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

Small Peering Location

Large Peering/Origin Location

AWS S3

Each cache has identical content = 80-100% offload

Sharded content ≈ 90+% offload

Sharded content 100% of active catalog

All downloadable 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)
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
Dude, no. This is serious. I just sharted.
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.
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
~3/30T of interconnect
Lowest Linecard cable connects to cassettes in module locations 1-6
Next Linecard 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 Servers will use 1x MTP-6666
(5 MTP by 5 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

1u Servers

4u Servers

4u Server Cabinet
Every site has the same layout

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

Reserved for 100g or 48x10G linecards

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

Reserved for 100g or 48x10G linecards

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

Flash rack a
Flash rack b
Storage rack A
Storage rack B
Future
Future
Demarc
Flash (1u) use SR4
Storage (4u) use PLR4
Demarc use PLR4
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
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
And a similar network architecture (add and remove pieces)
More info?

- [http://openconnect.netflix.com](http://openconnect.netflix.com)
- [dtemkin@netflix.com](mailto:dtemkin@netflix.com)

Questions?