ITNOG

THE ITALIAN NETWORK OPERATORS GROUP

Dive deep on AWS networking infrastructure

Colin Whittaker (he/him)

Principal Engineer
AWS

AWS networking

Infrastructure networking	Amazon EC2 networking	Edge networking
Routers/switches	Virtual private cloud (VPC)	Amazon Route 53
Copper/optical cables	Elastic network interface	AWS Global Accelerator
Data centers	AWS Hyperplane	Amazon CloudFront
Inter-Region backbone	Elastic Fabric Adapter (EFA)	AWS Direct Connect
Internet peering/transit	Placement groups	AWS Cloud WAN

Agenda

Choose your own adventure

Option A: Hardware innovation

- How and why we design our hardware for routing, encryption and transport

Option B: Software innovation

- Distributed vs centralized control, evolving out of self contained devices

Tenets

Tenets

Secure

Tenets

Secure

Available

Tenets

Secure

Available

Scalable

Tenets

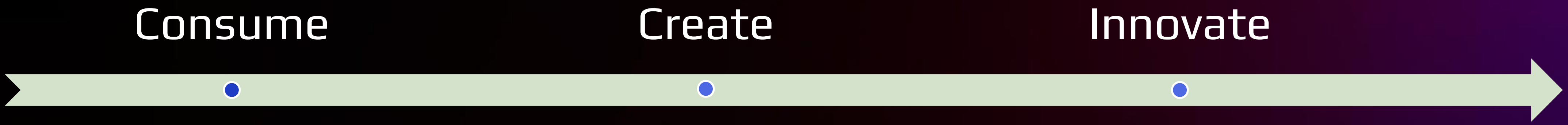
Secure

Available

Scalable

Performant

Phases of evolution



Consume

Industry hardware and software

Basic automation

Pushed beyond design intentions

Large chassis backplane/midplane

Core concepts into create

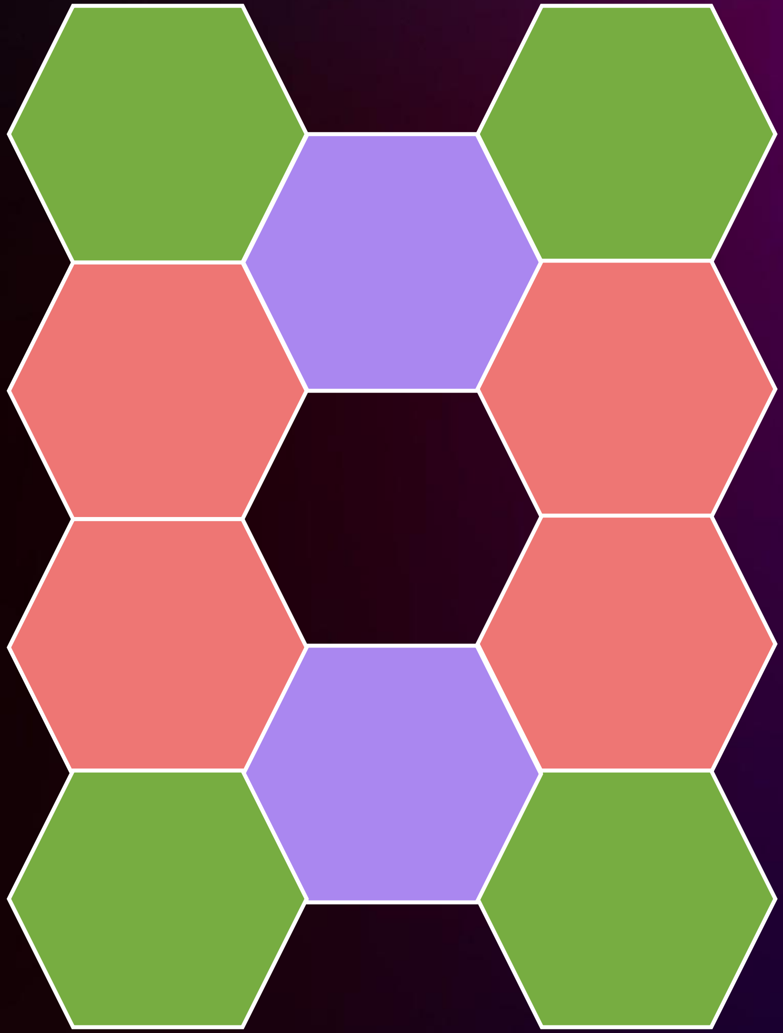
Embrace Moore's law

Own our destiny

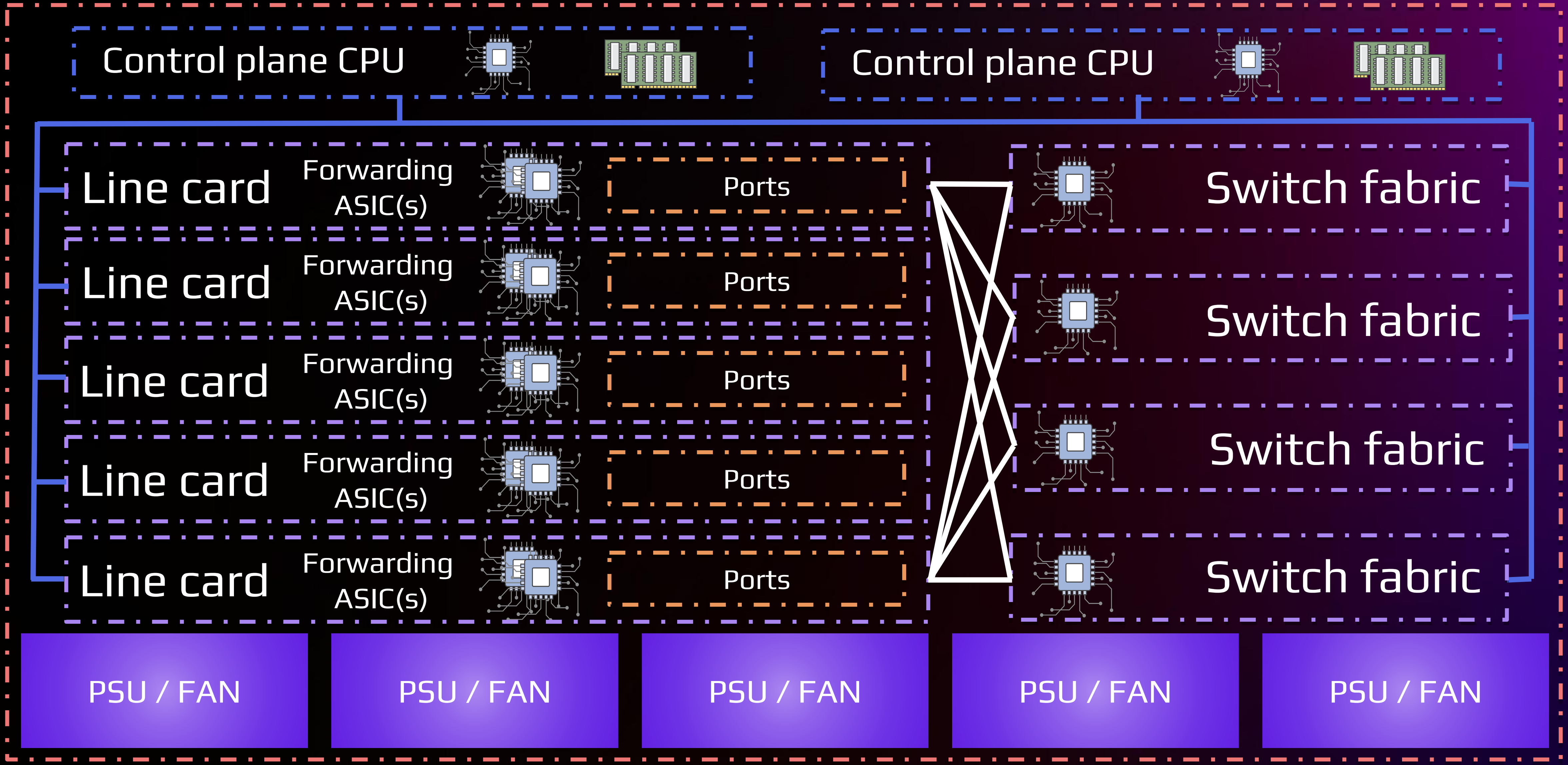
Use repeatable design patterns

Limit effect boundaries

Constantly iterate and evolve



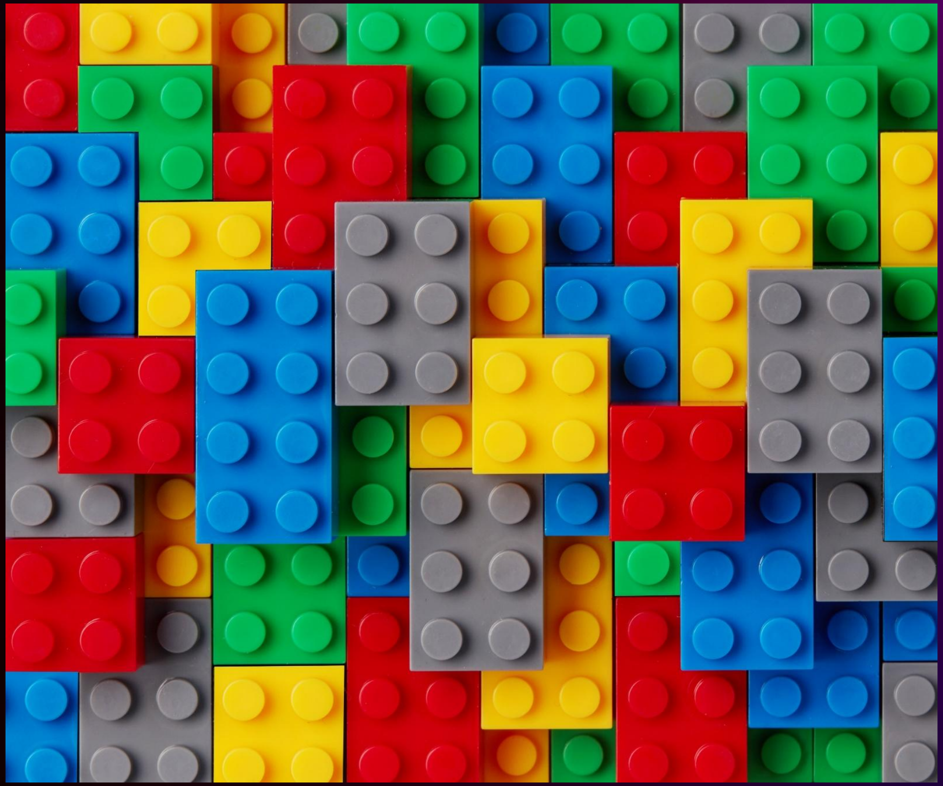
Chassis platforms



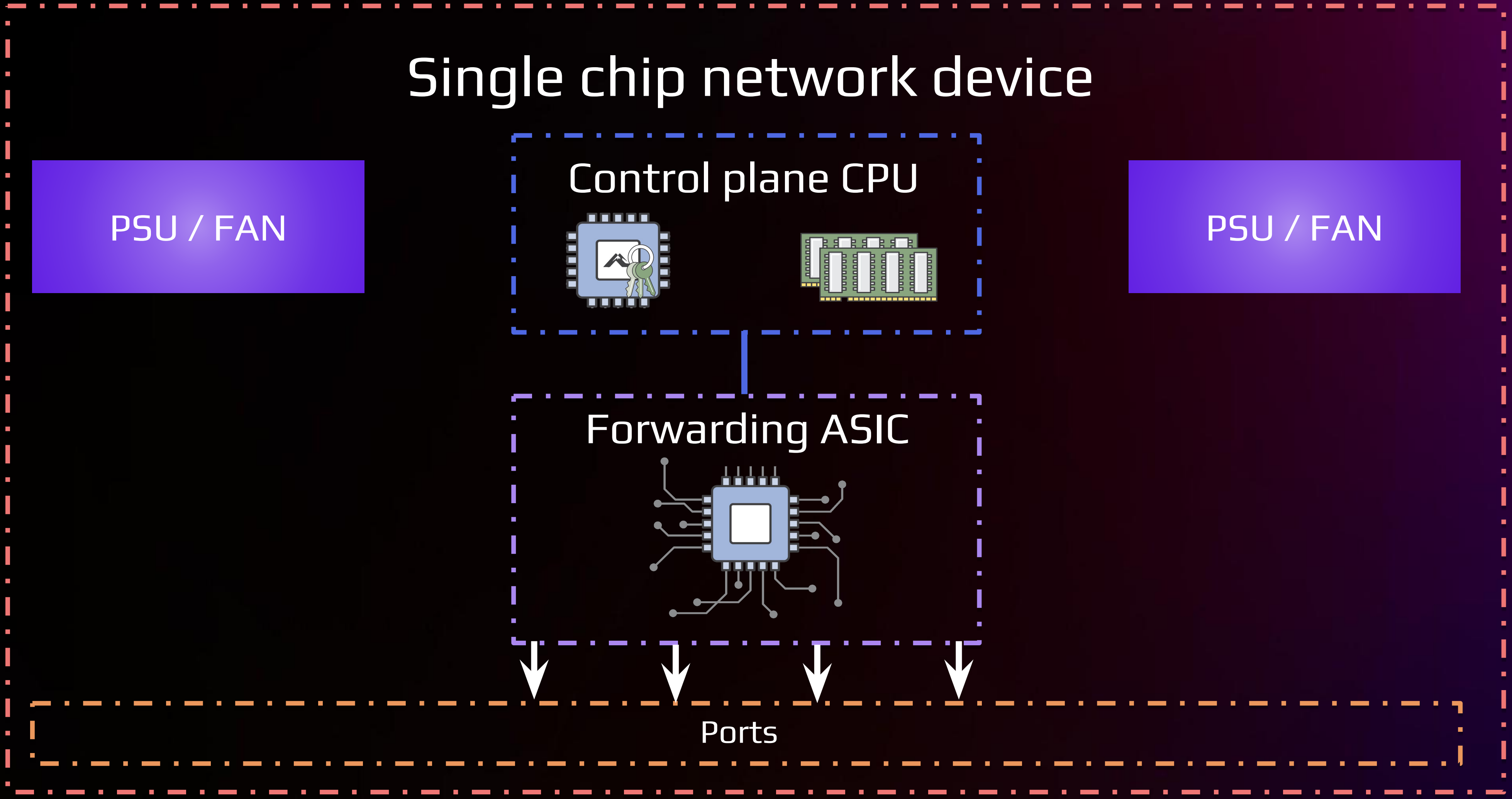
How we do it

"A complex system that works is invariably found to have evolved from a simple system that worked. The inverse proposition also appears to be true: A complex system designed from scratch never works and cannot be made to work. You have to start over, beginning with a working simple system."

John Gall, *General Systemantics: An essay on how systems work, and especially how they fail*, 1975



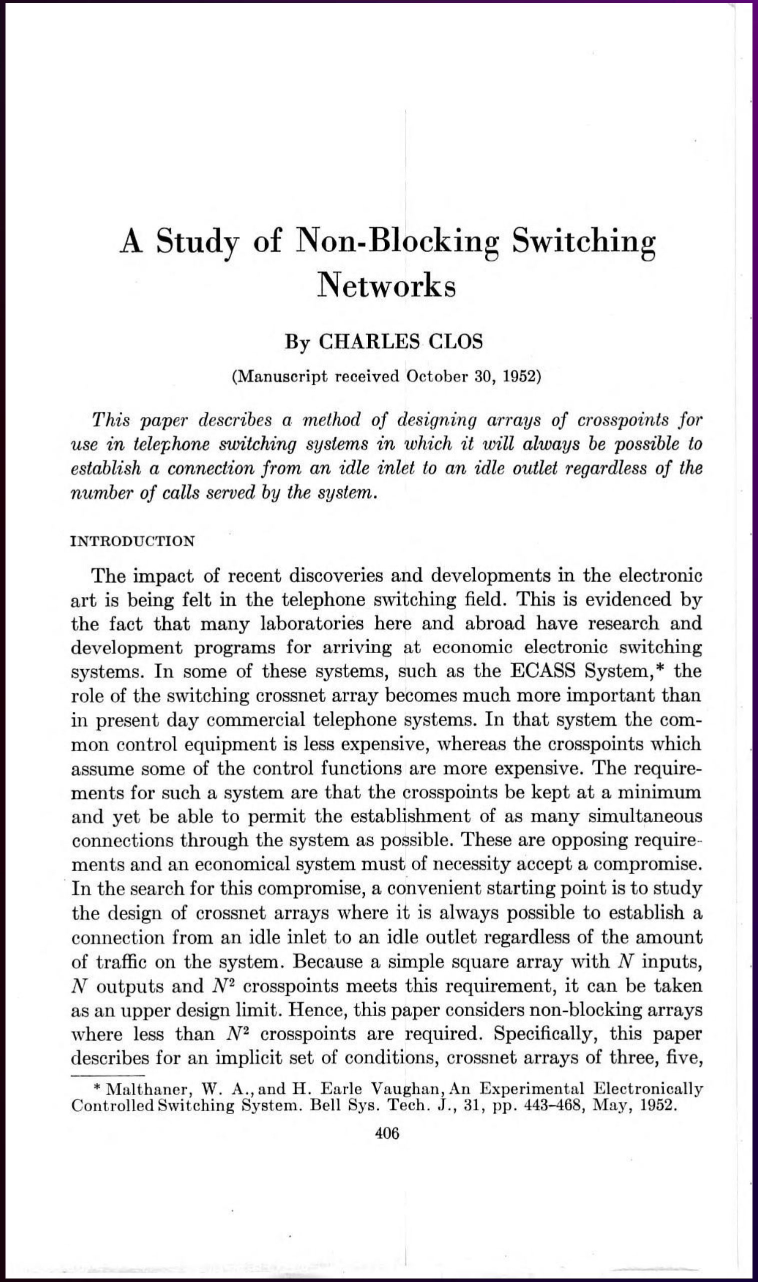
Single chip–based platforms



Create

TOPOLOGY AND HARDWARE

Clos fabric



A Study of Non-Blocking Switching Networks

By CHARLES CLOS

(Manuscript received October 30, 1952)

This paper describes a method of designing arrays of crosspoints for use in telephone switching systems in which it will always be possible to establish a connection from an idle inlet to an idle outlet regardless of the number of calls served by the system.

INTRODUCTION

The impact of recent discoveries and developments in the electronic art is being felt in the telephone switching field. This is evidenced by the fact that many laboratories here and abroad have research and development programs for arriving at economic electronic switching systems. In some of these systems, such as the ECASS System,* the role of the switching crossnet array becomes much more important than in present day commercial telephone systems. In that system the common control equipment is less expensive, whereas the crosspoints which assume some of the control functions are more expensive. The requirements for such a system are that the crosspoints be kept at a minimum and yet be able to permit the establishment of as many simultaneous connections through the system as possible. These are opposing requirements and an economical system must of necessity accept a compromise. In the search for this compromise, a convenient starting point is to study the design of crossnet arrays where it is always possible to establish a connection from an idle inlet to an idle outlet regardless of the amount of traffic on the system. Because a simple square array with N inputs, N outputs and N^2 crosspoints meets this requirement, it can be taken as an upper design limit. Hence, this paper considers non-blocking arrays where less than N^2 crosspoints are required. Specifically, this paper describes for an implicit set of conditions, crossnet arrays of three, five,

* Malthaner, W. A., and H. Earle Vaughan, An Experimental Electronically Controlled Switching System. Bell Sys. Tech. J., 31, pp. 443–468, May, 1952.

Phases of evolution

Consume
Create
Innovate

Innovate

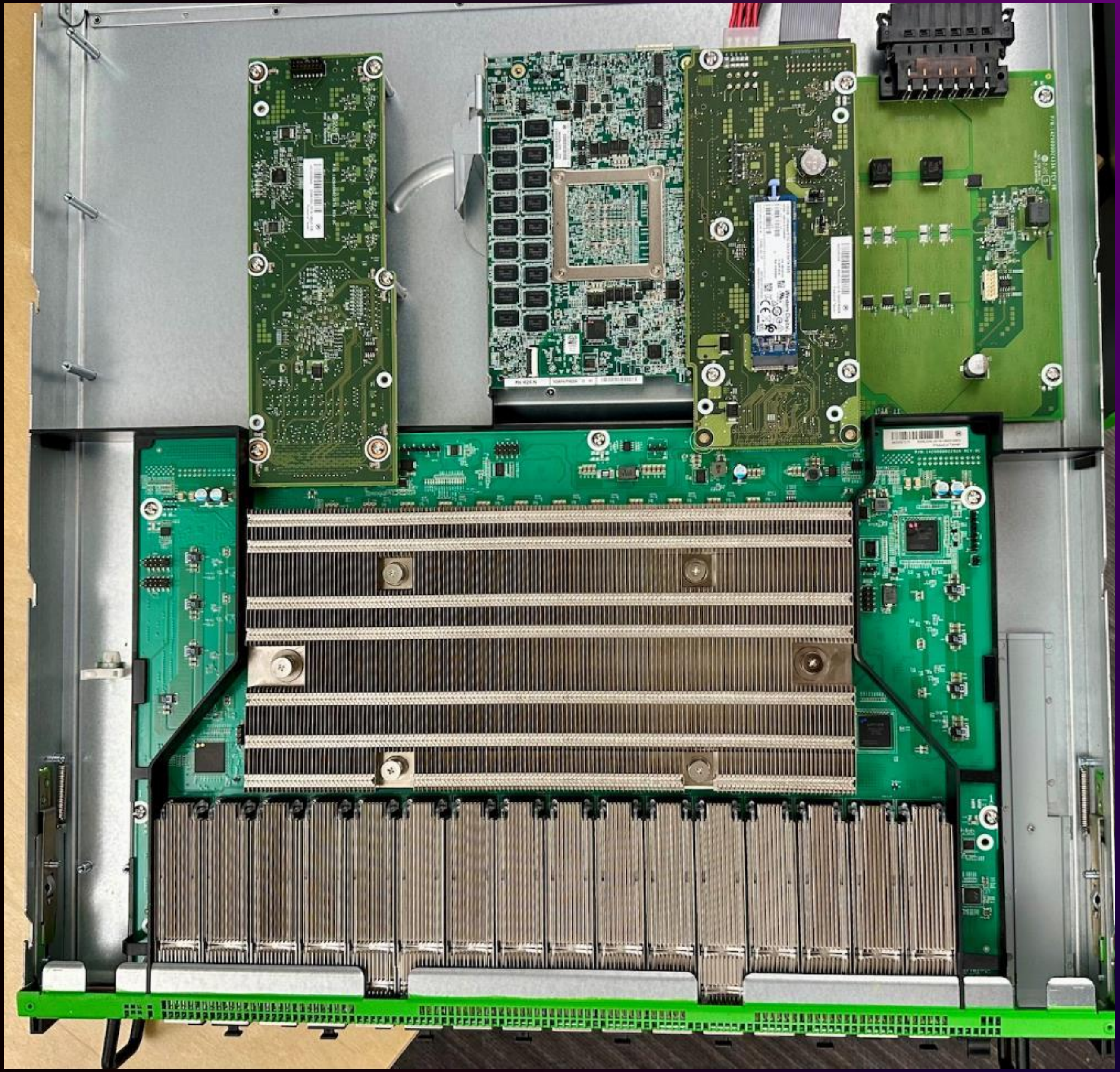
PRESENT DAY

Freedom to examine trade-offs

Custom hardware

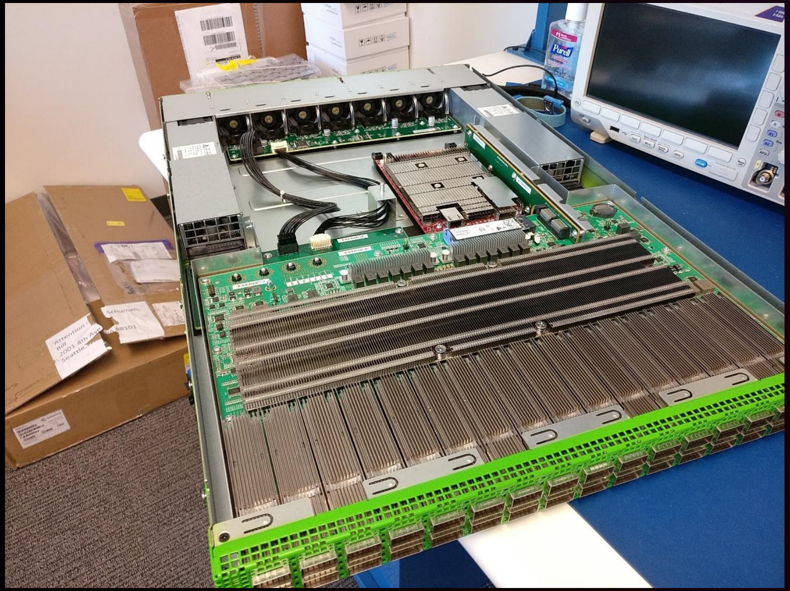
Multi-domain applications

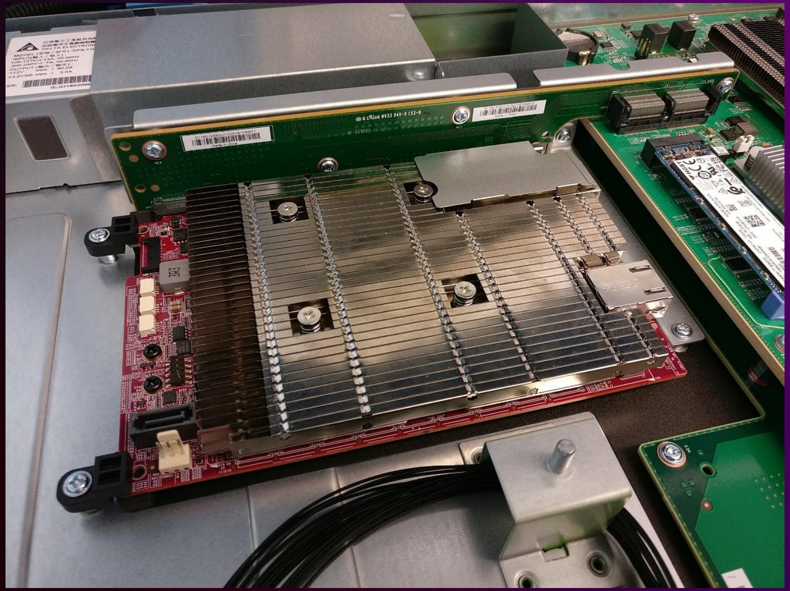
Focus on the benefit

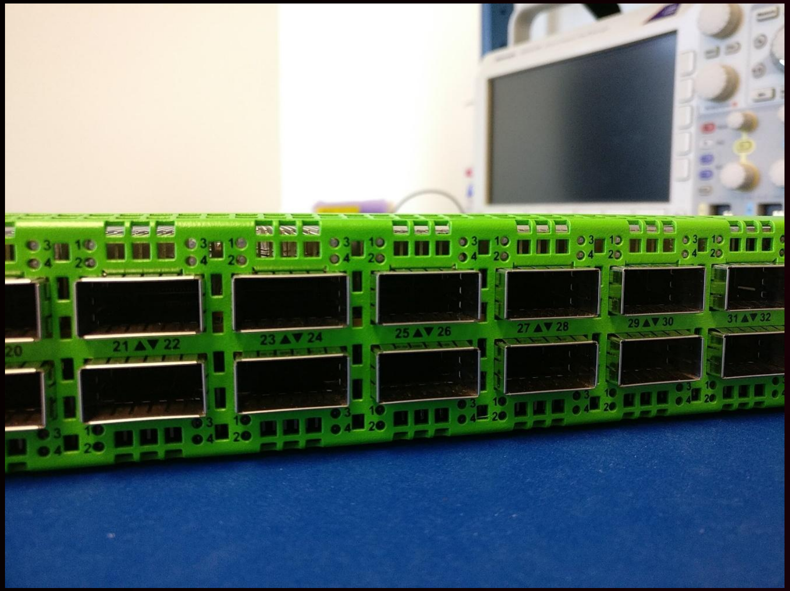


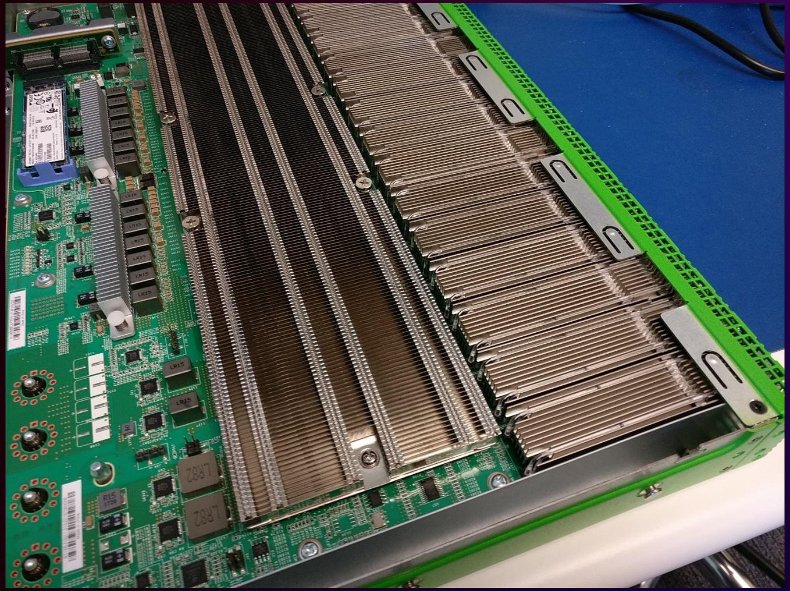
How we do it

WE'VE COME A LONG WAY

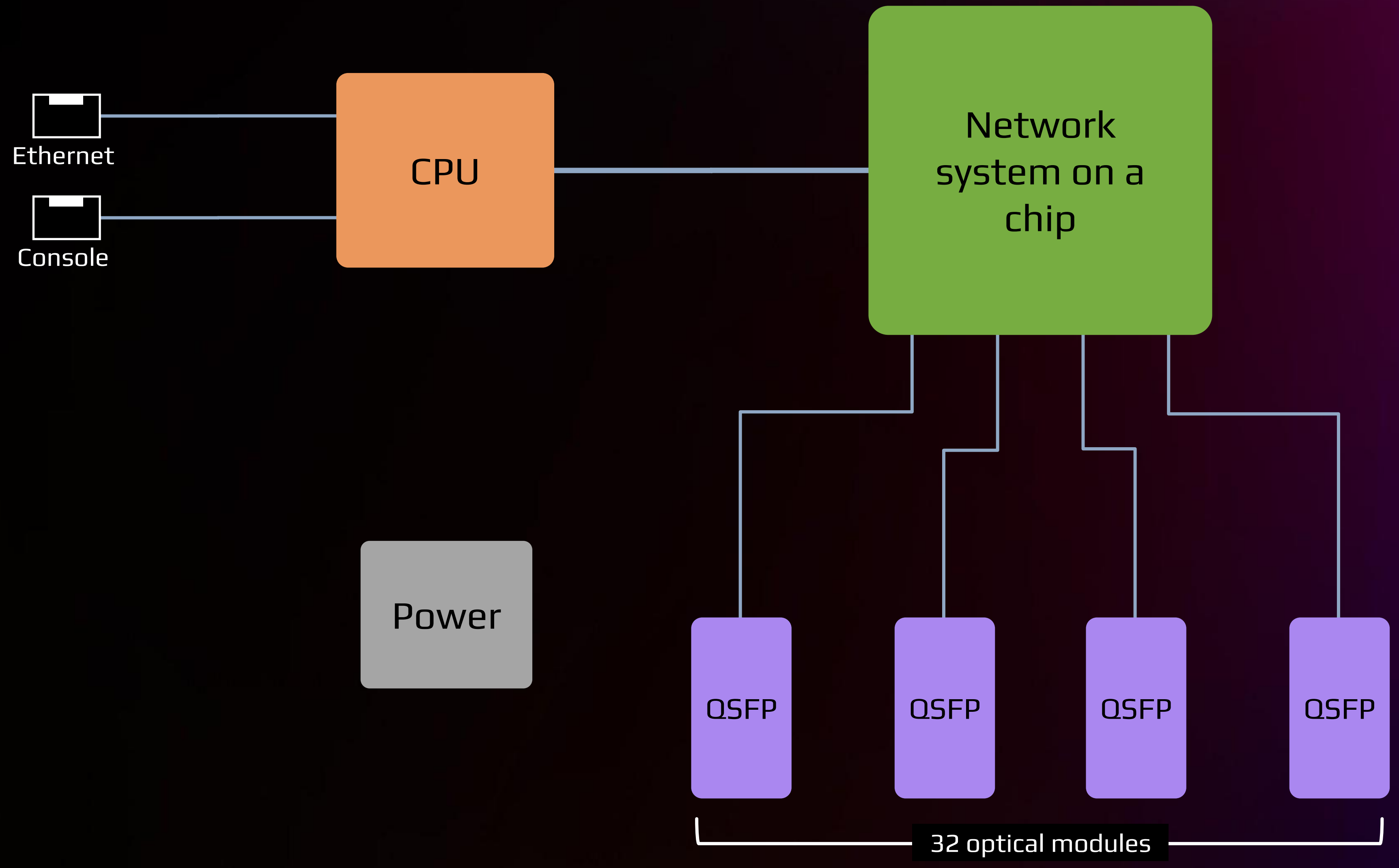




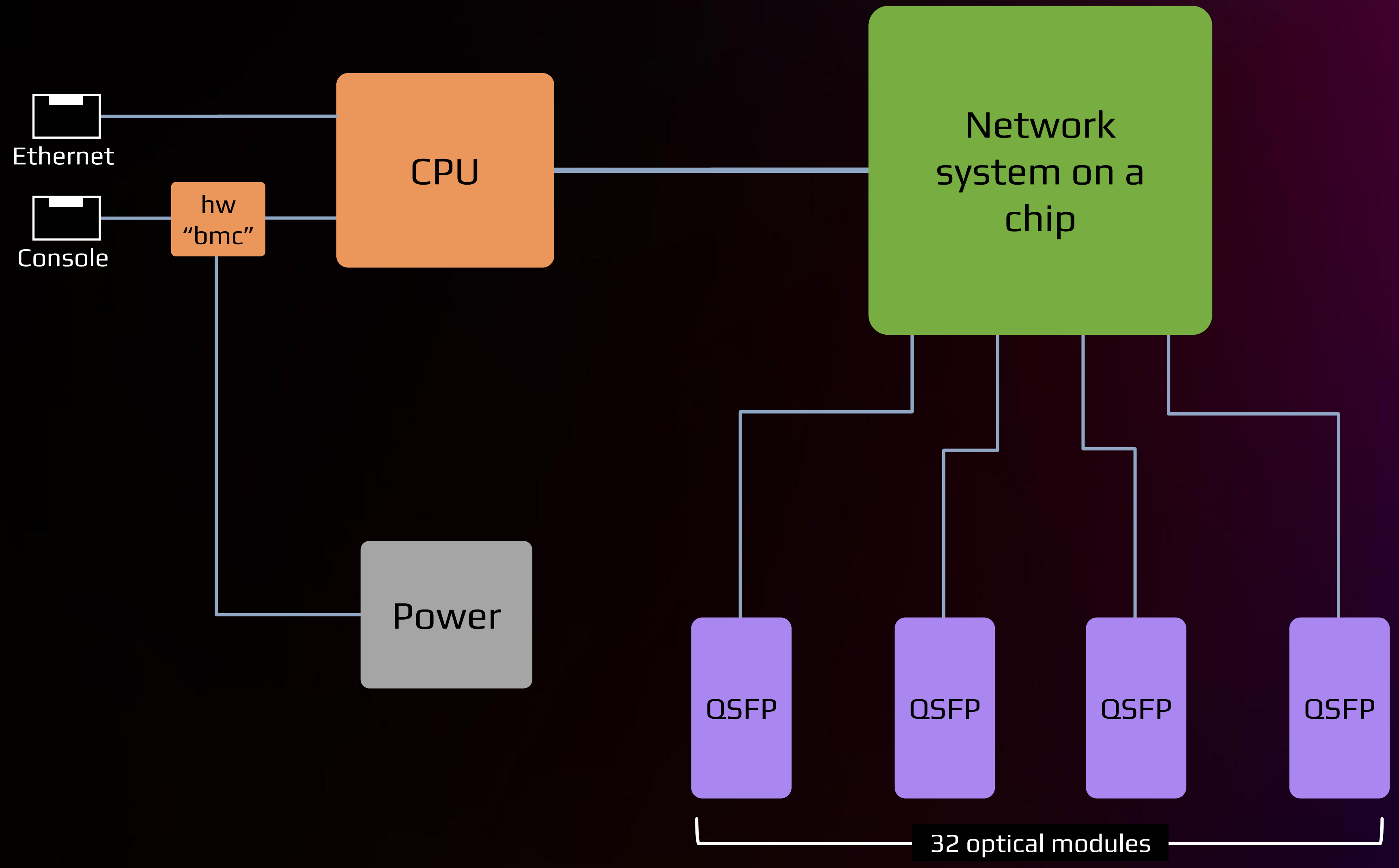




How we do it

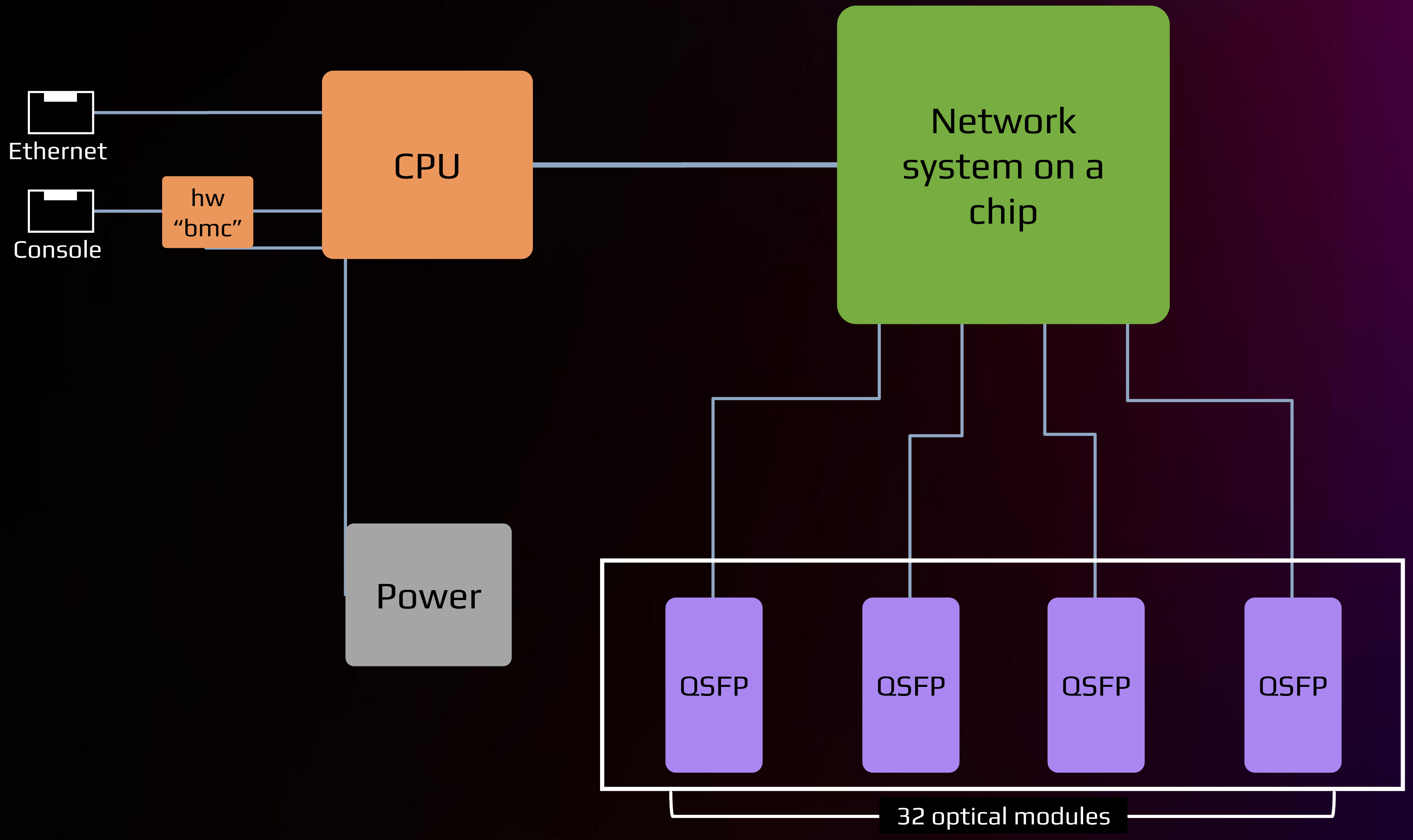

Ethernet
Console
CPU
Network system on a chip
Power
QSFP
QSFP
QSFP
QSFP
32 optical modules

How we do it

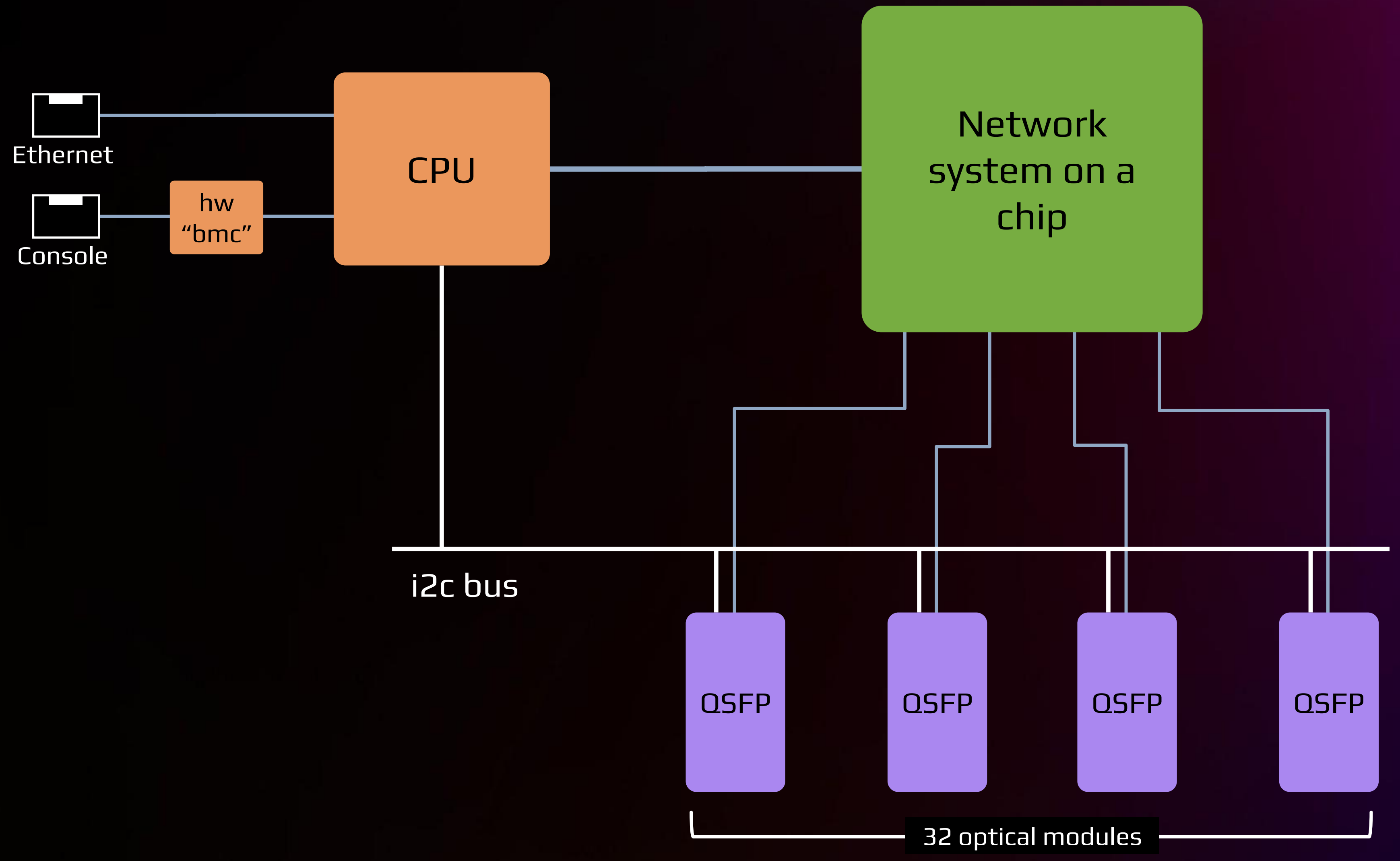


Ethernet
Console
hw “bmc”
CPU
Network system on a chip
Power
QSFP
QSFP
QSFP
QSFP
32 optical modules

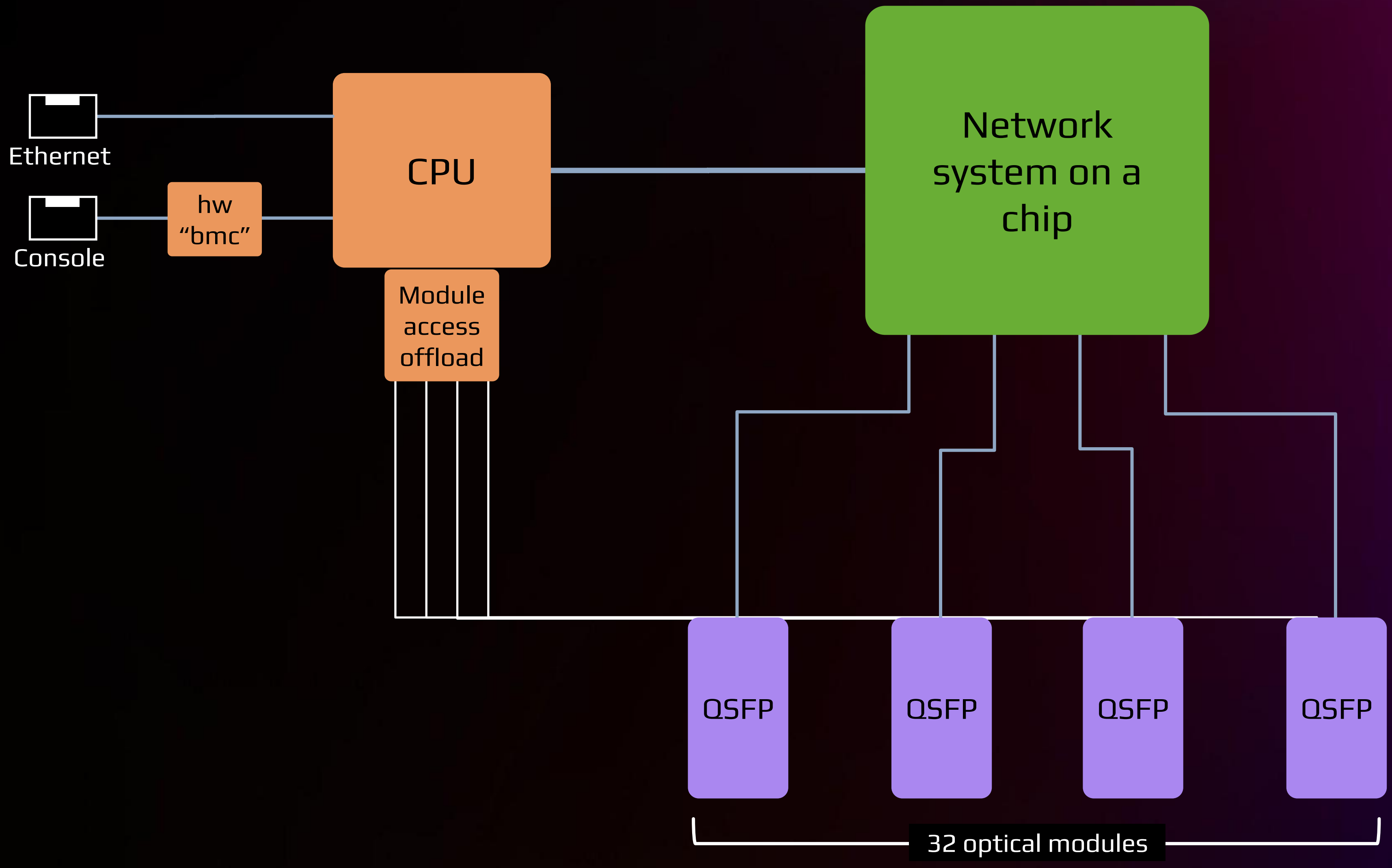
How we do it


Ethernet
Console
hw
“bmc”
CPU
Network system on a chip
Power
QSFP
QSFP
QSFP
QSFP
32 optical modules

How we do it


Ethernet
Console
hw
“bmc”
CPU
Network system on a chip
i2c bus
QSFP
QSFP
QSFP
QSFP
32 optical modules

How we do it


Ethernet
Console
hw “bmc”
CPU
Module access offload
Network system on a chip
QSFP
QSFP
QSFP
QSFP
32 optical modules

Innovate

PRESENT DAY

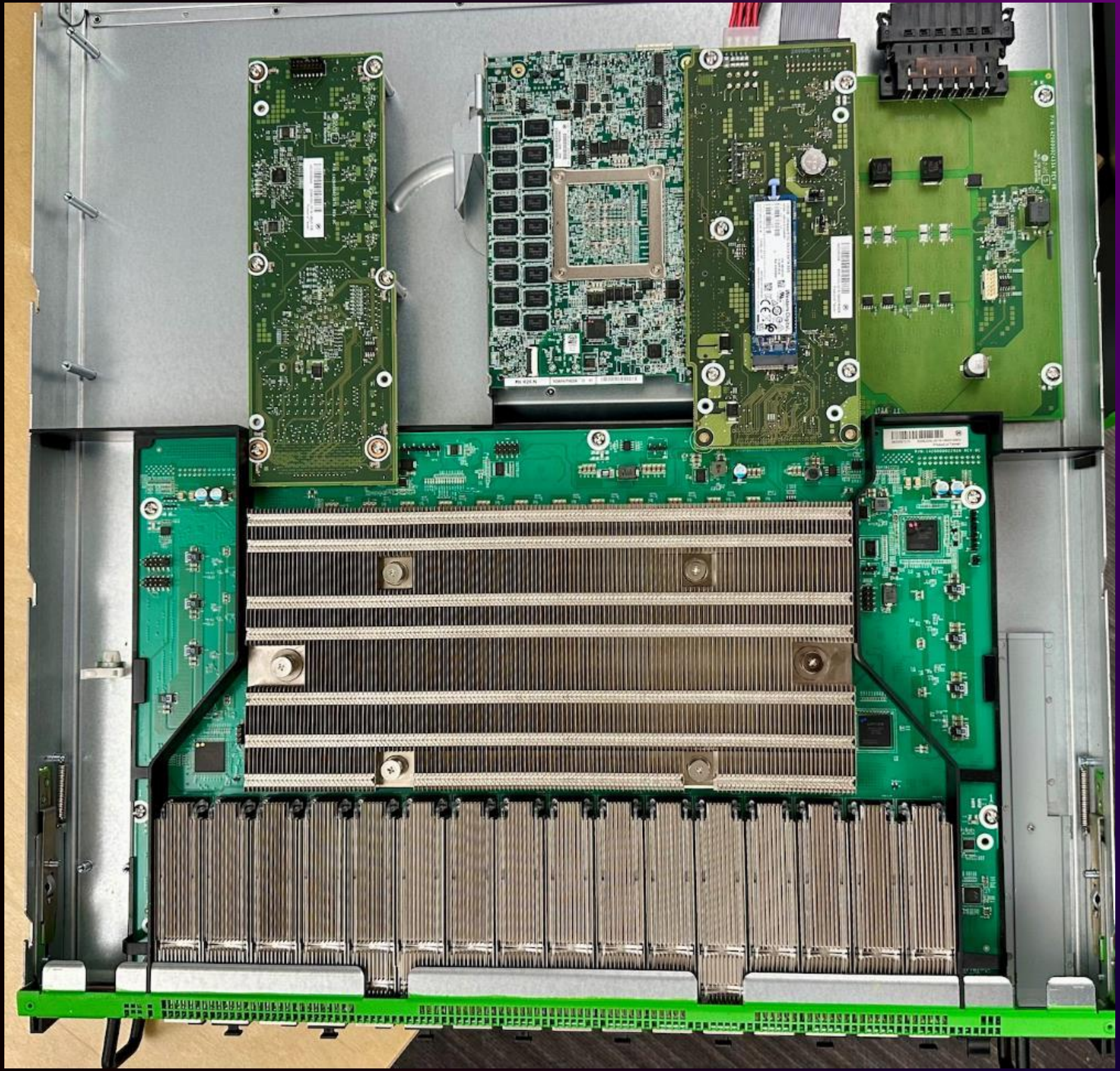
ASIC & Optics Board

CPU & Memory Card

Hardware BMC & Storage Board

I2C Offload Module

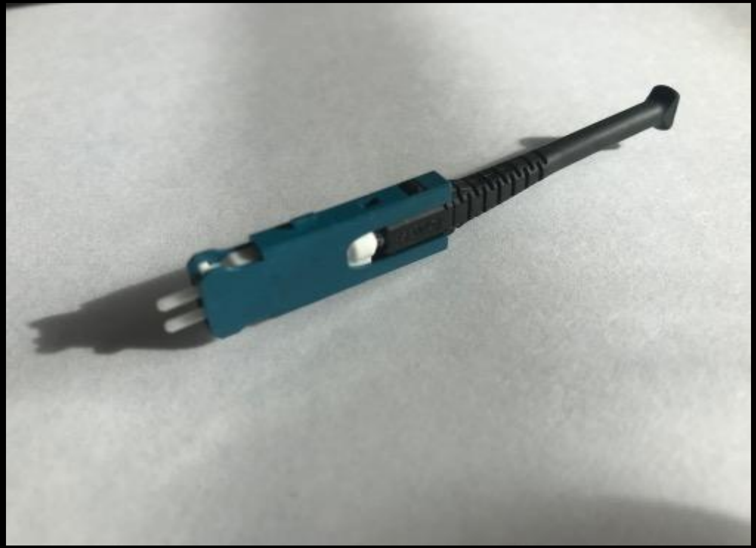
Power Delivery

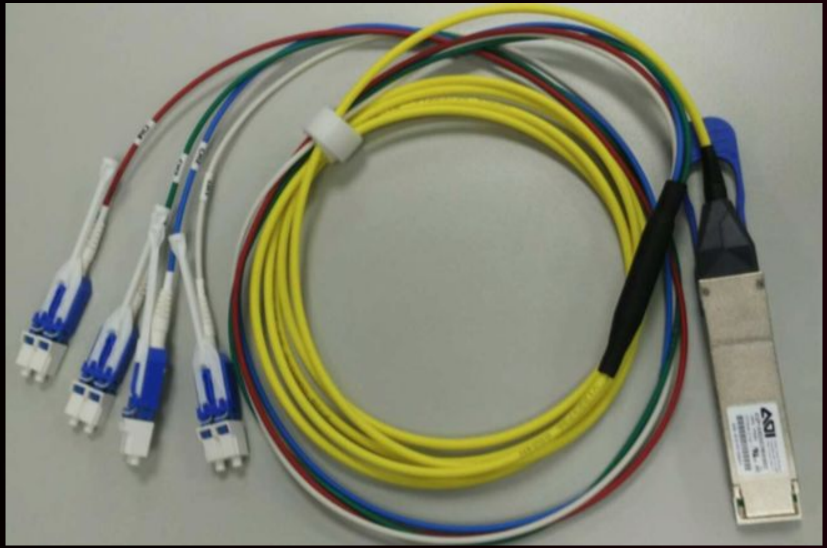


How we do it – 102.8T rack

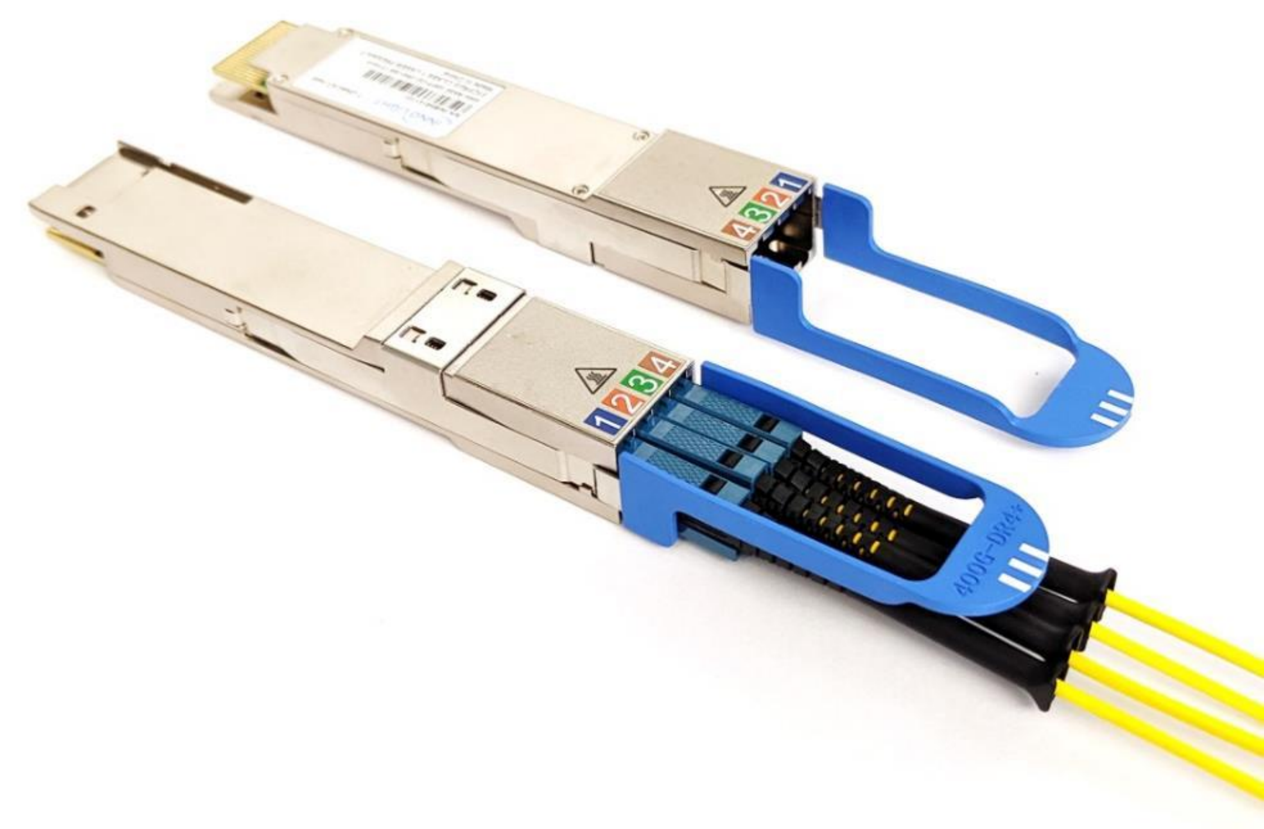
- 16+16 32x400G Devices
  - 1024x400G ports total
  - 256x400G ports for Consumers
  - Max 30.8kVA per rack
- Direct-attach copper (DAC) cabling
  - 100G 6.7mm OD at 2.5m
  - 400G 11mm OD at 2.5m
- Active DAC with retimers

How we do it – Short reach





How we do it – SN connector



How we do it

MEDIUM HAUL

Data center interconnect (DCI)

OIF 400G ZR

400G – ZR+ 400km

Integrated routing, DWDM, encryption

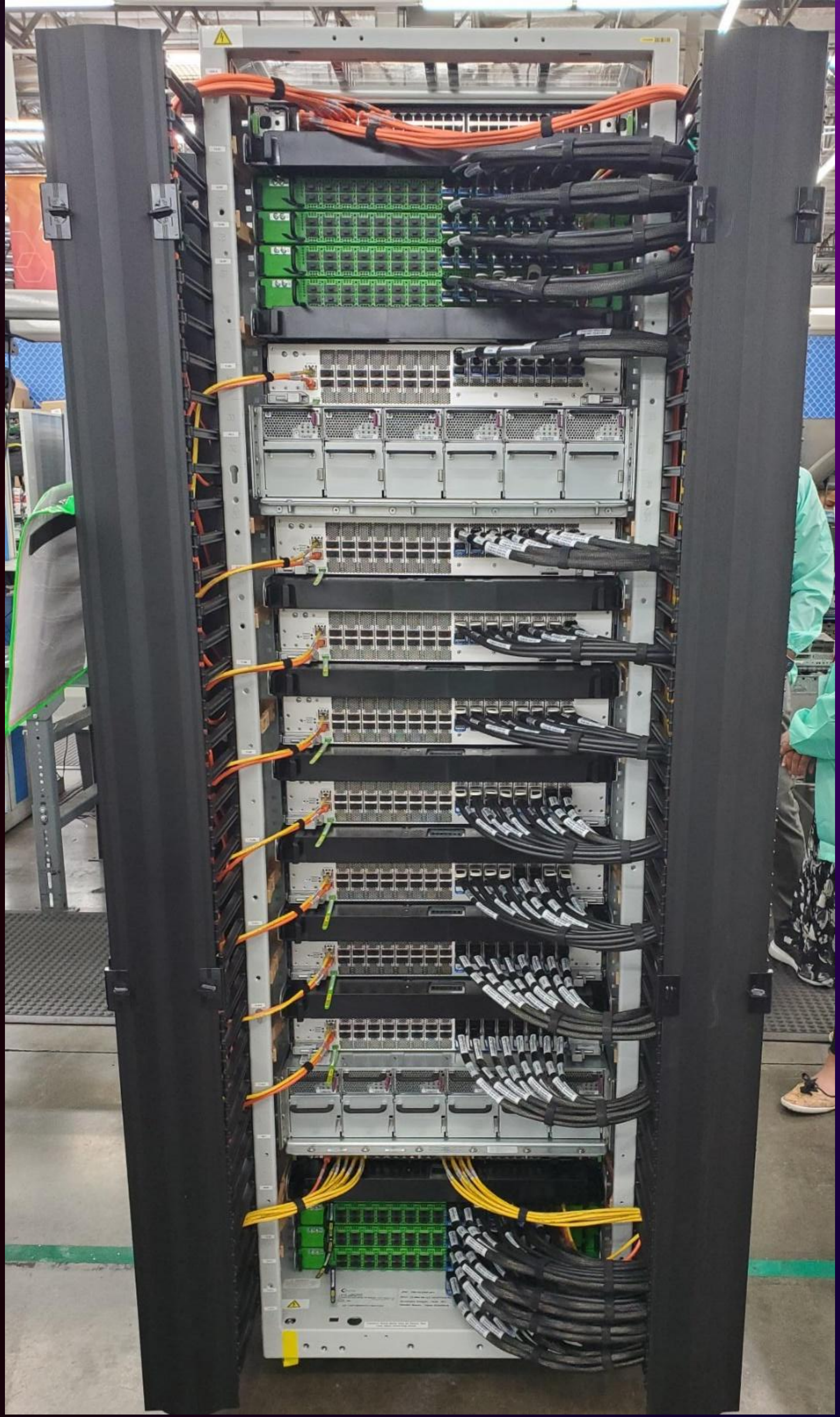
How we do it – 51.2T rack

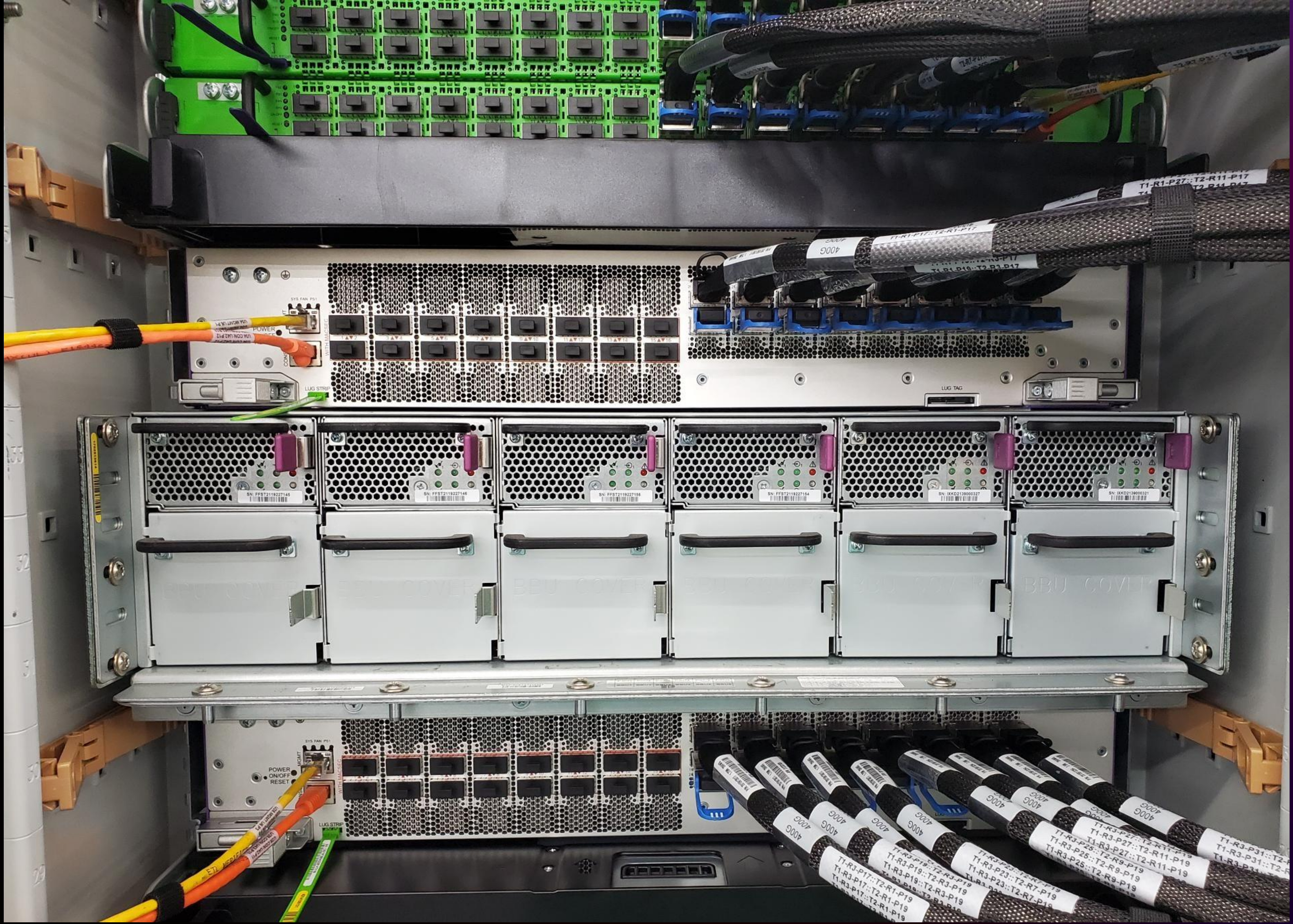
MEDIUM HAUL

8x 12.8Tbit/s T2 Devices

8x 12.8Tbit/s DWDM Switches

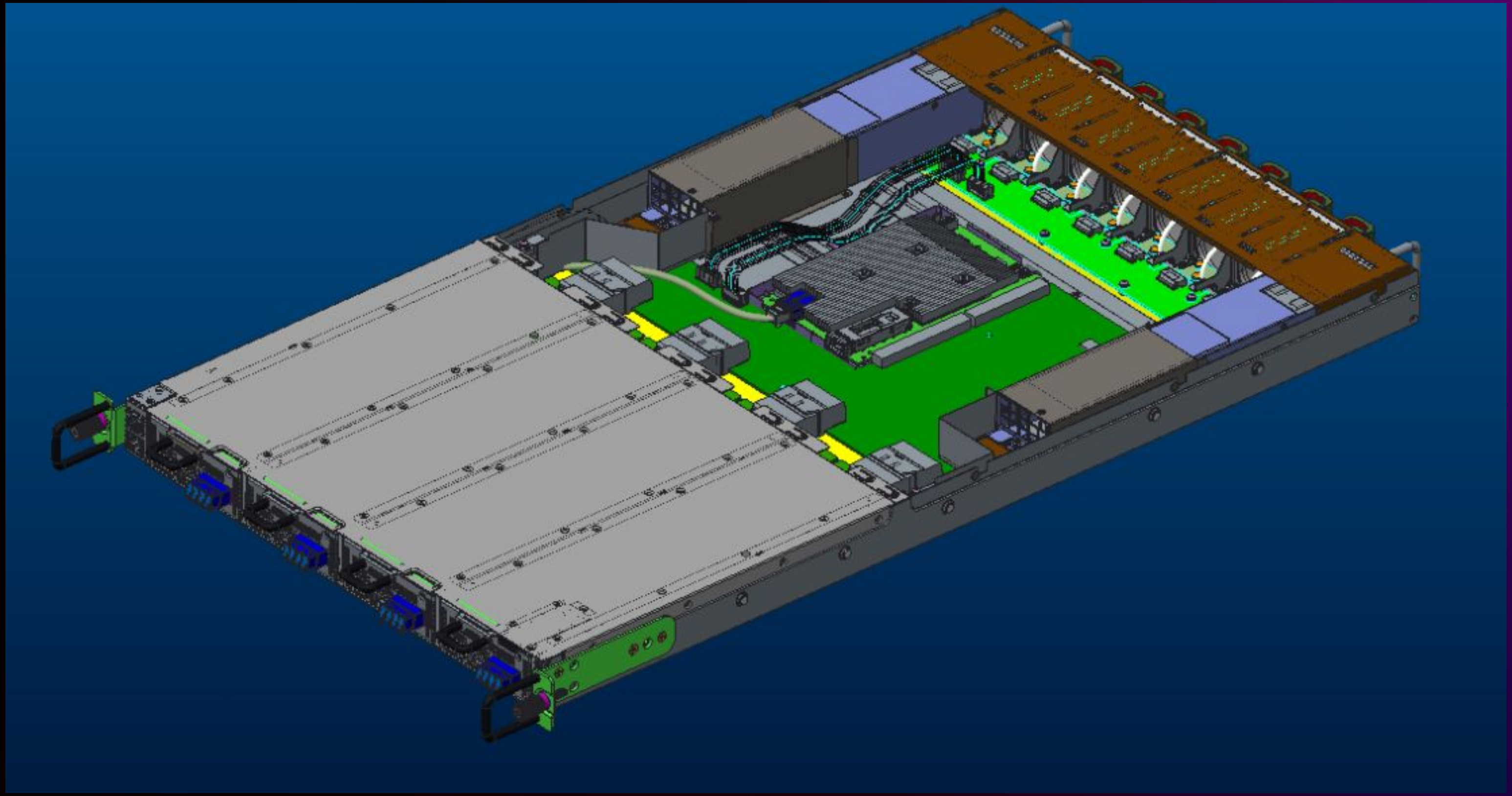
8x16x400G ZR(+) Ports




BBU COVER
400G
T1-R1-P27::T2-R11-P17
T1-R1-P17::T2-R1-P17
T1-R1-P19::T2-R1-P17
T1-R3-P17::T2-R1-P19
T1-R3-P19::T2-R3-P19
T1-R3-P23::T2-R7-P19
T1-R3-P25::T2-R9-P19
T1-R3-P27::T2-R11-P19
T1-R3-P31::T2-
LUG STRIP
LUG TAG
POWER
ON/OFF
RESET

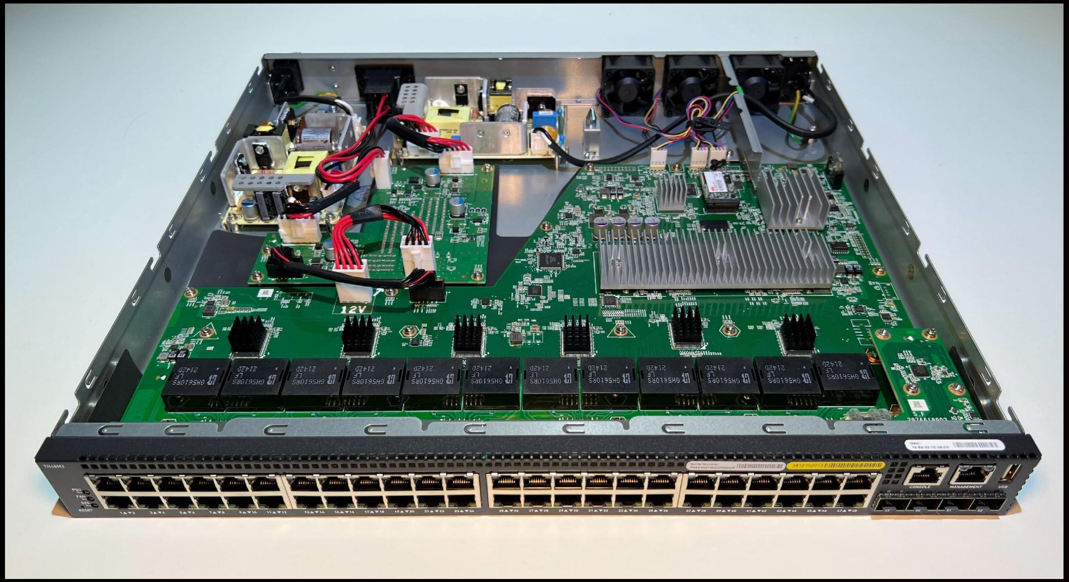
AWS DWDM Platform

- Four optical sleds
  - 2x400G QSFP-DD to DWDM
- Firmware upgradeable
  - fine tune link quality
- Layer 1 Encryption
  - AES-256

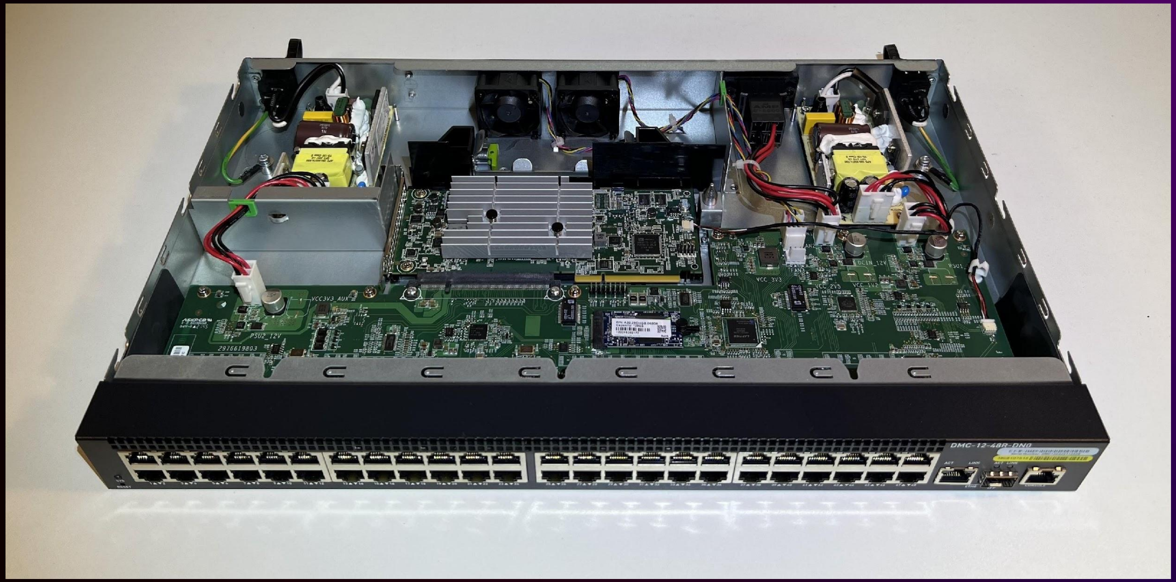


How we do it

MANAGEMENT NETWORKS



Out-of-band switch



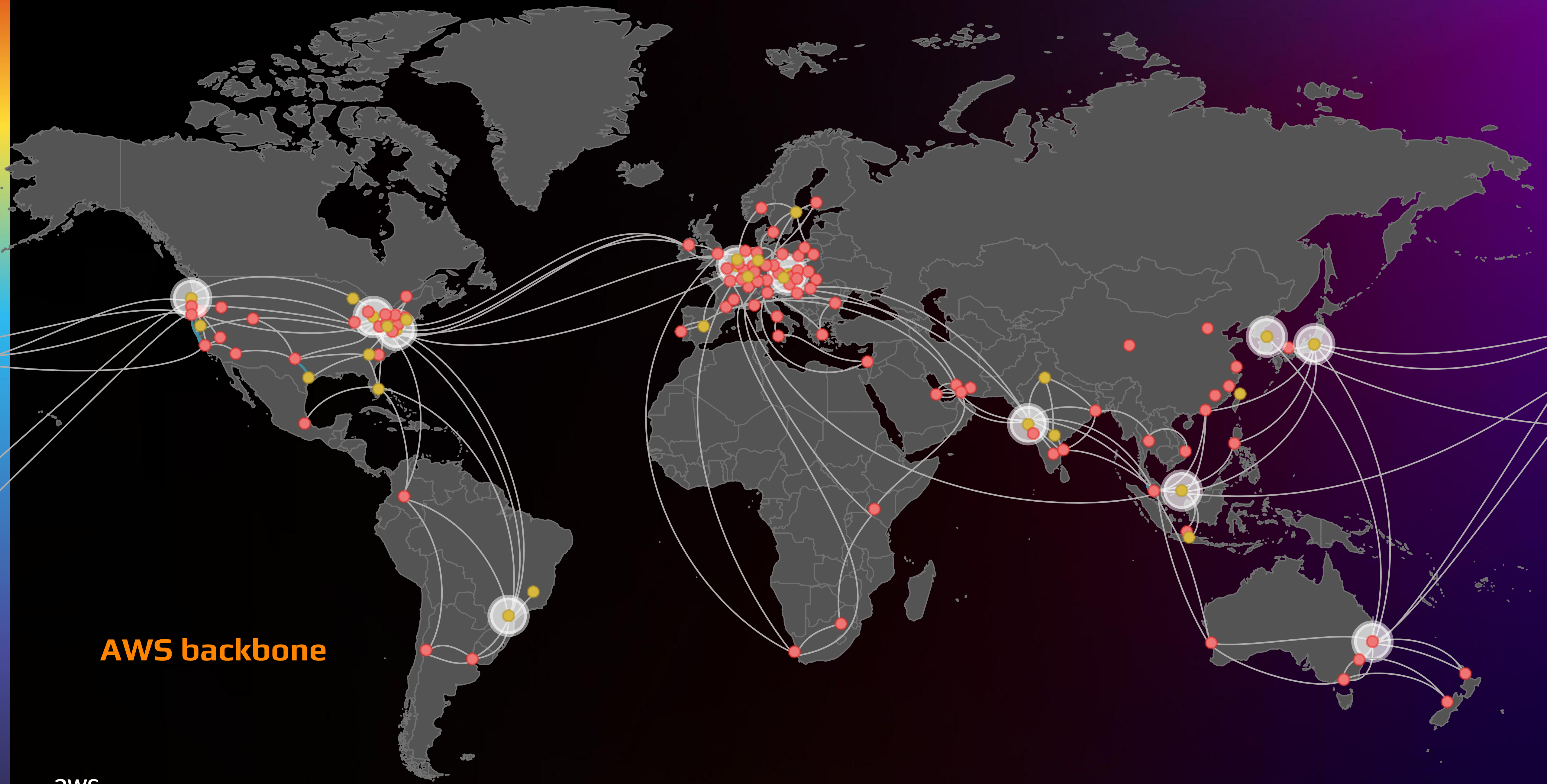
Console server

Thank you!

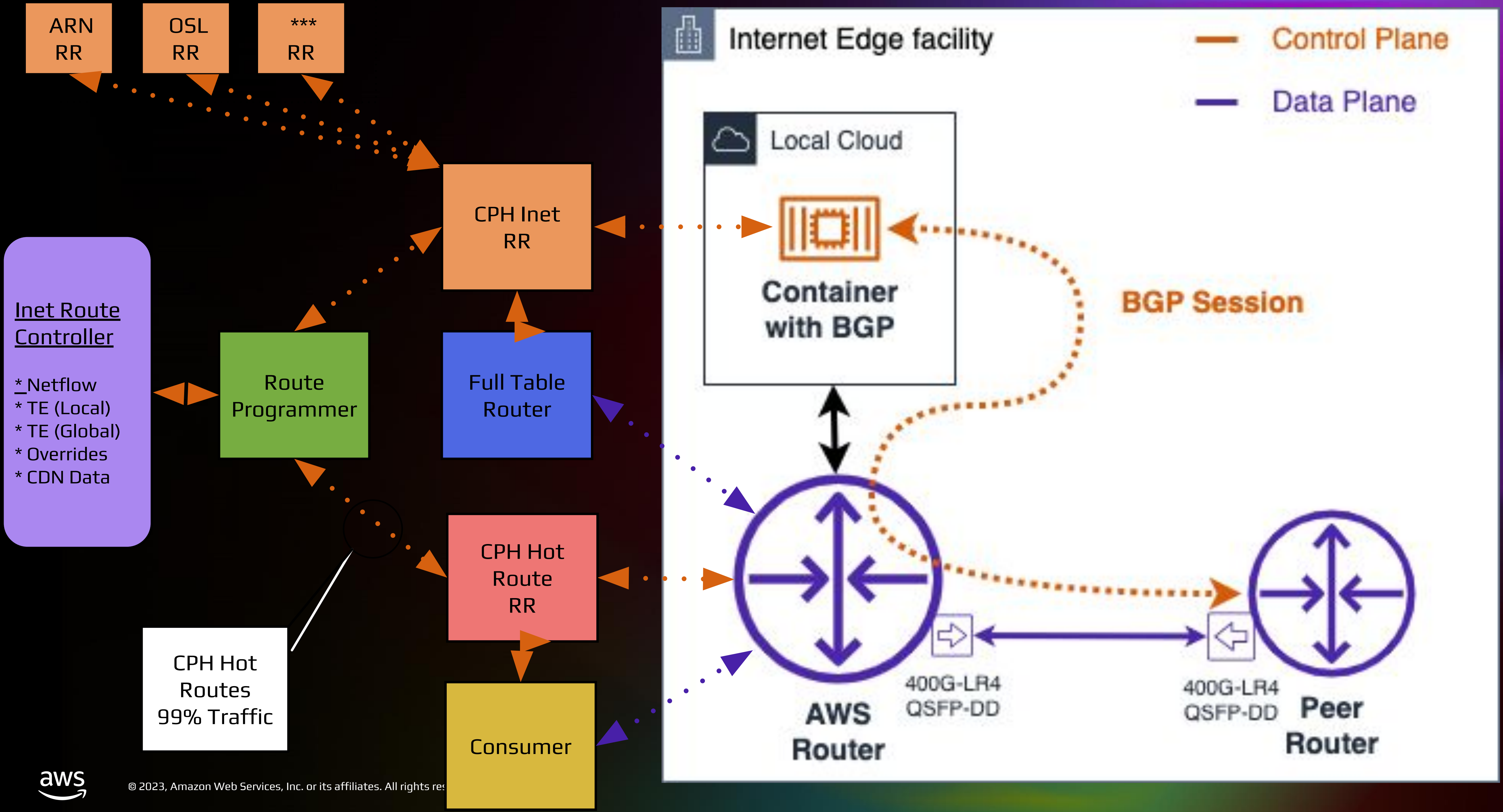
Colin Whittaker
colinwh@amazon.com

Giacomo Bernardi
giacombe@amazon.com

Giorgio Bonfiglio
bonfigg@amazon.com



AWS backbone



ARN RR

OSL RR

*** RR

CPH Inet RR

Inet Route Controller

* Netflow
* TE (Local)
* TE (Global)
* Overrides
* CDN Data

Route Programmer

Full Table Router

CPH Hot Route RR

CPH Hot Routes 99% Traffic

Consumer

Tenets

Tenets

Secure

Tenets

Secure

Available

Tenets

Secure

Available

Scalable

Tenets

Secure

Available

Scalable

Performant

Phases of evolution

Consume
Create
Innovate

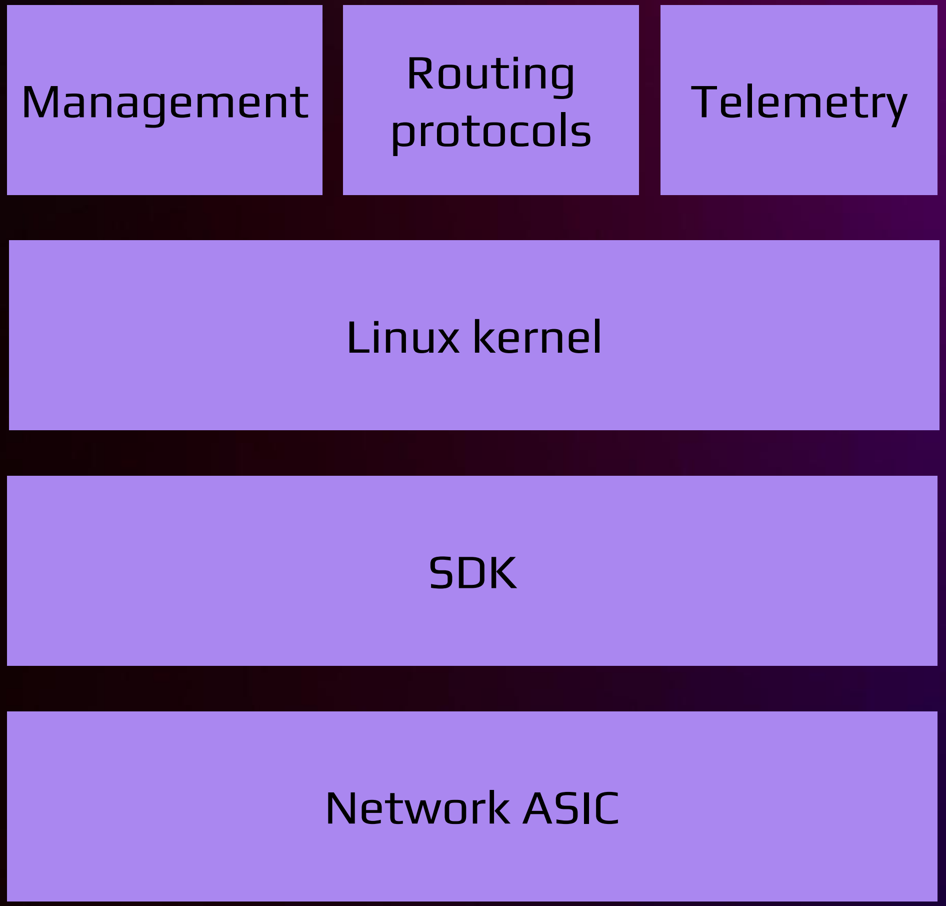
Create

NETWORK OPERATING SYSTEM

Linux-based

Multi-sourced manufacturing

Multi-ASIC



Create

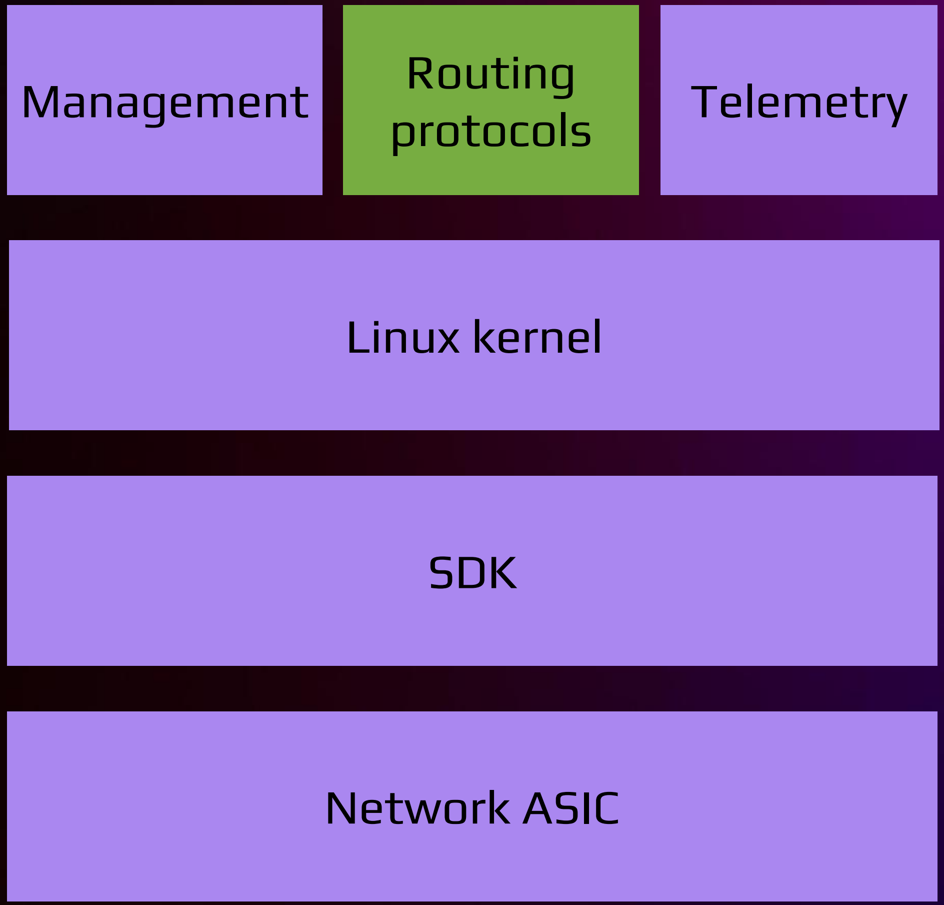
NETWORK OPERATING SYSTEM

Linux-based

Multi-sourced manufacturing

Multi-ASIC

OSPF/BGP ++



Management

Routing protocols

Telemetry

Linux kernel

SDK

Network ASIC

Create

Config generation

Deployment coordination

Active telemetry

Auto-remediation

NOC-less

Phases of evolution

Consume

Create

Innovate

Metal boxes and a lot of cables

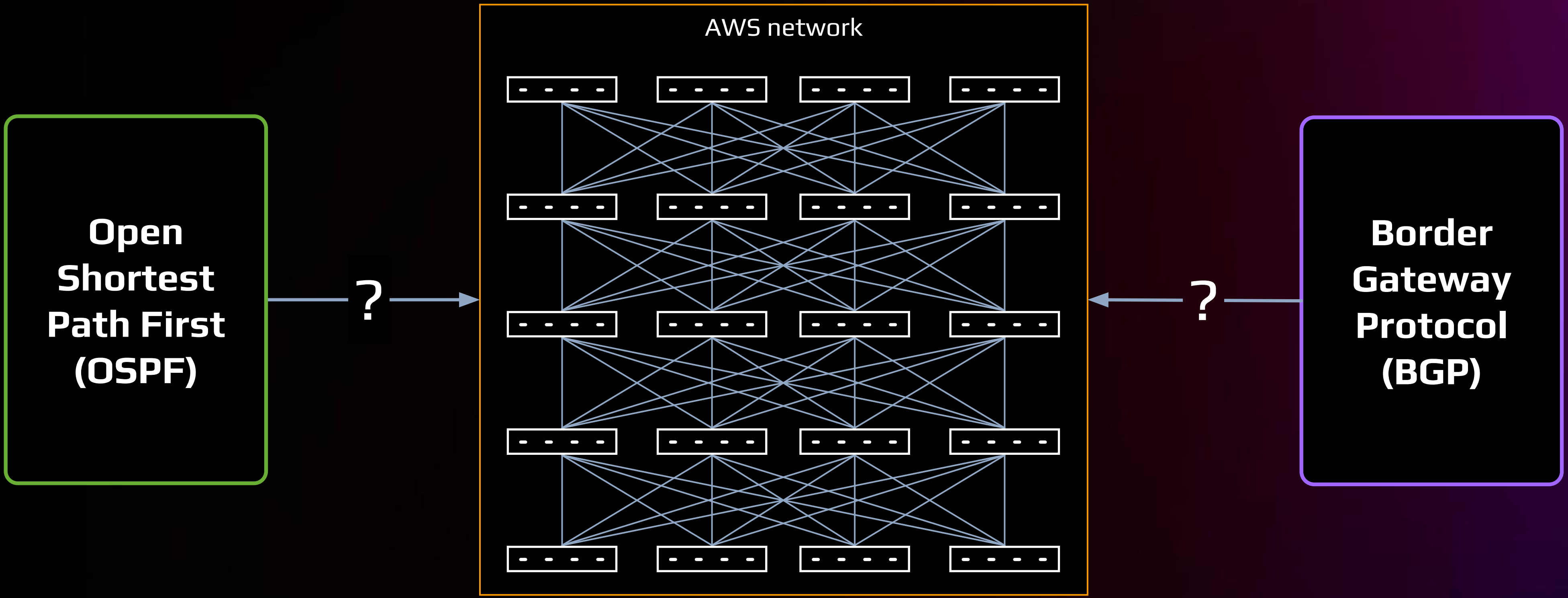
Small number of rack variations

Rack and cable switches for burn-in

Collect inventory and compare with bill of materials

Reprogram with AWS controlled binaries

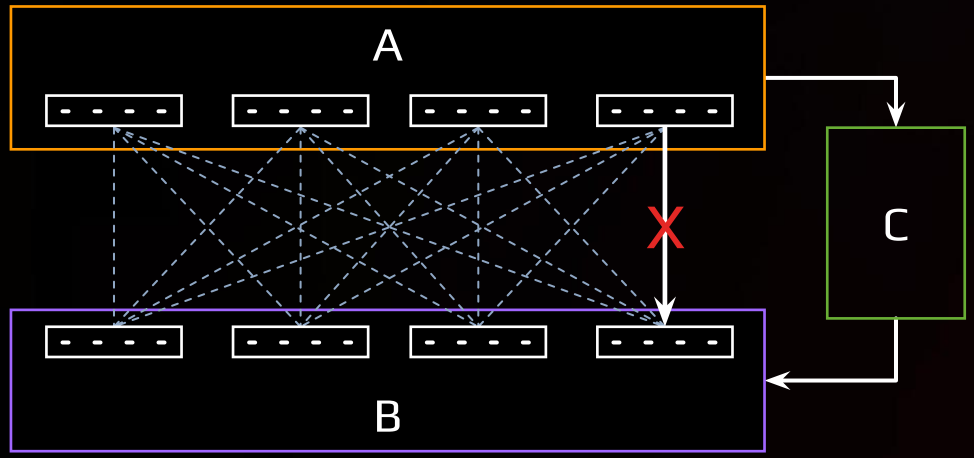
Which way do I go?



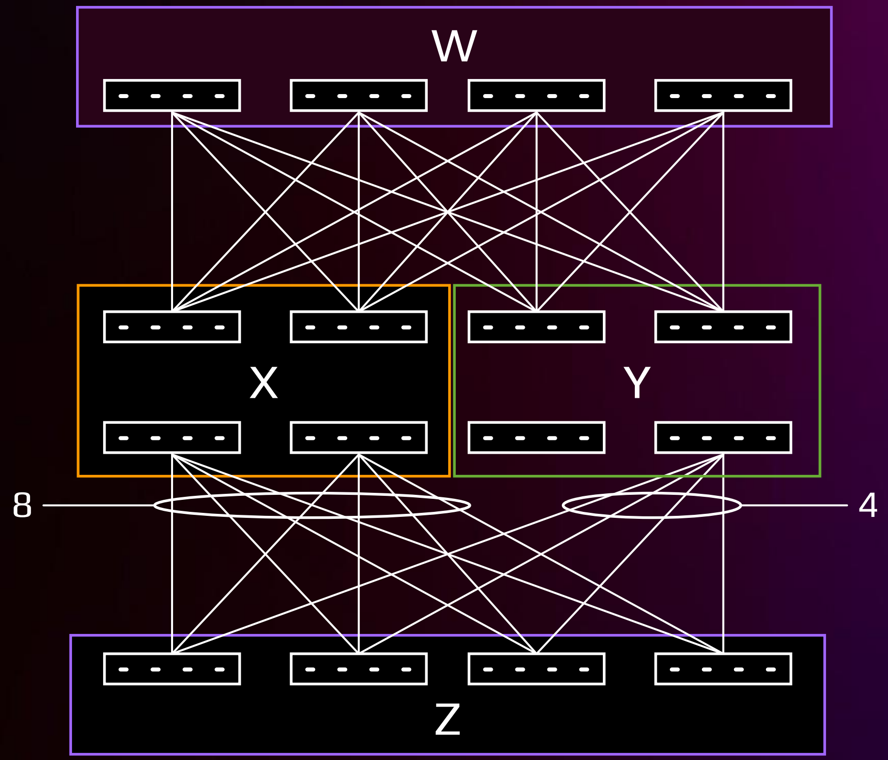
Open Shortest Path First (OSPF)
?
AWS network
?
Border Gateway Protocol (BGP)

Which way do I go?

ISSUES WITH OFF-THE-SHELF PROTOCOLS



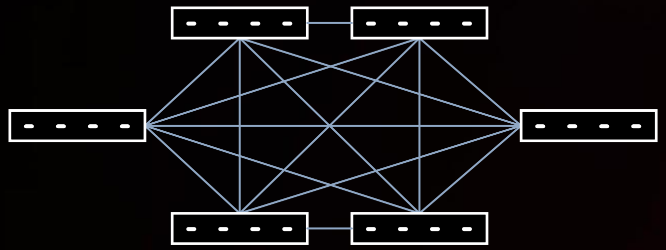
Last link standing

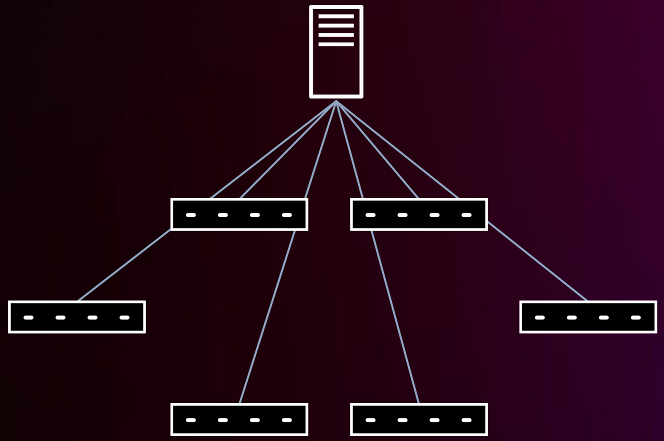


Cross-domain imbalance

Which way do I go?

DISTRIBUTED VERSUS CENTRAL





Statically stable   Low scope of impact

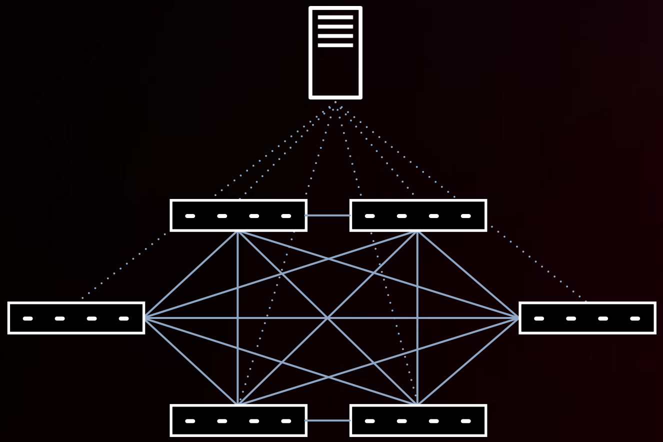
Distributed (classical)

High visibility   Deterministic

Centralized (SDN)

Which way do I go?

BEST OF BOTH WORLDS



Statically stable

Highly visible

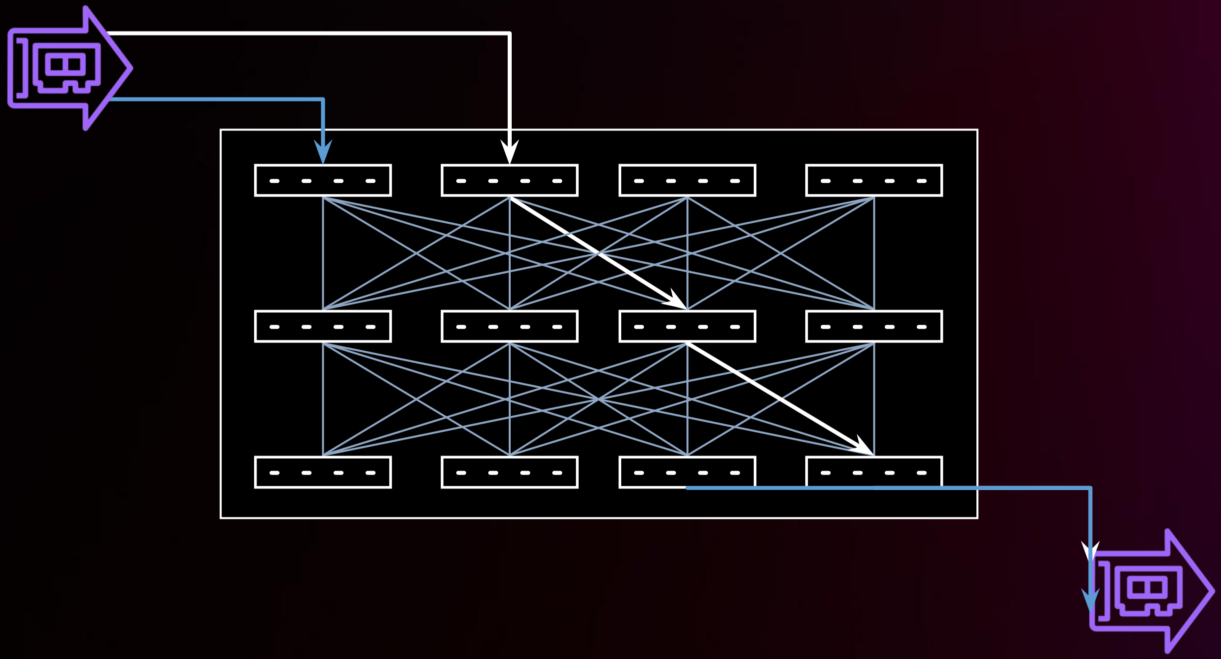
Deterministic

Low scope of impact

Hybrid

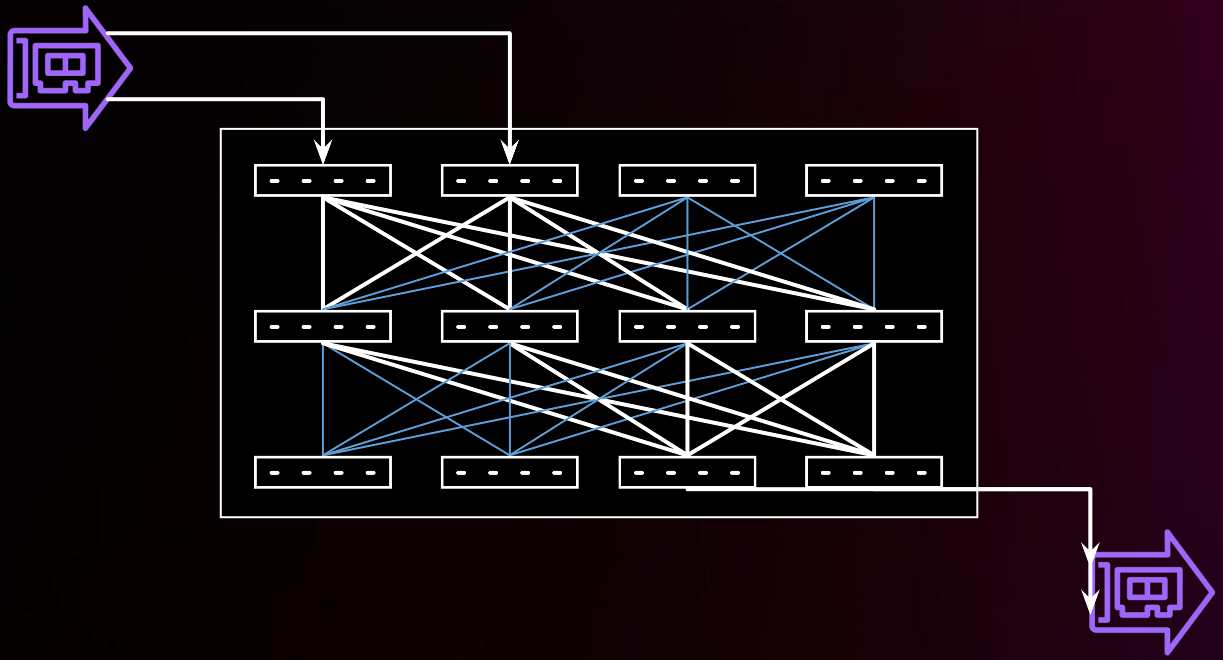
So many paths!

TRADITIONAL TCP BEHAVIOR



So many paths!

ELASTIC FABRIC ADAPTER (EFA) AND SCALABLE RELIABLE DATAGRAM



Dave Brown's Keynote

Session: NET211-L

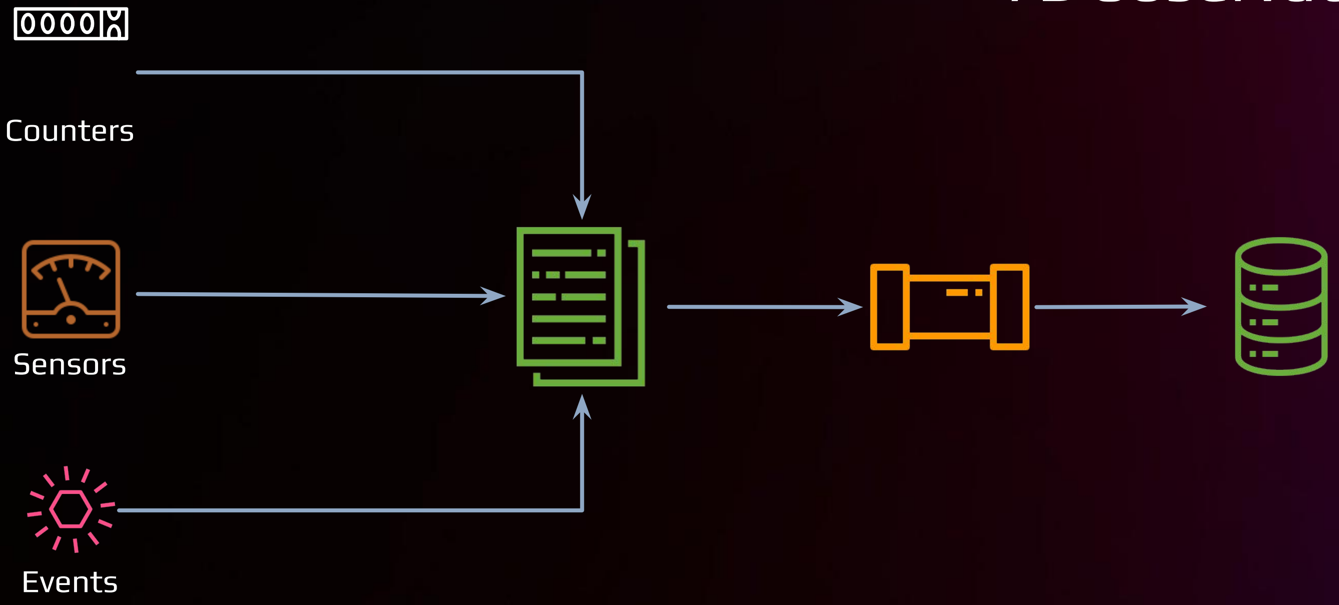
Monday Night Live with Peter DeSantis – 2018

Scaling HPC Applications on EC2 – 2018

Doctor, why does it hurt?

PASSIVE MONITORING

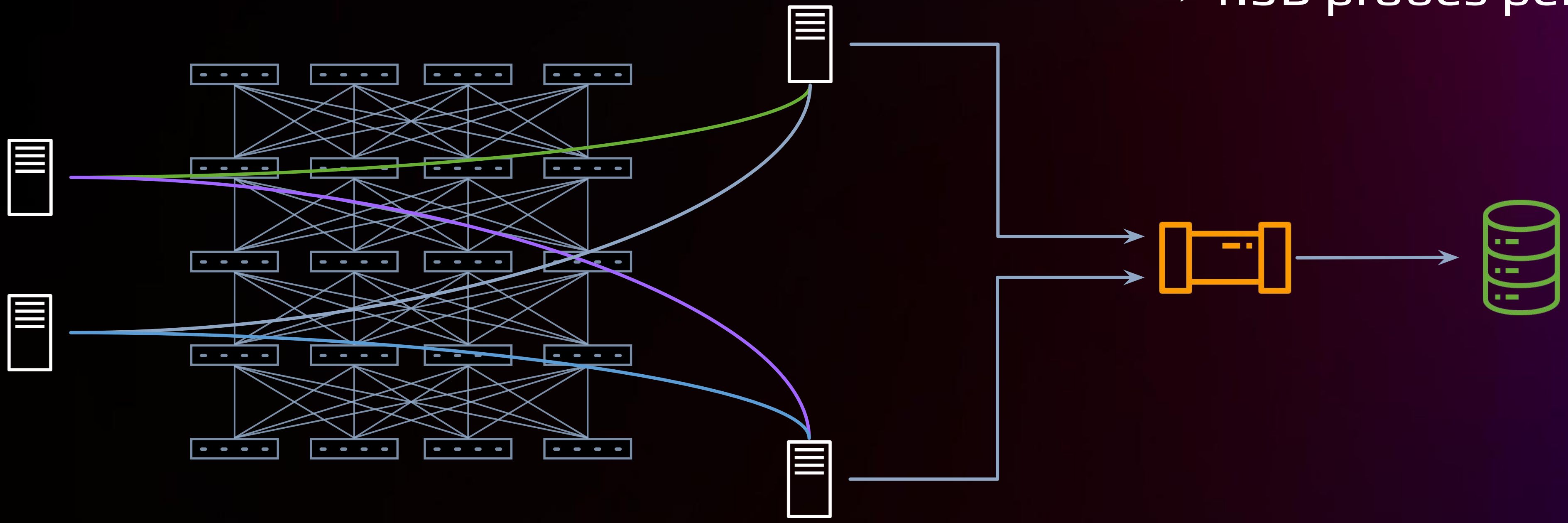
>7B observations per minute



Doctor, why does it hurt?

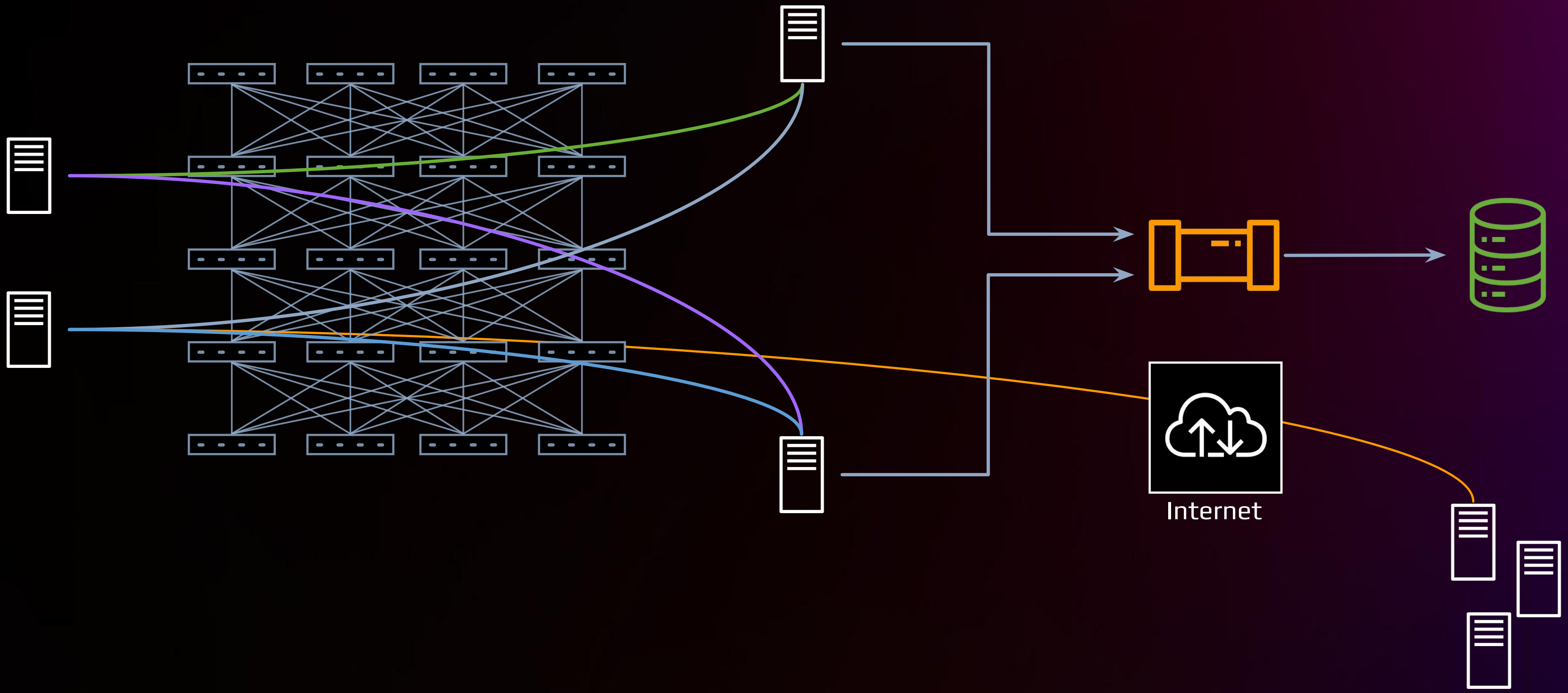
ACTIVE MONITORING

>1.5B probes per minute



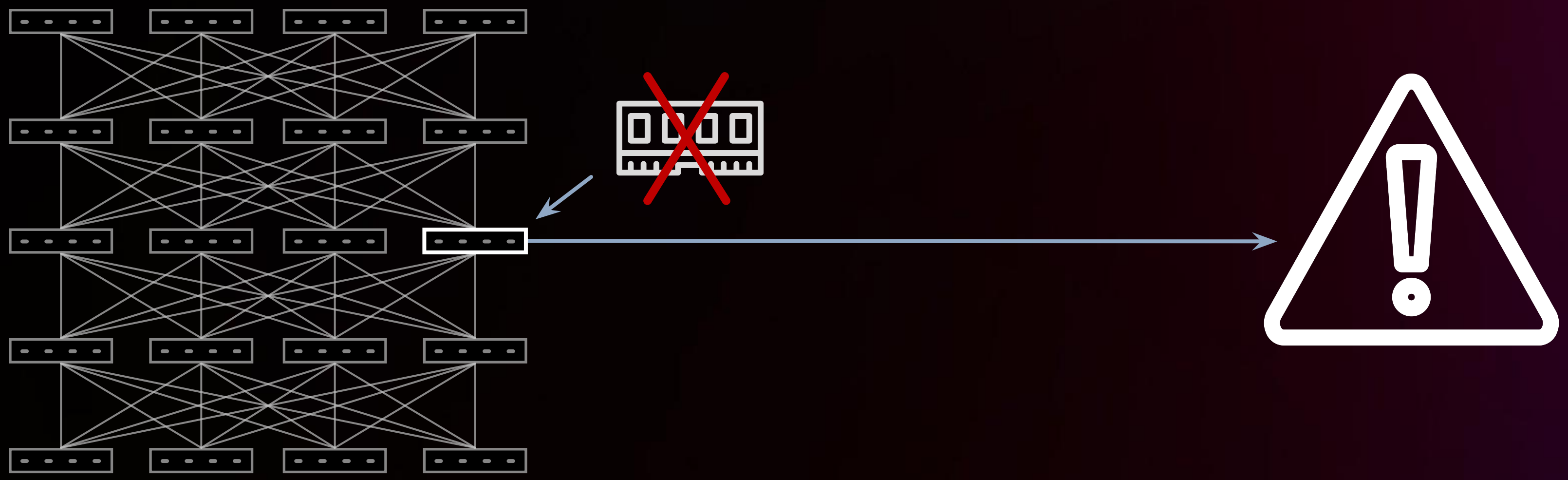
Doctor, why does it hurt?

ACTIVE MONITORING



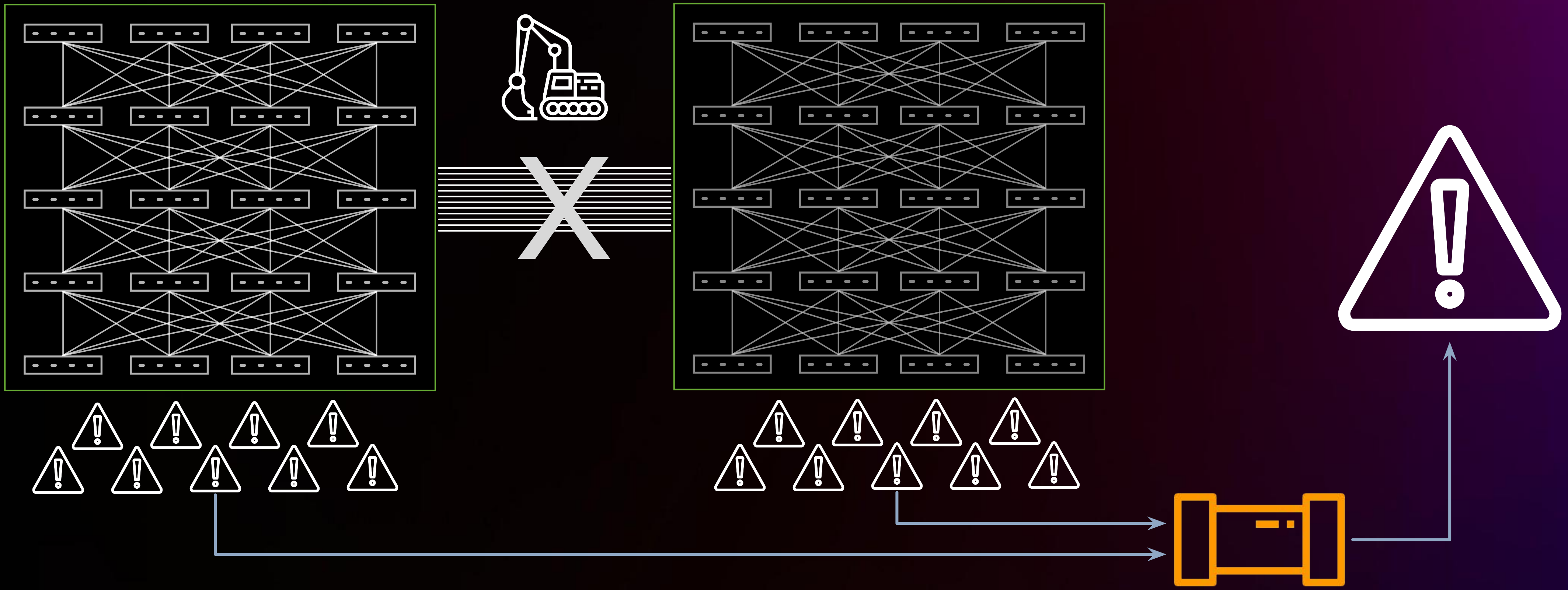
Doctor, why does it hurt?

CLEAR SIGNAL



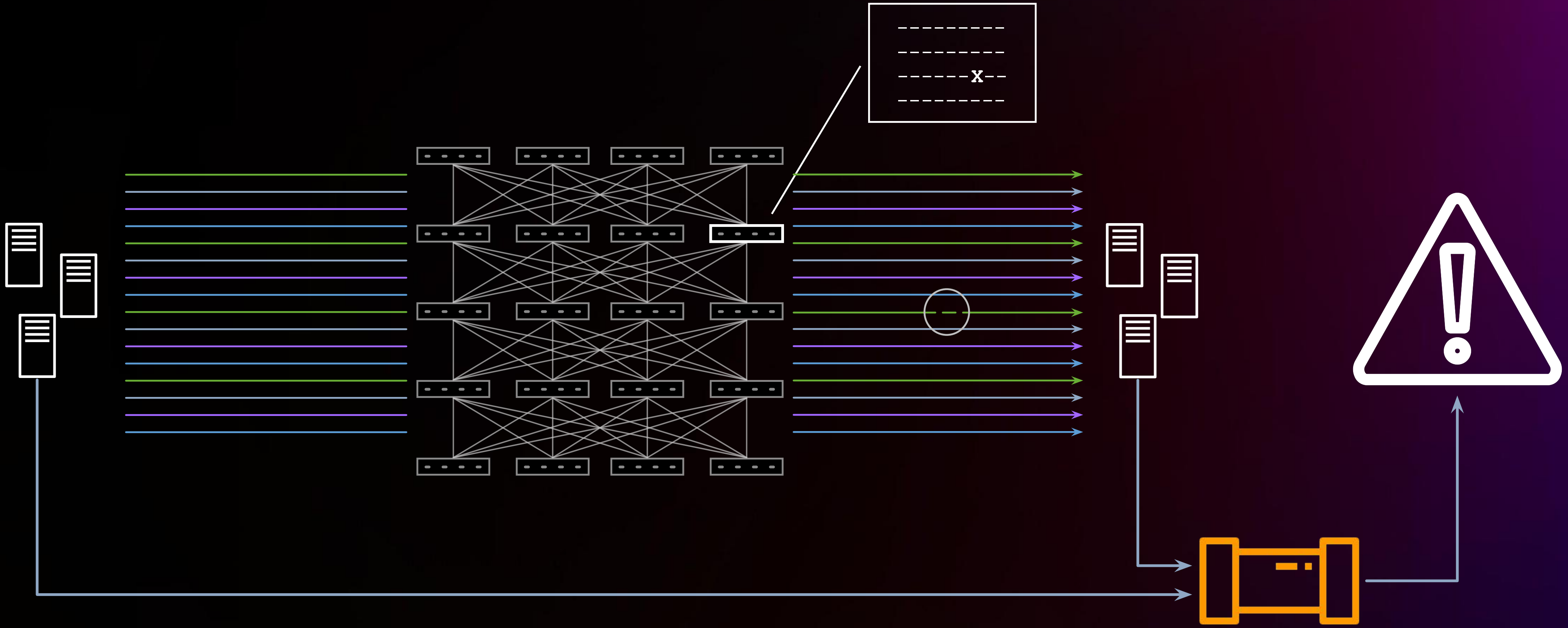
Doctor, why does it hurt?

CORRELATION



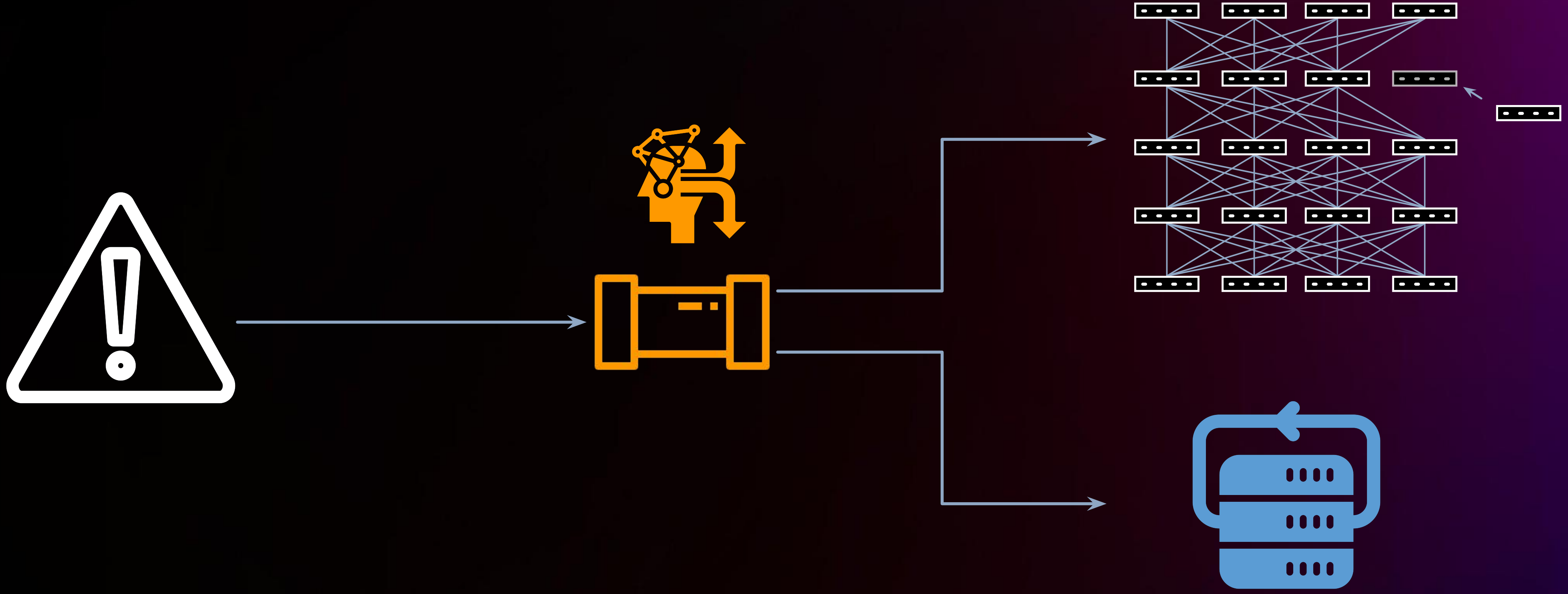
Doctor, why does it hurt?

TRIANGULATION



Ahhh . . . That’s better

AUTO-REMEDIATION



Layered control

Local for speed

Central for optimization

Hierarchical abstractions

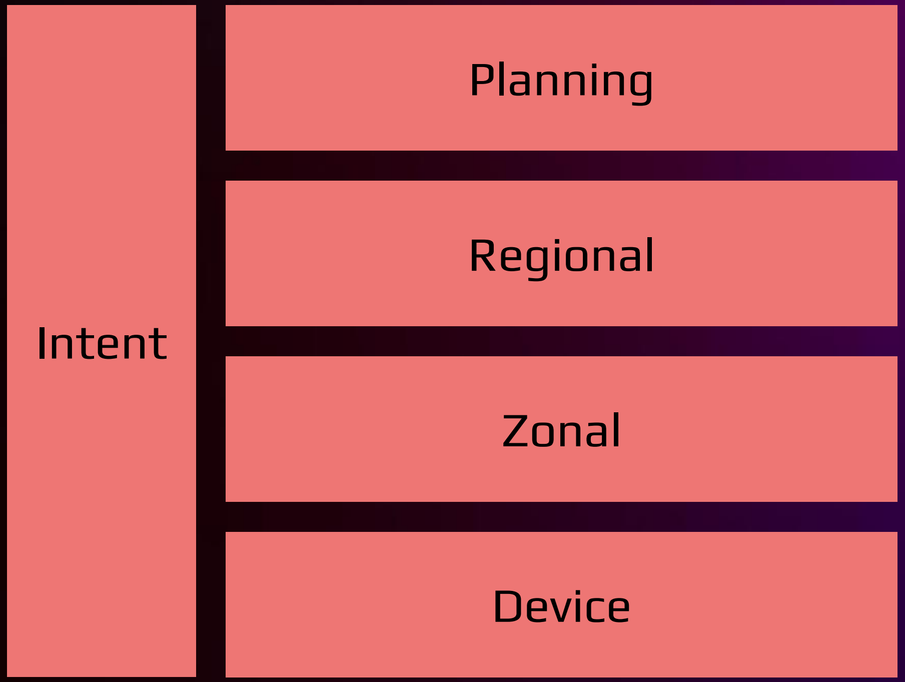
Future: Intentful management

Expected behaviors

Hierarchical

Multi-domain

Closed loop



Intent

Planning

Regional

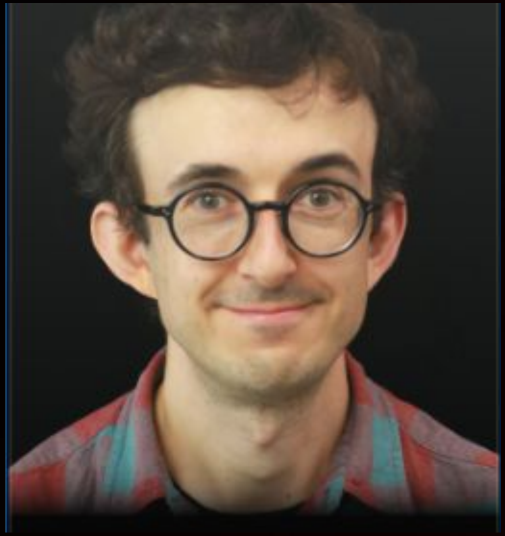
Zonal

Device

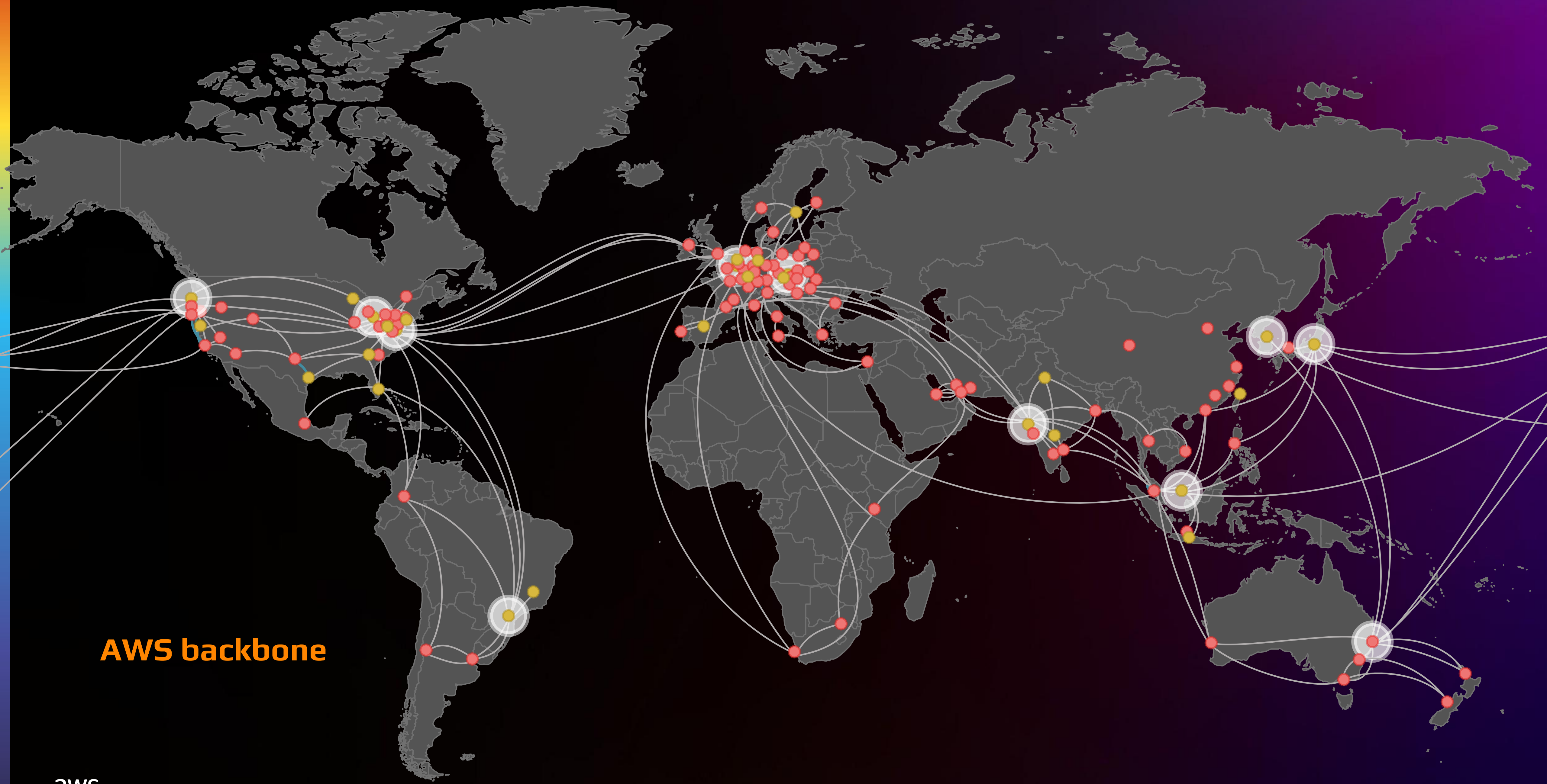
Thank you!

Colin Whittaker
colinwh@amazon.com

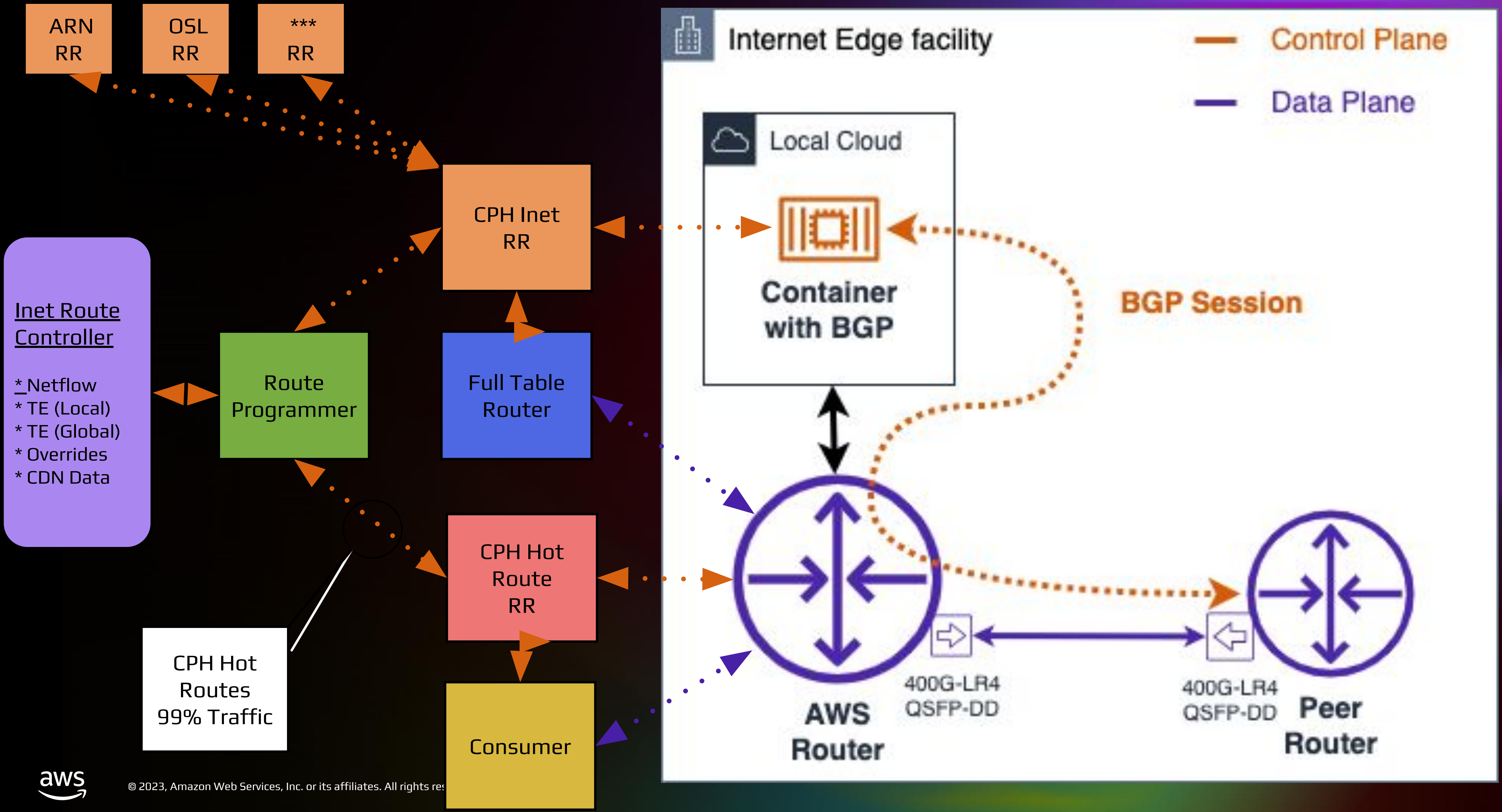
Giacomo Bernardi
giacombe@amazon.com



Giorgio Bonfiglio
bonfigg@amazon.com



AWS backbone



ARN RR

OSL RR

*** RR

CPH Inet RR

Inet Route Controller

* Netflow
* TE (Local)
* TE (Global)
* Overrides
* CDN Data

Route Programmer

Full Table Router

CPH Hot Route RR

CPH Hot Routes 99% Traffic

Consumer

Internet Edge facility