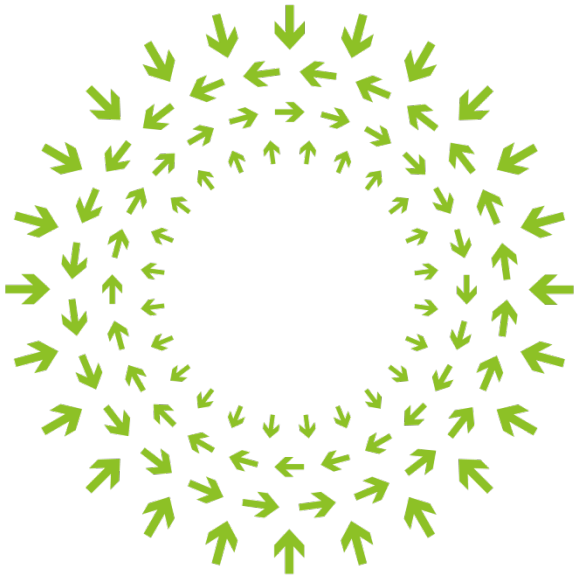

OPEN Compute Project

Facebook: Fabric Aggregator

Modular Design to Address our Traffic Demands

Authors:

João Ferreira, Network Engineer, Facebook

Naader Hasani, Network Engineer, Facebook

Jimmy Leung, Mechanical Engineer, Facebook

Zhiping Yao, Network Engineer, Facebook

Brian Taylor, Network Engineer, Facebook

Nina Schiff, Software Engineer, Facebook

Sree Sankar, Technical Program Manager, Facebook

1. Revision History

Date/Version	Name	Description
3/15/2018, v1.0	João Ferreira, Jimmy Leung, Naader Hasani, Nina Schiff, Brian Taylor, Sree Sankar, Zhiping Yao	1st public version of Fabric Aggregator specification

© 2018 Facebook.

As of March 15, 2018, the following persons or entities have made this Specification available under the Open Compute Project Hardware License (Permissive) Version 1.0 (OCPHL-P), which is available at http://www.opencompute.org/community/get-involved/spec-submission-process/.

Facebook, Inc.

Your use of this Specification may be subject to other third party rights. THIS SPECIFICATION IS PROVIDED "AS IS." The contributors expressly disclaim any warranties (express, implied, or otherwise), including implied warranties of merchantability, non-infringement, fitness for a particular purpose, or title, related to the Specification. The Specification implementer and user assume the entire risk as to implementing or otherwise using the Specification. IN NO EVENT WILL ANY PARTY BE LIABLE TO ANY OTHER PARTY FOR LOST PROFITS OR ANY FORM OF INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION OR ITS GOVERNING AGREEMENT, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND WHETHER OR NOT THE OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

2. Scope

This document defines a technical specification for the Fabric Aggregator modular solution for traffic aggregation in large-scale data center applications. This specification is being submitted to the Open Compute Foundation.

3. Contents

4. Overview

This document defines a specification for a distributed unit of capacity designed to support large-scale traffic demands between data centers. This specification is currently being deployed in Facebook's data center networks and is being presented for submission into the OCP Networking Group to be shared with the OCP community.

5. Fabric Aggregation Challenges in Data Center Operations

Recently we announced that Facebook will be breaking ground in Georgia for its 12th data center, worldwide, and will be tripling the size of its Papillion, Nebraska, data center (from two buildings to six). As our community continues to grow and we create more immersive experiences through videos, live video, 360-degree photos, and virtual reality experiences that require additional capacity, our ongoing challenge is scaling the Facebook network that interconnects more than 2 billion people.

The **fabric aggregation** layer in our network architecture is the tier that interconnects all the fabrics in all the data center buildings *within a region*. We refer to this traffic as *east/west traffic*.

This layer also acts as a point of aggregation for all traffic *exiting and entering* a region. We refer to this traffic as *north/south traffic*.

To address this situation, we set out to build a network system that could adapt to larger regions, changing services, and varied traffic patterns. A very large general-purpose network chassis – which had been the model for our first two fabric aggregation systems -- no longer met our needs in terms of scale, power efficiency, and flexibility.

Fabric Aggregator arose from our effort to build a completely distributed network system through assembling simple, open, and already available building blocks like Facebook's Wedge 100S switches and FBOSS software to meet the east/west and north/south traffic demands on the fabric aggregation tier.

The disaggregated strategy also allowed us to define a generic framework that can be reused in other parts of the network, thus avoiding the need to build a new chassis for each network tier.

Figure 1 illustrates that regions are a collection of data center facilities in the same geographic area. Each data center has its own fabric, serving all data center local traffic (intra data center flows), while all fabrics in a region are aggregated by a top-level system: the Fabric Aggregator.

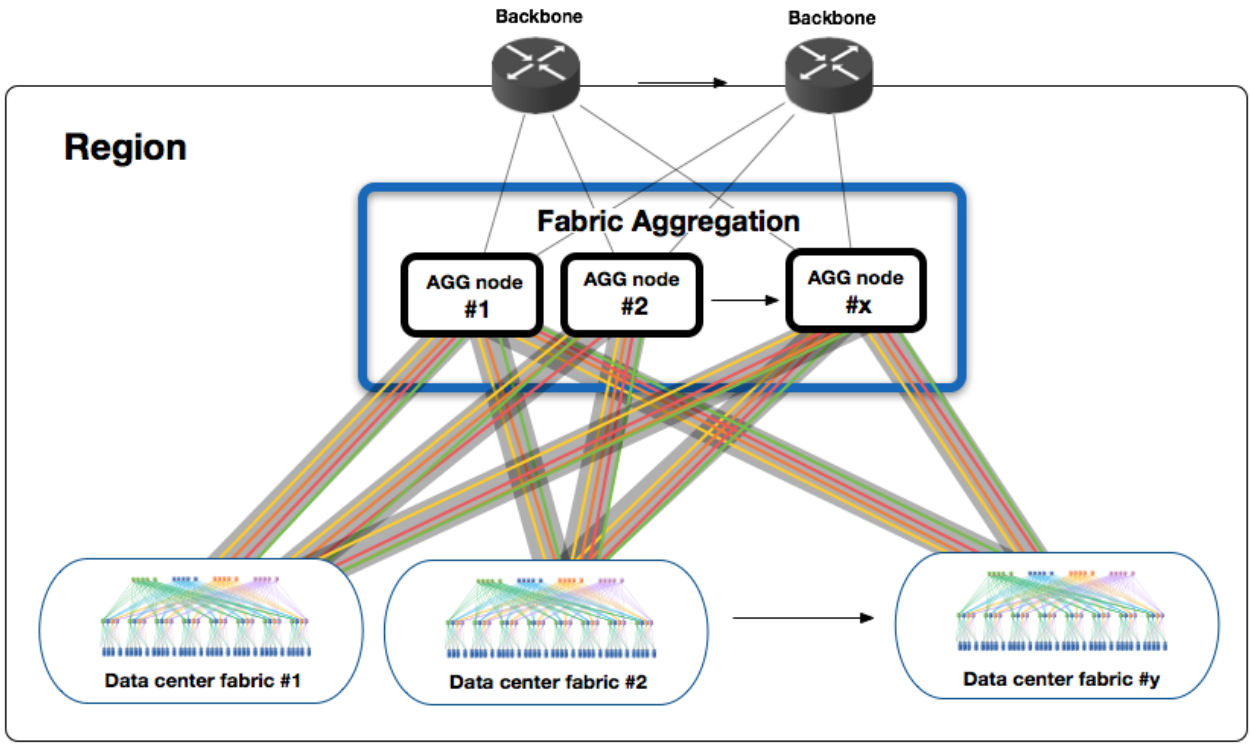

Figure 1: The Fabric Aggregator aggregates all data center fabrics within a region.

All traffic that leaves or enters any of the Facebook data centers is handled by the Fabric Aggregation system. There are multiple growth vectors and pressure points, primarily:

- The region size (the number of fabrics per region), and
- Capacity demand from each fabric.

Capacity is highly dependent on the service density in a particular data center, and its interdependency with other (remote) services. There is no convenient one-size-fits-all solution to apply to Facebook's data centers network.

6. A Modular Solution to Fabric Aggregation

The option that we chose was *integration*. By combining a group of Facebook Wedge 100S switches into a virtual unit, we were able to deliver a new, large-capacity distributed solution.

As noted above, this approach took advantage of Facebook's provisioning infrastructure, FBOSS software, operational metrics, and well-established in-house expertise designing rack solutions to support integrated server infrastructure.

Added benefits were future-proofing with the Wedge building blocks supporting exploitation of new ASIC and optics technologies, as they became available, without additional hardware resources. Other advantages included better power resiliency:

One feed failure wouldn't bring down the whole virtual chassis. If a Wedge fails for some reason within our distributed system, it's a very minimal failure domain. The parent node itself remains operational.

Finally, Facebook would have more control for the FA layer, as well as for its TOR and fabric-switch layers. With the same platform for all data center roles, Facebook could iterate quickly using just one basic building block, test in smaller portions of the network before roll-out, and respond more quickly to failure.

7. Architecting Node Implementation

The underlying architecture for implementing a node relies on a two-layer cross-connect topology. As diagrammed in Figure 2, a **downstream layer** "localizes" all East-west traffic demands – those that are inter-fabric and *intra*-region. East-west "localization" on the downstream layer is an important factor driving system efficiency. The upstream layer processes only north-south traffic demands – that is, those that are *inter*-region. Both layers can have a variable number of sub-switches, depending on region-specific needs.

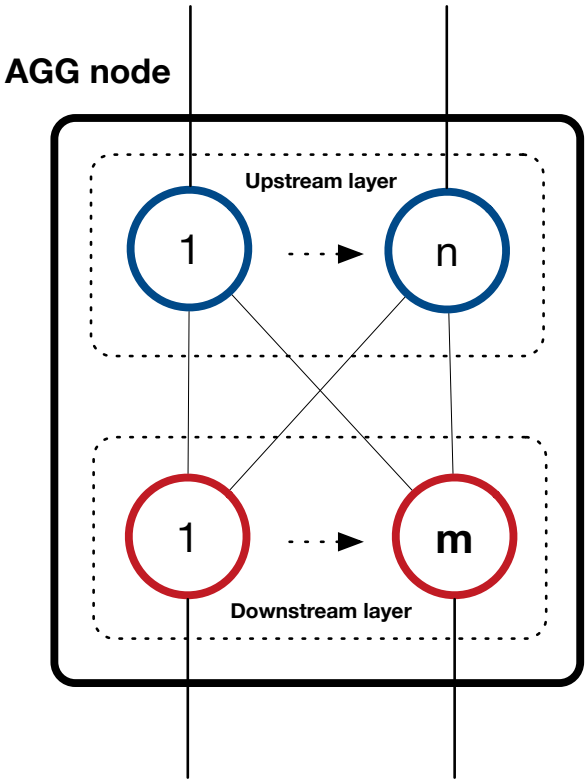

Figure 2: A two-layer cross-connect topology is an essential component of this scalable solution.

This is a simplified illustration of what turns out to be a very flexible, extensible geometry. Different iterations of the system can have different geometries, based on situational needs.

A single unit of the building block (Wedge 100S switch) in a Fabric Aggregator system can assume two different roles.

- **Downstream subswitch (DS)**: Responsible to interconnect all the buildings in the region
- **Upstream subswitch (US):** Responsible to switch traffic out of the region to another region or to the end customer.

Different regions may require different setups in terms of the downstream and upstream layer sizing.

Here are some sample configurations using Wedge 100S as a building block -- a 32x100G device, with a 3.2T forwarding capacity.

- **Conservative setup:** 8US + 16DS - Provides 3:1 oversubscription between downstream and upstream layers:
  - Total aggregated capacity per node facing DC Fabrics: **38.4T**
  - Total capacity per node facing EB/CBB layers: **12.8T**
- **Optimistic setup:** 8US + 24DS - Region can accept 9:1 oversubscription between downstream and upstream layers:
  - Total aggregated capacity per facing DC Fabrics: **57.6T**
  - Total capacity per rack facing EB/CBB layers: **6.4T**

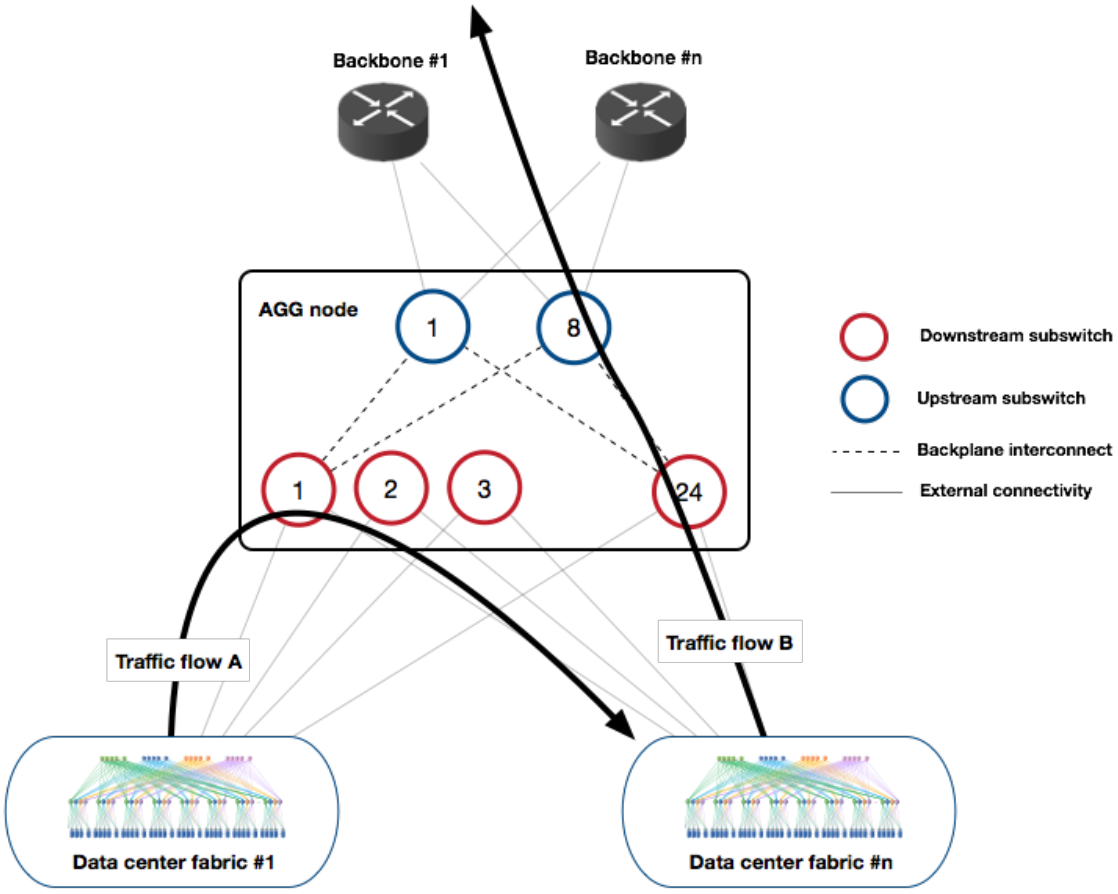

Figure 3: Traffic flow examples.

Examples of traffic flows:

- **Traffic flow 1:** Intra-region / Inter-DC traffic from Fabric to Fabric
    - In the example in Fig. 4, DS #1 takes the flow from Fabric 1. Since DS #1 has local connectivity to all DC in region, DS #1 directly switches the traffic back to Fabric 2. Every DS has direct connectivity to all DC Fabrics in region, thus DC Fabric to DC Fabric traffic switching is always kept local in the downstream layer.
- **Traffic flow 2:** Inter-region
    - In the example in Fig. 4, DS #24 takes a flow towards a different region/POP. In this case, DS #24 has a collection of US devices to forward the traffic. It selected US #8, which in turn delivers the flow to backbone layer.

This system can be operated with different levels of granularity -- from an individual Wedge 100S switch, to a full node level.

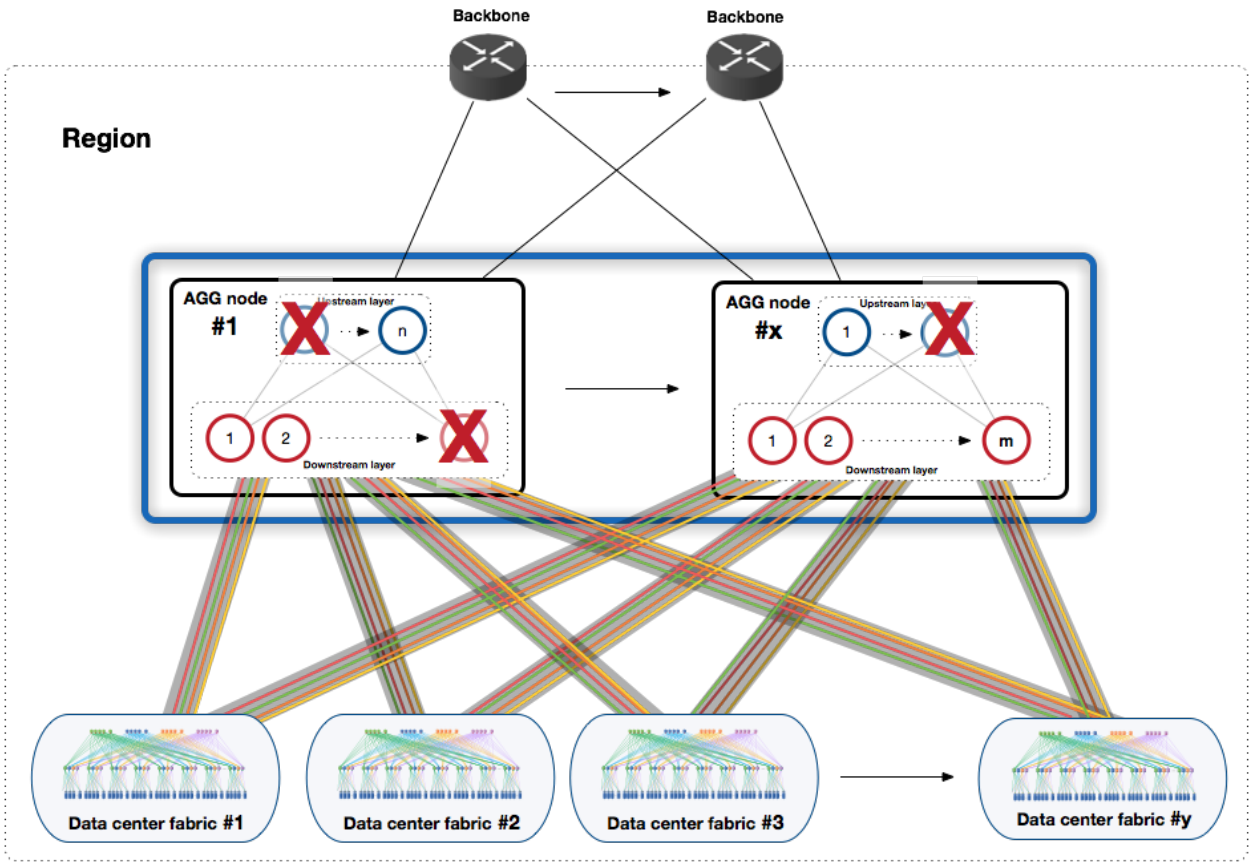

Figure 4: The system may be operated at the "subswitch" level, such that an individual subswitch failure degrades only a small percentage of total capacity. Or it may be operated at the "node" level, taking all Upstream and Downstream subswitches in a given node out of service.

7.1. Leveraging FBOSS

In the same way we were able to build on the established hardware we had already designed and deployed, we were able to leverage existing FBOSS software toward this solution. The existing FBOSS code was used almost entirely unchanged.

The only additions made to the code were around supporting the volume of traffic that could be handled by these devices. This included enabling Algorithmic Longest Prefix Matching (ALPM) to support the much greater route scale, as well as the addition of features to support Quality of Service controls so that we were able to better shape high-bandwidth links. Finally, we added support for Port Channels to ease configuration and interoperability of these devices.

Given these additions, it was not necessary to make the FBOSS agent aware of the position of its particular card in the larger device topology. As all cards perform the

exact same operations from a hardware perspective, we were able to isolate any differences to BGP policy alone. This made testing and deploying to such a topology much easier, as the cards truly are completely interchangeable, with no interdependencies between them during device operation.

8. Optimizing Physical Connections

This disaggregated architecture implies an additional cabling overhead between the downstream and the upstream layers. Minimizing the cabling complexity is one of the keys to drive the overall effectiveness of the solution.

This isn't a trivial matter. There are different cabling types to consider, and then there is the complex problem of facilitating the high-density 100G connections themselves. Factors of cost, cable management, operational ease, and power consumption all come into play.

Creating a backplane connection scheme for a device of this switching capacity was one of the challenges that drove up the price (and weight) on commercially offered switching products in this rarefied niche. We avoided the backplane problem and – to borrow a football term – instead ran a "lateral": We designed a "**sideplane**."

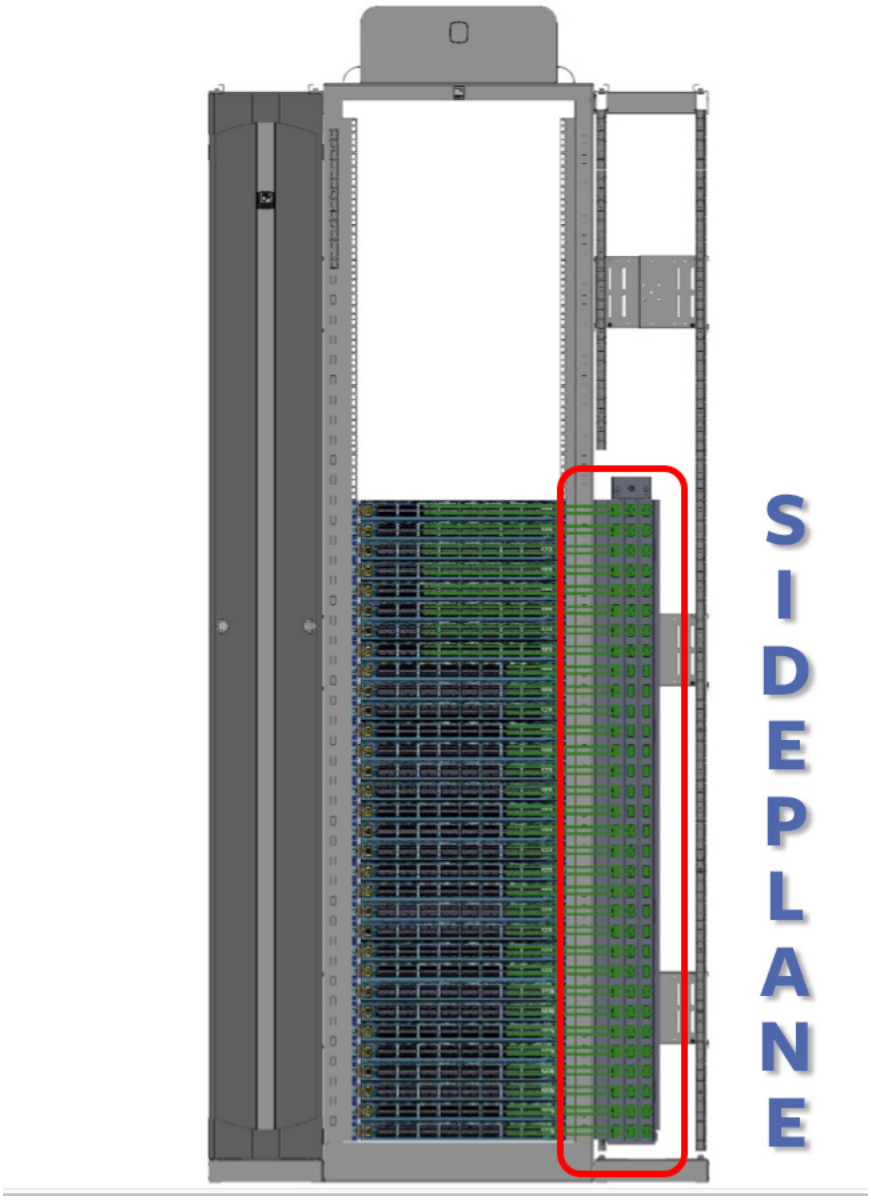

Figure 5: The sideplane system is highly compact and supports different cabling options.

The sideplane design offers multiple advantages. It appropriates a portion of the space of the Vertical Cable Manager (VCM) for intra cable management, taking only half of the VCM's 10" width, or less. It supports two types of cabling solutions: DAC, and pigtail AOC. Connections for both the Upstream Unit (UU—handling outbound traffic) and Downstream Unit (DU—handling inbound traffic) are all completed inside the sideplane, having no impact on the RU spaceof the rack. Further, the adapter bracket can be versioned for different types of racks.

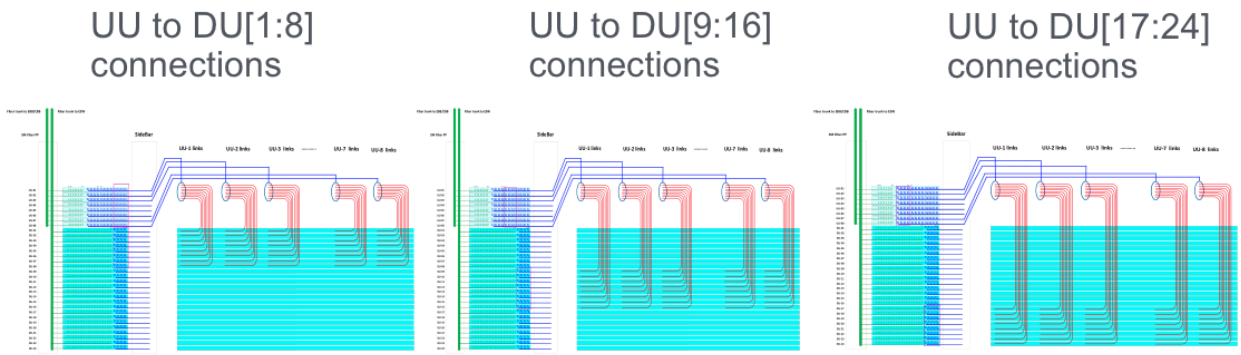

Figure 6: The sideplane approach offers a variety of connection schemes.

To create a distributed system from our Wedge 100S building blocks, we needed a way to interconnect them that was both operationally friendly and flexible enough to change as our demands change. Our "sideplane" solution is a cabling assembly to replace what the backplane in a large chassis provides. With this approach, we can change the cable assembly as our needs evolve. Some of the options that we explored are:

- Options that support Multi-rack deployments
    - Optical backplane:
        - CWDM4 + Shufflebox
        - PSM4 + Shufflebox
- Options that are confined to a single rack
    - Optical backplane:
        - Pig-tail AOC + Sideplane
    - Cable backplane:
        - DAC + Sideplane

Specifications for all of the above backplane ("sideplane") options are described below.

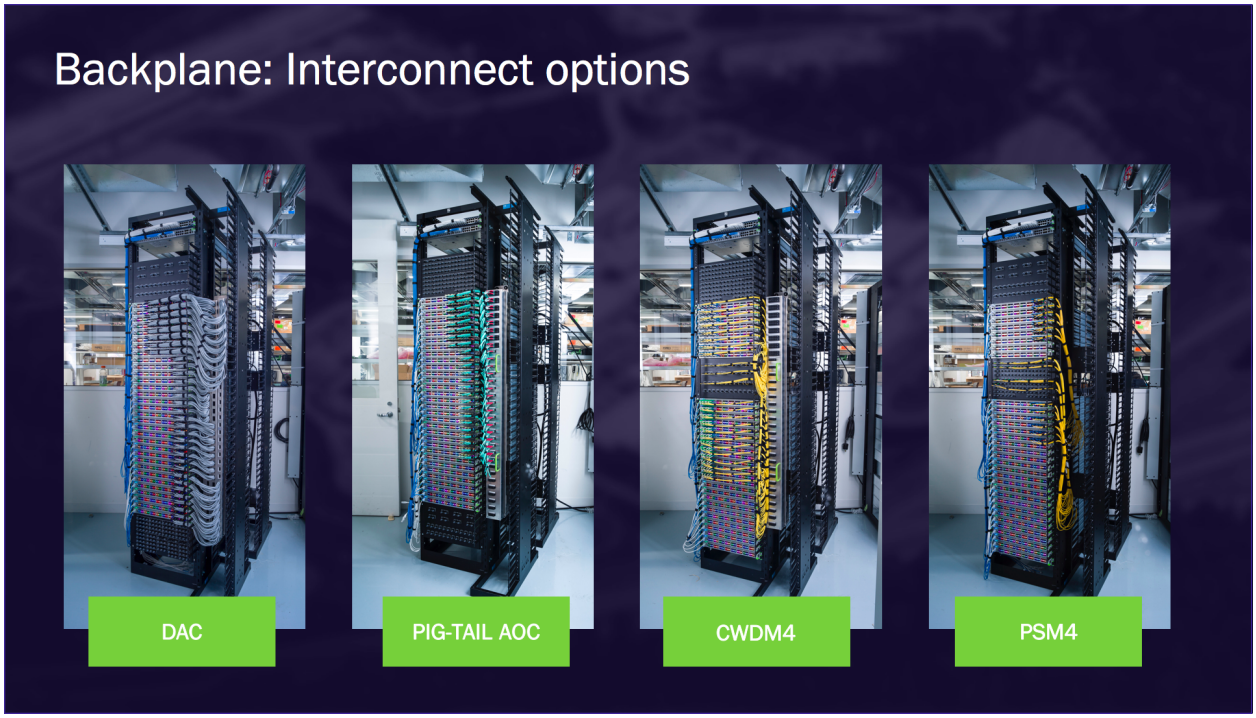

Figure 7: Different sideplane connections in place in a data center hall.

Three kinds of cabling are suitable for making the sideplane connections. Each offers advantages and disadvantages.

High-speed copper **Direct Access Cabling (DAC) often** offers advantages in price and power management. However, it often requires more overhead in terms of serviceability and operation, supports less density, and is less reliable.

An intermediate option is **Active Optical Cabling (AOC)**. Though more expensive than DAC cabling, it often is easier to work with and supports greater transmission distances.

CWDM4 optical cabling often offers the highest ease-of-use factors, if repurposing and reconnecting the cabling is likely for the installation. However, it often is also the most expensive choice.

The choice of cabling drives the design of the sideplane in a given rack configuration. In more detail, the factors to consider are:

DAC Solution

- Generally, very low cost
- Zero transceiver power, 0W power for internal connection
- Passive connection, higher MTBF
- Requires redesign for multi-rack (that is, control extended over multiple racks)

- Requires replacing a channel at a time to build; same process for maintenance

Pigtail AOC Solution

- 2w transceiver power, about 768W for internal connection BOL
- Easy installation and replacement using pigtail AOC
- OM3 multiple mode fiber, minimal fiber cleaning work
- Supports multi-rack installation, <100M distance
- Generally, least expensive option within the optical backplane category

PSM4 Solution

- Generally, lowest cost single-mode solution
- Highest MTBF of all connectivity options featuring high-reliability silicon-photonic based transceivers
- Supports multi-rack installation.
- 1.1kW internal connection
- Easy operation and maintenance minimizing number of optical matings that need cleaned or inspected from 1472 to 144 saving significant install time.
- Color-coded assembly allows easy visual installation, minimizing install time and reducing install errors.
- Backplane topology fixed in shuffle panel. No complex fiber routing.
- 500m distance, about 3.1W EOL transceiver power.

CWDM4 Solution

- Easy operation and maintenance. Unlike DAC or AOC, supports multi-rack installation.
- 1300W power for internal connections
- Easy operation and maintenance minimizing number of optical matings that need cleaned or inspected from 1472 to 776 saving significant install time.
- Backplane topology fixed in shuffle panel. No complex fiber routing.
- 500m distance with OCP CWDM4

9. DAC Sideplane Assembly

The dimensions for a DAC Sideplane Assembly are 32 RU tall, 1.3 in. wide, 4 in. deep, supporting 64 DAC connections. The sideplane is designed with three separate channels to better facilitate replacement and maintenance.

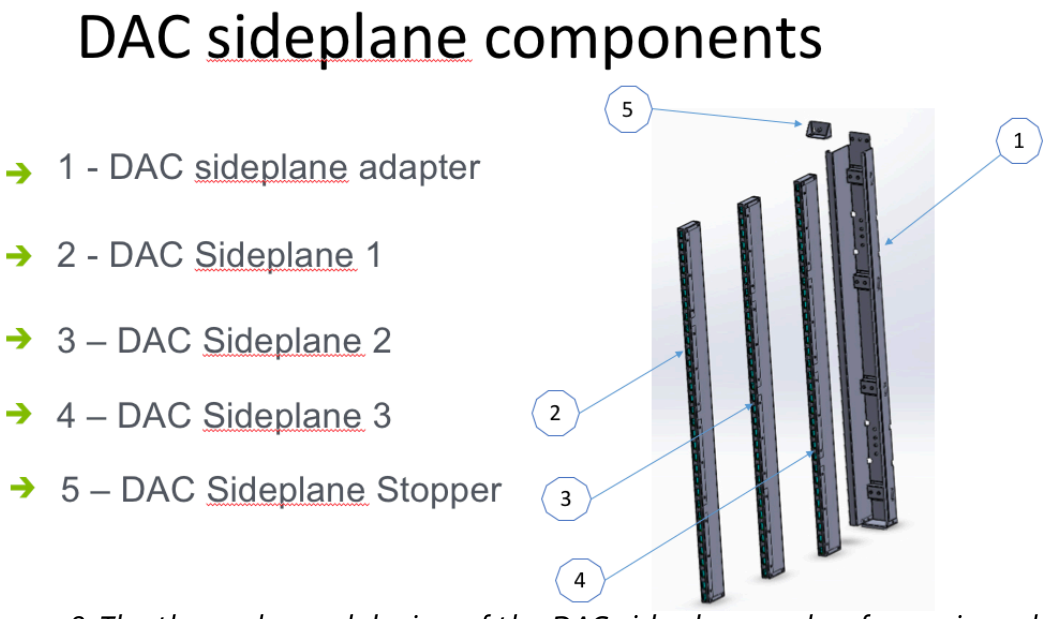

Figure 8: The three-channel design of the DAC sideplane makes for easier cable replacement and maintenance.

Against the economic advantages of the DAC solution must be balanced the limitations of the passive connection, and DAC's higher MTBF statistics.

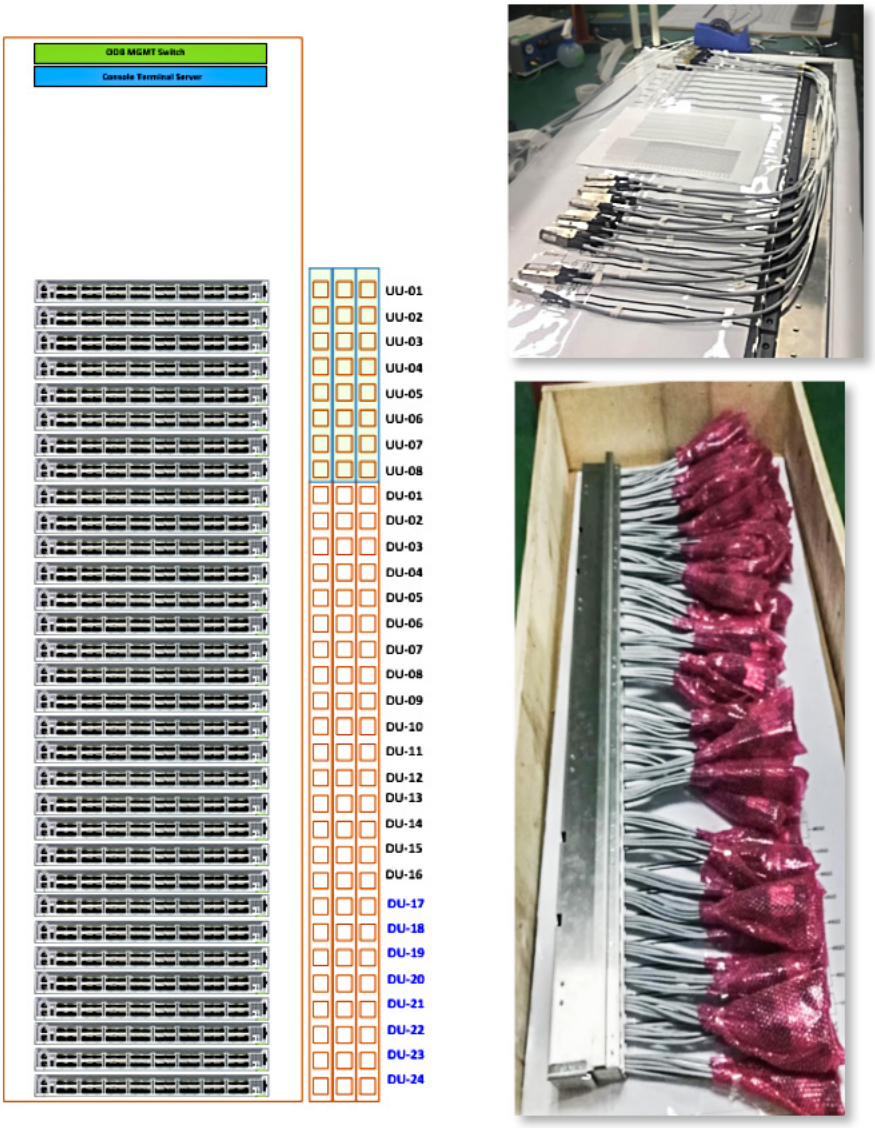

Figure 9: The three DAC sideplanes can be assembled and mounted individually.

Figure 10: An operational DAC sideplane assembly.

10. Pigtail AOC Sideplane Assembly

Pigtial AOC Sideplane Modules provide secure, high-density connections between MPO/MTP connectors. This solution is a quick, reliable, and efficient way to deploy cross-connectivity.

The Pigtail AOC Sideplane may be mounted between equipment chassis. Sideplane configurations include Multi-module, Single-module, and MPO/MPO 8~48 fibers connectors. The pigtail AOC itself affords easy cabling installation and replacement.

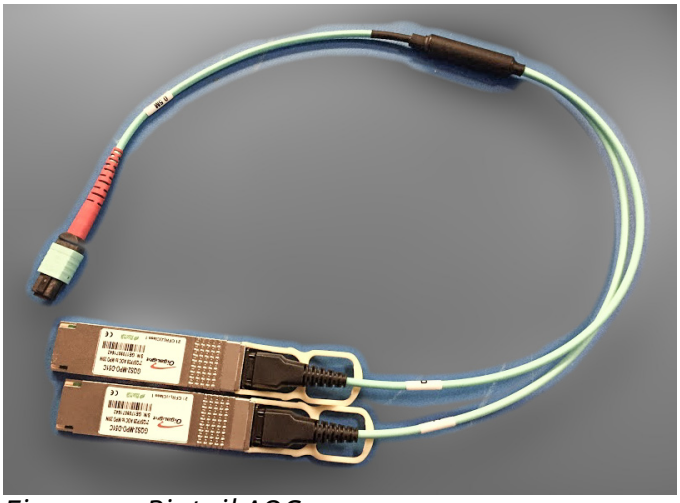

Figure 11: Pigtail AOC.

OM3 multiple-mode fiber requires minal fiber-cleaning work. Other advantages include 2W transceiver power, 768W of power for internal connections, and customized fiber polarity in the sideplane rack.

Two examples of possible Pigtail AOC Sideplane Configurations are:

192 link sideplane solution (24 DU SKU)

- 8 UU and 24 DU
- Limited to smaller size 2" x 4" and easy installation to VCM and rack
- Focus on Multi-mode fiber for low-cost pigtail AOC; can change to single mode if needed.
- Accessory for easy installation; cable length and slack is optimized for port connections, well-organized cabling.

192 MM link sideplane solution(24 DU SKU)

- 8 UU and 24 DU
- Multi-Mode fiber for low-cost pigtail AOC

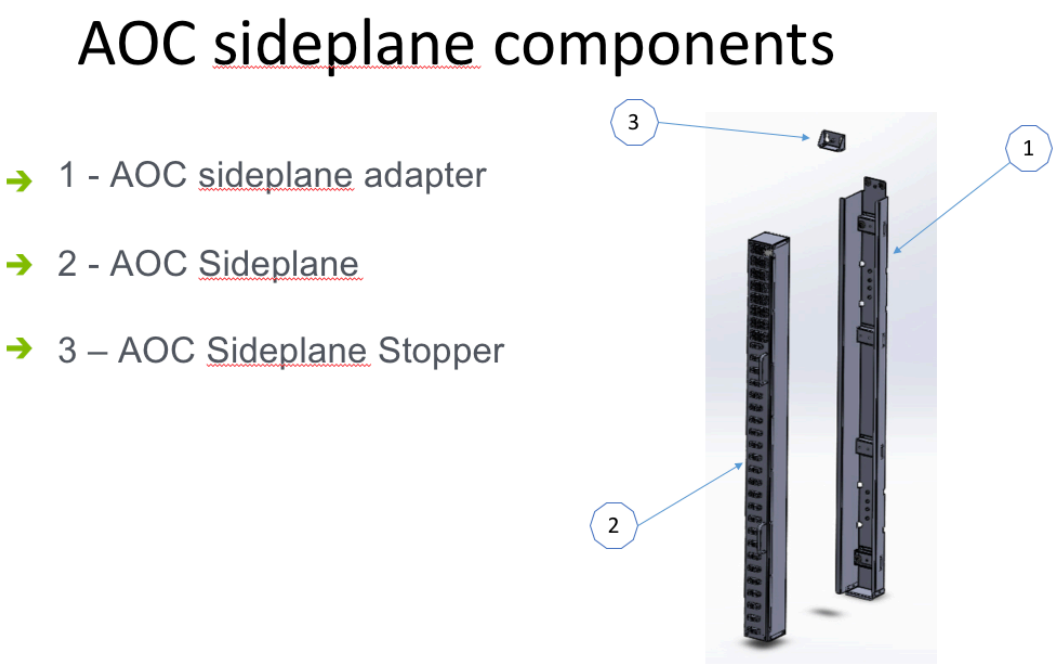

Figure 12: The simpler design of the AOC sideplane requires only an adapter, sideplane, and sideplane stopper.

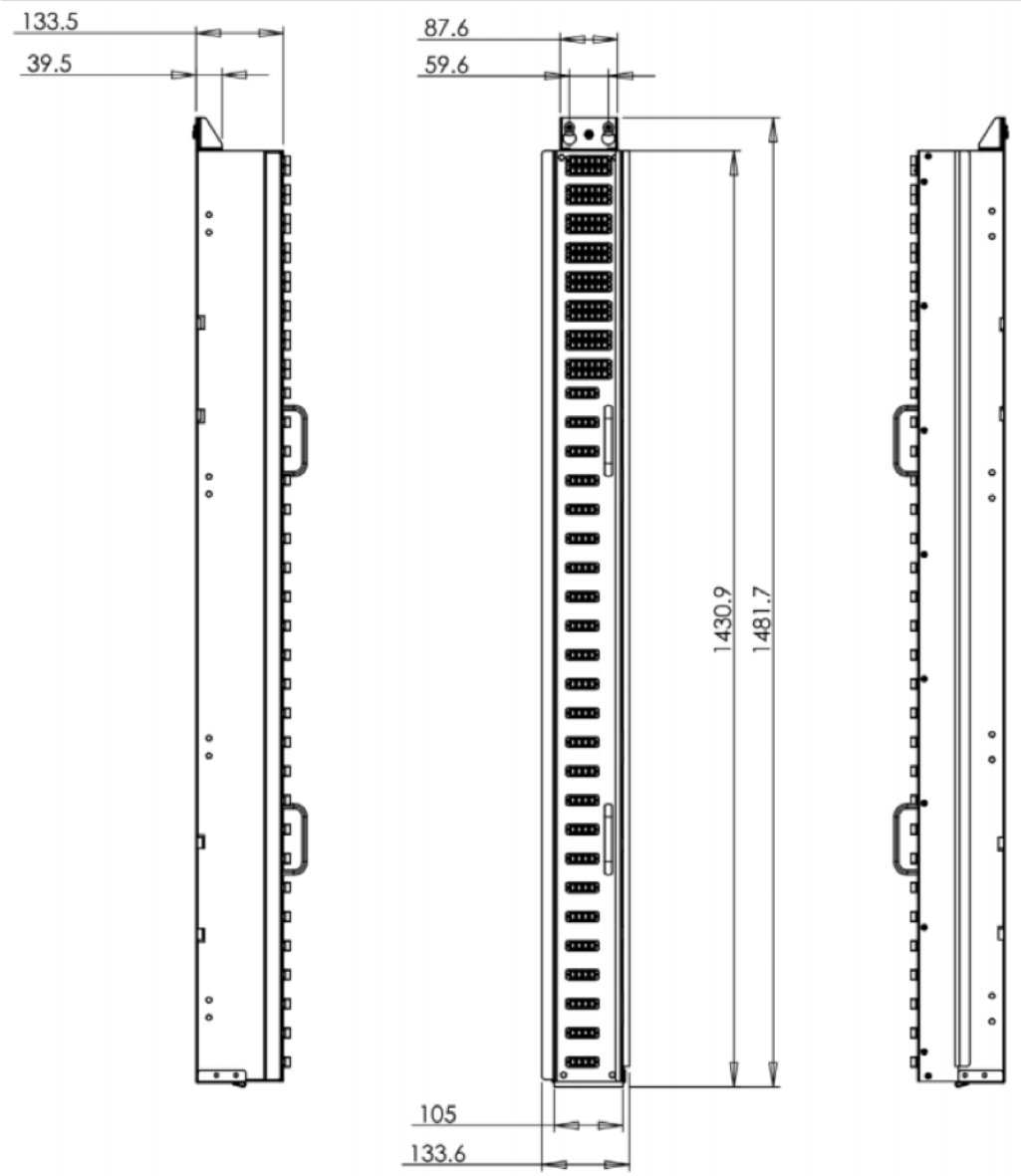

Figure 13: Dimensions for the AOC Sideplane assembly.

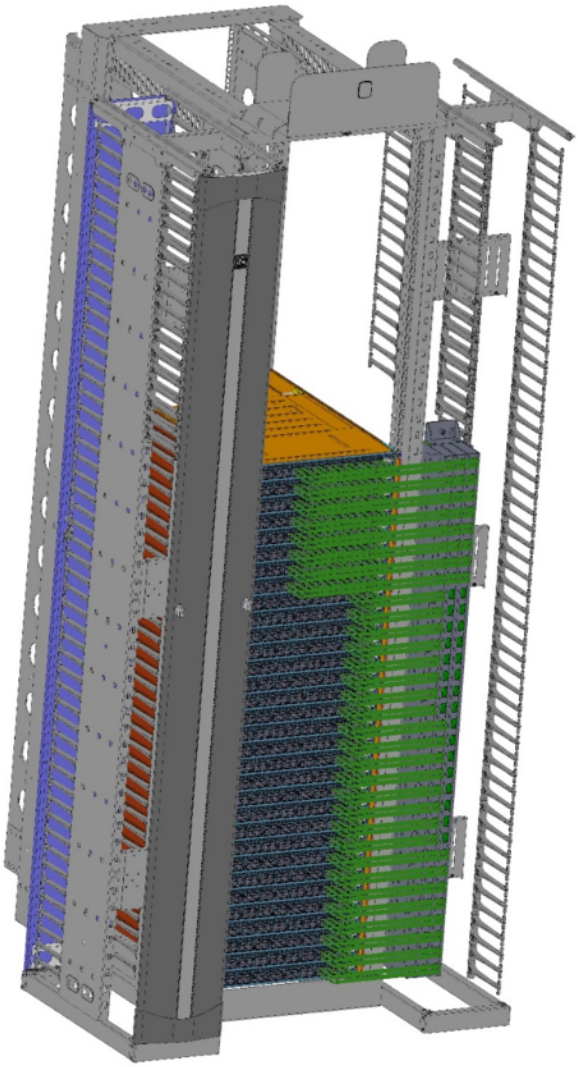

Figure 14: As noted previously, the vertical cable sideplane attaches inside the rack VCM.

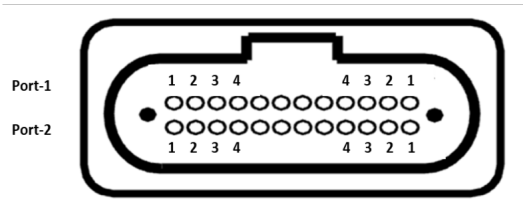

Figure 15: The MPO/MRP connector pinout for dual AOC.
Mapping the Wedge 100 switch ports to the panel ports is straightforward.

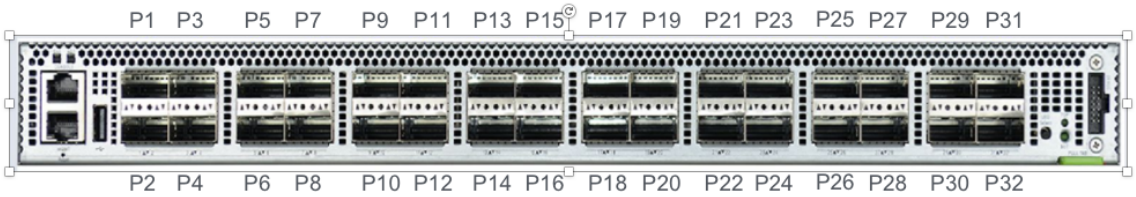

Figure 16: Port map for the Wedge 100 switch.

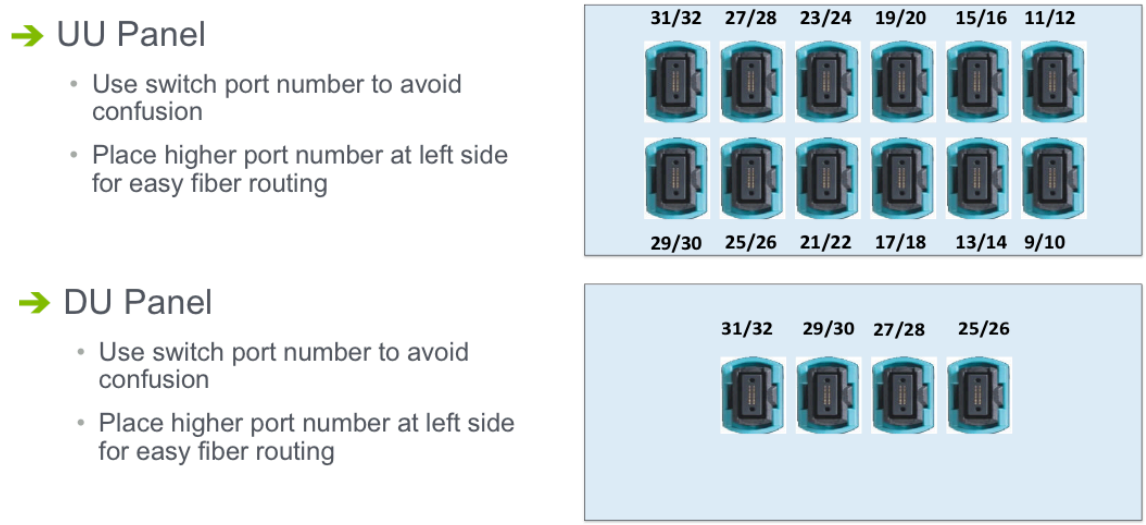

Figure 17: Corresponding relationships for sideplane panel connections..

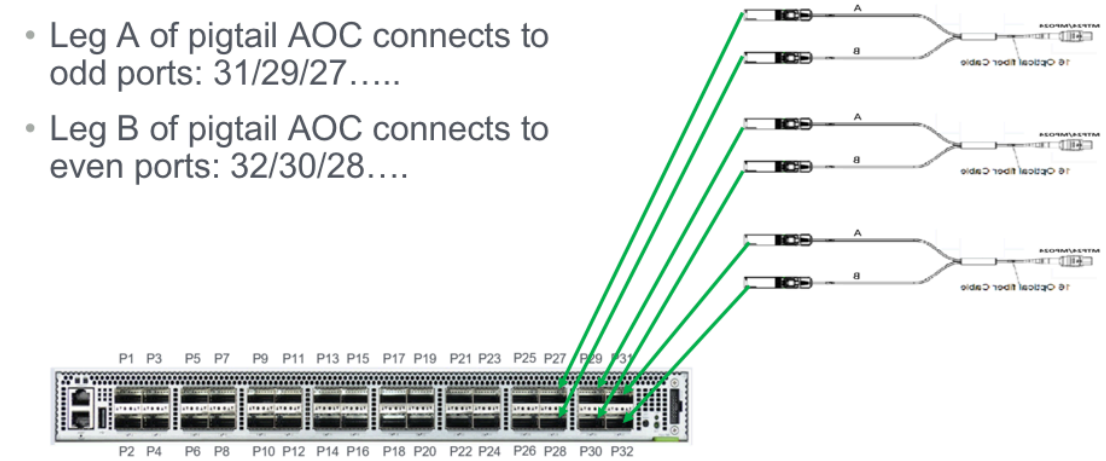

Figure 18: Switch port to sideplane panel pigtail connections.

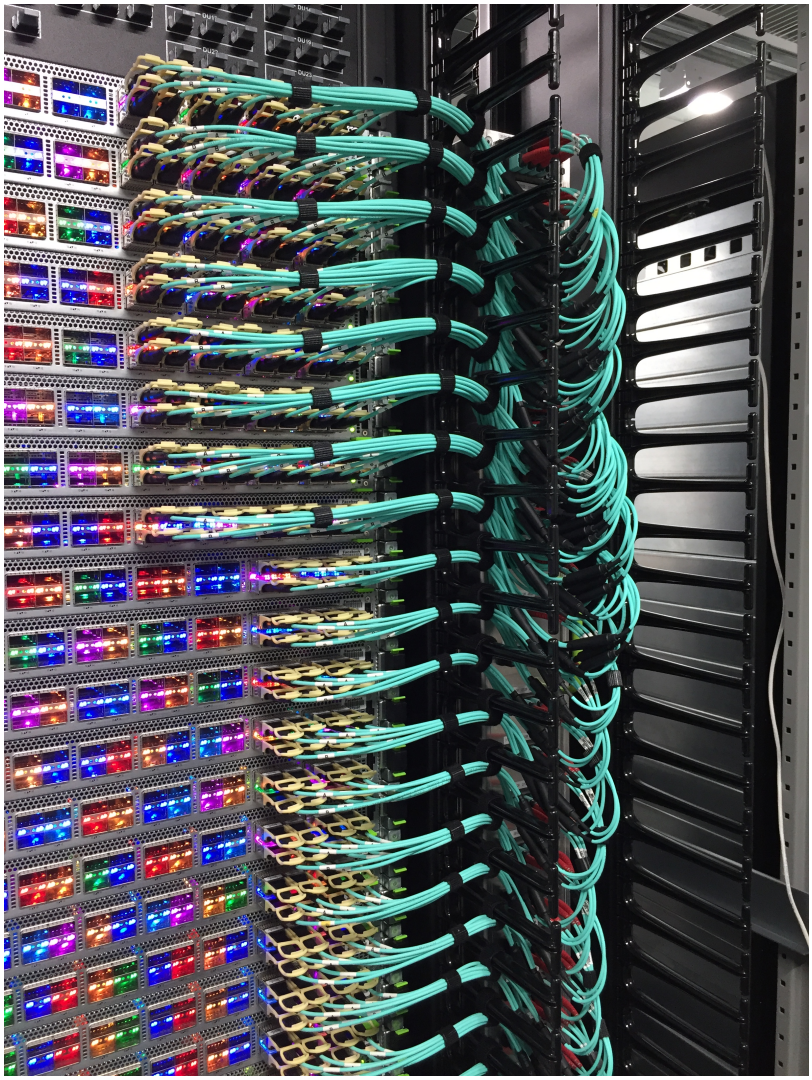

Figure 19: Sideplane panel and pigtail connections in an operational setting.

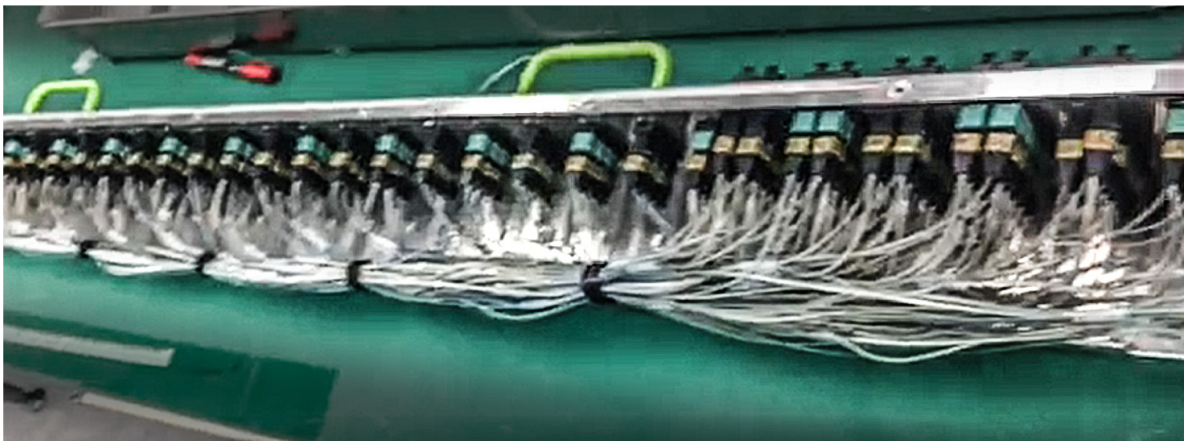

Figure 20: View of internal AOC sideplane connections.

11. PSM4 Shufflebox Assembly

Looking ahead to next-generation optics, optimizing the use of optical fiber was one of our primary goals.

We were chartered to support 100G/wave optics, tap into technological diversity (to avoid over-reliance on a single technology), and to find a scalable way to build highly dense, complex fiber topologies. Only Single Mode Fiber (SMF) is viable for the speeds required. Of the SMF choices, PSM4 represented for most applications a more attractive deployment option than CWDM4.

We elected to assemble Silicon Photonics PSM4 modules, which unify all high-speed functions in a single integrated die (improving reliability), into multi-pigtail units (Fig. 21) in order to reduce the number of fiber matings required by the system. Fewer connectors in the system meant fewer connectors to clean and manage.

The 3x100G and 2x100G fiber pigtails are color-coded for easy visual installation, which allows simple visual identification of mis-wiring. The pigtail assembly is captive and is not user serviceable. This ensures that the transceiver fiber mating remains uncontaminated, and eliminates the need for inspection and cleaning.

The panel was built as 6RU primarily to offer ease of serviceability and ample space for fiber management. The design can be scaled down to 4RU.

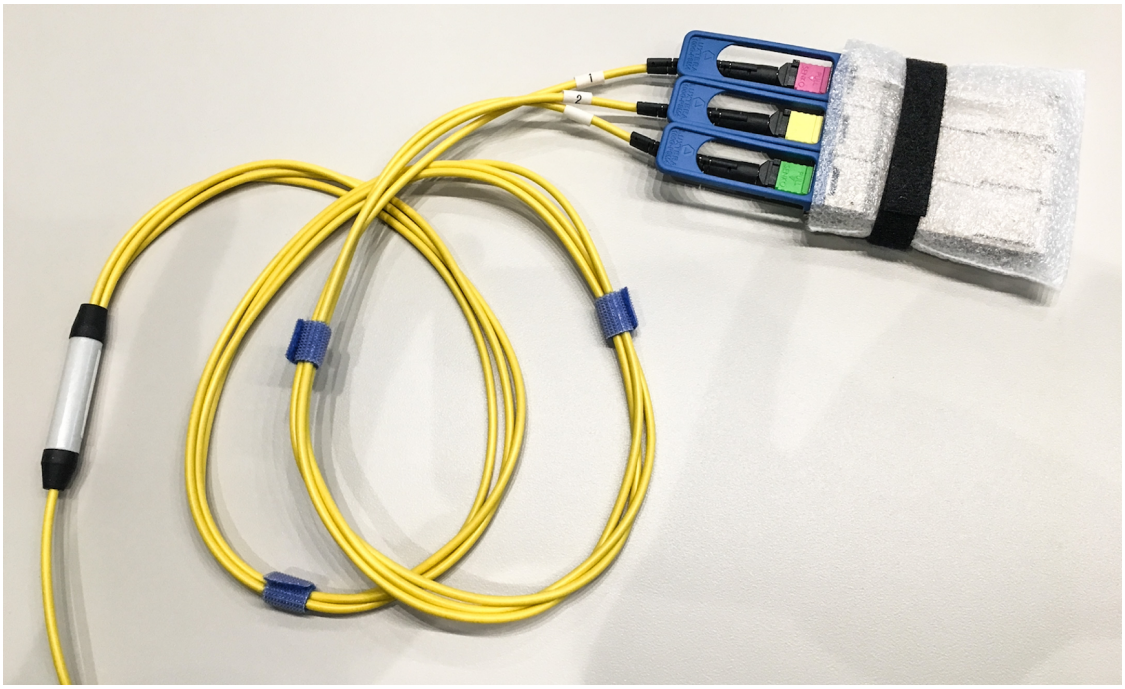

Figure 21: PSM4 ganged module.

For the PSM4 implementation we departed from the models of our DAC and AOC sideplanes, and built what we named a “Shufflebox.”

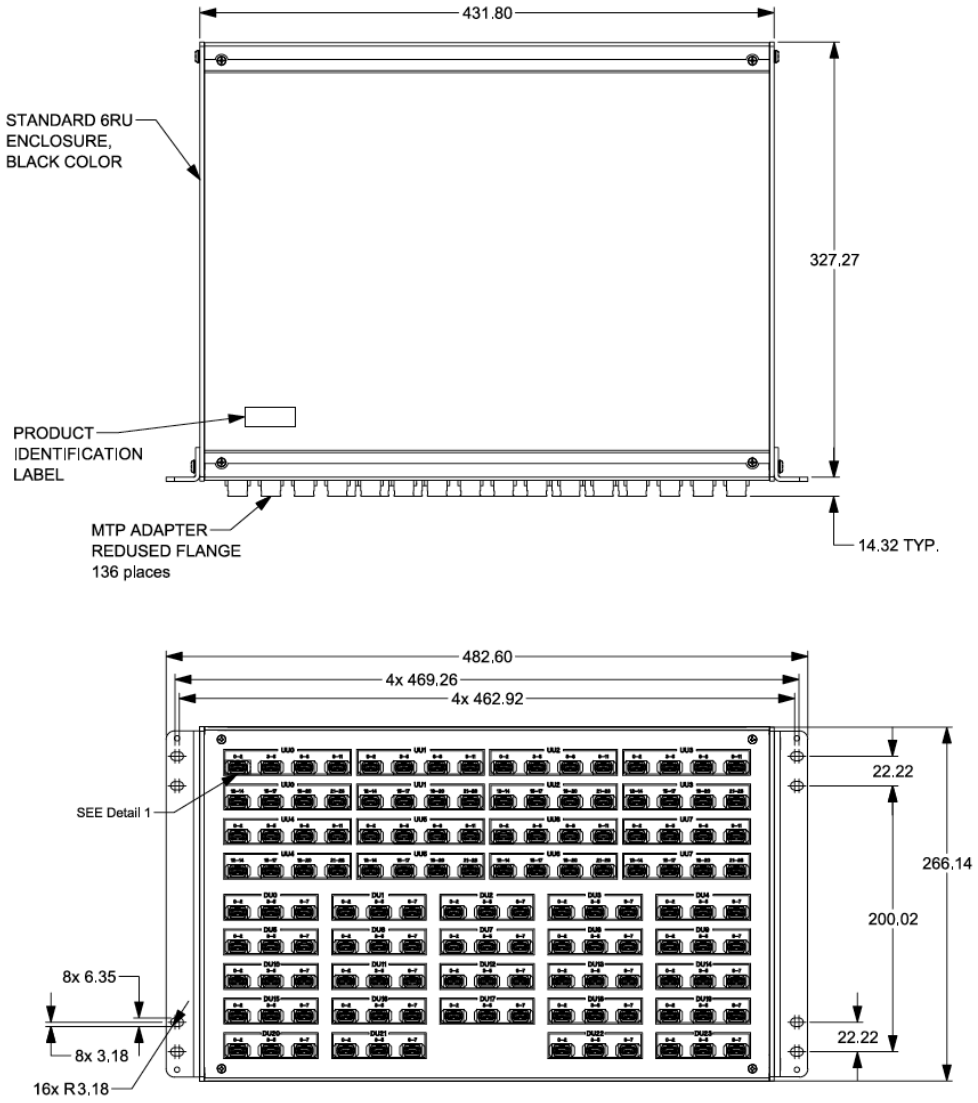

Figure 22: Shufflebox schematic.

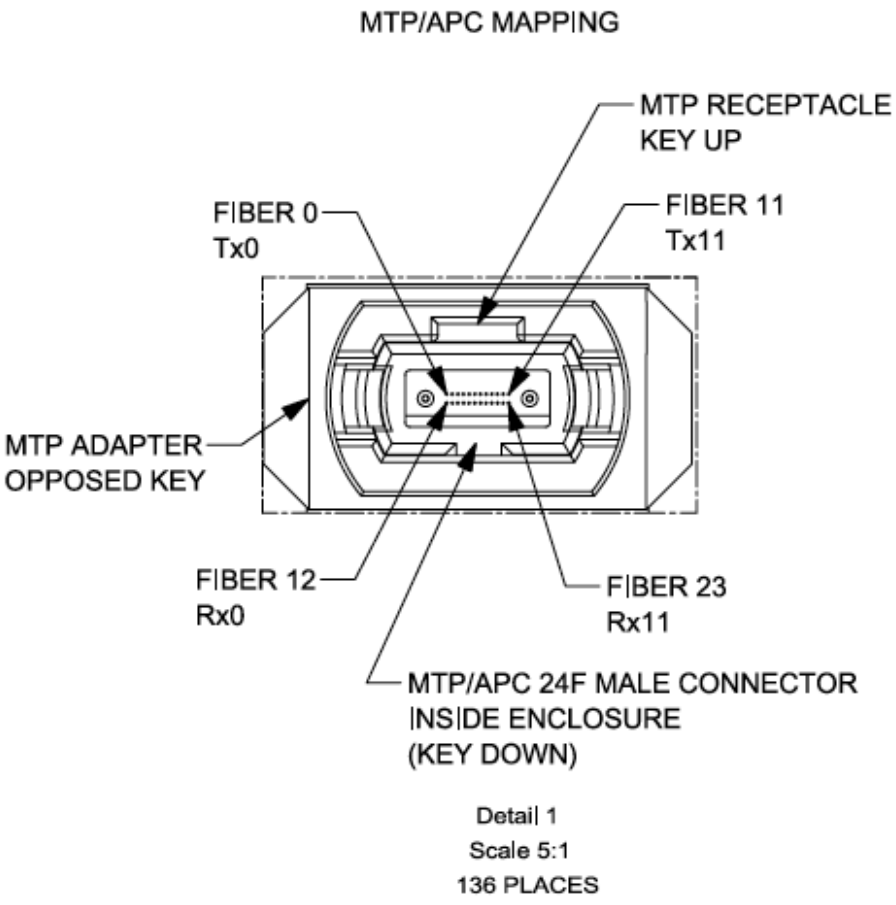

Figure 23: Shufflebox port pin-out.

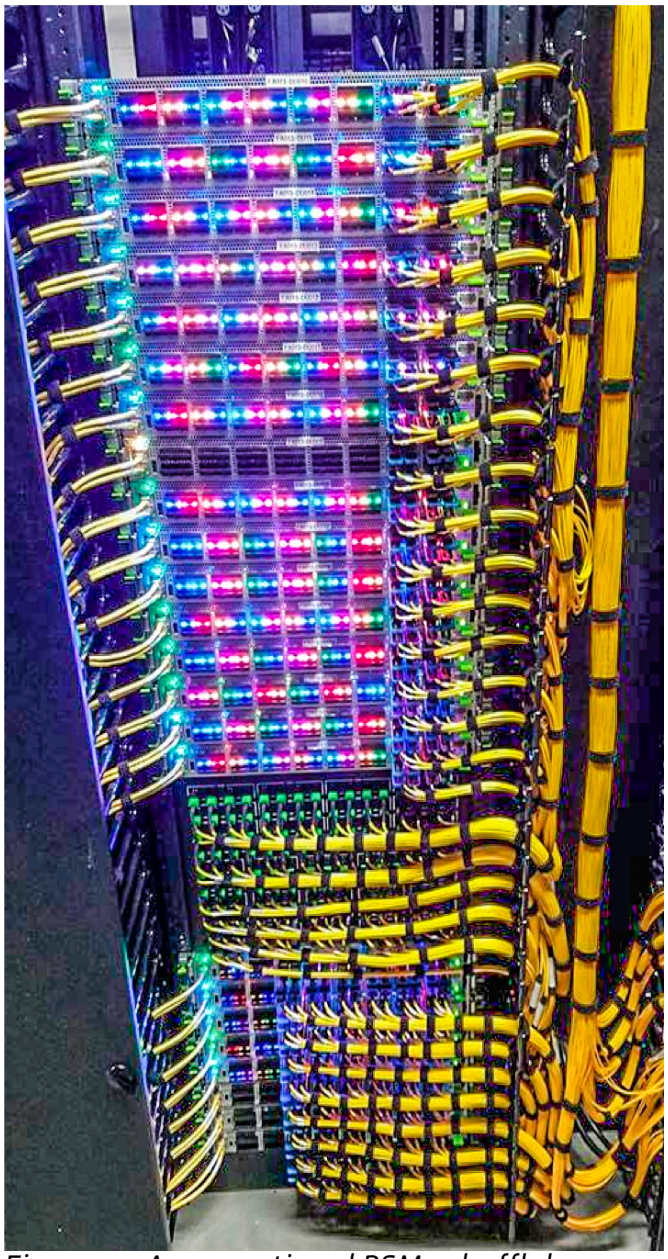

Figure 24: An operational PSM4 shufflebox assembly.

12. CWDM4 Shufflebox Assembly

Because duplex fibers represent a large reduction in the number of fibers required, a duplex option was designed. This option was intended to be forward-looking and to provide support for future low-cost dupex options, such as 100GBASE-DR1 or CWDM2. In that those optical PMDs are not yet widely available, while CWDM4 is standard in many data centers, this implementation was built around CWDM4.

A primary drawboack for CWDM4 is that while the number of fibers (and hence, the topological complexity) is lower, CWDM4 transceivers are not offered in a pigtail module. Consequently, the total number of optical matings and corresponding installation complexity is much higher.

The 6RU panel was retained for ease of serviceability, and to maintain consistency between the PSM-4 and CWDM4 footprints. This affords operational flexibility, permitting either optic to be used to allow for differences in lead times.

A possible configuration for a CWDM4 shuffle panel assembly appears in Figure 25.

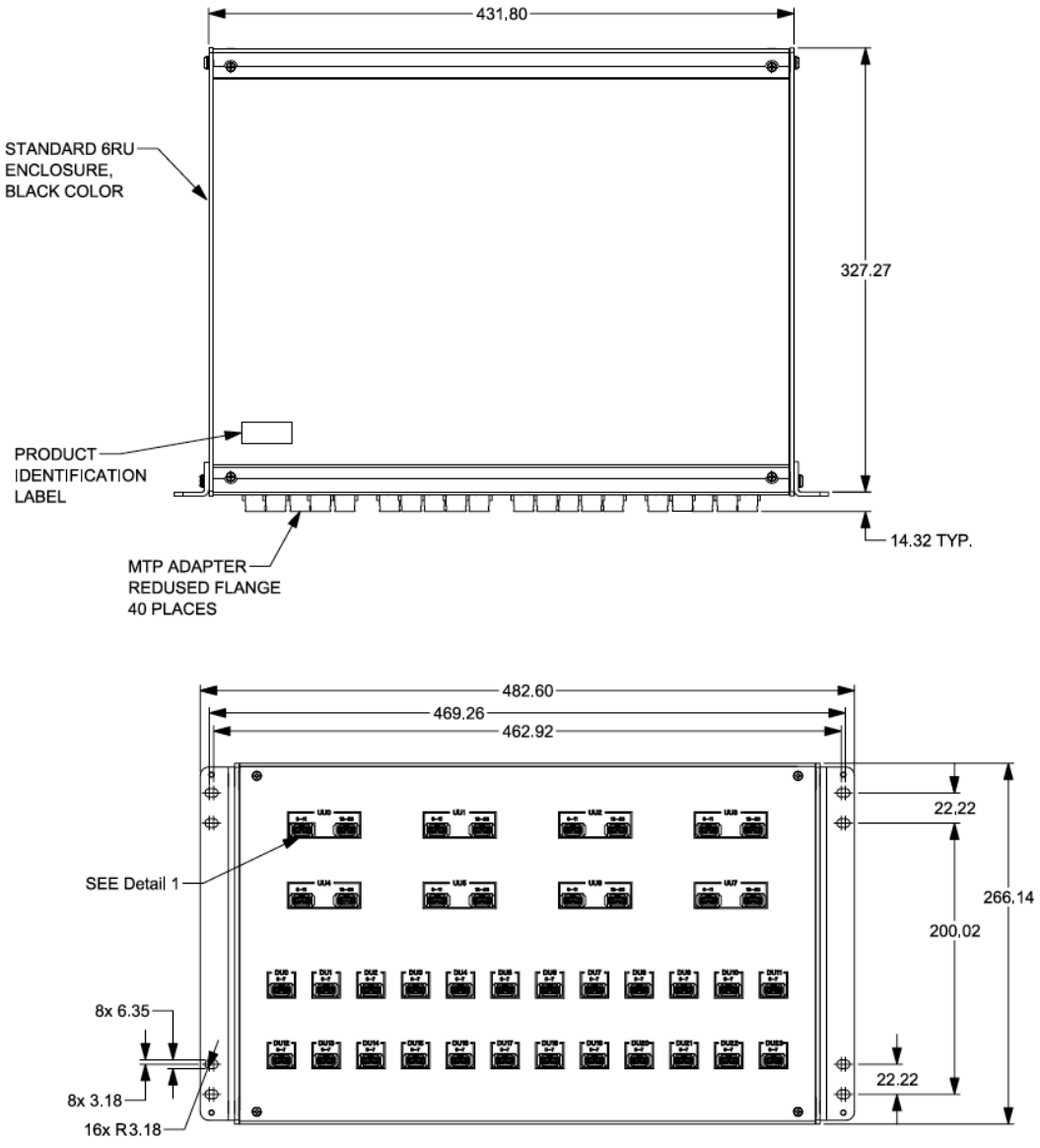

Figure 25: The 6RU shufflebox panel.

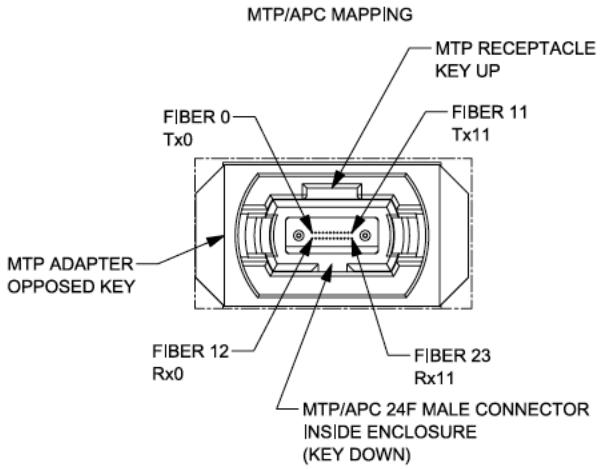

Figure 26: CWDM4 shufflebox port pin-out.

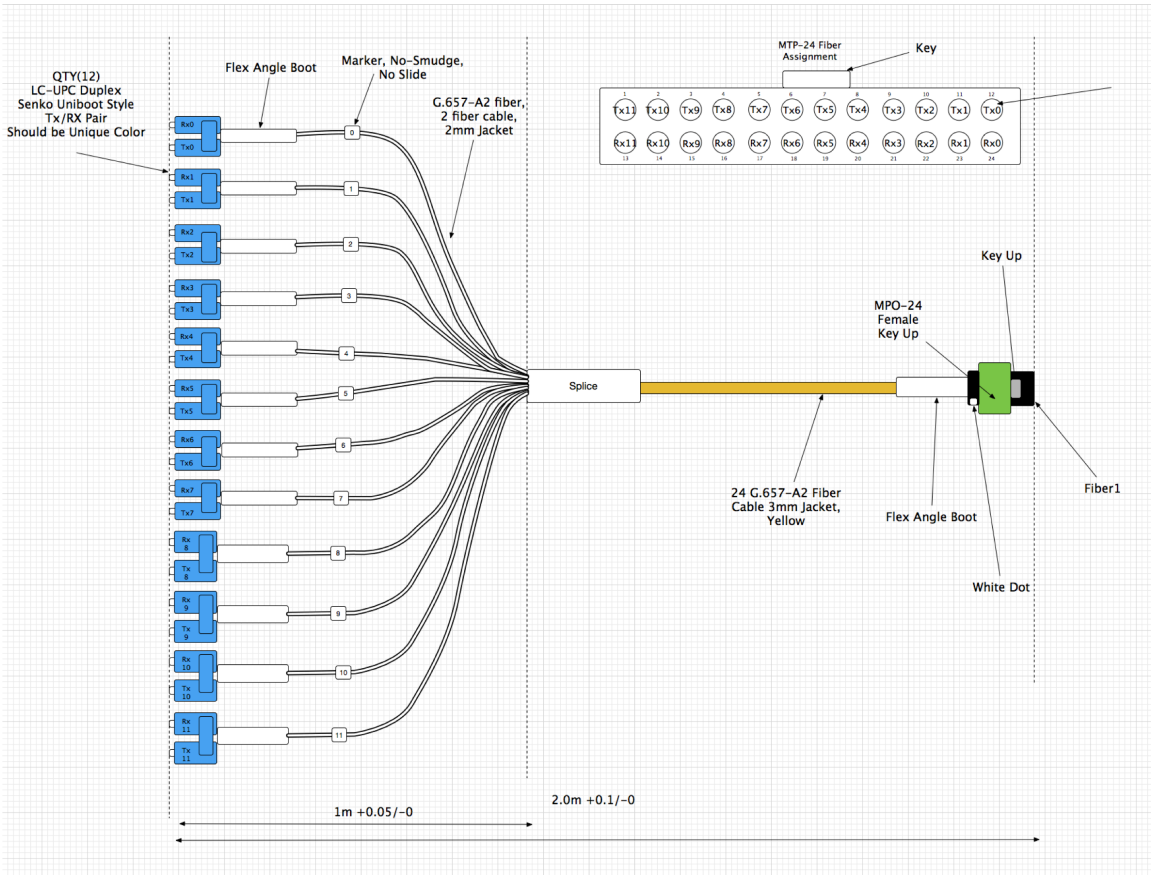

Figure 27: The UU hydra cable assembly featuring 12 LC duplex aggregating in a single MTP-24 connector.

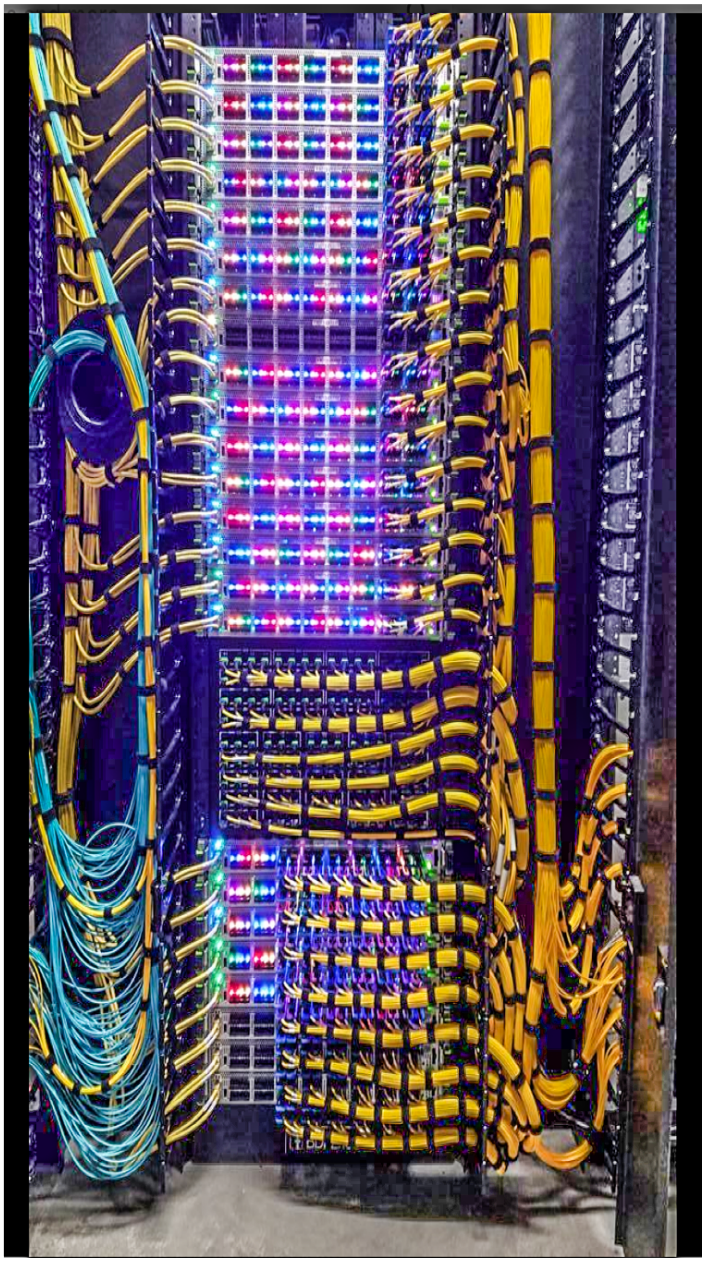

Figure 28: An operational CWDM4 shufflebox assembly.

13. Equipment Manufacturers

The ecosystem of vendors that supports this specification includes the companies listed in the table below.

Table 1: Vendors that support this specification include:

Supplier Name	Product Description
LOROM www.lorom.com	DAC Cabling. Very flexible, lighter weight due to aluminum backshell (rather than zinc).
Gigalight www.gigalight.com	AOC. Also passive fiber device vendor.
Molex www.molex.com	CWDM-4 Shufflebox 1065070032 – 6U BOX, DPLX, 40 PORTS
Molex www.molex.com	PSM-4 Shufflebox 1065070033 – 6U BOX,PARALLEL, 136 PORTS
Molex www.molex.com	UU Cable Hydra: 12 Duplex LC (24 fibers) to MPO-24: 106284-6167
Molex www.molex.com	DU Cable Hyrda: 8 Duplex LC (16 fibers) to MPO-24: 106284-6166
Luxtera www.luxtera.com	2x100Gb/s PSM-4 Pigtail LUXAP221M2A-0200, 0-70C Temperature Range
Luxtera www.luxtera.com	3x100Gb/s PSM-4 Pigtail LUXAP221M2A-0300, , 0-70C Temperature Range

14. Summary: Benefits of A Distributed Network System with a Building Block Approach

- **REUSE** of a building block in more than one network role creates operational efficiencies.
- Smaller building blocks lead to smaller failure domains, creating greater **RESILIENCY**.
- Sizing our system to our traffic needs allows for chip reduction and **POWER SAVINGS**.
- **MOVE FAST**: Using building blocks we know and understand helps us to **ITERATE** on a solution quickly.
- Having a **FLEXIBLE, ADAPTABLE,** and **SCALABLE** system enables us to change as our needs change.