A Web Company's View on Ethernet

IEEE 802.3 Plenary March 2007

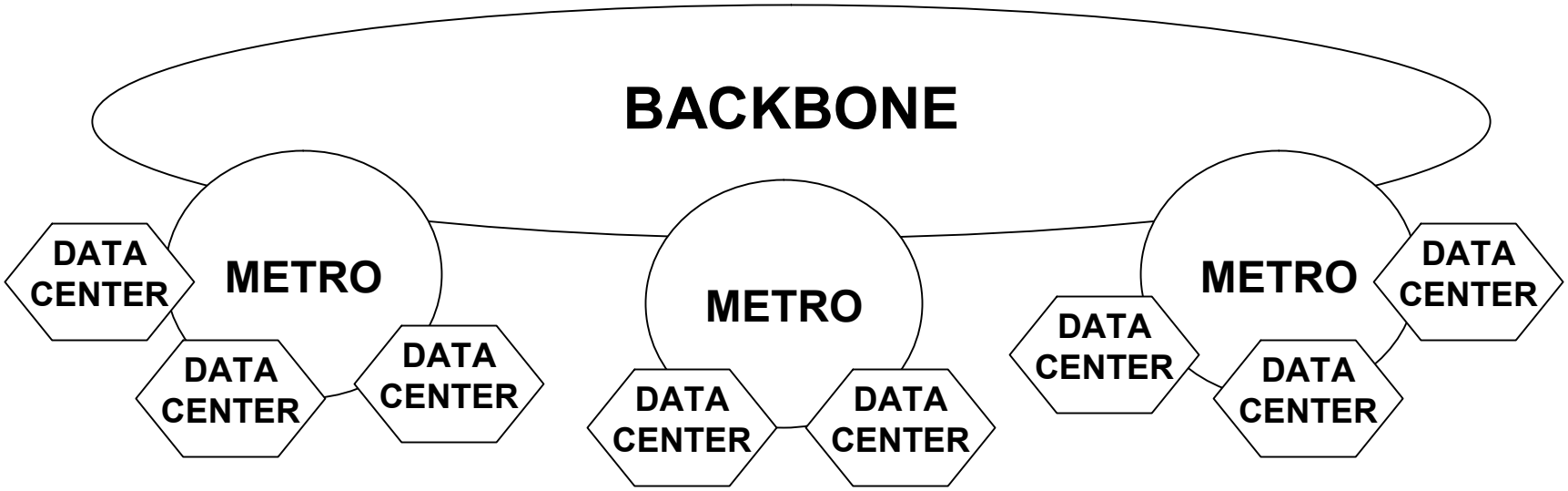
Adam Bechtel – Yahoo! Chief Architect

Trends from a Web Company

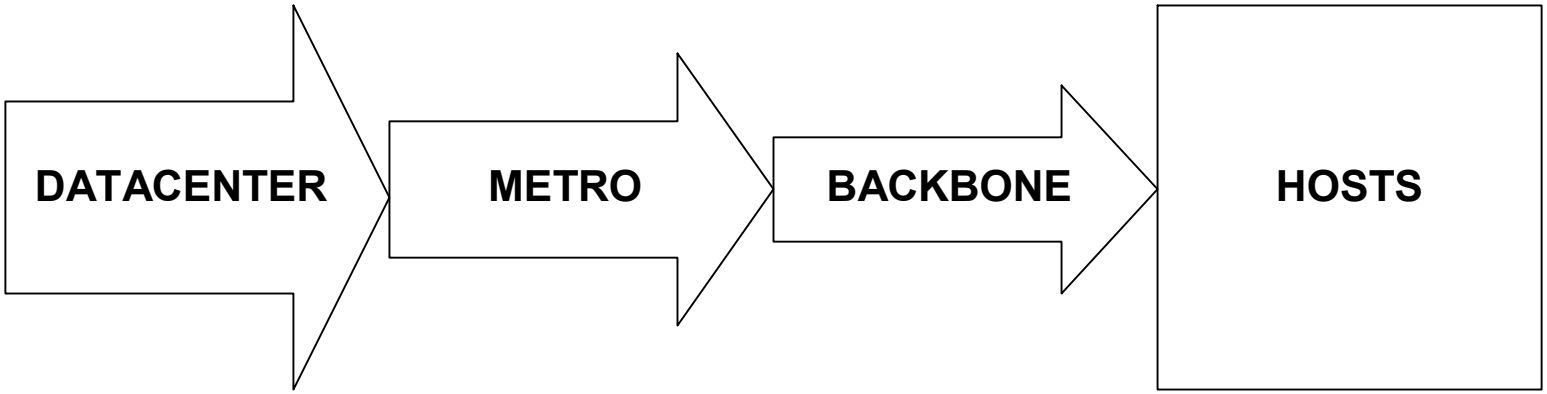
BW doubles <12 months
Many cheap servers vs few expensive servers
Prefer Ethernet vs specialty fabrics for HPC
Prefer NAS vs SAN
Encapsulate SAN into Ethernet
Metro's are Ethernet
Internet Exchanges are Ethernet
Long Haul Ethernet options

Ethernet is good enough because it is cheaper!

Web Company Network Architecture



Ethernet Adoption



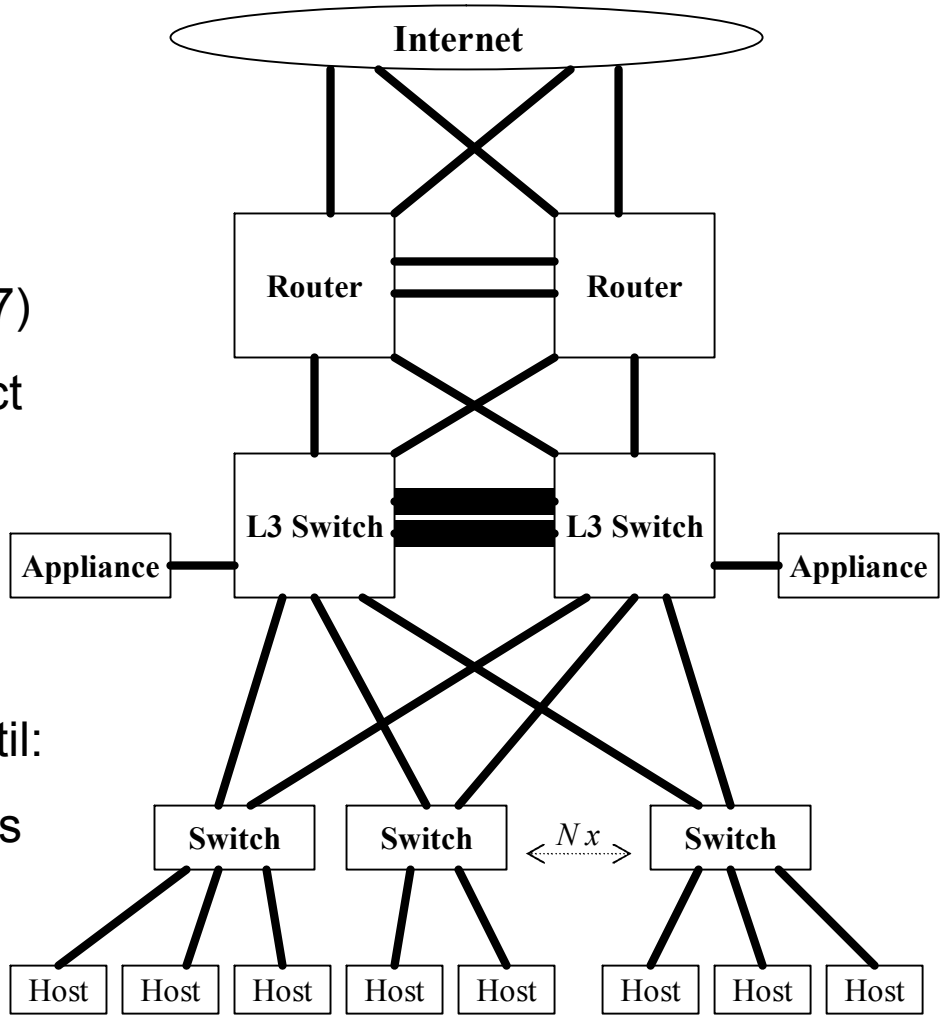
GE Cycle took 4 yrs for first ports, 7 yrs for massive penetration

We’re in year 4 of the 10GE cycle, host ports are still a ways out

Ethernet Evolution in a Typical Datacenter

10GE Evolution

1. Core device interconnects
2. Aggregation switch uplinks
3. Appliance connections (2007)
4. New core switch interconnect
5. Host NIC



Hosts will not move to 10GE until:

- Core density 10GE increases
- 10GBaseT products ship

10GE Adoption

** Percentage of ports deployed in Datacenters will continue to rise **

	2004(1)	2005(2)	2006(3)	2007(3)
Datacenter	80%	65%	59%	72%
Metro	20%	29%	33%	24%
Backbone	0%	6%	8%	4%
Hosts	0%	0%	0%	0%

(port order of magnitude)

Ethernet Composition 2007

** Don't forget the datacenter, that's where most of the ports end up **

	E(3)	FE(4)	GE(5)	10GE(3)
Datacenter	0%	1.6%	5%	72%
Metro	0%	0%	0%	24%
Backbone	0%	0.03%	0.5%	4%
Hosts	100%	98.4%	95%	0%

(port order of magnitude)

Challenges Scaling Beyond 10GE

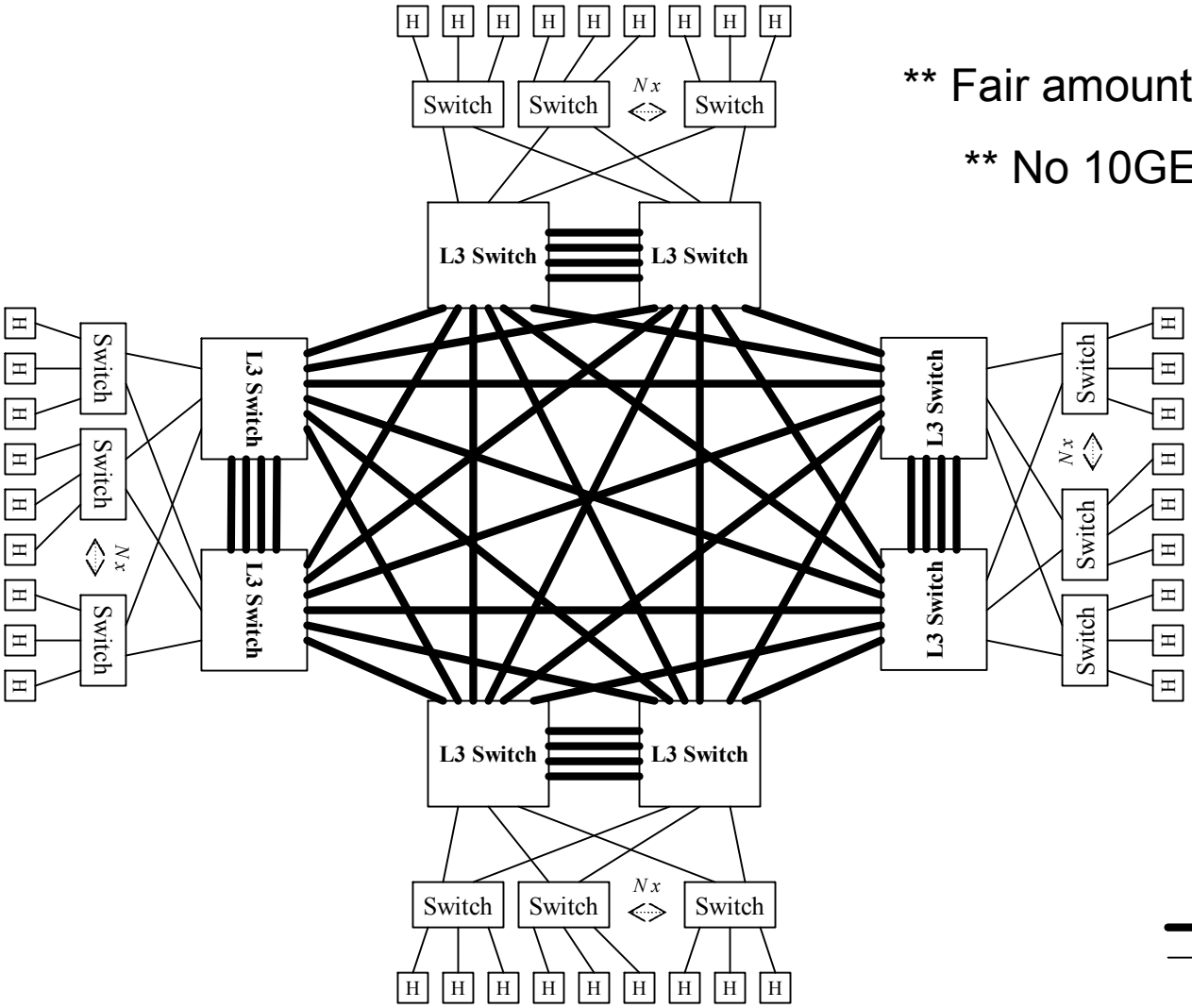
Why not LAG?
LAG is good, but…

- Large flow problem, difficult to capacity plan
- Unpredictable link removal and insertion
- LAG's fundamentally create a loop in layer2 networks
- Power of 2 hash problem (aim to stop at 4 links before upgrading speed)
- "Special" traffic usually traverses a single link (multicast, broadcast, control traffic, etc)

Why not ECMP?
ECMP is good, but…

- Large flow problem again, difficult to capacity plan
- FIB depletion as number of paths increase
- Better than LAG, but only works for layer3 environments
- Worst combination is to ECMP LAG's

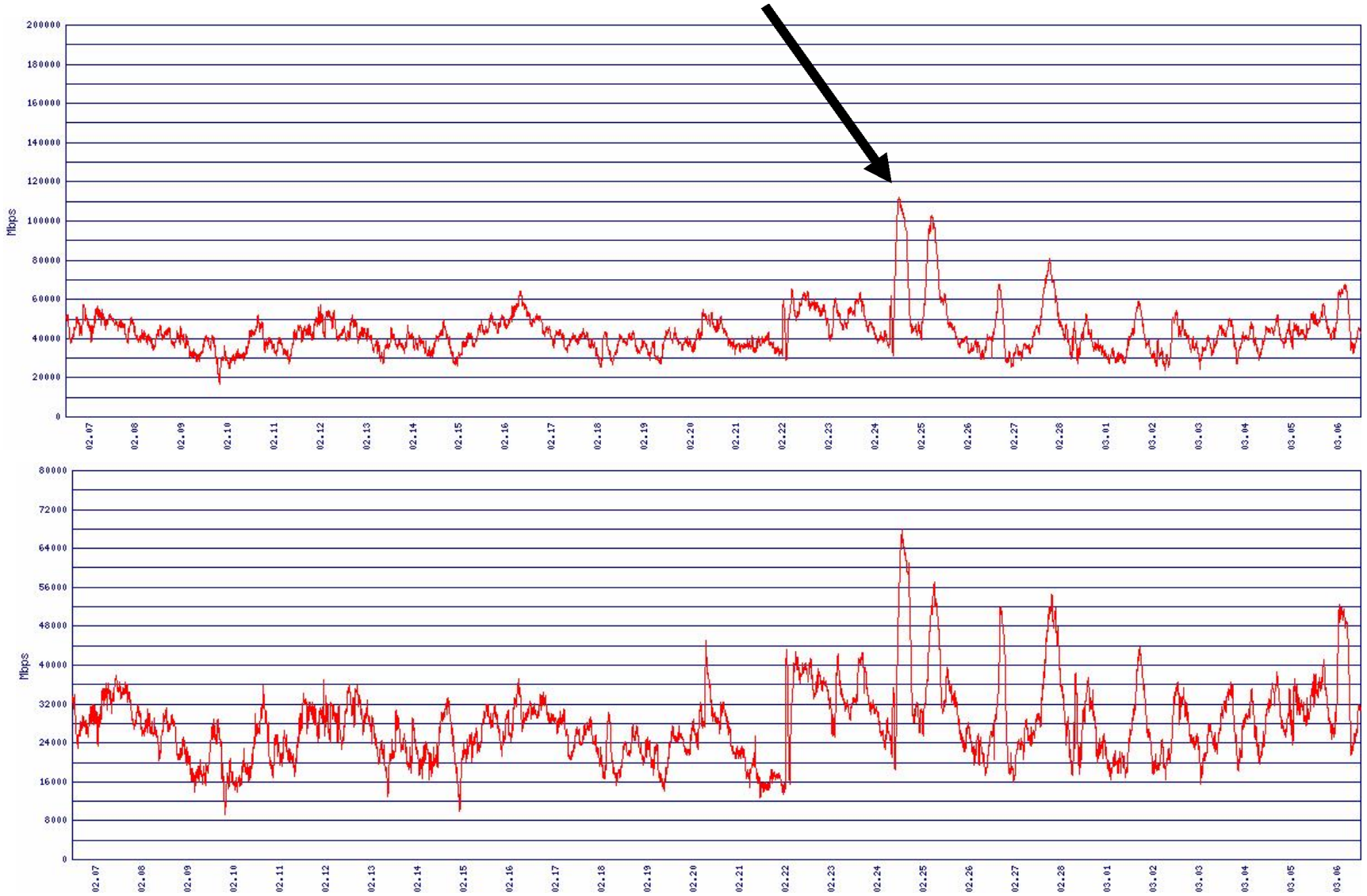
I have clusters built like this…



** Fair amount of ECMP **

** No 10GE LAGs **

…with interconnect utilization over 100Gbps

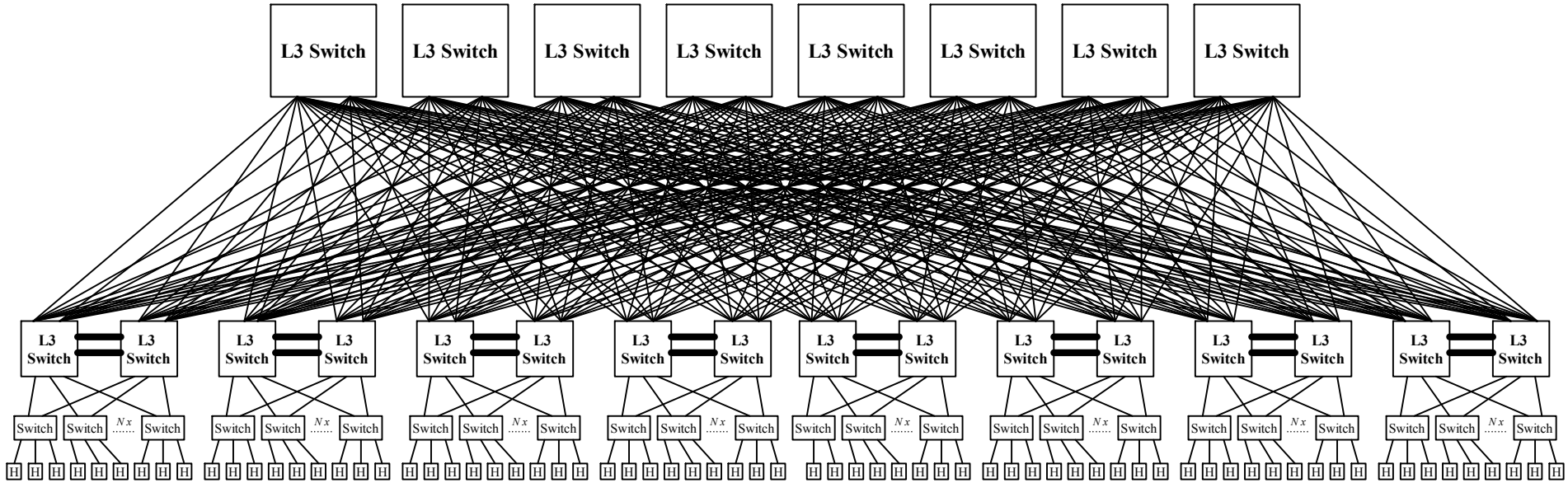


I’m currently building this…

** 8 way ECMP w/ 2x10GE LAGs**

** Way too many paths **

** Way too many cables **



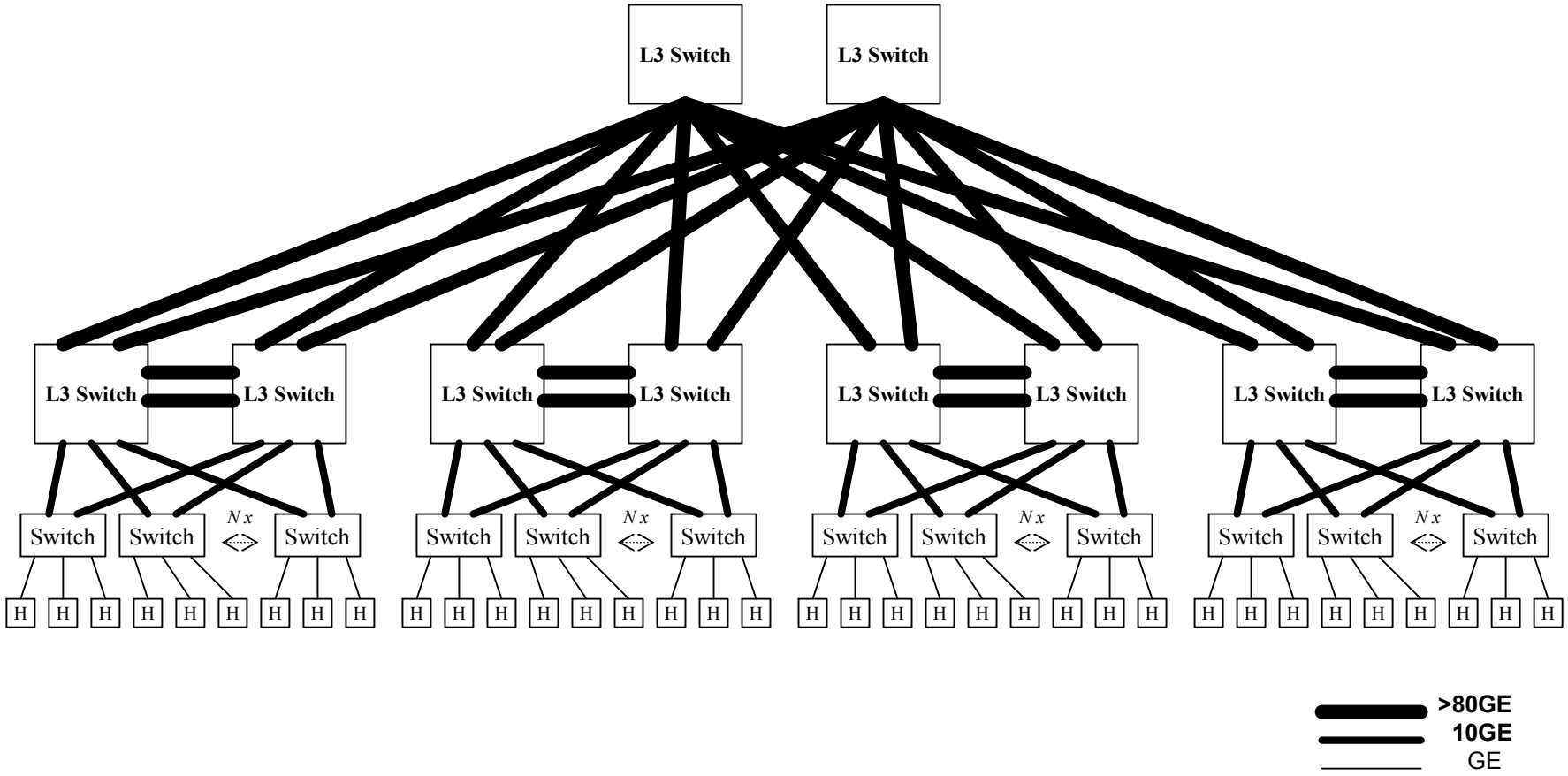
L3 Switch <10GE> L3 Switch
L3 Switch <GE> Switch
Host <GE> Switch

But what I'd like to build is this…

** Real need for at least 80GE today (ideally more) **

** Better mix of link speeds **

** Keep ECMP to 2 paths **



Supporters

Bill Trubey	TimeWarner Cable
Doug Wilson	Microsoft
Henk Steenman	AMS-IX
Jay Moran	AOL
Mark Kortekaas, Mark	CBS
Mike Bennett	Lawrence Berkeley National Laboratory
Peter Harrison	Netflix
Peter Schoenmaker	NTT America
Ted Seely	Sprint
Troy Sprenger	EDS
Vik Saxena	Comcast
Vish Yelsangikar	Netflix

Thanks