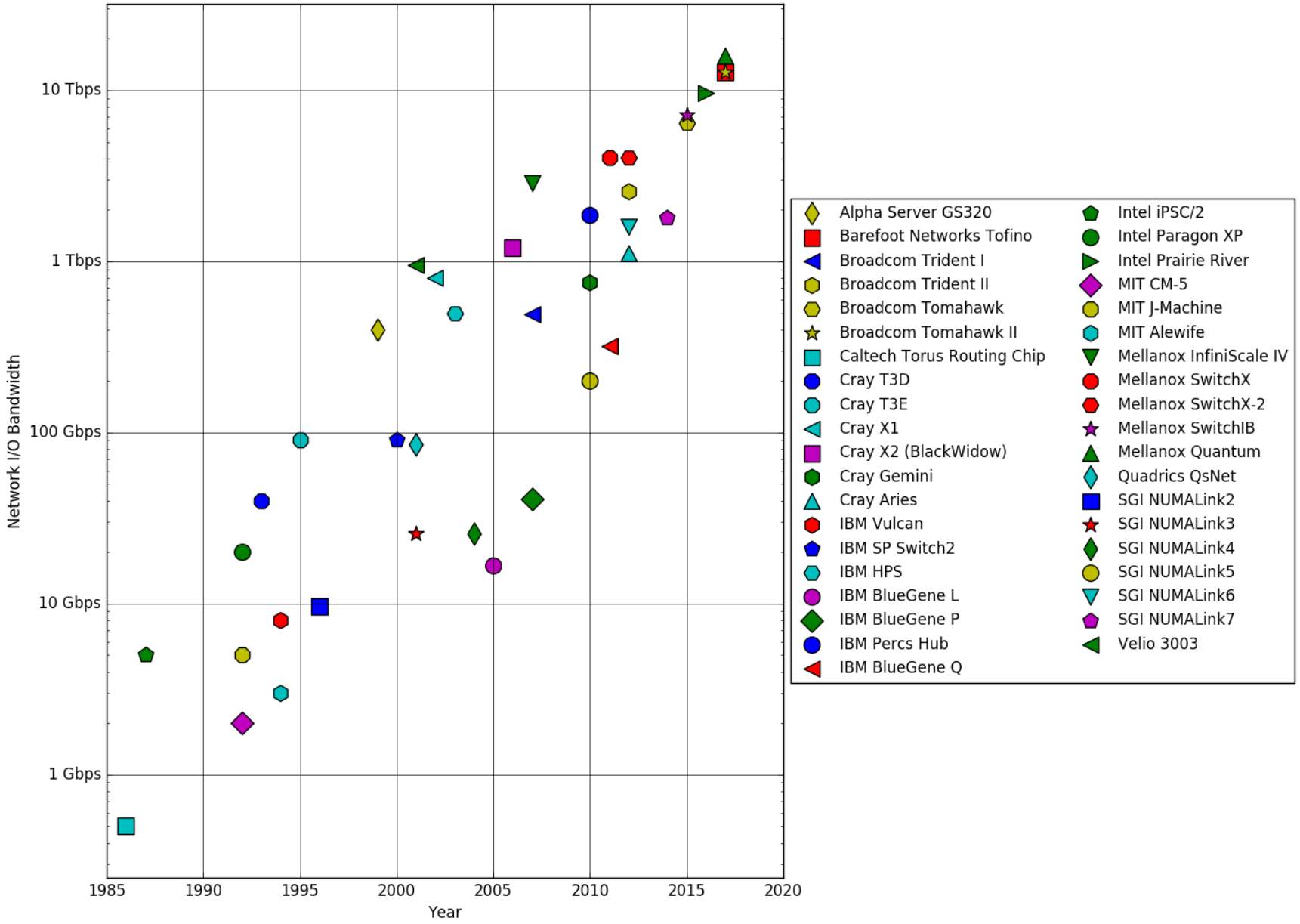
Hewlett Packard
Enterprise

Dear Photonics,
I love and hate you.
Sincerely, -Architect.

PHOTONICS Workshop – HPCA ‘17

Nic McDonald – February 4th, 2017
nicmcd@hpe.com

Chip Network Bandwidth



Hewlett Packard Enterprise

High-radix routers??

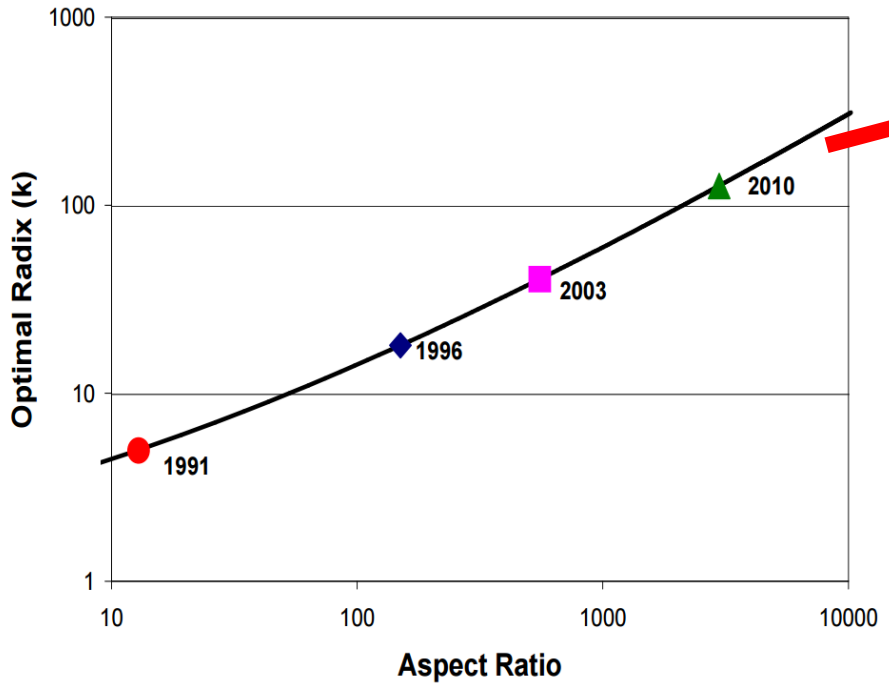


Figure 2. Relationship between optimal latency radix and router aspect ratio. The labeled points show the approximate aspect ratio for a given year's technology

$$k \log^2 k = \frac{B\, t_r \log N}{L}$$

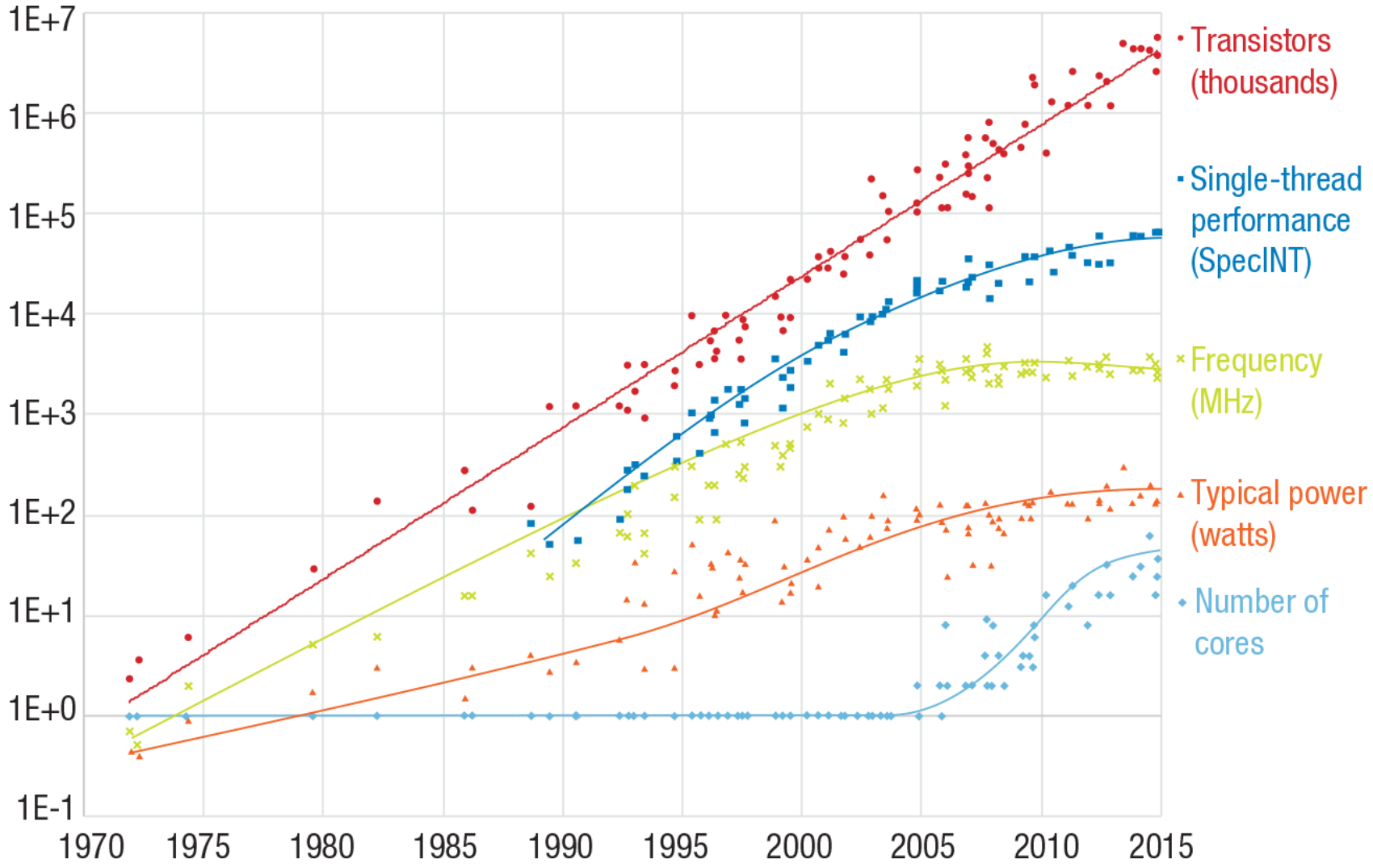
- Prediction made in 2005
  - "for 2010 technology the optimum radix is 127"
- Prediction made in 2009
  - "[radix 128] is typical of the high-radix switches we expect in the near future"
- Prediction made in 2011
  - 32nm (2012) will have radix 100 switches
  - 22nm (2014) will have radix 144 switches
- Reality:
  - 2006 (90nm) gave us the Cray YARC (radix 64)
  - …..

Massive bandwidths achieved through cheating

Technology	Lanes	SERDES Rate
10 G	4	2.5 Gbps
10 G	1	10 Gbps
40 G	4	10 Gbps
25 G	1	25 Gbps
50 G	2	25 Gbps
100 G	4	25 Gbps
50 G	1	50 Gbps
200 G	4	50 Gbps
400 G	8	50 Gbps
100 G	1	100 Gbps
400 G	2	100 Gbps
800 G	8	100 Gbps
1 T	10	100 Gbps
1.6 T	16	100 Gbps

Bandwidth is as useful of a metric as clock frequency is to processors

Processor Scaling



Hewlett Packard Enterprise

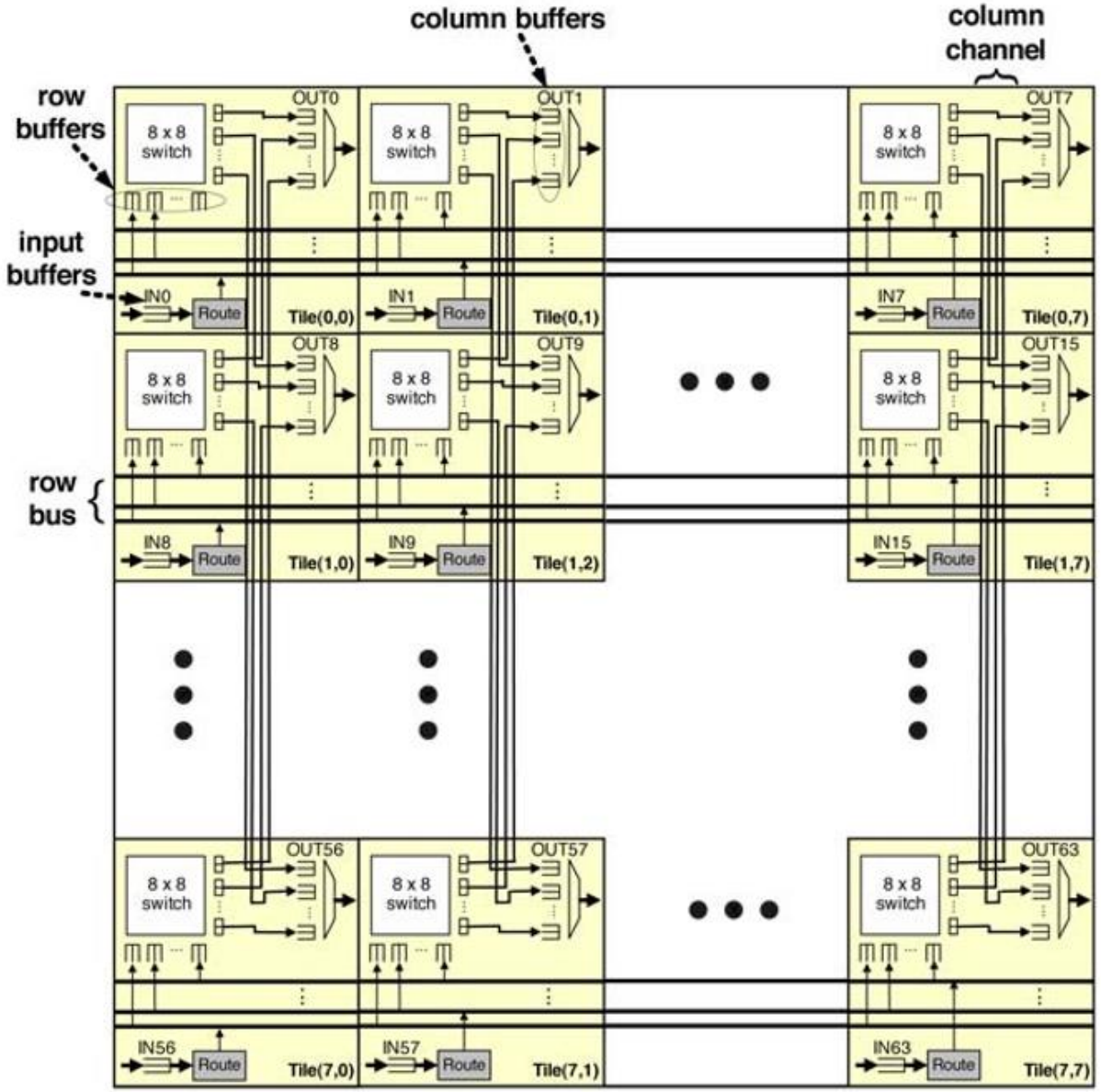
Overcoming long global wires

Cray YARC (ISCA 2006):

– “The router is organized hierarchically as an 8×8 array of tiles which simplifies arbitration by avoiding long wires in the arbiters.”

– Assumes ASICs have “abundant wiring but limited buffers”.

– Provides 8x the amount of wiring needed for uniformly distributed traffic.



Exploding datapath width

Technology	1 GHz	2 GHz	3 GHz	4 GHz
10 G	10	5	4	3
25 G	25	13	9	7
40 G	40	20	14	10
50 G	50	25	17	13
100 G	100	50	34	25
200 G	200	100	67	50
400 G	400	200	134	100
800 G	800	400	267	200
1 T	1000	500	334	250
1.6 T	1600	800	534	400

Example

Cray Aries has a 42 Gbps port rate (3x14G) and a 875 MHz core frequency, thus a 48-bit datapath, and 8x is 384 wires.

If Cray upgrades to 200G staying at 875 MHz, that yields a datapath of 229 bits, and 8x is 1832 wires.

Can we do better than a hierarchical crossbar design?

– We can organize the switch as its own embedded topology.

– This scales better than a hierarchical crossbar because it doesn't need to overprovision the datapath.

– However, you now have to deal with cyclic dependencies within EVERY router chip. You are likely to need at least one VC per router. Possibly two with non-minimal routing.

– Dragonfly needs 3-5 VCs (depending on the features you give it). Requests and responses must be isolated in many HPC systems so yields 6-10 VCs.

– The standard Dragonfly makes up to 6 router traversals, the Cray XC Dragonfly makes up to 9 router traversals.

– A Cray XC system using embedded networks might require 18 or 36 VCs just to alleviate network and protocol deadlock!

– What about using VCs for priorities and/or isolation?

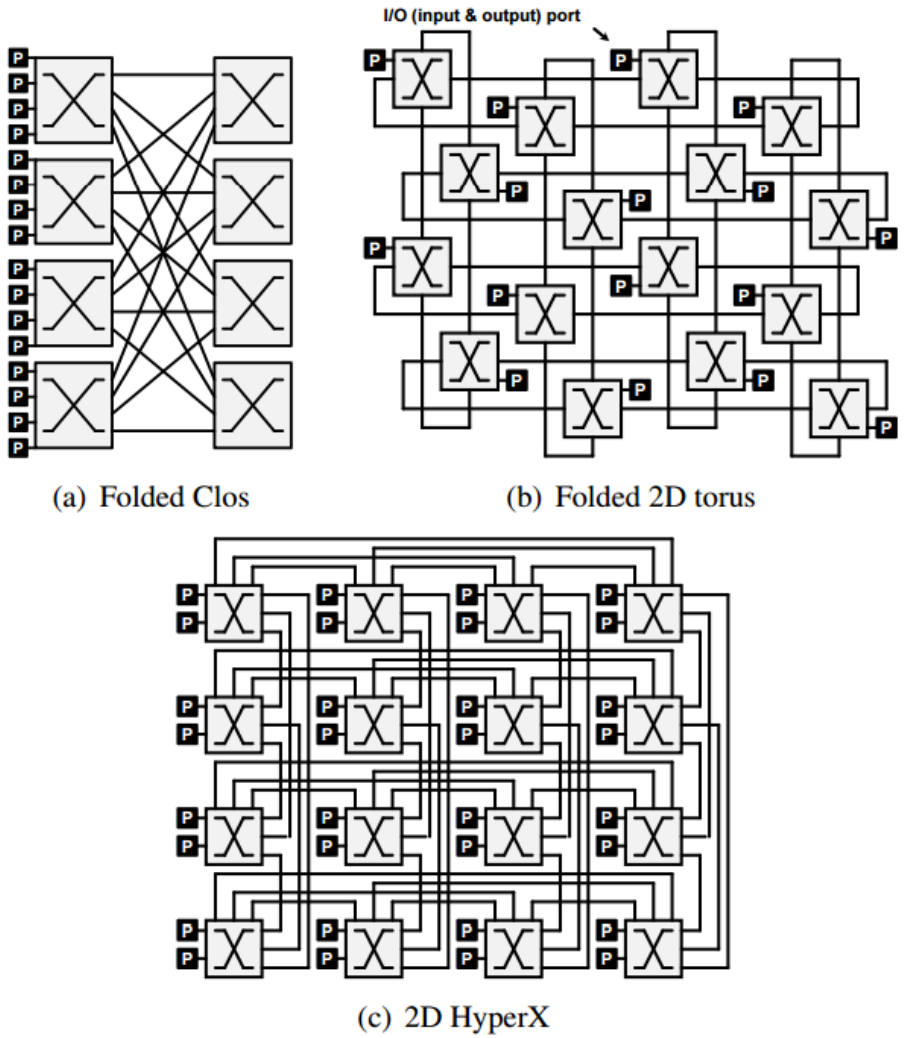
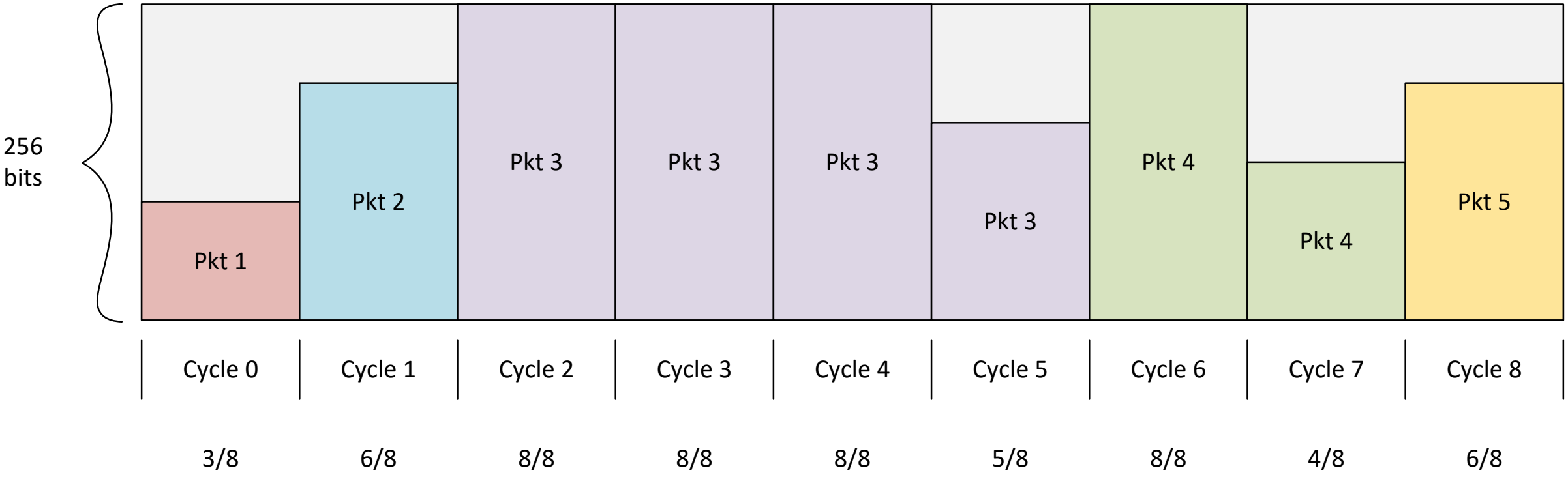


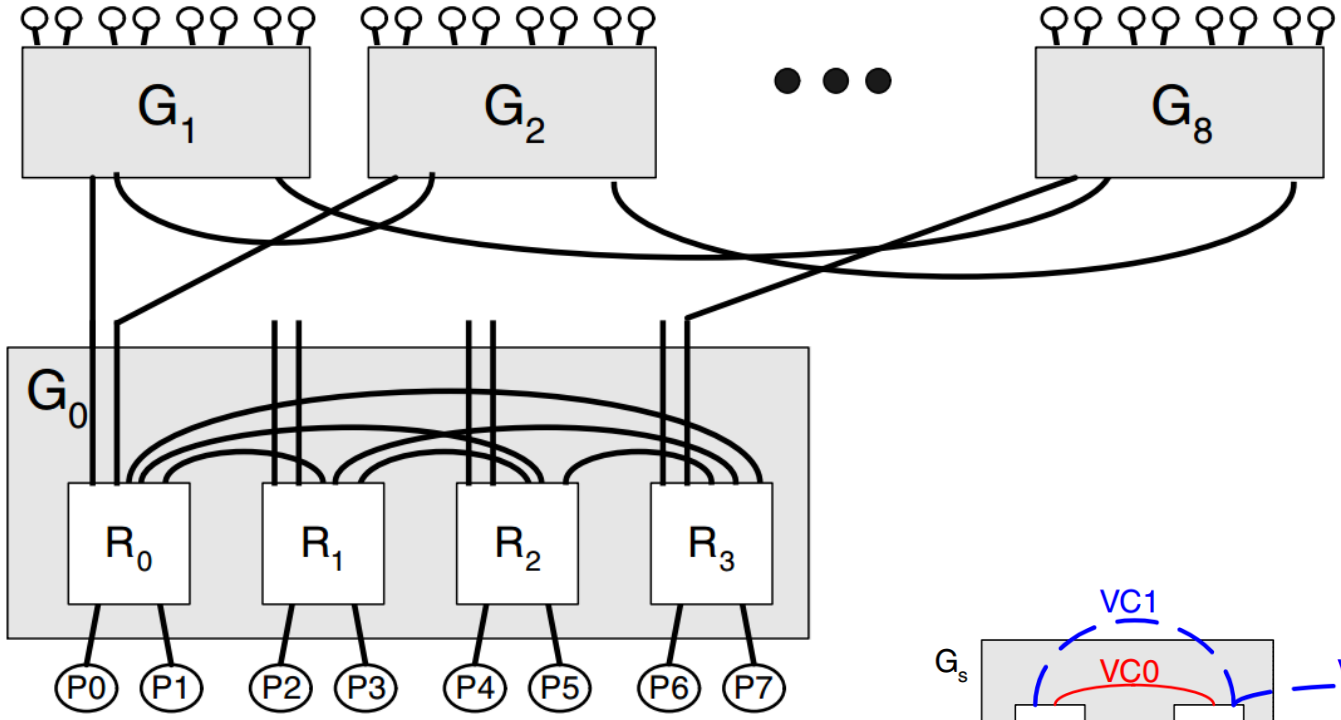
Figure 4: (a) A folded 2D torus, (b) a 2D HyperX, and (c) a folded-Clos switch organization.

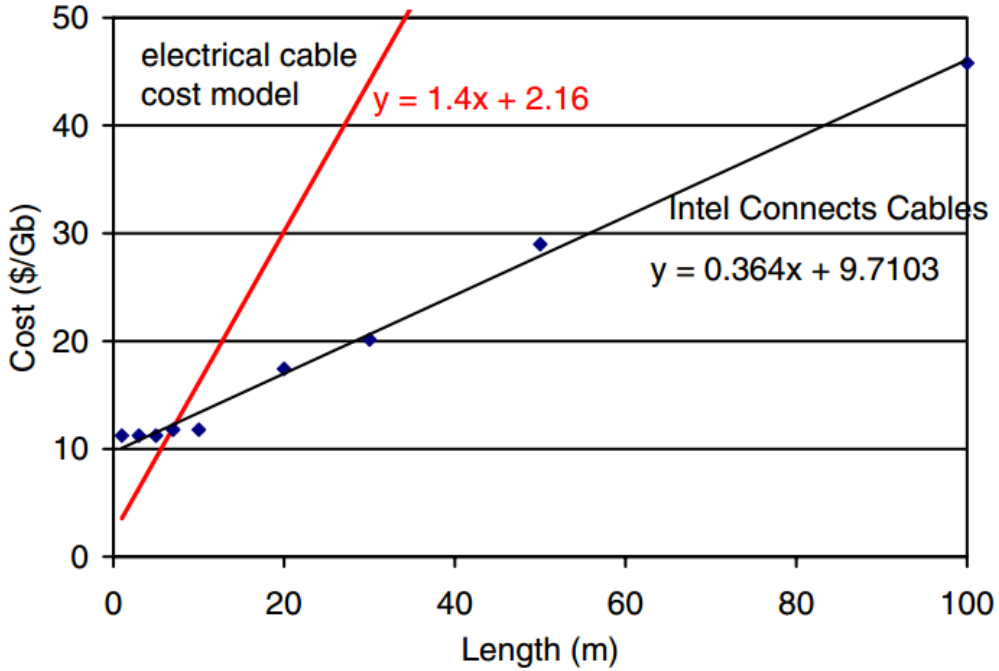
Protocol and datapath efficiency

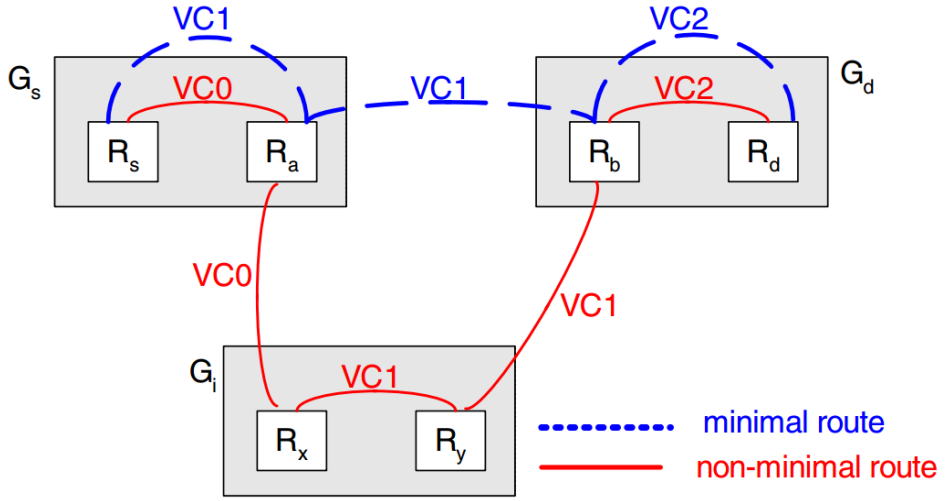


Only 77.7778% efficiency achieved

Scared of optics

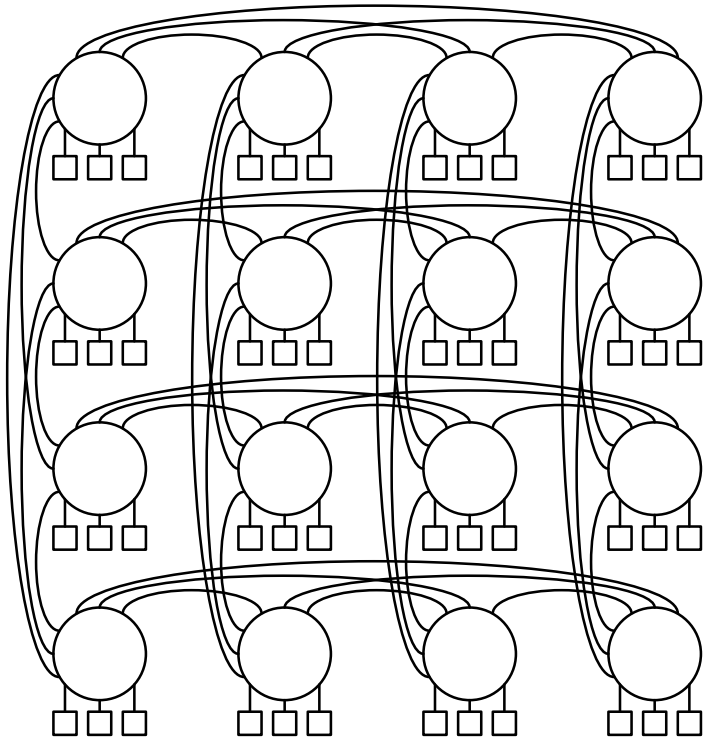






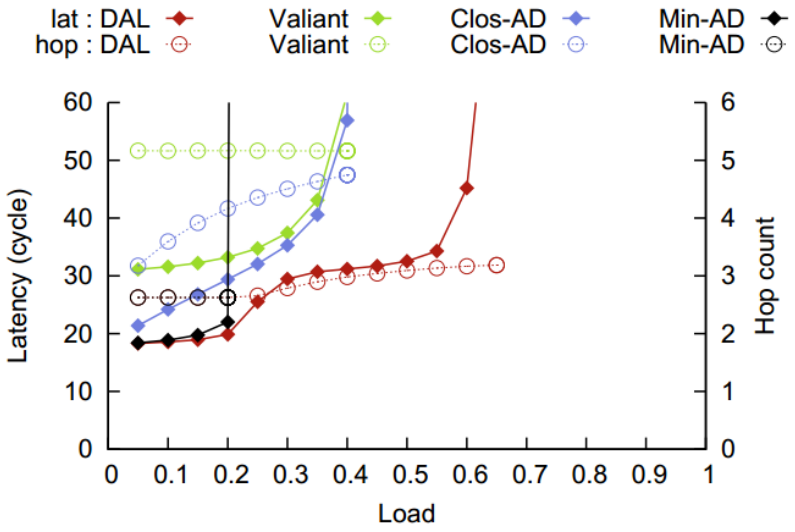
Hewlett Packard Enterprise

Embracing optics

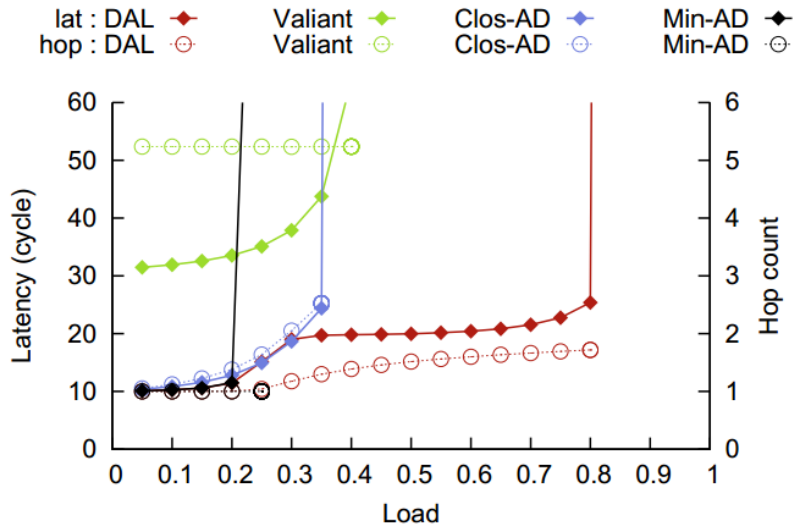


B	Best Regular	Switch Count	Best General	Switch Count
0.125	$L = 4,\ S = 7,\ K = 2,\ T = 55$	2401	$S = (5, 19, 19),\ K = (4, 1, 1),\ T = 76$	1805
0.25	$L = 3,\ S = 14,\ K = 2,\ T = 48$	2744	$S = (3, 27, 30),\ K = (9, 1, 1),\ T = 54$	2430
0.5	$L = 3,\ S = 16,\ K = 2,\ T = 32$	4096	$S = (3, 35, 36),\ K = (12, 1, 1),\ T = 35$	3780
1.0	$L = 4,\ S = 10,\ K = 3,\ T = 14$	10000	$S = (5, 38, 38),\ K = (8, 1, 1),\ T = 19$	7220

Table 1: Best regular and general HyperX networks for $N = 2^{17}$ and $R = 128$.



(c) Transpose

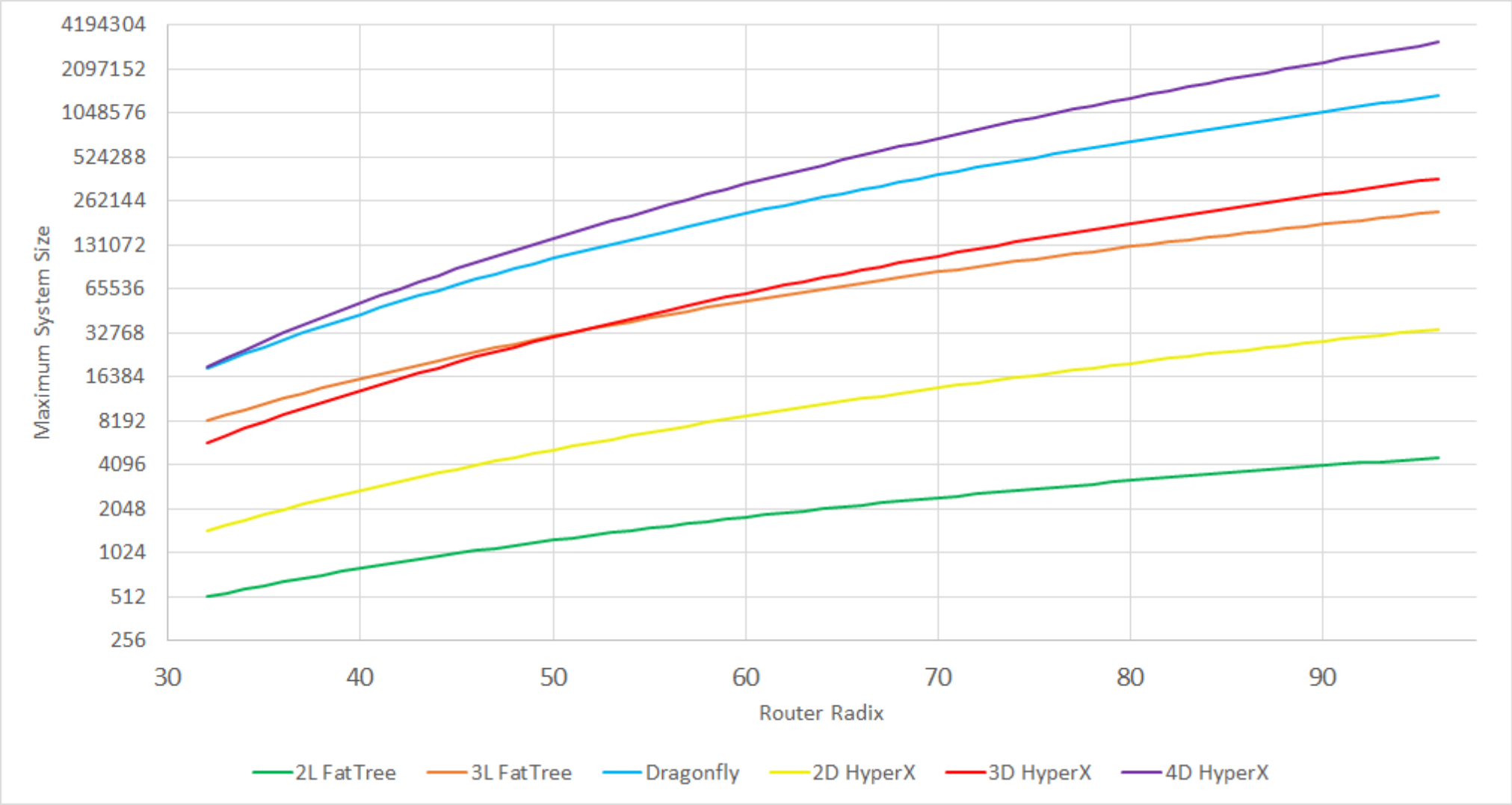


(d) Swap2

Hewlett Packard Enterprise

Topology Scalability

* topologies are designed for 100% UR throughput



D=3 D=5 D=4 D=3 D=4 D=5

Hewlett Packard Enterprise

Request #1 – Stay away from forward error correction

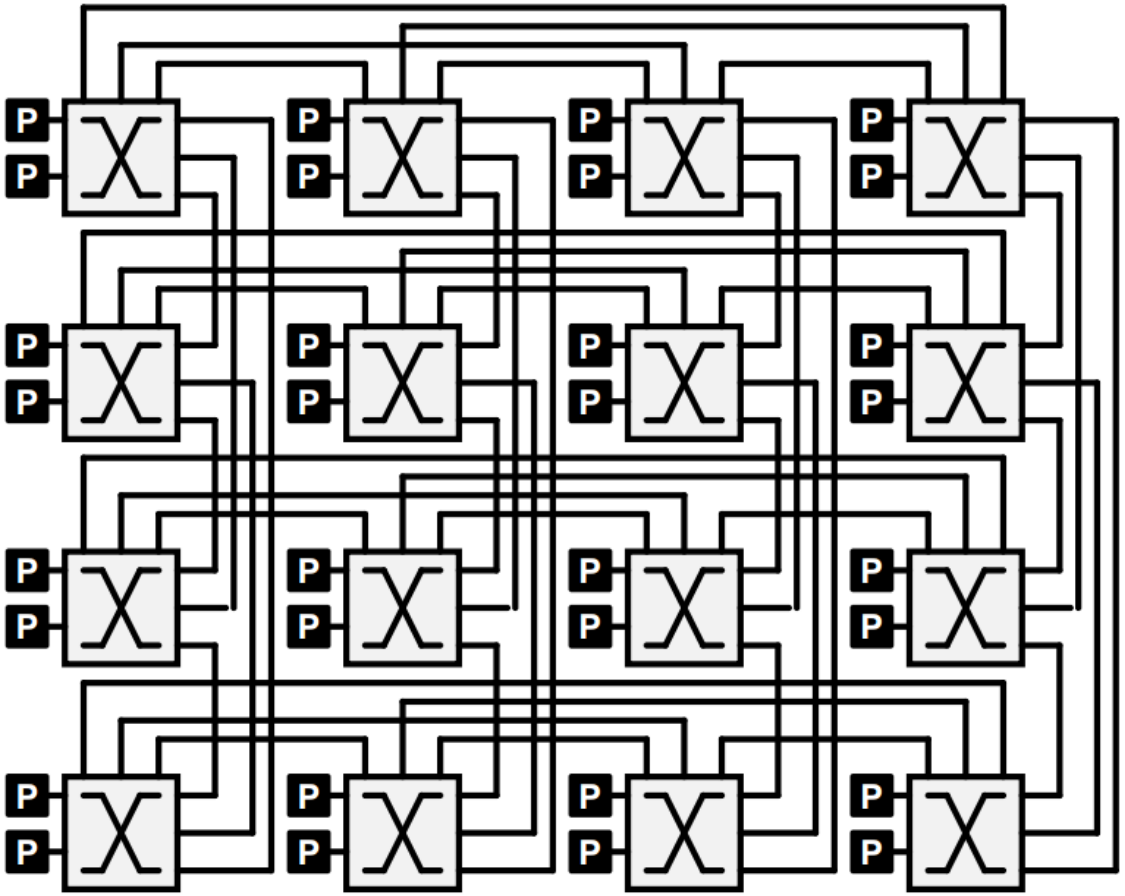
Forward Error Correction:

- Wastes time
  (~100ns of latency)
- Wastes bandwidth
  (extra bits sent)
- Wastes power
  (extra bits sent and correction logic)
- Wastes area
  (long pipelines of correction logic)

The 400G-PSM4 (4x100G) proposal:

- Uncorrected BER is < 2.1e-5
- Corrected BER is < 1e-15

Please make good channels without FEC



Request #2 – Design low latency SERDES

A quote from a fellow anonymous architect:

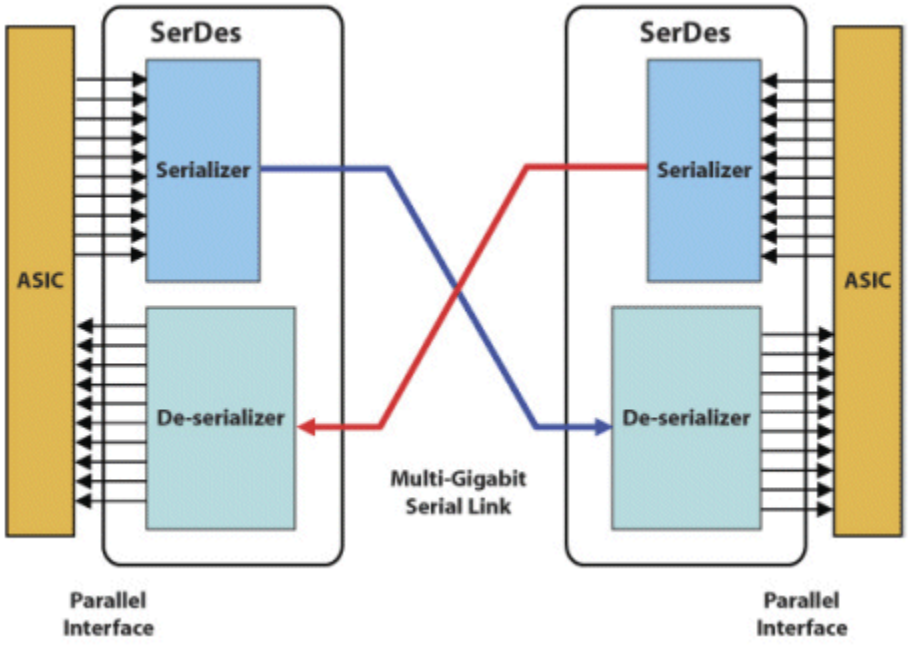
“the SERDES people seem to be adding cycles to their design like its going out of style :)”

Things I’ve noticed:

– The delay through a SERDES varies widely in industry. (a few cycles to many 10s of cycles)

– General purpose SERDES are slow and difficult to interface with the architecture.

– Special purpose SERDES are fast and interface well with the architecture.



Request #3 – Find efficiency in port aggregation

Architects don’t care whether the port is comprised of 1 lane or 100 lanes.
Architects do care about how the ASIC interfaces with the “port”.

Microring example:

– Two independent analyses of theoretical limits of microrings and DWDM show that maximum bandwidth in a fiber is approximately 1.25 Tbps.

– Practical limitations to microrings limit each ring to 25 GHz.

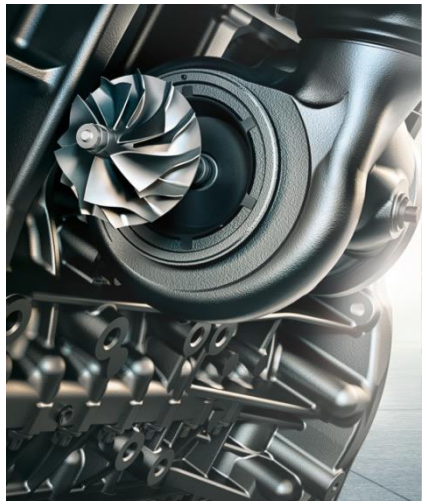
– With PAM4: 1.25 Tbps / 50 Gbps = 25 Tx rings, 25 Rx rings, 25 SER, 25 DES

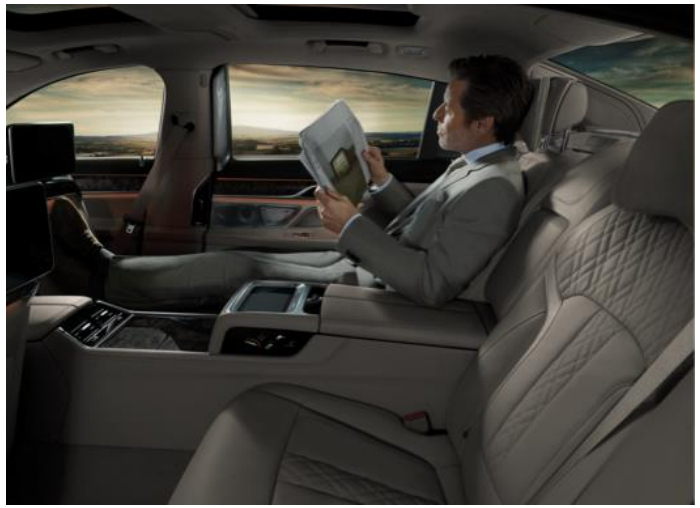
– Without PAM4: 1.25 Tbps / 25 Gbps = 50 Tx rings, 50 Rx rings, 50 SER, 50 DES

– At what point does the “small” microring still dominate the die area due to large numbers?

(Selfish) Request #4 – Drive development through HPC

4.4-liter turbocharged
8-cylinder engine
600 horsepower
590 lb-ft torque





Hewlett Packard Enterprise

(Selfish) Request #4 – Drive development through HPC

Hewlett Packard
Enterprise

Hewlett Packard
Enterprise

Thank you


Hewlett Packard
Enterprise