OPENFABRICS ALLIANCE

Innovative Topologies

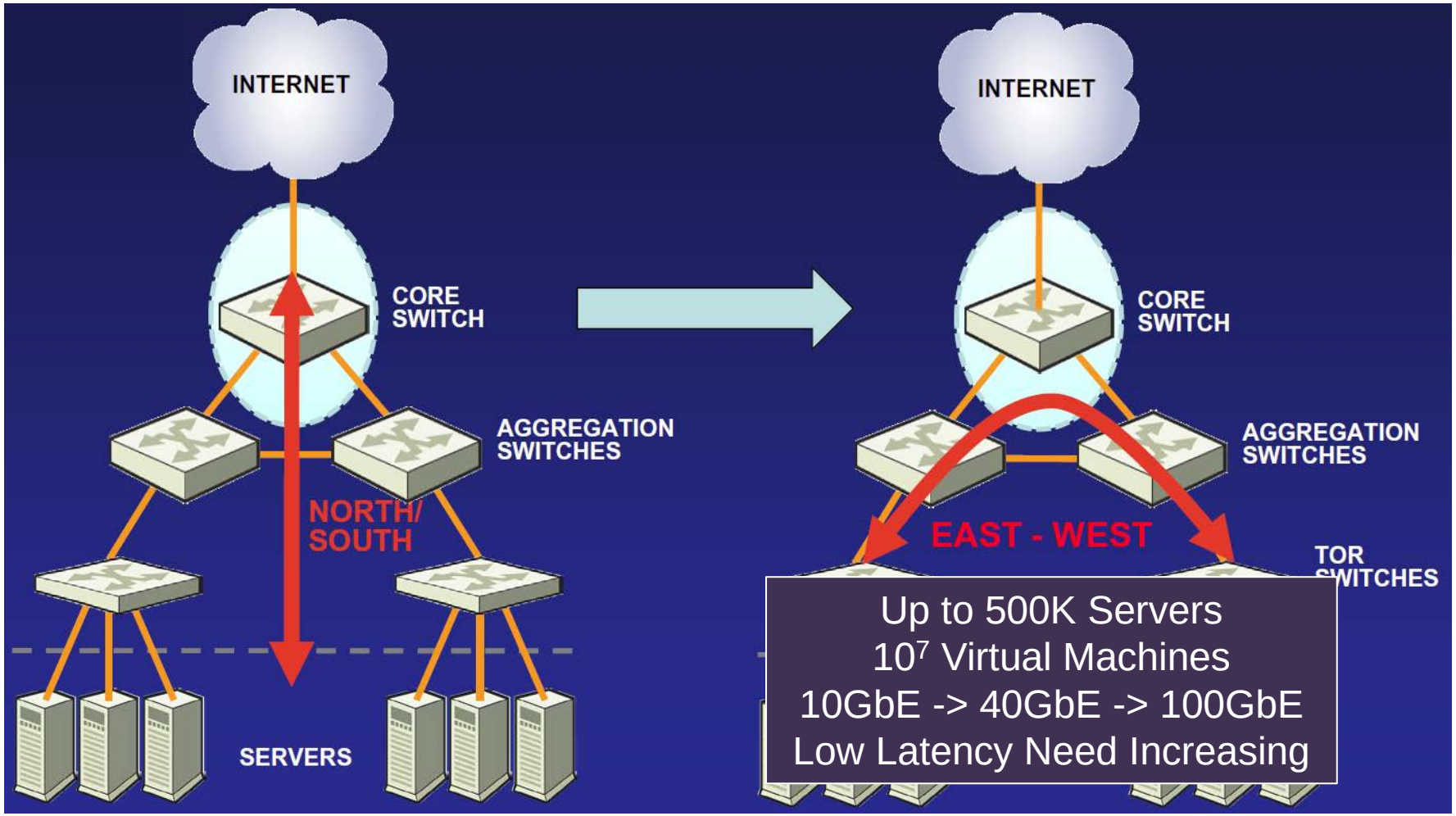
2013 OFA Developer Workshop

Authors: Harry V. Quackenboss, Ratko V. Tomic
Company: Infinetics Technologies, Inc.
Date: April 23, 2013

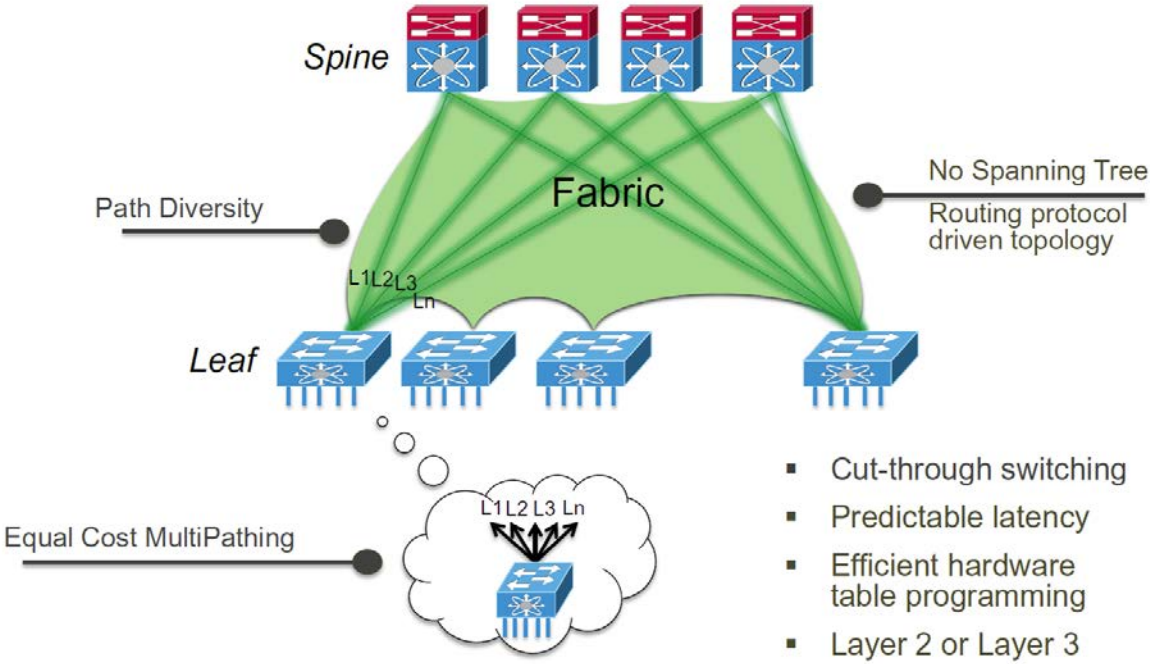
Network Throughput Optimization via Error Correcting Codes

Preprint: [arXiv 1301.4177 cs] Jan 17, 2013 http://arxiv.org/abs/1301.4177

Evolving Data Center Needs

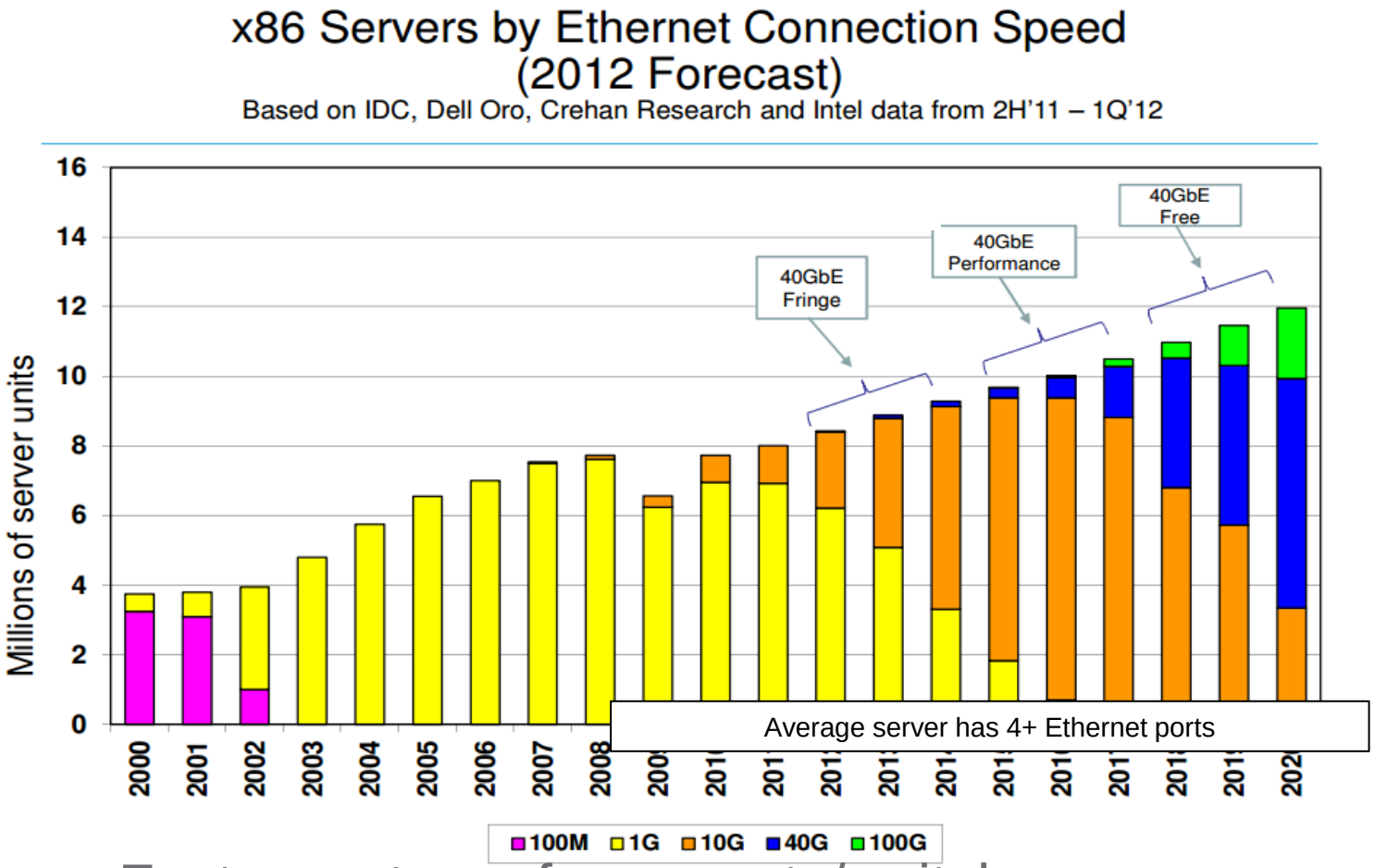


Fat Trees Also Have Challenges



- Congestion degradation when load > ~55%
- Can't take advantage of non-optimal paths under load surges
- Not practical for any-any connectivity across large data centers
- Max two stage edge ports: (# ports/chassis)^2/2
- Major re-cabling to move from two stages to three stages
- Faster port speeds mean more smaller switches

100GbE Server Connections < 3 Years



- Faster ports => fewer ports/switch
- Fat Tree gets worse

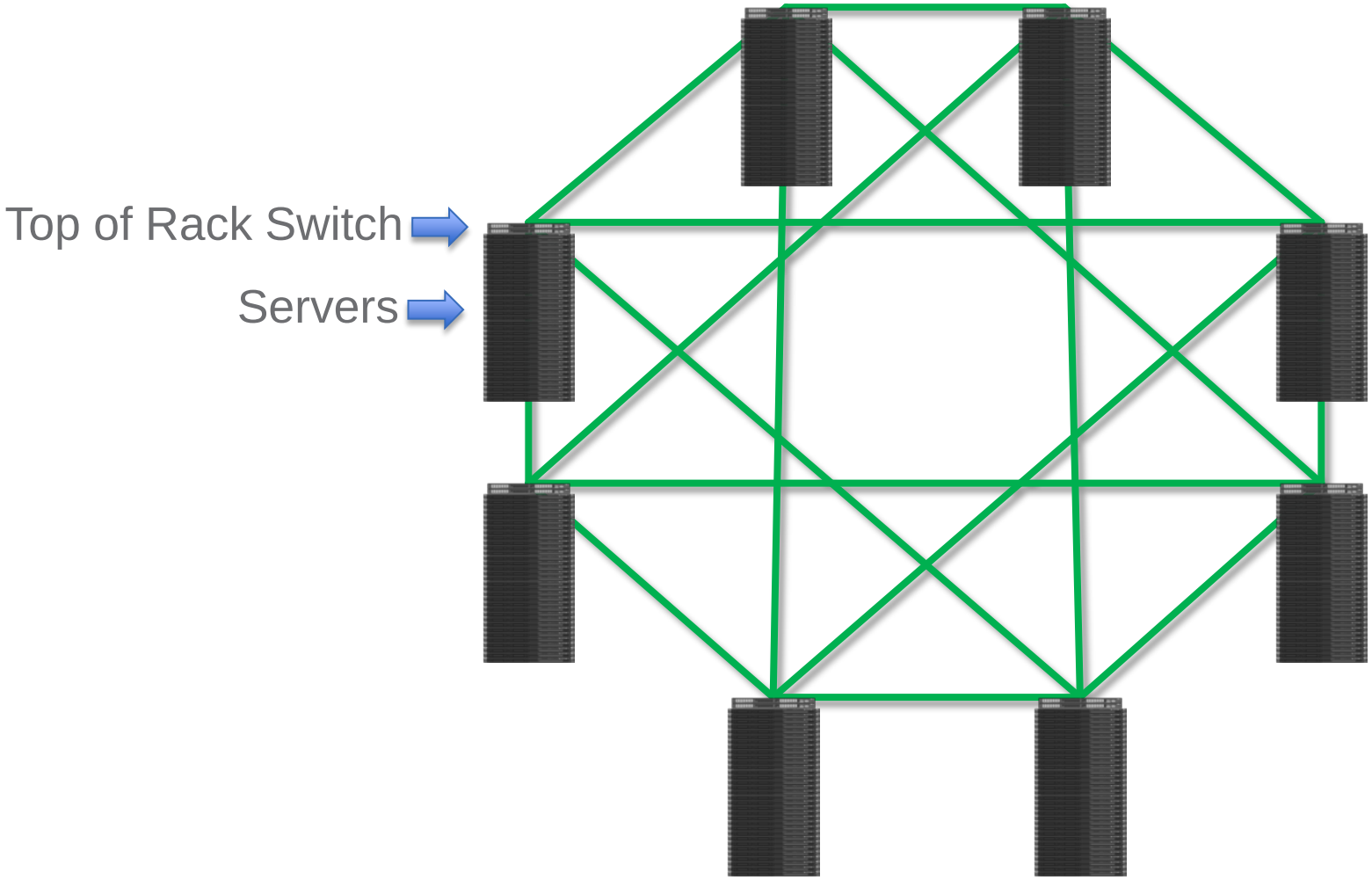
And What About SDN?

- SDN (HPC'ers definition)
  - *Concepts of InfiniBand Subnet Manager talking to Ethernet switches with learning and spanning tree disabled, using a new protocol (e.g., OpenFlow) to manually program the forwarding tables…*
- SDN has (almost) no traction for server fabrics
- *Same L2/L3 protocols on OpenFlow*
- *Same topology = same performance*
  - *At best…*

Infinetics Long Hop™ Alternative to FT

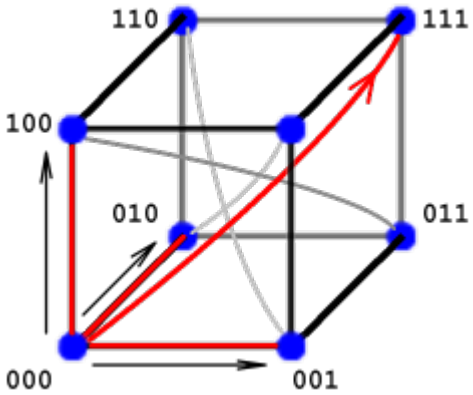
- More effective topology for any-to-any traffic
  - Lesser cost/complexity [100 - 1M] Switches
  - Larger bisection bandwidth [1.5 - 3X]
  - Larger # of edge ports
  - Fewer cables [1.7 – 3.5X]
- For given # of ports & oversubscription
  (any traffic pattern)
  - Higher average link utilization [1.5 – 1.9X]
  - Lesser packet loss
  - Fewer average hops/latency [1.5 – 2X]
- Current protocol friendly
  - Including RDMA, RoCE, etc.
  - Adaptable to InfiniBand, etc.

Smallest Long Hop Layout



Hypercube-like Networks

Folded Cube FC_3



Adjacency matrix A

	0	1	2	3	4	5	6	7
0	-	1	1	-	1	-	-	1
1	1	-	-	1	-	1	1	-
2	1	-	-	1	-	1	1	-
3	-	1	1	-	1	-	-	1
4	1	-	-	1	-	1	1	-
5	-	1	1	-	1	-	-	1
6	-	1	1	-	1	-	-	1
7	1	-	-	1	-	1	1	-

Construction

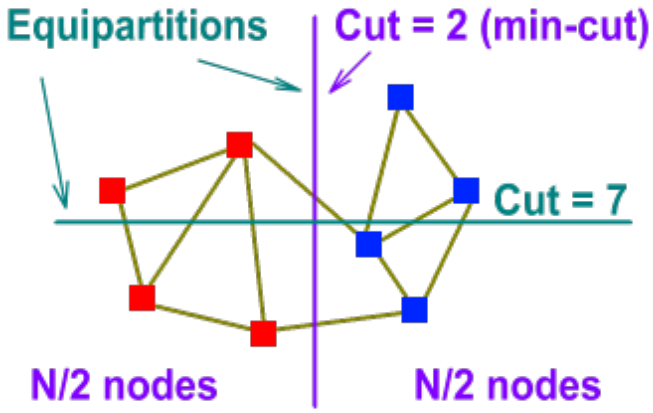
0	1	2	3	4	5	6	7	
-	1	2	-	3	-	-	4	0
1	-	-	2	-	3	4	-	1
2	-	-	1	-	4	3	-	2
-	2	1	-	4	-	-	3	3
3	-	-	4	-	1	2	-	4
-	3	4	-	1	-	-	2	5
-	4	3	-	2	-	-	1	6
4	-	-	3	-	2	1	-	7

- Nodes: $N = 2^d = 2^3 = 8$
- T. Radix: $m = d + 1 = 4$
- Bisection: $B = 2 \cdot \frac{N}{2} = 8$
- Diameter: $D = \lceil d/2 \rceil = 2$
- Avg. Hops: $A = 1.25$
- Ports/Switch: $p = 2$
- Fault Tolerance: $f = 1$

Hop List: *H*

from:	0	0	0	**0**
1.	**0**	**0**	**1**	1
2.	**0**	**1**	**0**	2
3.	**1**	**0**	**0**	4
4.	**1**	**1**	**1**	7

from:	0	1	1	**3**
1.	**0**	**1**	**0**	2
2.	**0**	**0**	**1**	1
3.	**1**	**1**	**1**	7
4.	**1**	**0**	**0**	4



Computing Bisection

OPENFABRICS ALLIANCE

Folded Cube FC_4

X	+	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-
+	-	1	1	-	1	-	-	-	1	-	-	-	-	-	-	1
+	1	-	-	1	-	1	-	-	-	1	-	-	-	-	1	-
+	1	-	-	1	-	-	1	-	-	-	1	-	-	1	-	-
+	-	1	1	-	-	-	-	1	-	-	-	1	1	-	-	-
-	1	-	-	-	-	1	1	-	-	-	-	1	1	-	-	-
-	-	1	-	-	1	-	-	1	-	-	1	-	-	1	-	-
-	-	-	1	-	1	-	-	1	-	1	-	-	-	-	1	-
-	-	-	-	1	-	1	1	-	1	-	-	-	-	-	-	1
+	1	-	-	-	-	-	-	1	-	1	1	-	1	-	-	-
+	-	1	-	-	-	-	1	-	1	-	-	1	-	1	-	-
+	-	-	1	-	-	1	-	-	1	-	-	1	-	-	1	-
+	-	-	-	1	1	-	-	-	-	1	1	-	-	-	-	1
-	-	-	-	1	1	-	-	-	1	-	-	-	-	1	1	-
-	-	-	1	-	-	1	-	-	-	1	-	-	1	-	-	1
-	-	1	-	-	1	-	-	-	-	1	-	-	1	-	-	1
-	1	-	-	-	-	-	1	-	-	-	1	-	1	-	1	-
-	1	-	-	-	-	-	1	-	-	-	1	-	-	1	1	-

$\mathbf{E}$ = { **X**: equipartitions }, $|\mathbf{E}|$= # of equipartitions

$$|\mathbf{E}| = \frac{1}{2}\binom{N}{N/2} \approx \frac{2^{N-1}}{\sqrt[2]{\pi \cdot N/2}} \qquad N{=}32 \Rightarrow 301\text{K}$$

$$\boldsymbol{Cut}(\mathbf{X}) = \frac{N}{4}\left(m - \frac{\langle \mathbf{X}|\mathbf{A}|\mathbf{X}\rangle}{\langle \mathbf{X}|\mathbf{X}\rangle}\right)$$

Rayleigh-Ritz ⇩

$$\min_{|\mathbf{X}\rangle \neq 0}\left\{\frac{\langle \mathbf{X}|\mathbf{A}|\mathbf{X}\rangle}{\langle \mathbf{X}|\mathbf{X}\rangle}\right\} = \lambda_{min}$$

Walsh Eigenbasis $\{W_r\}$

$$\boldsymbol{Cut}(W_k) = \frac{N}{2} \cdot \sum_{s=1}^{m} P(r\&h_s)$$

r = **1 0 0 1** **P**

h_1	0	0	0	1	**1**
h_2	0	0	1	0	**0**
h_3	0	1	0	0	**0**
h_4	1	0	0	0	**1**
h_5	1	1	1	1	**0**

$\boldsymbol{Cut}(W_r)/(N/2) \Rightarrow$ **2**

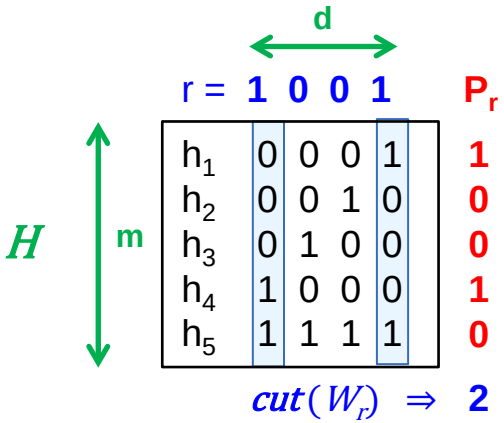
Walsh Functions: $W_r(x) = P(r\&x)$

	00	02	04	06	08	0A	0C	0E	
0:00	-	-	-	-	-	-	-	-	00
1:01	-	1	-	1	-	1	-	1	01
2:02	-	-	1	1	-	-	1	1	02
3:03	-	1	1	-	-	1	1	-	03
4:04	-	-	-	-	1	1	1	1	04
5:05	-	1	-	1	1	-	1	-	05
6:06	-	-	1	1	1	1	-	-	06
7:07	-	1	1	-	1	-	-	1	07
8:08	-	-	-	-	-	-	-	-	08
9:09	-	1	-	1	-	1	-	1	09
10:0A	-	-	1	1	-	-	1	1	0A
11:0B	-	1	1	-	-	1	1	-	0B
12:0C	-	-	-	-	1	1	1	1	0C
13:0D	-	1	-	1	1	-	1	-	0D
14:0E	-	-	1	1	1	1	-	-	0E
15:0F	-	1	1	-	1	-	-	1	0F
	00	02	04	06	08	0A	0C	0E	

Complexity: $O(Nm)$, **FWT:** $O(Nlog(N))$

Maximizing Bisection

$$b = \frac{B}{N/2} = \min_{r\in[1,N)} \sum_{s=1}^{m} P(r \& h_s)$$



$$b_{opt} = \max_{\{h_1..h_m\}} \left\{ \min_{r\in[1,N)} \sum_{s=1}^{m} P(r \& h_s) \right\}$$

$$|\{h_1..h_m\}| = \binom{N-1}{m} \sim O(N^m)$$

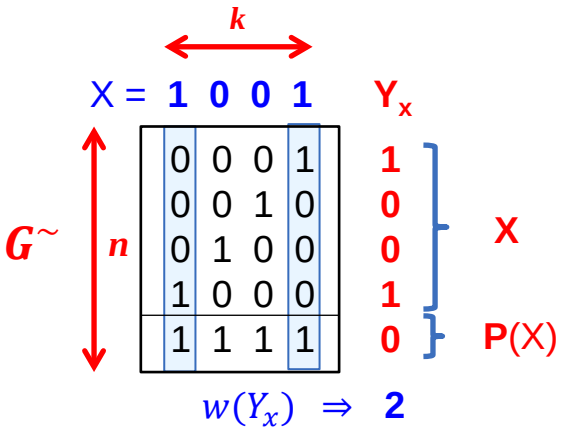
Translation Recipe

H	$G^{\sim}$
d	k
m	n
b	Δ, w_{min}
r	X
P_r	Y_x

Δ_{opt}

Long Hop Networks

Error Correcting Code [n,k,Δ]

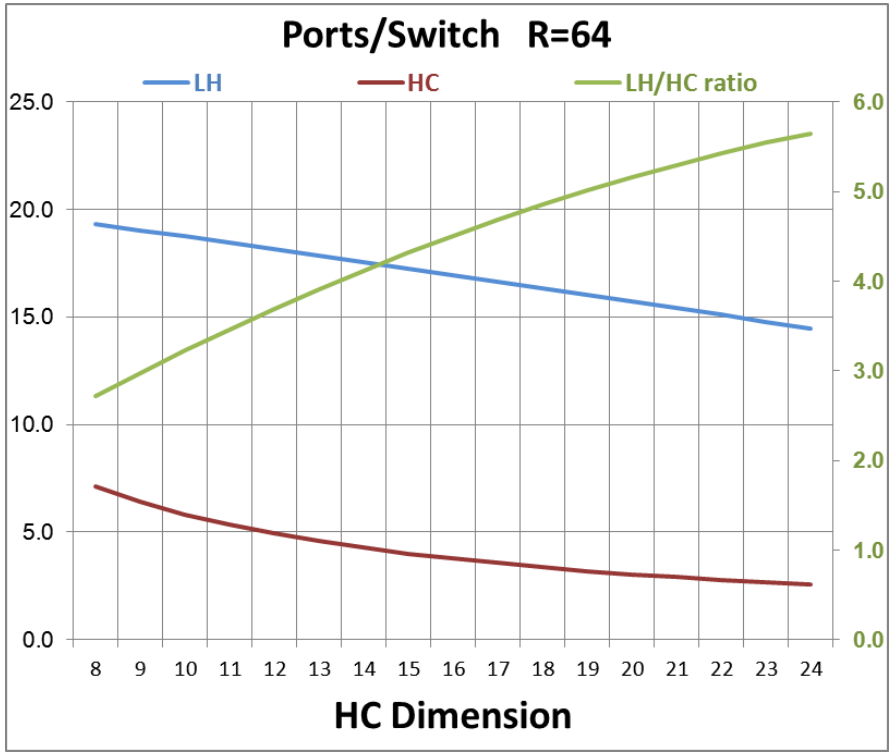


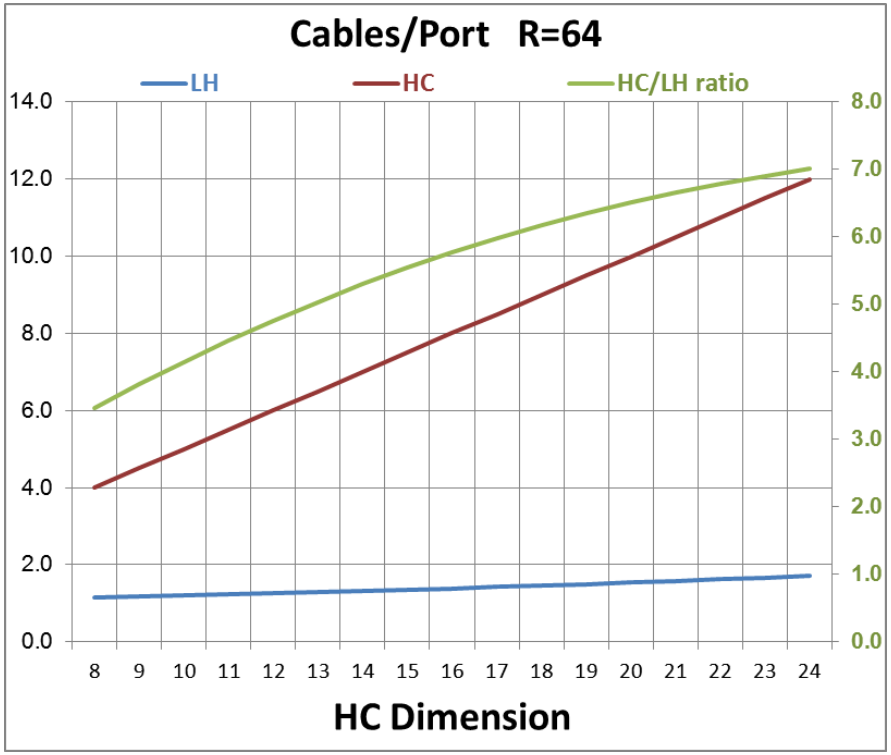
$$\Delta = w_{min} = \min_{x\in[1,N)} \{w(Y_x)\}$$

$$\Delta_{opt} = w_{opt} = \max_{\{G\}} \left\{ \min_{x\in[1,N)} \{w(Y_x)\} \right\}$$

Trunking	Repetition code
Multipartite graphs	Reed-Muller code
Fully connected	Hadamard code

Effects of LH Optimization





- LH Solutions Database (ECC tables + LH solver)
- 3364 configs, N ≤ 10^6 switches, m ≤ 256 ports
- Max size: over 117·10^6 non-oversubscribed ports

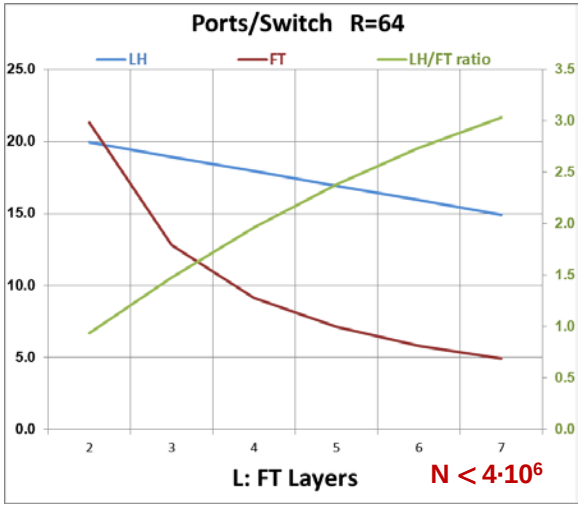
- Max fault tolerance f=b-1 faulty links or dim
- Max number of independent d-cube subgraphs
- Fast ~N·log(N), exact graph partition via Walsh functions

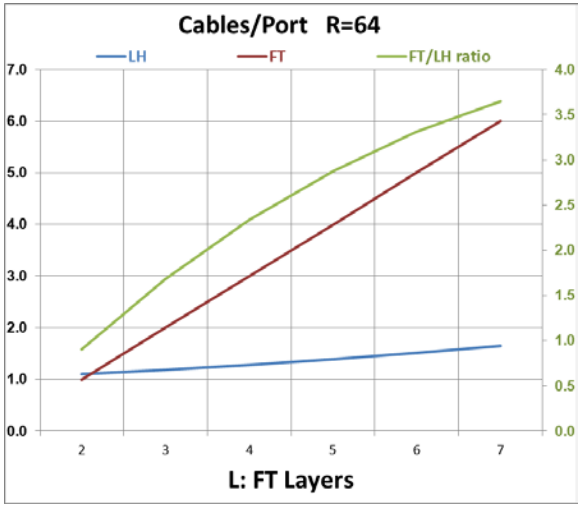
LH Solutions Data Base

dim=5, N=32 switches

Hops	Radix	#Ports	<ECC>	%	LIN	Bisection Optimal ST	MinC	MaxC	Wmin	Wmax	Diam	AvgDist	Distance Optimal ST	MinC	MaxC	Wmin	Wmax	Diam	AvgDist
5.	6	32	1-1	0	1	84	1	5	00001	0001F	5	2.5000000	84	1	5	00001	0001F	5	2.5000000
6.	8	64	2-2	0	1	84	2	6	00001	0001F	3	2.0625000	84	2	6	00001	0001F	3	2.0625000
7.	9	64	2-2	0	2	84	2	6	00001	0001F	3	1.9062500	84	2	6	00001	0001F	3	1.9062500
8.	10	64	2-2	0	2	84	2	6	0000C	0000F	3	1.7812500	84	2	6	0000C	0000F	3	1.7812500
9.	12	96	3-3	0	3	84	**3**	7	00001	0000E	3	1.6875000	84	**2**	6	0000C	0000D	2	1.6562500
10.	14	128	4-4	0	3	84	4	8	00001	0000E	2	1.6250000	84	4	8	00001	0000E	2	1.6250000
11.	15	128	4-4	0	4	84	4	9	00002	0001B	2	1.5937500	84	4	9	00002	0001B	2	1.5937500
12.	16	128	4-4	0	4	84	4	9	00002	0001B	2	1.5625000	84	4	9	00002	0001B	2	1.5625000
13.	18	160	5-5	0	5	84	5	10	00001	0001B	2	1.5312500	84	5	10	00001	0001B	2	1.5312500
14.	20	192	6-6	0	5	84	6	10	00001	0000D	2	1.5000000	84	6	10	00001	0000D	2	1.5000000
15.	22	224	7-7	0	6	84	**7**	15	00001	0001F	3	1.5000000	84	**6**	12	00001	00018	2	1.4687500
16.	24	256	8-8	0	6	84	8	16	00001	0001F	2	1.4375000	84	8	16	00001	0001F	2	1.4375000
17.	25	256	8-8	0	7	84	8	16	00001	0001F	2	1.4062500	84	8	16	00001	0001F	2	1.4062500
18.	26	256	8-8	0	7	84	8	16	00007	0001F	2	1.3750000	84	8	16	00007	0001F	2	1.3750000
19.	27	256	8-8	0	8	84	8	16	00007	0001F	2	1.3437500	84	8	16	00007	0001F	2	1.3437500
20.	29	288	9-9	0	8	84	9	16	00007	0001F	2	1.3125000	84	9	16	00007	0001F	2	1.3125000
21.	31	320	10-10	0	9	84	10	16	00003	0001F	2	1.2812500	84	10	16	00003	0001F	2	1.2812500
22.	32	320	10-10	0	9	84	10	16	00003	0001F	2	1.2500000	84	10	16	00003	0001F	2	1.2500000
23.	34	352	11-11	0	10	84	11	16	00001	0001F	2	1.2187500	84	11	16	00001	0001F	2	1.2187500
24.	36	384	12-12	0	10	84	12	16	00001	0000F	2	1.1875000	84	12	16	00001	0000F	2	1.1875000
25.	37	384	12-12	0	11	84	12	16	00003	0000F	2	1.1562500	84	12	16	00003	0000F	2	1.1562500
26.	38	384	12-12	0	11	84	12	16	00003	0000F	2	1.1250000	84	12	16	00003	0000F	2	1.1250000
27.	40	416	13-13	0	12	84	13	16	00001	0000F	2	1.0937500	84	13	16	00001	0000F	2	1.0937500
28.	42	448	14-14	0	12	84	14	16	00001	00007	2	1.0625000	84	14	16	00001	00007	2	1.0625000
29.	43	448	14-14	0	13	84	14	16	00001	00007	2	1.0312500	84	14	16	00001	00007	2	1.0312500
30.	45	480	15-15	0	13	84	15	16	00001	00003	2	1.0000000	84	15	16	00001	00003	2	1.0000000
31.	47	512	16-16	0	14	84	16	16	00001	00001	1	0.9687500	84	16	16	00001	00001	1	0.9687500

Long Hop vs. Fat Tree





Fat Tree Overload on Random Traffic

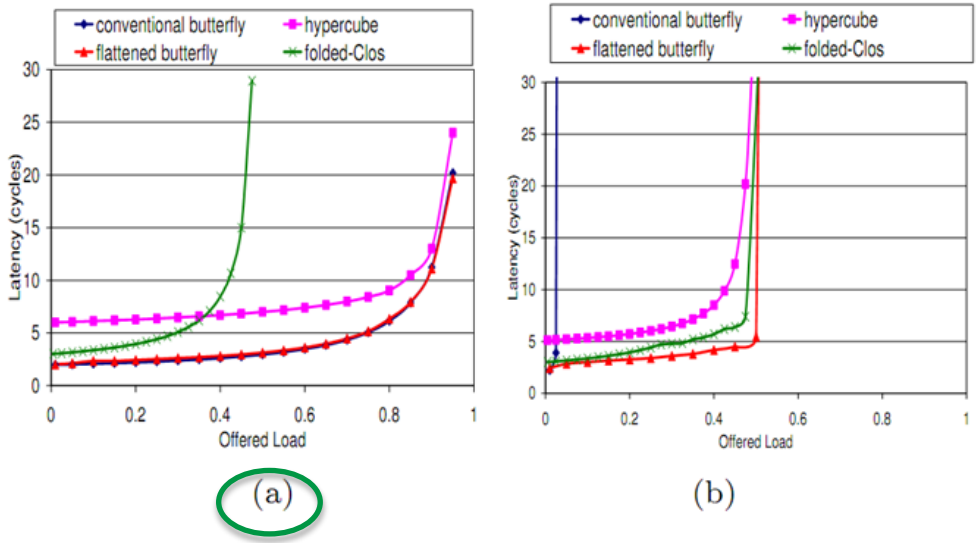
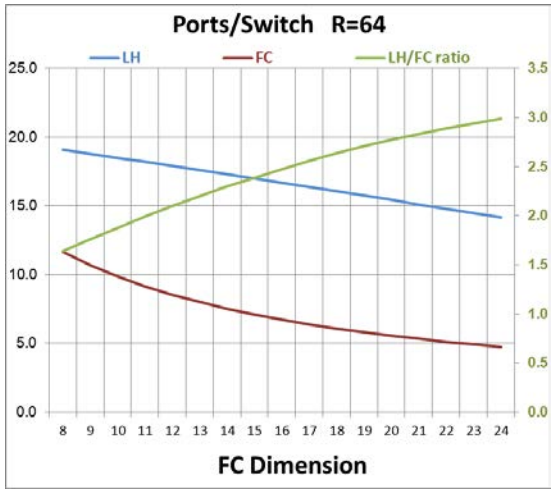


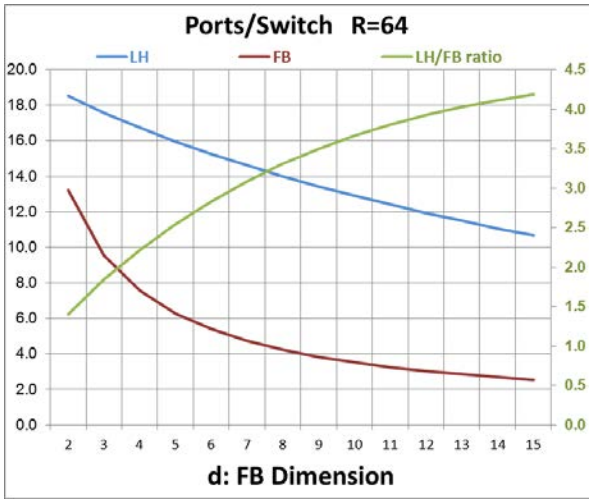
Figure 6: Topology comparisons of the flattened butterfly and the folded Clos with (a) uniform random traffic and (b) worst case traffic pattern.

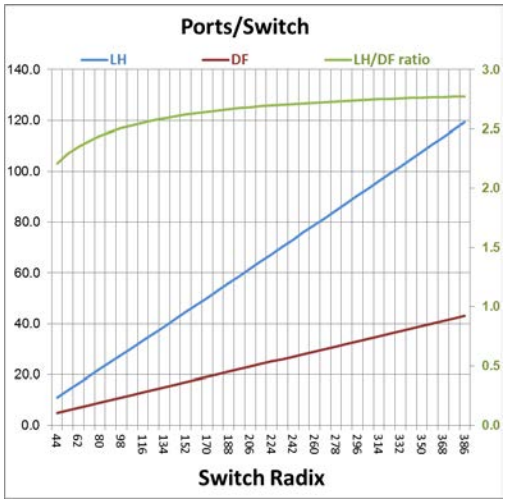
Flattened butterfly: a cost-efficient topology for high-radix networks
J. Kim, W. J. Dally, D. Abts (Stanford, Google)
Proc. ISCA`07, May 2007, pp. 126-137

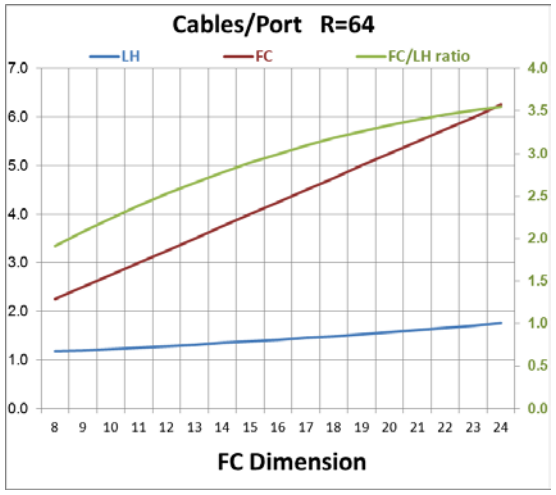
High-Radix Interconnection Networks
J. Kim, PhD thesis, Stanford University, 2008.

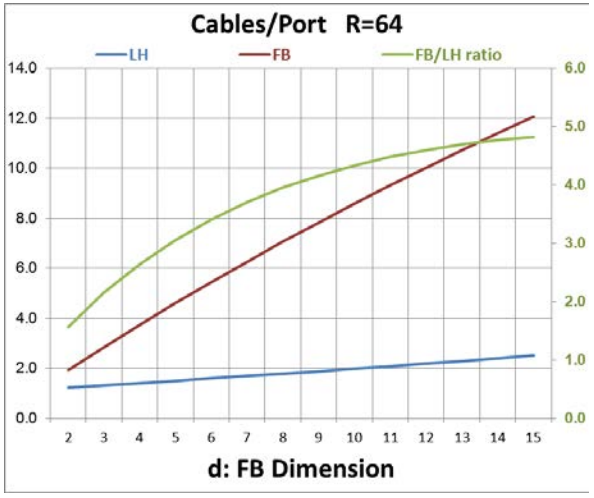
LH vs. FC, FB, Dragonfly

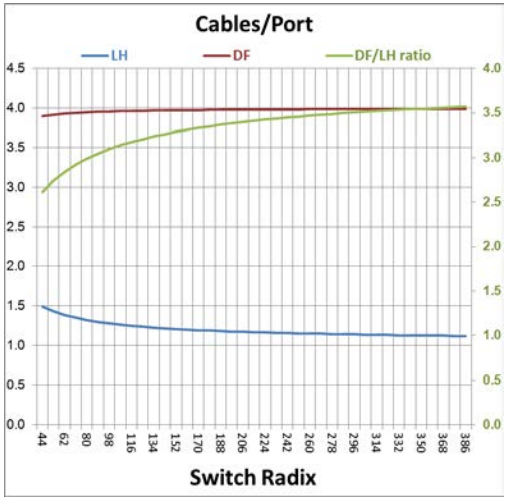












TCALC Comparisons

NETWORKS COMPARED WITH THE LONG HOP (LH) NETWORK

Fat Tree (FT): FT Levels L=4 Q=2.519842 (trunking factor)
Flattened Butterfly (FB): FB(k:17.428, n:4.365, c:8.714)
Dragonfly (DF): DF(p:7.24, a:37.85, h:12.62, g:478.44) Q=1.148
Folded Hypercube (FC): FC(dimension 14.096) Q=3.744
Hypercube (HC): HC(dimension 15.000) Q=4.000

TARGET: Ports P=131072, Switch radix R=64, Oversubscription OVS=1

##	Switches	Ports/Sw.	Cost/Pt.	Cost Gb/s	Cables/Pt.	Cabling	Max	Avg Hops	Latency
LH	8192	16.000	100	100	1.500	100	4	2.915039	100
FT	14336	9.143	175	358	3.000	200	6	5.968750	205
FB	15042	8.714	184	238	3.172	211	4	3.777778	130
DF	18107	7.239	221	221	3.921	261	3	2.916464	100
FC	17506	7.487	214	447	3.774	252	8	6.100012	209
HC	32768	4.000	400	1029	7.500	500	15	7.500000	257

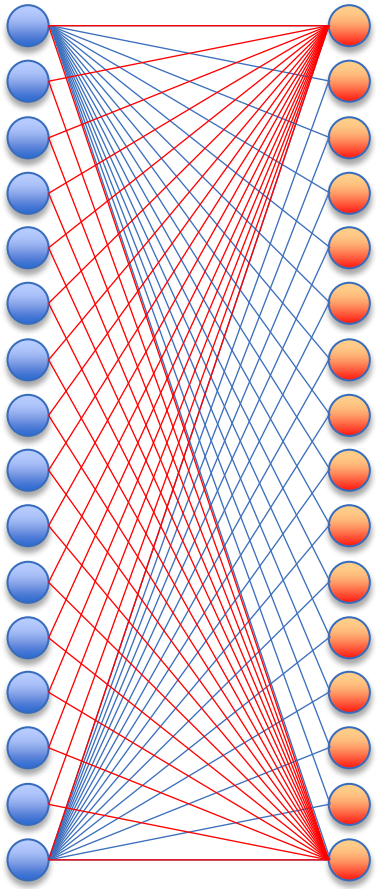
- Includes ~20 sample Long Hop networks
- Compares LH to 5 alternative topologies
- Obtain **P** non-oversubscribed ports via radix **R** switches
- Cost Gb/s = (Cost/pt.) × (Latency) / 100

Wiring Patterns

Low Density

PORT	1	2	3	4	5	6
SWITCH						
0	1	2	4	8	16	31
1	0	3	5	9	17	30
2	3	0	6	10	18	29
3	2	1	7	11	19	28
4	5	6	0	12	20	27
5	4	7	1	13	21	26
6	7	4	2	14	22	25
7	6	5	3	15	23	24
8	9	10	12	0	24	23
9	8	11	13	1	25	22
10	11	8	14	2	26	21
11	10	9	15	3	27	20
12	13	14	8	4	28	19
13	12	15	9	5	29	18
14	15	12	10	6	30	17
15	14	13	11	7	31	16
16	17	18	20	24	0	15
17	16	19	21	25	1	14
18	19	16	22	26	2	13
19	18	17	23	27	3	12
20	21	22	16	28	4	11
21	20	23	17	29	5	10
22	23	20	18	30	6	9
23	22	21	19	31	7	8
24	25	26	28	16	8	7
25	24	27	29	17	9	6
26	27	24	30	18	10	5
27	26	25	31	19	11	4
28	29	30	24	20	12	3
29	28	31	25	21	13	2
30	31	28	26	22	14	1
31	30	29	27	23	15	0

High Density



Each node connects to all nodes of the other color
Density is the same as 2 level Fat Tree

Functional Architecture

Hypercube: X → Y
X^Y = 0 1 0 1 1 ⇨ L=3 #Paths = L! = 6

Long Hop: d=5 m=9	
1. 10000 2. 01000 3. 00100 4. 00010 5. 00001 6. 00101 7. 10011 8. 01111 9. 11111	b = 3, D = 3 X^Y = 0x13 Distance L=3 #Path Sets: 4 #Paths=4*3!=24

hop	1	2	3	4	5	6	7	8	9	
1.	1	1	.	.	1	.	.	.	.	6
2.	.	1	.	1	.	.	1	.	.	6
3.	.	.	.	.	.	1	1	1	.	6
4.	.	.	1	1	.	.	.	.	1	6

Static Forwarding Tables

- Q aliases per Y node, Dst=(**PS**:Y) → port, path
- Path Selector: VID, **TMA**
- Paths are ordered by length, edge disjoint
- L2 sw. → sw. (2 level TCAM used for large N)
- L3 egress hop to server

Topological MAC Address: **TMA**

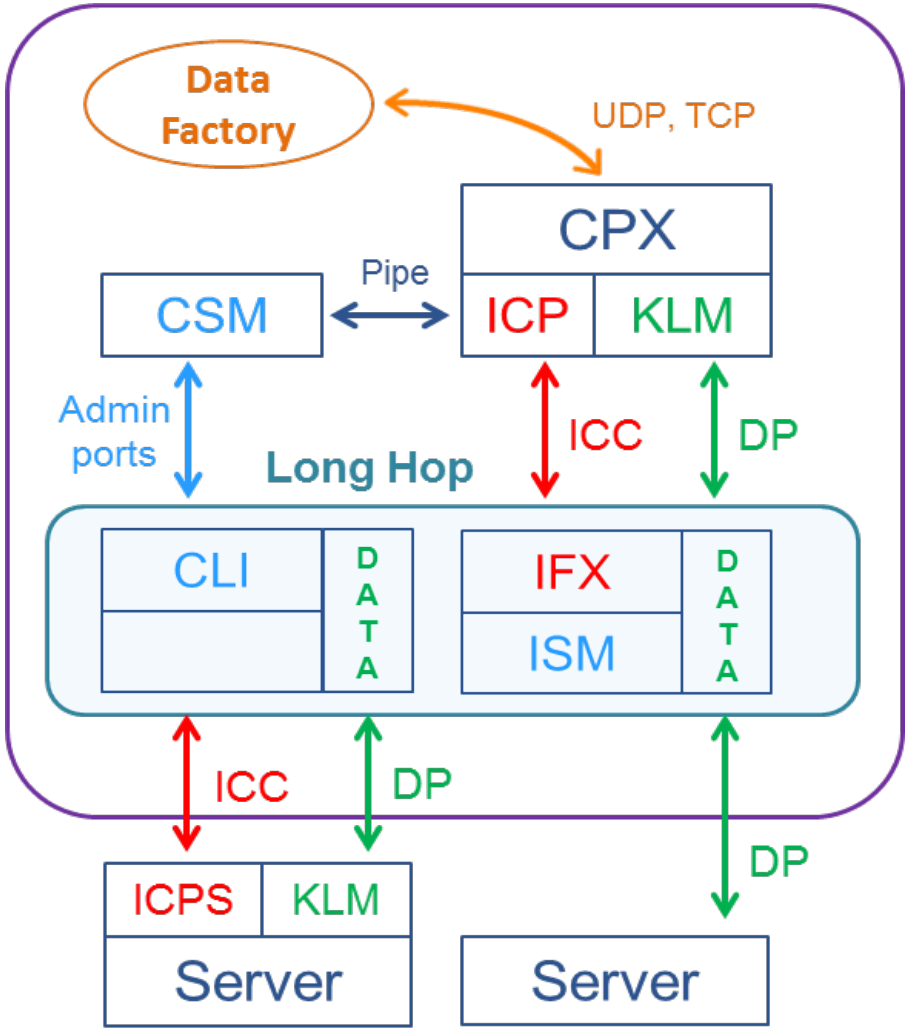
Cluster #	Node Index

Server/Hypervisor Stack, KLM

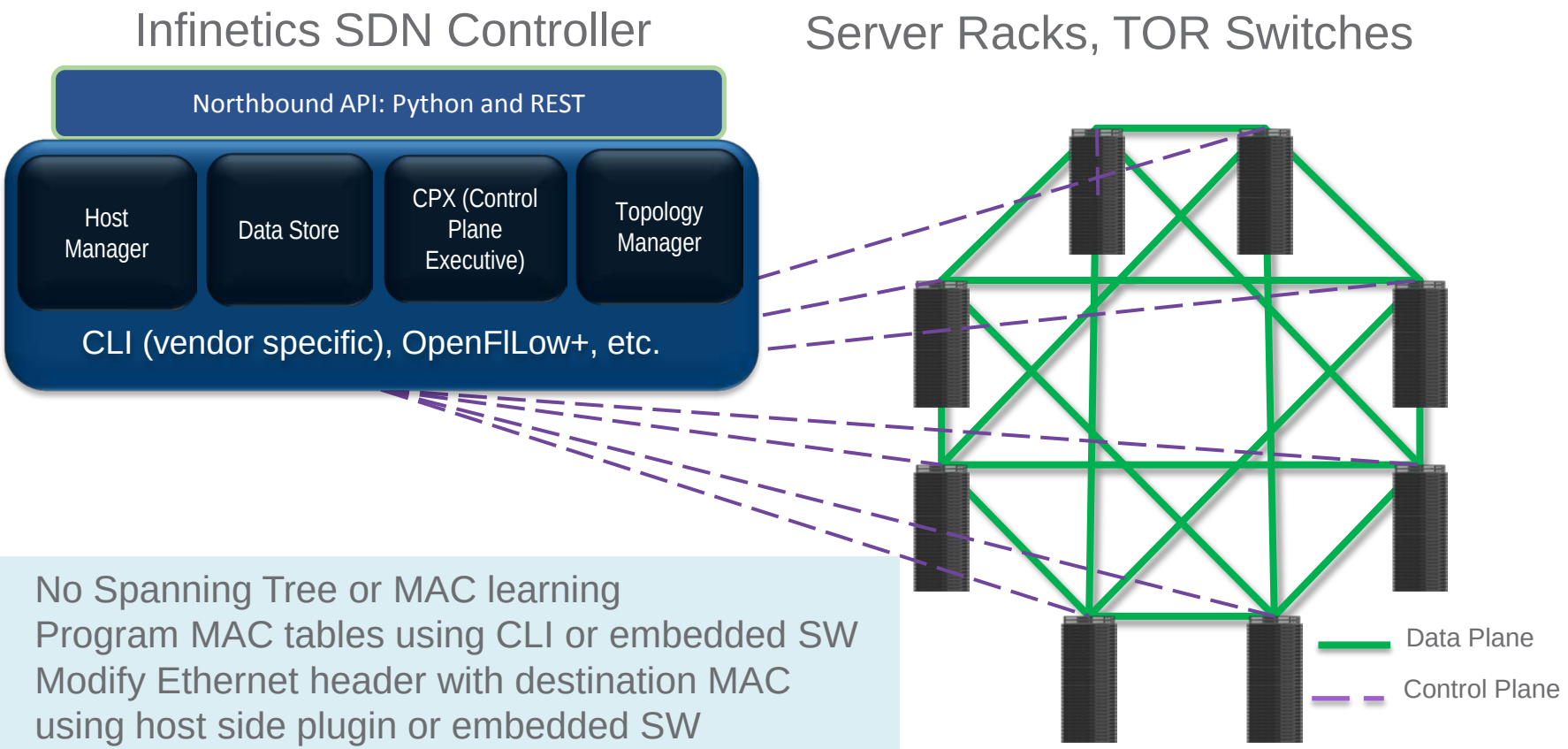
- Server ARP disabled
- KLM inserts L2 headers only for known dest. IPs
- L3 flows (LB, QOS, FW)
- Paths assigned per flow

Flexible Radix Switch

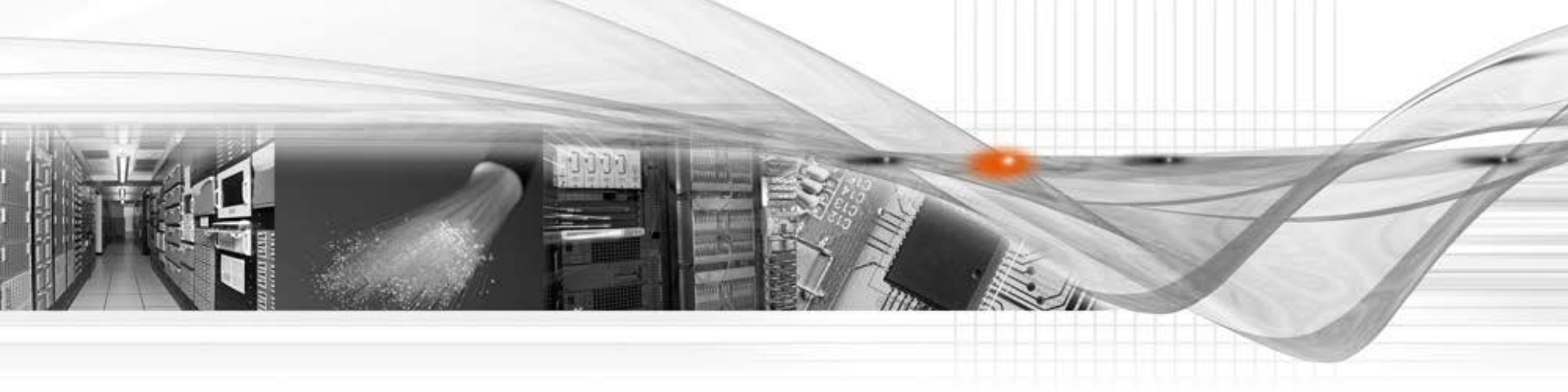
- Flat Layer 2 Network
- Wire-speed
- Scales to 10^7



Management & Control Plane



- No Spanning Tree or MAC learning
- Program MAC tables using CLI or embedded SW
- Modify Ethernet header with destination MAC using host side plugin or embedded SW

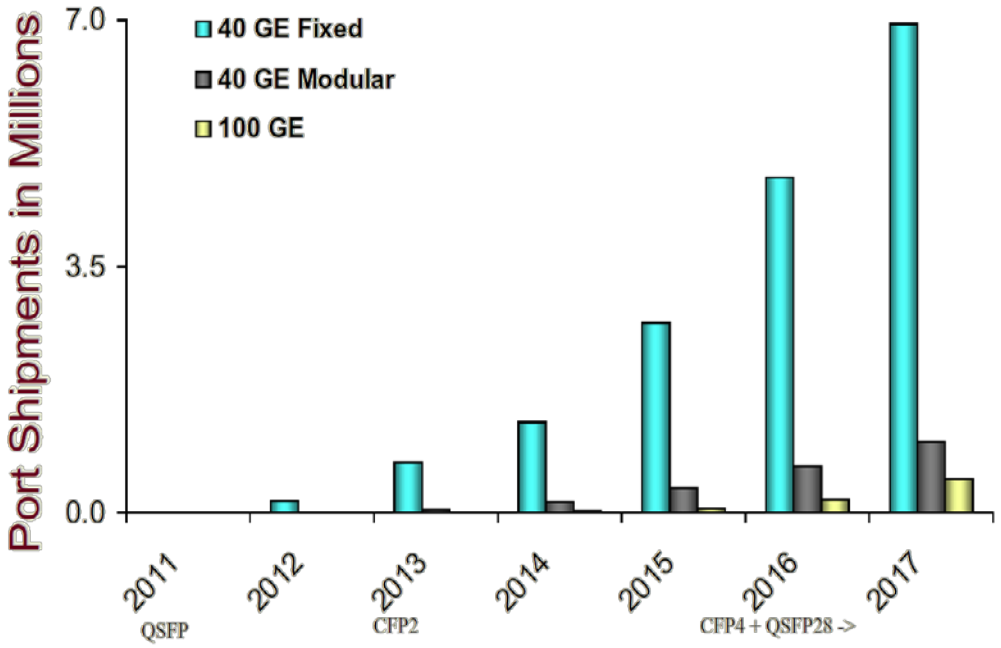


Thank You!

Fixed Config Ethernet Ports

Key Points

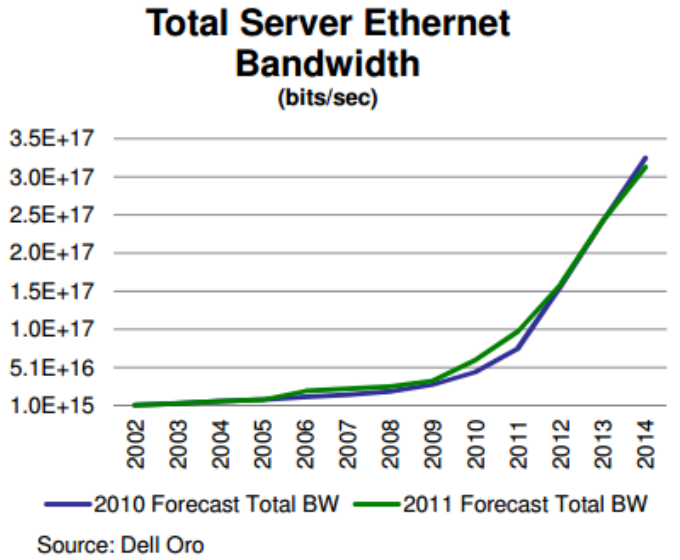
- Fixed configuration switches dominate
- Copper cable interconnects still prevalent
  - => Mesh topologies
- 2016 -100GbE ~= 2012 40GbE



Source: Dell'Oro Group 2013

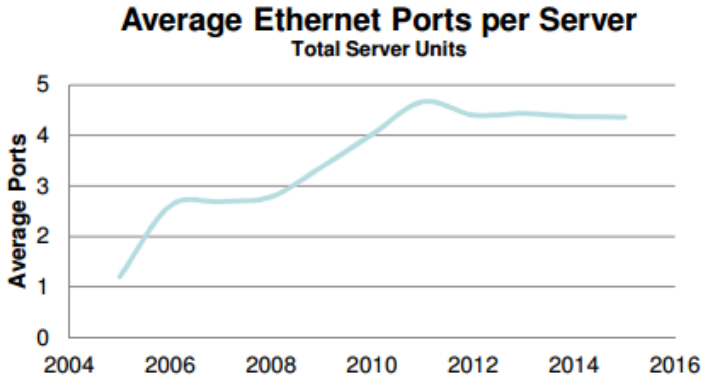
Average Server: 4+ Ethernet Ports

- Total Server BW shipped increased from 2010 forecast
  - 10GbE grew, but
  - 1000BASE-T shipments far exceeded earlier forecast



Source: Dell Oro

- Average Ports/Server grew to >4 in 2010 and expected to maintain



Causal Web of DC Problems