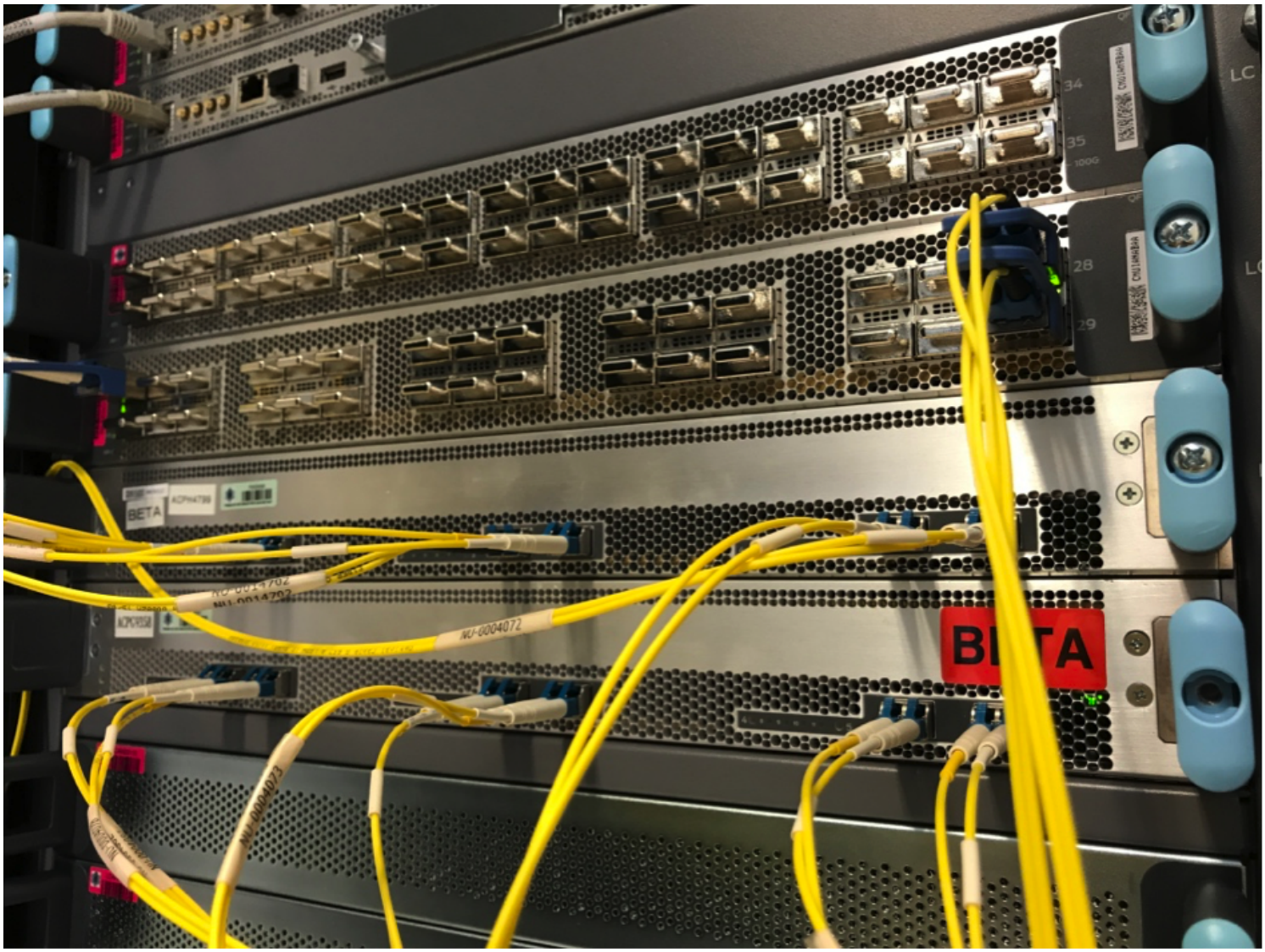
SUNET

NEW TOYS – 200G DWDM IN JUNIPER QFX10000

VOODOO IN SUNETC



BACKGROUND

Sunet got the opportunity to test Junipers new DWDM card for the QFX switching platform. It was on beta testing run in Europe and we got to steal it for a week before it was going back to the US. There might be some dark magic going on as the internal Juniper project is called Voodoo. The Voodoo card can run different line speeds of 100G, 150G and 200G on the same optical spectrum with different type of modulation/encoding using SD-FEC (Soft Decision Forward Error Correction) with 25% overhead to achieve better reach. Compared to the 100G DWDM cards that Sunet has deployed today in SunetC that are are running QPSK modulation with differential encoding and SD-FEC with 15% overhead. For Sunet it was very interesting to see how far we could push these cards on the SunetC network at those new speeds (modulations). The card is a 6x200G card so succesfully deploying this gives you 1.2Tbit/s of DWDM interconnect between the switches.

We were a bit surprised when Juniper dropped of a huge road-case at the office.

It took four guys to lift it into the datacenter, as our office-DC has two steps and no ramp.

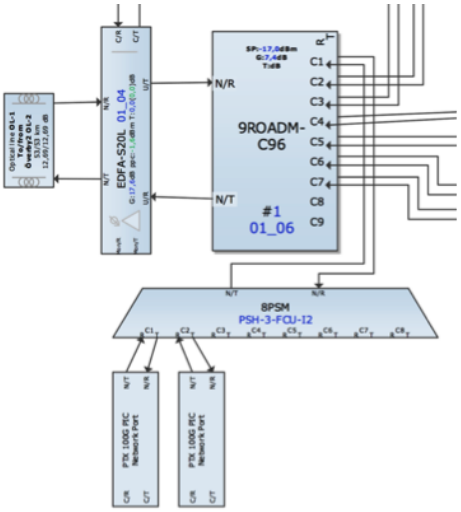

JUNIPER
NETWORKS

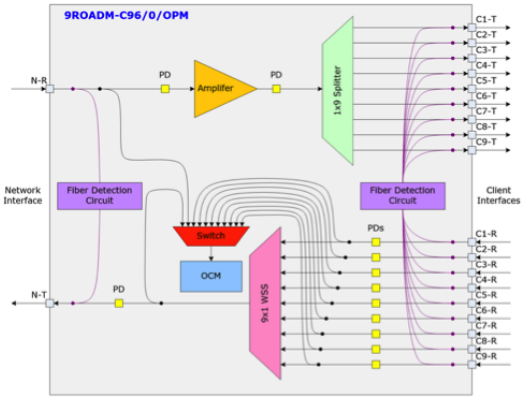
The cards when they arrived...

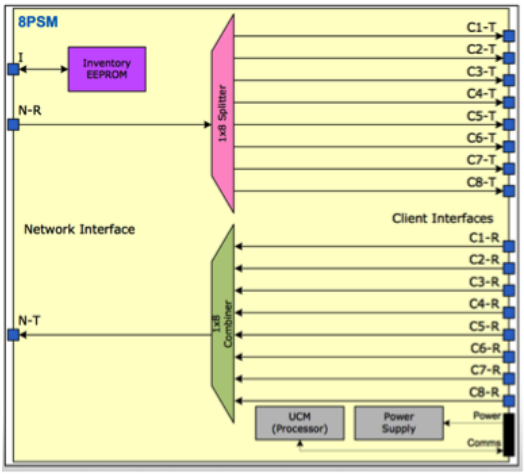
SUNET NETWORK

The SunetC optical network consists of DWDM boxes from Adva. With ROADM at drop sites, EDFA and Raman for amplification on the spans. It's optimized for high OSNR by putting Ramans on all fibers with more than 16dB loss. The system is colorless, so there are no passive or active filters, only passive splitters and combiners. That means that all the optical waves that comes into a site from a degree goes directly to the receiver on the DWDM card. The receiver needs to tune to its specific wave.

A line degree at a drop site looks like this:







TEST SETUP

From Juniper we got one QFX10008 and two Voodoo cards. As we only had one QFX we needed to connect it to two sites to be able to ”snake” an optical wave around the network as a wave can only pass the same Optical node once. That problem was solved by a 10km darkfiber to the other Sunet pop in Stockholm. Unfortunately we did not have any extra fiber pair between those sites but we did manage to find an optical splitter so we could move over two production links to share the same fiber during the test. To cope with the extra loss of that dark fiber we had to put two EDFA at the end where the Juniper switch was located, one as pre amp and the other as a post amp.

The QFX can do layer-3 but I guess it is mostly a switch as all configured IP-interfaces got the same mac-address. Try to ping another side and that side has the same mac-address. ARP doesn’t like that.

The default configuration of the QFX is to run per port spanning tree. If you connect all 6 ports back to back, five ports are going to go in to block. All solvable, but you tend to forget those things when you normally work with real routers.

We had the Voodoo channels going into the Adva system as alien/external – waves/channels (different terminology is used by different vendors but they all mean the same thing). Each wave was 50GHz wide, we would like to have had the capability to run them at 75GHz to have less impact from all the cascading filters in ROADMs. But Adva is a bit late on that (software) feature so we cant do it yet on our current software.

Since we have OTDR data on all the fibers in the network we know the exact length of each span between each amplifier. Together with the output from the ADVA NMS inventory report for each external channel and a small python script we can combine that data and get the total fiber length for each wave and also the number of amplifier and ROADM-nodes each channel passes.

The Juniper have three different line encodings to play with and six interfaces in each card so we configured the optical network for three paths and ran two channels on each path.

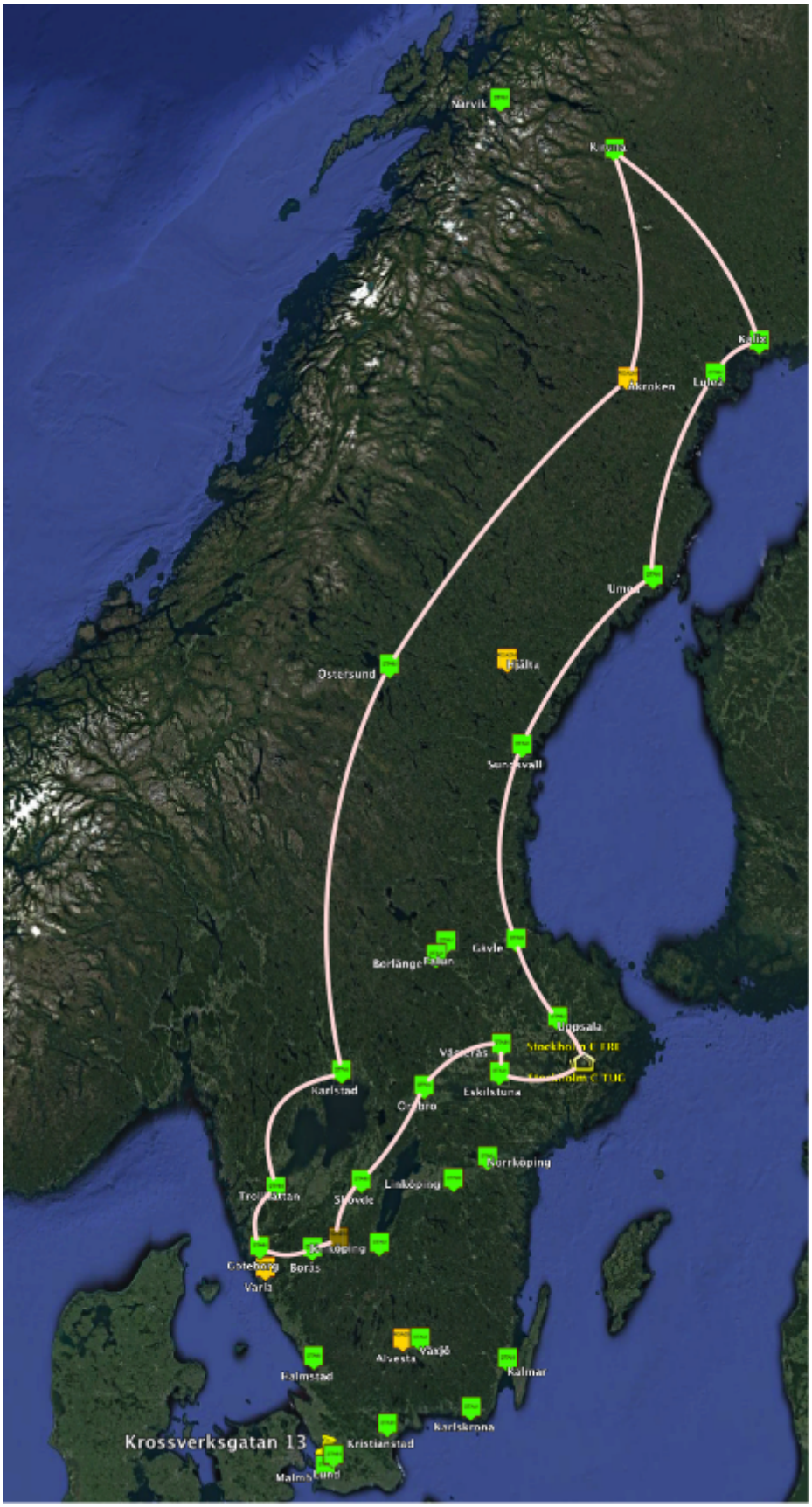
One new thing compared to the old DWDM cards we used is that Juniper separated the optical interface from the Ethernet interface. All the optical part was configured under ot-x/y/z and that interface could have the speed of 100, 150 or 200Gbps. That was then mapped to one, two or three et-x/y/z interfaces. For example, when running at 150Gbps (8QAM) you needed to have two ot- interface going the same path and those together was mapped to three et- interfaces. A bit confusing at first but it makes sense once you have typed the appropriate commands a few times. Also, it helps if you actually read the manual before starting to configure and set up different combinations of encoding and modulation to see what is actually supported or not.

QPSK

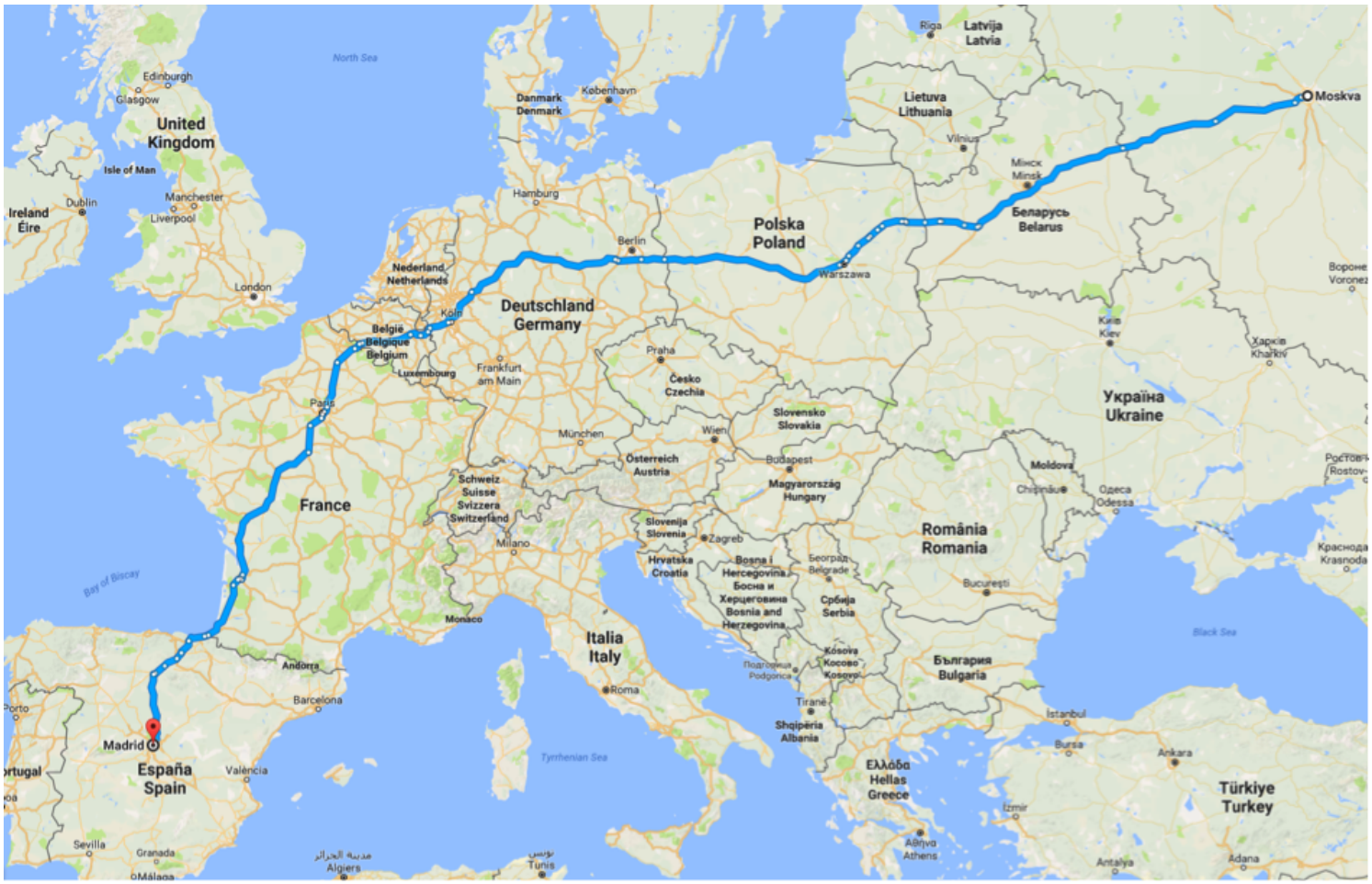
Line side encoding QPSK gives a bit rate of 100Gbps and you could configure it with the default SD-FEC at 25% overhead or the legacy of 15% as well as differential or non-differential encoding. We went for the best with 25% and non-differential encoding. We started on a 300km path through the SunetC network and tested different paths until we ended up with a path of ~4000km. The Adva NMS takes up to 10min to setup a new wave with all the steps of tearing down the old and creating a new with a lot of point and click so it was a bit of tedious work, but compared to configuring it manually in every ROADM its fairly simple.

	Length	Loss	Gain	Raman Gain	Nr RODAMs	Nr EDFA	Nr RAMAN	Nr Spans
A-Z	4073,2	890,8	1348,8	251,3	18	61	26	61
Z-A	4073,2	890,8	1358,2	259,8	18	61	26	61

The Length are in km, the Loss is the combined loss of the fiber from the OTDR measurement in dB. The Gain is the sum of the configured gain of the EDFA and Ramans amps on the path in dB. Raman Gain is the sum of all the Raman amps and that number is part of the Total Gain. The numbers show how many ROADMs, EDFAs and Ramans the span passes. Last column shows the total number of fiber spans on the path, where a span is the fiber between two amps.



This is the actual path through our network with its correct stretches, this might not look like much but Sweden is a long country, if we lay out 4000km on another map it gives a better perspective of scale.



Optical read out from the switch...

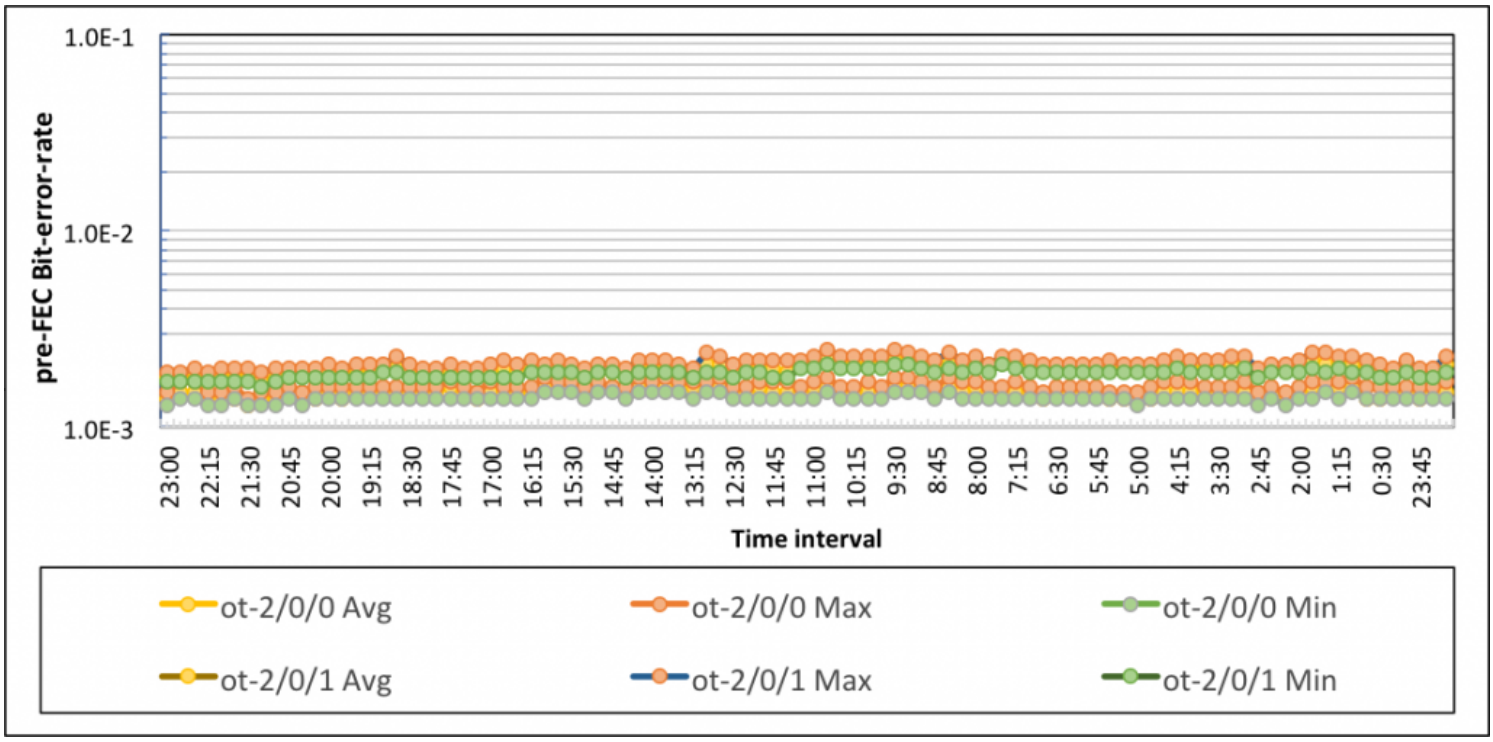
```
root@voodoo-1# run show interfaces transport pm optics current ot-2/0/1
Physical interface: ot-2/0/1, SNMP ifIndex 545
  10:30-current
  Suspect Flag:False          Reason:Not Applicable
    PM                              CURRENT   MIN       MAX       AVG       THRESHOLD       TCA-ENABLED       TCA-RAISED
                                                                            (MIN) (MAX)     (MIN) (MAX)       (MIN) (MAX)
  Lane chromatic dispersion(ps/nm)  -1890     -1910     -1886     -1887     0     0         NA    NA          NA    NA
  Lane differential group delay(ps) 11        5         18        6         0     0         NA    NA          NA    NA
  Lane Q2 factor(0.1dB)             91        90        93        91        0     0         NA    NA          NA    NA
  SNR(0.1dB)                        93        91        95        92        0     0         NA    NA          NA    NA
  Carrier frequency offset(MHz)     -155      -214      5         -39       -3600 3600      No    No          No    No
  Tx output power(0.01dBm)          -1        -1        1         0         -1200 1000      No    No          No    No
  Rx input total power(0.01dBm)     -3        -8        -1        -2        -3200 1000      No    No          No    No
  Module temperature(Celsius)       53        53        54        53        -5    75        No    No          No    No
  FEC        Suspect Flag:False          Reason:Not Applicable
    PM                  COUNT           THRESHOLD       TCA-ENABLED       TCA-RAISED
  FEC-CorrectedErr      14588083693302  0               NA                NA
  FEC-UncorrectedWords  0               0               NA                NA
  BER        Suspect Flag:False          Reason:Not Applicable
    PM                  MIN       MAX       AVG       THRESHOLD       TCA-ENABLED       TCA-RAISED
  BER                   1.6e-3    2.4e-3    2.0e-3    0.0e-0          No                No
```

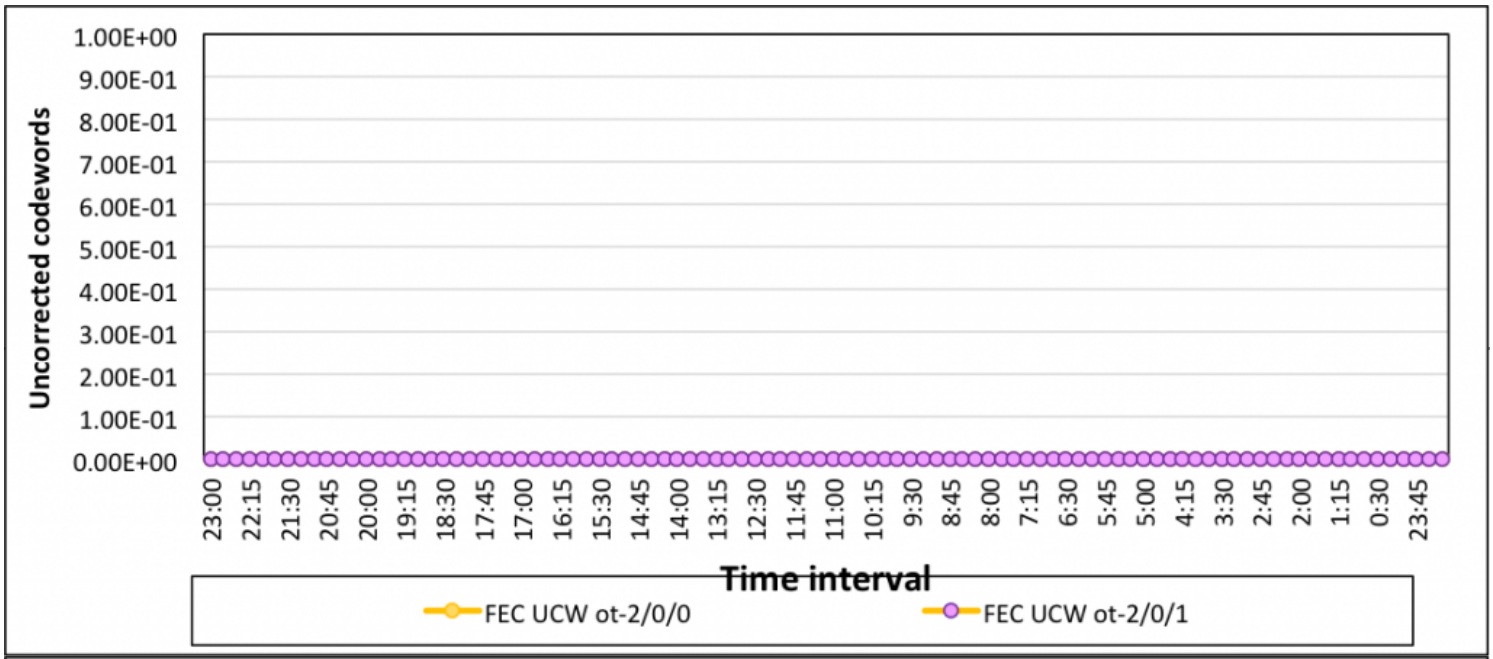
The Chromatic dispersion is not negative it's just that Juniper think it's a signed INT so it wraps around and becomes negative.... this problem existed in MX aswell before we reported it, now we did it again.

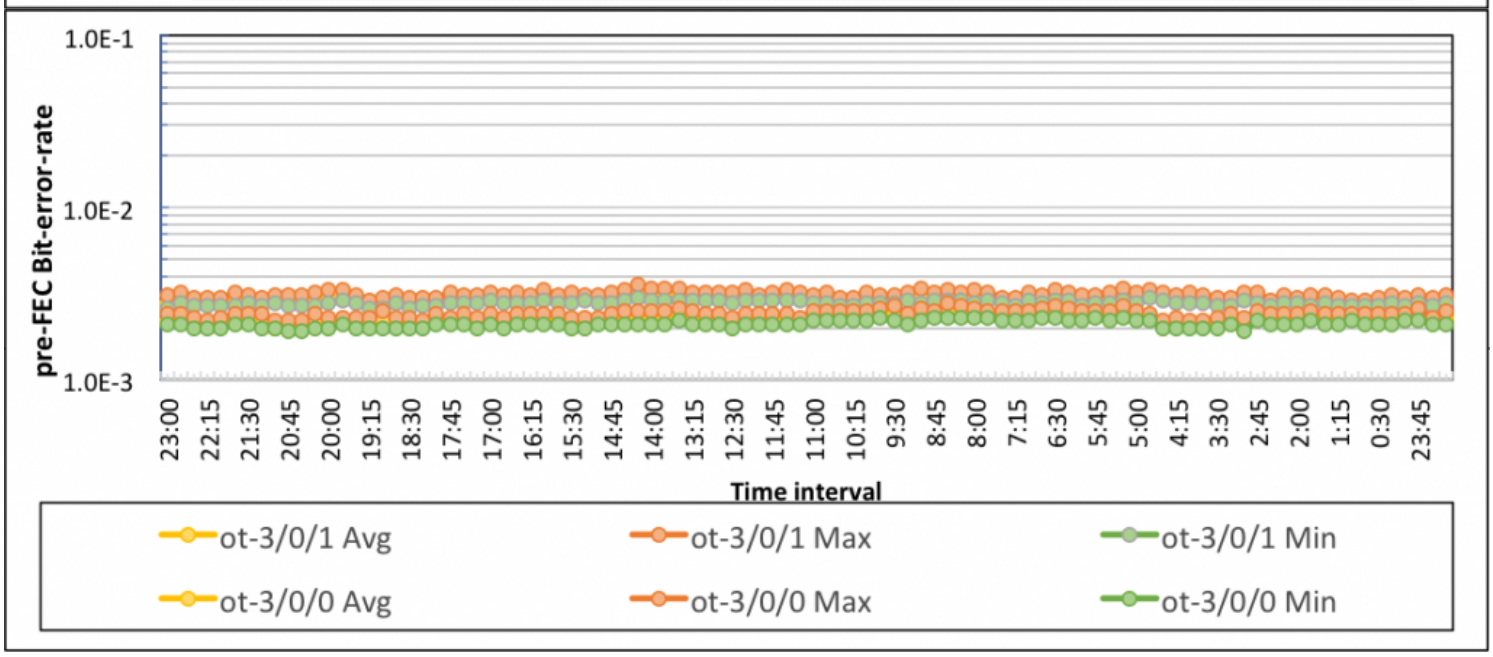
The SNR is the electrical Signal to Noise Ratio and not the optical. The optical is harder to figure out but you can get a good estimate by looking at the Spectrum analyzer graphs below.

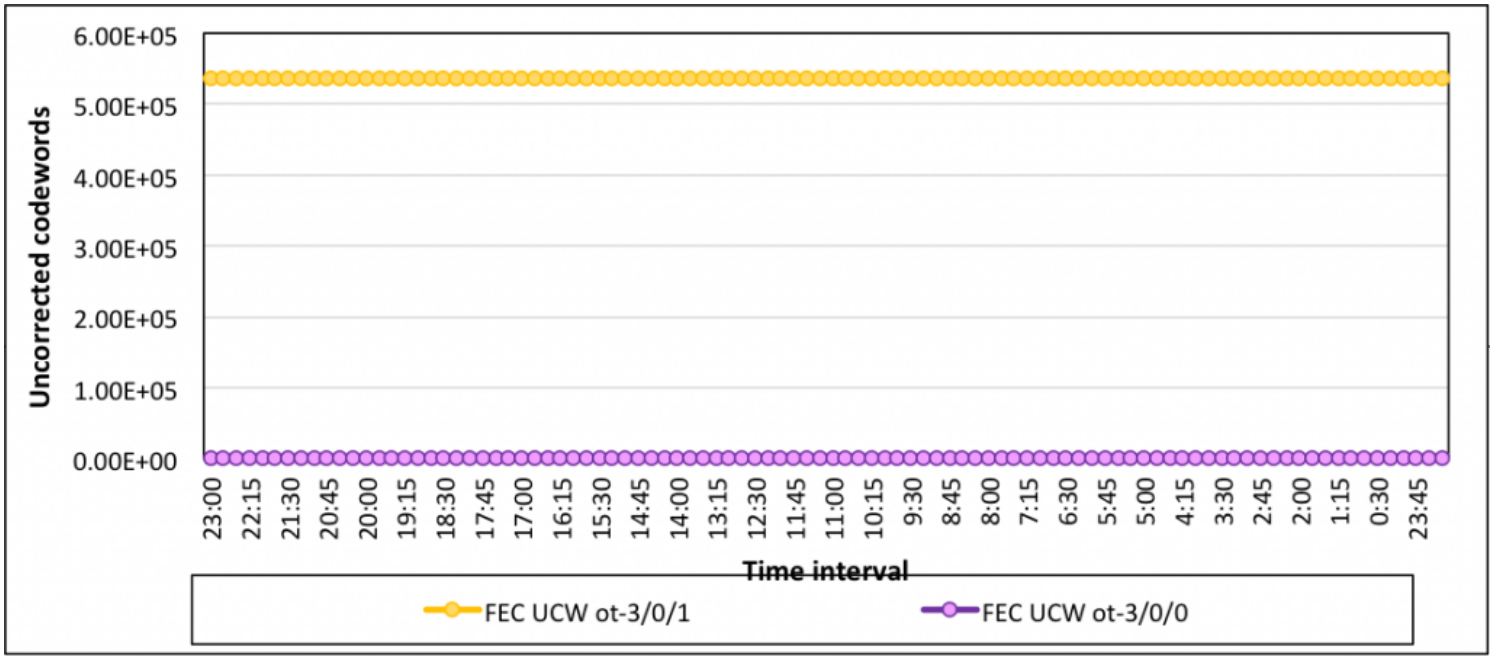
We got help from Juniper with collecting the data from the QFX (custom event-scripts) along with some excel magic.

As you can see on the FEC values there are still some margin to the FEC limit on 3.4e-2. In one direction on one of the links we were cutting it a lite bit to close, as you can see the Uncorrected codewords counting up. We didn't have the time to dig down and find the issue for that, might be as simple as a dirty patch, we did scope all connectors but it's not always enough.


1.0E-1
1.0E-2
1.0E-3
pre-FEC Bit-error-rate
23:00
22:15
21:30
20:45
20:00
19:15
18:30
17:45
17:00
16:15
15:30
14:45
14:00
13:15
12:30
11:45
11:00
10:15
9:30
8:45
8:00
7:15
6:30
5:45
5:00
4:15
3:30
2:45
2:00
1:15
0:30
23:45
Time interval
ot-2/0/0 Avg
ot-2/0/0 Max
ot-2/0/0 Min
ot-2/0/1 Avg
ot-2/0/1 Max
ot-2/0/1 Min


Uncorrected codewords
1.00E+00
9.00E-01
8.00E-01
7.00E-01
6.00E-01
5.00E-01
4.00E-01
3.00E-01
2.00E-01
1.00E-01
0.00E+00
23:00 22:15 21:30 20:45 20:00 19:15 18:30 17:45 17:00 16:15 15:30 14:45 14:00 13:15 12:30 11:45 11:00 10:15 9:30 8:45 8:00 7:15 6:30 5:45 5:00 4:15 3:30 2:45 2:00 1:15 0:30 23:45
Time interval
FEC UCW ot-2/0/0
FEC UCW ot-2/0/1


pre-FEC Bit-error-rate
1.0E-1
1.0E-2
1.0E-3
23:00 22:15 21:30 20:45 20:00 19:15 18:30 17:45 17:00 16:15 15:30 14:45 14:00 13:15 12:30 11:45 11:00 10:15 9:30 8:45 8:00 7:15 6:30 5:45 5:00 4:15 3:30 2:45 2:00 1:15 0:30 23:45
Time interval
ot-3/0/1 Avg
ot-3/0/1 Max
ot-3/0/1 Min
ot-3/0/0 Avg
ot-3/0/0 Max
ot-3/0/0 Min

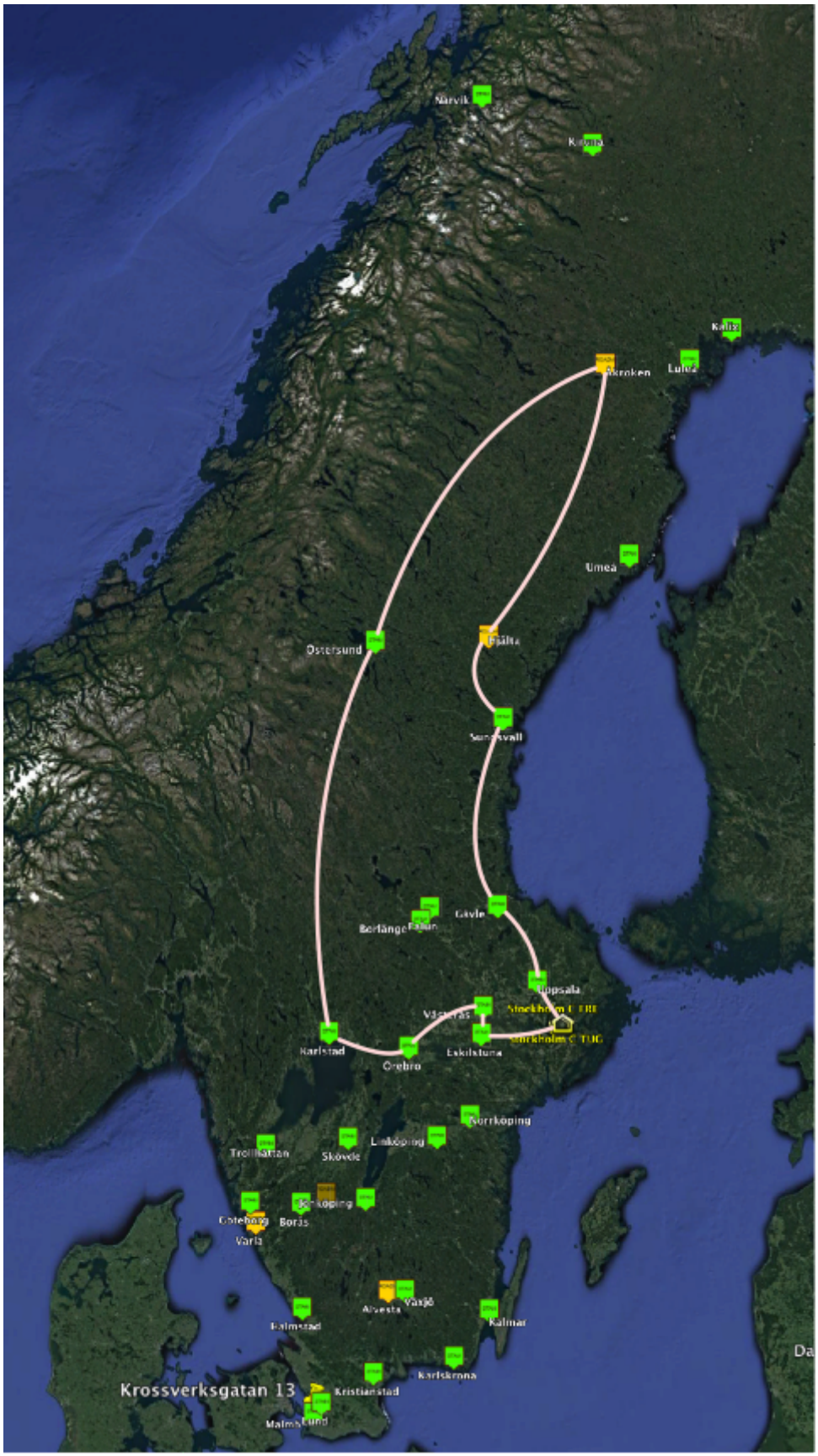

Uncorrected codewords
6.00E+05
5.00E+05
4.00E+05
3.00E+05
2.00E+05
1.00E+05
0.00E+00
23:00 22:15 21:30 20:45 20:00 19:15 18:30 17:45 17:00 16:15 15:30 14:45 14:00 13:15 12:30 11:45 11:00 10:15 9:30 8:45 8:00 7:15 6:30 5:45 5:00 4:15 3:30 2:45 2:00 1:15 0:30 23:45
Time interval
FEC UCW ot-3/0/1
FEC UCW ot-3/0/0

We tried running QPSK over 4600km but the Voodoo card was not able to sync up, so we are pretty sure we hit maximum feasible distance

8QAM

This line encoding gives a transfer rate of 150Gbps. As expected was the reach shorter but still impressive at 2338km.

	Length	Loss	Gain	Raman Gain	Nr ROADMs	Nr EDFA	Nr RAMAN	Nr Spans
A-Z	2338,2	503,4	775,6	142	10	37	15	37
Z-A	2338,2	503,4	788,9	145,5	10	37	15	37



This setup had worse FEC than QPSK but was running totally error free. Indicating that the QPSK with errors where due to something else

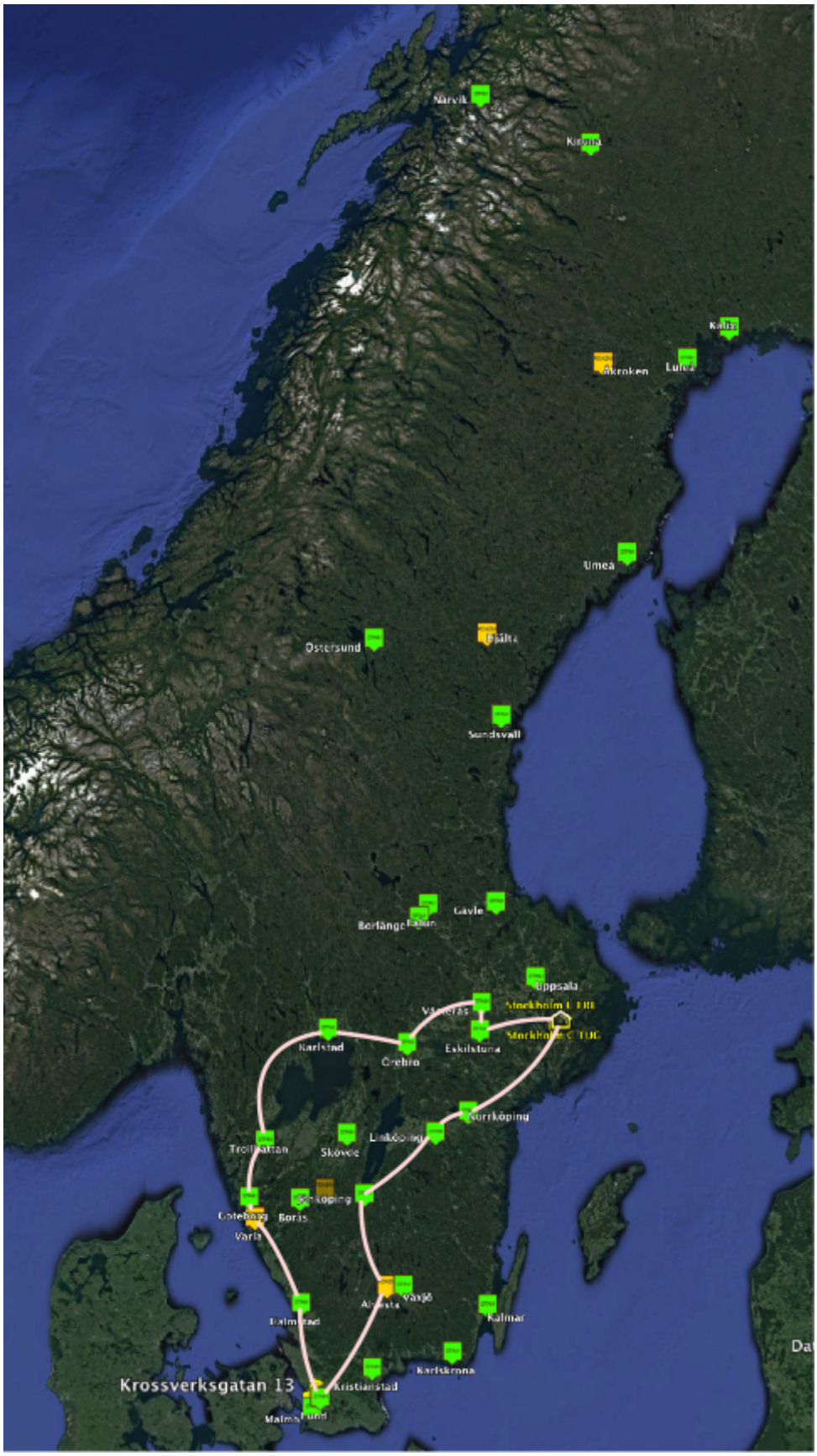

pre-FEC Bit-error-rate
1.0E-1
1.0E-2
1.0E-3
23:00 22:15 21:30 20:45 20:00 19:15 18:30 17:45 17:00 16:15 15:30 14:45 14:00 13:15 12:30 11:45 11:00 10:15 9:30 8:45 8:00 7:15 6:30 5:45 5:00 4:15 3:30 2:45 2:00 1:15 0:30 23:45
Time interval
ot-2/0/2 Avg
ot-2/0/2 Max
ot-2/0/2 Min
ot-2/0/3 Avg
ot-2/0/3 Max
ot-2/0/3 Min

Uncorrected codewords
1.00E+00
9.00E-01
8.00E-01
7.00E-01
6.00E-01
5.00E-01
4.00E-01
3.00E-01
2.00E-01
1.00E-01
0.00E+00
23:00 22:15 21:30 20:45 20:00 19:15 18:30 17:45 17:00 16:15 15:30 14:45 14:00 13:15 12:30 11:45 11:00 10:15 9:30 8:45 8:00 7:15 6:30 5:45 5:00 4:15 3:30 2:45 2:00 1:15 0:30 23:45
Time interval
FEC UCW ot-2/0/2
FEC UCW ot-2/0/3

16QAM

This is for us the most interesting line encoding where we get a bit rate of 200Gbps in a 50GHz channel window. The distance is even shorter than 8QAM but still long to reach most places in Sweden from Stockholm.

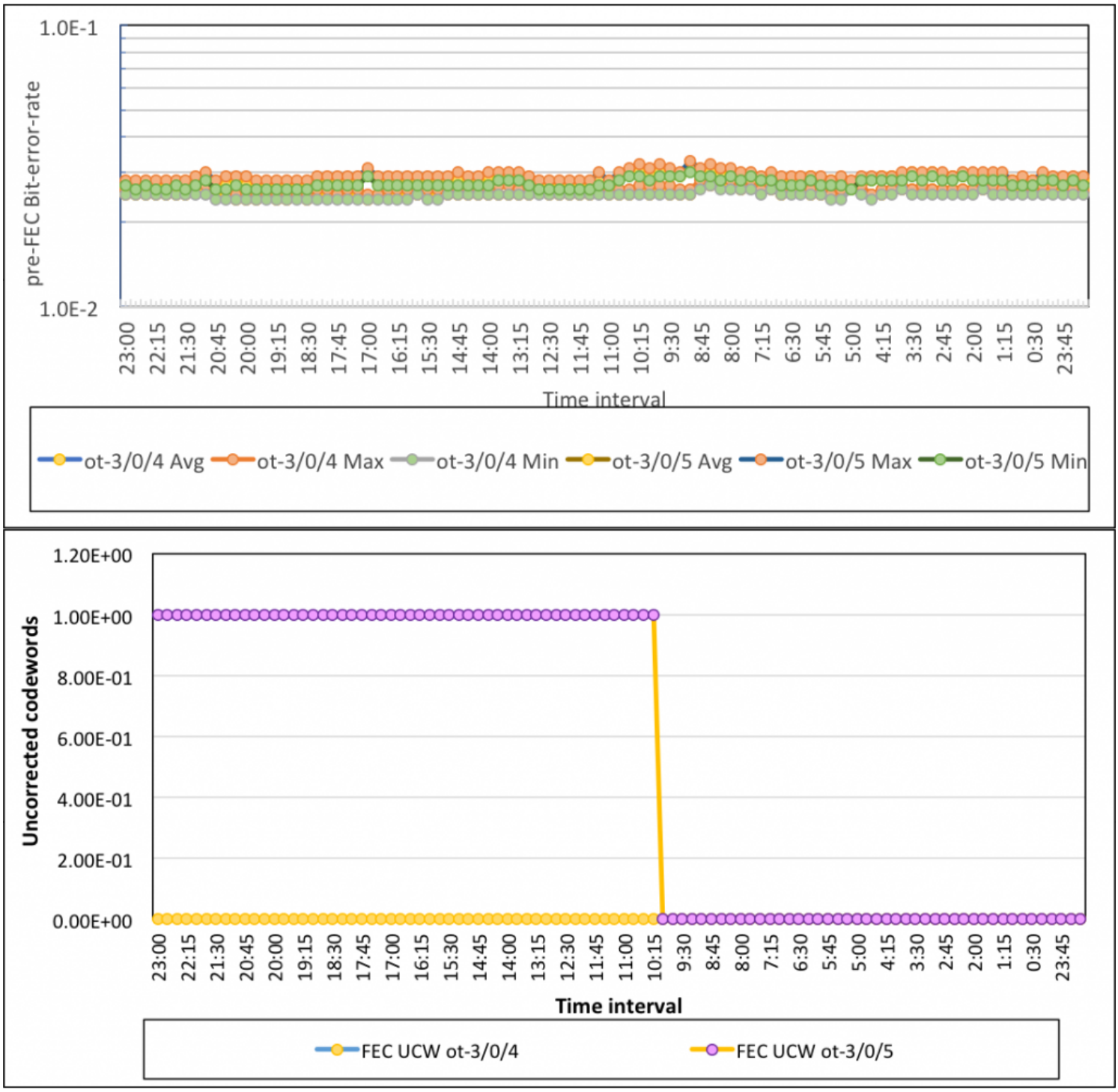
	Length	Loss	Gain	Raman Gain	Nr WSS	Nr EDFA	Nr RAMAN	Nr Spans
A-Z	1843,3	422,91	731,8	99	14	30	10	30
Z-A	1843,3	422,91	732,6	105,8	14	30	10	30



This was cutting it very close to what the FEC could handle. At least one of the two links where running error free. We had 16QAM running at 1700km on both links error free with a FEC of 2,4E-2 but it is more fun to find the limit.


1.0E-1
1.0E-2
pre-FEC Bit-error-rate
23:00
22:15
21:30
20:45
20:00
19:15
18:30
17:45
17:00
16:15
15:30
14:45
14:00
13:15
12:30
11:45
11:00
10:15
9:30
8:45
8:00
7:15
6:30
5:45
5:00
4:15
3:30
2:45
2:00
1:15
0:30
23:45
Time interval
ot-2/0/4 Avg
ot-2/0/4 Max
ot-2/0/4 Min
ot-2/0/5 Avg
ot-2/0/5 Max
ot-2/0/5 Min

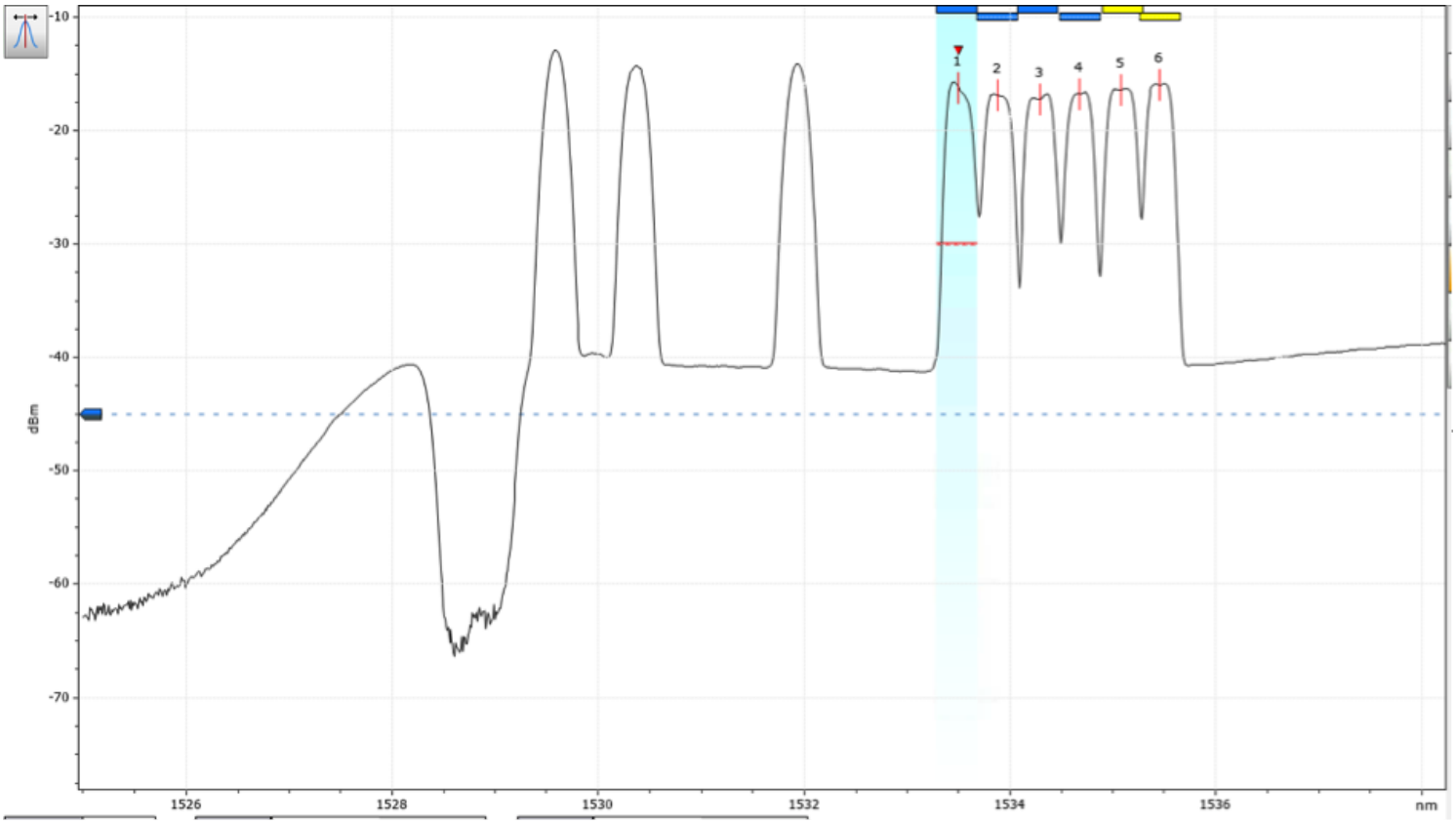
3.00E+06
2.50E+06
2.00E+06
1.50E+06
1.00E+06
5.00E+05
0.00E+00
Uncorrected codewords
23:00
22:15
21:30
20:45
20:00
19:15
18:30
17:45
17:00
16:15
15:30
14:45
14:00
13:15
12:30
11:45
11:00
10:15
9:30
8:45
8:00
7:15
6:30
5:45
5:00
4:15
3:30
2:45
2:00
1:15
0:30
23:45
Time interval
FEC UCW ot-2/0/4
FEC UCW ot-2/0/5



SPECTRUM

We connected an EXFO OSA to the splitter that combined all the QFX interfaces and this is the graph of the waves going in to the QFX.

The first three channels are production traffic. The dip in the beginning around 1528nm is caused by the OSC channel filter at 1528nm.



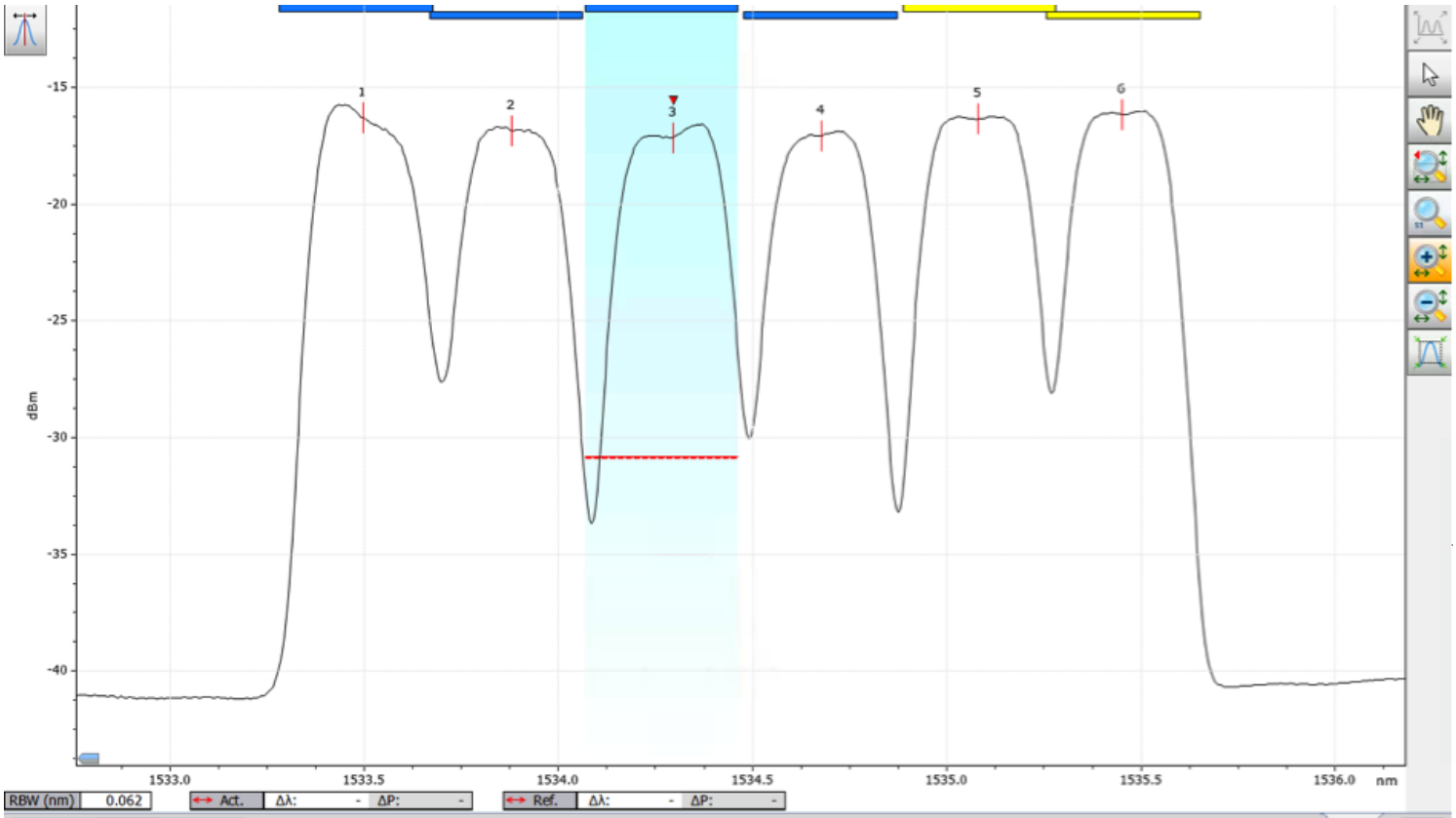
The six channels you see below in the zoomed in graph are

Channel 1+2: QPSK ~4000km @ 2x 100G

Channel 3+4: 8QAM ~2300km @ 2x 150G

Channel 5+6: 16QAM ~1800km @ 2x 200G

Zoom! ENHANCE!



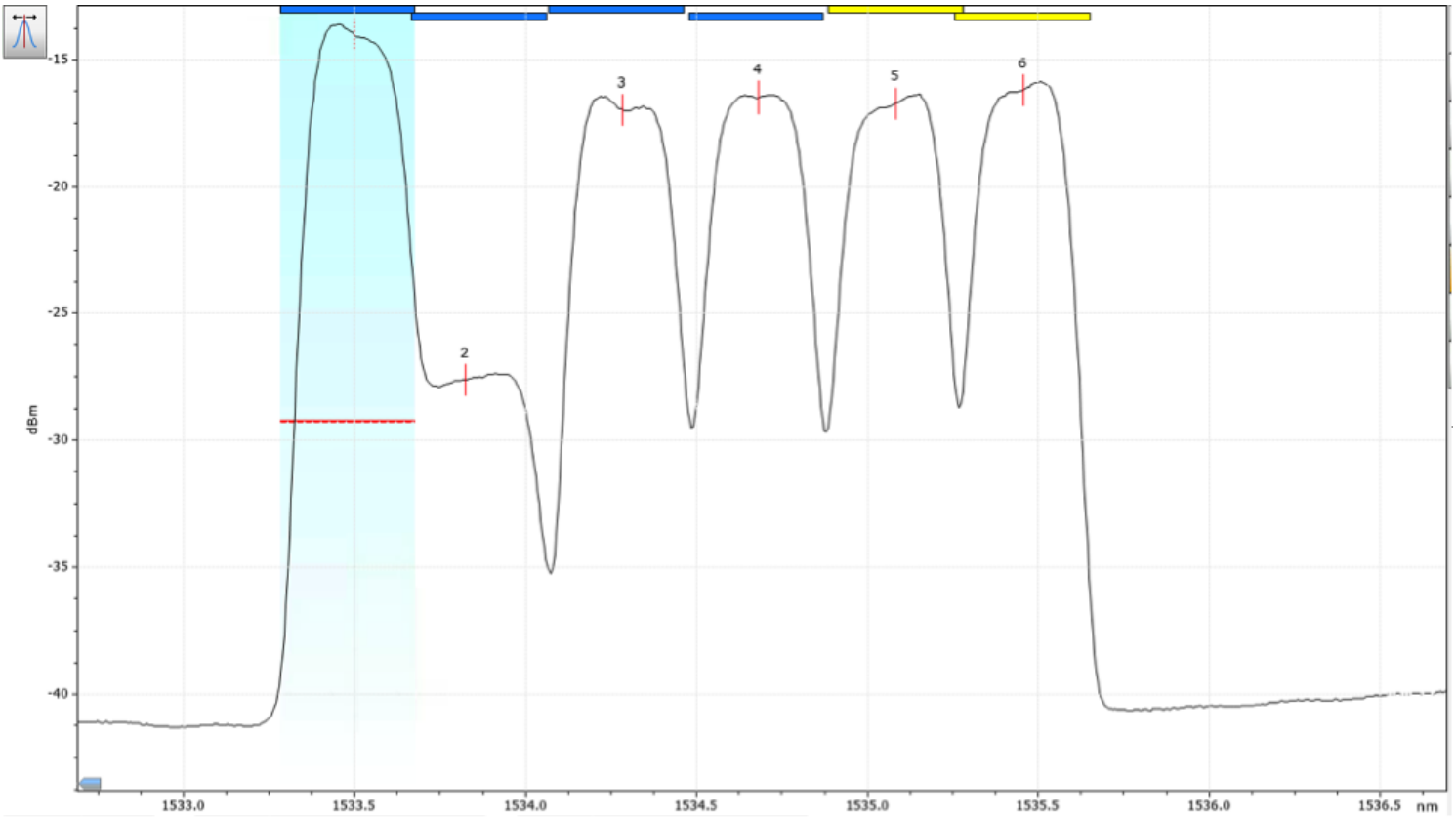
Acquisition | Results | Trace Info.

Ch. #	λ (nm)	Power (dBm)	OSNR (dB)	Noise (dBm)	BW 3.00 dB (nm)	BW 20.00 dB (nm)
1	1533.497	(i)-10.48	17.42	(IEC)-27.90	0.238	-
2	1533.878	(i)-11.38	14.62	(IEC)-25.99	0.237	-
3	1534.294	(i)-10.87	17.91	(IEC)-28.78	0.259	-
4	1534.678	(i)-11.51	16.64	(IEC)-28.15	0.234	-
5	1535.080	(i)-10.44	16.21	(IEC)-26.65	0.259	-
6	1535.451	(i)-10.51	16.05	(IEC)-26.56	0.241	-

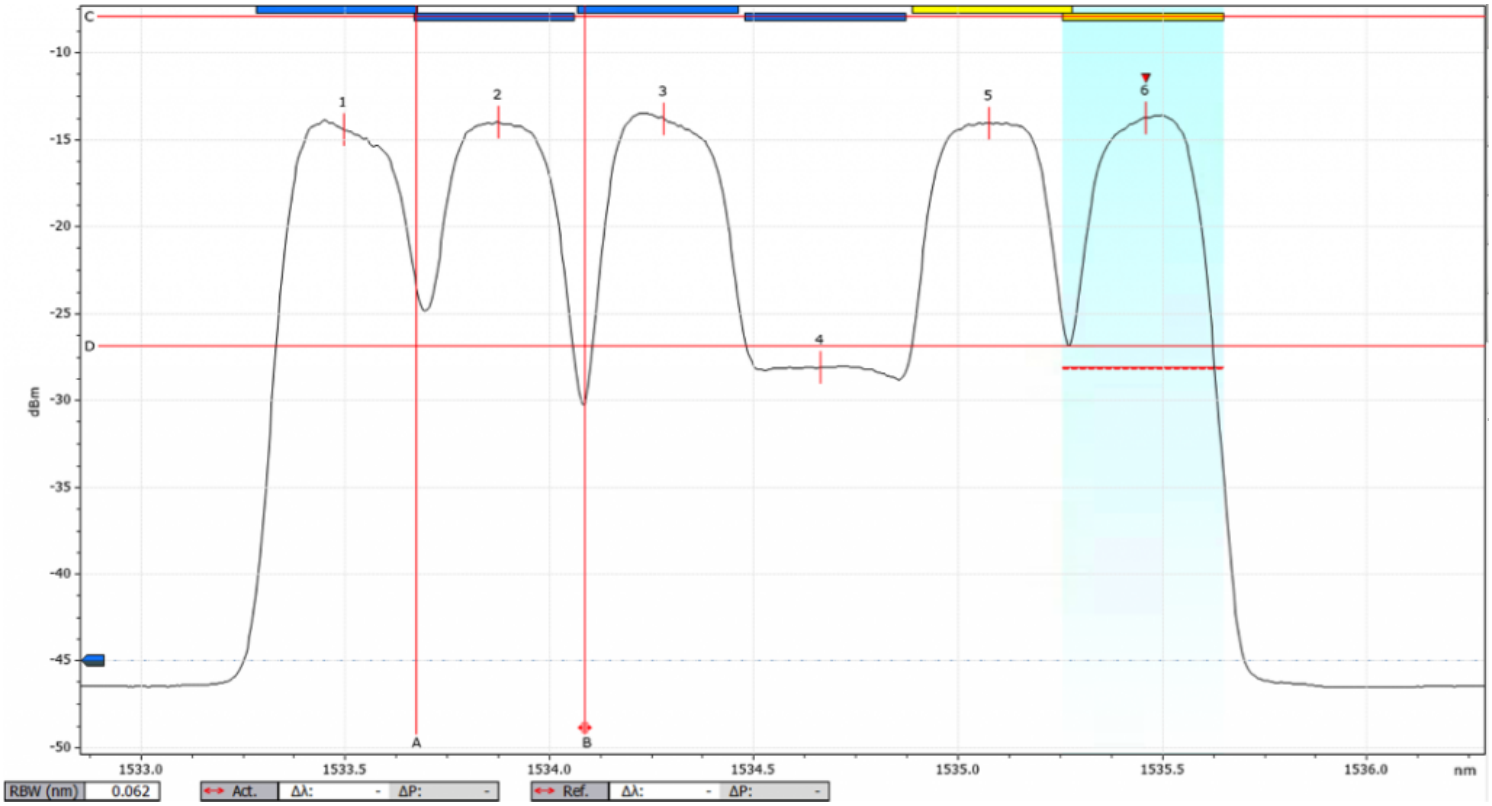
NOISE, OSNR…

We pulled out the fiber to the second port in the switch, this is much faster than turning the laser off and on as the interface is very slow to turn on the laser and tune to its wavelength. The DWDM system have all the amps and ROADM filters open so the thing that we see in the graph are the background noise from all the amps on that channel. The difference between the background noise and the signal power becomes the OSNR (Optical Signal to Noise Ratio).

One interesting effect is that the other channel that goes the same path gets an extra amplification when the (other) channel is removed. This can probably be explained in great detail by some physicists but in short, the Raman light boost the light it hits and with one wave instead of two there is a higher probability that the light from the first wave gets hit by the Raman light and gets an extra boost. Makes sense right? Maybe Juniper is onto something here when calling the productline Voodoo.....



Channel 3 and 4 are shorter (2300km) than 1 and 2 (4000km) and therefore a lower background noise and a better OSNR is observed. As a result, we can run a better or more advanced encoding.



TEST AGAINST THE CURRENT CARD SUNET RUN, CORDOBA

The Juniper 100G DWDM MIC (Codename Cordoba) for the MPC3E MX card should work with the Voodoo card if you set it to SDFEC 15% overhead (25% is default) and QPSK modulation. It didn't work at the first try and we punched the switch hard for being naughty. The Voodoo card didn't even send out light and any channels configured for 8QAM went down, I hope that this is just a beta

software issue, we punched it again to make it behave. Turns out that you need to change the encoding to “differential” since the standard is non-differential. After that it came up without any problems and the 8QAM channels came up again. Here is a low level command that confirms that we now have Cordoba compatible mode as well.

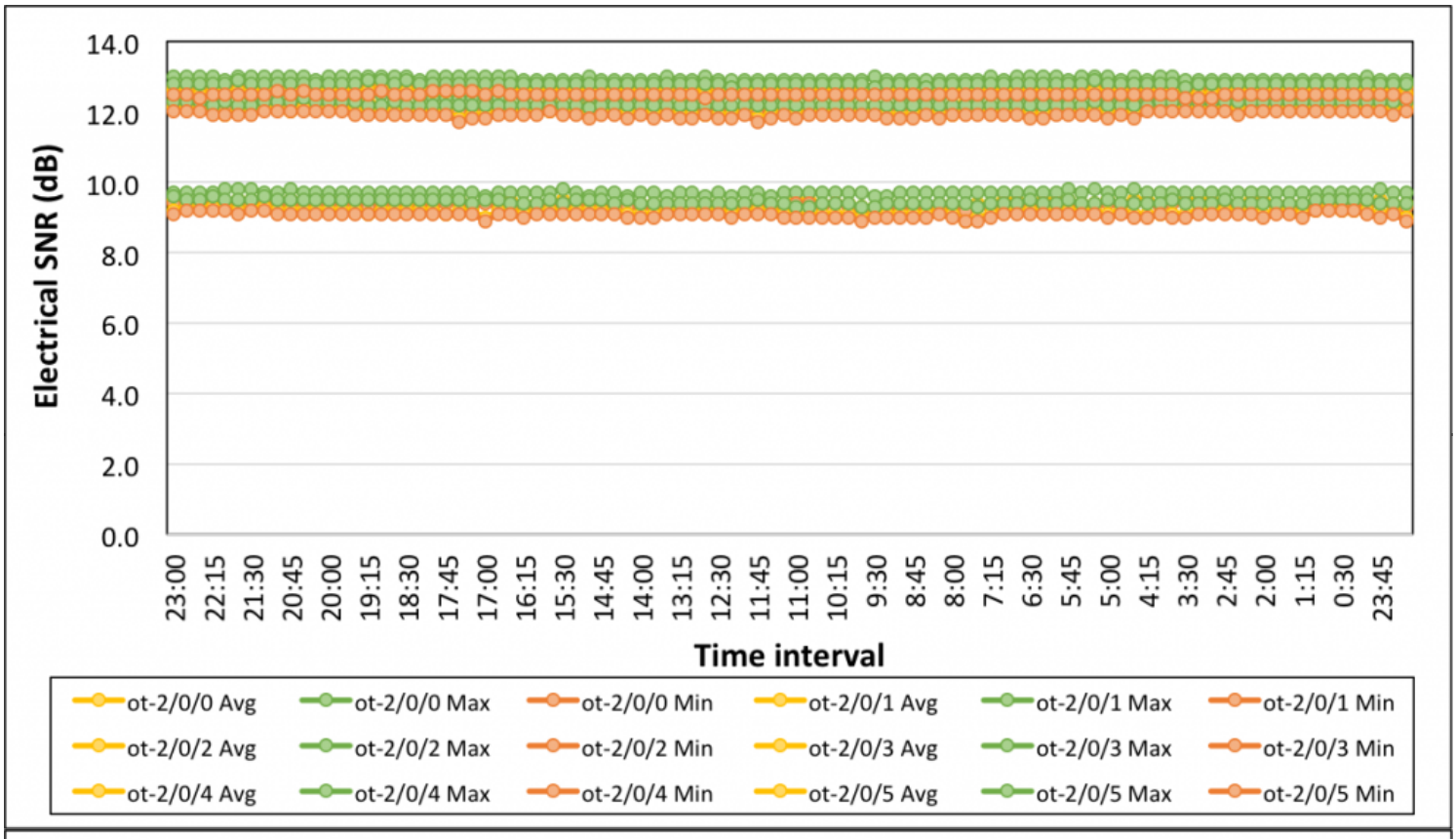
```
FPC2(voodoo-1 vty)# show ac400 3 modulation-format
Network Lane0: cordoba-compatible
Network Lane1: 16qam
```

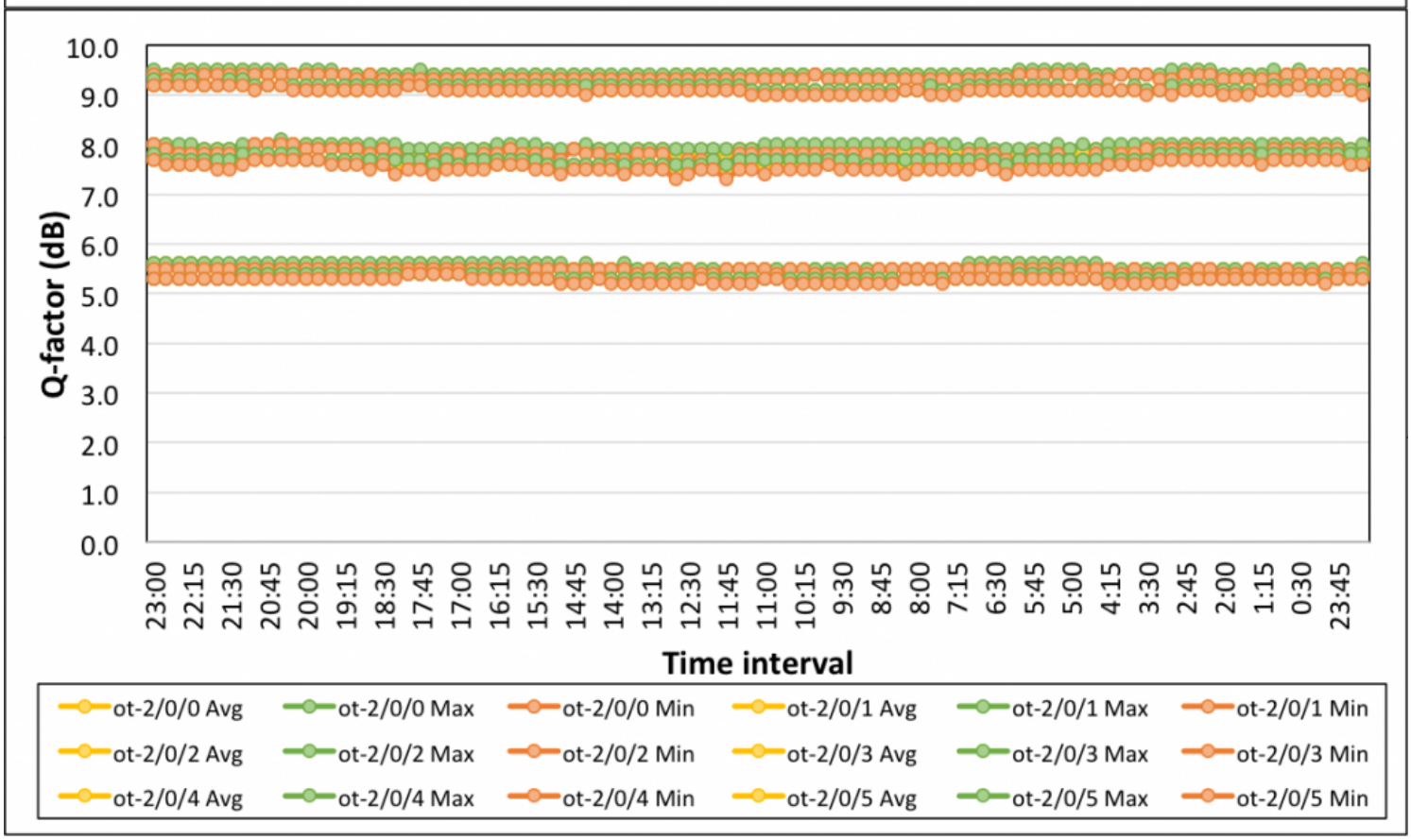
Why AC400? What does it mean? Its the name of the DSP that Juniper buys to realize their Coherent portfolio and its a Commercial Off The Shelf product from Acacia. This chipset can be found in many other DWDM-equipments aswell as these type of DCI-style switchcards but also in more traditional transponders. Acacia is essentially the optical equivalent of what Broadcom is in the switching-world.

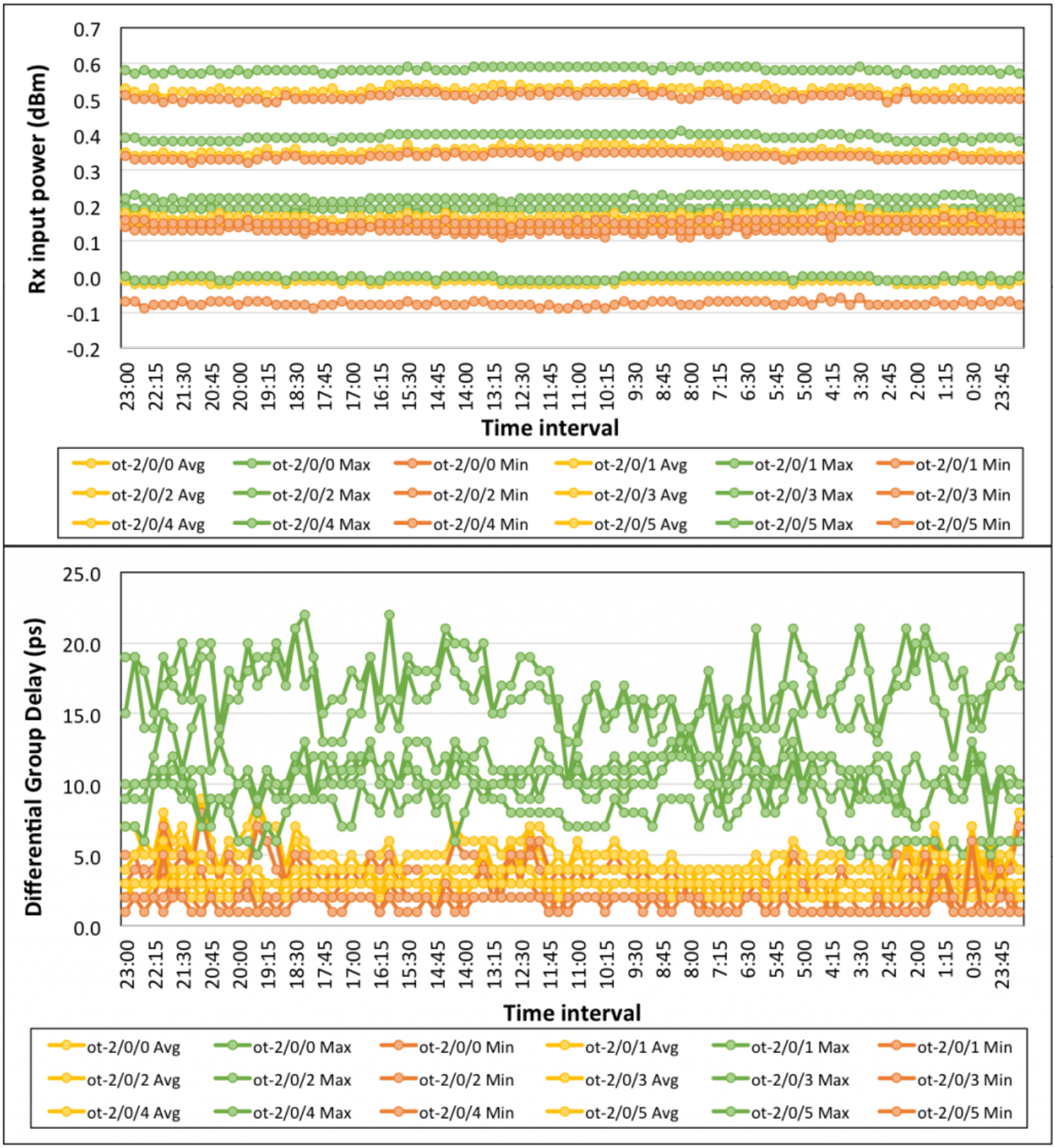
```
root@voodoo-1# run show interfaces transport pm optics current ot-3/0/0
Physical interface: ot-3/0/0, SNMP ifIndex 564
  13:15-current
  Suspect Flag:False          Reason:Not Applicable
    PM                              CURRENT   MIN       MAX       AVG       THRESHOLD       TCA-ENABLED       TCA-RAISED
                                                                            (MIN)  (MAX)     (MIN)  (MAX)      (MIN)  (MAX)
  Lane chromatic dispersion(ps/nm)  4905      4902      4910      4903      0      0         NA     NA         NA     NA
  Lane differential group delay(ps) 1         1         3         1         0      0         NA     NA         NA     NA
  Lane Q2 factor(0.1dB)             159       157       159       158       0      0         NA     NA         NA     NA
  SNR(0.1dB)                        165       164       167       165       0      0         NA     NA         NA     NA
  Carrier frequency offset(MHz)     -878      -878      -771      -785      -3600  3600      No     No         No     No
  Tx output power(0.01dBm)          1         -2497     2         -106      -1200  1000      No     No         No     No
  Rx input total power(0.01dBm)     24        8         24        9         -3200  1000      No     No         No     No
  Module temperature(Celsius)       52        52        52        52        -5     75        No     No         No     No
  FEC        Suspect Flag:False          Reason:Not Applicable
    PM                    COUNT          THRESHOLD       TCA-ENABLED     TCA-RAISED
  FEC-CorrectedErr         3590            0               NA              NA
  FEC-UncorrectedWords     0               0               NA              NA
  BER        Suspect Flag:False          Reason:Not Applicable
    PM                  MIN       MAX       AVG       THRESHOLD       TCA-ENABLED       TCA-RAISED
  BER                   0.0e-0    5.0e-9    2.3e-10    0.0e-0          No                No

bergroth@fre-r1-re0> ...rt pm optics current et-11/0/0
Physical interface: et-11/0/0, SNMP ifIndex 662
  13:15-current
  Suspect Flag:False          Reason:Not Applicable
    PM                              CURRENT   MIN       MAX       AVG       THRESHOLD       TCA-ENABLED       TCA-RAISED
                                                                            (MIN)  (MAX)     (MIN)  (MAX)      (MIN)  (MAX)
  Lane chromatic dispersion(ps/nm)  4731      0         5553      4803      0      0         NA     NA         NA     NA
  Lane differential group delay(ps) 4         3         5         4         0      0         NA     NA         NA     NA
  Lane Q2 factor(0.1dB)             144       143       145       144       0      0         NA     NA         NA     NA
  SNR(0.1dB)                        150       144       150       145       0      0         NA     NA         NA     NA
  Carrier frequency offset(MHz)     -740      -860      -740      -763      -3600  3600      No     No         No     No
  Tx output power(0.01dBm)          0         -3        3         0         -1100  300       No     No         No     No
  Rx input total power(0.01dBm)     -66       -216      -55       -68       -3000  300       No     No         No     No
  Module temperature(Celsius)       33        33        33        33        -5     75        No     No         No     No
```

We take these cross-compability issues quite seriously and would really like to prefer vendors to stop playing games. This was a important test for us to make sure that it actually worked and it did after we did some....voodoo on it. In theory any vendors box that has this same chipset should work. And even better is if the vendor also implement staircase-fec which is a generic line-standard that anyone can use and it is being used. Read more about that on Kristian Larssons blog from how Terastream is testing 100G DWDM lineside interop. CLICK ME

MORE NICE GRAPHS FOR THE VOODOO LINKS.


Electrical SNR (dB)
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
23:00 22:15 21:30 20:45 20:00 19:15 18:30 17:45 17:00 16:15 15:30 14:45 14:00 13:15 12:30 11:45 11:00 10:15 9:30 8:45 8:00 7:15 6:30 5:45 5:00 4:15 3:30 2:45 2:00 1:15 0:30 23:45
Time interval
ot-2/0/0 Avg
ot-2/0/0 Max
ot-2/0/0 Min
ot-2/0/1 Avg
ot-2/0/1 Max
ot-2/0/1 Min
ot-2/0/2 Avg
ot-2/0/2 Max
ot-2/0/2 Min
ot-2/0/3 Avg
ot-2/0/3 Max
ot-2/0/3 Min
ot-2/0/4 Avg
ot-2/0/4 Max
ot-2/0/4 Min
ot-2/0/5 Avg
ot-2/0/5 Max
ot-2/0/5 Min


Q-factor (dB)
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
23:00 22:15 21:30 20:45 20:00 19:15 18:30 17:45 17:00 16:15 15:30 14:45 14:00 13:15 12:30 11:45 11:00 10:15 9:30 8:45 8:00 7:15 6:30 5:45 5:00 4:15 3:30 2:45 2:00 1:15 0:30 23:45
Time interval
ot-2/0/0 Avg
ot-2/0/0 Max
ot-2/0/0 Min
ot-2/0/1 Avg
ot-2/0/1 Max
ot-2/0/1 Min
ot-2/0/2 Avg
ot-2/0/2 Max
ot-2/0/2 Min
ot-2/0/3 Avg
ot-2/0/3 Max
ot-2/0/3 Min
ot-2/0/4 Avg
ot-2/0/4 Max
ot-2/0/4 Min
ot-2/0/5 Avg
ot-2/0/5 Max
ot-2/0/5 Min



These tests are always fun to make. It proves to us that we did our due diligence when we built our optical network from the bottom-up. The network is optimized for highest possible OSNR and it pays off now when we add signals with advanced modulation, we are able to actually run these stuff up to the maximum spec more or less and whenever we need to add 200G signals we know we are already in a good position to add these without to much hassle.

With TeliaCarrier press-release about running TIP Voyager opto-whitebox from Stockholm to Hamburg on a 1089km stretch between Stockholm and Hamburg on the Acacia 16QAM modulation format (voyager and this card uses the same DSP) we just have one reaction up here in the 1900km club.

// Dennis, Magnus and hugge from SUNET test-labs

Skriven av

FREDRIK “HUGGE” KORSBÄCK

Network architect and chaosmonkey for AS1653 and AS2603. Fluent in BGP hugge@nordu.net