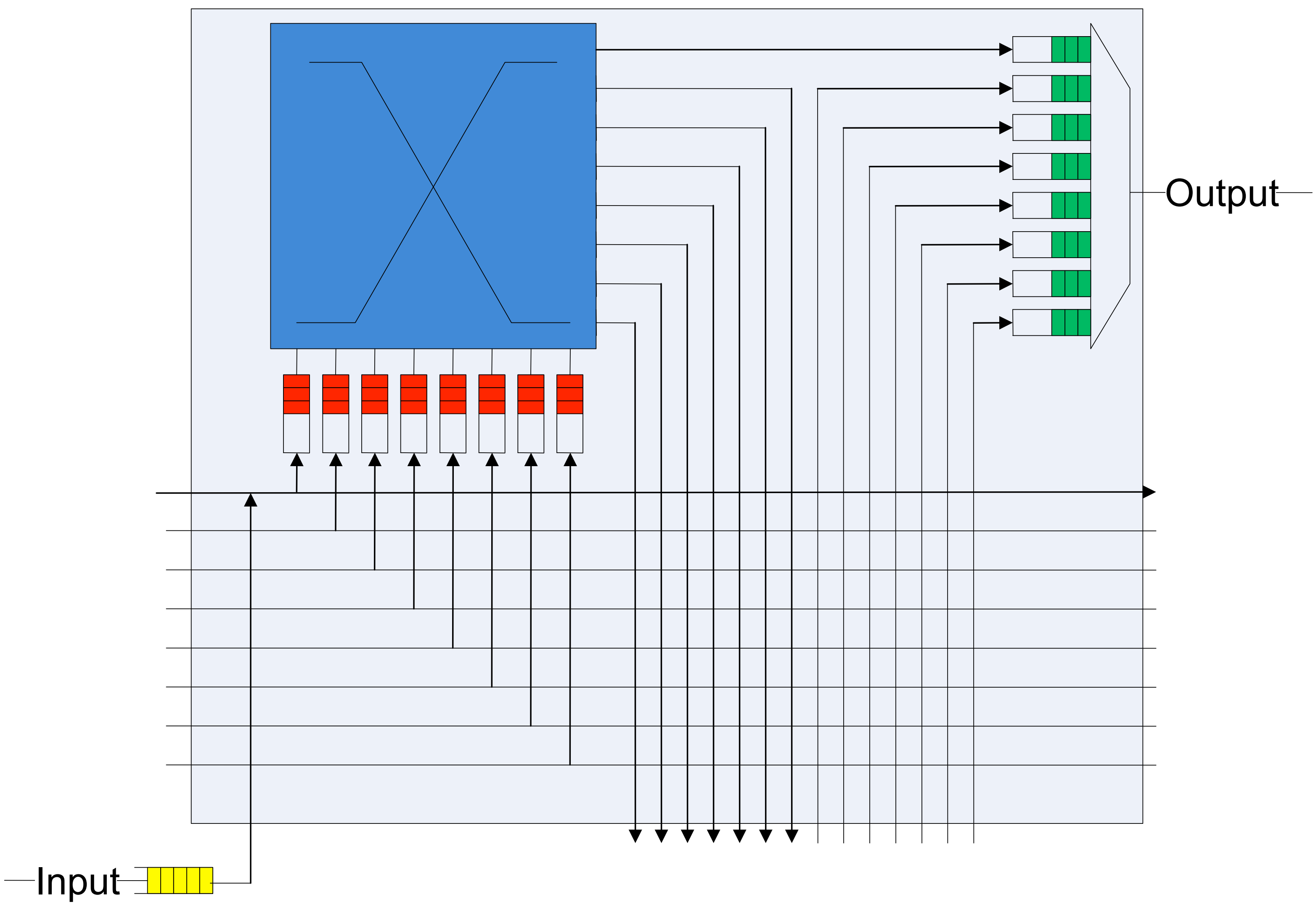
High-Radix Router Architecture



YARC Die Photo

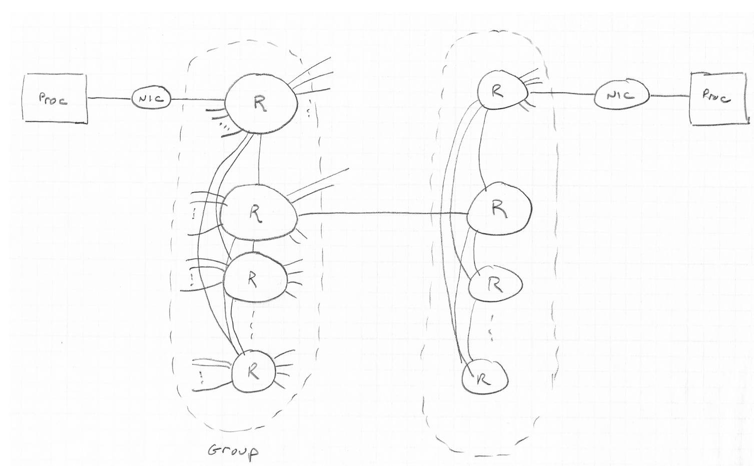


One Tile



End-To-End Latency (814ns, mostly wire)

TNIC	13 ns	NIC Send delay
C(N-R)	5 ns	Channel NIC to Router
Router	31 ns	Router delay
C(L)	50 ns	Local intra-group channel
Router	31 ns	Router delay
C(G)	500 ns	Global channel
Router	31 ns	Router delay
C(L)	50 ns	Local intra-group channel
Router	31 ns	Router delay
C(R-N)	5 ns	Channel Router to NIC
TNICR	65 ns	Receive NIC delay
Ser	51 ns	Serialization Latency
TOTAL	814	
TWIRE	610 ns	
TOVH	204 ns	
OVH	25.02%	



Router Traversal

ROUTER		
Link	1	
Sync	5	
In	2	
Horiz	3	
Xbar	3	
Vert	3	
Mux	2	
Sync	5	
Link	1	
Subtotal	25	cycles
Subtotal	25	ns
Ser	6.4	ns
TROUTER	31.4	ns

NIC Packet Launch

NIC		
Send	1	clock
Sync	5	clocks
Link	1	clock
Subtotal	7	clocks
Subtotal	7	ns
Ser	6.4	ns
TNIC	13	ns To first word transmitted

PCIe Overhead

- About 400ns to write an 8-word burst into the NIC registers.
- Another 400ns to read the NIC on the far end – processor can be waiting on read.
- 800ns total – comparable to the rest of the end-to-end latency
- Eliminate this by integrating the NIC

Cost Estimate ($344 per endpoint) $310 channel, $34.29 router and NIC

NIC Chip	10	
NIC PCB	10	
Total NIC	20	
Router Chip	100	
Router PCB	30	
Router Box	20	
Total Router	150	
Routers	0.10	per endpoint
Router Cost	14.29	per endpoint
Electrical Channel	10	
AOC	100	
Channel Cost	310	
Total Cost	344.29	per endpoint

This is cost, not price

Conclusion
Lightspeed Networking

- Supercomputing network technology
  - Dragonfly topology, indirect adaptive routing, speculative reservation flow control, high-radix routers

- Gives packet delivery close to time-of flight over the wire (814ns, 610ns wire, 204ns NIC+Router)

- Cost of ~$350 per endpoint – dominated by channels ($34.29 for routers + NIC per endpoint)