Small is Beautiful

Running Smaller Exchanges with Smaller Equipment

Nick Hilliard

CTO

nick@inex.ie

Some Background

- Prior to 2009
  - INEX operated dual LANs in two locations
  - 4 x Cisco 6500: GE-TX, SFP, 10G/XENPAK
- Opened up 1.5 new PoPs since 2009
  - Procurement process indicated that C6500 was too expensive
  - Problems with 10G support
    - 6704: low density, high cost, XENPAKs, port contention
    - 6708: same density for non-contended ports, X2
  - Created a tech wish-list
  - After beauty contest, settled on Brocade TI24X and FES-X624 fixed-configuration switches

This is a Picture of Some Switches

inex
internet neutral exchange

Financial Analysis

- Existing C6500 are expensive to run
  - Power charges of €0.29 per kWh (€2.50 / W / year)
  - Support costs based on high initial capex
  - Cost per port of 10G was too high for new 10G members and core links
- Plugged these figures into 5Y analysis spreadsheet
  - Assumed sale of existing C6500 at 20% less than eBay
  - Lower support costs for fixed config switches
  - Third party transceivers
  - By doing complete equipment swap-out right now, we could end up with significant 5Y savings.
- Cost per port of a Brocade TI24X 10G port is about 10% of cost of C6500 10G port

inex
internet neutral exchange

Well, that's all very interesting

But Will It Blend?

Due Diligence

Wire Speed on all ports	Unicast flood control
Wire Speed L2, L3 filtering	OEM Optics
IPv6 ACLs on L2 interfaces	TDR support on TX ports
DHCP Snooping	Link aggregation with full features
IPv6 RA Guard	Port mirroring
PIM Snooping	Remote port mirroring
IGMP Snooping	Rapid spanning tree
MLD Snooping	BPDU guard
Dynamic ARP inspection	Bridge management other than than STP
Port security (mac address counting)	SSH CLI management
Sflow / Netflow	UDLD
Mac address accounting using ACL counters	Environmental monitoring
Broadcast / multicast storm control	Dual Hotswap PSU

IxAutomate Report Run0001.res

rfc2544 - Throughput - Aggregate Results

Trial	Frame Size	Agg Throughput (fps)	Agg Throughput (Mbps)	Max Throughput (fps)	Max Throughput (Mbps)	Agg Throughput (%)	Rx Sequence Errors
1	64	29,761,904.76	15,238.10	14,880,952.38	7,619.05	100.00	0
1	128	16,891,891.90	17,297.30	8,445,945.95	8,648.65	100.00	0
1	256	9,057,971.02	18,550.73	4,528,985.51	9,275.36	100.00	0
1	512	4,699,248.12	19,248.12	2,349,624.06	9,624.06	100.00	0
1	1024	2,394,636.02	19,616.86	1,197,318.01	9,808.43	100.00	0
1	1280	1,923,076.92	19,692.31	961,538.46	9,846.15	100.00	0
1	1518	1,623,376.14	19,714.28	811,688.07	9,857.14	100.00	0
2	64	29,761,904.76	15,238.10	14,880,952.38	7,619.05	100.00	0
2	128	16,891,891.90	17,297.30	8,445,945.95	8,648.65	100.00	0
2	256	9,057,971.02	18,550.73	4,528,985.51	9,275.36	100.00	0
2	512	4,699,248.12	19,248.12	2,349,624.06	9,624.06	100.00	0
2	1024	2,394,636.02	19,616.86	1,197,318.01	9,808.43	100.00	0
2	1280	1,923,076.92	19,692.31	961,538.46	9,846.15	100.00	0
2	1518	1,623,376.14	19,714.28	811,688.07	9,857.14	100.00	0

10G input, 8 x Snake ports

i.e. 80G full duplex throughput / 160G overall throughput

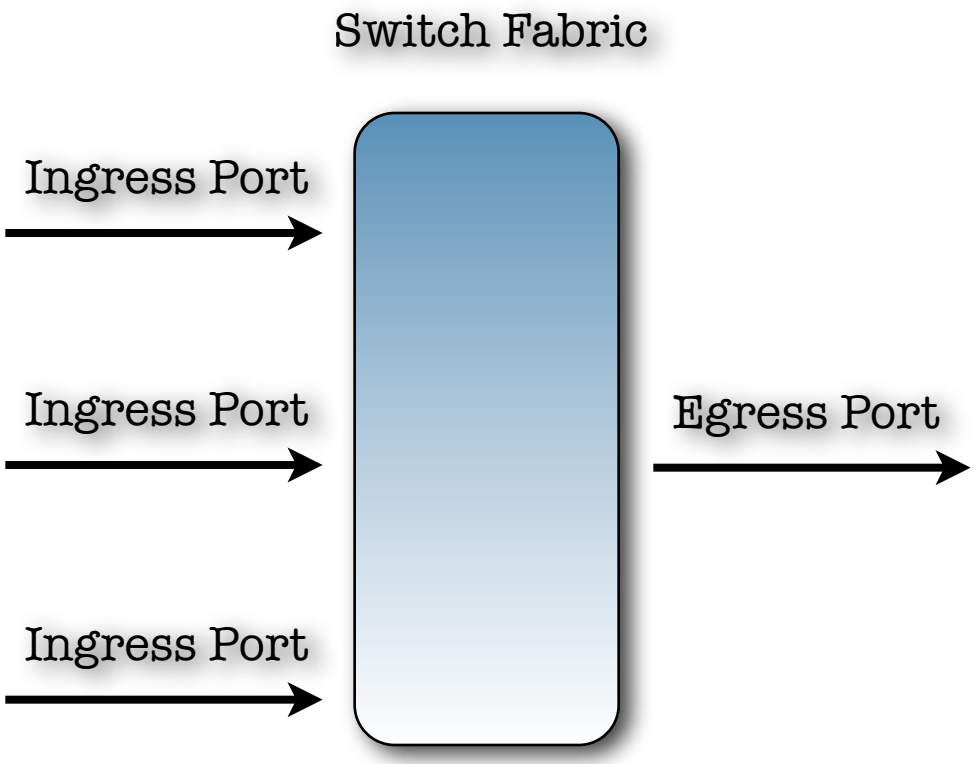
Switch Features

- Feature compatibility results were good
  - FES-X6xx: lacks L2 ethertype filtering
  - TI24X: lacks sflow5
  - both currently lack RA Guard, mld snooping but support pim/igmp snooping
  - Doesn't look like there are hardware limitations
  - Features are on road-map
  - INEX doesn't need L3 functionality or fancy features
- But the really interesting questions surround switch architecture
  - New generation of ToR switches are cut-through rather than store-n-forward
  - Specifically queueing and buffering

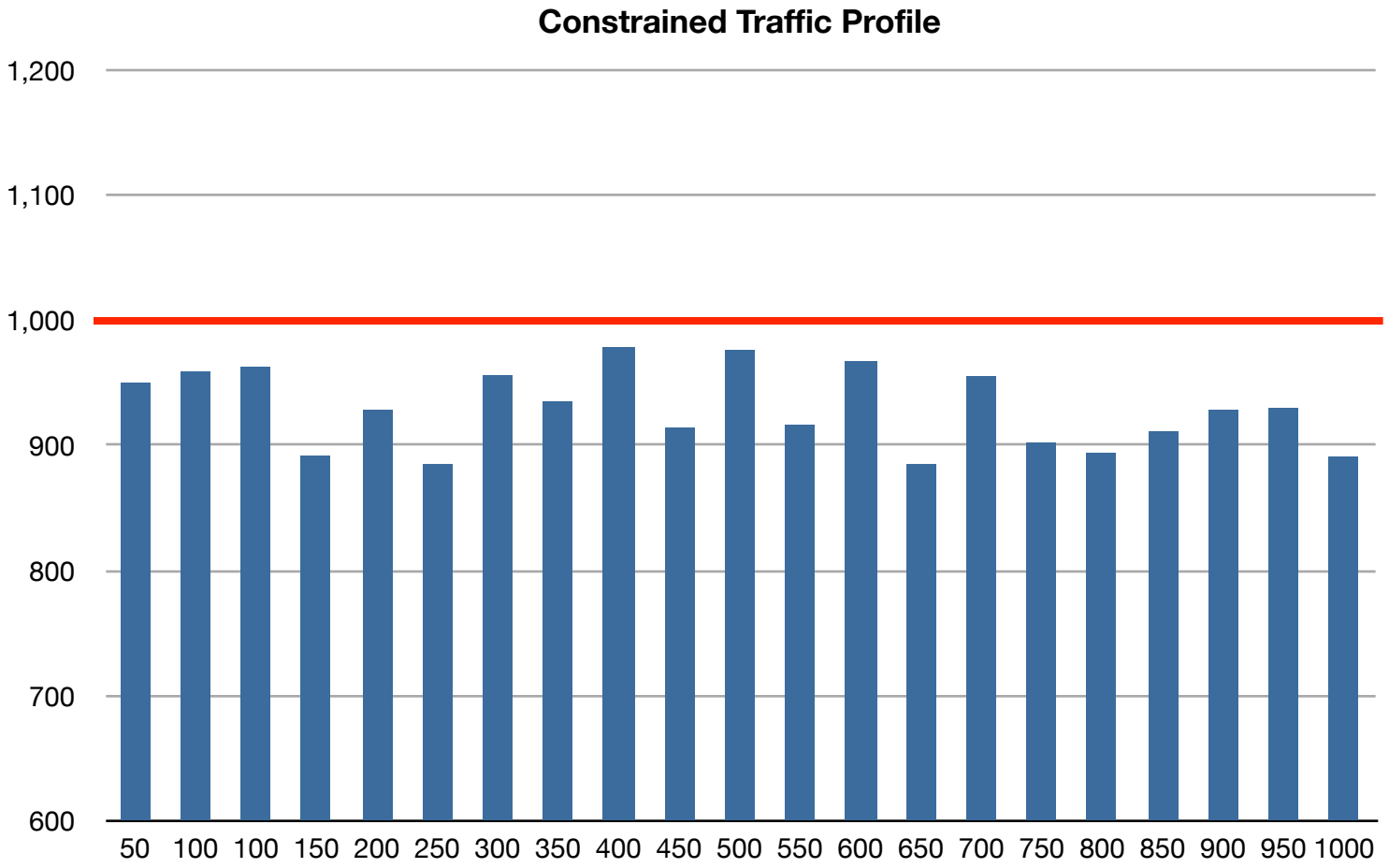
Buffering

- Store-n-forward: switch receives the entire frame before forwarding to destination port
- Cut-through switches
  - starts forwarding packet to destination port as soon as it receives destination mac address
  - requires less buffer space
    - WS-X6704-10GE: 16Mb per port
    - TI24X: 2Mb shared between 24 10G ports
  - recommended not to mix port speeds on the same box
- Buffers
  - Shared vs per-port
  - Queueing mechanism specifies how buffers are used

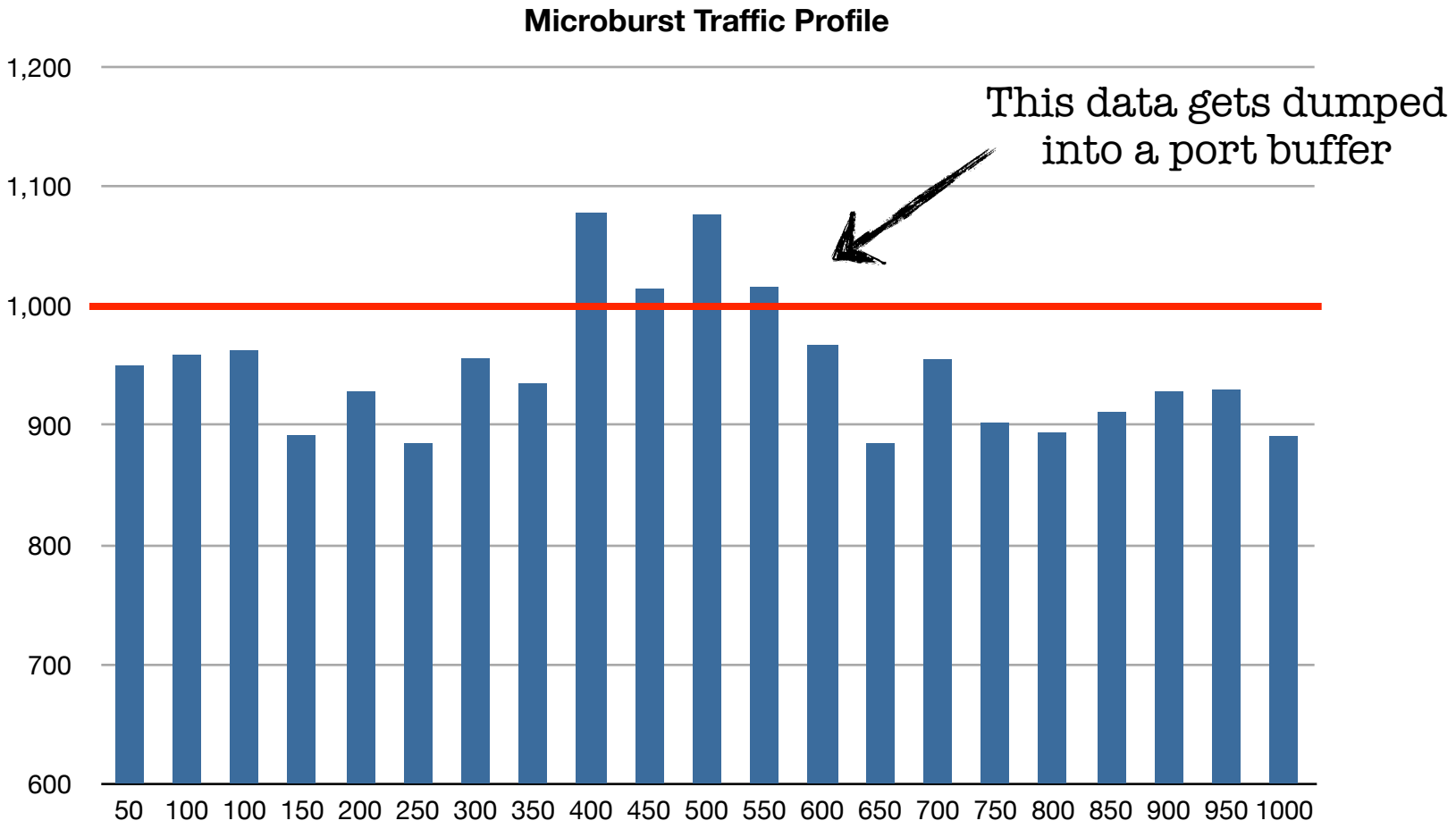
Simplistic Introduction to Buffering



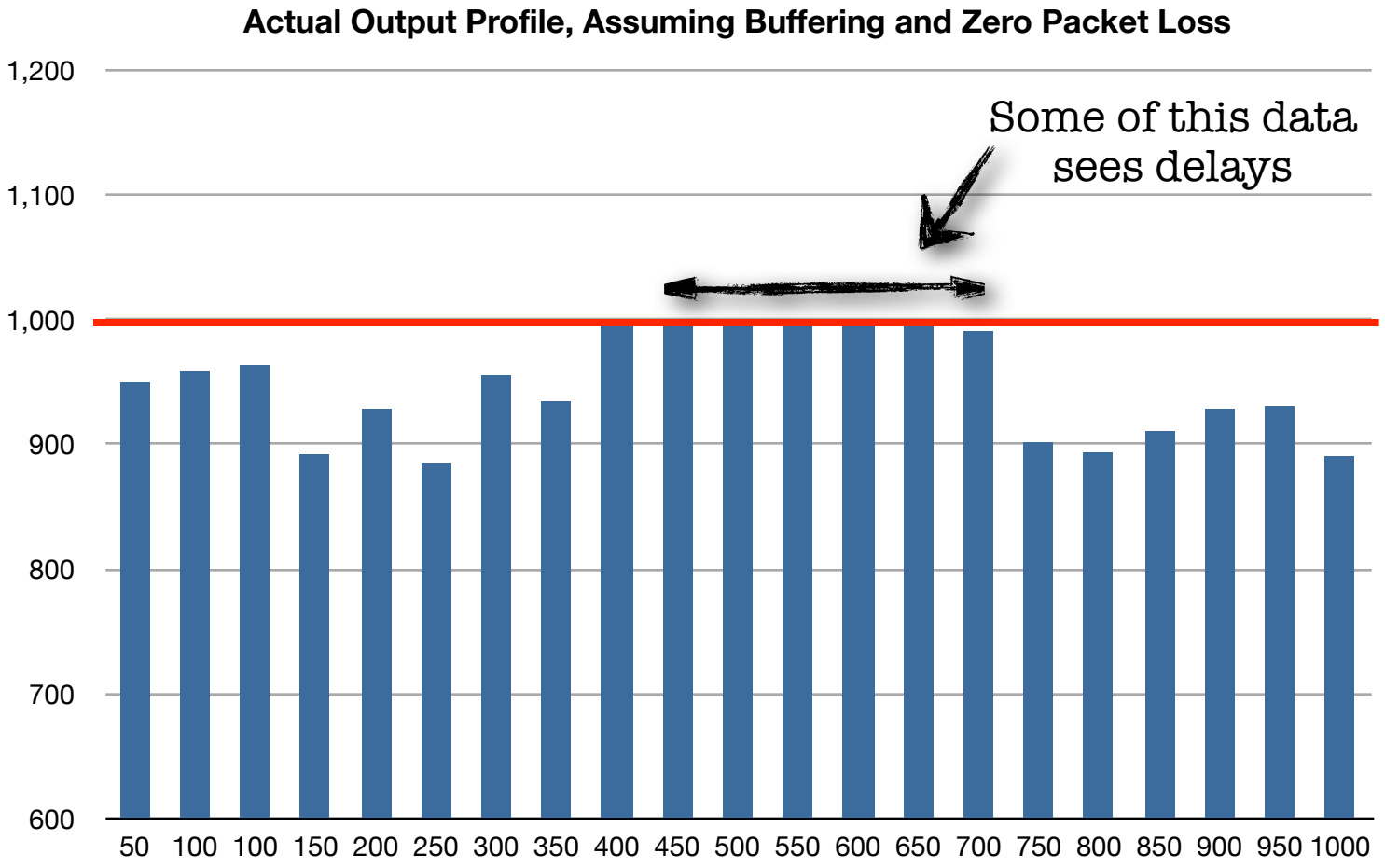
Microbursts



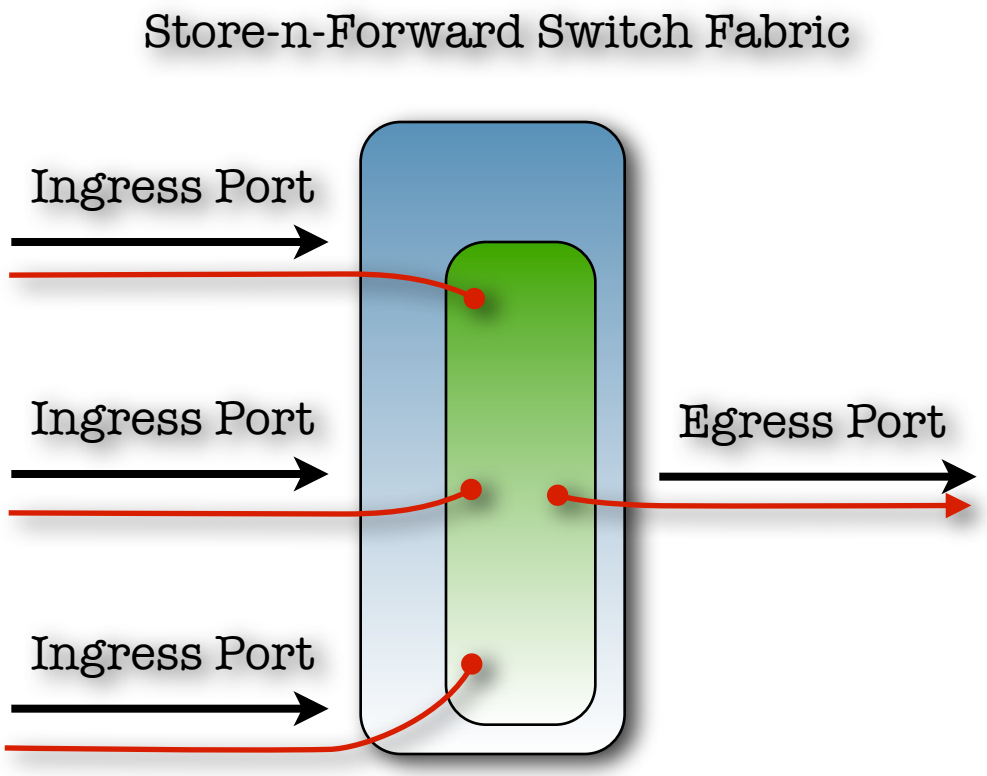
Microbursts



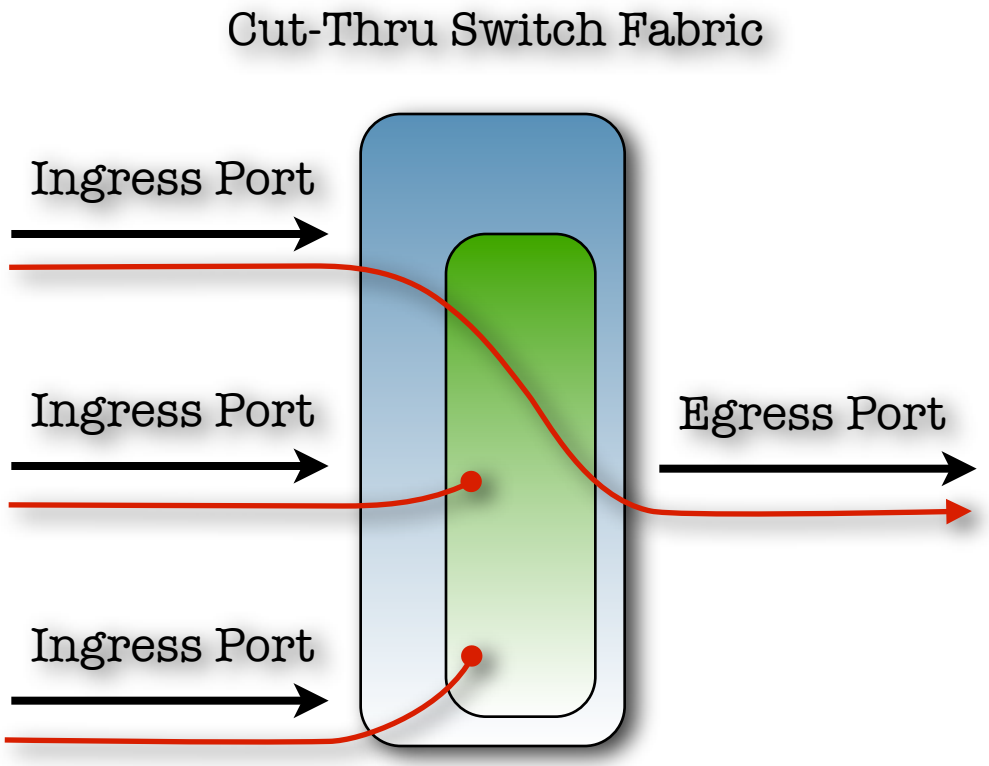
Microbursts



Buffering on Store-n-Forward Fabric



Simplistic Buffering on Cut-Thru Fabric



Observations on Buffering

- Different fabric forwarding architectures require different buffering mechanisms
  - e.g. microcell architecture vs whole packet switching
- Store-n-forward switches always require much bigger buffers than cut-thru switches
  - So, cut-thru switches generally built with smaller buffers
  - In some situations you may see more packet loss than on big buffer switches
  - Heavy outbound traffic will cause packets drops sooner on cut-thru switches than on big buffer switches
  - This can be avoided by implementing 10G to 1G step-down on different switches (e.g. core / edge separation
  - Lab setups can be invented to show that each methodology will work better in specific cases

Conclusion

- Will it work?
  - Yes, for INEX, but will not work for large IXPs
  - Certain limitations exist
    - Need aggressive monitoring of frame drops to find out why and where those frames are dropped
- Will it break?
  - "Big switch with big buffers" model scales much further
  - We look forward to having an exchange large enough for cut-thru model to break
- Recommendations:
  - Critical to understand buffering and queueing
  - Critical to implement extensive packet drop monitoring