Interoperable 400ZR Deployment at Cloud Scale

Chuan Qin*, Binbin Guan, Kyle Edwards, Jian Kong, Ryan Morgan, Yawei Yin, Avinash Pathak, Mounika Banda, Sridharan J, Govardan Chandrababu, Jeetesh Jain, Jamie Gaudette

Azure Networking, Microsoft Corporation, Redmond, Washington 98052, USA
Azure Networking, Microsoft Corporation, India.
Email: * qinchuan@microsoft.com

Abstract: We report 400ZR deployment data from the Microsoft private network, highlighting module interoperability, performance stability and availability, and parallel module firmware upgrade at cloud scale.

OCIS codes: (060.1660) Coherent communications, (060.2360) Fiber optics links and subsystems, (060.4250) Networks.

1. Introduction

From individual customers to fortune 500 companies, people and organizations across the globe rely on the cloud to live their lives and run their businesses. The cloud must always be available. To accommodate the near-perfect service availability, we leverage a distributed datacenter architecture, using long-haul fiber to connect different regions globally, and using metro fiber to connect server clusters spread across the same region as if they were in one contiguous mega-campus. The center of the distributed architecture is clean, diverse, and cost-effective metro optical fiber running DWDM optics. MSFT has leveraged the PAM4 technology to build such a DWDM metro network [1,2]. However, the Microsoft Network experienced 40 times peak demand-growth since the COVID-19 outbreak in 2020 and the hybrid workspace is being adopted worldwide [3]. The accelerated need for an efficient 400G transport is driving Azure to deploy 400G-ZR technology [4]. In this paper, we review the performance of 400G-ZR from the following three perspectives. 1) Interoperability among module vendors, 2) Parallel module firmware upgrade by automation, and 3) Networking stability inspected from key optical and DSP parameters.

2. 400ZR deployment

Fig. 1 (a) describes a typical unprotected 400ZR point-to-point system. The packet routers host 36x400G linecards supporting QSFP-DD pluggable modules. The 400G line systems feature 64 channels with 75-GHz grid spacing enabling 25.6 Tb/s on a single pair of fibers. The number of routers at A and Z side depends on the application of the span, and multiple routers are often combined (striped) across this single fiber pair. Fig. 1 (b) shows the distribution of estimated fiber span loss in Azure metro networks.

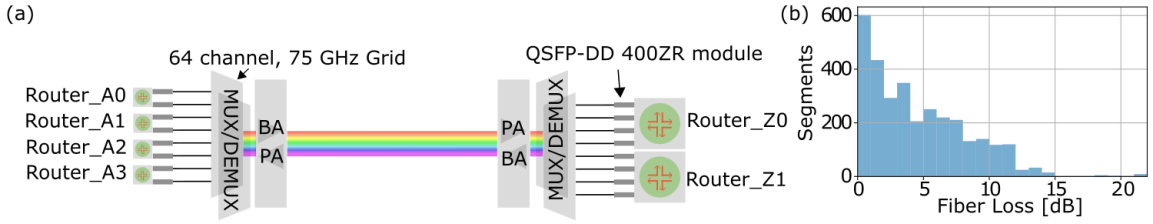

Fig. 1 (a) 400ZR use case in Microsoft Azure Networking. BA: Booster Amplifier; PA: Pre-Amplifier. (b) Estimated span loss histogram in Microsoft's metro networks

Table 1 Interoperability test with TX to RX variations in an 80-km G.652 fiber transmission.

TX to RX	PreFEC-BER	CD [ps/nm]	PDL [dB]	DGD [ps]	CFO [MHz]	Estimated OSNR [dB]	Actual OSNR [dB]	Case temp [ºC]
X to X	1.27e-3	1354	1.3	2.0	-221	35.0	35.6	38.0
X to Y	4.16e-4	1296	0.5	3.6	108	29.9	35.6	52.0
Y to X	1.87e-3	1392	0.5	2.0	-127	34.4	35.2	40.0
Y to Y	5.73e-4	1203	0.5	4.4	15	28.9	34.6	55.7

1.1. Interoperability among different module vendors

After years of standardization efforts from Optical Internetworking Forum (OIF), industry witnesses the first interoperable coherent pluggable optics - 400ZR, and its implementation agreement [5]. Interoperability provides more options on module selection and drives down the cost. Recent Plugfest result mostly focused on optical link-up and rOSNR when testing interoperability [6]. However, the optical and DSP metrics reported by the 400ZR modules are also critical for network operation and troubleshooting. The requirement is non-trivial that a receiver correctly understand and estimate the performance metrics of signal from an alien transmitter.

Table 1 shows the DSP reported values from the RX module when we have four bookended or interoperable scenarios exist in one span with line systems and 80-km G.652 fiber transmission. These values show that there is little difference between bookended and interoperable scenarios, not only network performance-wise, but also with monitoring ability in a vendor-agnostic way.

1.2. Parallel Module Firmware Upgrade Automation for All Module Vendor and All Router Platforms

To achieve fast deployment goal, we take minimalist approach, abstract domain specific complexity for all routers, and apply zero touch provisioning (ZTP) during a span turn up. However, in a span of up to 128 modules, it can be extremely time-consuming and difficult to manage if firmware upgrade is done in a serial manner. The automation tool for parallel firmware upgrade serves the purpose of reliable and fast deployment.

Fig. 2 (a) depicts how we automate and expedite the process by running parallel upgrade among different routers, linecards, and controller buses. In the current 36-slot linecards, 5 controller buses can run firmware upgrade on 5 modules simultaneously. This limits the entire upgrade process to complete within 120 minutes assuming 15 minutes per module and at least one controller bus is shared by 8 400ZR modules. The parallel upgrade scheme vastly reduces the time to 1/8 compared to firmware upgrade done in a complete serial order within a linecard. The entire parallel upgrade process is fully common management interface specification (CMIS) compliant regardless of module vendors. The automation also features an abstraction layer based on Python netmiko package to neglect router vendor dependency by only calling router specific driver functions in the lower layer. These design features of automation enable full interoperability at both module and router levels. Fig. 2 (b) shows the upgrade diagram. It starts by querying all the 400ZR modules in the span of interest, and then categorizes modules into different upgrade groups by their location in the router linecard and ports. After a complete set of firmware for all module vendors gets delivered to all the routers, parallel process begins from Step 4, pre-upgrade check and runs throughout the whole process. The pre-upgrade check step inspects module presence, obtains the active running firmware and confirms traffic drainage on the module if it needs an upgrade. The firmware upgrade step is the lengthiest process, and the bulk data block transfer fully occupies a controller bus. Between different controller buses, these processes run in parallel. After firmware upgrade finishes, a post-upgrade check guarantees the active running firmware matches the desired version. Link up check is optional if the span is at the deployment stage. It would be necessary if the span is in production. It confirms all links are up before shifting back traffic after the maintenance window for firmware upgrade.

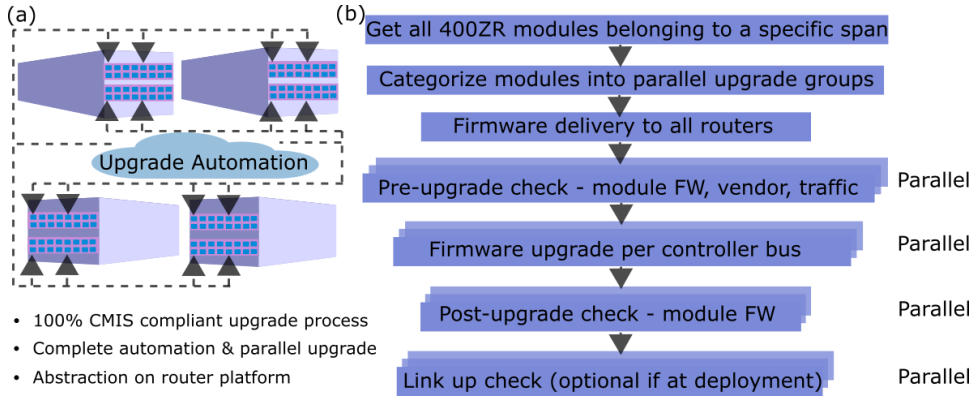

Fig. 2 Parallel 400ZR module firmware upgrade on different router, linecard and controller bus.

1.3. Network stability

Network stability and availability is the highest priority for a cloud service provider. Metric polling for all the 400ZR modules in production is performed proactively via CLI about every 20 minutes. Metrics for line systems are polled via REST API approximately every 10 minutes. All the data from routers and line systems are categorized and

converted to a universal, vendor-agnostic format in the central database. Alerts are generated based on Syslog via UDP streaming and are stored into the same central database. The performance metrics and alerts jointly endow network management with simplicity and visibility, making the troubleshooting process happen promptly and without the need of optical expertise.

Fig. 3 (a) shows the violin plots on Pre-FEC BER, TX and RX power distribution over 5 months on 14 ports of one router. Red dots on each violin plot show the initial values. Each port has 7490 points on Pre-FEC BER, TX and RX power, respectively. One module TX and another module RX power show only one point far from the central distribution due to data logging happening at the time when linecard restarted for maintenance and modules turn up. The distribution is very tight overall: the average standard deviation for Q^2-factor is 0.06 dB, for TX power 0.035 dBm, and for RX power 0.047 dBm. Fig. 3 (b) shows the Q^2-factor variation on 64 routers, approximately 700 400ZR modules over a 1.5-month period. The 0.1 and 99.9 percentile are at -0.63 and 0.53 dB, respectively. Fig. 3 (c) shows the module case temperature on these ~700 modules in real time, covering different modules, router vendors and various operation environment around the world. The median case temperature is below 50 ºC. Only 8 modules are currently operating above 70 ºC (~ 1% of all modules) and all of them are below 72 ºC. While we only display a few most important items here, we can also collect other optical and DSP PMs to guarantee excellent monitor ability on the network and facilitate network operation effort on fault behavior identification and troubleshooting.

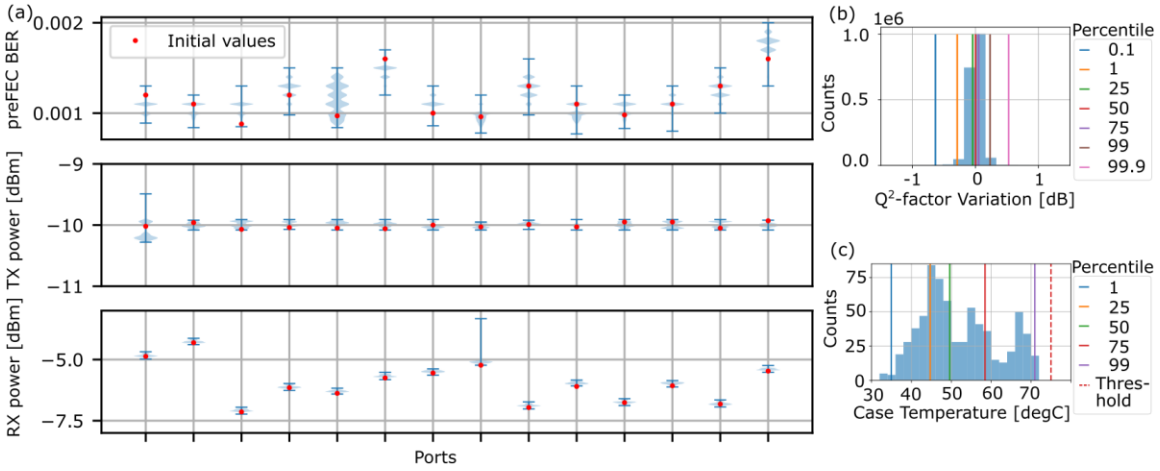

Fig. 3 (a) Five-month production data for Pre-FEC BER, TX and RX power monitoring on 14 ports from one router. Red dots: Initial values. (b) 1.5-month Q^2-factor variation histogram on each port from 64 routers, ~700 modules and over 1.87 million data points. (c) Module case temperature on ~700 modules.

3. Summary

We shared the experience in deployment of 400ZR and summarized it into three topics highlighting interoperability, tooling and automation on parallel firmware upgrade, and network stability. The optical and DSP performance metrics show stable performance over a few months' timeframe. 400ZR paves the way to the ultimate goal featuring high deployment velocity, interoperability, quality, stability and availability.

4. References

[1] M. Filer, S. Searcy, Y. Fu, R. Nagarajan, and S. Tibuleac, "Demonstration and performance analysis of 4 Tb/s DWDM metro-DCI system with 100G PAM4 QSFP28 modules," 2017 Optical Fiber Communications Conference and Exhibition (OFC), 2017, pp. 1-3.
[2] M. Filer, Jamie Gaudette, Yawei Yin, Denizcan Billor, Zahra Bakhtiari, and Jeffrey L. Cox, "Low-margin optical networking at cloud scale," J. Opt. Commun. Netw. **11**, C94-C108 (2019).
[3] Azure responds to COVID-19, Azure responds to COVID-19 | Azure Blog and Updates | Microsoft Azure
[4] P. Wright, R. Davey and A. Lord, "Cost Model Comparison of ZR/ZR+ Modules Against Traditional WDM Transponders for 400G IP/WDM Core Networks," 2020 European Conference on Optical Communications (ECOC), 2020, pp. 1-4.
[5] OIF, 400ZR Implementation Agreement: https://www.oiforum.com/wp-content/uploads/OIF-400ZR-01.0_reduced2.pdf.
[6] E. Pincemin et al., "End-to-End Interoperable 400-GbE Optical Communications through 2-km 400GBASE-FR4, 8x100-km 400G-OpenROADM and 125-km 400-ZR Fiber Lines," in Journal of Lightwave Technology, 2022