Meta Data Centers Network

Mohsen Dalil
Network Engineer

Meta

AGENDA

01

Meta’s European data centre fleet

02

DC Architecture

03

F16 DC Fabric

04

HGRID

05

Open Compute Project

06

Hardware Platform

07

FBOSS

08

GenAI Network

SWEDEN
Luleå

2013 break ground

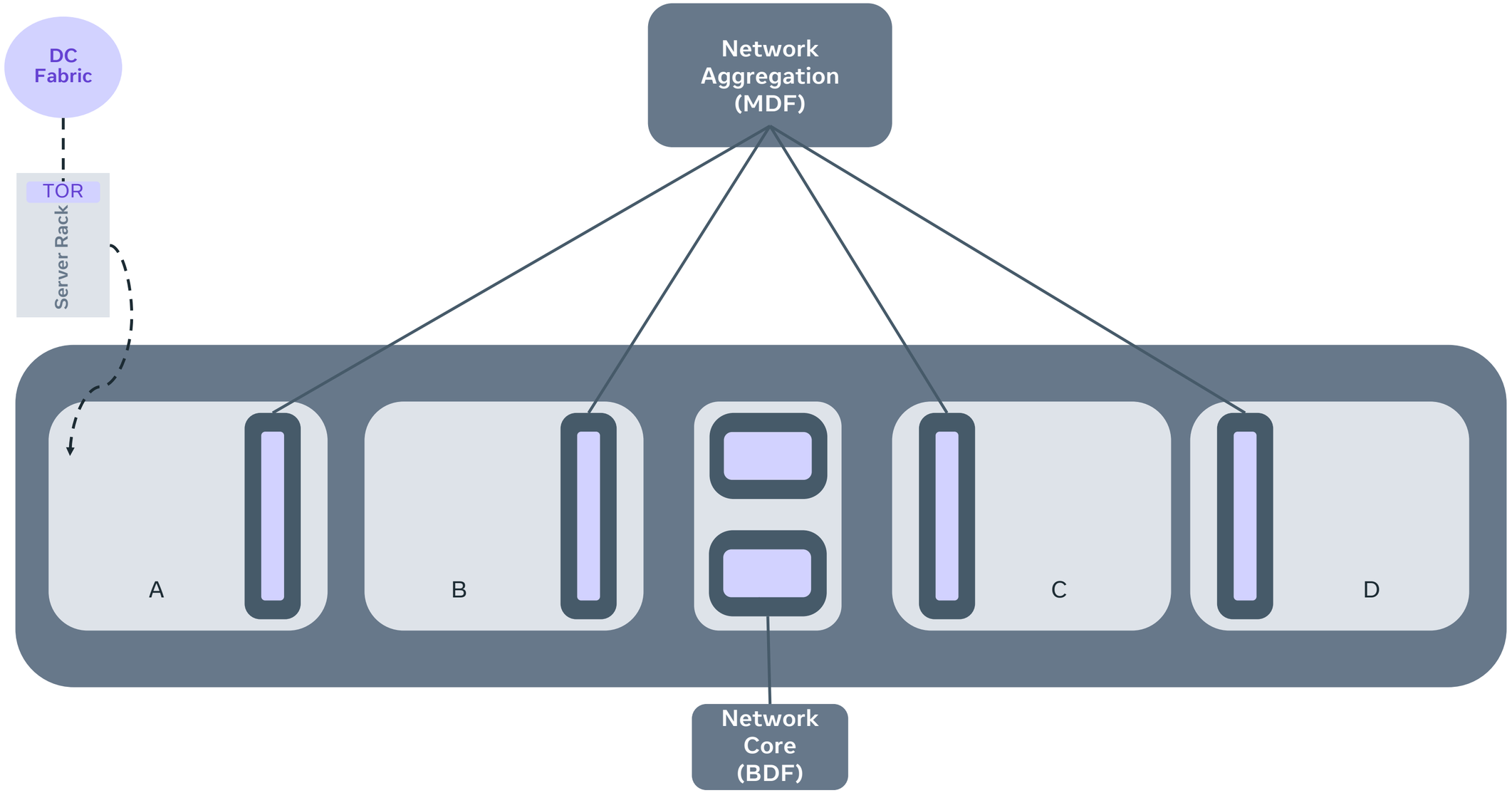
IRELAND
Clonee

2016 break ground

Denmark
Odense

2017 break ground

DC Building Architecture

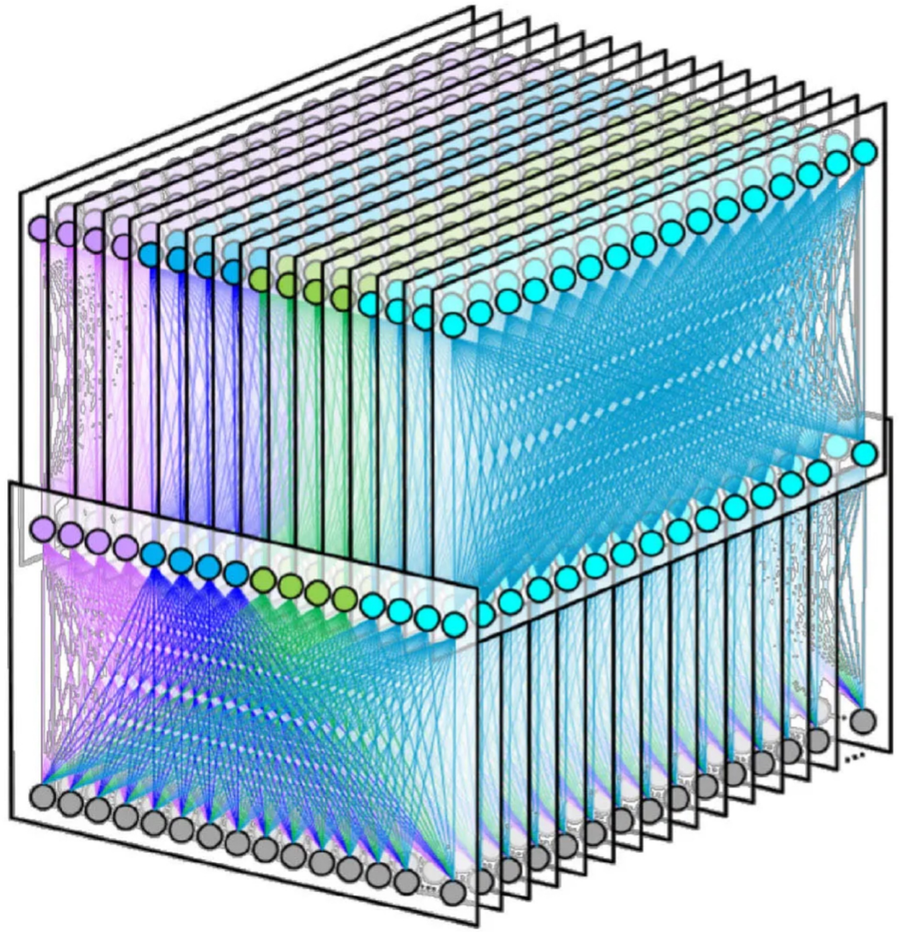


F16 (DC Fabric)
A next-generation data center fabric

Each rack connect to 16 separate planes

The plane above the rack comprise 16 fabric switches

And Spine layer



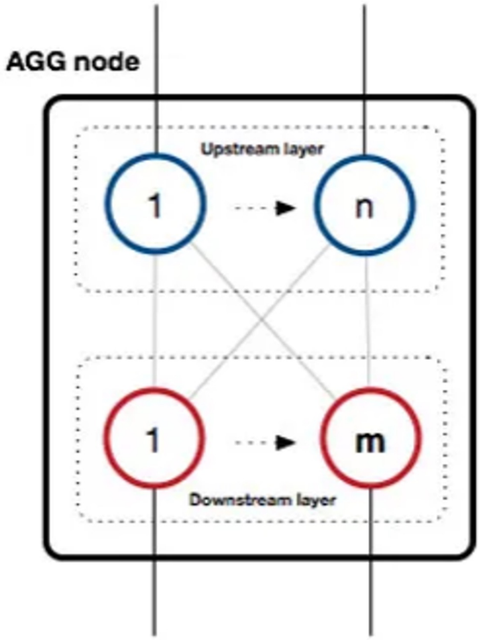
HGRID

Fabric Aggregator to handle the 6 buildings per region

All traffic that leaves or enters Meta’s data centers is handled by the HGRID layer.

HGRID is a 2-layer cross-connect architecture.

Traffic that flows between buildings is referred to as **east/west** traffic.
Traffic exiting d entering a regions is known as **north/south** traffic.

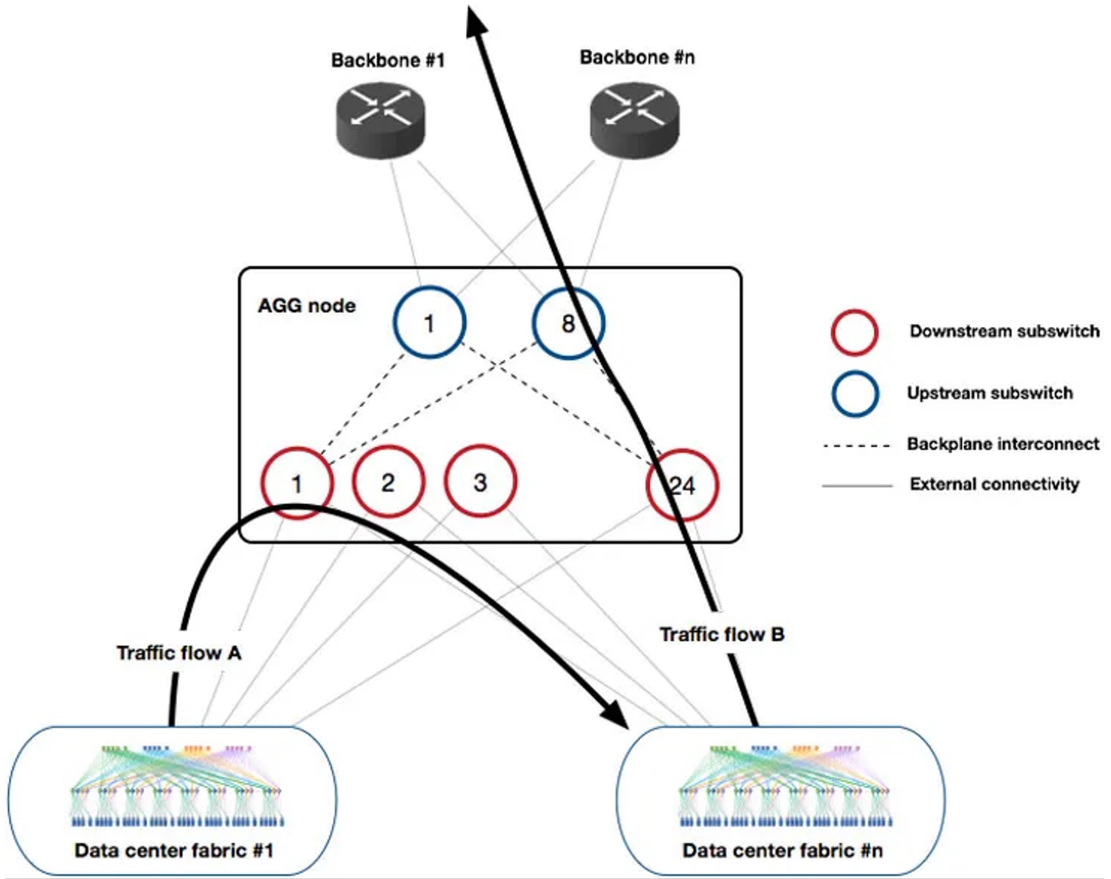


HGRID

Fabric Aggregator to handle the 6 buildings per region

The downstream layer is responsible for:
switching regional traffic (fabric to fabric inside the same region, designated as east/west).

The upstream layer is responsible for:
switching traffic to and from other regions (north/south direction).



The Open Compute Project (OCP) is a collaborative community focused on redesigning hardware technology to efficiently support the growing demands on compute infrastructure.

Wedge 400/400C

It utilizes Broadcom’s **Tomahawk 3** ASIC

Wedge 400C uses **Cisco’s Silicon One**

Both TORs offer higher front panel port density and greater performance for AI and machine learning applications

Switching capacity **12.8 Tbps**

field-replaceable CPU subsystem

They are manufactured by **Celestica** and are **open platforms**

Top row interfaces

- Uplink ports
  - Port 1-16 : 100/200/400G
- Downlink ports
  - Port 17-48 : 4x25/2x50/100/200G

Minipack 2

Meta has developed next-generation 200G fabric switches

Minipack2 is based on the Broadcom **Tomahawk 4, 25.6Tbps** switch ASIC

High-performance switches that transmit up to 25.6 Tbps and 10.6 Bpps with modular line cards.

Port Interface Module (PIM)

- PIM-16Q 16x QSFP56 200G
  - Backward compatibility: Can support a 100G QSFP28 module.
  - Forward compatibility: Two QSFP56 ports can be “combined” to support a 400G QSFP-DD module.
- More options:
  - PIM for 8x QSFP-DD 400G ports with MACSec encryption/decryption.

Arista 7388X5

Meta also developed next-generation 200G fabric switches Arista 7388X5, in partnership with Arista Networks.

7388X5 is also based on the Broadcom **Tomahawk 4, 25.6Tbps** switch ASIC

Port Interface Module (PIM)

- 16 x 200G QSFP56 / 100G QSFP
- 8 x 400G QSFP-DD
- 8 x 400G QSFP-DD with MACSec

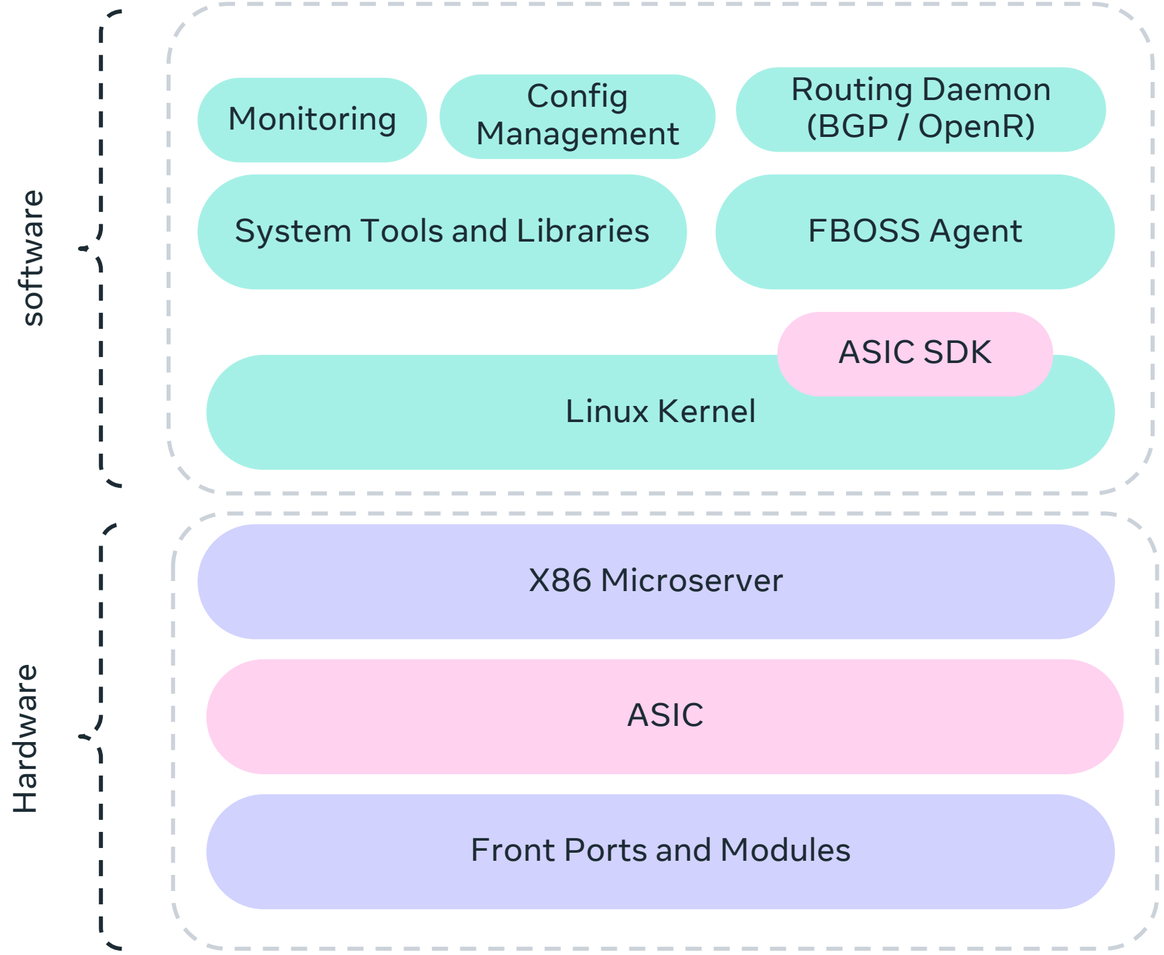
FBOSS
the unifying software that binds together our data centers

Facebook open switching system (FBOSS)

Code is leaner than standard network switch

FBOSS is not tied to a specific hardware or feature set

FBOSS is also an open source



GenAI Network

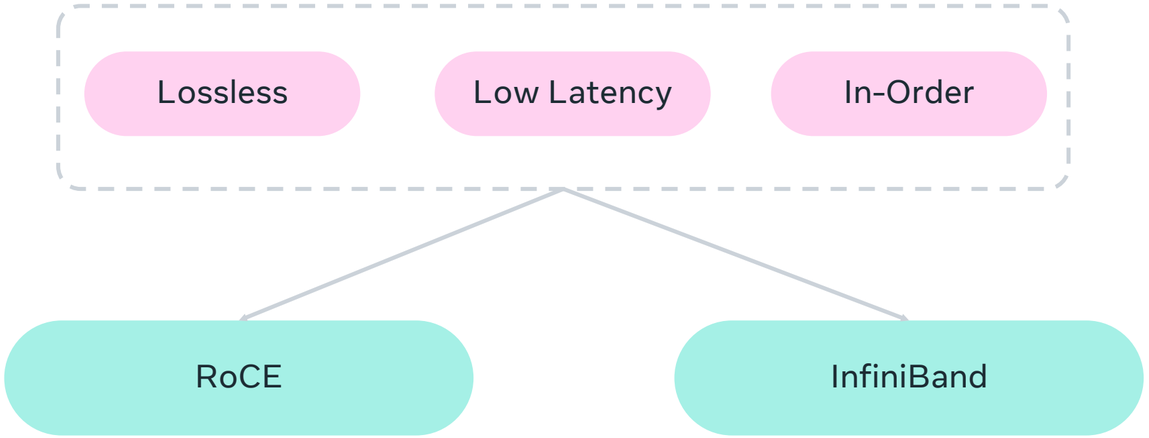
Two 24K GPU clusters

Meta built one cluster with a remote direct memory access (RDMA) over converged Ethernet (RoCE) network fabric solution based on the Arista 7800 with Wedge400 and Minipack2 OCP rack switches.

The other cluster features an NVIDIA Quantum2 InfiniBand fabric.

Both of these solutions interconnect 400 Gbps endpoints.

350,000 NVIDIA H100s GPUs by end of 2024

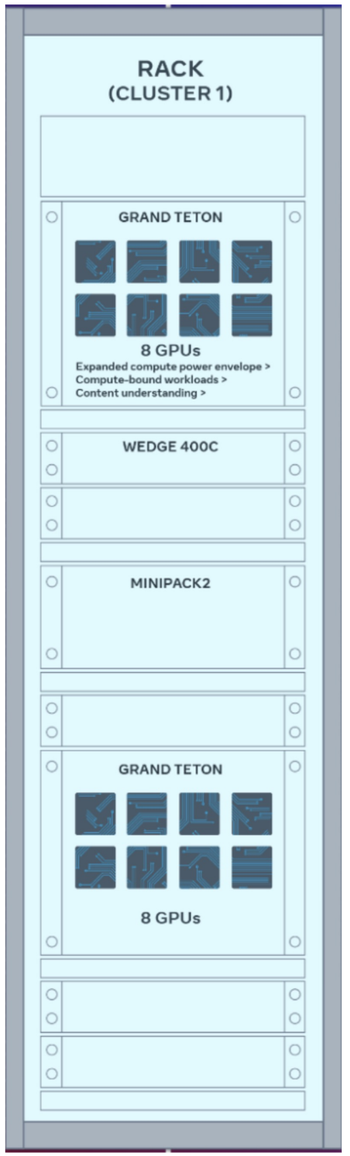


AI Racks

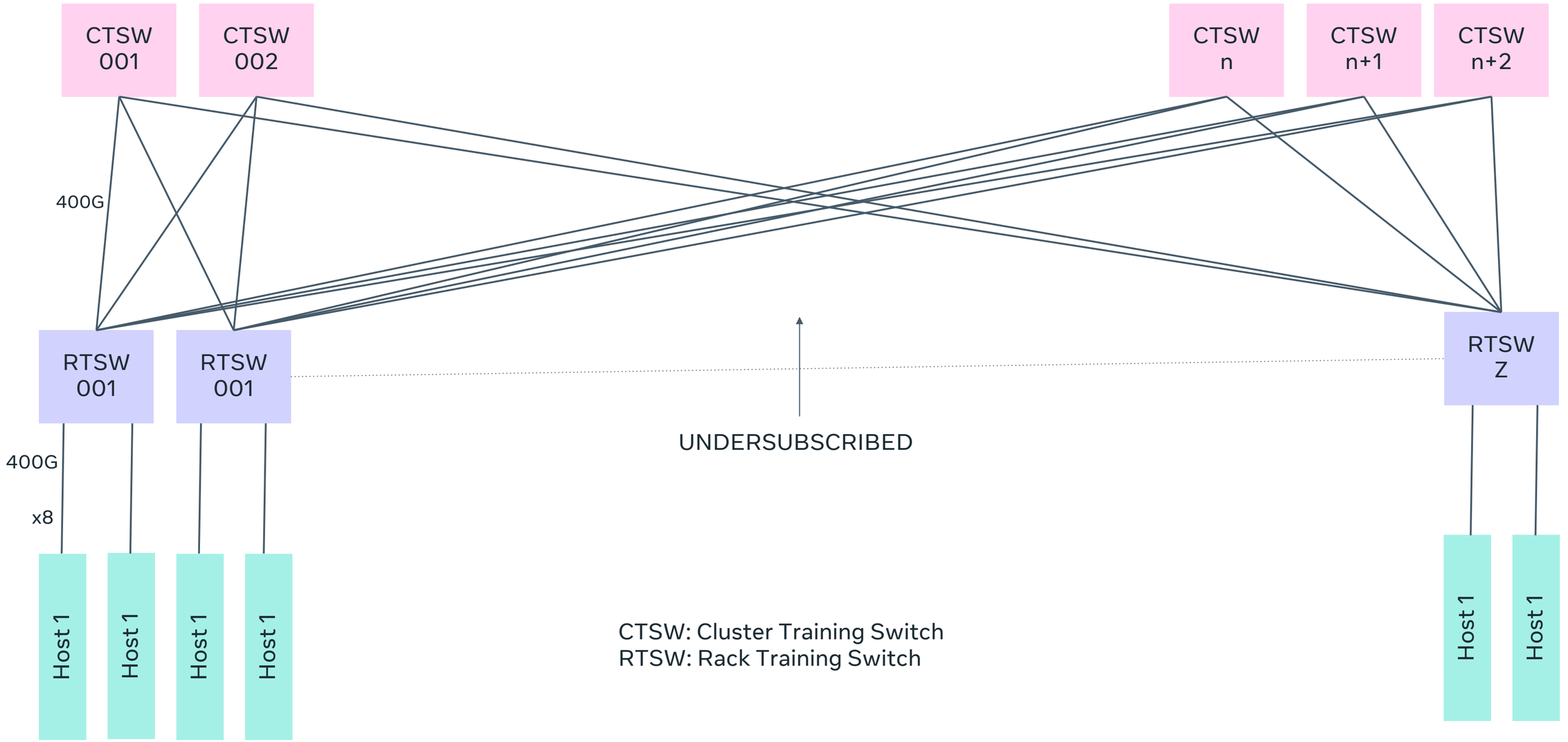
Grand Teton (in-house-designed) open GPU hardware platform and Open Rack v3 (ORV3)

Frontend = Data Ingestion

Backend = GPU - to - GPU



AI Fabric


CTSW 001
CTSW 002
CTSW n
CTSW n+1
CTSW n+2
400G
RTSW 001
RTSW 001
RTSW Z
UNDERSUBSCRIBED
400G
x8
Host 1
Host 1
Host 1
Host 1
Host 1
Host 1
CTSW: Cluster Training Switch
RTSW: Rack Training Switch