A Web Company's View on Ethernet: **The need for 100+GbE**

NANOG 40

Igor Gashinsky

Yahoo! Principal 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

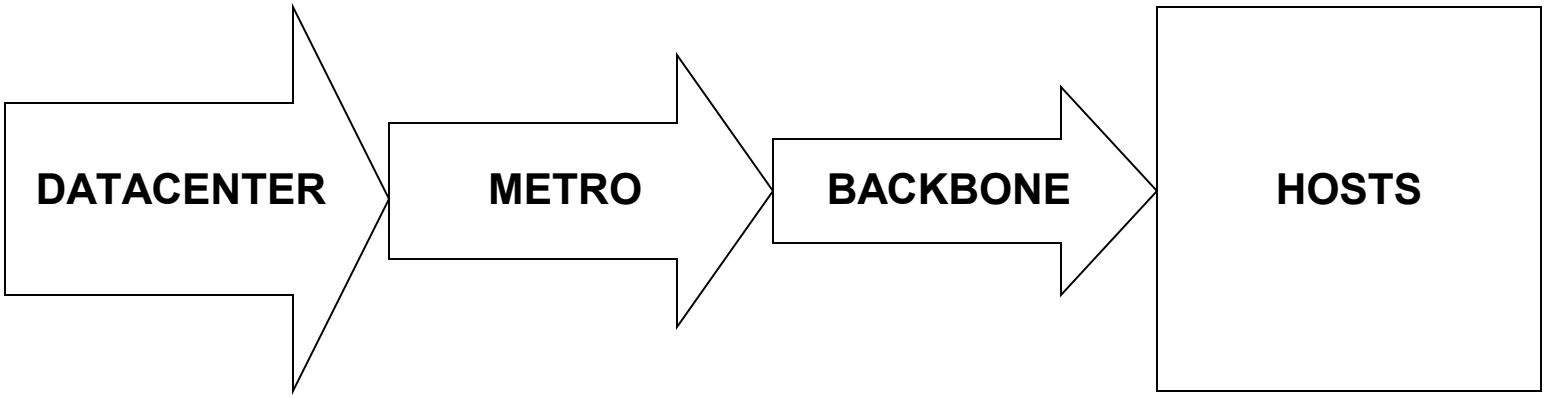
Metro's are Ethernet

Internet Exchanges are Ethernet

Long Haul Ethernet options

Ethernet is good enough because it is cheaper!

Ethernet Adoption



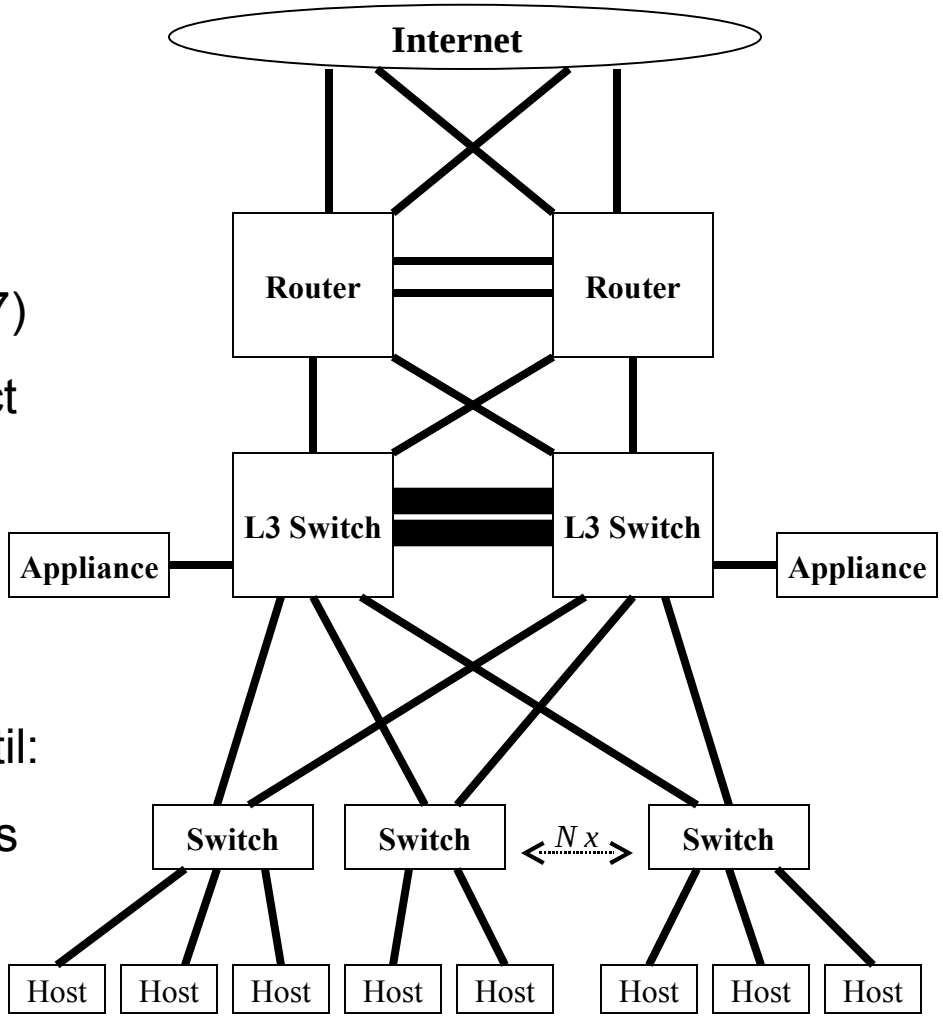
GE Cycle took 4 yrs for first ports, 7 yrs for massive penetration

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

Ethernet Evolution in a Typical Datacenter

10GE Evolution

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



Hosts will not move to 10GE until:

- Core density 10GE increases
- 10GBaseT products ship

Challenges Scaling Beyond 10GE

Why not LAG?

LAG is good, but…

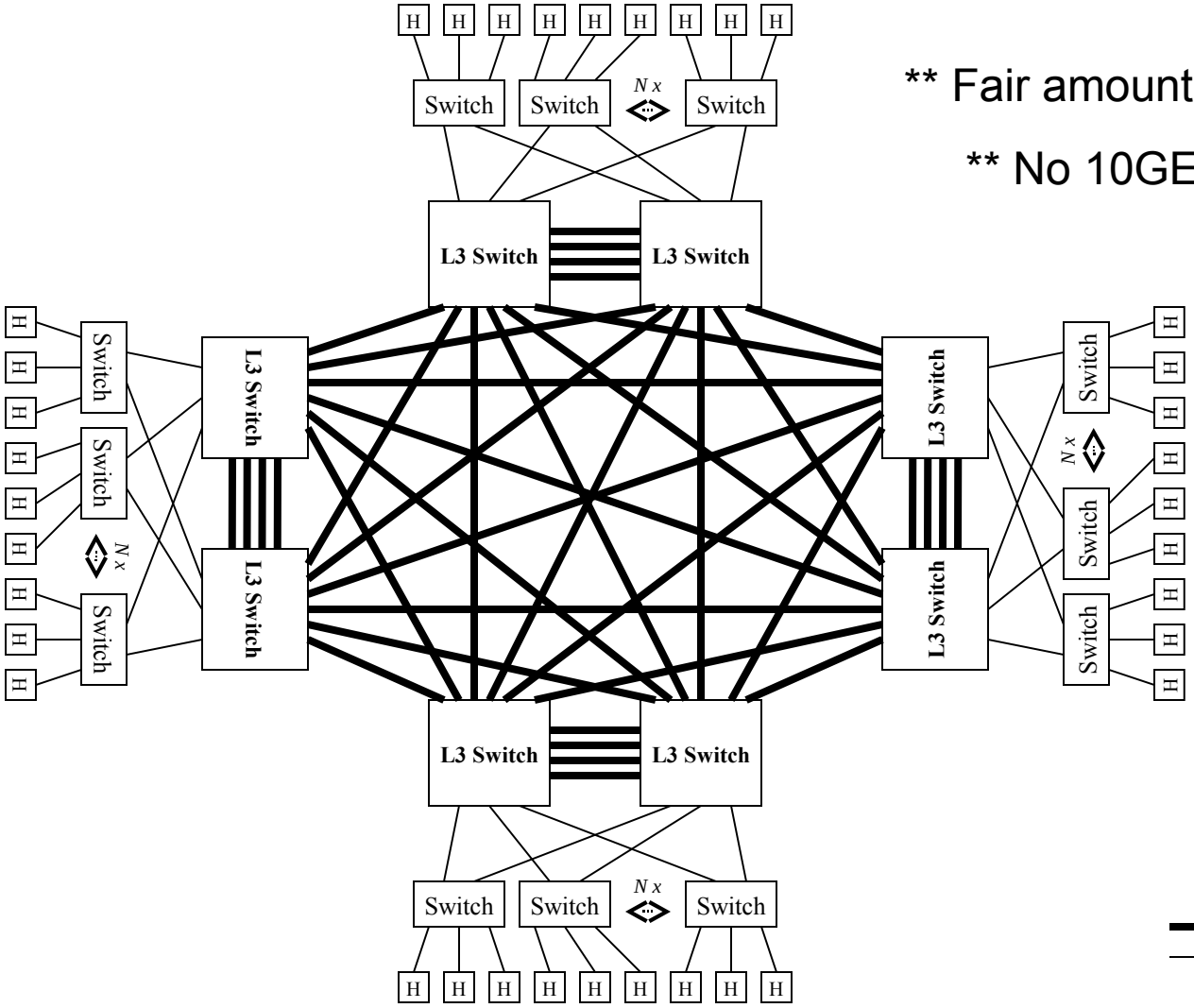
- Large flow problem, difficult to capacity plan
- Unpredictable link removal and insertion
- LAG's fundamentally create a loop in layer2 networks
- Power of 2 hash problem (aim to stop at 4 links before upgrading speed)
- "Special" traffic usually traverses a single link (multicast, broadcast, control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

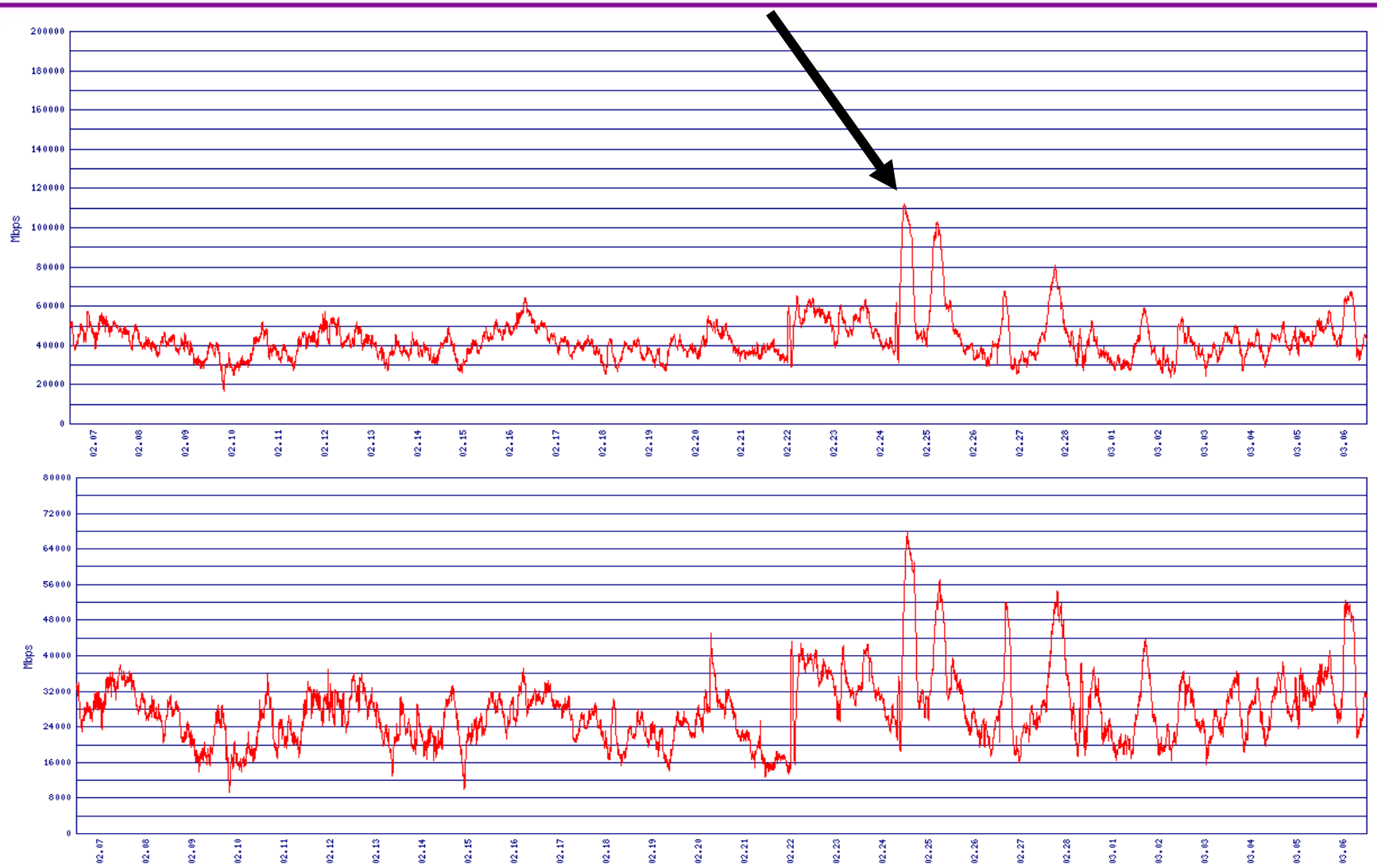
We have clusters built like this…



** Fair amount of ECMP **

** No 10GE LAGs **

…with interconnect utilization over 100Gbps

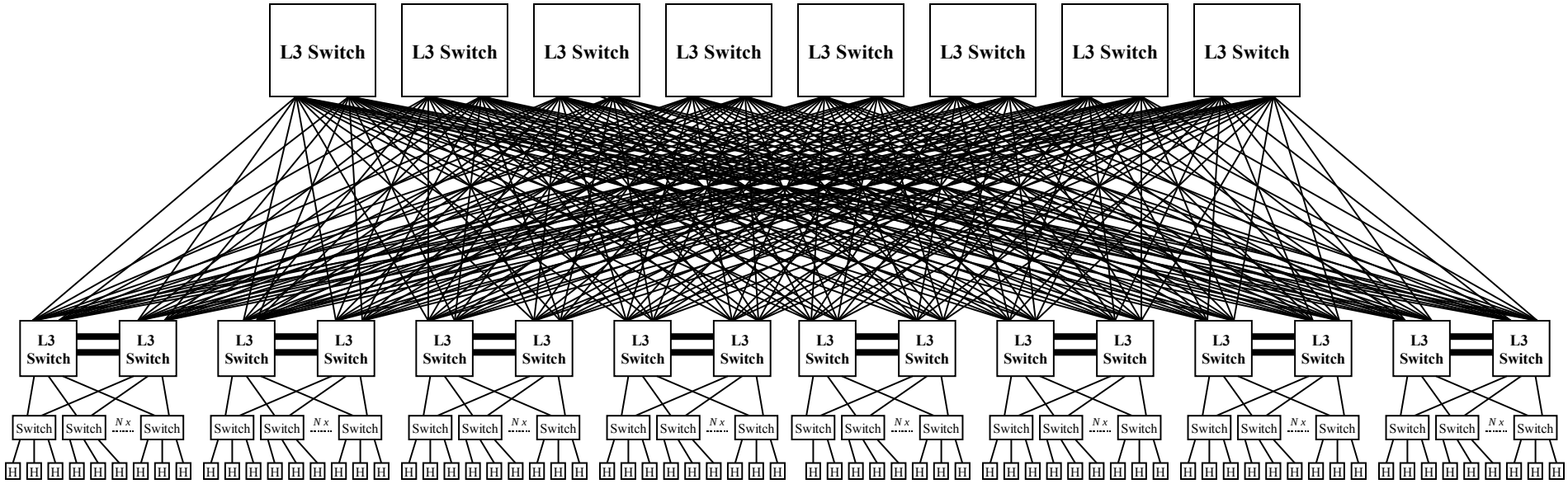


We are currently building this…

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

** Way too many paths **

** Way too many cables **



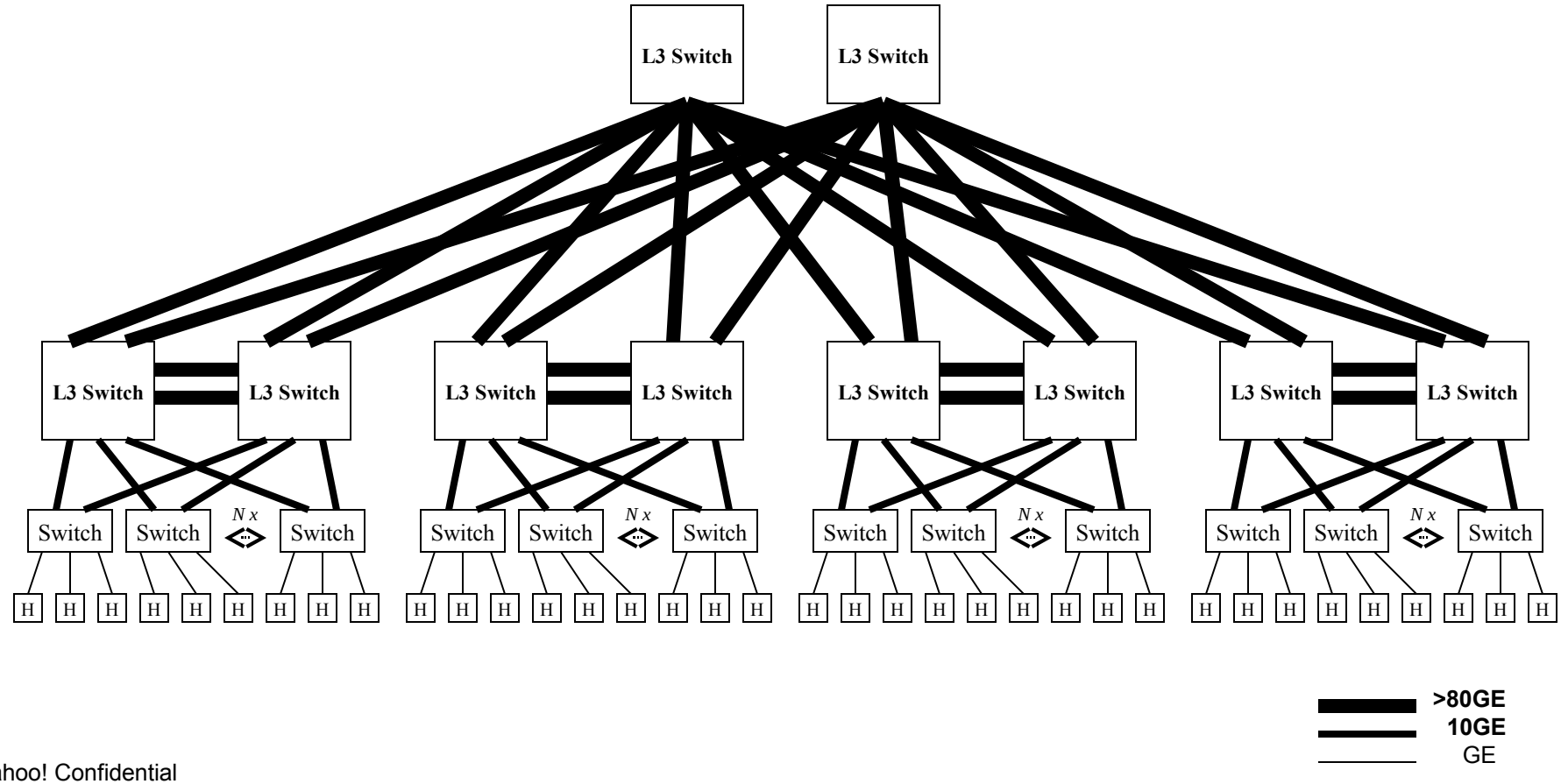
L3 Switch <10GE> L3 Switch
L3 Switch <GE> Switch
Host <GE> Switch

But what we'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 **



Why Skip 40GbE and go to 100+GbE?

- The IEEE standards take 4-5 years
- We are in the 2nd year of the process, so another 2-3 years to go
  - We don't have a way to accelerate it
- We need a standards based solution for higher speed links
  - It is the only way to achieve Vendor Interoperability
- 100+GbE is achievable
  - Vendors are already shipping chips that are capable of doing 66% of 100GbE packet rates
  - NTT has successfully tested 111Gbps x 140 channels for 160km (09/06)
- There is a real need for 100+GbE by 2010
  - We have a need for it today already.
  - We are not alone – many enterprise datacenter, content provider, IXP, HPC, and other end-users are in the same situation
  - LAG, ECMP, PLM, APL are either insufficient, or too complex
  - 40GbE is simply too little, too late in 2010!
- We need to be designing a standard now for technology shipping in '09-'10, not for technology we already have today!


LIFE ENGINE™