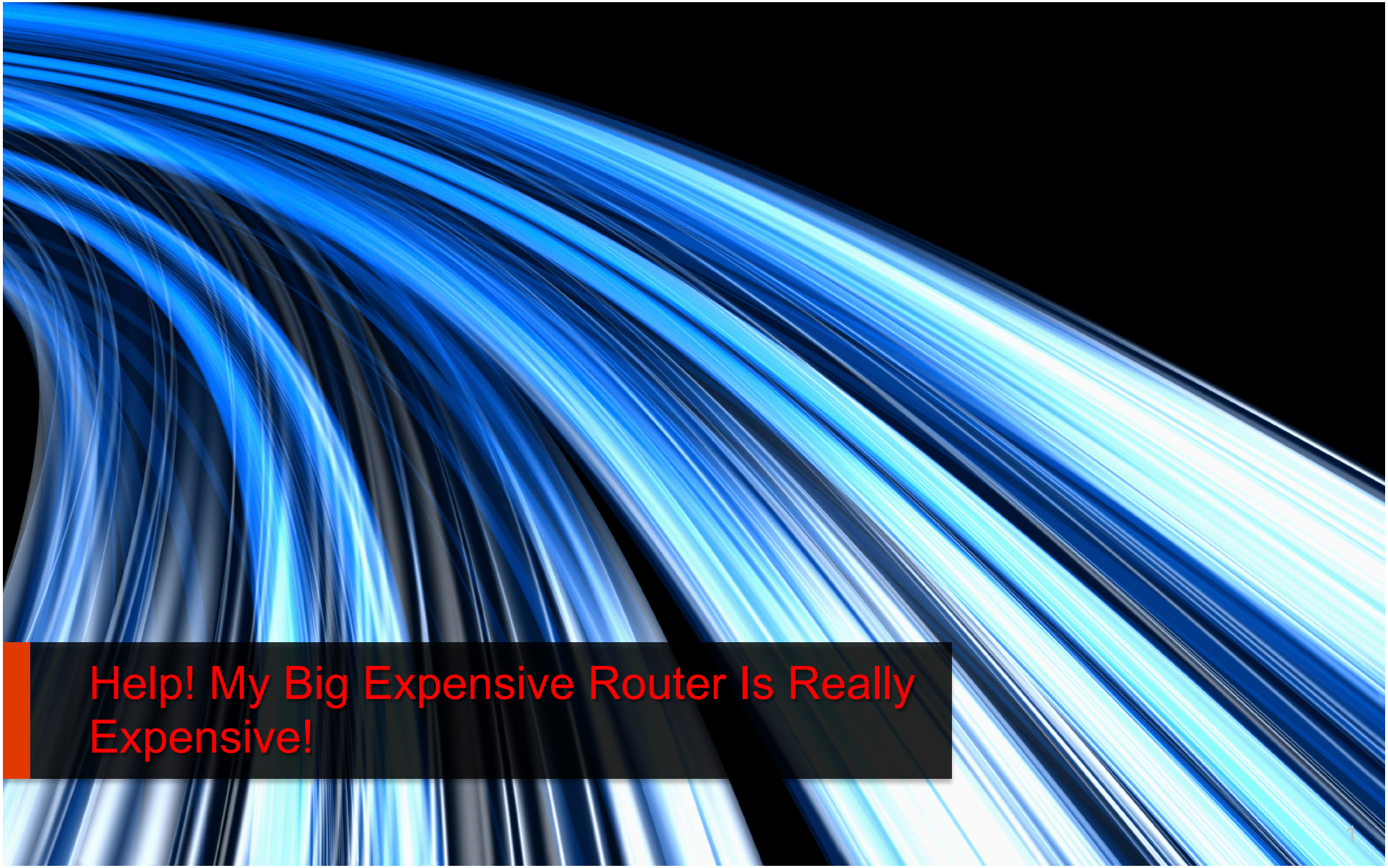

Help! My Big Expensive Router Is Really Expensive!

Panelists

David Temkin, Netflix

Craig Pierantozzi, Microsoft

Richard Turkburgen, GTT

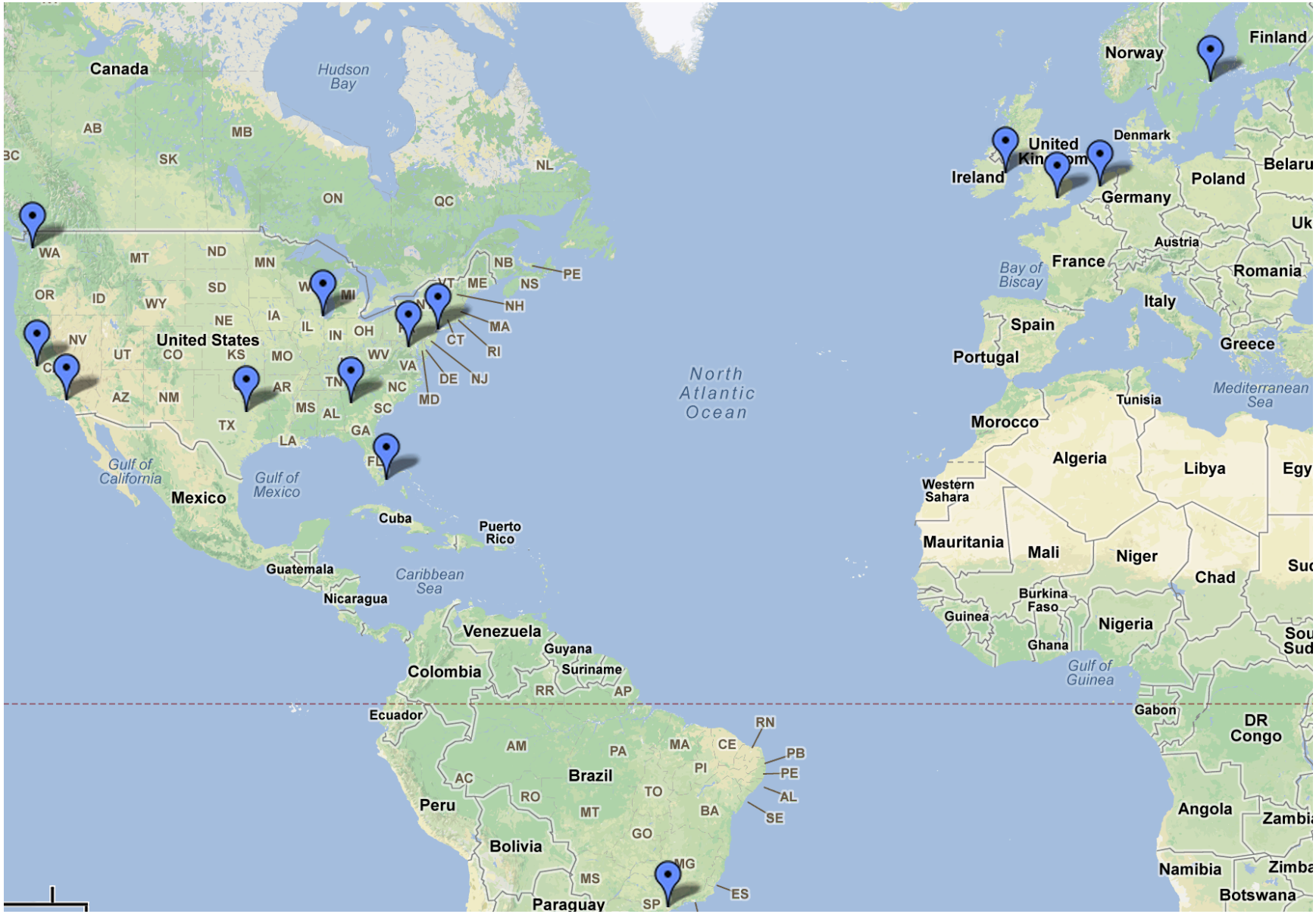
Mark Berly, Arista

Kevin Wollenweber, Cisco

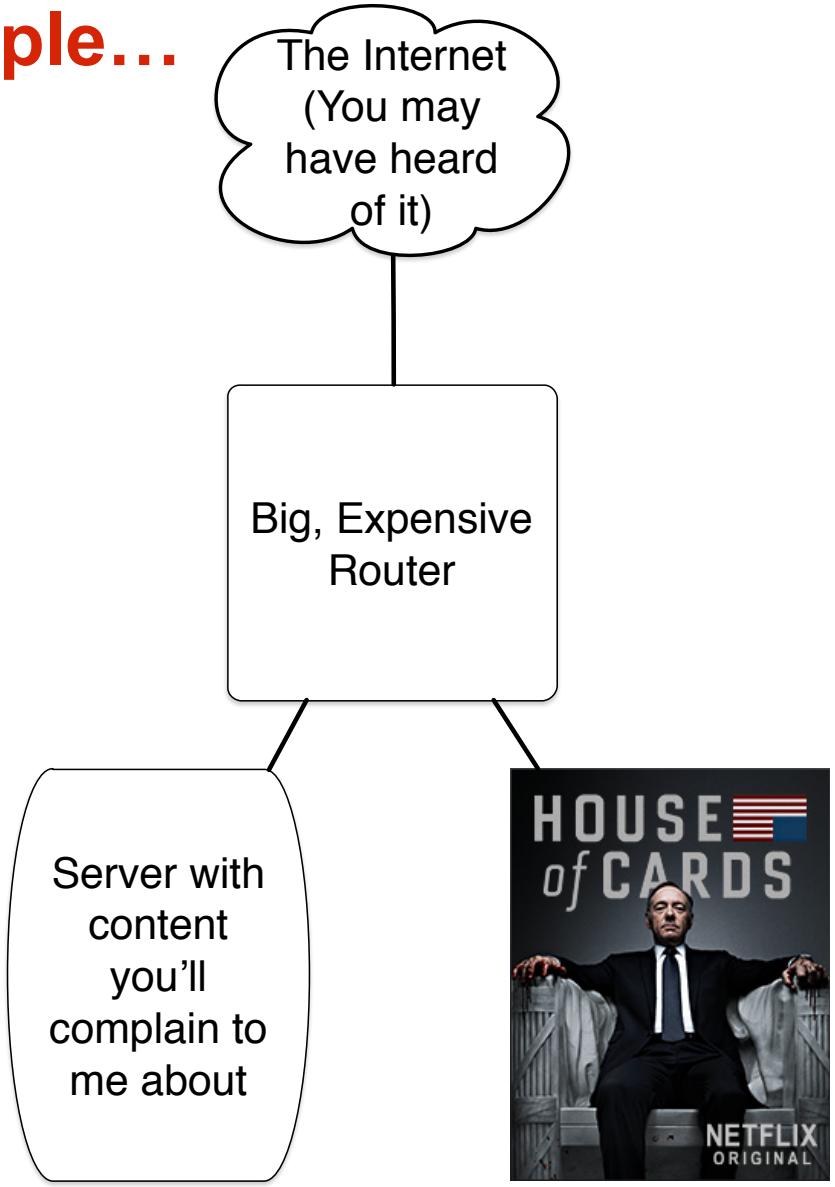
Let me be clear…

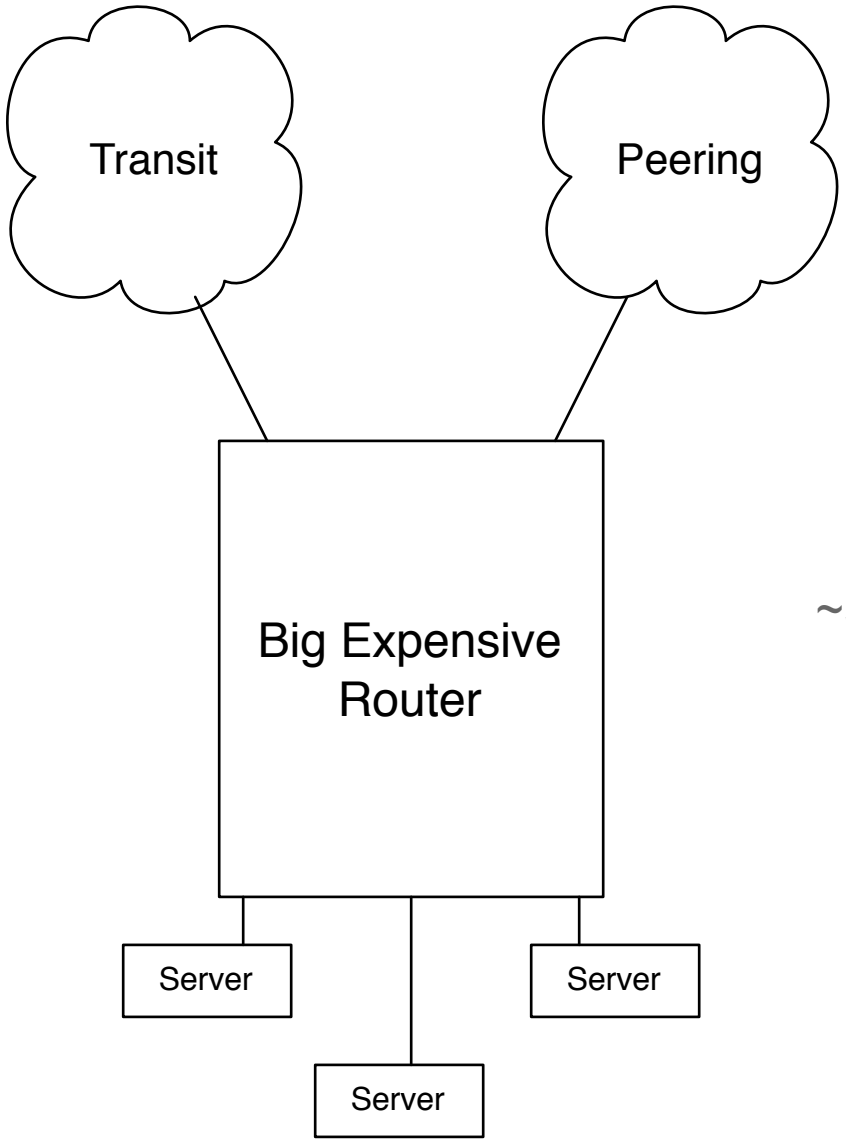
Netflix Background

- We began the Open Connect project approximately two years ago
- We are at 100% of our traffic served from the Open Connect platform
- We have 18 Terabits of network and server capacity located around the world


Canada
Hudson Bay
AB
MB
SK
NL
ON
QC
WA
MT
ND
MN
OR
ID
WY
SD
NE
IA
IL
IN
OH
NV
UT
CO
KS
MO
WV
VA
NC
SC
GA
AZ
NM
TX
AR
MS
AL
LA
TN
DE
NJ
MD
NB
PE
NS
ME
VT
NH
MA
RI
CT
United States
Gulf of California
Gulf of Mexico
Mexico
Cuba
Puerto Rico
Guatemala
Nicaragua
Caribbean Sea
North Atlantic Ocean
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
AM
RR
AP
PA
AC
RO
TO
MT
GO
MS
MG
SP
RN
CE
MA
PB
PE
AL
SE
PI
BA
ES
Finland
Norway
Ireland
United Kingdom
Denmark
Germany
Poland
Belarus
Uk
Austria
France
Bay of Biscay
Romania
Italy
Spain
Portugal
Greece
Mediterranean Sea
Tunisia
Morocco
Algeria
Libya
Egy
Western Sahara
Mauritania
Mali
Niger
Chad
Suc
Burkina Faso
Guinea
Nigeria
Ghana
Gulf of Guinea
Sou
Sud
Gabon
DR Congo
Angola
Zambia
Namibia
Zimba
Botswana

My network is very simple… (drawing to scale)





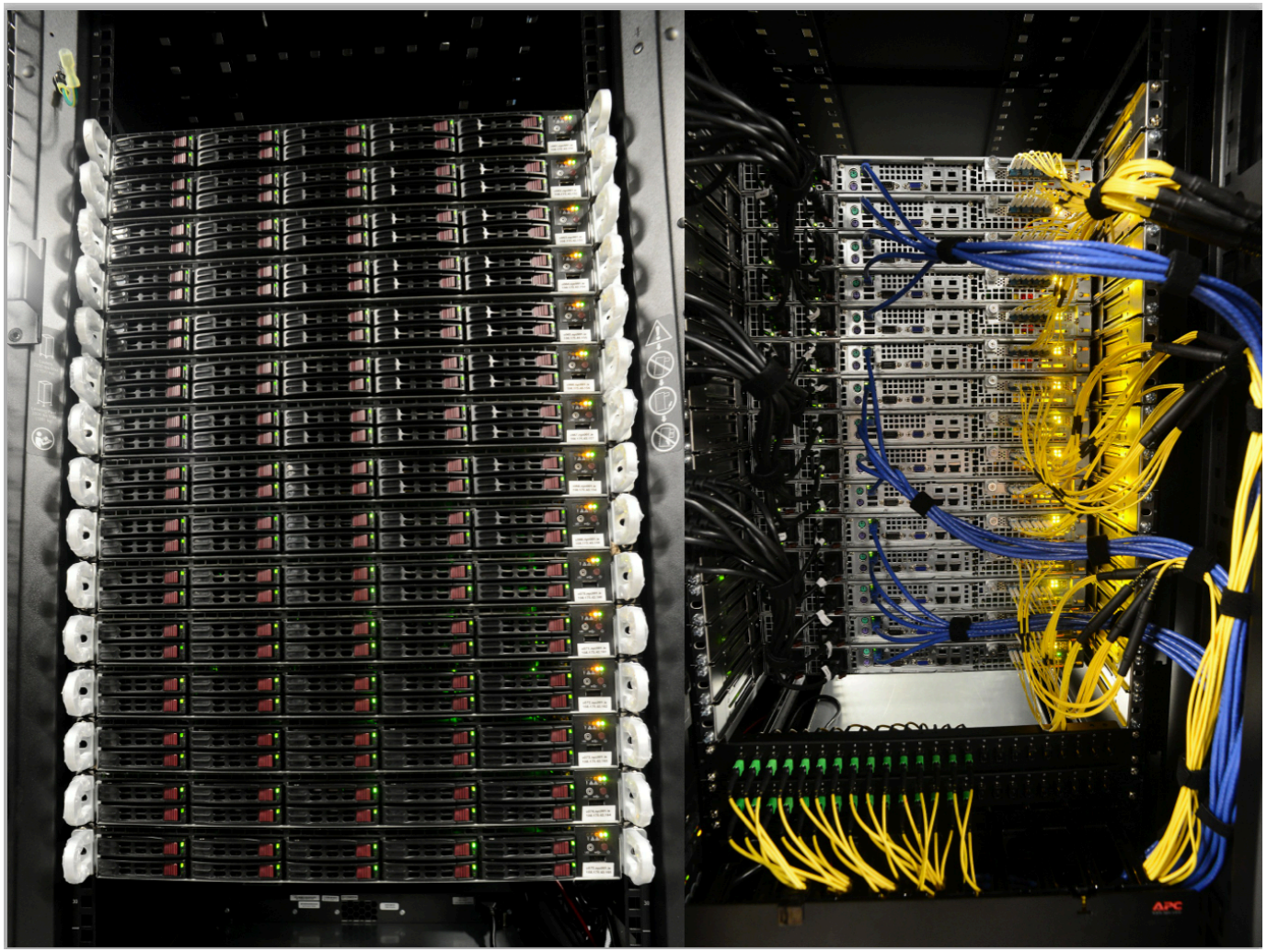
~5000 ports deployed

Enter SSD-based Open Connect Appliances..

In sites with 1+ Tbps of Netflix traffic at peak:

- 28 TB of storage per 1U system
  - Commodity SSD (< 60c/GB, Micron m500)
  - 2 TB in 2.5" form factor
- 4x 10 Gbps SFP+ NIC
- Total system power 150W per 1U

900Gbit/sec in another rack



This scale us unprecedented

Our goal is to deliver 2Tbit from 2 clusters split across 6 racks

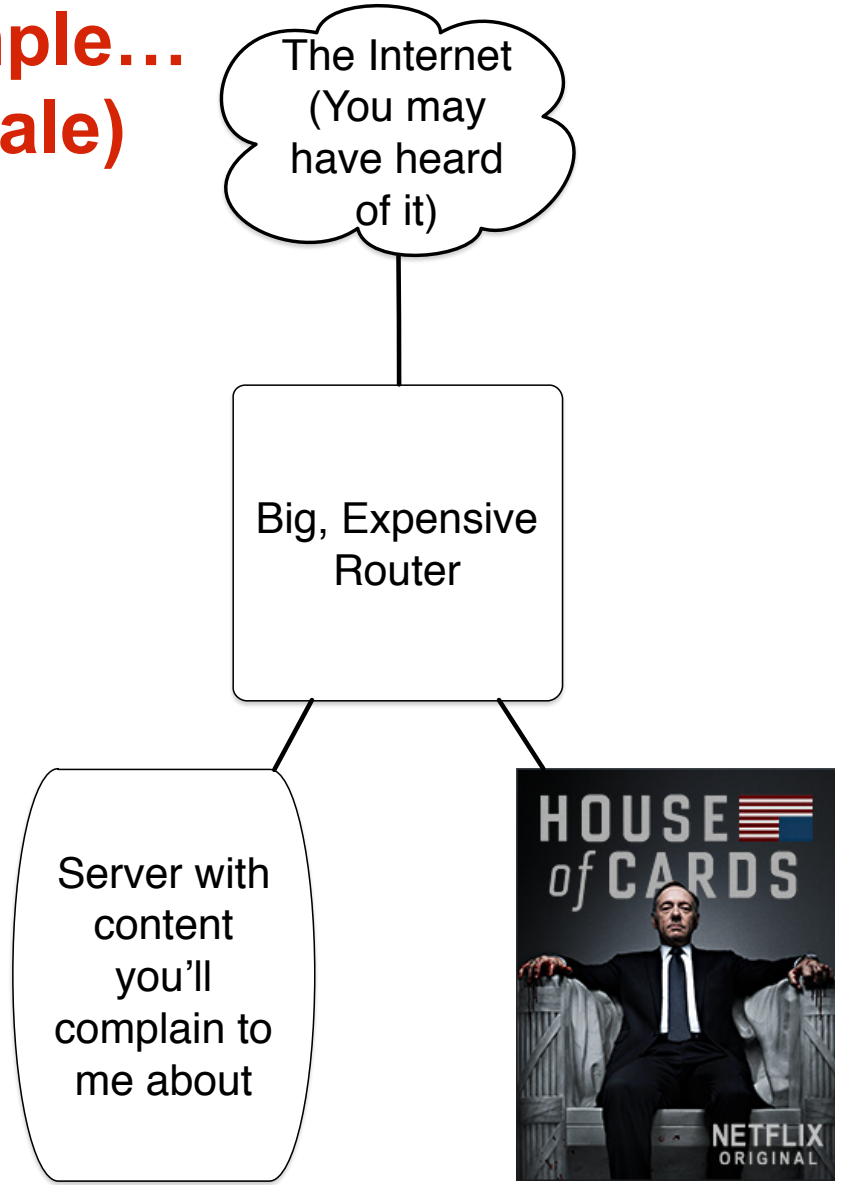
40 servers @ 2 ports each	=	80 10G ports
60 servers @ 4 ports each	=	240 10G ports
Uplinks	=	320 10G ports
	=	640 routed 10G ports (X2)

This is not optimal

- Routers are expensive
  - List price of $10-15,000 per 10GE
- They do lots of things I don't need them to
- They do a few things I absolutely need them to
- I don't run a data center
  - Little opportunity for aggregation
- I have a Layer 7 content routing engine
  - How can I apply that to Layer 3 routing decisions?
- I make my own server hardware
  - I can buy most of it at Fry's
  - I don't want to make my own network hardware

So now what?

What we do is pretty simple… (the complexity is the scale)



So how do I bring this all together?

- Routers are terrible at making routing decisions
  - We haven't improved the BGP path selection algorithm in 20 years
    - Segment Routing gives me hope
  - It's doubtful that anything comes along at the edge that factors actual path performance into the decision matrix anytime soon
  - My control plane knows whether a path is performing well or poorly
  - Most of my routing decisions are based on path performance, not path cost

Routers are really expensive

My network cost is >50% of my server cost.

- I don’t need MPLS
- I don’t need Carrier Ethernet
- I don’t need IPv6
  - (is anyone still paying attention to this presentation?)
- I don’t need L3VPN
- I posit that I don’t even need a full-scale FIB if I do this right

So, audience

- Your turn. What now?
  - Content providers: You have to feel the same pain to some extent

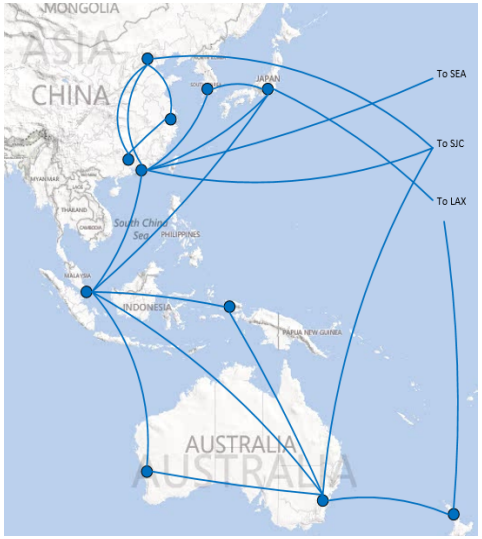
- Please refrain from using acronyms and other buzzwords to describe your solution
  - SDN
  - Big Data
  - The Cloud
  - Big Data In the Cloud

Craig Pierantozzi

Global Network Services, Microsoft Corporation

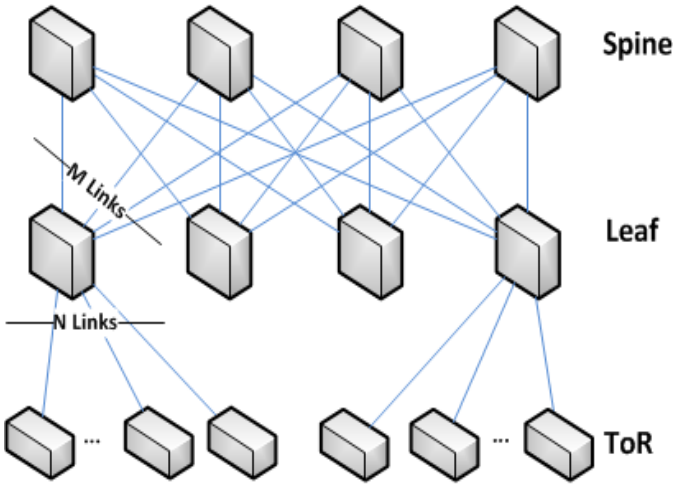
Microsoft

Microsoft Network – AS8075

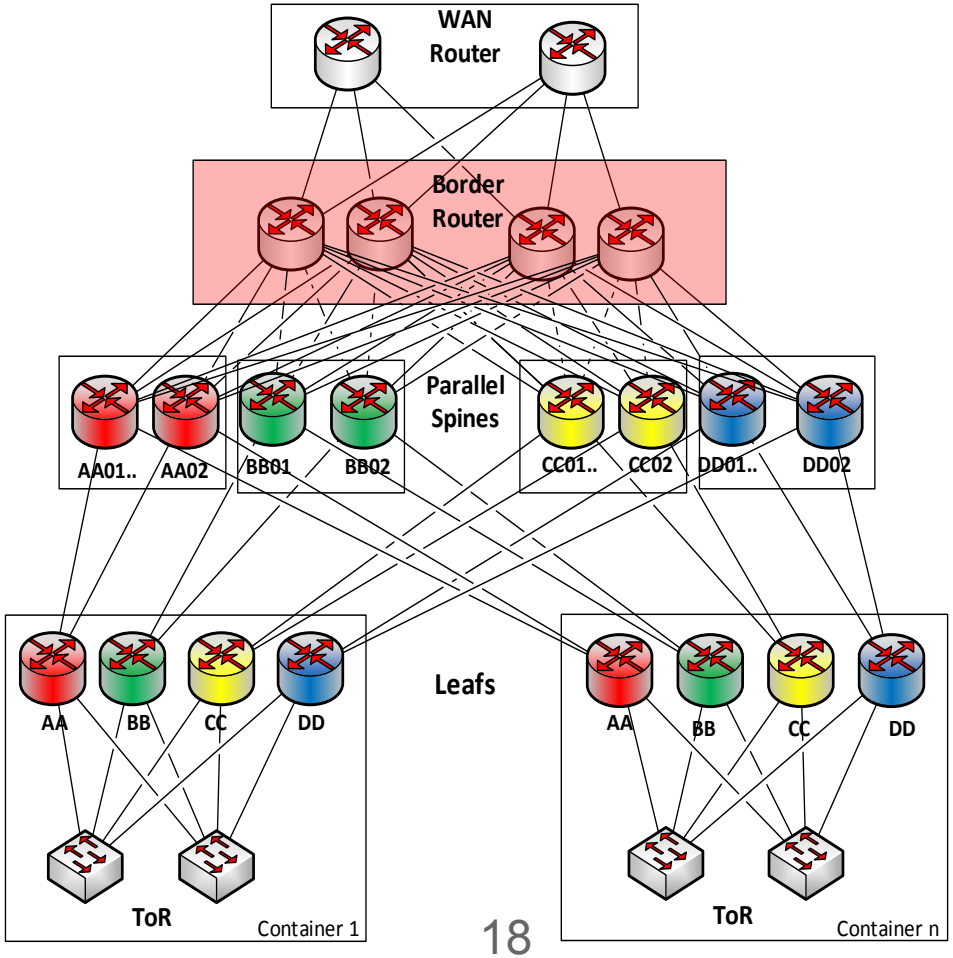


Network Design: Topology

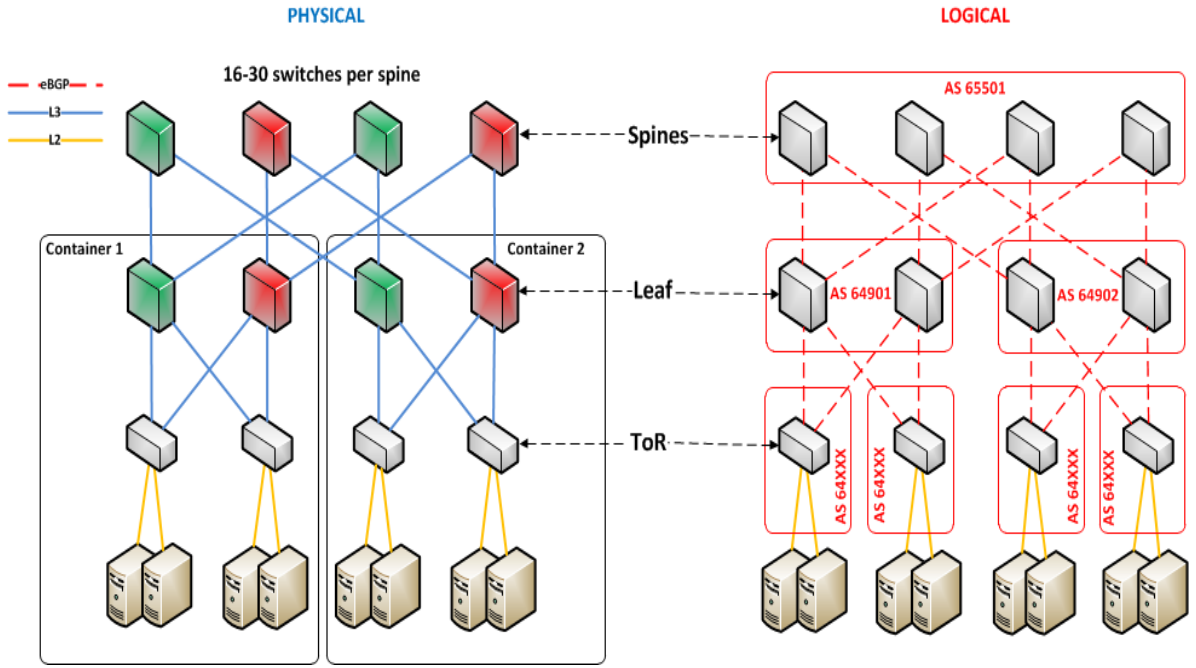
- 3-Stage Folded CLOS.
- Full bisection bandwidth (m ≥ n) .
- Horizontal Scaling (scale-out vs. scale-up)
- Service control plane managed workloads
- Viable with dense commodity hardware.
  - Build large “virtual” boxes out of small components



3-Stage Folded Clos Topology



Data Center Design: Requirements



~100 Spines

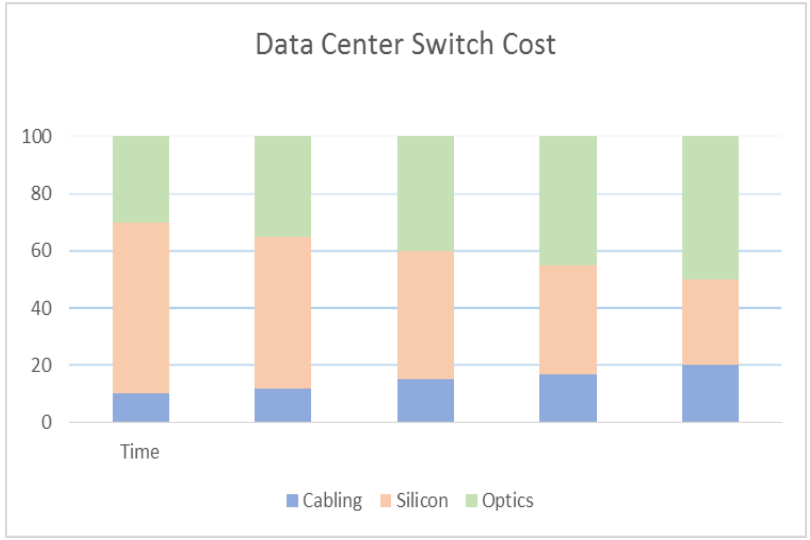
~200 Leafs

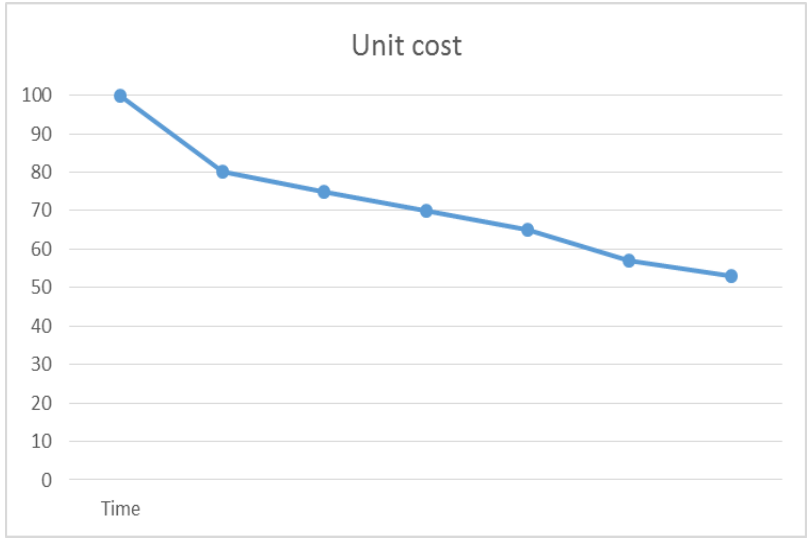
~2K ToR

~100K

- Applications:
  - Map/Reduce: Social Media, Web Index and Targeted Advertising
  - Public and Private Cloud Computing: Elastic Compute and Storage
  - Real-Time Analytics: Low latency computing leveraging distributed memory across discrete nodes
- East ←→ West traffic profile drives need for large bisectional bandwidth

Data Center Trends





- Costs shifting over time from silicon to optics and cabling infrastructure
- Switching costs continue to decline with silicon economics but not reflected in large, feature-rich “core” routing devices
- Power and cooling on large network elements continues to be a major concern

What can we do

- Optimizations around service resiliency in software and not solely reliant on network resiliency
  - End to end optimization of WAN, DC networks as well as server resources
  - Equipment failure is an operating condition and software opportunity
  - Removes requirements such as fast boot, graceful-restart etc.
- Increase scale at a lower cost
  - Cheap label switching
  - Lower cost, integrated WDM solutions
  - Fewer or no protocols on the network elements
  - Lower power consumption and trade off density/capacity
  - Hardware ubiquity - Leverage same chipsets across network layers
  - “Smaller, stupider, cheaper. But not too small or too stupid.”

Richard A Steenbergen <ras@gtt.net>
GTT Communications, Inc.

How To Make Big Expensive Routers Less Expensive (but probably still big)

Ethernet CAN BE Really Freaking Cheap

- “Simple” hardware is actually really, really cheap.
  - Half of the planet now makes a Broadcom derived box.
  - Now with 640Gbps+ in 1U for hundreds of dollars.

So Why Is My Expensive Router Still Expensive?

- So why does my MX960 blade cost a couple of orders of magnitude more than a 1U HP switch?
  - It's not ***ALL*** about your router vendor trying to bilk you.
  - Some of these are complex technical problems to solve.
  - But then again, sometimes those complex technical problems never needed to exist in the first place.
- You probably want your big expensive router to do some fancy things that a small cheap router doesn't.
  - But is that because it CAN'T?
  - Or is there some other reason that it doesn't?

Example: Core MPLS Switching

- Consider the case of Core MPLS switching.
  - MPLS switching is actually **really** freaking easy in HW.
    - This is one of the reasons it was invented in the first place.
  - Simple exact-match lookups, very little state to maintain, very simple headers to parse, etc.
  - In fact, most commodity hardware can do it today.
- So where is my cheap MPLS-only core platform?
  - The Juniper PTX is barely any better than the Juniper MX.
  - Why do I need a million-dollar “core” box to do something a thousand-dollar box CAN do today?

Core MPLS Switching

- The answer: They don't have the software.
- It turns out MPLS is actually pretty complex in SW.
  - Signaling, bandwidth reservations, fast reroute, etc.
  - Only the incumbent router vendors actually understand it.
    - Which is no surprise, considering they wrote it in the first place.
  - And none of the guys who know how to make cheap hardware know the first thing about RSVP/LDP/etc.
- So why would Cisco/Juniper make a 1U MPLS core?
  - They'd only be cannibalizing from their own "carrier" biz.
  - And they face no competition in this market segment.

SDN Big Data Cloud Hadoop!!!!

But Before We Fire Up The Hype Machine

- Software Defined Networking?
  - I've been using software to define my network for years.
  - If you don't, you must have a lot of unmanaged switches.
- Is SDN just another FAD?
  - As Avi would say, "Funding Augmentation Device".
  - Or is there actually a ray of hope in there somewhere?
- But First…
  - What did you actually spend your money on?

Some Fundamental Truths

Blowing Your Mind Without CDP or MAC Addresses

- So what did you just spend all that money on?
  - If the hardware needed to forward 1.21 jiggabits is actually commodity…
  - That means what you're actually buying is software.
- Software is hard.
  - Routing protocols, CLIs, network management platforms, and feature after feature after feature after feature…
- Software is what you're actually buying.
  - The hardware is just a delivery vehicle, so you don't feel so bad for spending millions on invisible electrons.
  - But the software is what you actually care about.

So Why Do We Care About SDN?

- It turns out that some people are good at one thing but not another (try not to act shocked).
  - Commodity silicon manufacturers can produce hardware that forwards untold numbers of packets for pennies, but they can't write routing protocols or CLIs.
  - Somewhere, there exists people who know how to write routing protocols, but who can't fab an ASIC.
- SDN is about the threat of getting these two groups of people together in a way that ***isn't*** an incumbent router vendor.

Take My Money!!!

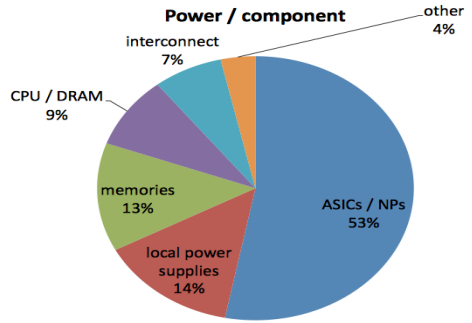
- So is there an SDN “product” out that that will revolutionize anything?
  - If it exists, I haven’t seen it yet.
- But as a concept, SDN has a lot of merit.
- And maybe some day soon, your big expensive router will be a little bit less expensive.

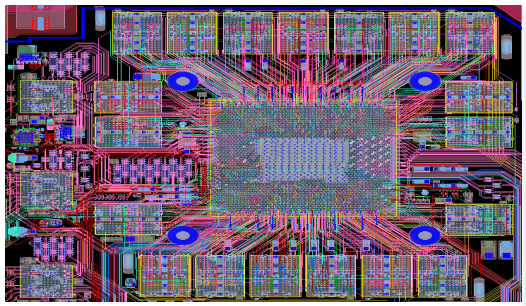
Send questions, comments, complaints to: Richard A Steenbergen <ras@gtt.net>

Challenges in High End Routing

Technology Optimization Point

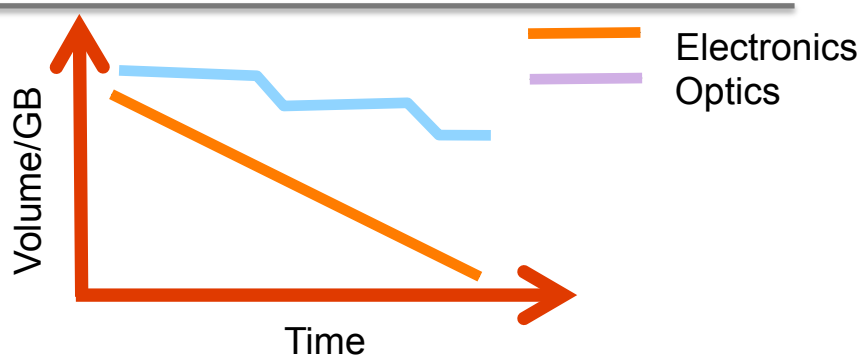
Power Consumption and Cooling





- BW vs PPS
- FIB Scale
- Queuing
- Buffering

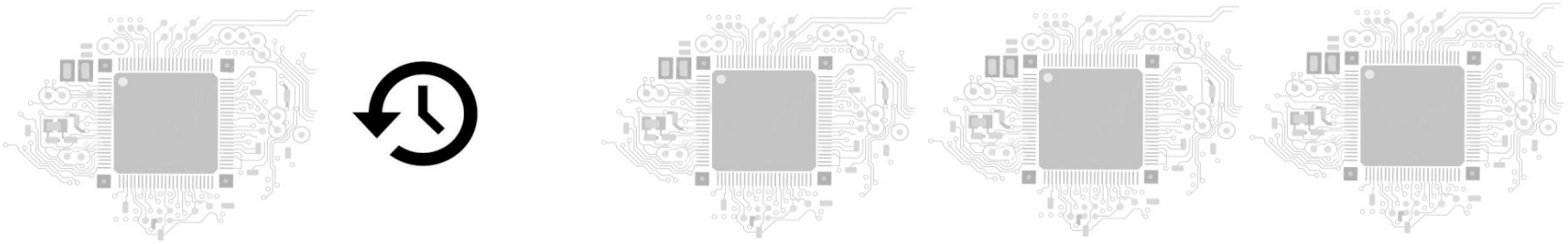
ASIC Technology and Moore's Law



Optics Size and Cost

Power Consumption Dilemma

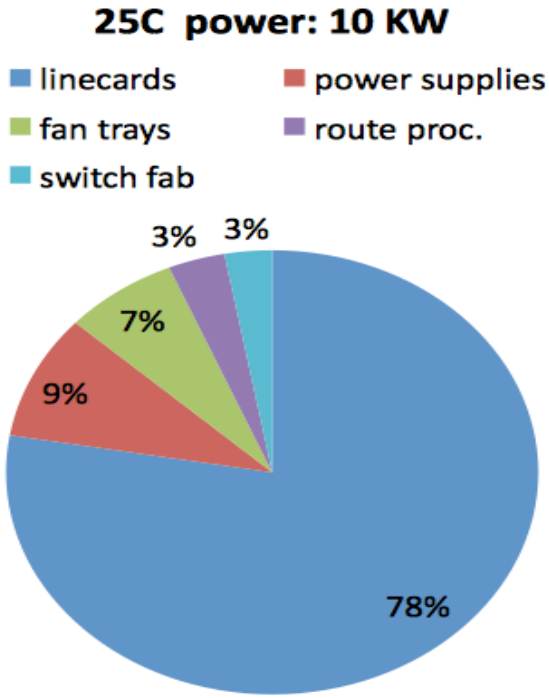
- Silicon capability has grown more rapidly than the decrease of power usage.
- This creates a paradigm where unit power decreases, but total power increases.

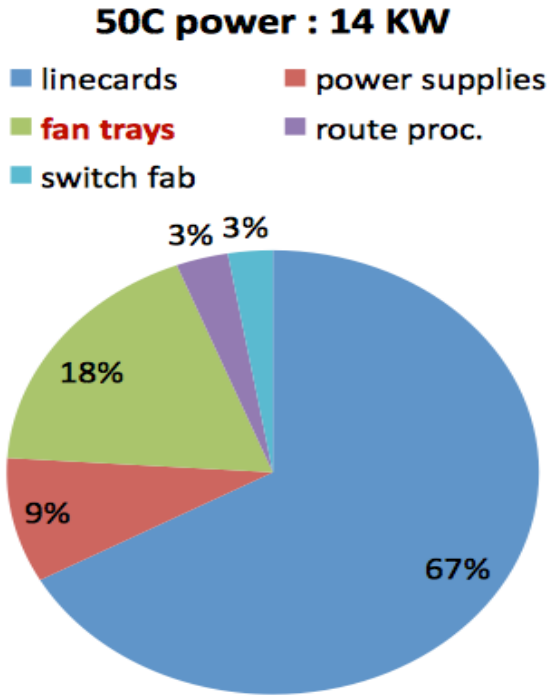


- Because bandwidth in exploding, although the silicon is faster and consumes less power, we use more of it which increases the overall power consumption of devices.

Device Power Consumption
Example

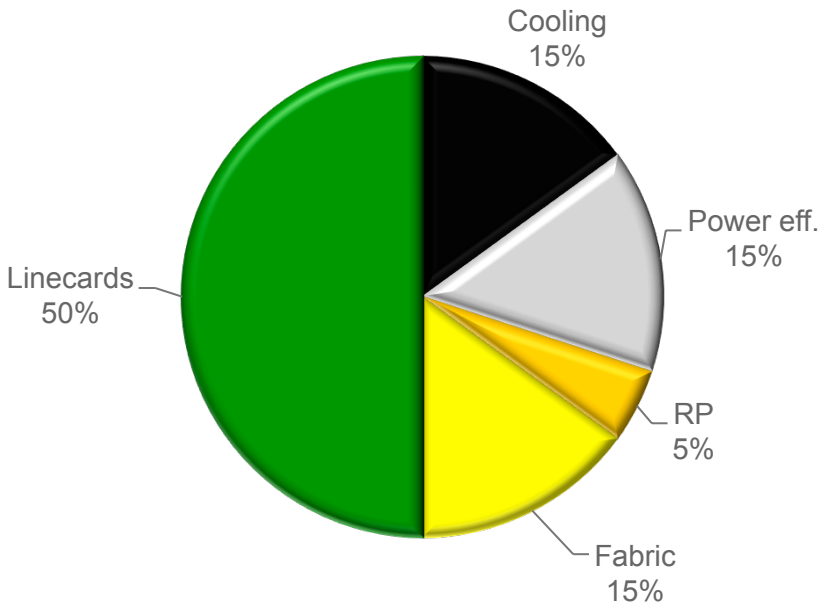
- The increase of fan trays at 50 C represents a significant shift in the amount of power draw from the line card power.





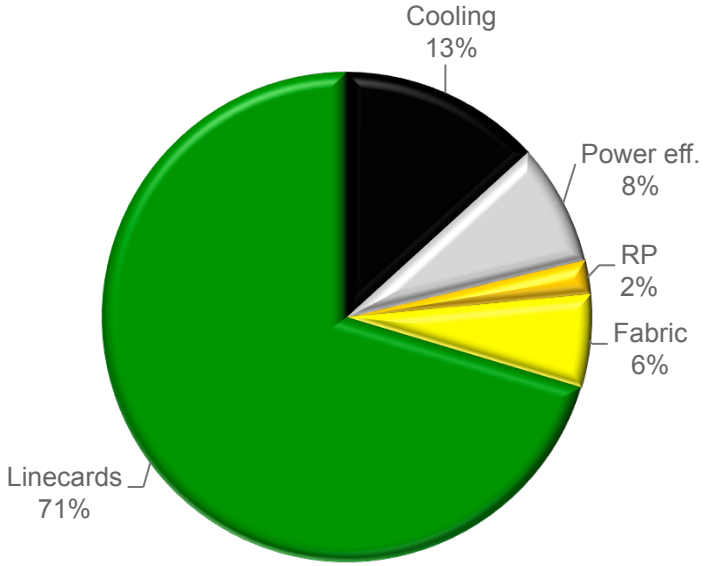
Power Utilization – Historical Perspective
One Decade Forwarding Increase x 12.5

CRS-1(2004)



Total ~12KW at 50C

NCS – 6008 (50C)



Total ~13.4KW at 50C

Cisco nPower Silicon

New Cisco Network Processor Family

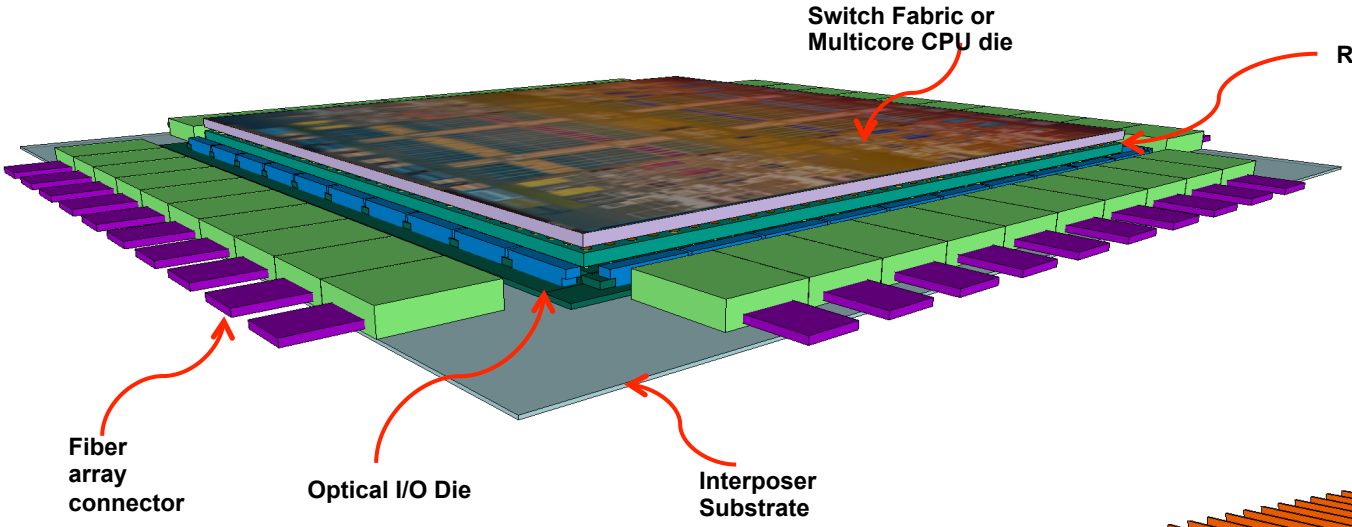
- Functionality and Optimization
  - High-scale, multiservice IPv4/IPv6/MPLS forwarding
  - LSR for label switching
  - Integrated Ethernet MACs and OTN framers
  - Advanced Memory Management System
  - ZPL/ZTL smart support
- Flexibility
  - Fully programmable with 336 on-chip multi-threaded packet processors
  - Support for a wide range of Ethernet interfaces (10 GE, 40 GE, 100 GE)
- Integration and Low Power Consumption
  - High integration (MAC, NPUs, OTN) provides low power consumption and reduced footprint

4 Billion Transistors

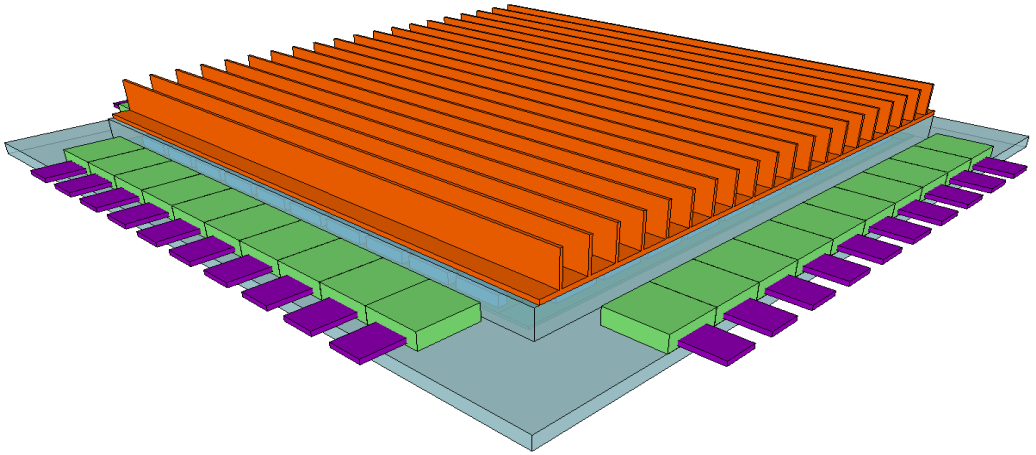
Relative Cost of Optics & Electronics

Optics ~25 years behind Electronics – level of integration, manufacturability etc & Gap is widening due to orders of magnitude larger investment in infra-structure

Si Photonics for Efficient Integration



Only 1 full duplex SERDES required



3D stacking of Opto-electronics ICs with passive alignment to external WGs

Potential Datapath Evolution