NETFLIX

Open Connect: Starting from a Greenfield (a mostly Layer 0 talk)

Dave Temkin
06/01/2015

The story

- It all started with a discussion...

- After much debate, the “Netflix CDN” team was launched in mid-2011

- We built caches. They were red. Now they serve lots of Terabits and we call our CDN “Open Connect”

A decision had to be made...

- **Do we use caches inside of ISP networks that want them..**
  - And then continue to use third party CDNs

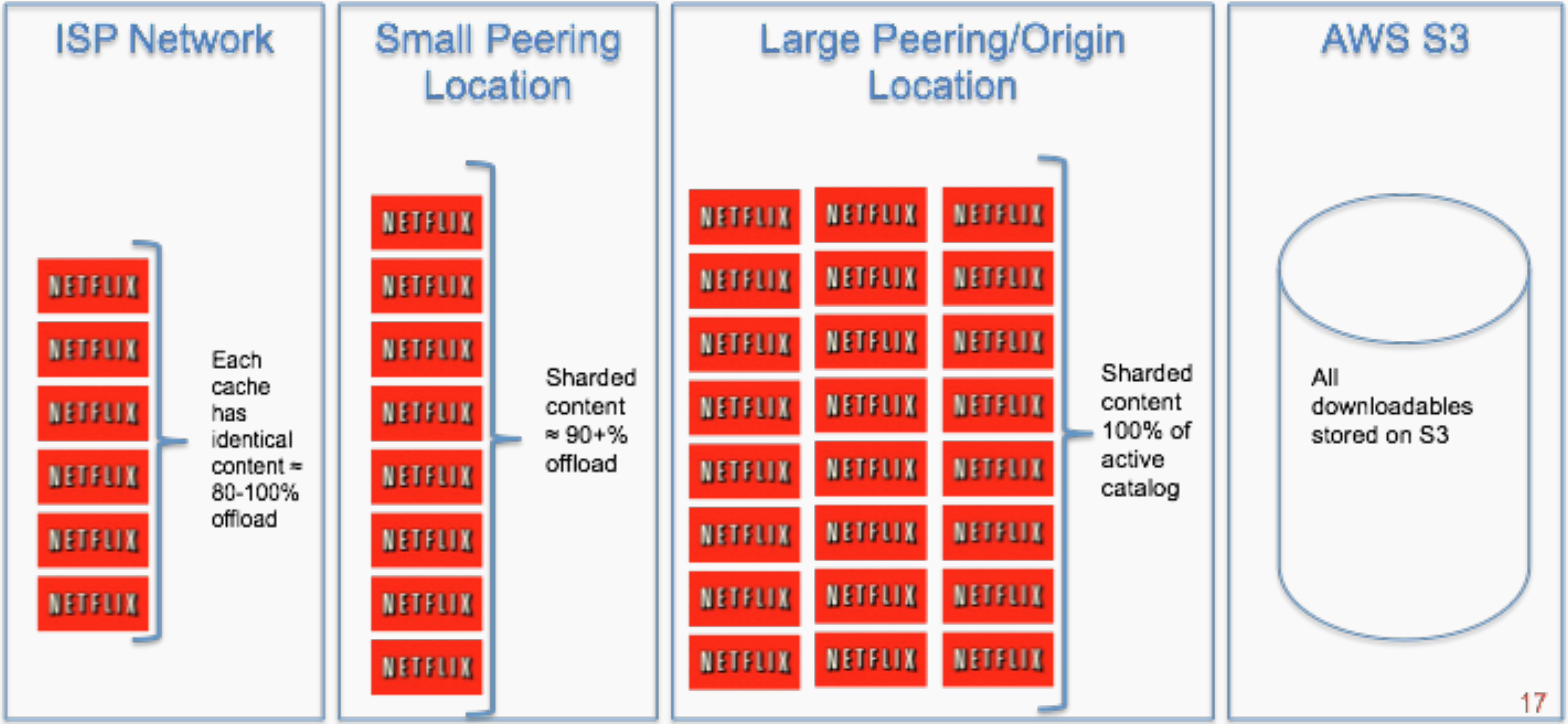
- **Or build a fully functional, standalone CDN**

OpenConnect Appliance

I'm afraid I have no choice but to sell you all for scientific experiments

A single purpose system

- Deliver video at the highest quality possible while allowing operators the ability to manage the traffic on their network



ISP Network

Each cache has identical content ≈ 80-100% offload

Small Peering Location

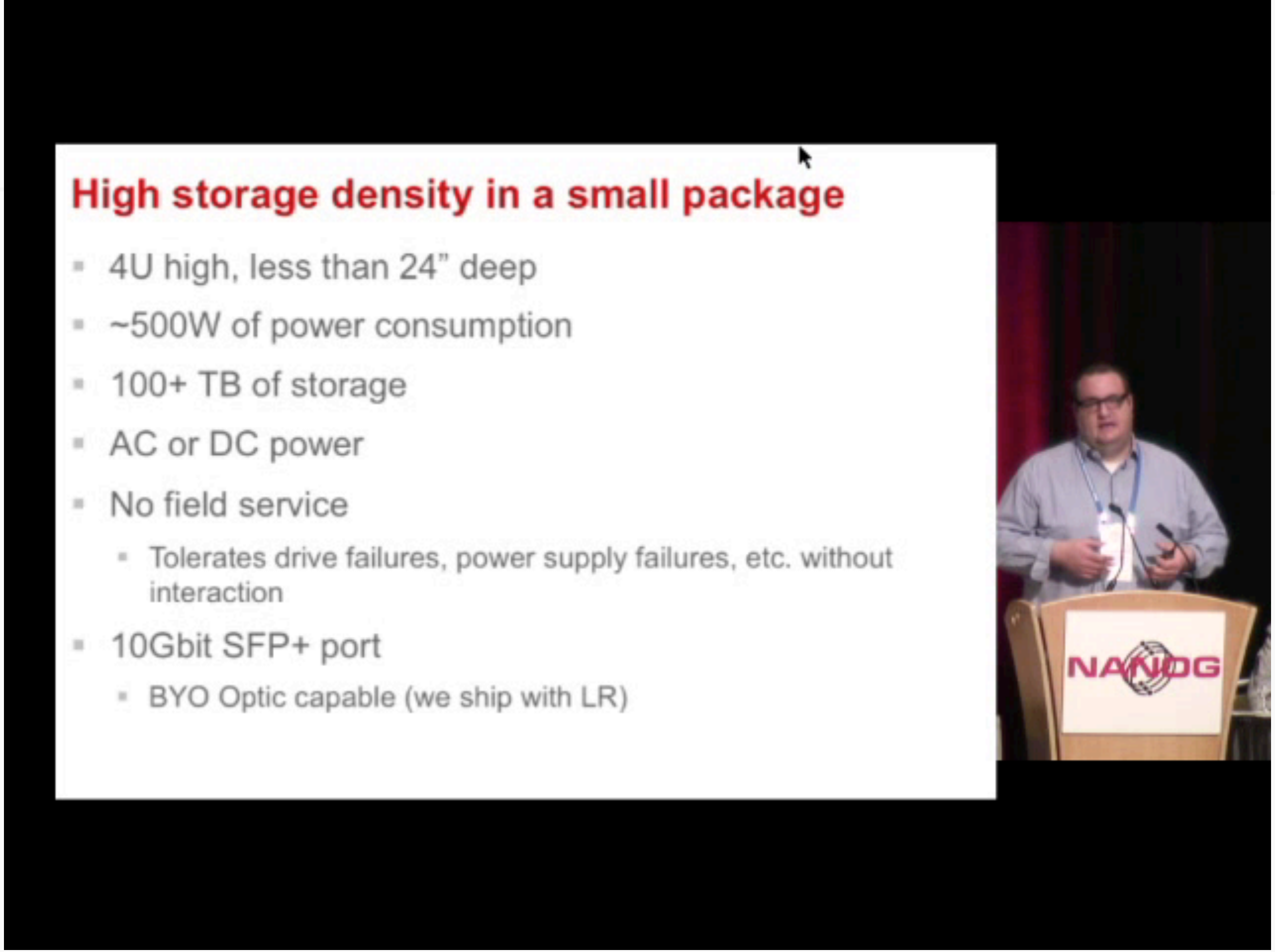
Sharded content ≈ 90+% offload

Large Peering/Origin Location

Sharded content 100% of active catalog

AWS S3

All downloadables stored on S3



High storage density in a small package

- 4U high, less than 24” deep
- ~500W of power consumption
- 100+ TB of storage
- AC or DC power
- No field service
  - Tolerates drive failures, power supply failures, etc. without interaction
- 10Gbit SFP+ port
  - BYO Optic capable (we ship with LR)

2012

Now

Storage Appliance

- Still 4U high
- ~550 watts
- 288 TB of storage
- 2x 10G ports
- 20Gbit/s delivery

and...

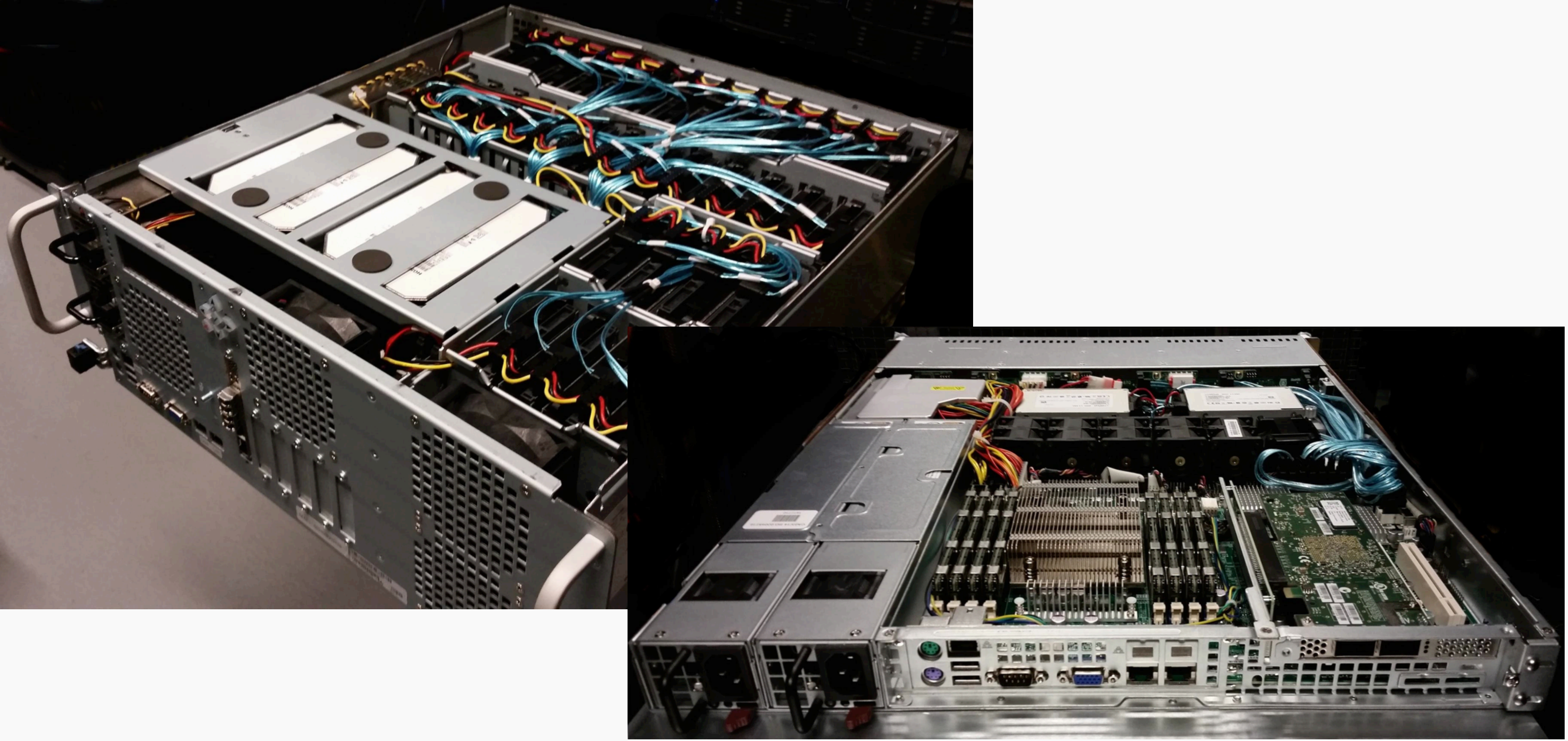
Flash Appliance

- 1U
- ~175 watts
- 24 TB of flash
- 2x 40G ports
- 40Gbit/s delivery

Cache Types

- **We have two main types of Netflix Caches**
  - Rev H: 36 8TB spinning drives, up to 20Gbit
    - Used for catalog offload
  - Rev I: 24 1TB SSD’s, up to 40Gbit
    - Used for high speed popular content serving

- **Our mantra has been to use the same hardware that we would expect an ISP to install in their network**
  - Consistent software stack



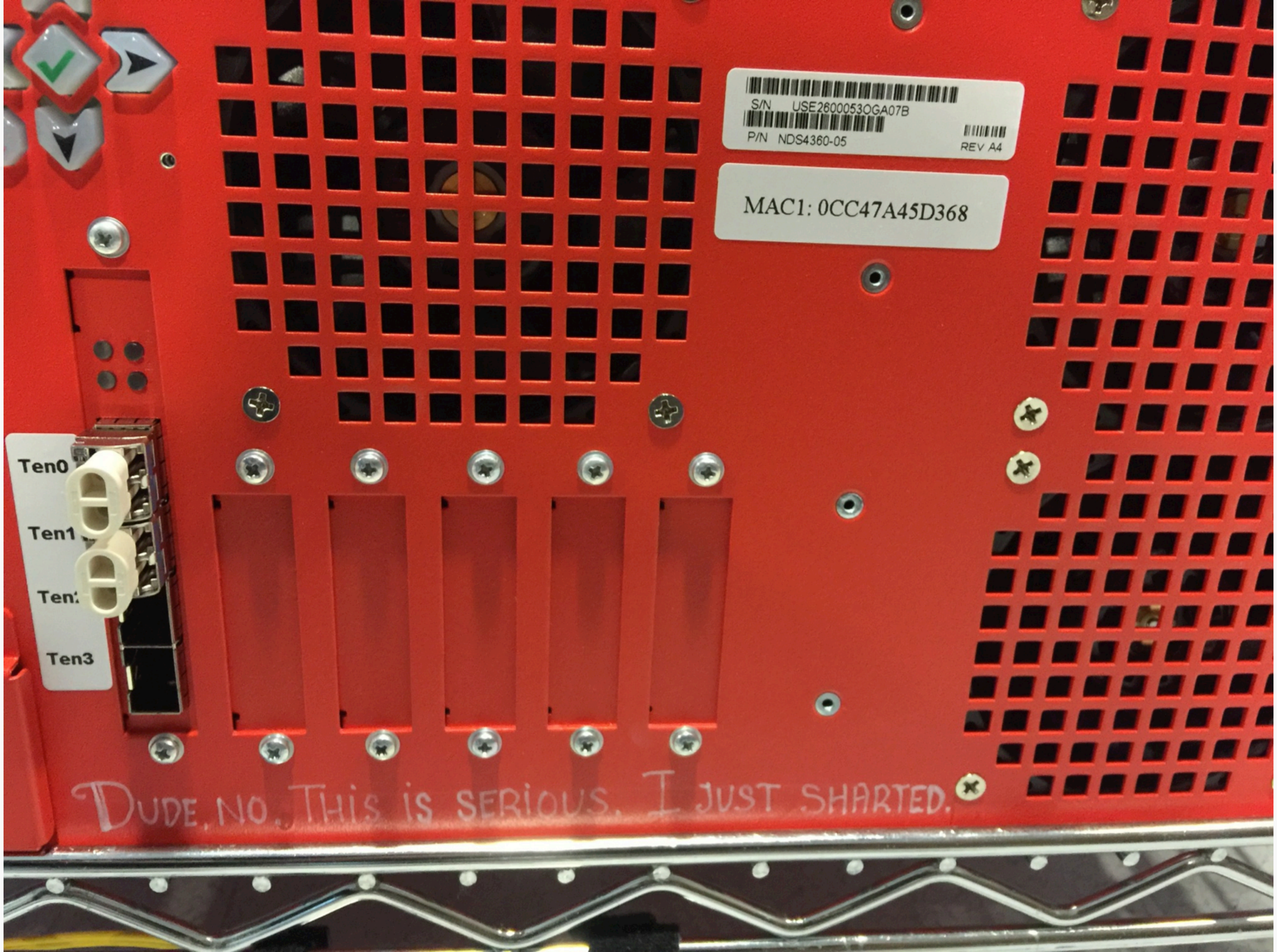
Left: Storage OCA
Right: Offload OCA

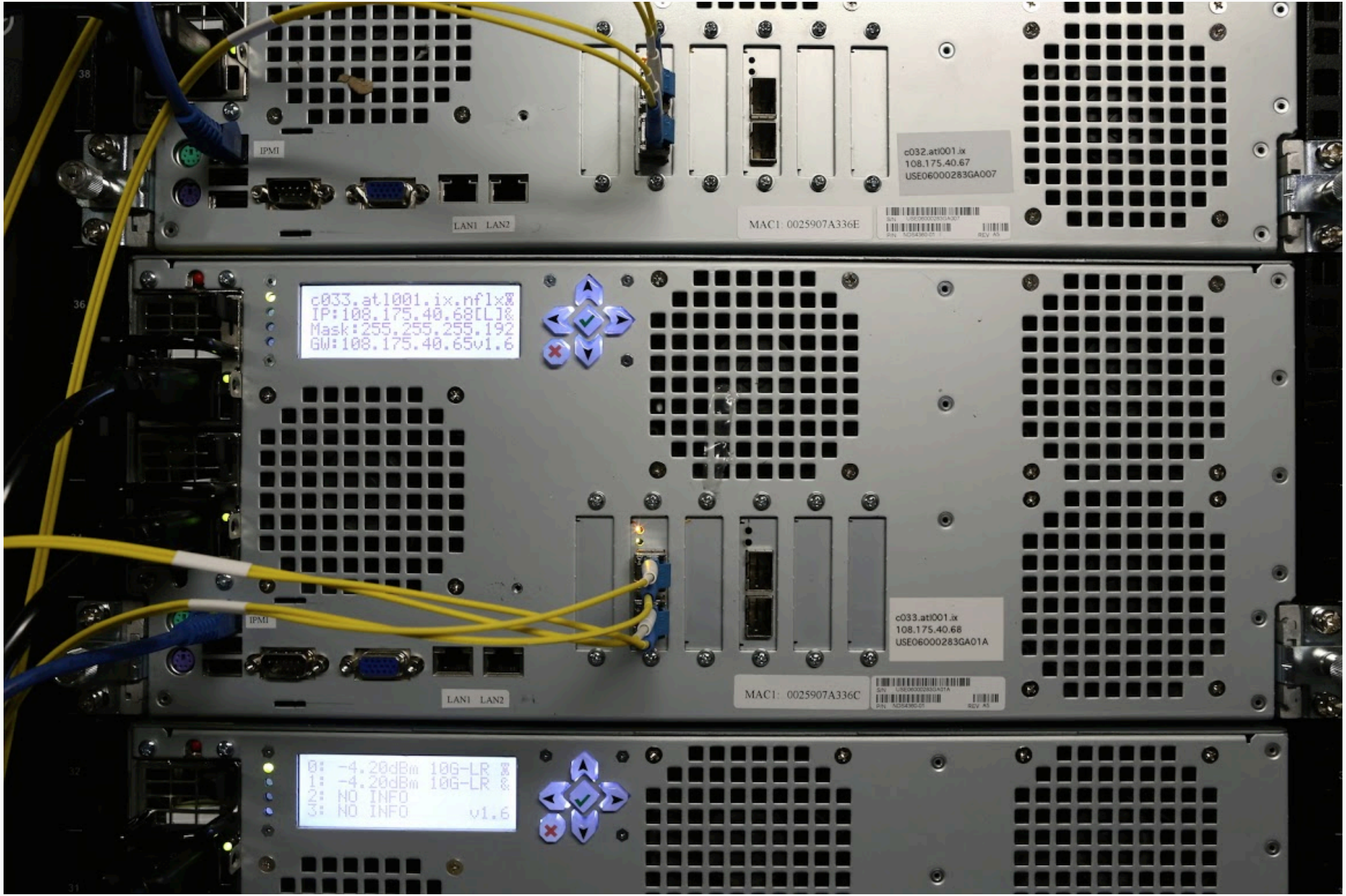
Power Utilization and Footprint

- **Rev H: 560 watts**
  - .31 watts per megabit
- **Rev I: 250 watts**
  - .006 watts per megabit

Our standard deployment has been 10 Rev H's per rack and 30 Rev I's, or a 5.6kW/7.5kW deployment

OpenConnect Appliance


S/N USE26000530GA07B
P/N NDS4360-05
REV A4
MAC1: 0CC47A45D368
Ten0
Ten1
Ten2
Ten3
DUDE, NO. THIS IS SERIOUS. I JUST SHARTED.


IPMI
LAN1
LAN2
c032.atl001.ix
108.175.40.67
USE06000283GA007
MAC1: 0025907A336E
c033.atl001.ix.nflx
IP:108.175.40.68[L]
Mask:255.255.255.192
GW:108.175.40.65v1.6
IPMI
LAN1
LAN2
c033.atl001.ix
108.175.40.68
USE06000283GA01A
MAC1: 0025907A336C
0: -4.20dBm 10G-LR
1: -4.20dBm 10G-LR
2: NO INFO
3: NO INFO v1.6

Today’s focus

Layer Zero: Peering and Interconnection Locations

A big challenge

- **Goal was to be off third party CDNs by 2013**
    - At a reasonable cost compared to what we were paying CDNs
        - (Remember, Vertical Integration)
    - At better quality
    - Scalable for global growth

Our first attempt… ~24G of capacity

To ~4Tbits of edge capacity in 4 racks...

2.4T of serving capacity at a “small” peering location
35kW of power

A note on site selection

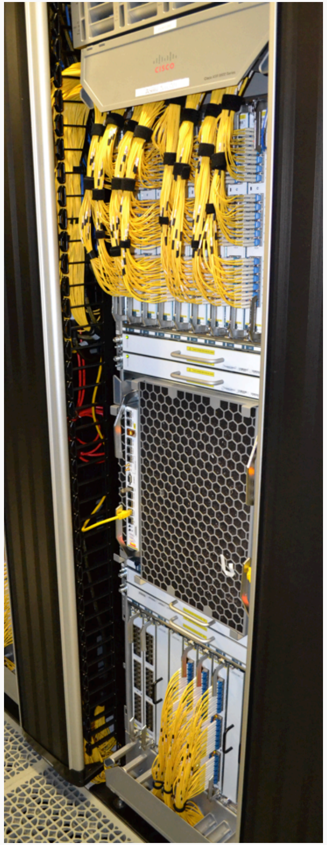
- Pick the most popular site in a metro
- Can’t find space and power?
  - Maybe the second most popular

Challenges

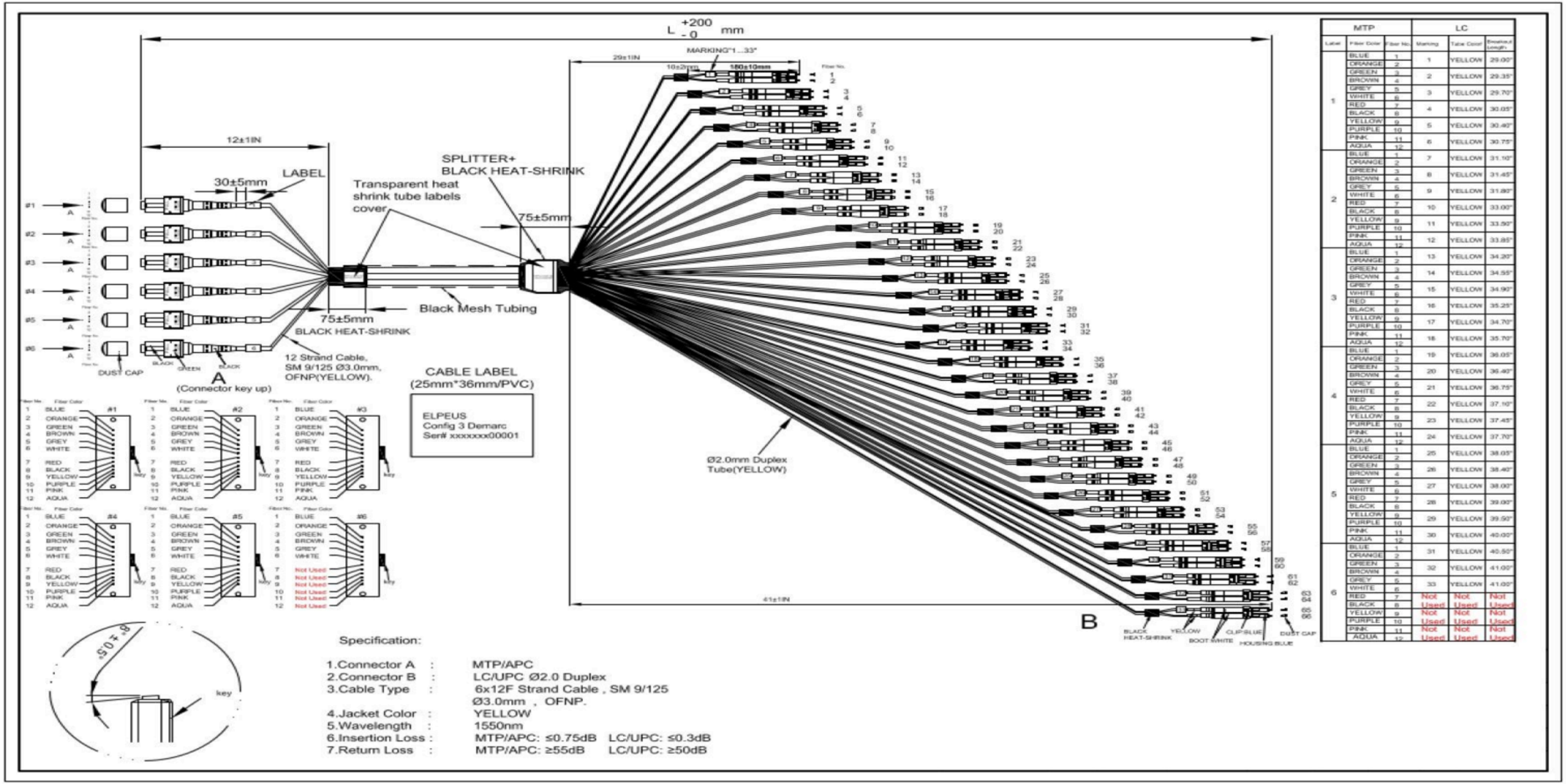
- **5.6kW for Rev H's is relatively easy to get**
  - Fully loaded, we ask for a 6.5kW per rack footprint
- **7.5kW is stretching the limits of most legacy data centers**
- **Rev H's are space limited (most sites can accommodate a 42U rack, so we engineer to that)**
- **Rev I's are power limited**
  - Would like to go from 250 watts to 300 watts
  - End up with a 9kW rack
    - Easy when you own the data center, not so easy when you lease

Cabling

- **Something taken for granted for many years**
  - “Call up a contractor, have them run some fiber, plug it all in”
- **Not so simple anymore**
- **10G in the data center is still the most affordable**
  - 40G mostly there
  - 100G still prohibitive beyond interconnect



This doesn’t happen by accident… but takes an hour to do.



CABLE LABEL
(25mm*36mm/PVC)

ELPEUS
Config 3 Demarc
Ser# xxxxxxx00001

MTP			LC		
Label	Fiber Color	Fiber No.	Marking	Tube Color	Breakout Length
1	BLUE	1	1	YELLOW	29.00"
	ORANGE	2			
	GREEN	3	2	YELLOW	29.35"
	BROWN	4			
	GREY	5	3	YELLOW	29.70"
	WHITE	6			
	RED	7	4	YELLOW	30.05"
	BLACK	8			
	YELLOW	9	5	YELLOW	30.40"
	PURPLE	10			
	PINK	11	6	YELLOW	30.75"
	AQUA	12			
2	BLUE	1	7	YELLOW	31.10"
	ORANGE	2			
	GREEN	3	8	YELLOW	31.45"
	BROWN	4			
	GREY	5	9	YELLOW	31.80"
	WHITE	6			
	RED	7	10	YELLOW	33.00"
	BLACK	8			
	YELLOW	9	11	YELLOW	33.50"
	PURPLE	10			
	PINK	11	12	YELLOW	33.85"
	AQUA	12			
3	BLUE	1	13	YELLOW	34.20"
	ORANGE	2			
	GREEN	3	14	YELLOW	34.55"
	BROWN	4			
	GREY	5	15	YELLOW	34.90"
	WHITE	6			
	RED	7	16	YELLOW	35.25"
	BLACK	8			
	YELLOW	9	17	YELLOW	34.70"
	PURPLE	10			
	PINK	11	18	YELLOW	35.70"
	AQUA	12			
4	BLUE	1	19	YELLOW	36.05"
	ORANGE	2			
	GREEN	3	20	YELLOW	36.40"
	BROWN	4			
	GREY	5	21	YELLOW	36.75"
	WHITE	6			
	RED	7	22	YELLOW	37.10"
	BLACK	8			
	YELLOW	9	23	YELLOW	37.45"
	PURPLE	10			
	PINK	11	24	YELLOW	37.70"
	AQUA	12			
5	BLUE	1	25	YELLOW	38.05"
	ORANGE	2			
	GREEN	3	26	YELLOW	38.40"
	BROWN	4			
	GREY	5	27	YELLOW	38.00"
	WHITE	6			
	RED	7	28	YELLOW	39.00"
	BLACK	8			
	YELLOW	9	29	YELLOW	39.50"
	PURPLE	10			
	PINK	11	30	YELLOW	40.00"
	AQUA	12			
6	BLUE	1	31	YELLOW	40.50"
	ORANGE	2			
	GREEN	3	32	YELLOW	41.00"
	BROWN	4			
	GREY	5	33	YELLOW	41.00"
	WHITE	6			
	RED	7	Not Used	Not Used	Not Used
	BLACK	8			
	YELLOW	9	Not Used	Not Used	Not Used
	PURPLE	10			
	PINK	11	Not Used	Not Used	Not Used
	AQUA	12			

Specification:

1.Connector A : MTP/APC
2.Connector B : LC/UPC Ø2.0 Duplex
3.Cable Type : 6x12F Strand Cable , SM 9/125 Ø3.0mm , OFNP.
4.Jacket Color : YELLOW
5.Wavelength : 1550nm
6.Insertion Loss : MTP/APC: ≤0.75dB LC/UPC: ≤0.3dB
7.Return Loss : MTP/APC: ≥55dB LC/UPC: ≥50dB

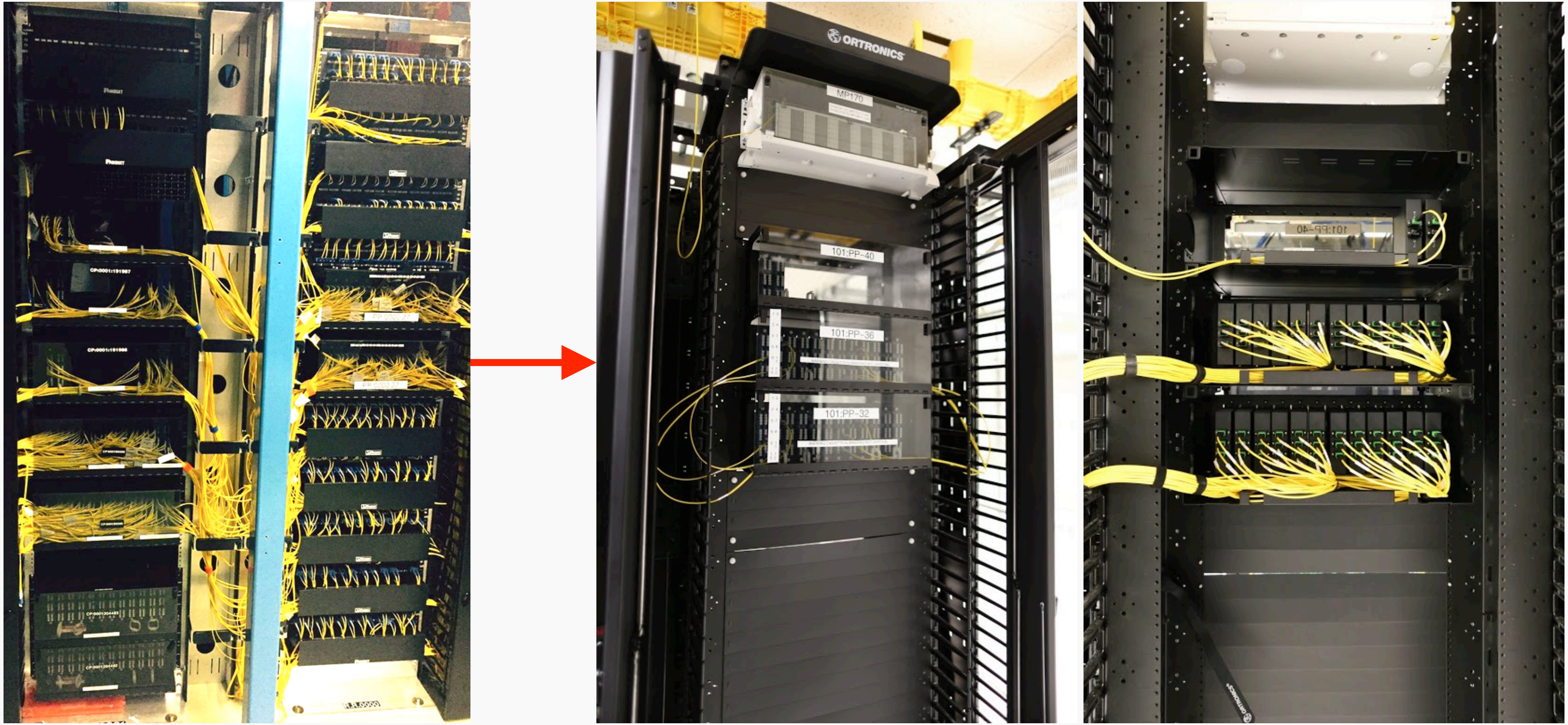
1 of 18 custom cable types...

This next slide was originally going to have a witty GIS’ed image for “Cable Porn”

But I had turned SafeSearch off and quickly abandoned that idea

Moving on...

- Somehow we need to get the data out of here..



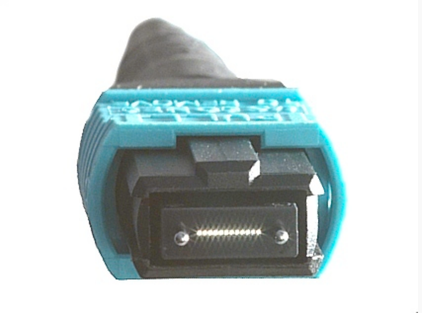
Right: 1440 cross connects per rack
288x10G or 100G = 2.88 or 28.8 Terabits
Left: 192 cross connects per rack

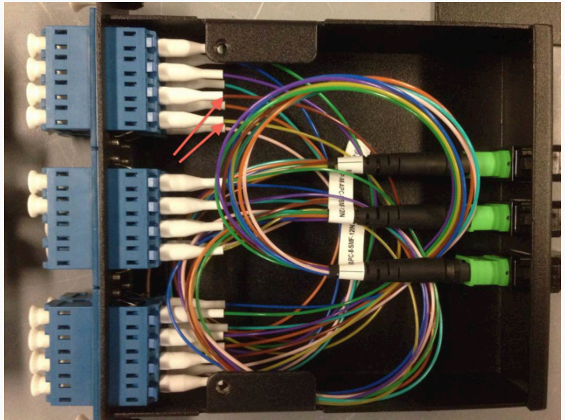
Rapid Deployment

- Having custom cables to fit our deployment allows for rapid implementation
- Only levers are lengths and types
- Allows for a complete solution to be shipped to site
- If everything goes well, we can have a multi-Terabit site online in less than two days
- Never underestimate the value of not having $colo vendors touch anything other than your patch panels

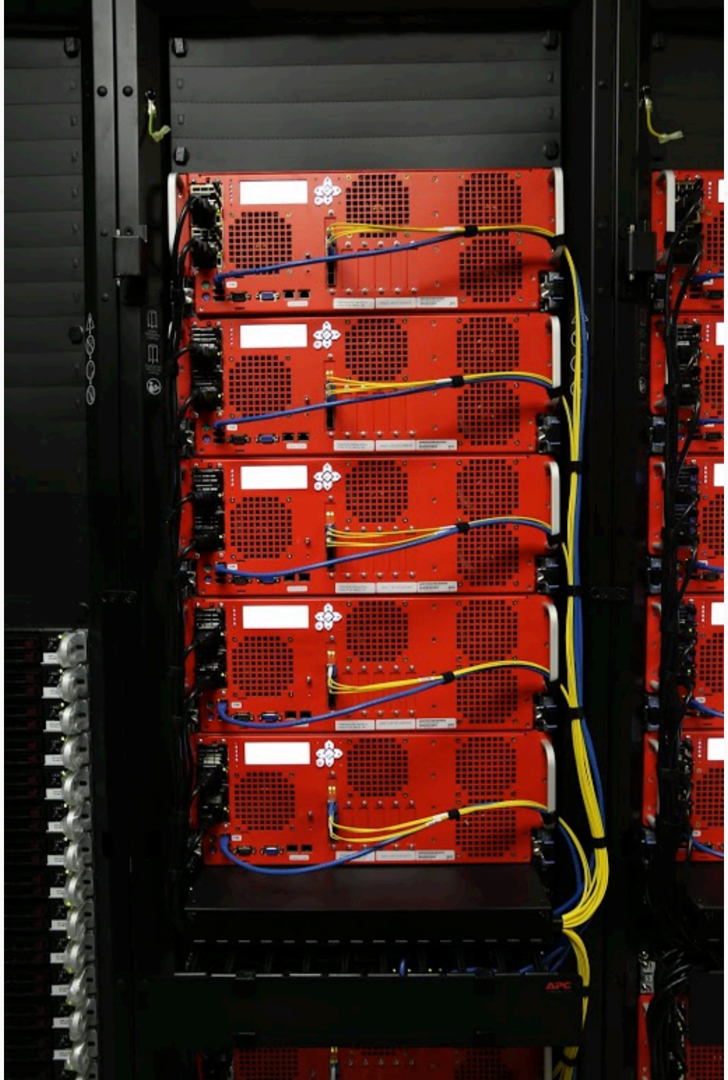
Leveraging MTP

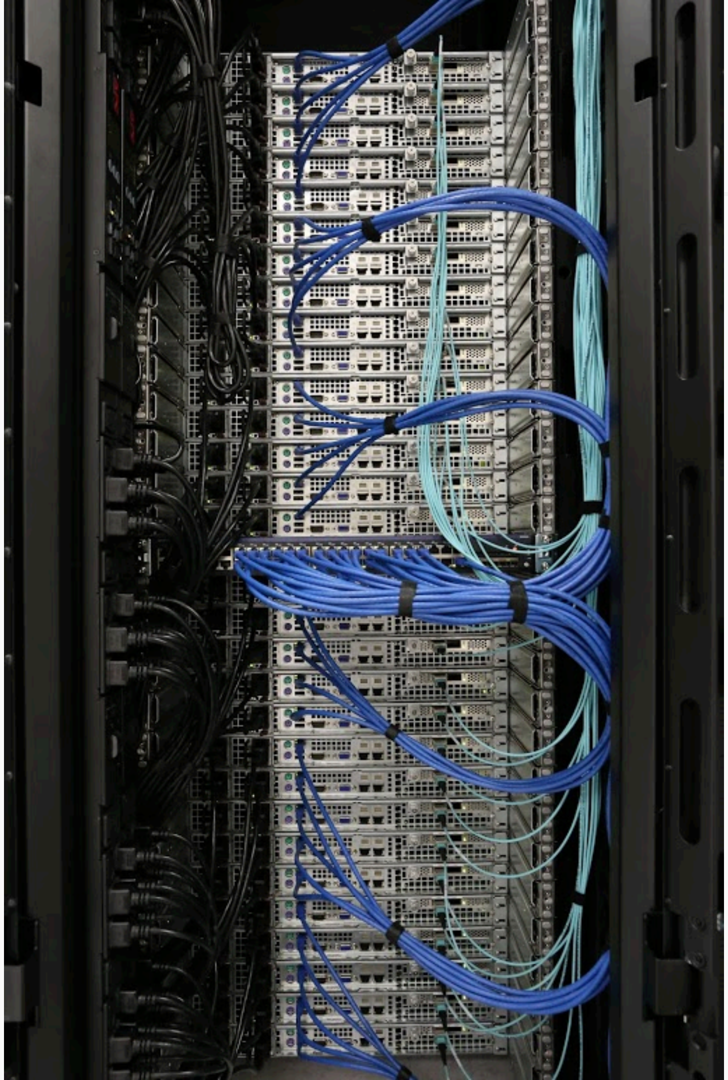
- **MTP connectors on everything..**
  - Servers (40G)
  - Switches (QSFP), including PLR/PSM 4x10G
  - Patch panels
  - Cassettes
- **Allows for rapid field deployment**
- **Reconfigurable - direct path to 100G**
- **Inexpensive**

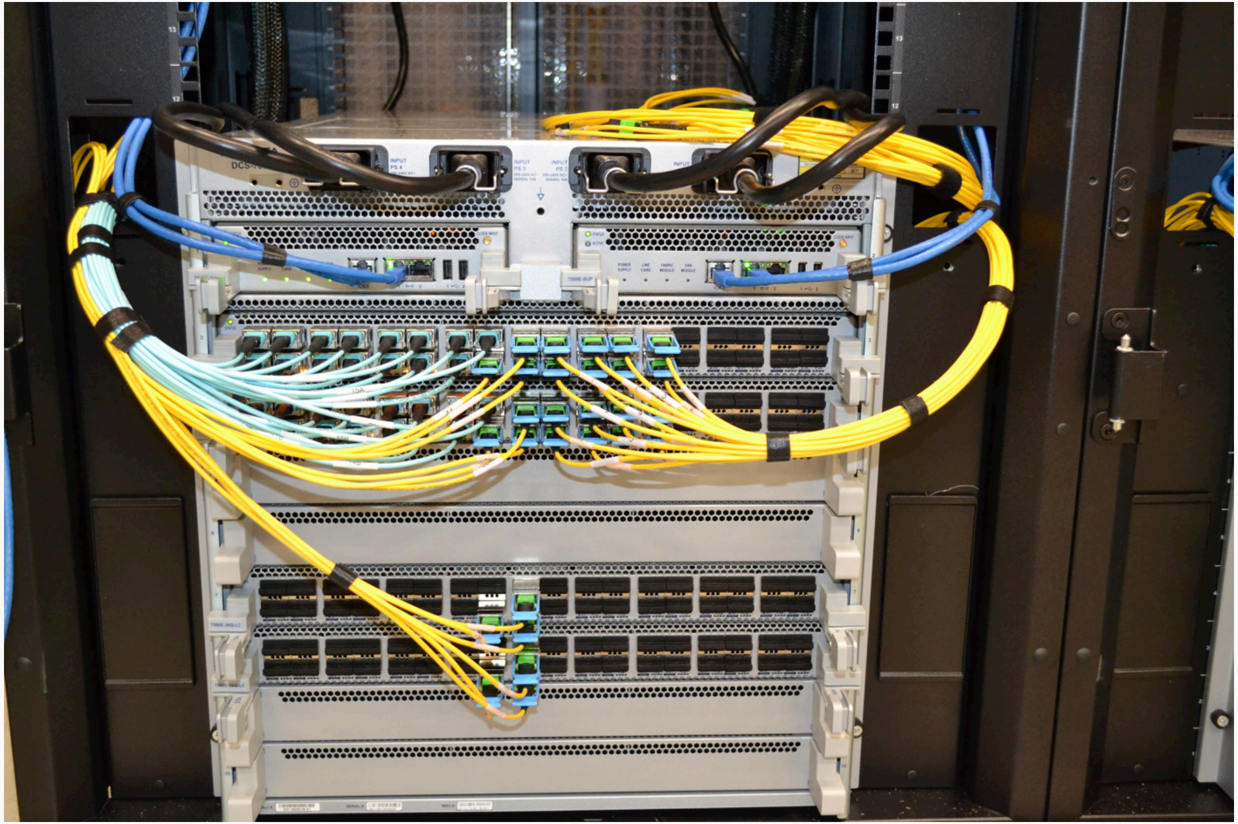




LAN1 LAN2
Hostname: c006.ewr002.ix
Public IP: 23.246.30.135
MAC1: 0CC47A426372
3
4
7
8
11
12
3
4
7
8
11
12
APC
by Schneider


APC

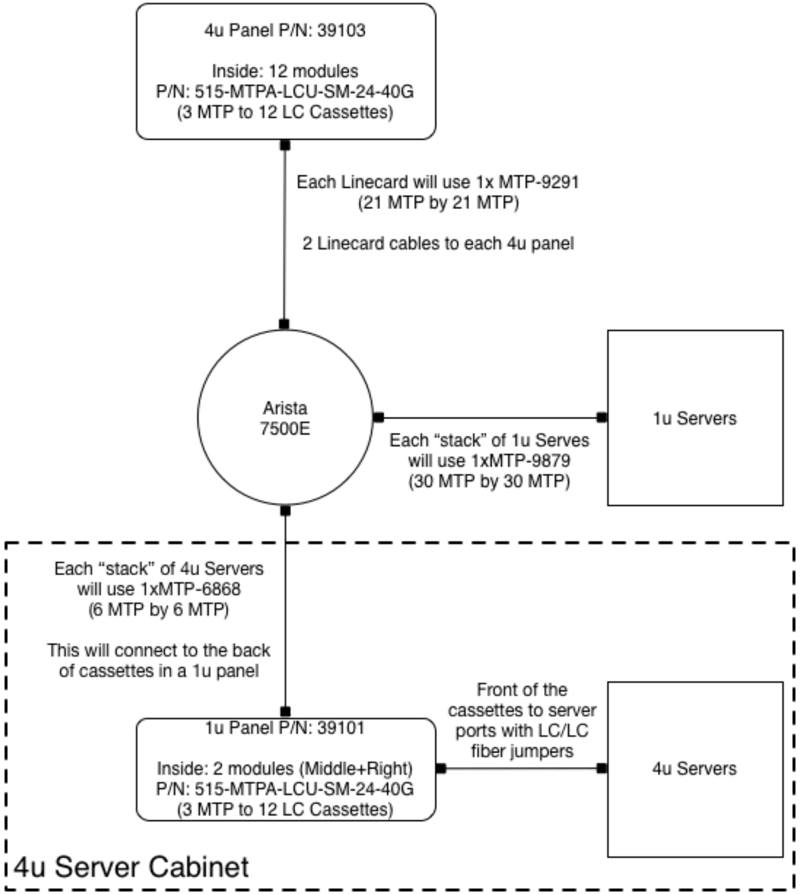




~3/30T of interconnect

Lowest Linecard cable connects to cassettes in module locations 1-6
Next Linecard cable connects to cassettes in module locations 7-12

Cables will have serial numbers on both ends to identify them



4u Panel P/N: 39103

Inside: 12 modules
P/N: 515-MTPA-LCU-SM-24-40G
(3 MTP to 12 LC Cassettes)

Each Linecard will use 1x MTP-9291
(21 MTP by 21 MTP)

2 Linecard cables to each 4u panel

Arista
7500E

Each “stack” of 1u Serves
will use 1xMTP-9879
(30 MTP by 30 MTP)

1u Servers

Each “stack” of 4u Servers
will use 1xMTP-6868
(6 MTP by 6 MTP)

This will connect to the back
of cassettes in a 1u panel

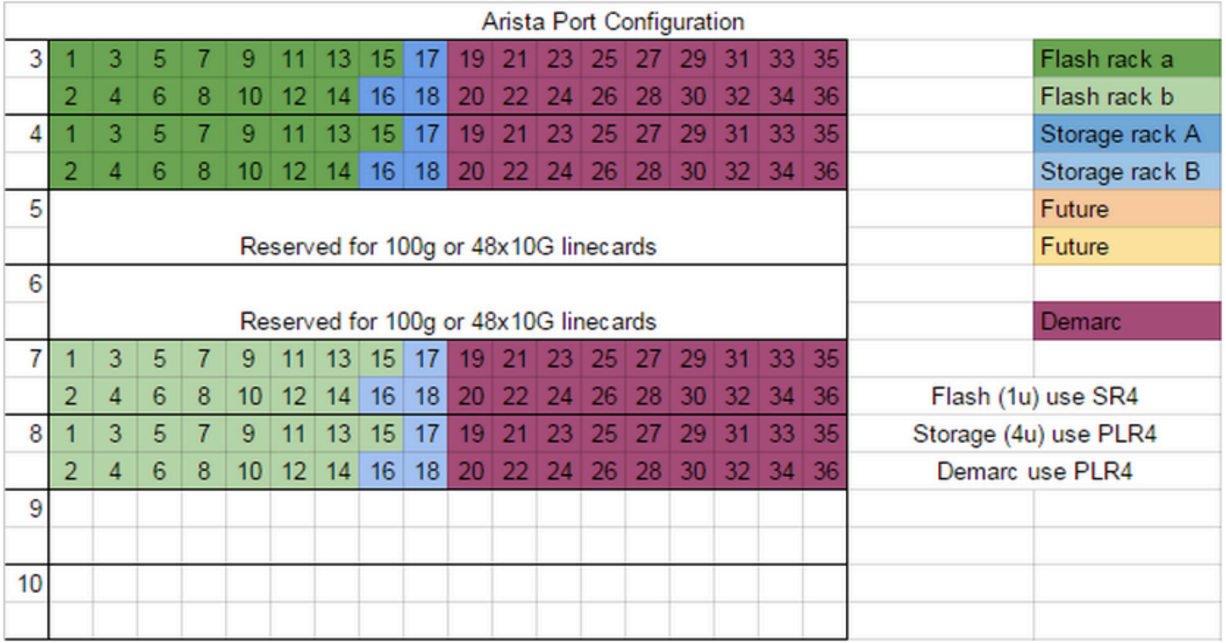
1u Panel P/N: 39101

Inside: 2 modules (Middle+Right)
P/N: 515-MTPA-LCU-SM-24-40G
(3 MTP to 12 LC Cassettes)

Front of the
cassettes to server
ports with LC/LC
fiber jumpers

4u Servers

4u Server Cabinet



Arista Port Configuration

3	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	Flash rack a
	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	Flash rack b
4	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	Storage rack A
	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	Storage rack B
5																			Future
	Reserved for 100g or 48x10G linecards																		Future
6																			
	Reserved for 100g or 48x10G linecards																		Demarc
7	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	
	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	Flash (1u) use SR4
8	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	Storage (4u) use PLR4
	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	Demarc use PLR4
9																			
10																			

Every site has the same layout

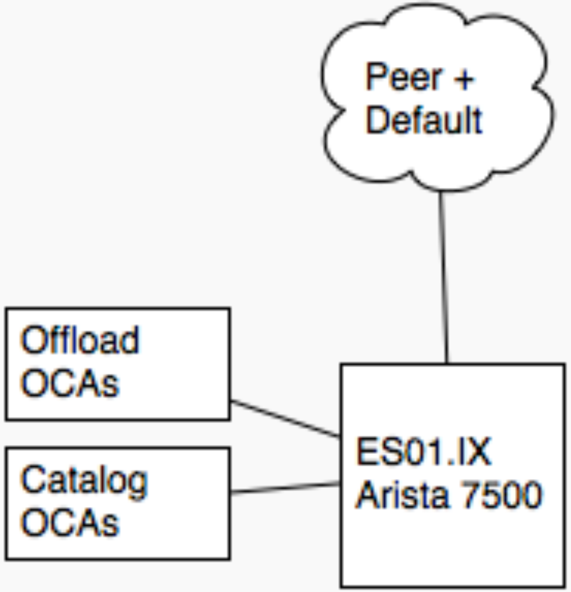
Homogeny

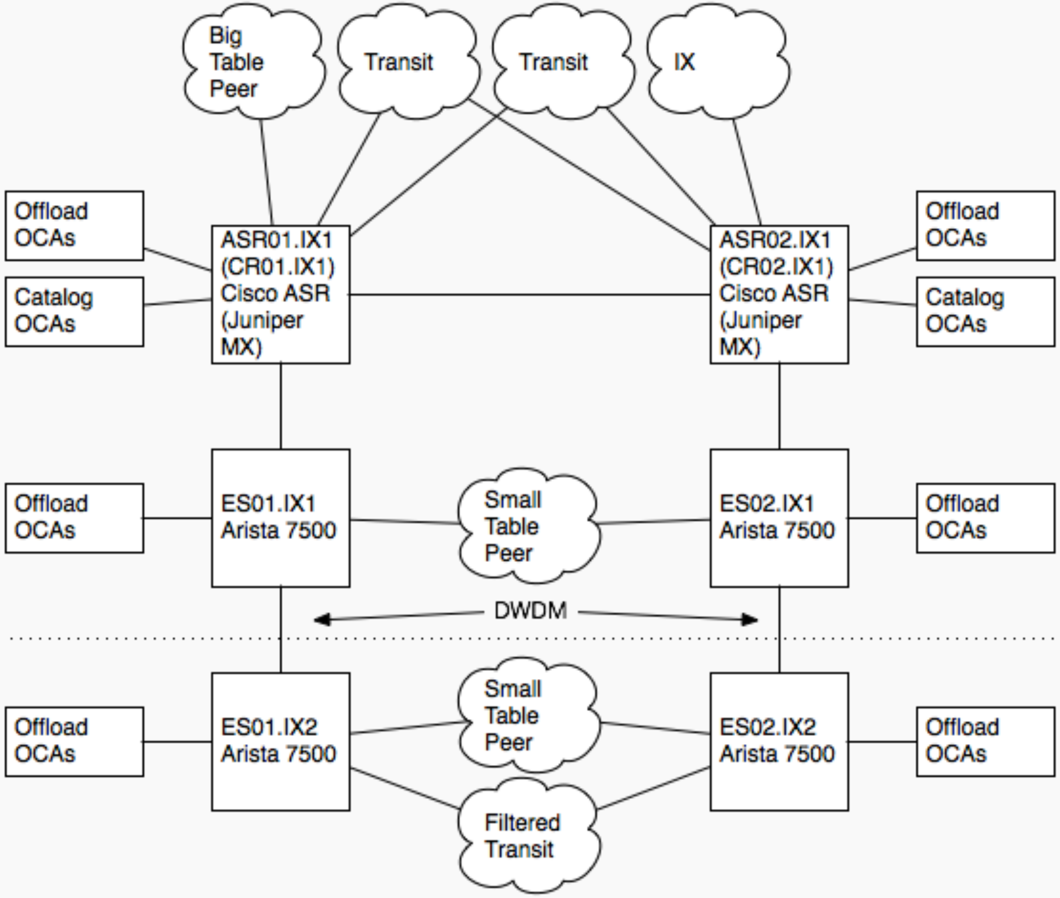
- **Allows us to make rapid deployment decisions**
  - Standardized negotiating for space and power depending on forecast
  - Quick Bill of Material generation
  - Signed to live in less than 30 days

2013

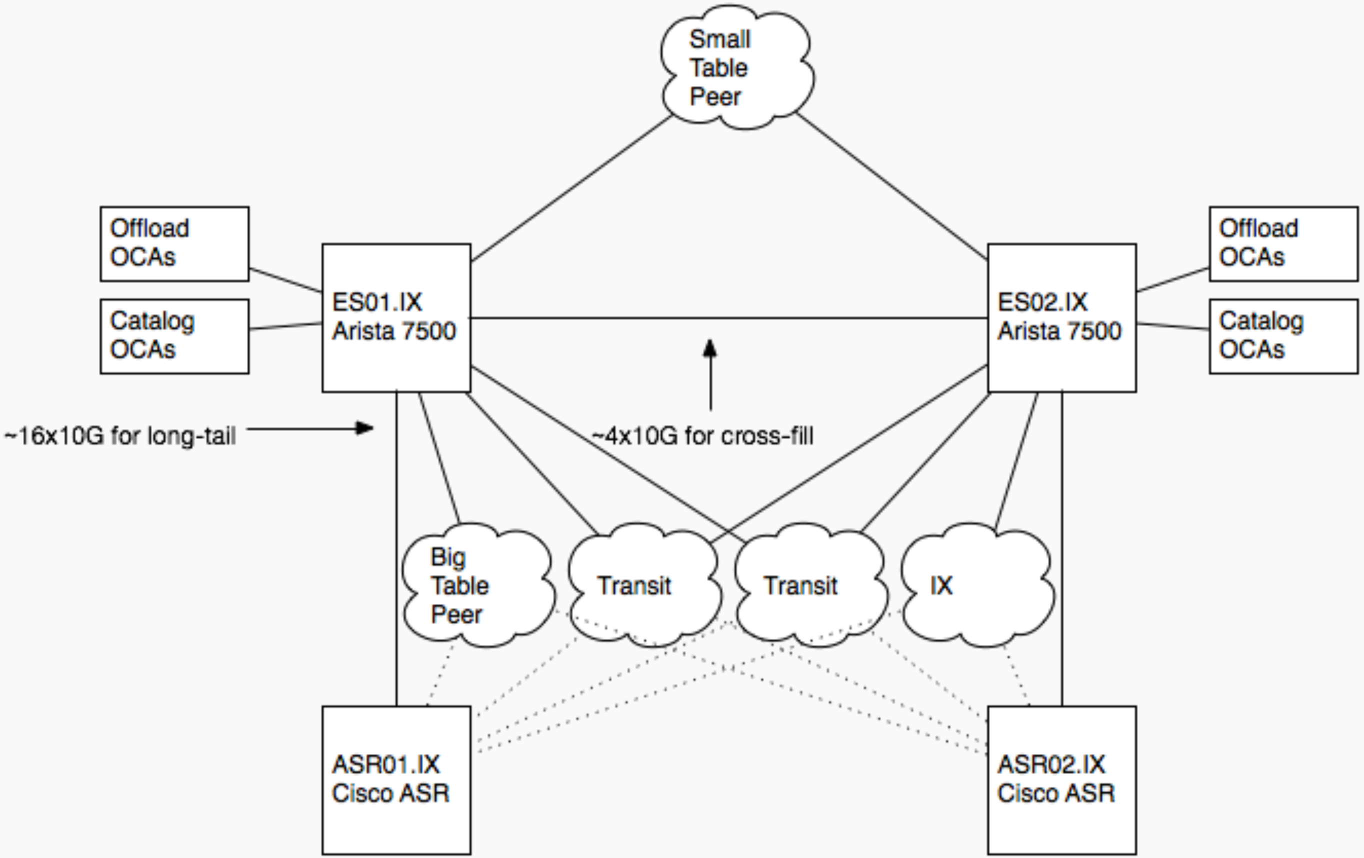
Layer 2/3

- **No opportunity for aggregation**
- **Big chassis is best**
- **Sticking to off-the-shelf platforms (for now!)**
  - Better to focus on software
- **Developing our own routing platform**
  - No longer buying big expensive routers
  - We’ve had a traffic management platform for 8+ years


Peer +
Default
Offload
OCAs
Catalog
OCAs
ES01.IX
Arista 7500



And a similar network architecture (add and remove pieces)


Small Table Peer
Offload OCAs
Catalog OCAs
ES01.IX Arista 7500
ES02.IX Arista 7500
Offload OCAs
Catalog OCAs
~16x10G for long-tail
~4x10G for cross-fill
Big Table Peer
Transit
Transit
IX
ASR01.IX Cisco ASR
ASR02.IX Cisco ASR

More info?

- http://openconnect.netflix.com
- dtemkin@netflix.com

Questions?