On Growth of Parallelism within Routers and Its Impact on Packet Reordering

A. A. Bare[1], A. P. Jayasumana and N. M. Piratla[2]

Electrical & Computer Engineering, Colorado State University

Ft. Collins, CO

[1]Agilent Technologies, Inc., Loveland, CO

[2]Deutsche Telekom Laboratories, Berlin, Germany

Outline

- Growth trends
- Impact on router architectures
- Metrics for reordering
- Simulation results
- Conclusions

Growth Trends

- "Internet traffic continues to grow vigorously, approximately doubling each year, as it has done every year since1997." (annual growth between 70 and 150%.)

[**Internet traffic growth: Sources and implications**, A. M. Odlyzko. *Optical Transmission Systems and Equipment for WDM Networking II*, B. B. Dingel, W. Weiershausen, A. K. Dutta, and K.-I. Sato, eds., Proc. SPIE, vol. 5247, 2003, pp. 1-15]

Growth Trends

- Progress in transmission technology appears sufficient to double network capacity each year for about the next decade.

[**Internet growth: Is there a "Moore's Law" for data traffic?**, K. G. Coffman and A. M. Odlyzko. *Handbook of Massive Data Sets*, J. Abello, P. M. Pardalos, and M. G. C. Resende, eds., Kluwer, 2002, pp. 47-93]

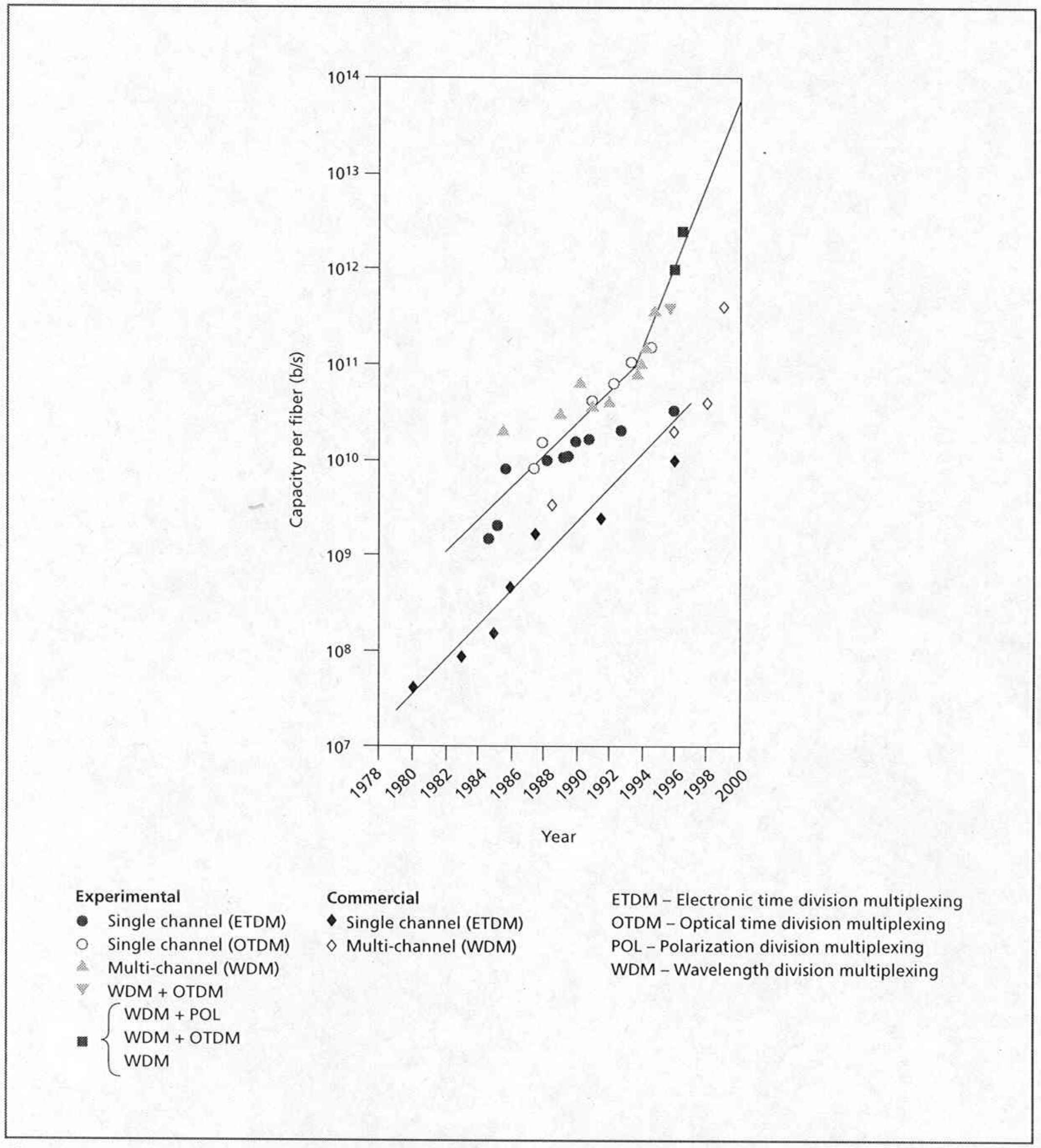


Figure 1.
Progress in lightwave transmission capacity.

R. C. Alferness, H. Kogelnik, T.H. Wood, The Evolution of Optical Systems: Optics Everywhere – Bell Labs Tech Journal, Jan-March 2000

Growth Trends

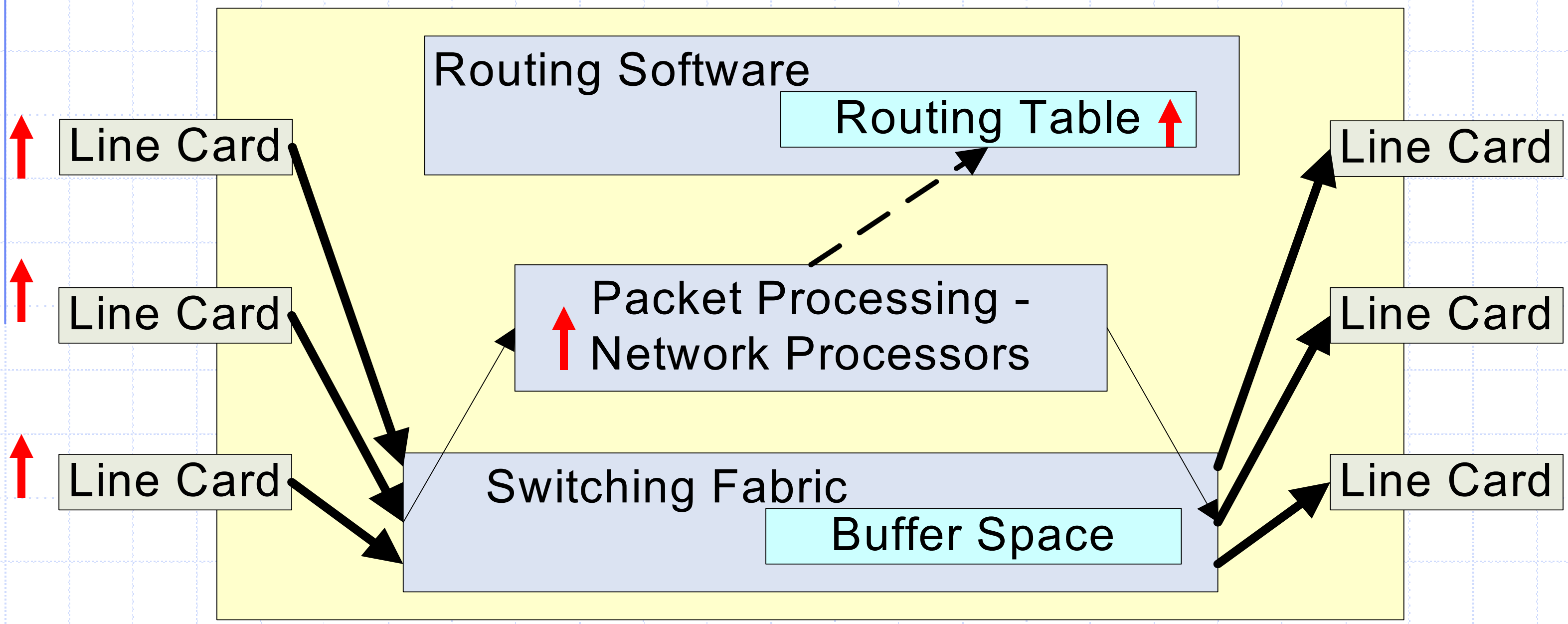
- “For the first time in history, performance improvements are required at a rate faster than 18-month doubling of semiconductor performance that Moore’s Law predicted in 1975.”

[**Beyond Moore’s Law: Internet Growth Trends**, Lawrence G. Roberts, IEEE Computer, January 2000.]

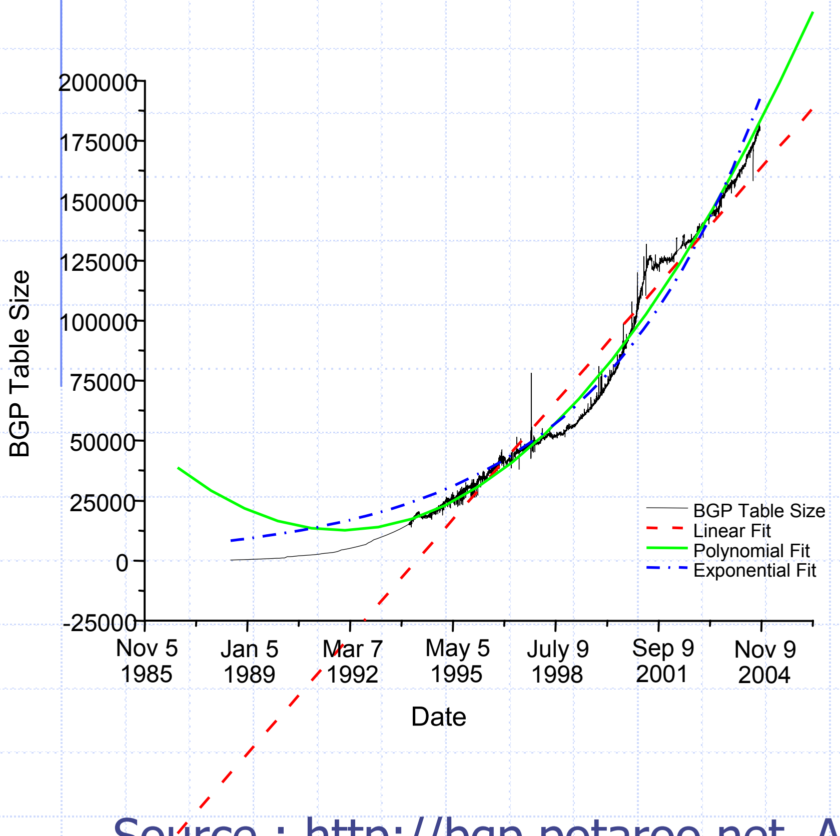
System	Fiber Capacity	Wide Deployment
8x2.5 Gbps	20Gbps	1996
16x2.5 Gbps	40Gbps	1996
32x2.5 Gbps	80Gbps	1996
80x2.5 Gbps	200 Gbps	2000
40x10 Gbps	400 Gbps	2000
160x2.5 Gbps	400 Gbps	Late 2000
80x10 Gbps	800 Gbps	2002
160x10 Gbps	1.6 Tbps	2003
40x40 Gbps	1.6 Tbps	2003
80x40 Gbps	3.2 Tbps	2003/4
100x40 Gbps	4 Tbps	2005
160x40 Gbps	6.4 Tbps	2007

Source: K. G. Coffman and A. M. Odlyzko, “Internet Growth: Is there a “Moore’s Law” for Data Traffic?” July 11, 2000

A Generic Router



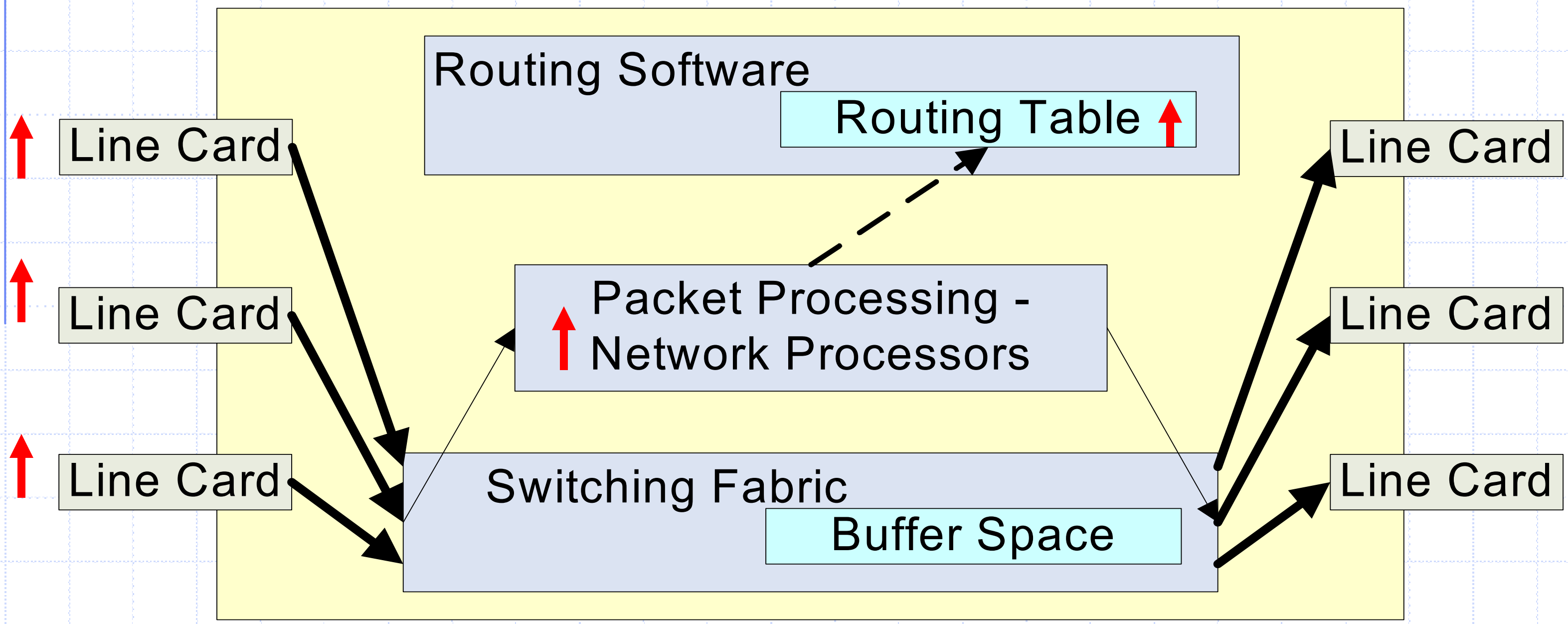
BGP Table Size



$$S = 7.804e\text{-}013 * T_u^2 - 0.00076 * T_u + 9.603e{+}04$$

Source : http://bgp.potaroo.net, AS1221 (Telstra) router

A Generic Router



Growth Factors

- α - increase in network link speed
- β - increase in processing speed
- γ - increase in routing table size

- Computations for packet processing

$$\log_2(\gamma)$$

- Number of NPUs

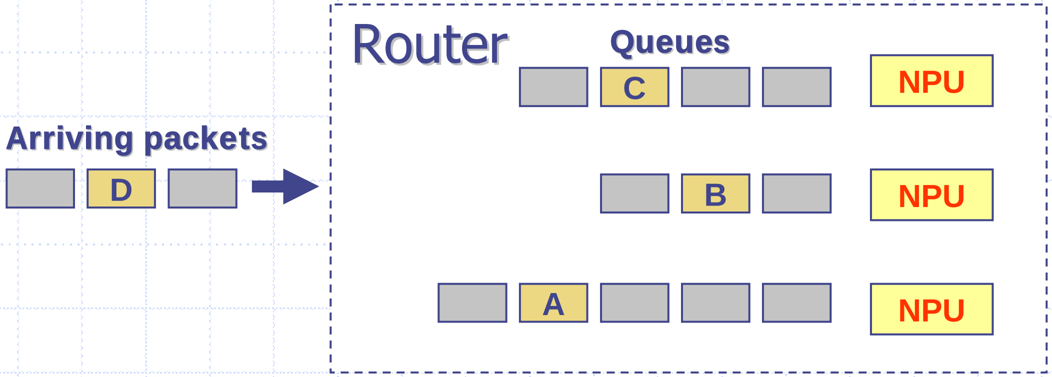
$$\alpha \log_2(\gamma)/\beta$$

Parallelism within routers

Link change OC-3 to	α	β	γ	ω	*n*	*Mean Packet Processing time*
OC-12	4	2.00	1.60	2.71	3	0.00722 * L + 8.476
OC-24	8	2.83	1.87	7.22	5	0.00680* L + 7.977
OC-48	16	4.00	2.16	17.77	9	0.00592 * L + 6.944
OC-96	32	5.66	2.46	41.56	15	0.00489 * L + 5.736
OC-192	64	8.00	2.77	94.07	24	0.00391 * L + 4.593
OC-384	128	11.3	3.10	208.9	37	0.00307 * L + 3.608
OC-768	256	16.0	3.44	456.3	57	0.00237 * L + 2.785

Table i. parameters used in simulations for different link speeds

Parallelism within Routers => Reordering ?



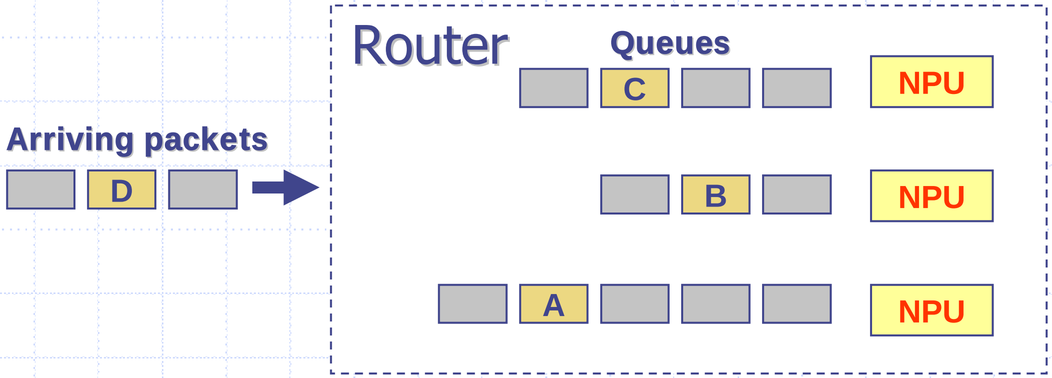
Ex: Juniper M160

Impact of Reordering

- Ex: Degradation of performance in TCP
  - No of unnecessary transmissions increase (drop in throughput)
  - Congestion window becomes small
  - RTT estimate degrades
  - Receiver performance degrades
  - Detection of lost packets delayed
  - Forward/reverse path causes loss of self-clocking

[Laor & Gendel, IEEE Network 2002]

Parallelism within Routers => Reordering ?



Ex: Juniper M160

Counter Measures : input flow tracking
output buffering

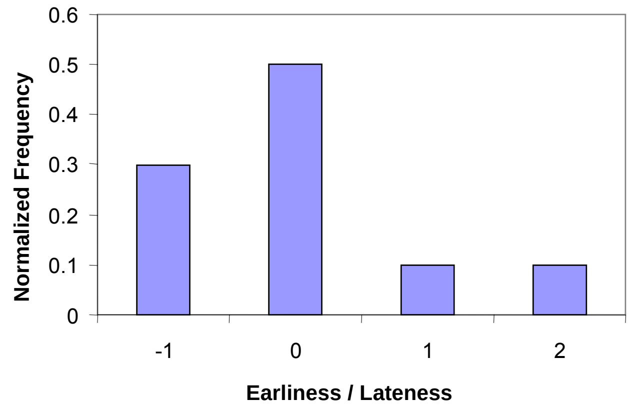
Reorder Density (RD)

- EX: Received sequence (1,3,4,2,5,6,8,7,9,10)

Sequence number	1	3	4	2	5	6	8	7	9	10
Receive index	1	2	3	4	5	6	7	8	9	10
Displacement	0	-1	-1	2	0	0	-1	1	0	0

  - Each arrival is assigned a receive index
  - Displacement of a packet = (receive index – sequence number)
- RD is computed as

Displacement	-1	0	1	2
Frequency	3	5	1	1
Normalized Frequency	0.3	0.5	0.1	0.1



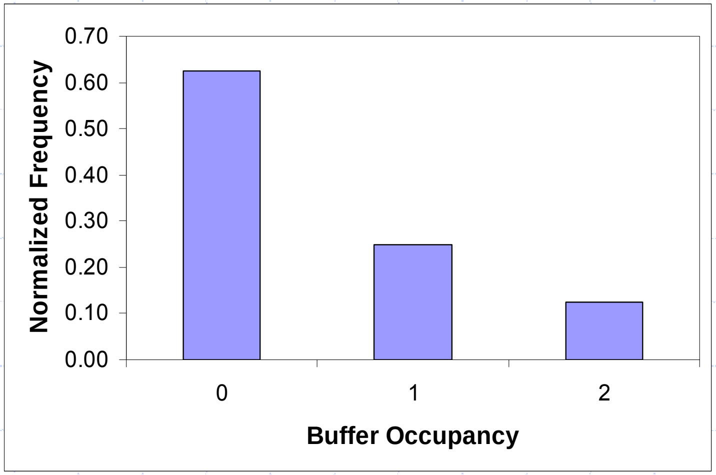
Reorder Buffer Density RBD

- Ex: Received sequence: 1,3,4,2,5,7,6,8
  - Compute the buffer occupancy after each arrival

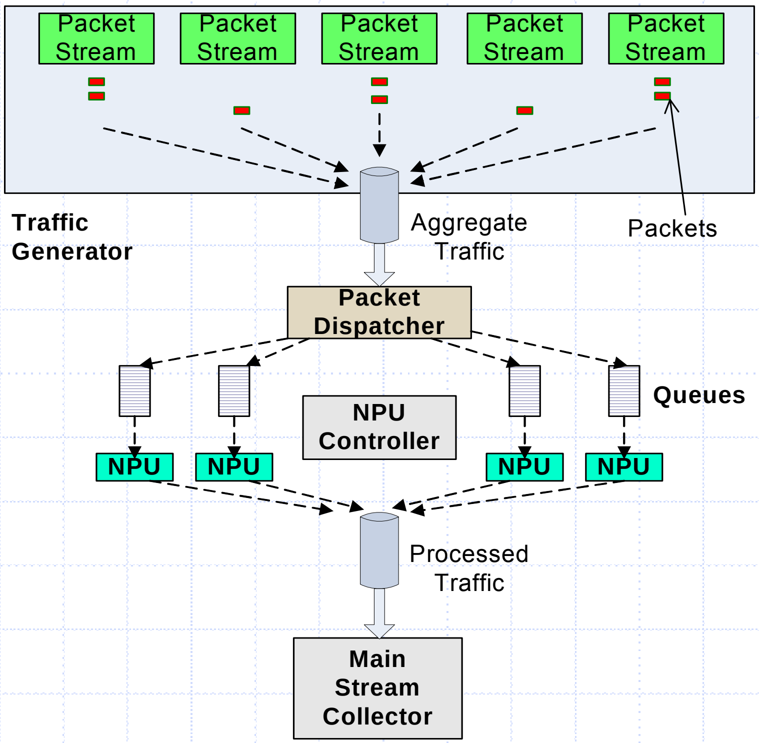
Arrived sequence number (S)	1	3	4	2	5	7	6	8
Sequence number expected (E)	1	2	2	2	5	6	7	8
Buffer contents after arrival	-	3	3,4	-	-	7	-	-
Buffer occupancy (B)	0	1	2	0	0	1	0	0

- RBD is the normalized frequency of each buffer occupancy

Buffer occupancy	Frequency	Normalized frequency
0	5	5 / 8
1	2	2 / 8
2	1	1 / 8



Router Simulation



Router Parameters

- Time taken for processing a packet arriving over OC-3 link*

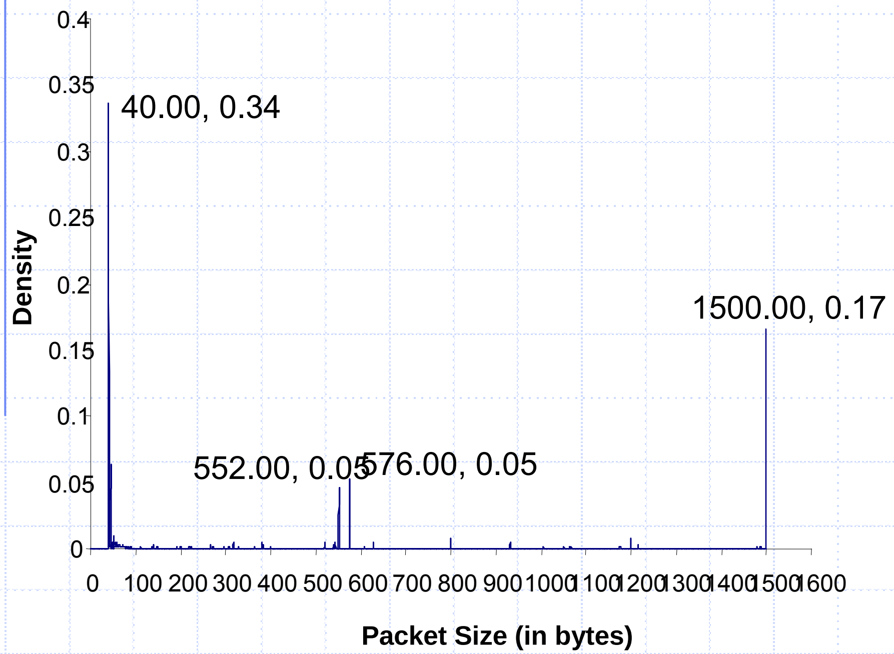
  $$d(L) = (0.0213 * L + 25) \text{ in } \mu s$$

  where L is length of a packet in bytes
- For simulation, packet processing time is

  $$P(L) = d(L) * \log_2(\gamma) / \beta$$
- Standard deviation in packet processing time: 5% used unless specified

* Source: K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, F. Tobagi, and C. Diot, "Analysis of measured single-hop delay from an operational backbone network," *Proc. IEEE INFOCOM*, June 2002

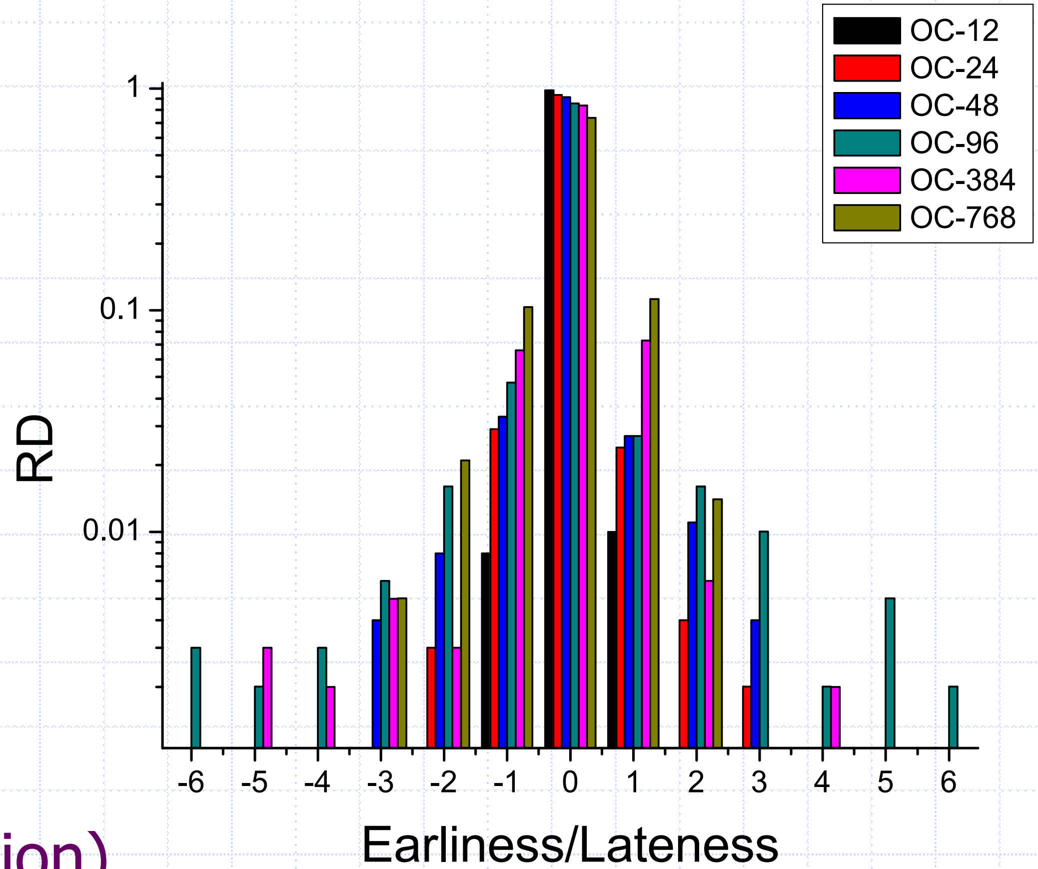
Simulation Parameters



Source: Packet Sizes and Sequencing, CAIDA, http://traffic.caida.org/TrafficAnalysis/Learn/Size/.

- Multiple streams combined per input to provide the required utilization
  Ex: 4000 streams
- Reordering measured on one of the streams
  Ex: 1000 packets
- Self-similar traffic arrivals in each stream
- Constant packet size in each stream, aggregate follows given distribution
- Round-robin scheduling (Conservative choice)
- No buffer overflow

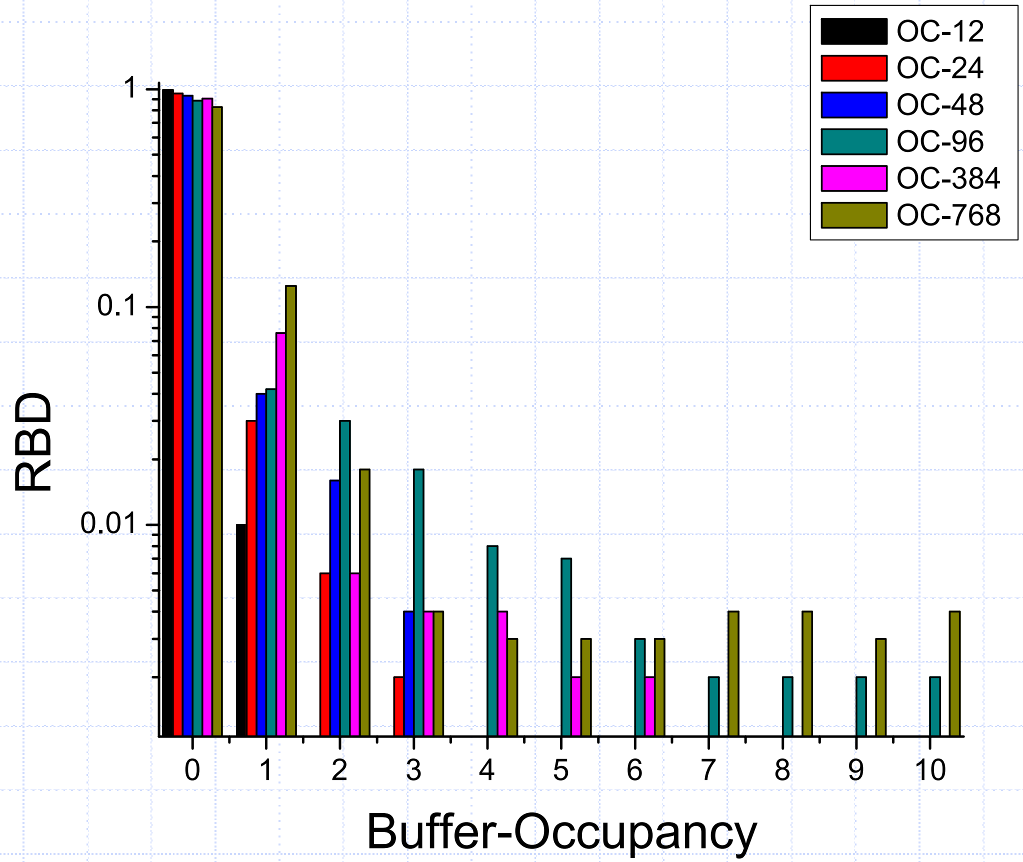
RD vs. Link Speed



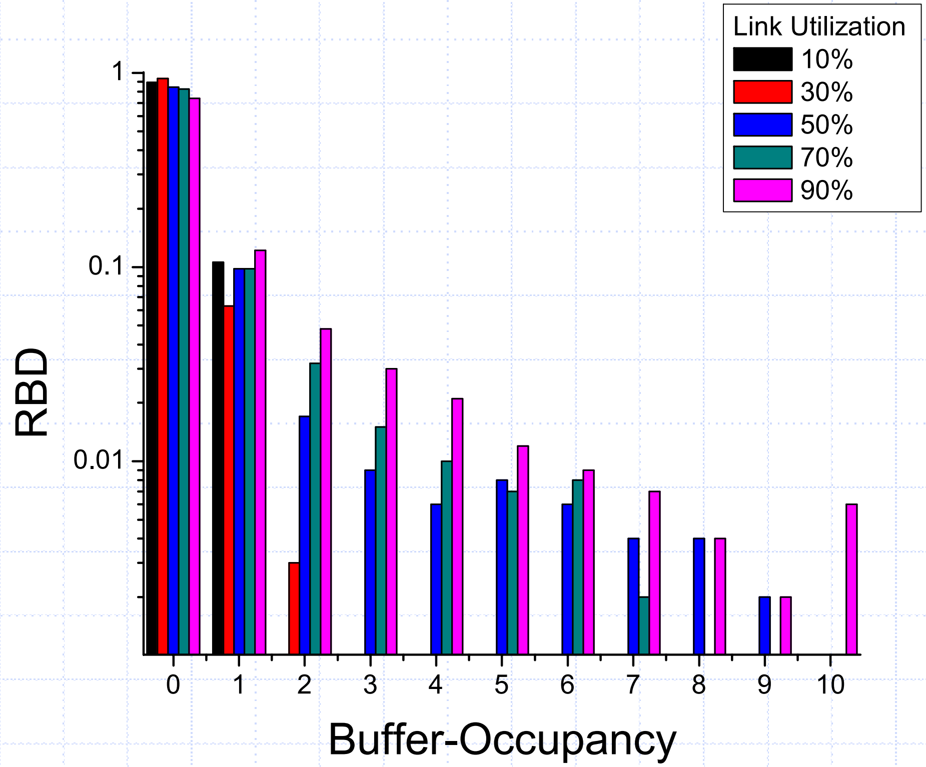
(at 50% utilization)

RBD vs. Link Speed

(at 50% utilization)

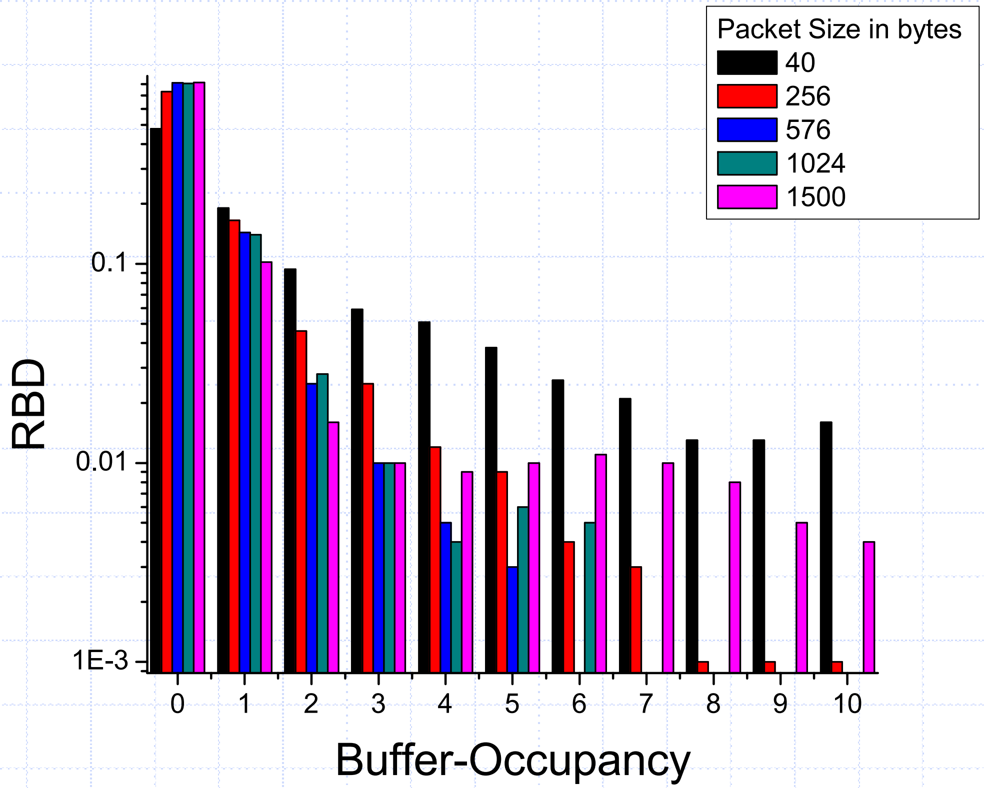


RBD vs. Link Utilization



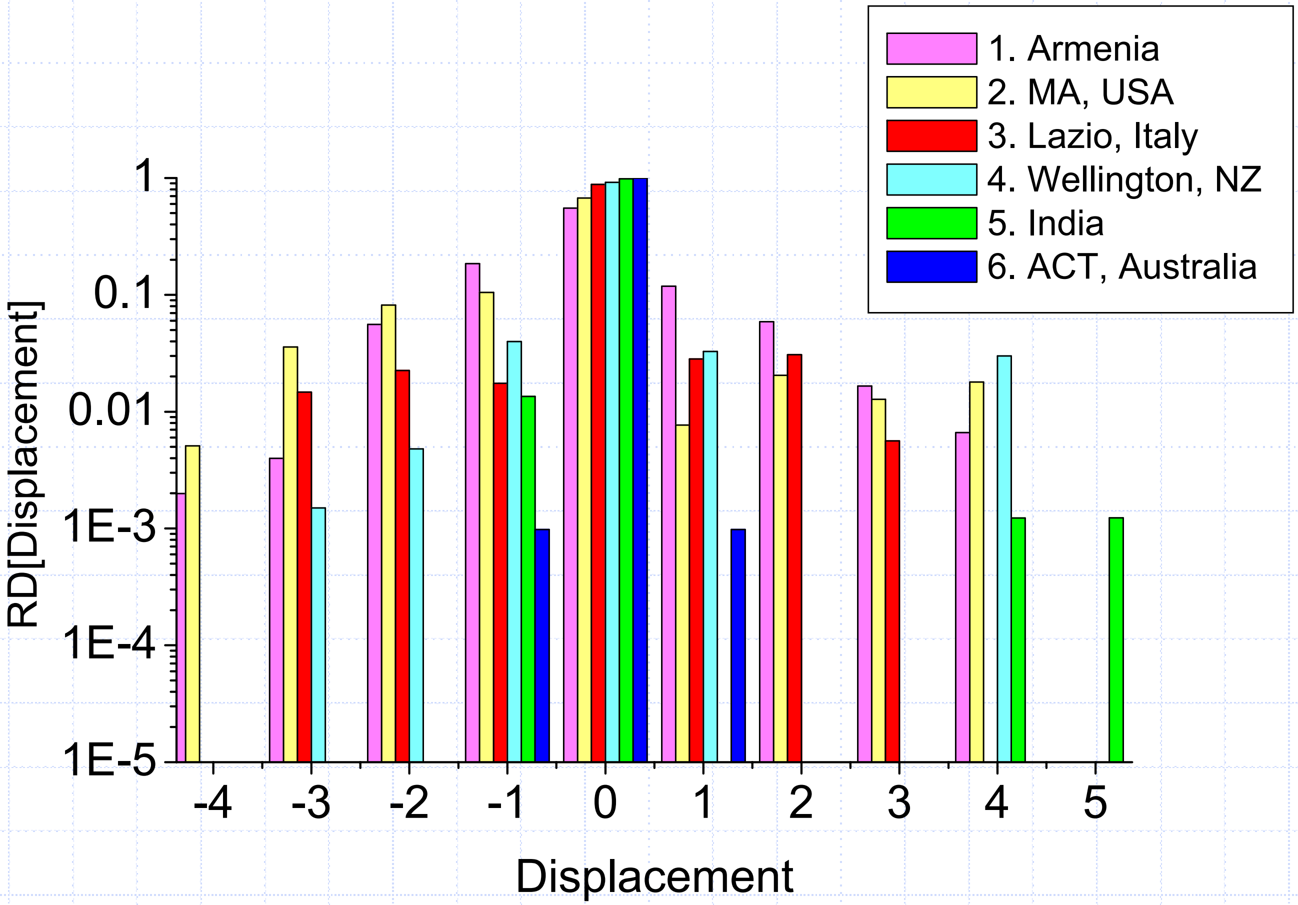
(For OC768, packet size 1500 bytes)

RBD vs. Packet Size



(with stream of interest occupying 10Mbps on a 50% utilized OC768)

Internet Measurements

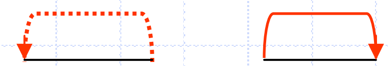


(a)

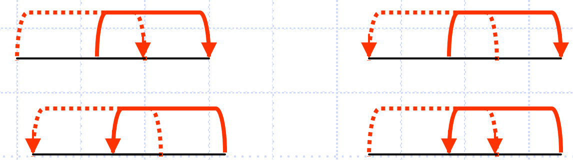
Internet Measurements

Net	*IR%*	*OR%*	*ER%*	*Other*
1	64	24	0	12
2	45	38	17	0
3	86	12	1	1
4	84	2	4	10
5	100	0	0	0
6	100	0	0	0
7	100	0	0	0
8	100	0	0	0
9	100	0	0	0

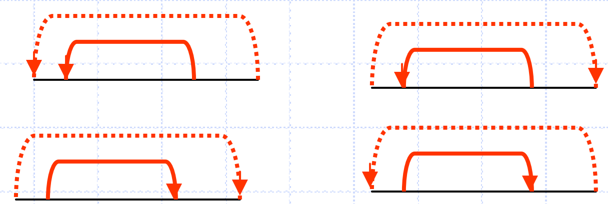
IR – Independent Reordering



OR – Overlap Reordering



ER – Embedded Reordering



Basic Patterns form significant amount of reordering patterns

Solutions?

- Deal with it at end nodes
- Input tracking and output buffering
- Creative designs – pipelined solutions
- Switch bursts of packets, use longer packets, …

Summary

- Increasing gap between link speeds and processing speeds demands increased parallelism within routers resulting in increased packet reordering
- Need to deal with packet reordering proactively to prevent it from negating some of the performance gain resulting from link and processing speeds.
- Throughput and delay characteristics are no longer sufficient to characterize network performance. Need to pay attention to secondary measures such as packet reordering