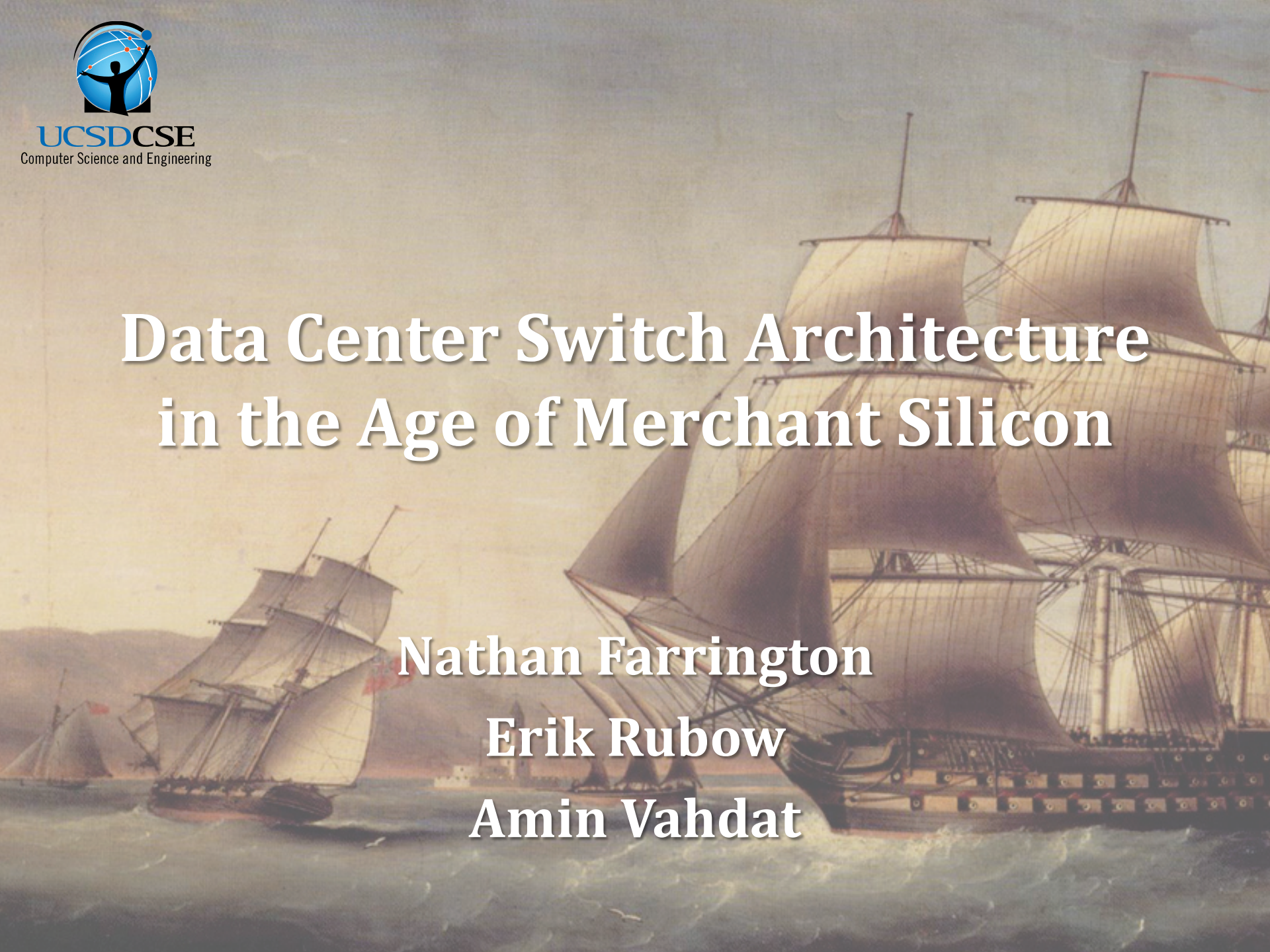


Data Center Switch Architecture in the Age of Merchant Silicon

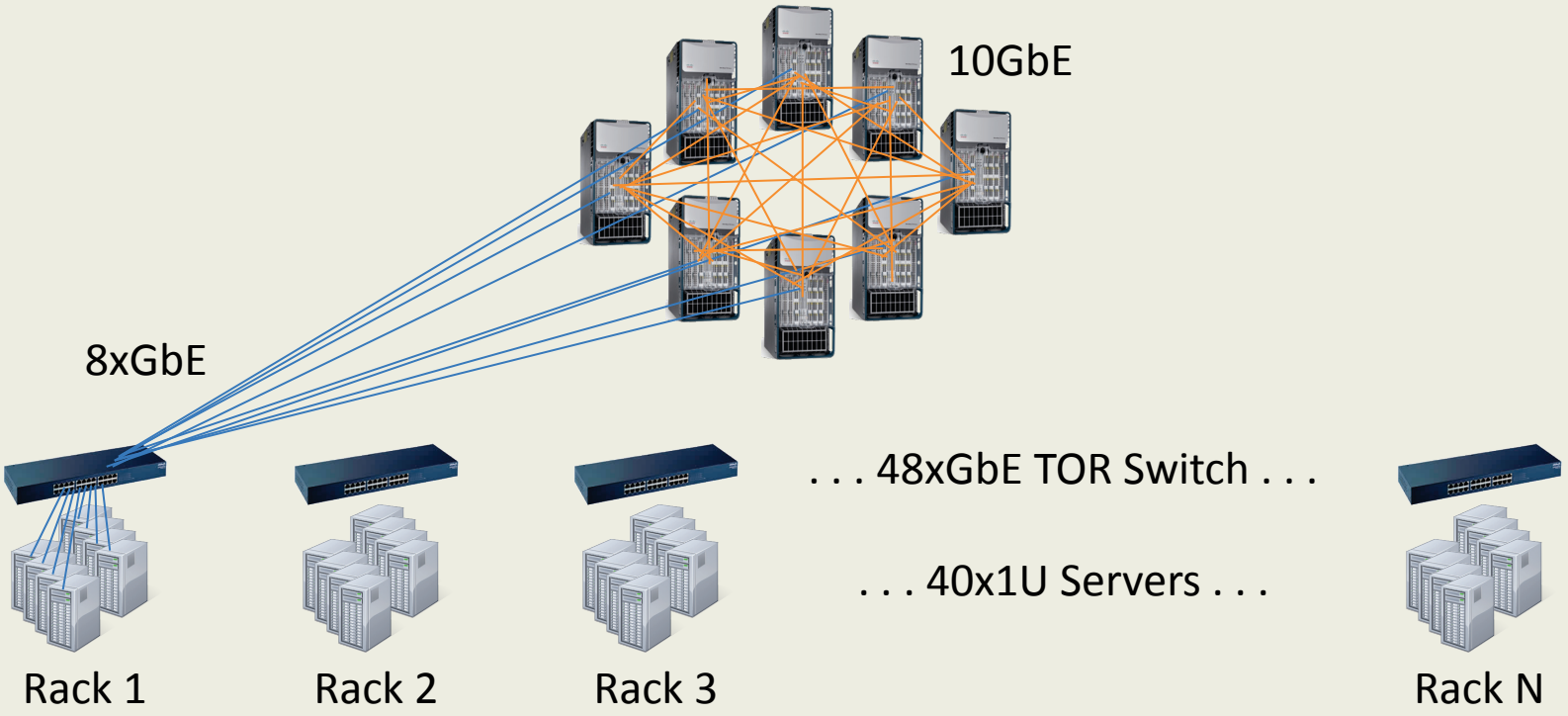
Nathan Farrington
Erik Rubow
Amin Vahdat

The Network is a Bottleneck

- HTTP request amplification
  - Web search (e.g. Google)
  - Small object retrieval (e.g. Facebook)
  - Web services (e.g. Amazon.com)
- MapReduce-style parallel computation
  - Inverted search index
  - Data analytics
- Need high-performance interconnects

Nathan Farrington
farrington@cs.ucsd.edu

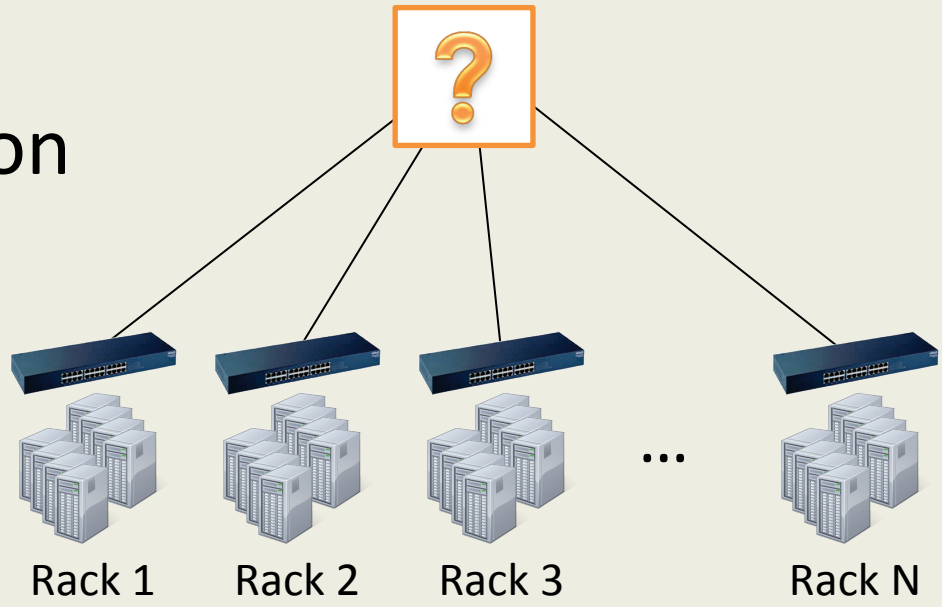
The Network is Expensive



Nathan Farrington
farrington@cs.ucsd.edu

What we really need: One Big Switch

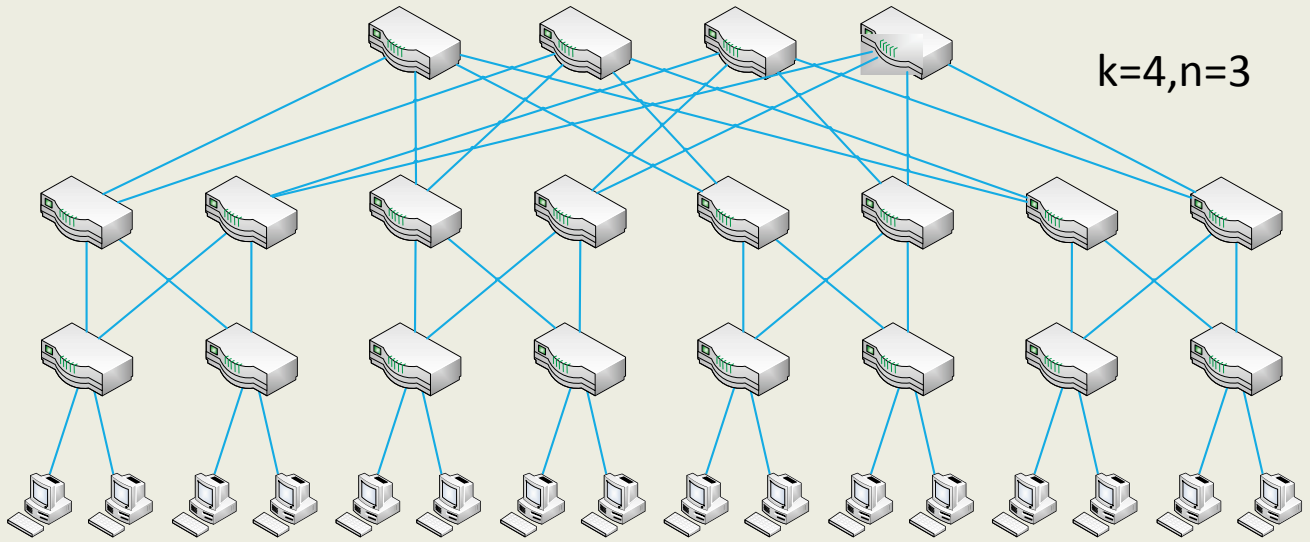
- Commodity
- Plug-and-play
- Potentially
  no oversubscription



Nathan Farrington
farrington@cs.ucsd.edu

Why not just use a fat tree of commodity TOR switches?

M. Al-Fares, A. Loukissas, A. Vahdat. A Scalable, Commodity Data Center Network Architecture. In SIGCOMM ’08.

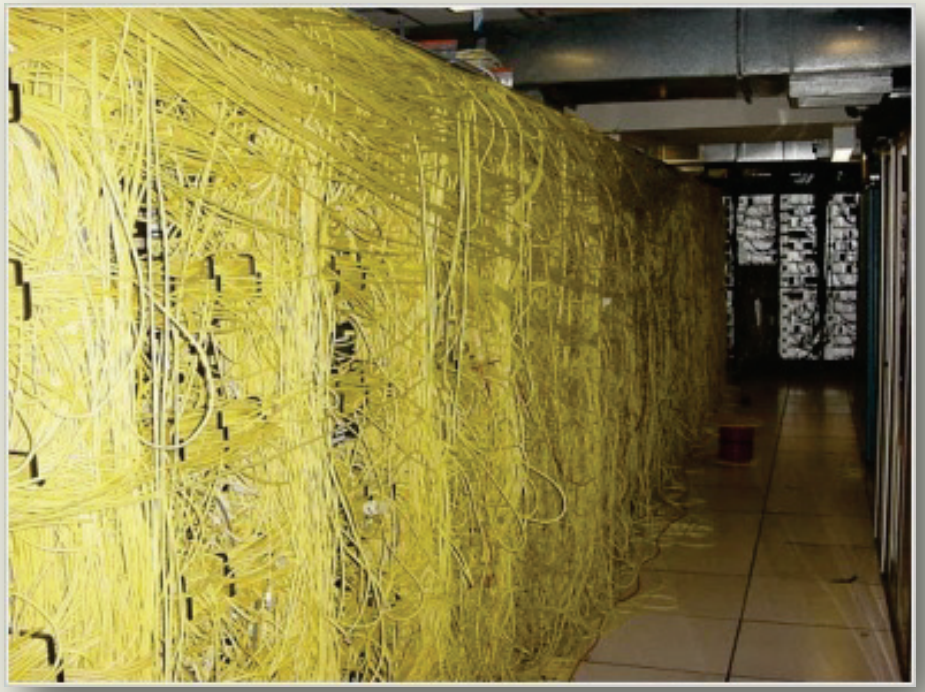


Nathan Farrington
farrington@cs.ucsd.edu

10 Tons of Cable

- 55,296 Cat-6 cables
- 1,128 separate cable bundles

The “Yellow Wall”



Nathan Farrington
farrington@cs.ucsd.edu

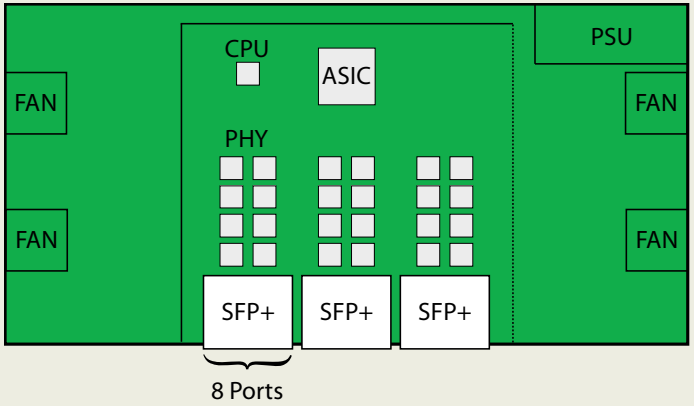
Merchant Silicon gives us Commodity Switches

Maker	Broadcom	Fulcrum	Fujitsu
Model	BCM56820	FM4224	MB86C69RBC
Ports	24	24	26
Cost	NDA	NDA	$410
Power	NDA	20 W	22 W
Latency	< 1 µs	300 ns	300 ns
Area	NDA	40 x 40 mm	35 x 35 mm
SRAM	NDA	2 MB	2.9 MB
Process	65 nm	130 nm	90 nm

Nathan Farrington
farrington@cs.ucsd.edu

Eliminate Redundancy

- Networks of packet switches contain many redundant components
  - chassis, power conditioning circuits, cooling
  - CPUs, DRAM
- Repackage these discrete switches to lower the cost and power consumption

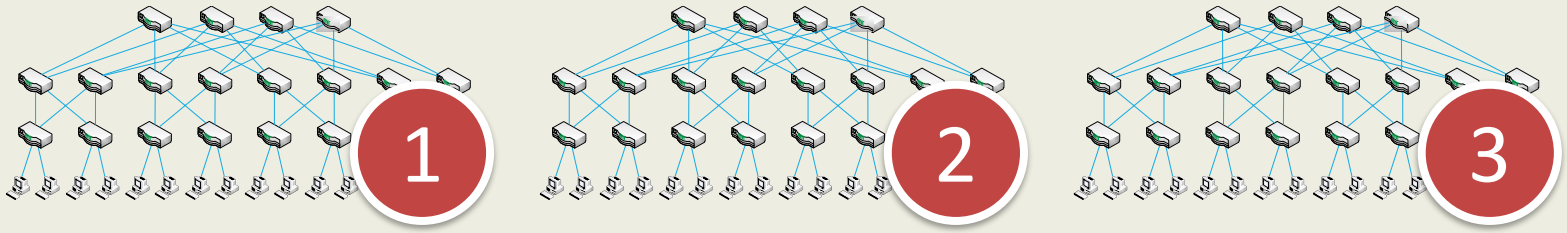


Nathan Farrington
farrington@cs.ucsd.edu

Our Architecture, in a Nutshell

- Fat tree of merchant silicon switch ASICs
- Hiding cabling complexity with PCB traces and optics
- Partition into multiple pod switches + single core switch array
- Custom EEP ASIC to further reduce cost and power
- Scales to 65,536 ports when 64-port ASICs become available, late 2009

3 Different Designs



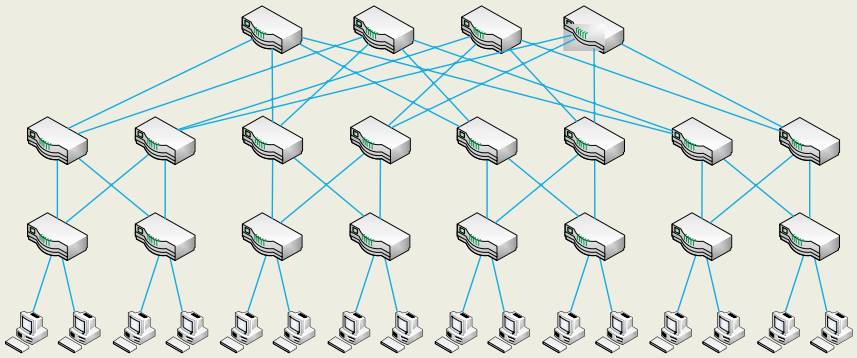
- 24-ary 3-tree
- 720 switch ASICs
- 3,456 ports of 10GbE
- No oversubscription

Nathan Farrington
farrington@cs.ucsd.edu

Network 1: No Engineering Required

- 720 discrete packet switches, connected with optical fiber

Cost of Parts	$4.88M
Power	52.7 kW
Cabling Complexity	3,456
Footprint	720 RU
NRE	$0



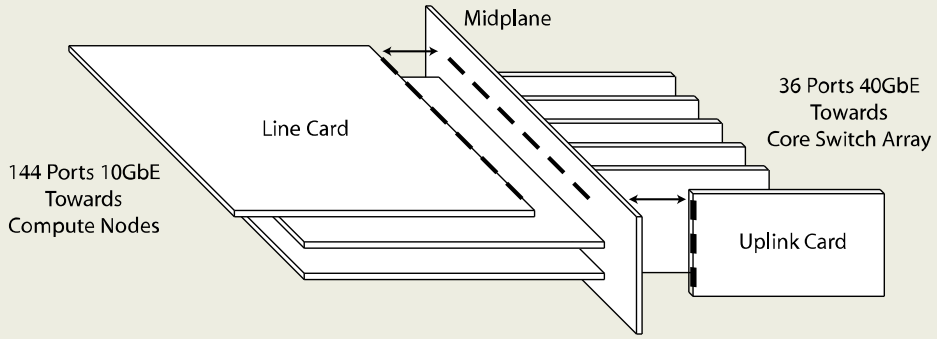
Cabling complexity (noun): the number of long cables in a data center network.

Nathan Farrington
farrington@cs.ucsd.edu

Network 2: Custom Boards and Chassis

- 24 “pod” switches, one core switch array, 96 cables

Cost of Parts	$3.07M
Power	41.0 kW
Cabling Complexity	96
Footprint	192 RU
NRE	$3M est



This design is shown in more detail later.

Nathan Farrington
farrington@cs.ucsd.edu

Switch at 10G, but Transmit at 40G

	SFP	SFP+	QSFP
Rate	1 Gb/s	10 Gb/s	40 Gb/s
Cost/Gb/s	$35*	$25*	$15*
Power/Gb/s	500mW	150mW	60mW

* 2008-2009 Prices

Nathan Farrington
farrington@cs.ucsd.edu

Network 3: Network 2 + Custom ASIC

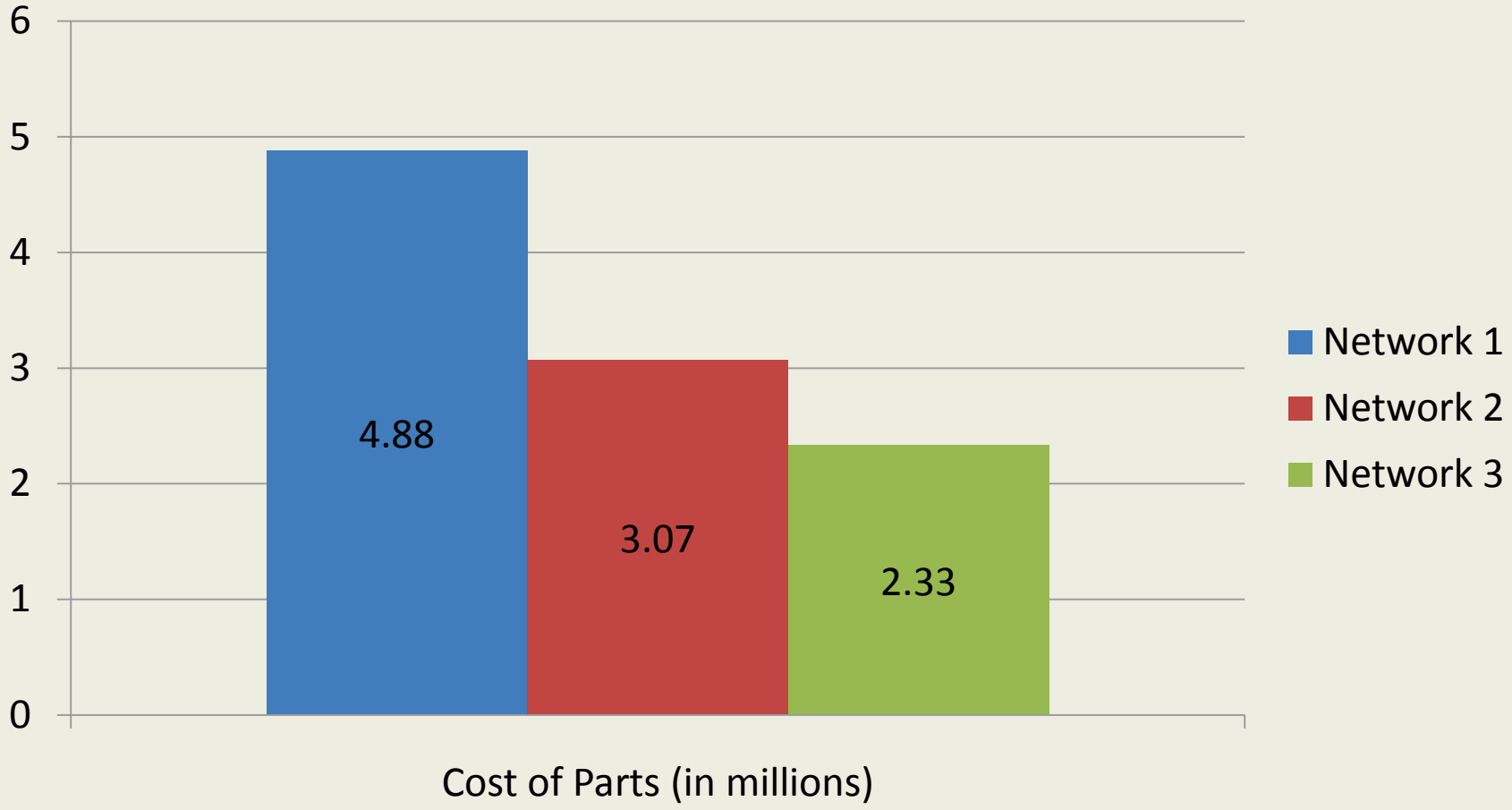
- Uses 40GbE between pod switches and core switch array; everything else is same as Network 2.

Cost of Parts	$2.33M
Power	36.4 kW
Cabling Complexity	96
Footprint	114 RU
NRE	$8M est

This simple ASIC provides tremendous cost and power savings.

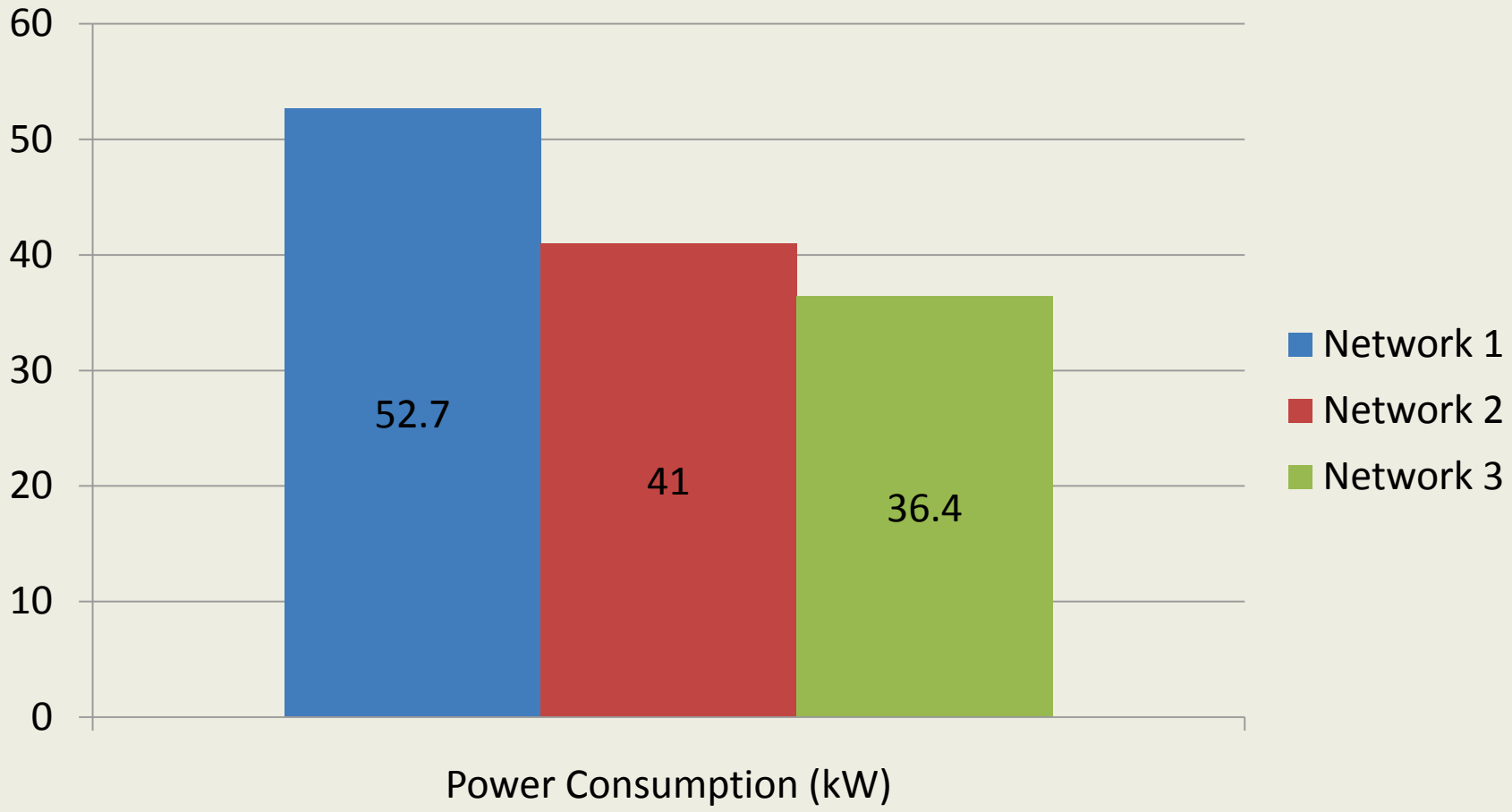
Nathan Farrington
farrington@cs.ucsd.edu

Cost of Parts



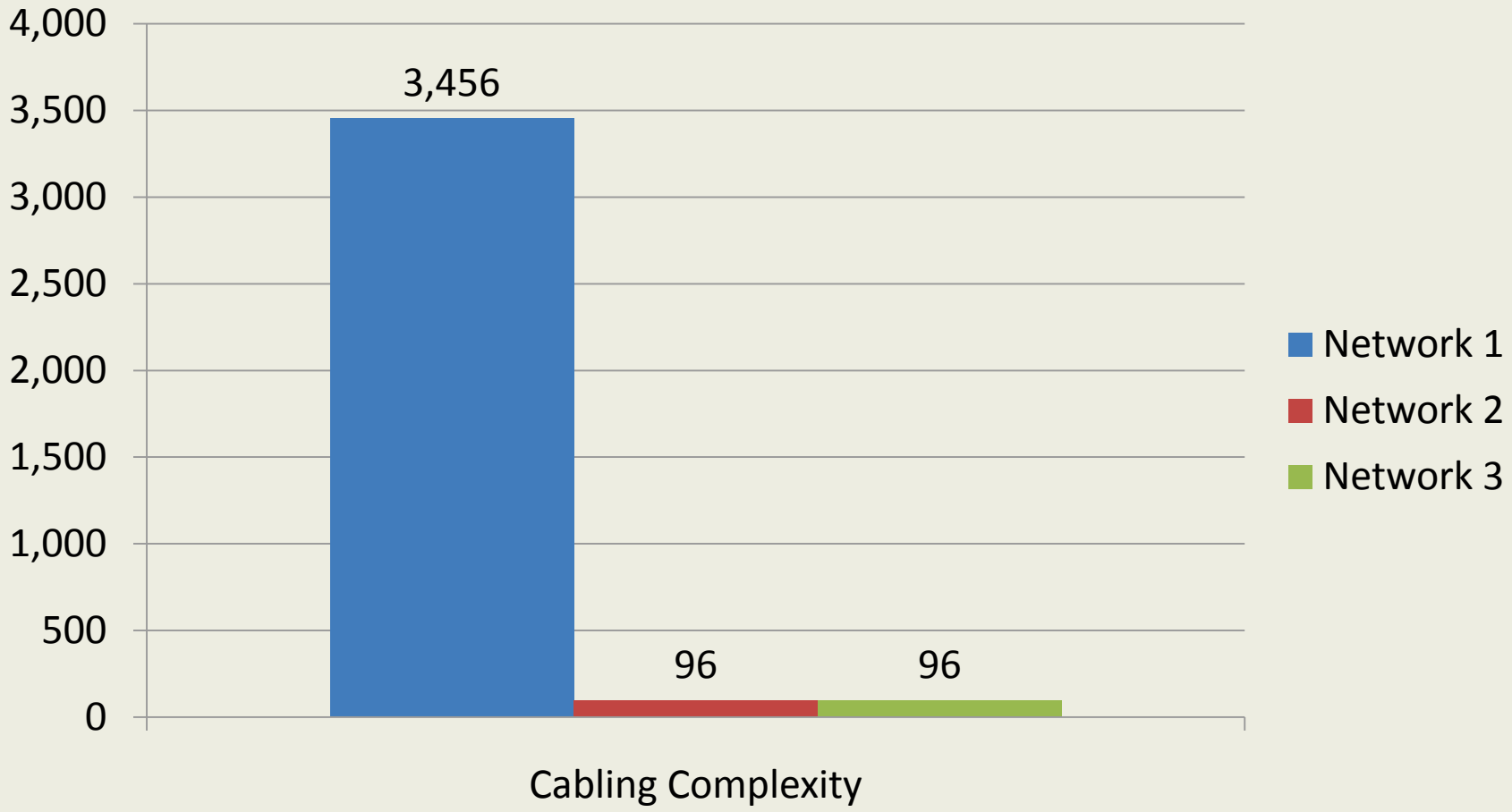
Nathan Farrington
farrington@cs.ucsd.edu

Power Consumption



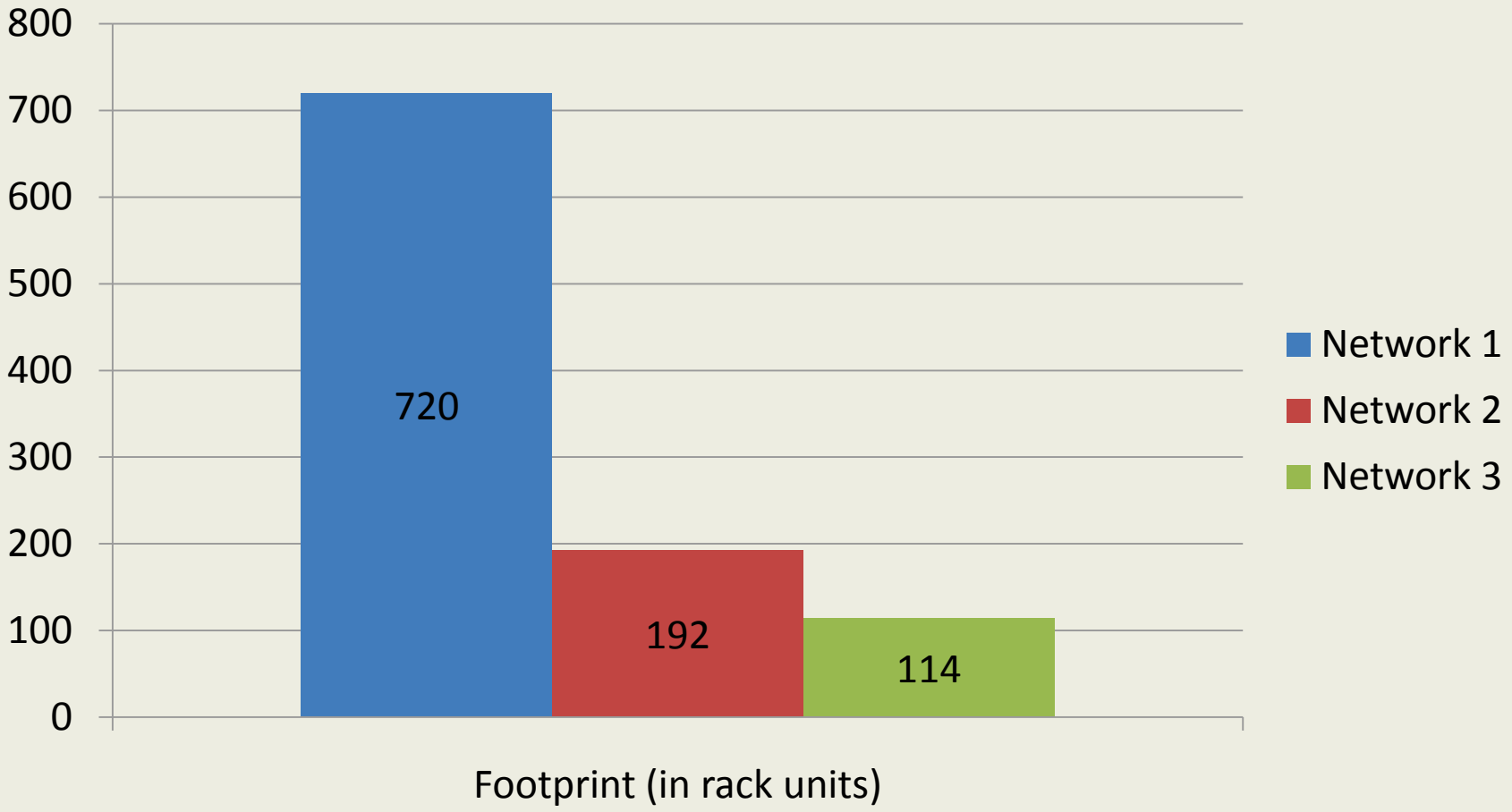
Nathan Farrington
farrington@cs.ucsd.edu

Cabling Complexity



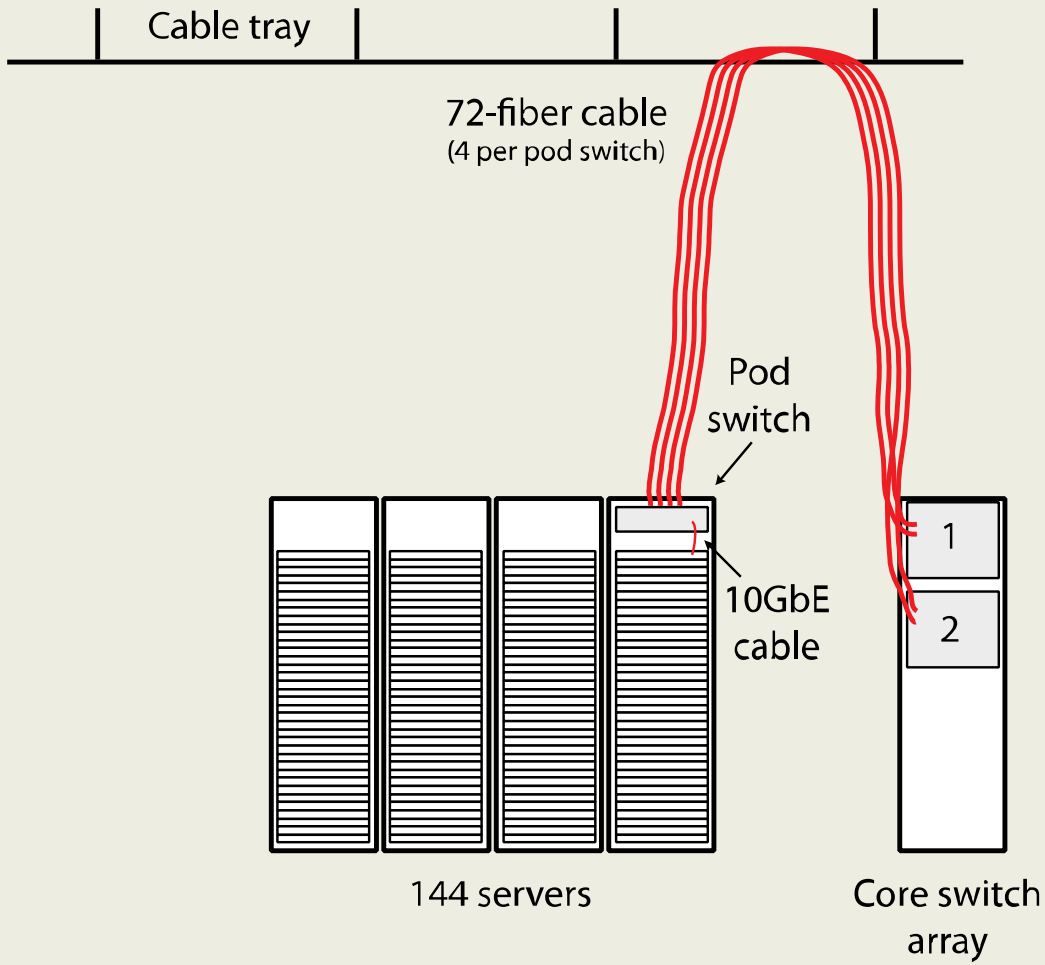
Nathan Farrington
farrington@cs.ucsd.edu

Footprint



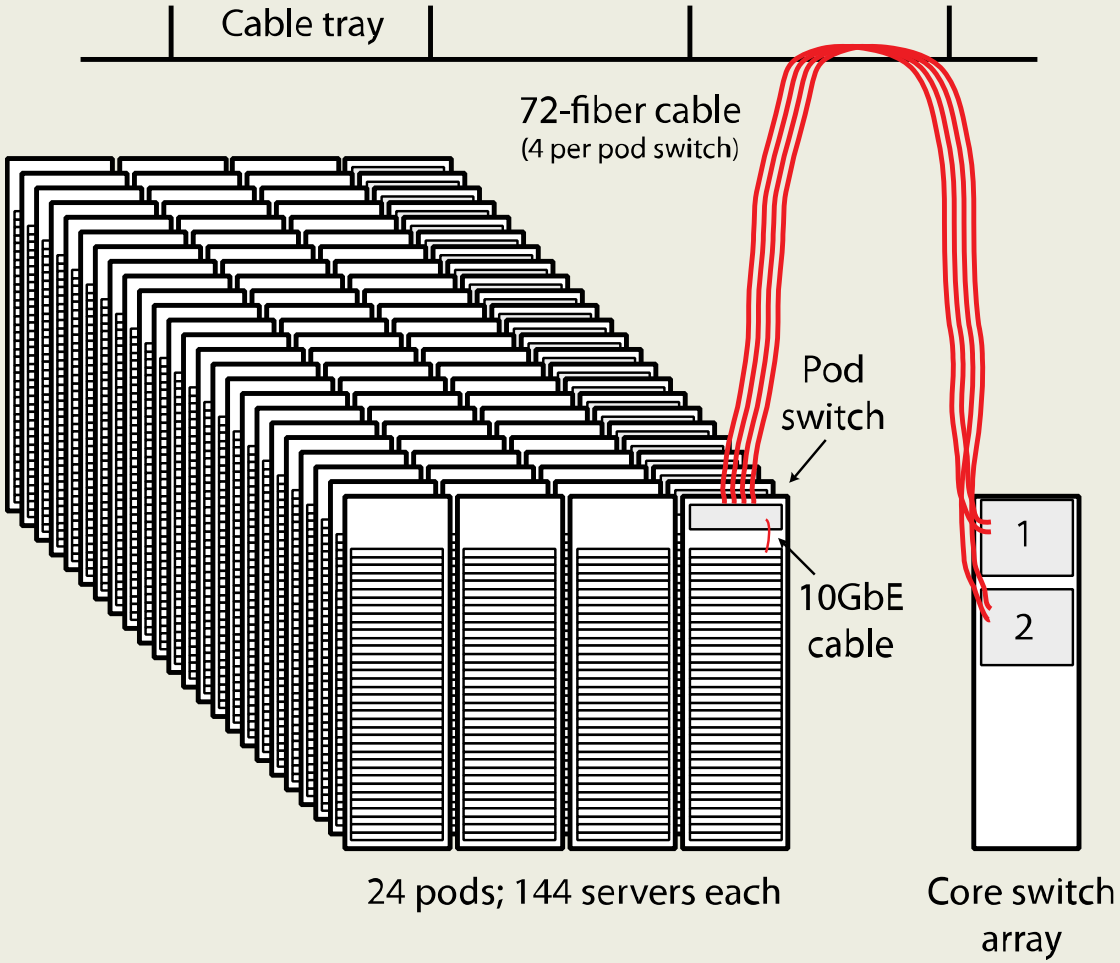
Nathan Farrington
farrington@cs.ucsd.edu

Partially Deployed Switch



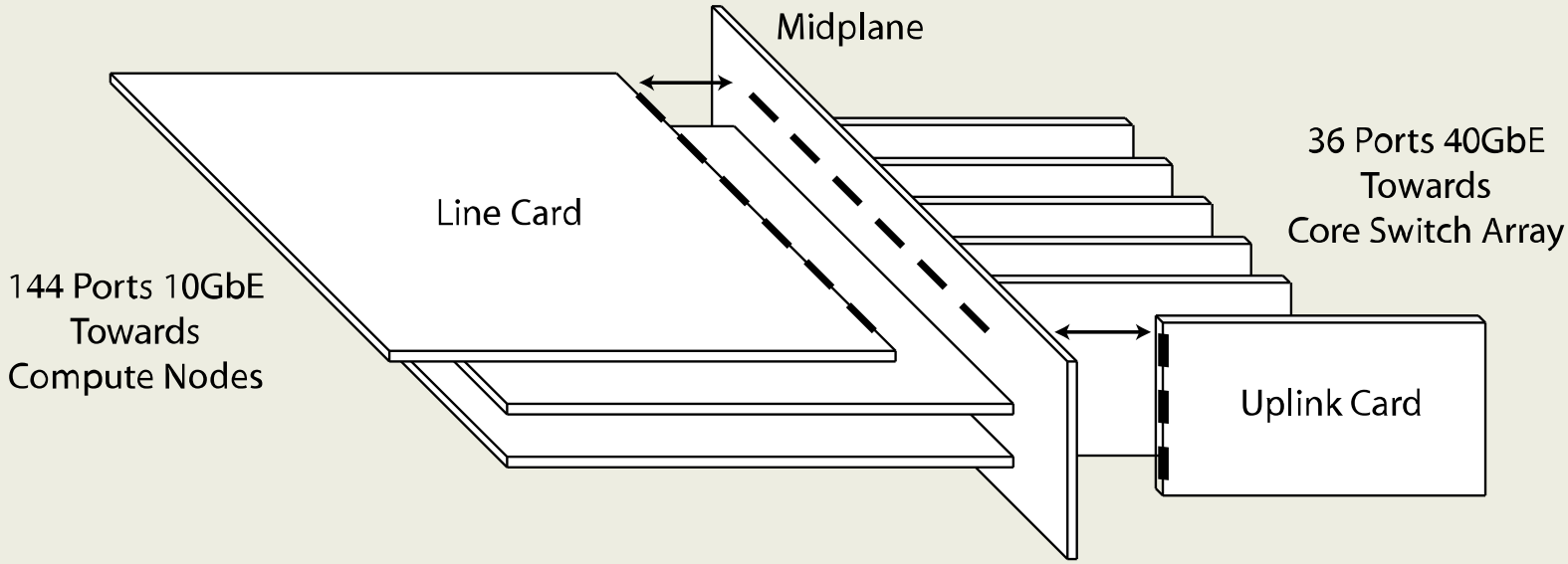
Nathan Farrington
farrington@cs.ucsd.edu

Fully Deployed Switch



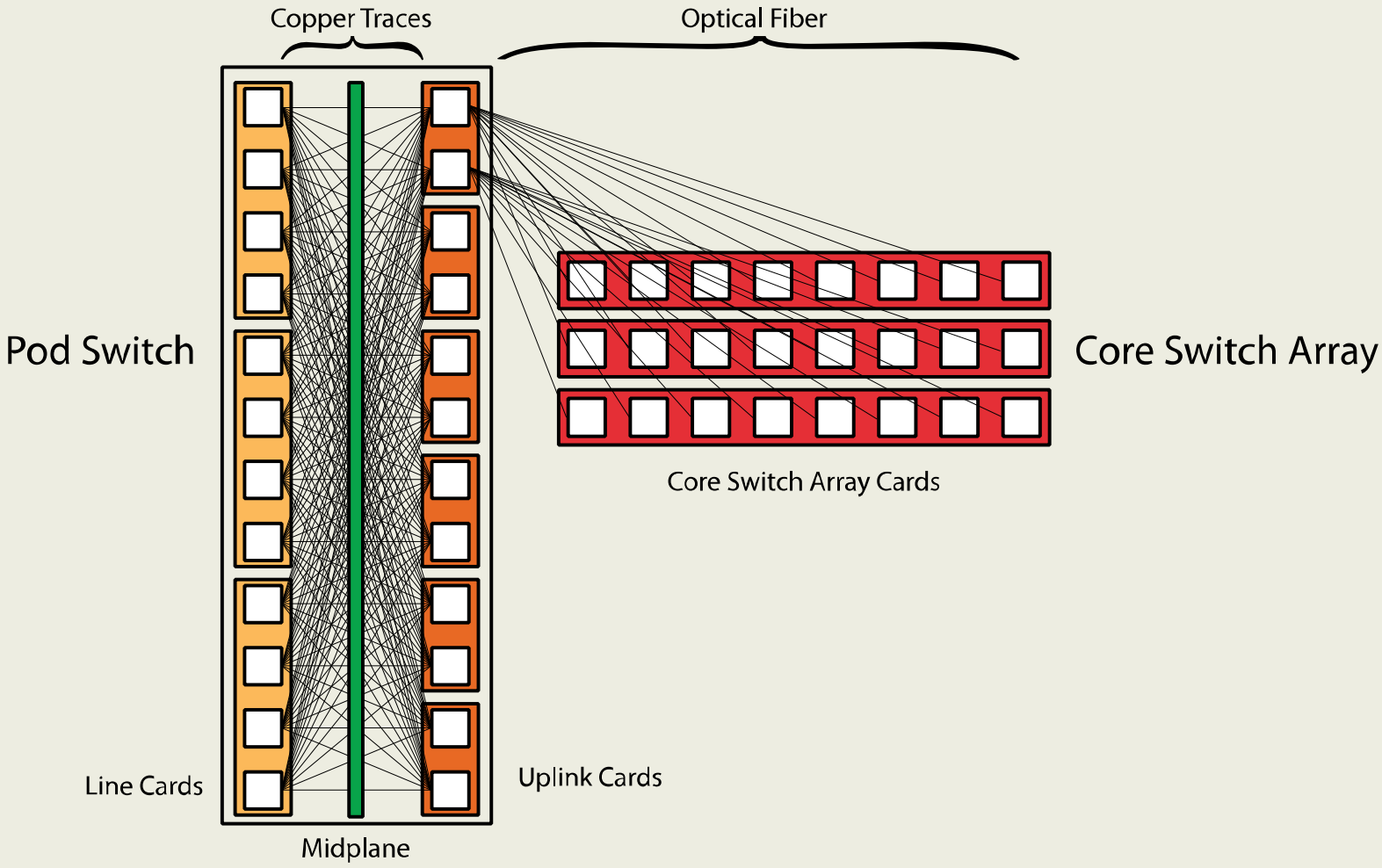
Nathan Farrington
farrington@cs.ucsd.edu

Pod Switch



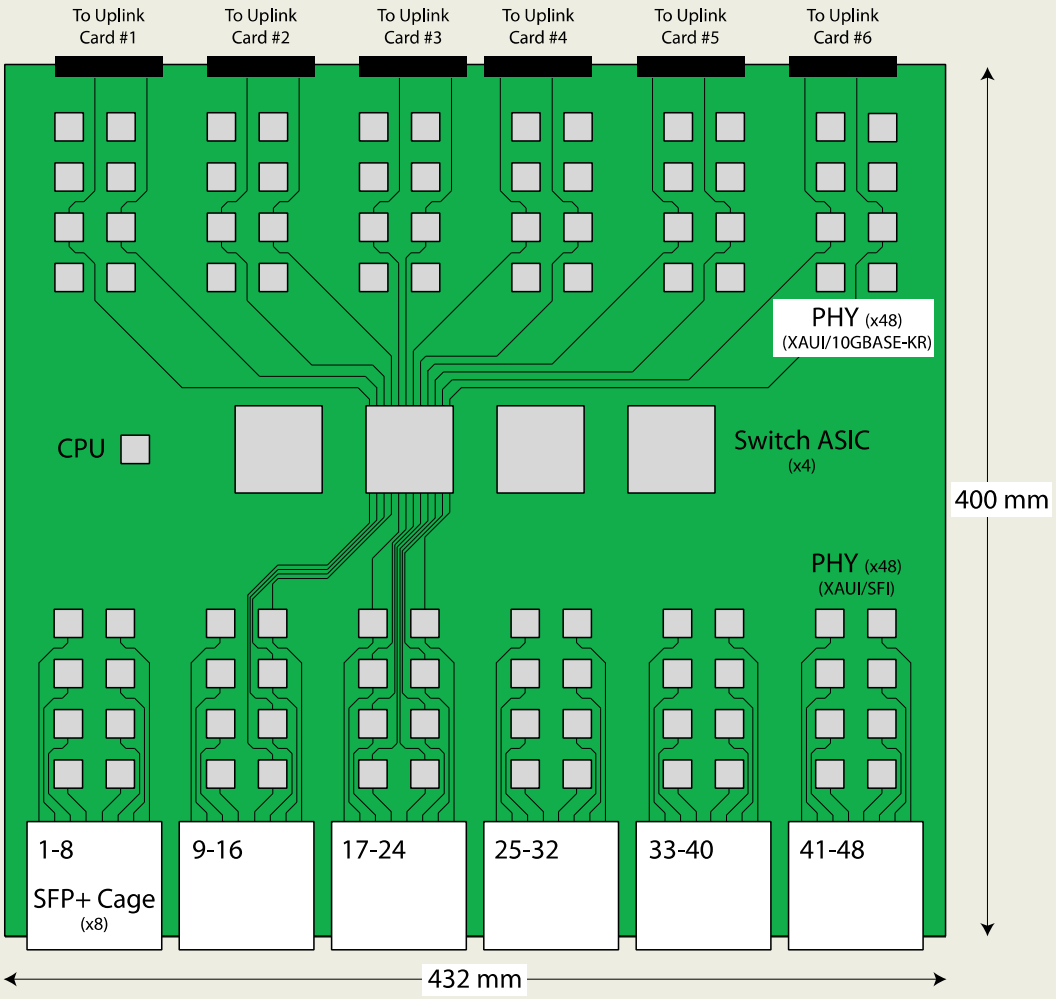
Nathan Farrington
farrington@cs.ucsd.edu

Logical Topology



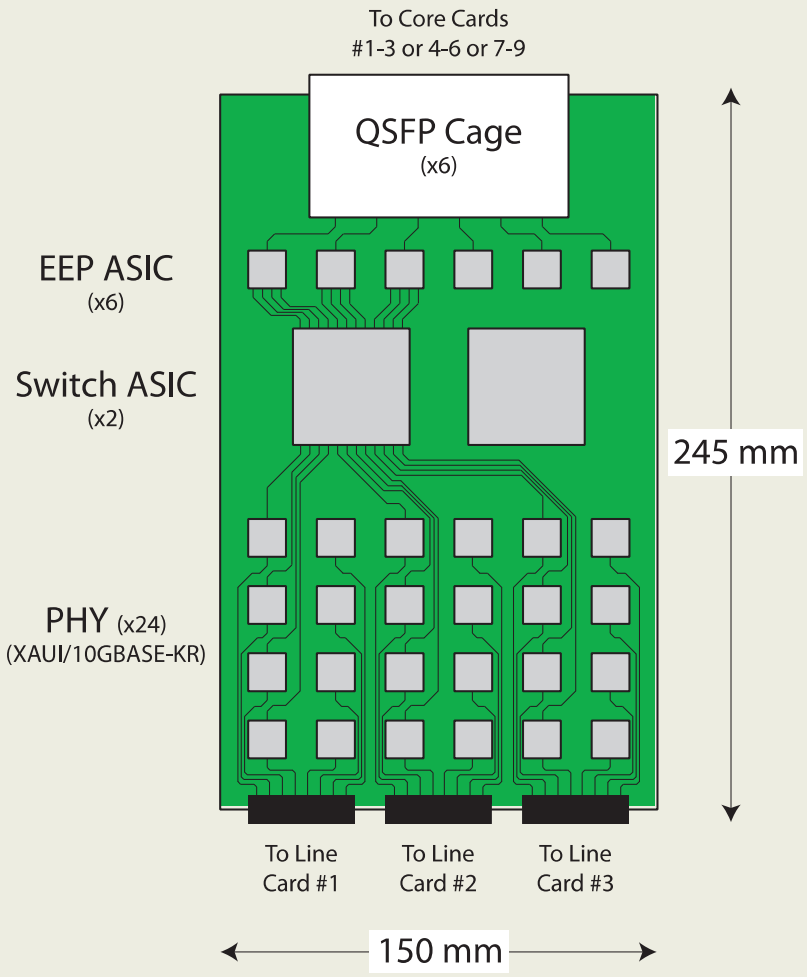
Nathan Farrington
farrington@cs.ucsd.edu

Pod Switch Line Card

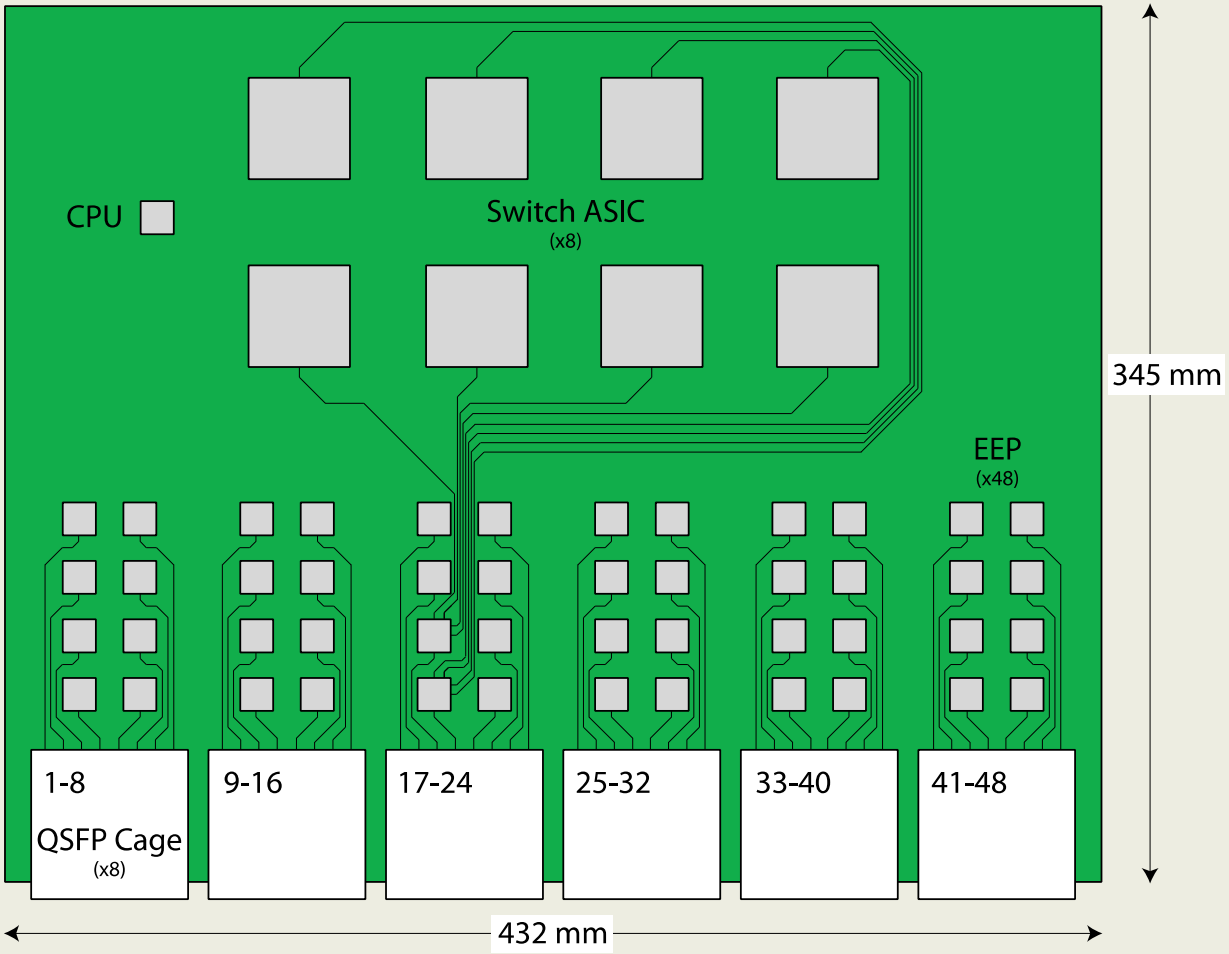


Nathan Farrington
farrington@cs.ucsd.edu

Pod Switch Uplink Card



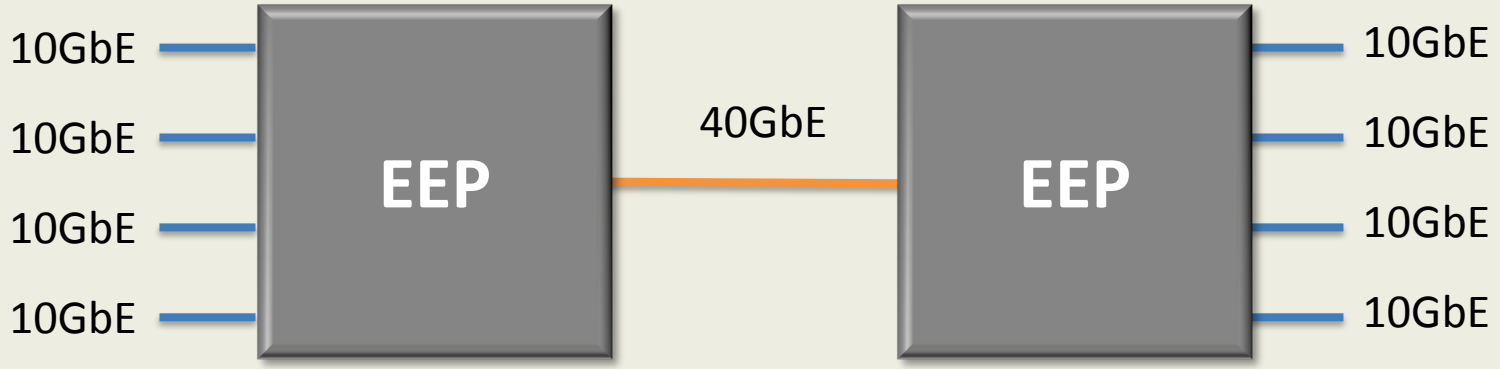
Core Switch Array Card



Nathan Farrington
farrington@cs.ucsd.edu

Why an Ethernet Extension Protocol?

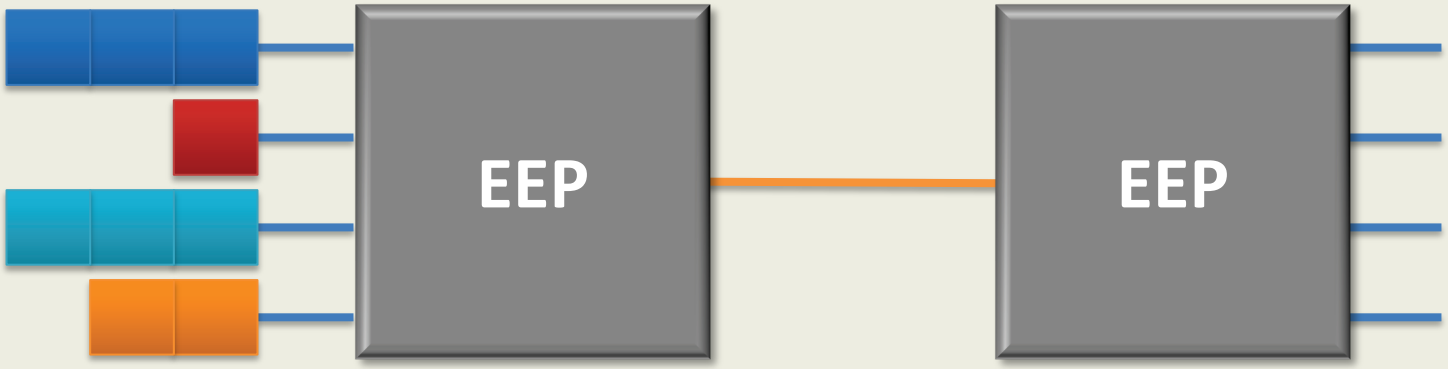
- Optical transceivers are 80% of the cost
- EEP allows the use of fewer and faster optical transceivers



Nathan Farrington
farrington@cs.ucsd.edu

How does EEP work?

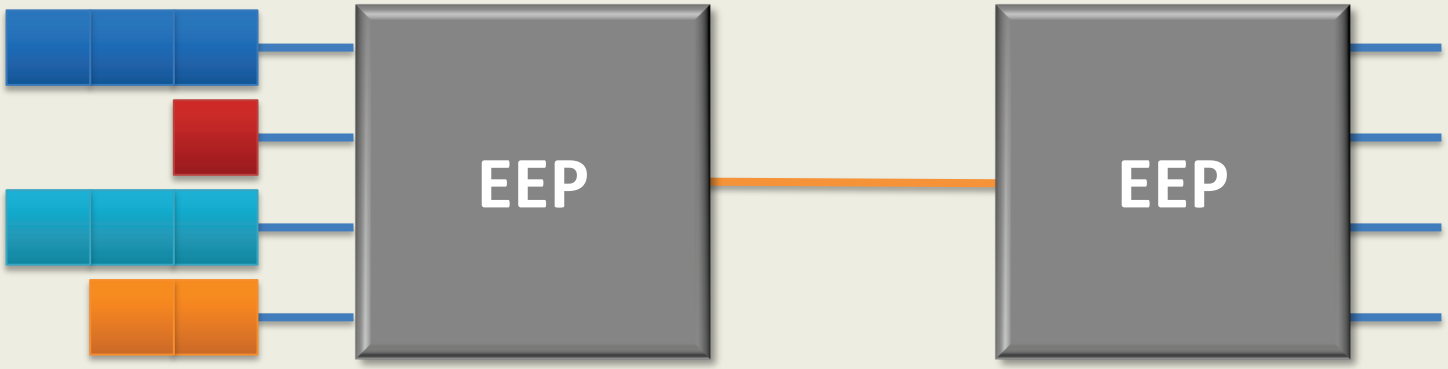
- Ethernet frames are split up into EEP frames
- Most EEP frames are 65 bytes
  - Header is 1 byte; payload is 64 bytes
- Header encodes ingress/egress port



Nathan Farrington
farrington@cs.ucsd.edu

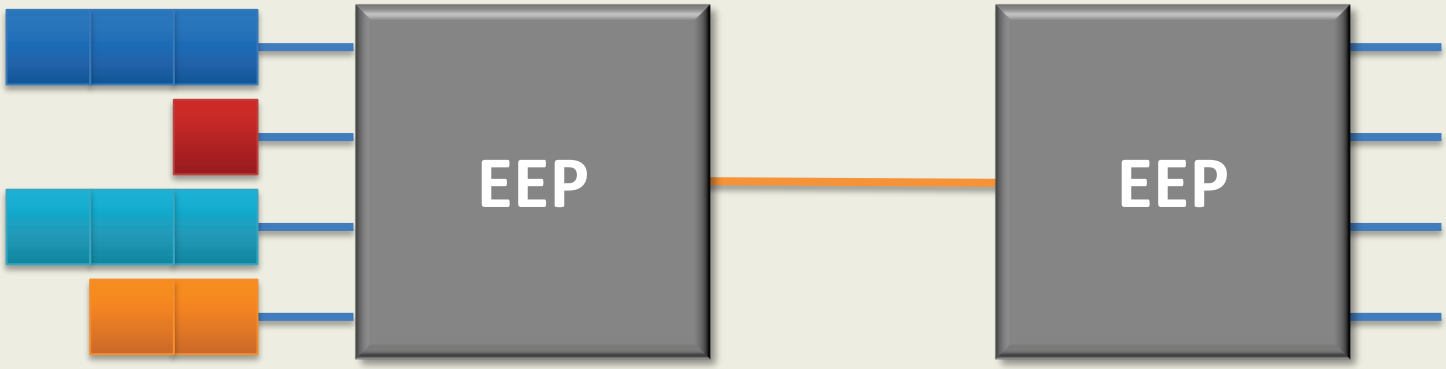
How does EEP work?

- Round-robin arbiter
- EEP frames are transmitted as one large Ethernet frame
- 40GbE overclocked by 1.6%



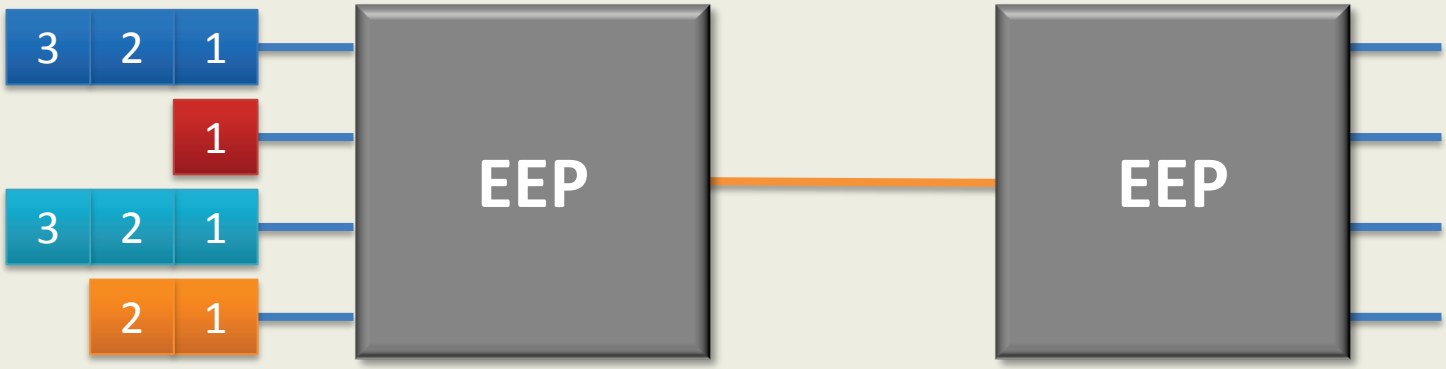
Nathan Farrington
farrington@cs.ucsd.edu

Ethernet Frames

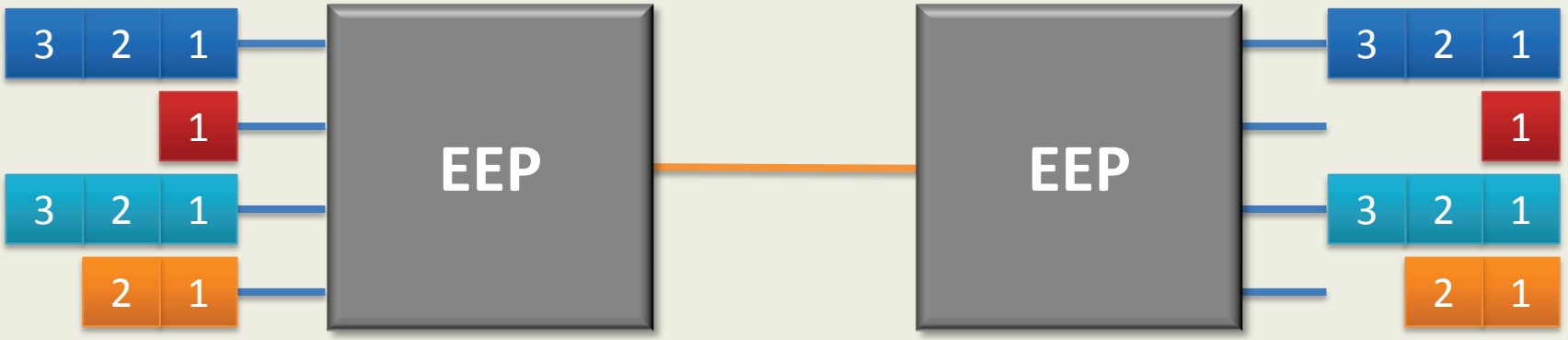


Nathan Farrington
farrington@cs.ucsd.edu

EEP
Frames

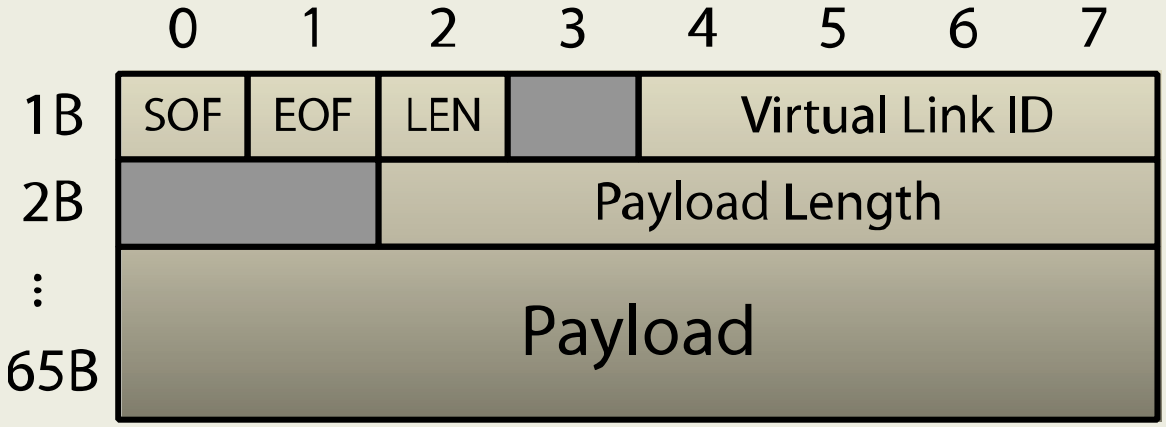


Nathan Farrington
farrington@cs.ucsd.edu



Nathan Farrington
farrington@cs.ucsd.edu

EEP Frame Format



SOF: Start of Ethernet Frame

EOF: End of Ethernet Frame

LEN: Set if EEP Frame contains less than 64B of payload

Virtual Link ID: Corresponds to port number (0-15)

Payload Length: (0-63B)

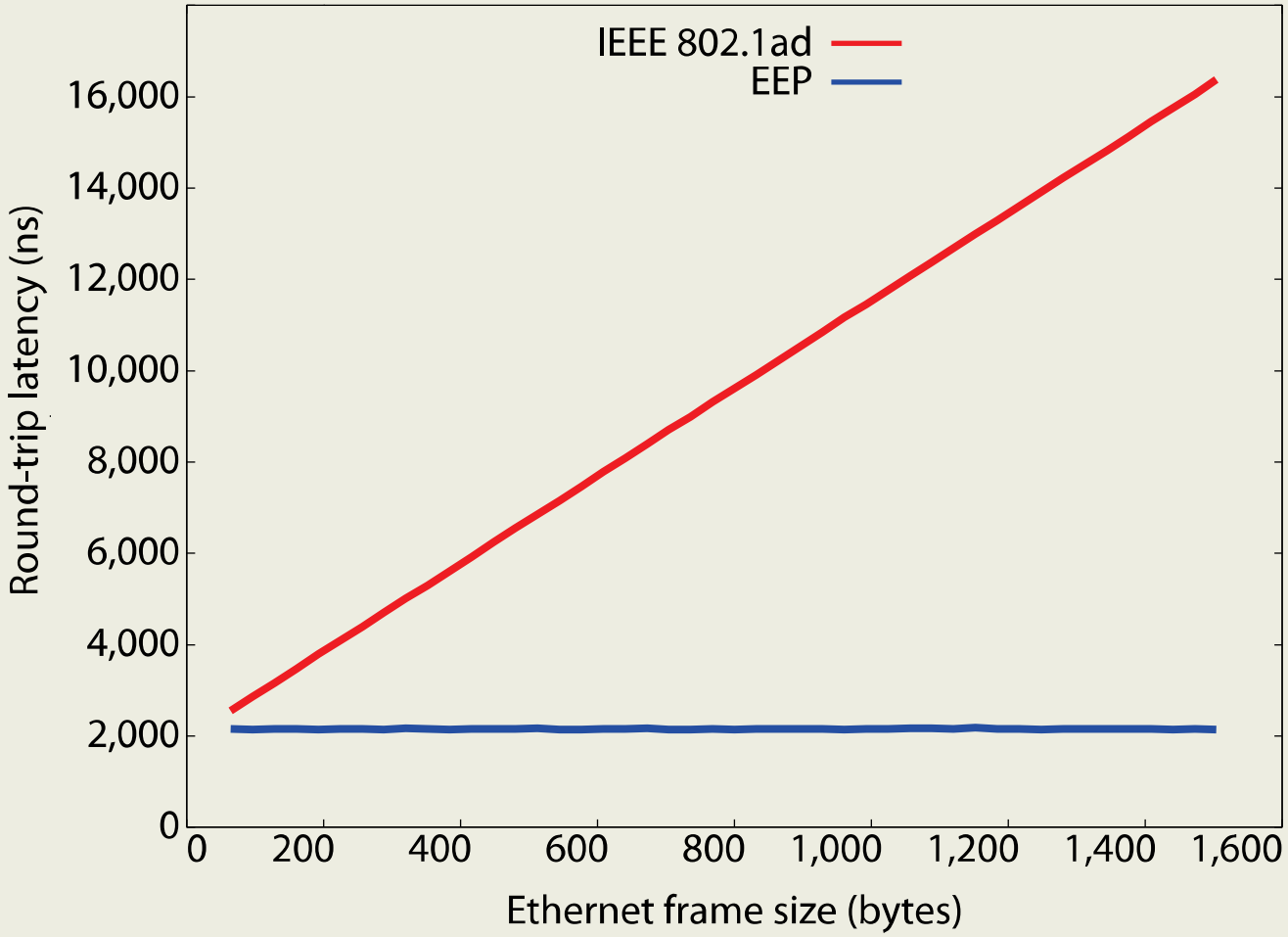
Nathan Farrington
farrington@cs.ucsd.edu

Why not use VLANs?

- Because it adds latency and requires more SRAM
- FPGA Implementation
  - VLAN tagging
  - EEP

Nathan Farrington
farrington@cs.ucsd.edu

Latency Measurements



Nathan Farrington
farrington@cs.ucsd.edu

Related Work

- M. Al-Fares, A. Loukissas, A. Vahdat. A Scalable, Commodity Data Center Network Architecture. In SIGCOMM ’08.
  - Fat trees of commodity switches, Layer 3 routing, flow scheduling

- R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat. PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric. In SIGCOMM ’09.
  - Layer 2 routing, plug-and-play configuration, fault tolerance, switch software modifications

- A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: A Scalable and Flexible Data Center Network. In SIGCOMM ’09.
  - Layer 2 routing, end-host modifications

Conclusion

- General architecture
  - Fat tree of merchant silicon switch ASICs
  - Hiding cabling complexity
  - Pods + Core
  - Custom EEP ASIC
  - Scales to 65,536 ports with 64-port ASICs
- Design of a 3,456-port 10GbE switch
- Design of the EEP ASIC