The Dawn of Channelized Ethernet

Nathan Farrington

Hot Interconnects 22, Mountain View, CA, USA, August 26-27, 2014

Outline

This talk is mostly about parallelism.

EtherChannel: Parallelism Going Up

- Kalpana's 7-port 10 Mb/s EtherSwitch EPS-700 (1989)
  - Allows 2 or more Ethernet ports to be grouped into a single logical port of a higher data rate
  - Ethernet frames load balanced among physical ports using a hash algorithm
- Kalpana acquired by Cisco in 1994; still called EtherChannel
- IEEE 802.3ad Link Aggregation Group (LAG) (2000)

Problems with EtherChannel

- Requires software to configure both endpoints, e.g. LACP
    - Misconfiguration can lead to forwarding loops
    - At Facebook, we experimented with a design that used EtherChannel and no software. Result: unstable network.
- Imperfect hashing yields lower throughput
- Later protocols such as PCI Express solve these problems in hardware (2004)

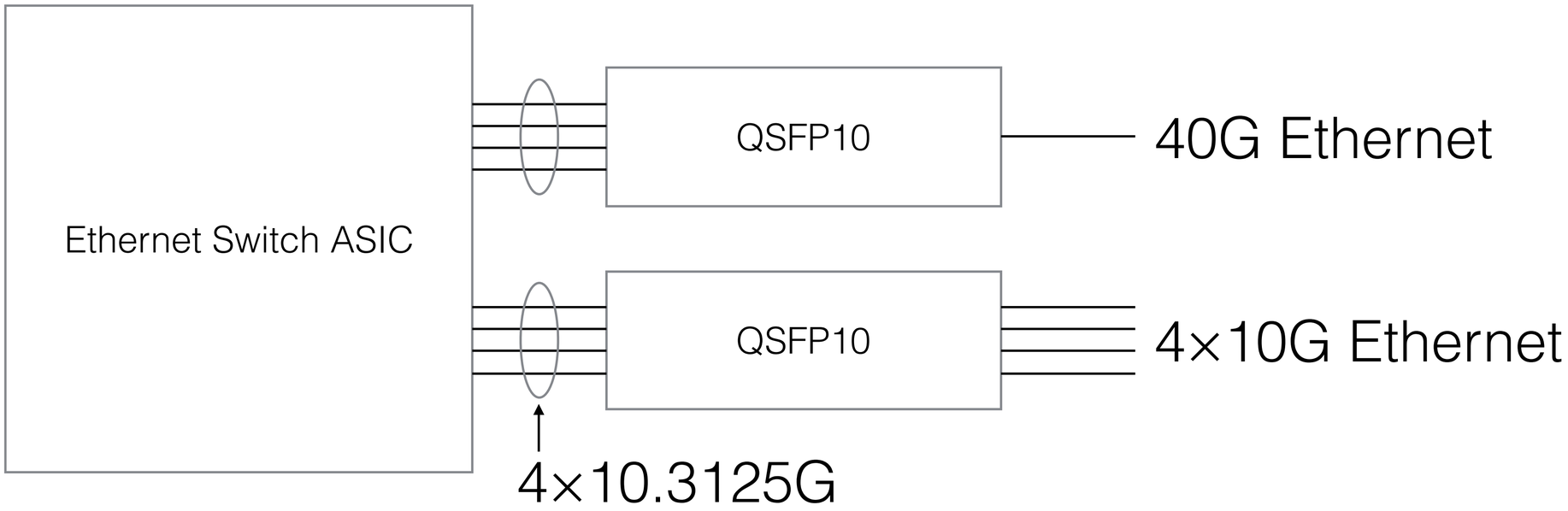
Ethernet Parallelism Going Down

Physical Layer	Rate	Media	Channels	Rate per Channel
1000BASE-T	1G	Twisted Pair (Cat 5e)	4	
10GBASE-CX4	10G	InfiniBand CX4	4	3.125G
10GBASE-KX4	10G	PCB Copper Trace	4	3.125G
10GBASE-LX4	10G	WDM over MMF	4	3.125G
10GBASE-T	10G	Twisted Pair (Cat 6)	4	
40GBASE-KR4	40G	PCB Copper Trace	4	10.3125G
40GBASE-CR4	40G	Twin Ax	4	10.3125G
40GBASE-SR4	40G	MMF	4	10.3125G
40GBASE-LR4	40G	WDM over SMF	4	10.3125G
40GBASE-T	40G	Twisted Pair (Cat 8)	4	
100GBASE-CR10	100G	Twin Ax	10	10.3125G
100GBASE-SR10	100G	MMF	10	10.3125G
100GBASE-CR4	100G	Twin Ax	4	25.78125G
100GBASE-SR4	100G	MMF	4	25.78125G
100GBASE-LR4	100G	WDM over SMF	4	25.78125G
100GBASE-KR4	100G	PCB Copper Trace	4	25.78125G

Ethernet's Approach to Parallelism is Not So Good

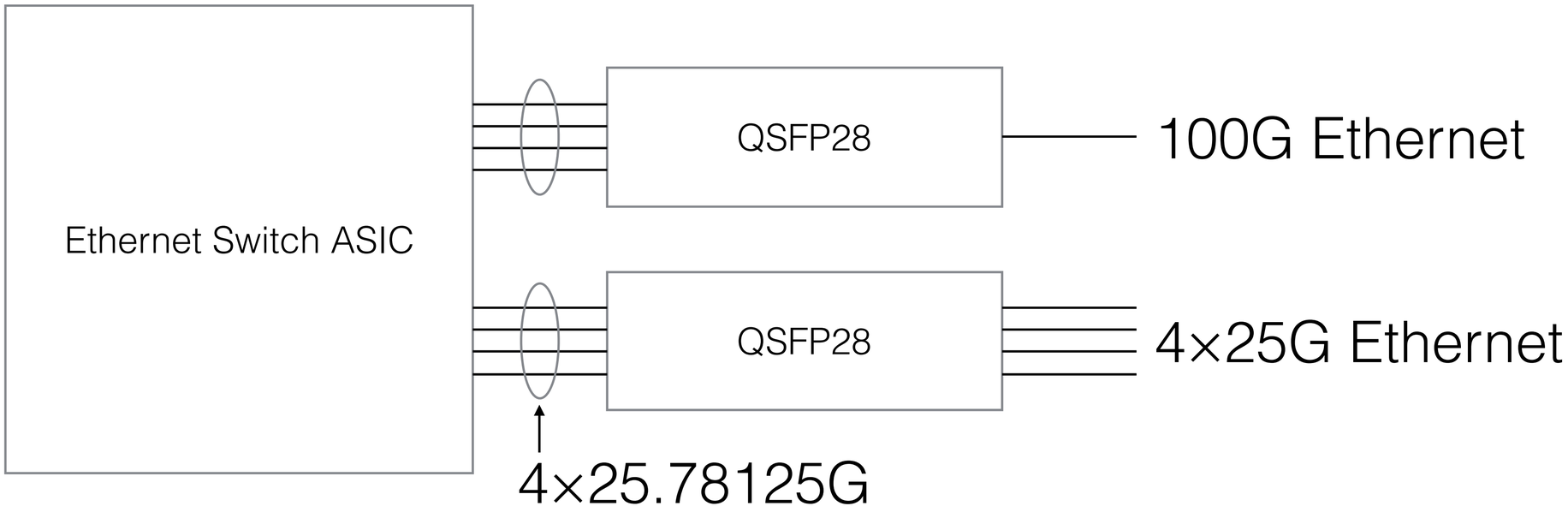
- Ethernet uses EtherChannel to combine 2 to 8 Ethernet links into one logical link.
  - EtherChannel is a flawed protocol.
- Ethernet combines 4 (sometimes 10) parallel communication channels to go from lower SERDES rates (e.g. 3.125G, 10.3125G, 25.78125G) to higher link rates (e.g. 10G, 40G, 100G).
  - This fixed configuration is wasteful if we need less (or more) capacity, e.g. servers.

Parallelism and QSFP10



The same technology can be used for 4 servers or 1 fabric link.

Parallelism and QSFP28



Problem: There is no such thing as 25G Ethernet.

25G Ethernet Consortium

http://25gethernet.org

IEEE 802.3 25G Ethernet Study Group

- IEEE 802 LMSC July 2014 Plenary Meeting
- Mark Nowell, Chair (Cisco)
- Most of the work was already done for 100G Ethernet

http://www.ieee802.org/3/cfi/request_0714_1.html

http://www.enterprisetech.com/2014/07/23/ieee-gets-behind-25g-ethernet-effort/

Technology Exploration Forum 2014: The Rate Debate

- Hosted by The Ethernet Alliance
- October 16, 2014 at Santa Clara Convention Center
- Discussion on 2.5G, 25G, 50G, and 200G Ethernet
  - In addition to 40G, 100G, and 400G Ethernet

Quick Overview of PCI Express

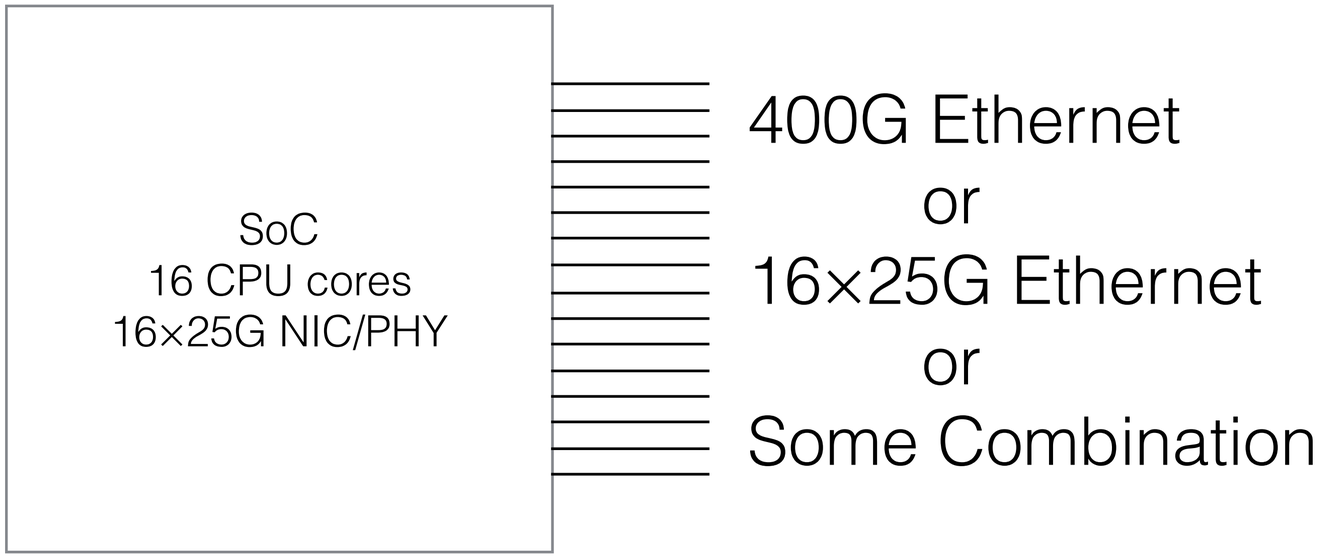
- 1, 2, 4, 8, 12, 16, or 32 physical channels (lanes) can be grouped into one logical channel (link). Just like Ethernet, the clock is recovered from the signal (asynchronous).

	Gen 1	Gen 2	Gen 3	Gen 4
Year	2003	2007	2010	2014?
Code	8b/10b	8b/10b	128b/132b	128b/132b
Layer 1 Rate	2.5 GT/s	5 GT/s	8 GT/s	16 GT/s
Layer 2 Rate	2 Gb/s	4 Gb/s	7.877 Gb/s	15.754 Gb/s

PCIe–Ethernet Mismatch

PCI Express	Ethernet	Waste
PCIe Gen1 ×1 Lane (2.0 Gb/s)	1 Gb/s	50%
PCIe Gen1 ×8 Lanes (16 Gb/s) PCIe Gen2 ×4 Lanes (16 Gb/s) PCIe Gen3 ×2 Lanes (15.754 Gb/s)	10 Gb/s	37.5%
PCIe Gen1 ×32 (64 Gb/s) PCIe Gen2 ×16 (64 Gb/s) PCIe Gen3 ×8 (63.016 Gb/s)	40 Gb/s	37.5%
PCIe Gen2 ×32 (128 Gb/s) PCIe Gen3 ×16 (126.032 Gb/s)	100 Gb/s	21.875%

SoC with Embedded NIC

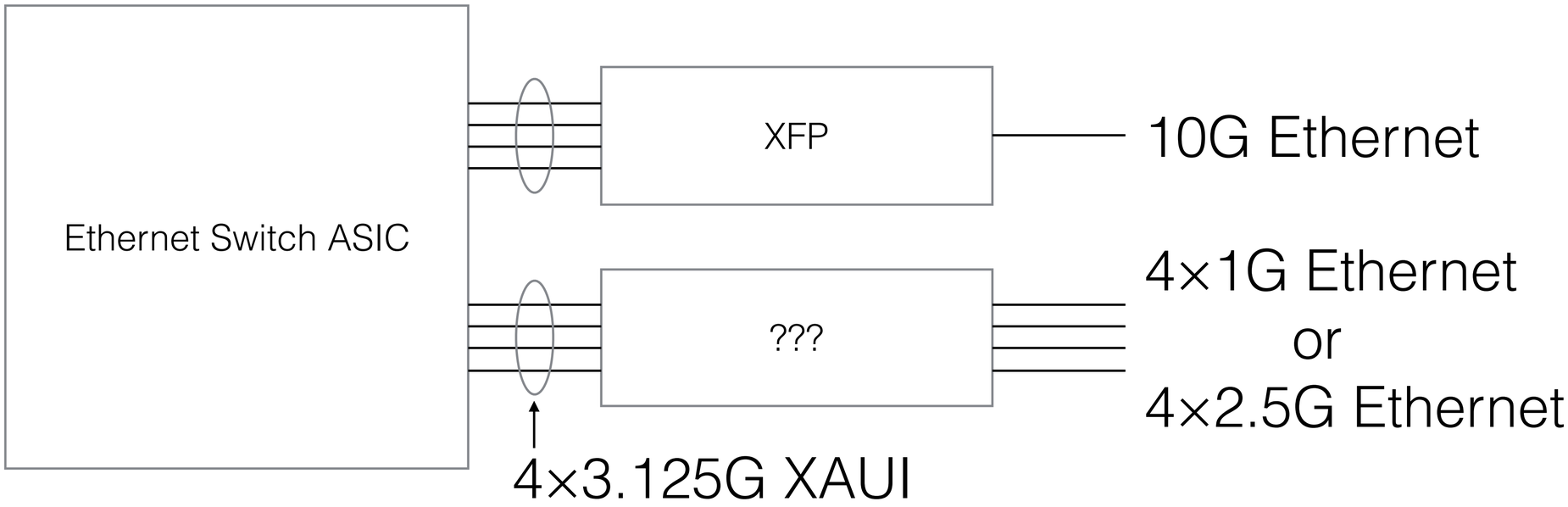


Embedded NIC/PHY removes PCIe waste.

What if Ethernet was like PCI Express?

Channels	25G SERDES	50G SERDES	100G SERDES
1	25G Ethernet	50G Ethernet	100G Ethernet
2	50G Ethernet	100G Ethernet	200G Ethernet
4	100G Ethernet	200G Ethernet	400G Ethernet
8	200G Ethernet	400G Ethernet	800G Ethernet
16	400G Ethernet	800G Ethernet	1.6T Ethernet

But what about 2.5G Ethernet?



Both Broadcom and Fulcrum had developed 2.5G Ethernet based on 10G Ethernet back when 10G SERDES were not available. Why did this technology not catch on, and why is 25G different today?

Summary

- EtherChannel tries to solve a hardware problem with a software solution, a bad thing
- PCI Express has shown us what Ethernet should be: powers-of-2 based on the fundamental SERDES rate
- Embedded NICs and PHYs solve the PCI Express–Ethernet mismatch waste problem
- Emerging 25G Ethernet standardization is just the beginning of the channelization of Ethernet; powers of 10 are on the way out

And Now For Something Completely Different

Is Ethernet a Layer 2 Protocol?

ARP, DHCP, STP, OSPF, …

Layer 7: Application

Layer 6: ??? (mythical)

Layer 5: ??? (mythical)

PAUSE, FCoE, priority, …

Layer 4: Transport

DST, SRC, broadcast, LAG, …

Layer 3: Network

preamble, ethtype, FCS, …

Layer 2: Data Link

NRZ, 8b/10b, 64b/66b, …

Layer 1: Physical Coding

coax, twisted pair, fiber, …

Layer 0: Physical Channel

That’s like saying the Odyssey is about a cyclops.