SUNET

FINAL POC AND ACCEPTANCE OF ROUTERS

Now it's just a matter of days until we start the rollout of the new network. All optical equipment is in-transit to Sweden and the routing-gear is being sent out in small batches whenever it rolls off the assembly line somewhere in China. We are expecting to receive items more or less every week from now to midsummer. To be fully prepared for the buildout and also to make sure all bugs and concerns is being adressed accordingly a small team from SUNET was sent out to Juniper WorldWide Proof-of-Concept lab in Sunnyvale California for a week to make a full-scale proof-of-concept with the exact gear the new network will consist off.

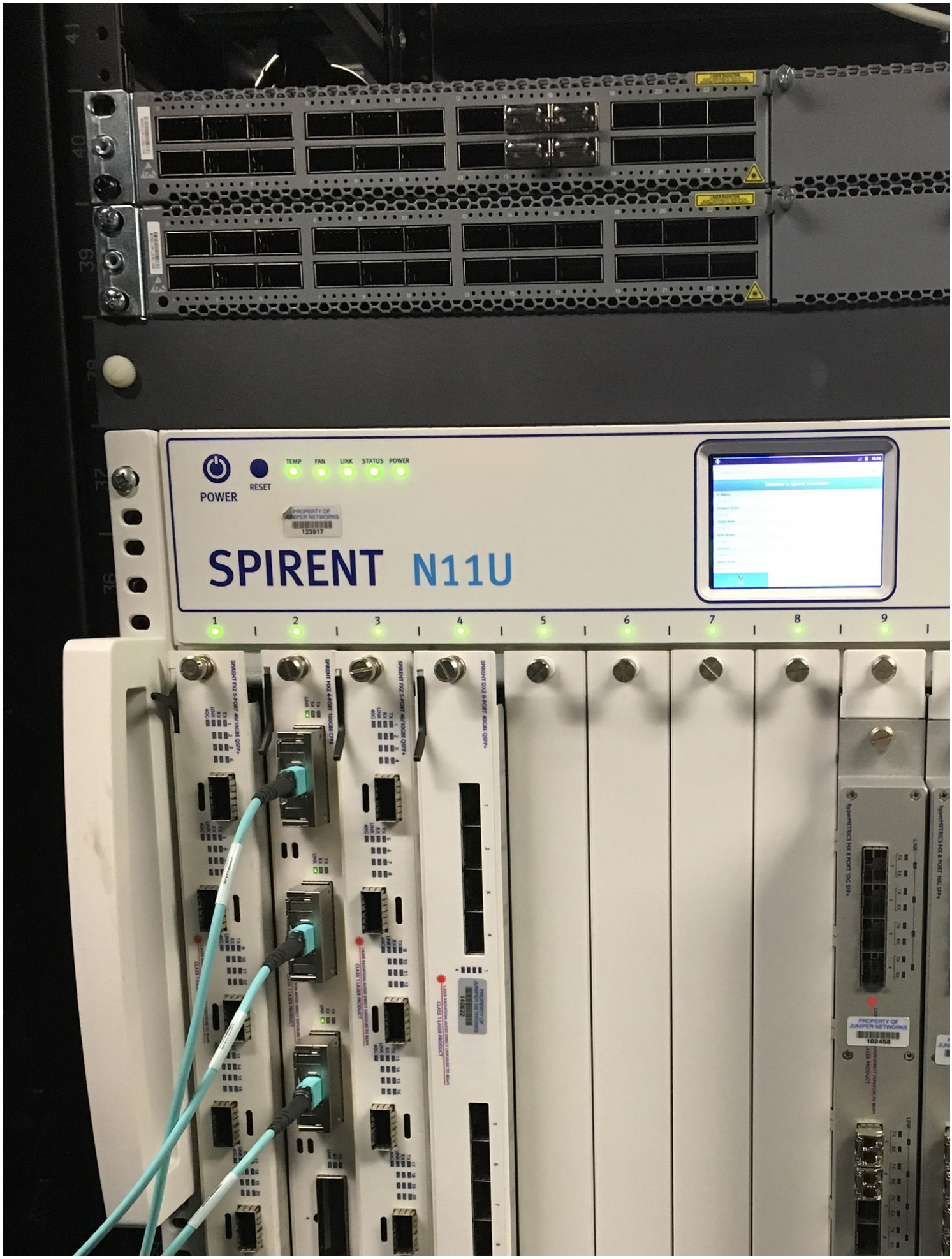
The idea behind doing a a real full-scale acceptance-test is that we are very cautious of bugs and misbehaviours, especially from the Coherent 100G cards which only work with unreleased beta-software currently. We are the first customer in the world that will run this card in a production network so we are expecting a bumpy start. We had decent success with the engineering-sample cards we tested a few months ago and we got a production-worthy link up and running between Västerås and Stockholm without that much hassle, this was on a nightly custom build of Junos. Naturally there were some bugs and weird behaviours noticed that we documented thoroughly and sent in as feedback back to Juniper.

To build our PoC-setup we used 7 routers, all of which we will have in the new network. 2xMX480, 3xMX960, 1xMX2020 and then a MX80. We populated this with the linecards we will have in the new network to make sure we get hit by all bugs at this stage already and have a chance of avoiding them in the production-ready network.


CABINET G08
CABINET G07
juniper
MX2020

MX2020, one of the biggest routers on th

A proof-of-concept lab like this consists of many moving parts, we spent a major part of december writing together a set of tests which is supposed to make sure and validate all of our past and present configuration, making sure that our old stuff works and that our new ideas is also possible to actually build with new platforms. This is mainly to verify full protocol functionality and also to verify new functions and making sure we understand the implications of new features. The second major part of a PoC-lab is to verify hardware. Making sure that various versions of hardware work together and that all announced specifications is actually being honored.


POWER
RESET
TEMP
FAN
LINK
STATUS
POWER
SPIRENT N11U
1
2
3
4
5
6
7
8
9

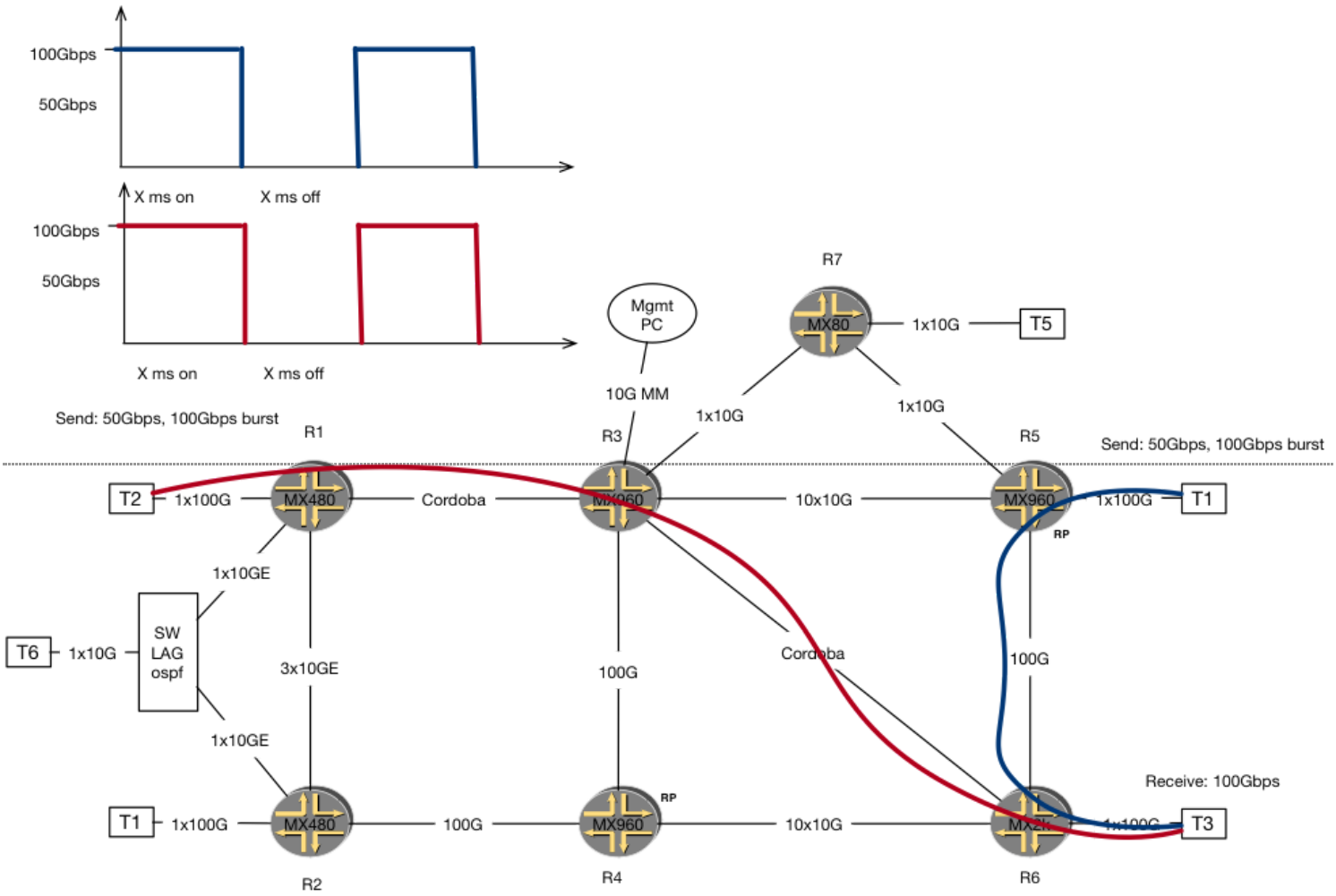
Knowing the limitations of our hardware is of uttermost importance, marketed specifications is as with everything, very optimistic numbers. Think of linecard performance specifications the same way as we think nowadays of german diesel-car emissions 🙂 It's always stated under the best possible conditions with the optimal type of traffic and trafficflows. To get to know the hardware better we naturally requested to get access to the biggest packet-cannons they have and try to get the boxes to perform as horrible as possible so we got access to IXIA and Spirent packet-generators. Pasted below is an example of one of many buffer-tests we did for the MX2020 router.

————————————————————————————-

1 Traffic tests

1.1 MX2k buffering

1.1.1.1 Traffic flows



T2 sends traffic towards T3 at 100Gbps, to verify lowest packet size without dropping packets. Use that packet size for the rest of the test.

T2 sends traffic to T3. 50Gbps average burst 100Gbps. X ms at 100Gbps 2.X ms at 0Gbps and so on.

T1 sends traffic to T4 with the same pattern as T2. T1 and T2 should start sending at exactly the same time.

Test with both ipv4 and ipv6.

1.1.1.2 Setup

All routers in as1653, ip lookup on all routers

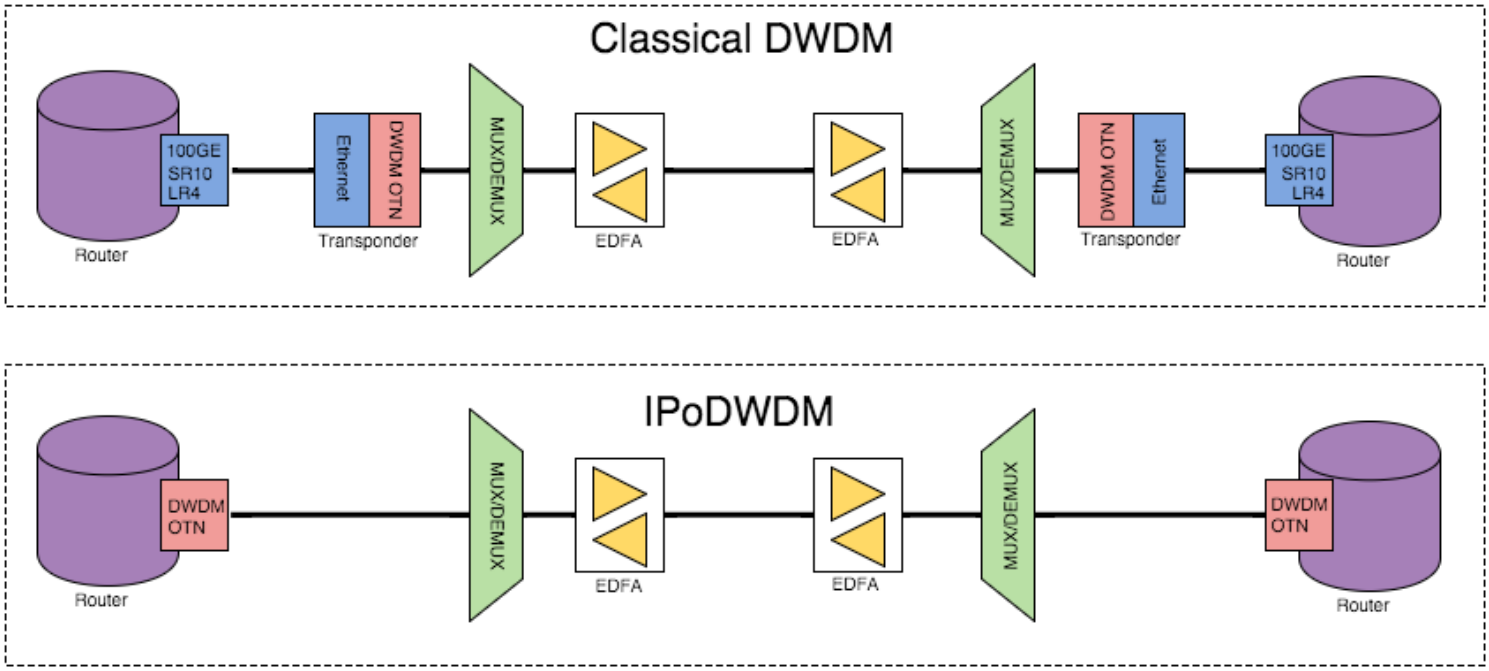
1.1.1.3 Tests

1. Verify lowest packet size for 100Gbps without dropping packets – ***174 bytes***
2. Test with X=100ms duty cycle.
3. Find largest burst that could be sent without loosing packets at receiver T3. ***With TOS 0, no loss. 7,4ms burst gap***

First look into the lab, lots of eag

What differs this proof-of-concept from many others is that we we will be the first customer that tries the coherent 100G cards for the MX-platform out. The coherent cards (codename Cordoba) isn't just "a new card" such as you expect any vendor to release every once in awhile. These types of cards fundamentally change how you think and build a nationwide (or continental) network, a coherent IPoDWDM-card as a concept is not really new, however having these types of card in a PE-type of device has not been seen before. Core-routing platforms such as Juniper PTX, Alcatel-Lucent XRS 7950, Cisco CRS and NCS6000 and also Huawei NetEngine has had these style of cards for a little while but the use-case is slightly different then what we see here, these aforementioned boxes is more or less pure core-devices. Getting IPoDWDM down to the PE-level enables most operators to actually

cut out the need for transponders in the legacy-style DWDM completely (yes, i nowadays call regular DWDM-setups for legacy, hopefully it offends someone). This is obviously a good thing for everyone (except the DWDM-vendors) since you can cut out a product completely from your optical network, this thing you are cutting out is also by far the most expensive component. We have seen trends of DWDM-vendors ”sorting” this out with a licensing-fee when using external transponders, which will probably a short-lived way of trying to earn money.



There is alot of buzz in the market that ”DCI” will be one of the biggest expanding markets the next few years and now that we have already seen vendors (See Arista 7500) start to put DWDM-linecards in their datacenter-switches, really interesting things will definitely happen to the packet-transport business now that it’s not a isolated segment anymore.

Right, so testing Coherent cards. We are not real optical engineers at this company so hopefully these tests does not make real optical engineers angry by being completely wrong.

———————————

1 Cordoba

1.1.1 Verify BER and uncorrected words.

Have traffic flowing over both Cordoba links in both directions. Degrade with VOA until uncorrected words starts counting up. Verify that interface errors doesn’t start before.

1.1.2 Verify output power limitations.

Configure CLI max and min values and measure using the DOM on the input interface.

1.1.3 Verify, optical counters

- Combined input power
- Wave input power
- CD
- SNR
- Group delay

1.1.4 Verify Wave center freq and width in GHz

Provide Optical spectrum analyzer and measure, both ends of the c-band.

1.1.5 Telemetry

Verify that we are able to extract statistics and values for the Cordoba-card through SNMP and streaming Telemetry

1.1.6 Configuration

Verify that we can configure and set Cordoba specific parameters through SNMP and most importantly through Netconf

1.1.7 Wavelength Sweep

Verify that the card is not sweeping or crossing other wavelengths at any point in time when tuning or adjusting levels.

————————————————

The card performed as per specification more or less which is good, this post is long enough anyway so i wont go into every detail, we even managed to run the card at -37dbm of input power and still not seeing packetloss (FEC was working hard though)

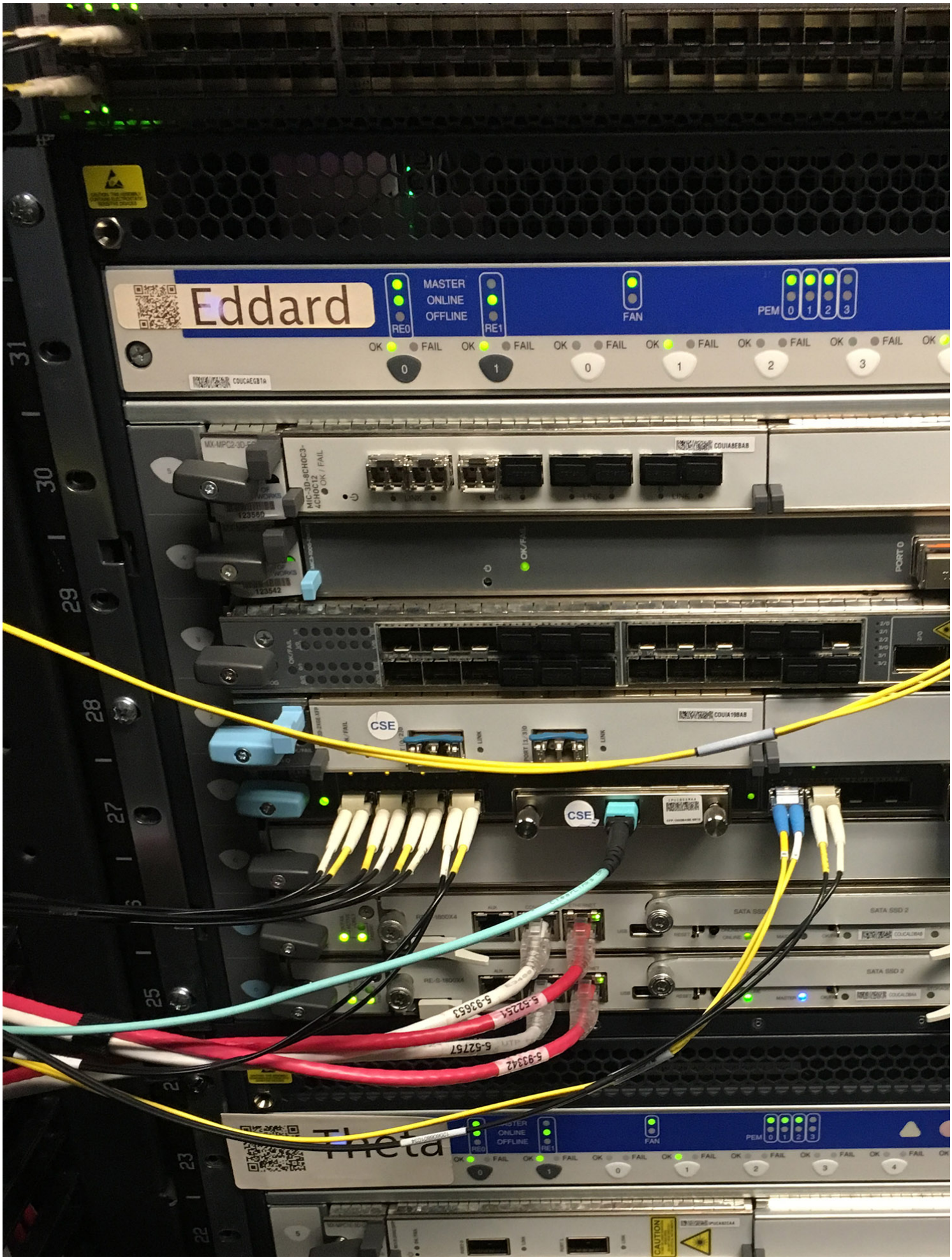
There is two things i would like to point at however which some of you might think ”why does this matter?”

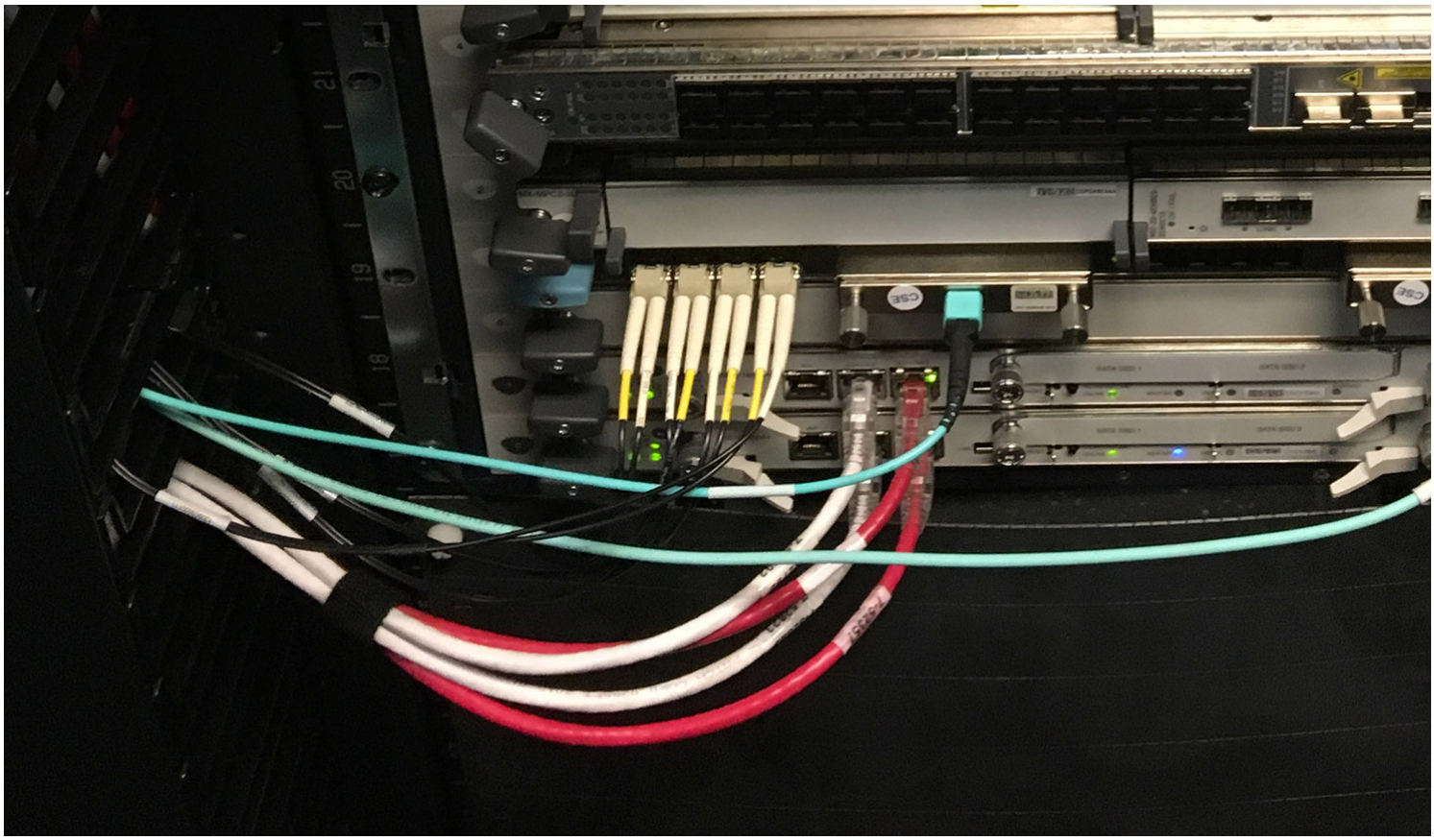
1.1.7 *Wavelength sweep?* ”Why does this matter, you will have a MUX/DEMUX with a grid anyway that takes care of this”. Yes. This is usually the case, however we will not build this network with any fixed-grid devices at all. All of our client-facing filters (not actually a filter) will be completely gridless, just splitters and combiners linked together in a box, a optical hub. In some iterations of these IPoDWDM-cards we have seen a behaviour where the transciever would sweep over the spectrum when tuning itself into the correct channel. Not a problem when running with fixed-grid but a huge problem when running gridless since it will kill all channels for a short while when sweeping.

The reason why we build with gridless is many. It’s cheaper, it’s cooler and also it enables us to run any size (in terms of how wide the channel is) wavelength on the network. If it happens to be that the latest and greated terabit-cards will need to use 127.5Ghz of spectrum, we have the possibility to do so.

1.1.3 *Wave Input Power.* This is tightly related with **1.1.7**, since our optical network will function very much alike a FM-radio we will at any given time and point in the network have multiple waves hitting every receiver in the network, the receiver needs to tune to it’s specified wavelength and just ignore the rest of the light coming in. This is where a classical fixed grid mux/demux comes into play since it will only pass through waves of the correct frequency. We are running gridless so the receiver needs to do it job here. This is why it’s crucial for us to be able to measure lightlevels on the tuned wavelength of the receiver, and not only total light coming in. In a troubleshooting-scenario it will be very hard to find out where the fault could actually be when all interfaces in the network is able to see light almost all the time, except if the fault is very local.

This did not unfortunately NOT work in the PoC so this was bounced back to software engineering to present a solution for us before the equipment is shipping. We know the optics has the hardware in place to be able to accommodate this so we hope for a quick fix. The idea is that the beta-software we have been running for both the PoC in Sunnyvale but also on the live-fiber in Stockholm will turn into official Junos release 15.1F5 when first cards is shipped out to us. Which is in two weeks from now… well see how that goes.


Eddard
MASTER
ONLINE
OFFLINE
RE0
RE1
FAN
PEM
0 1 2 3
OK FAIL
CSE
CAUTION



All in all the PoC went great. We did not encounter any major problems, except the thing with wavelength-power on the coherent-cards. Last time we tested the cards before christmas between Stockholm and Västerås we found 16 different bugs and all of them got solved and verified during this PoC so that was both surprising and good.
We hope that we have made the correct choice of going IPoDWDM full-on for this network, immense time and effort has been spent from the engineering-department to make sure this could actually work.

The first equipment has arrived from ADVA so we have started building two amplifier-sites outside of Stockholm and also the main dropnode (yes, we will still be running traditional ROADMs) inside Tulegatan is being built out at this very moment. Hopefully we can also build Fredhäll this week finalizing both the core-nodes here in Stockholm.

No equipment from Juniper has arrived yet but the preparation for housing the core MX2020 in Stockholm is almost done, the carpenters has reinforced the flooring to withstand 1000kg/m2 in the area where the router is to be located and the electricians has been working hard with sourcing enough power feeds to get the box to start.

Not looking forward carrying this up to the third floor without an elevator

Every week from now on we expect to roll new sites out and our idea is that all core and longhaul-links is to be done before summer. So we can start installing and swapping customers over to the new network gradually.

Skriven av

FREDRIK “HUGGE” KORSBÄCK

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