
Open. Together.
OCP
SUMMIT

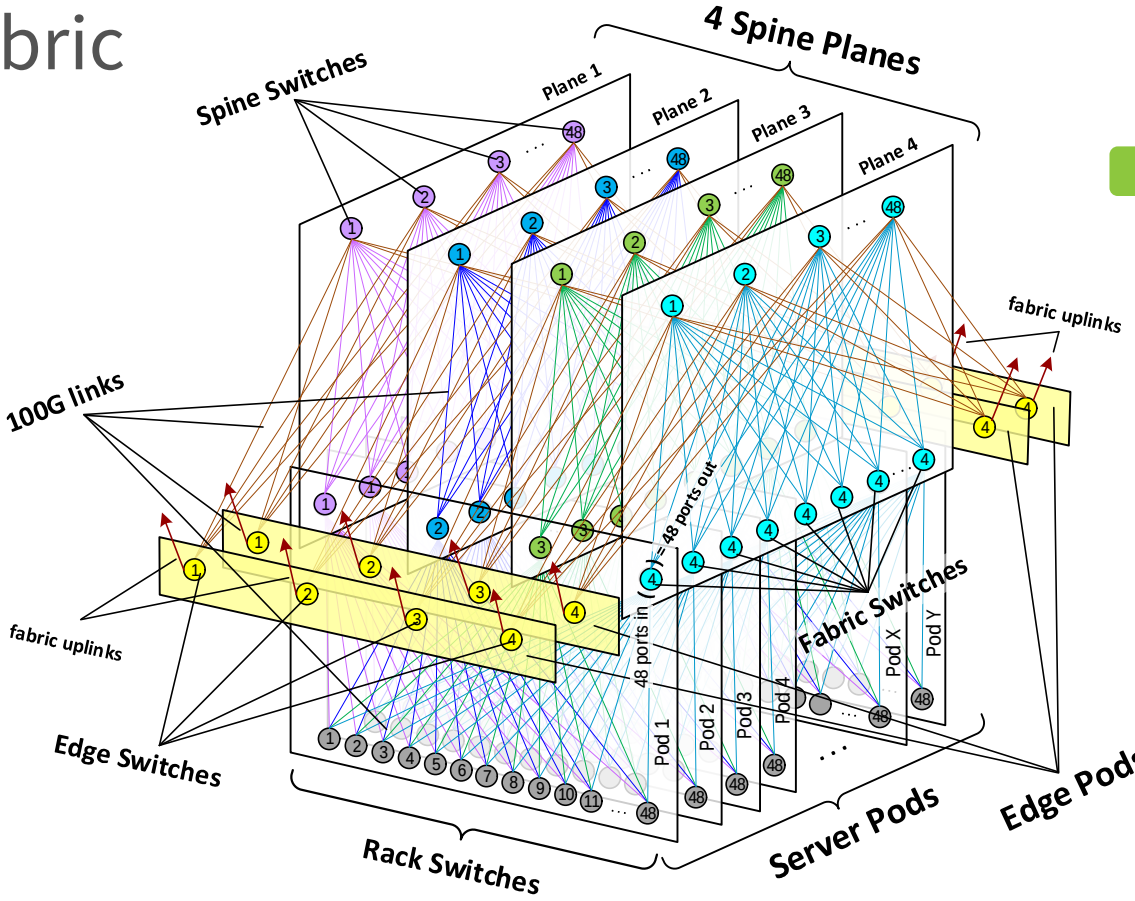
Networking

F16: the next-generation fabric

Alexey Andreyev, Network Engineer
Facebook, Inc.

Classic Facebook Fabric

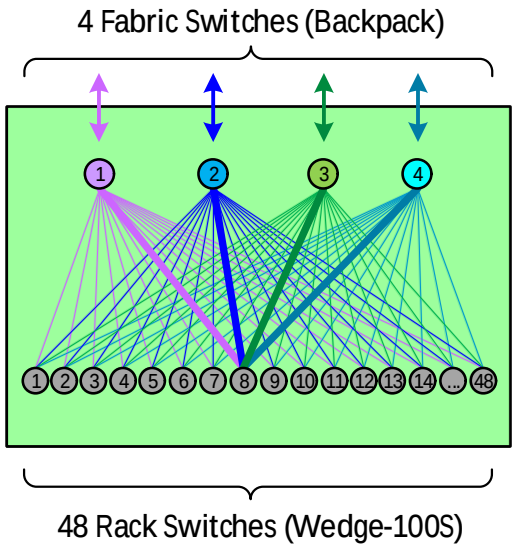
- Server Pods: racks
- 4 Parallel Spine Planes
- Edge Pods: uplinks
- Up to 1:1 Racks:Spine (non-blocking)
- Practical so far: 2:1
- Links: 100G, Fiber: SMF
- Routing: BGP

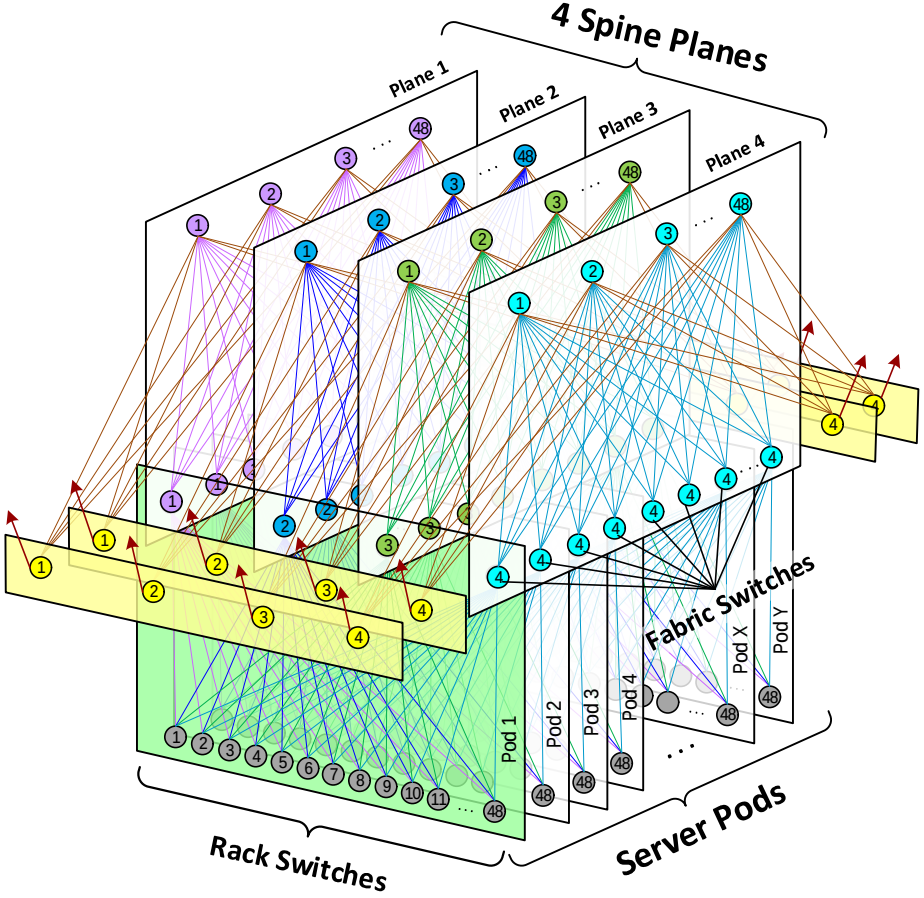


NETWORKING

Unit of Deployment: Pod

- Server Pod: 48 racks
- 4 x 100G per rack (400G)

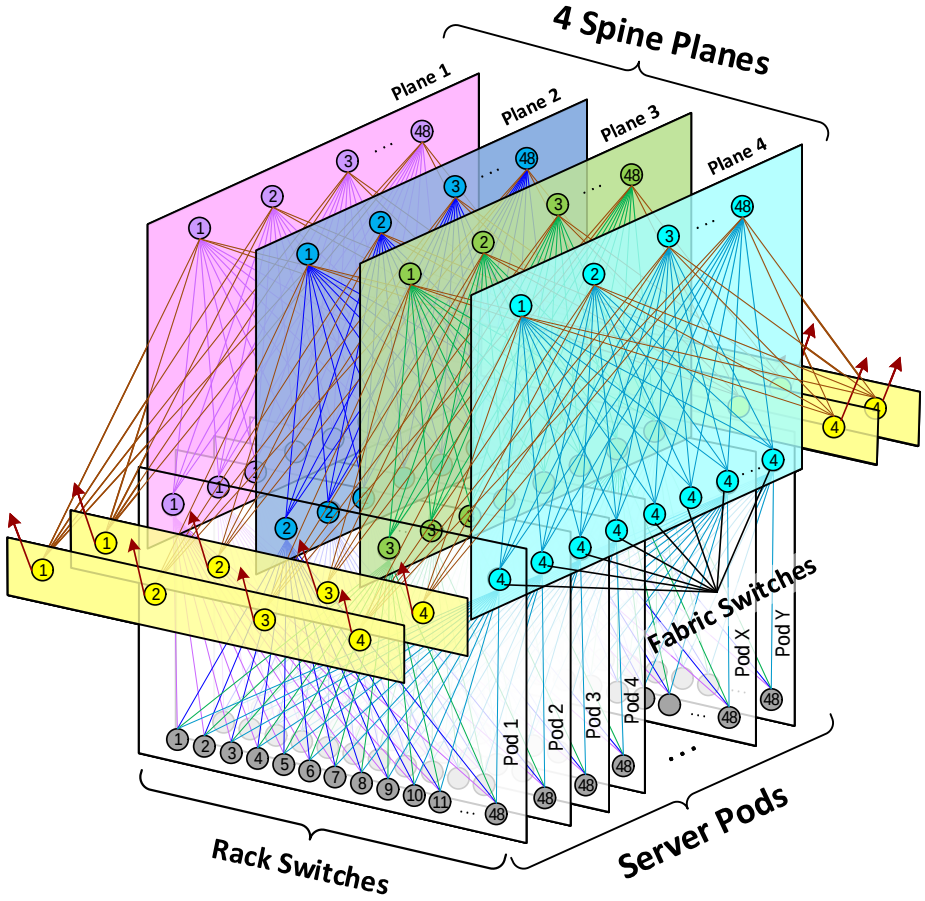




NETWORKING

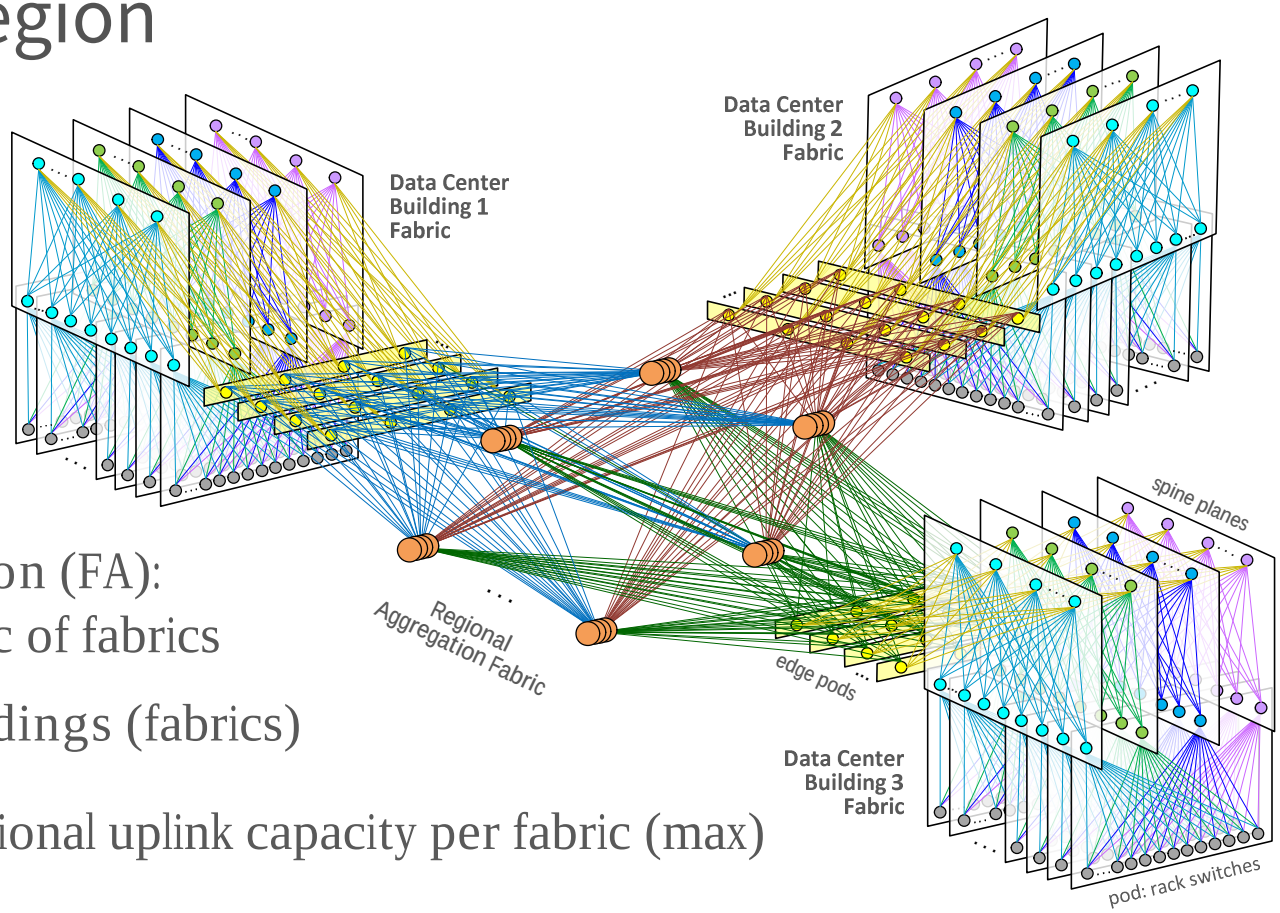
Fabric Spine Planes

- Scalability – without large boxes
- Capacity – load balanced between and within the planes
- Reliability – contained failure domains and large-scale ops
- Flexibility – independent planes



NETWORKING

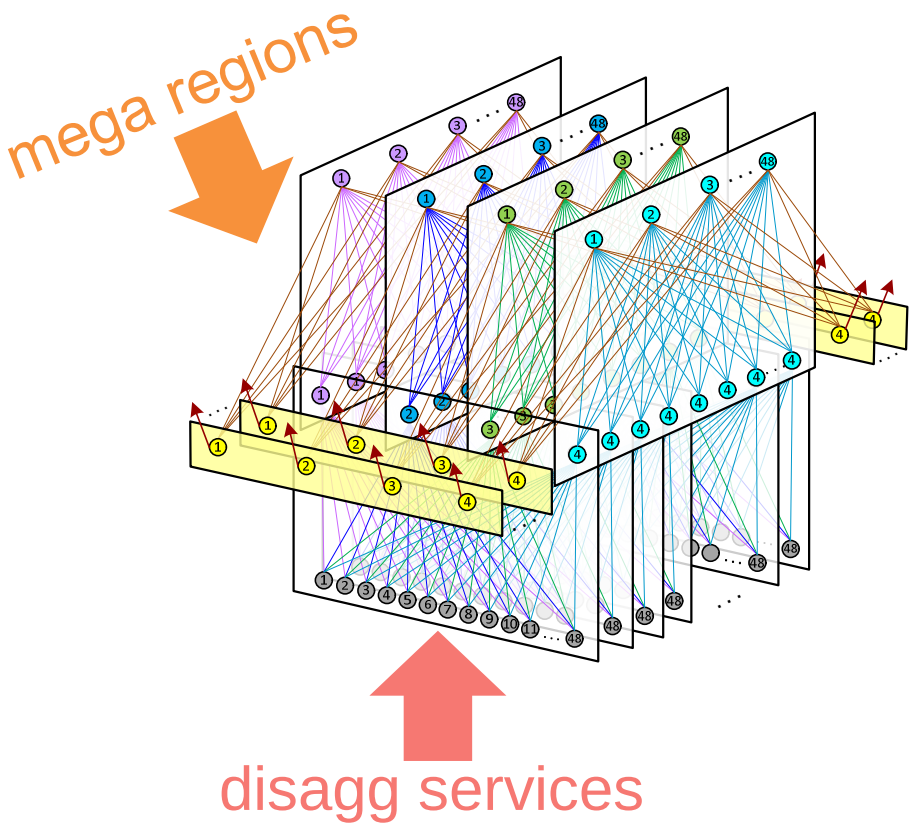
Data Center Region



- Fabric Aggregation (FA): inter-building fabric of fabrics
- Up to 3 large buildings (fabrics)
- 100Ts level of regional uplink capacity per fabric (max)

NETWORKING

Growing pressures



NETWORKING

- Expanding Mega Regions (5-6 buildings) = accelerated fabric-to-fabric East-West demand
- Compute-Storage and AI disaggregation requires near-Terabit capacity per Rack
- Both require larger fabric Spine capacity (by 2-4x) …

DC network – a system with many parameters

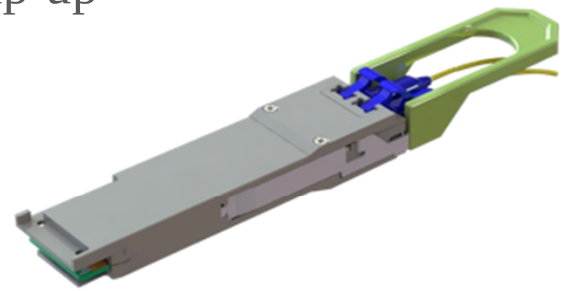
NETWORKING

- Bandwidth capacity
- Scale and scalability
- Topology and routing
- Regional composition
- Lifecycle: deployment and retrofits
- Automation and management
- Servers and Services
- Switch ASICs
- Optics and link speeds
- Power and cooling
- Fiber infrastructure
- Physical space
- Timelines: need by 2019, deployed vs. technology availability and development

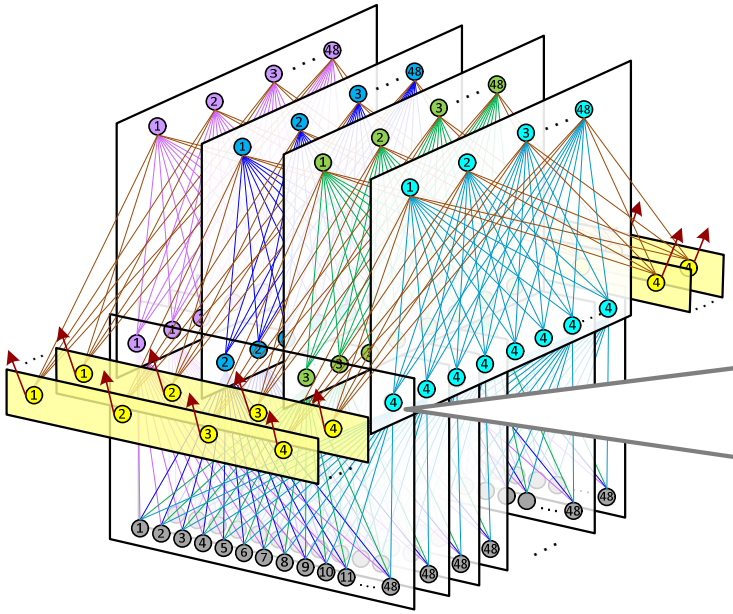
Optics

NETWORKING

- Concerns: 400G availability @ scale
- We start large – no time for new tech to ramp-up
- Risky dependency on bleeding-edge tech
- High cost of early adoption
- Interop for upgrade / retrofit paths
- Large-scale ISP and OSP structured fiber plants

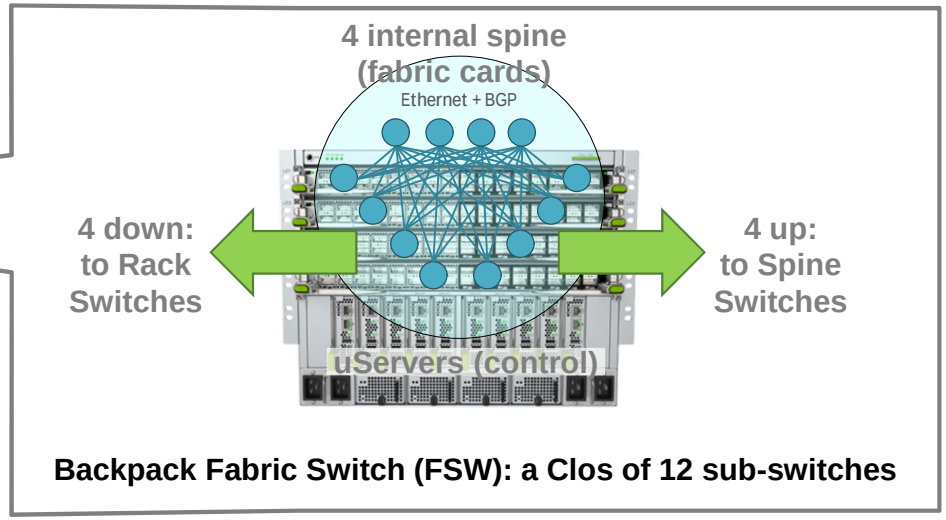


Networking power & efficiency



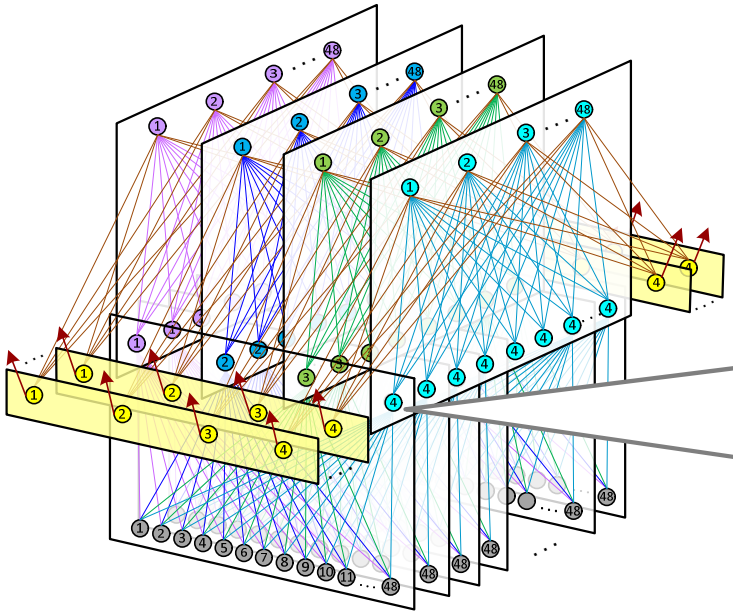
- Node radix-128 – best fit at our scale
- Achieved by building intra-node topologies from radix-32 sub-switches (ASIC+uServer)

NETWORKING



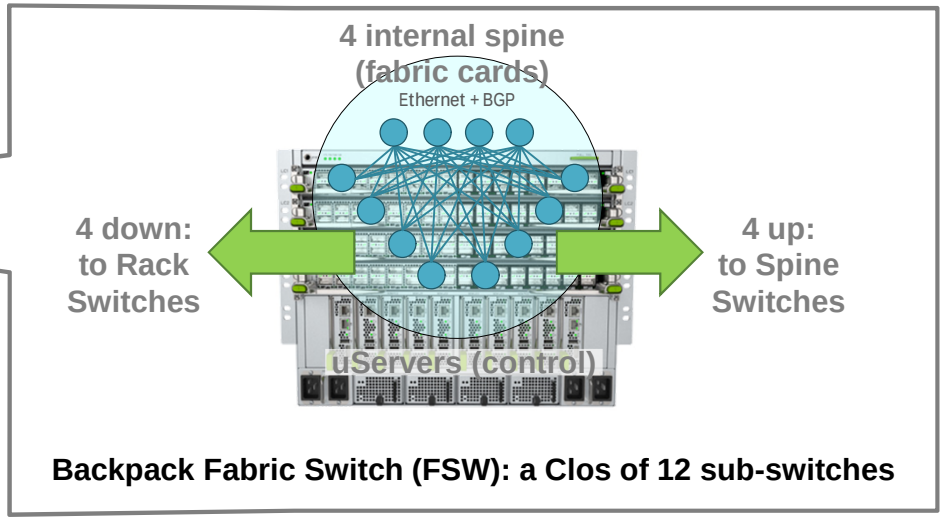
Backpack Fabric Switch (FSW): a Clos of 12 sub-switches

Networking power & efficiency



- 12 small-radix subsystems – Ok @100G
- At higher speeds + growing scale the efficiency starts declining …

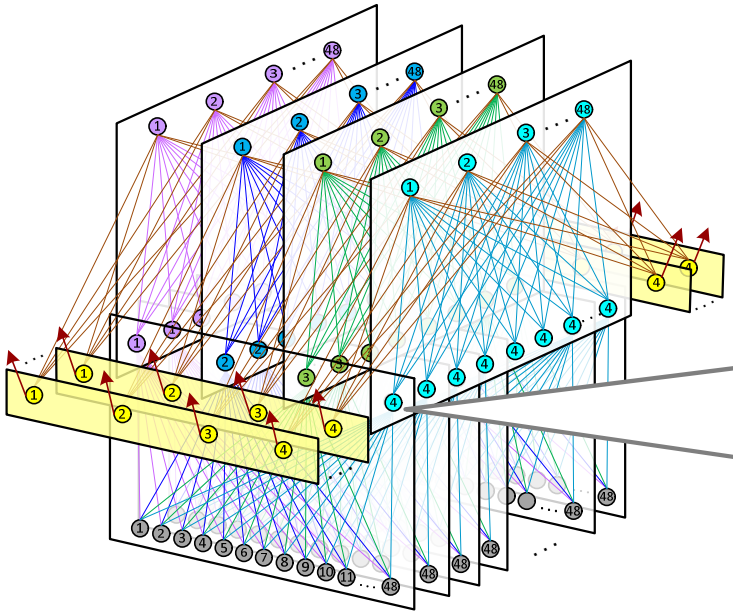
NETWORKING



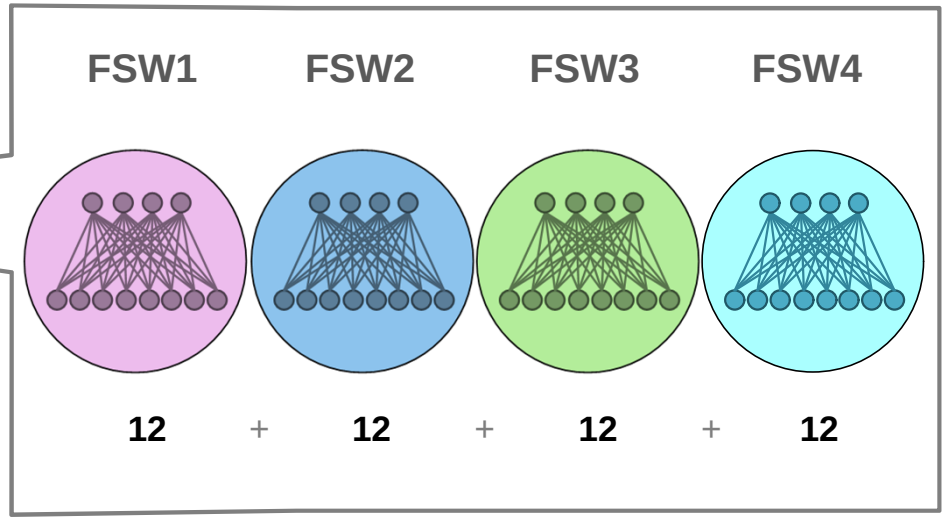
Backpack Fabric Switch (FSW): a Clos of 12 sub-switches

Networking power & efficiency

NETWORKING

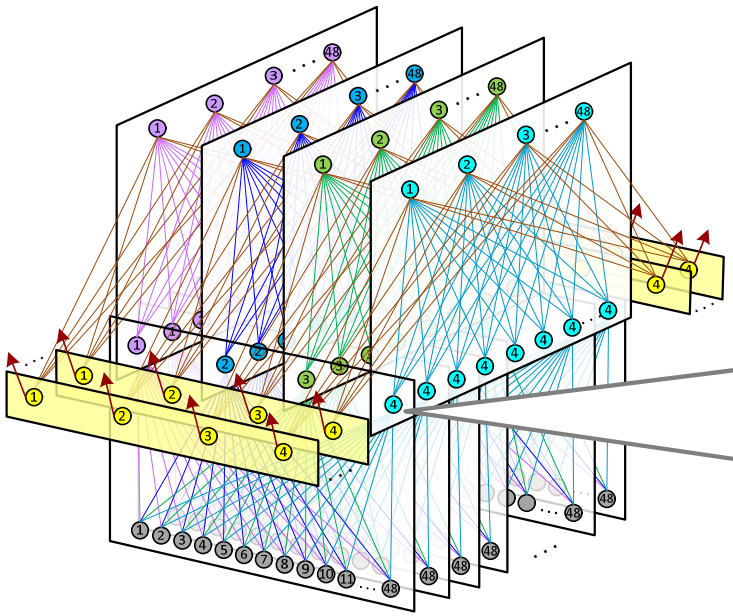


- This is 48 FSW ASICs per Pod
- Also, multi-chip Spine-tier nodes
- +Optics dependency for every next generation

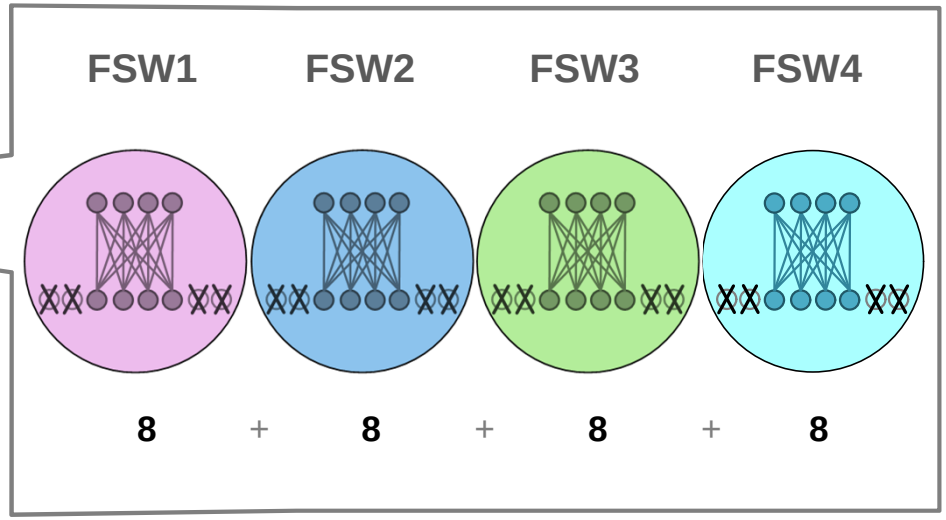


Networking power & efficiency

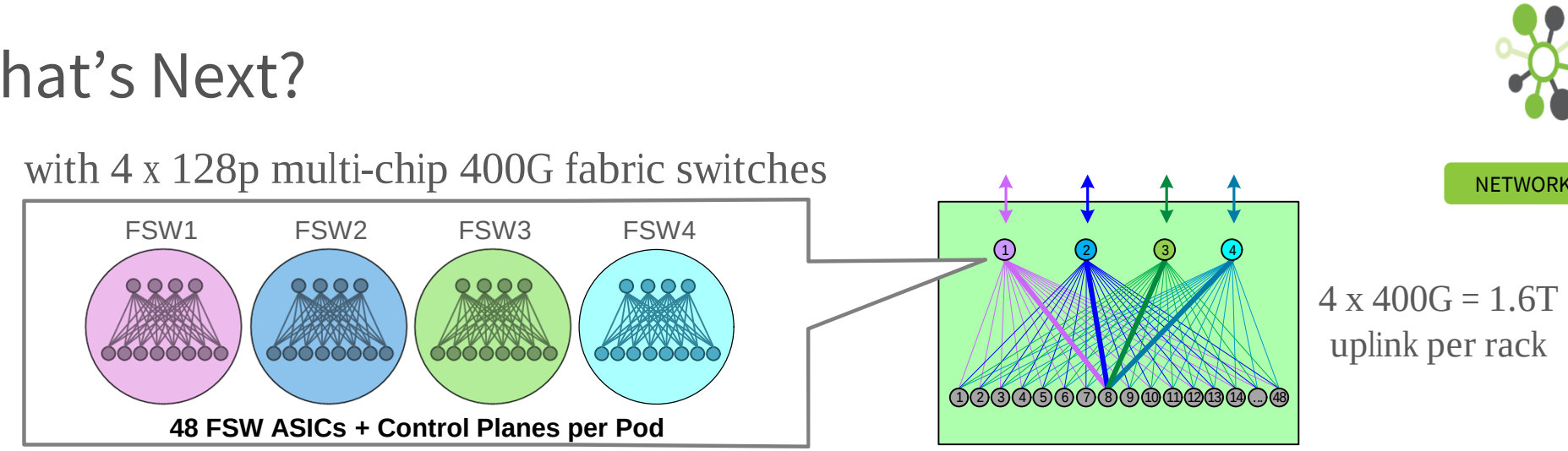
NETWORKING



- Alternative internal topologies (e.g., butterfly) – still not much better with 75% capacity protection (3+1)



What's Next?



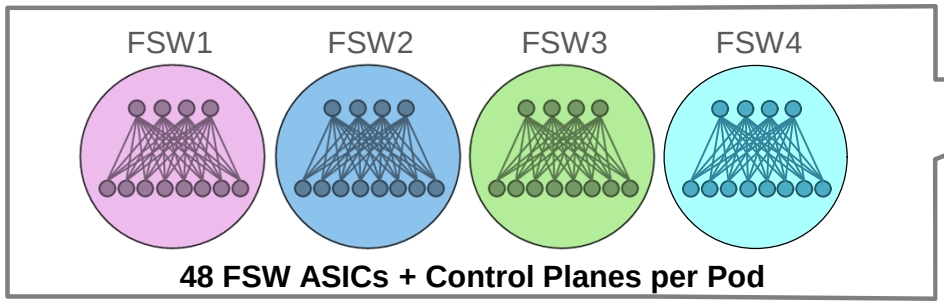
How would we achieve the next 2-4X after 1.6T?

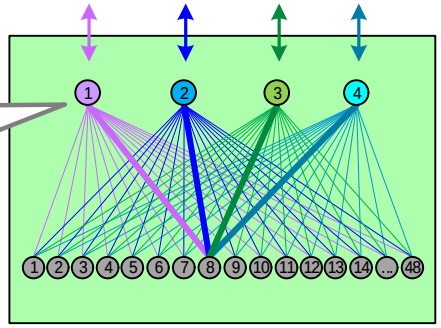
- Adding more fabric planes on multi-chip hardware = too much power...
- Increasing link speeds = would need 800G or 1600G optics in 2-3 years...

Introducing F16 fabric

NETWORKING

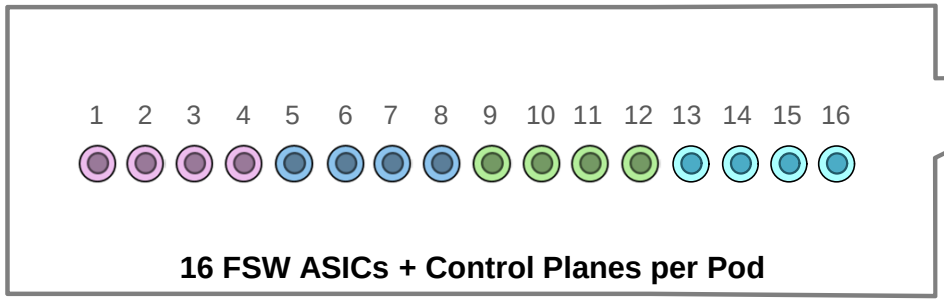
- from 4 x 128p multi-chip 400G fabric switches



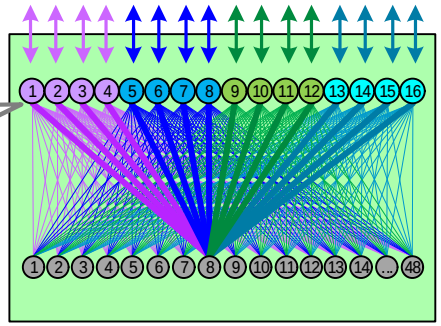


4 x 400G = 1.6T
uplink per rack

- to 16 x 128p single-chip 100G fabric switches



sample Server Pod

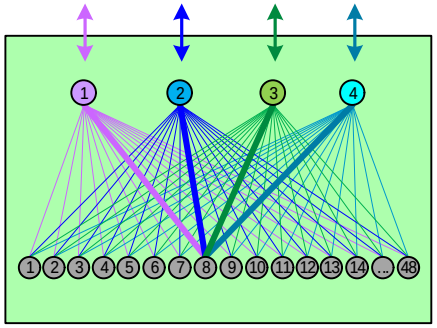


16 x 100G = 1.6T
uplink per rack

Introducing F16 fabric

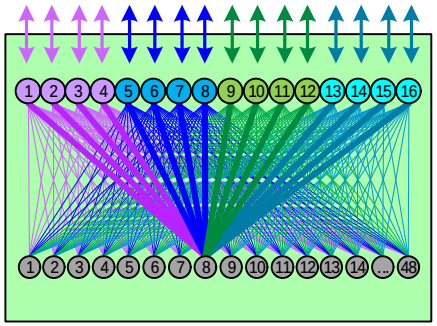
NETWORKING

- Same ASIC building block as multi-chip candidate: Broadcom Tomahawk-3
- Same rack uplink bandwidth capacity as 4 x 400G: up to 1.6T per TOR
- 3X+ less chips and control planes = TCO and Ops efficiency



4 x 400G = 1.6T uplink per rack

sample Server Pod

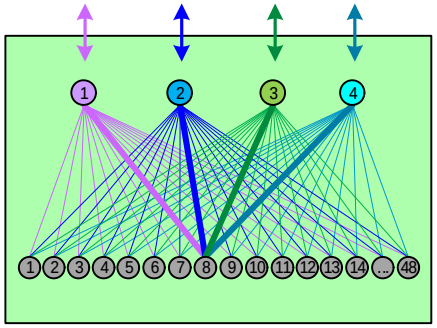


16 x 100G = 1.6T uplink per rack

Introducing F16 fabric

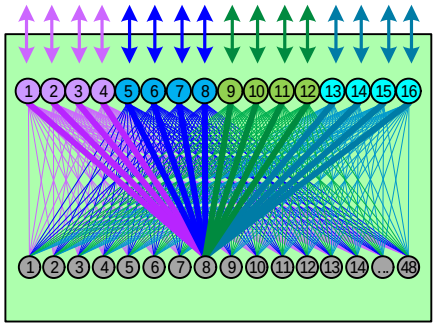
NETWORKING

- 2X+ less power/Gbps than 100G F4 fabrics
- Mature and available optics, instead of high-volume bleeding edge ramp-up: OCP 100G CWDM4
- Realistic next-steps scalability:
  - optimized for power in current and future generations
  - 200G or 400G optics as the way to achieve the next 2x or 4x



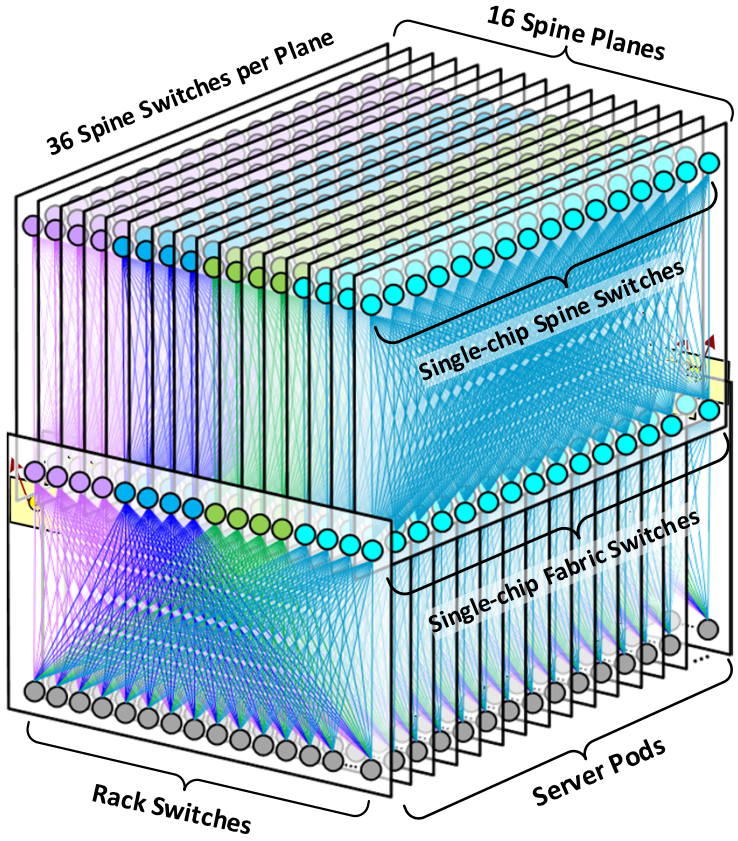
4 x 400G = 1.6T uplink per rack

sample Server Pod



16 x 100G = 1.6T uplink per rack

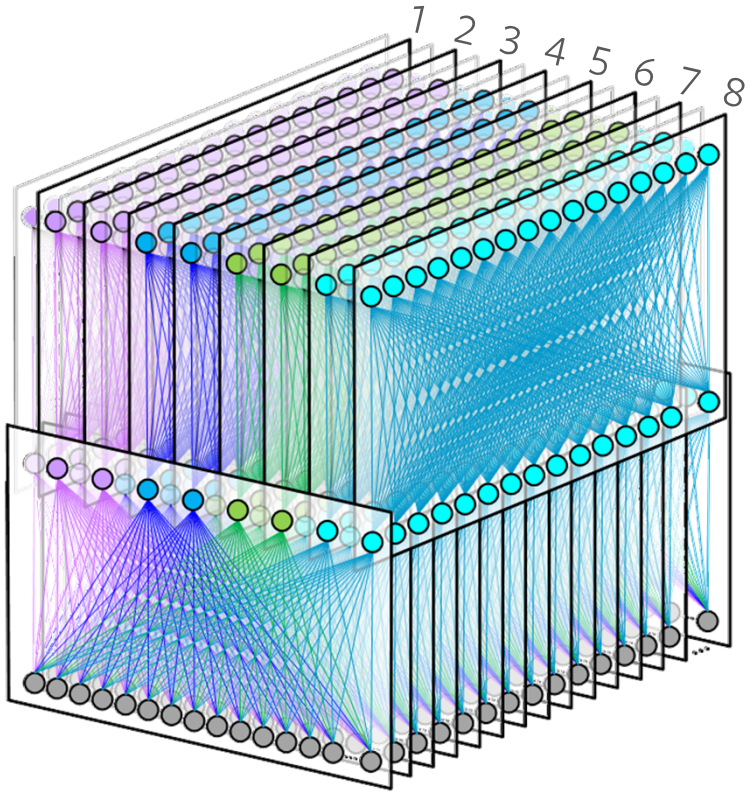
F16 fabric design



NETWORKING

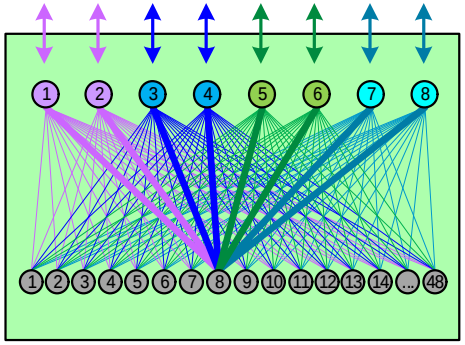
- Up to 16-plane architecture: achieving 4X capacity with 100G links
- Up to 1.6T capacity per rack
- Single-chip radix-128 building blocks
- Locked Spine scale at 1.33:1 from start (36 FSW-Spine uplinks for 48 Racks/Pod)
- No Edge Pods – replaced with direct Spine uplinks to new large-scale Disaggregated FA

F16.8P: 8-plane variant



NETWORKING

- Physical Infra and fiber designed and built for full F16
- Starting number of parallel planes: 8
- 800G capacity per rack (8 x 100G)

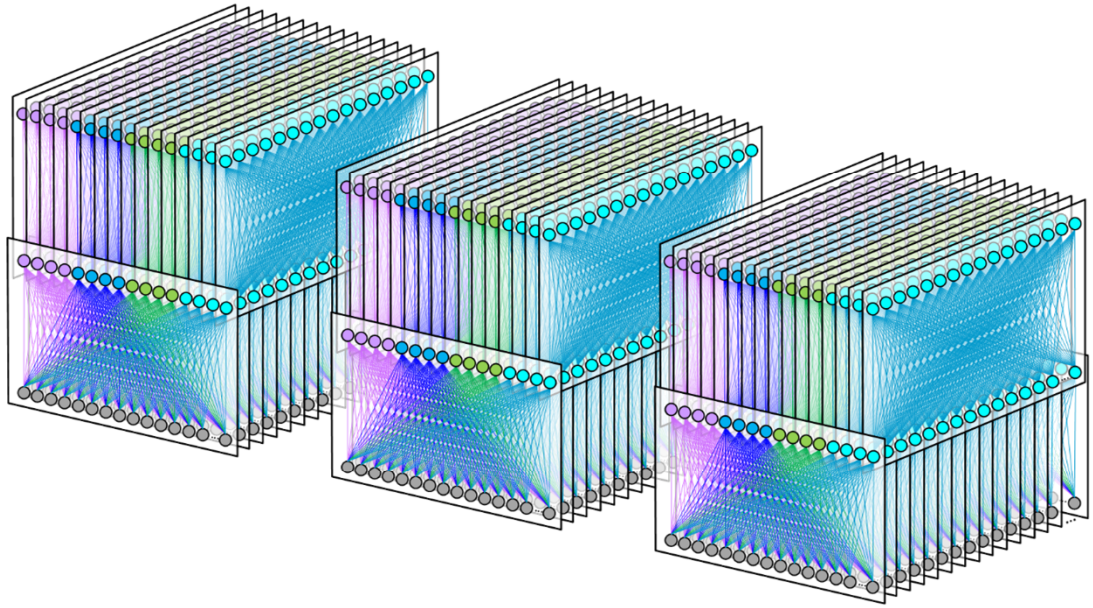


F16 region evolution: HGRID

NETWORKING

- Edge Pods → direct Spine-FA uplinks
- No device is big enough to mesh F16 fabrics – disaggregated solution required
- Goal: mega-region – beyond 3 fabrics

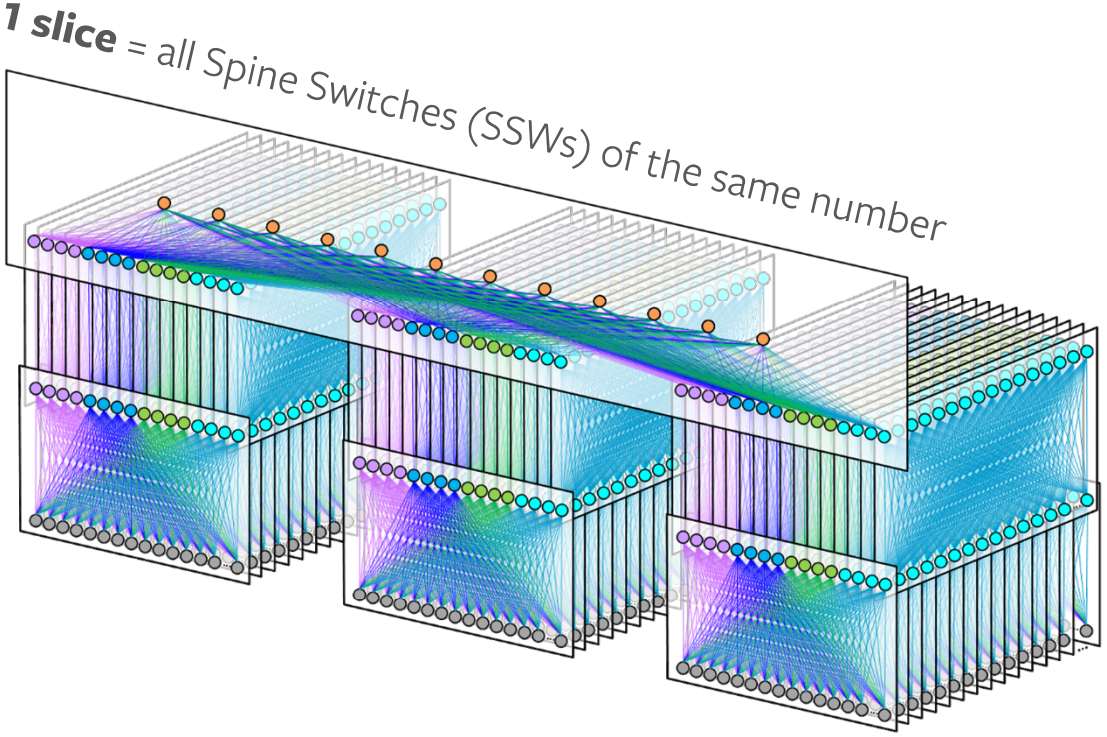
each F16 fabric =
576 Spine Switches (SSWs)



F16 region evolution: HGRID

NETWORKING

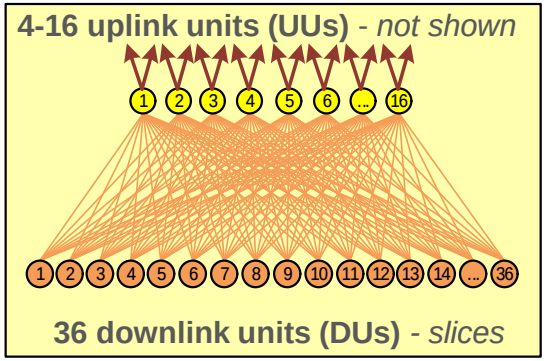
- HGRID – connecting slices of matching Spine Switches across F16s
- Partial Mesh = additional routing and reachability considerations

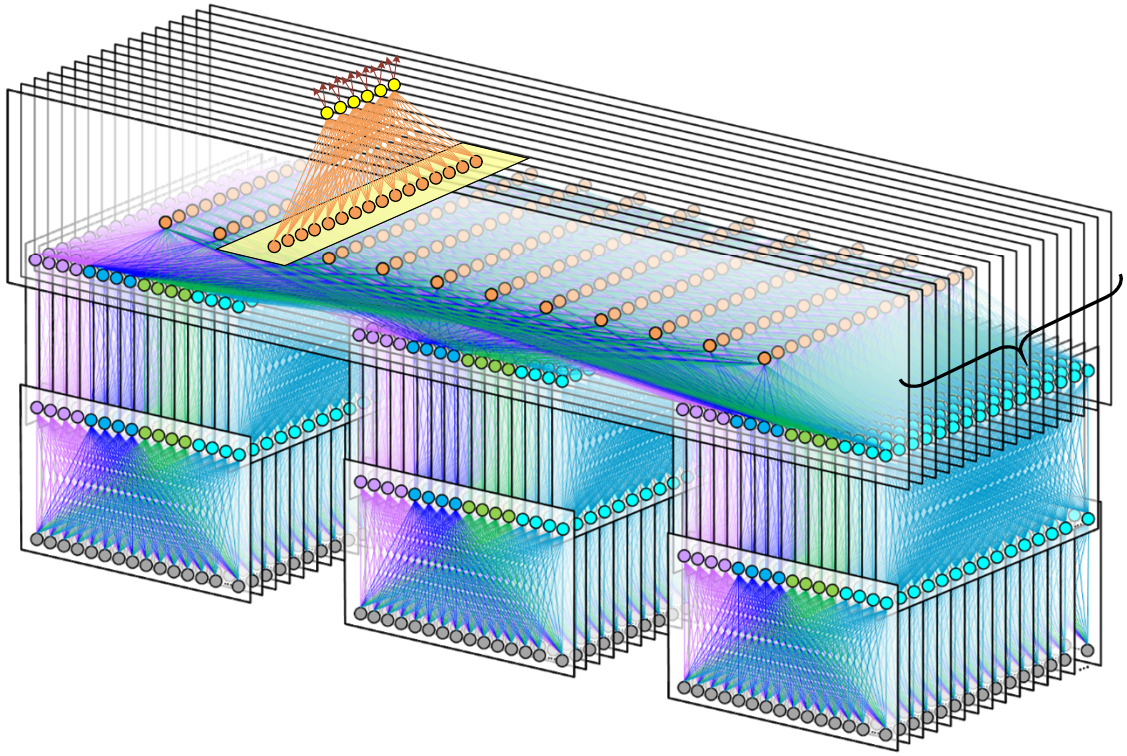


F16 region evolution: HGRID

NETWORKING

HGRID entity composition:

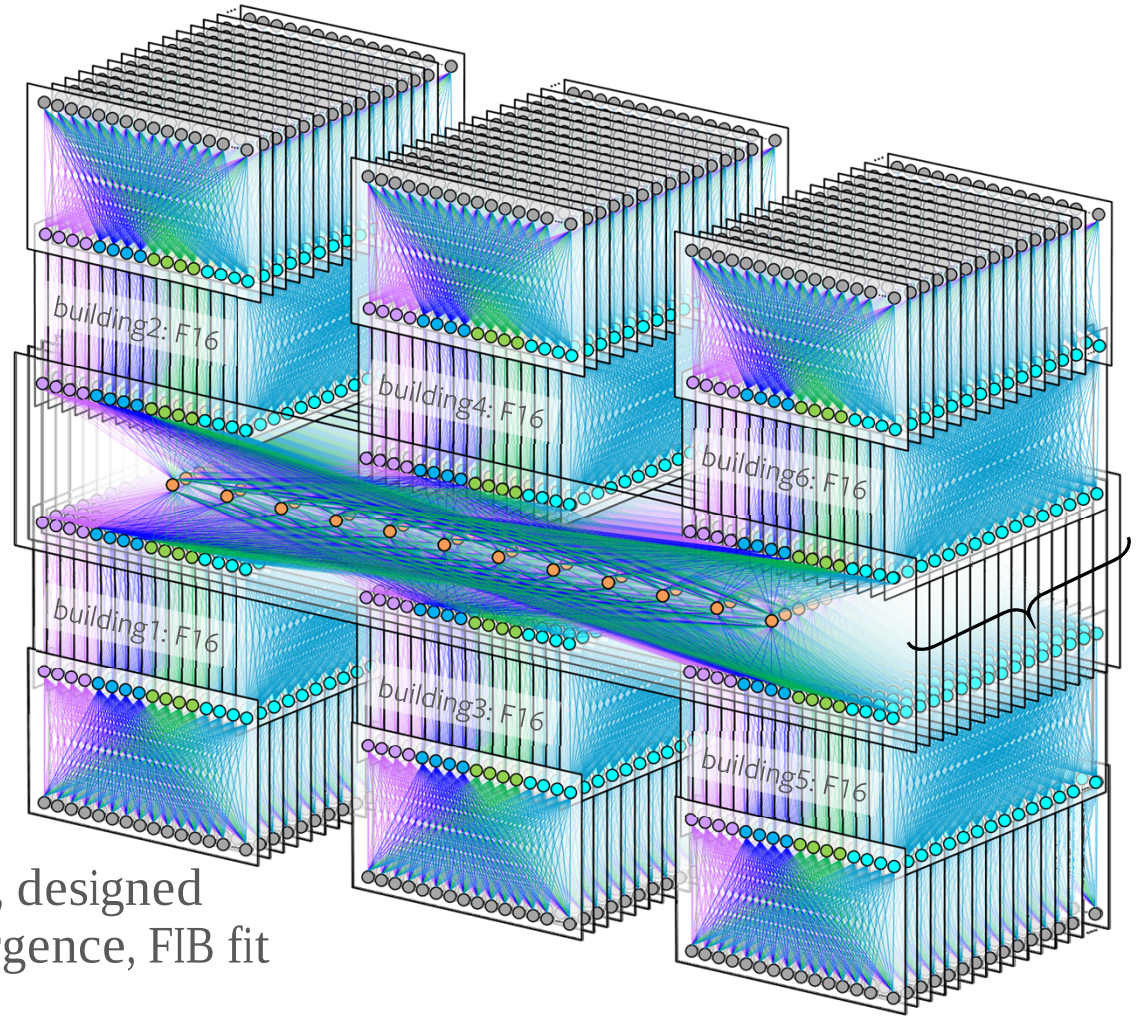




HGRID:
36-slice
Disagg-FA
architecture

F16 mega-region

- Sample 6-building region with full-size F16 fabrics
- Petabit-level regional uplink capacity, per fabric
- Evolution of our Fabric Aggregator with new building blocks
- BGP routing end-to-end, designed for reliability, fast convergence, FIB fit

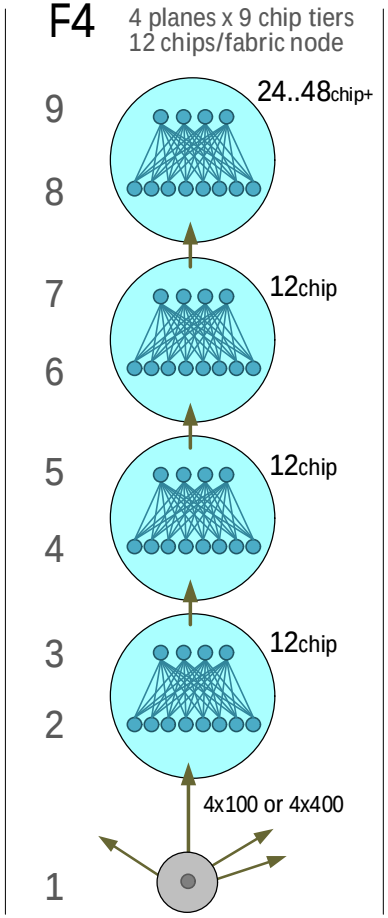


HGRID: 36-slice Disagg-FA architecture

NETWORKING

Simpler and Flatter

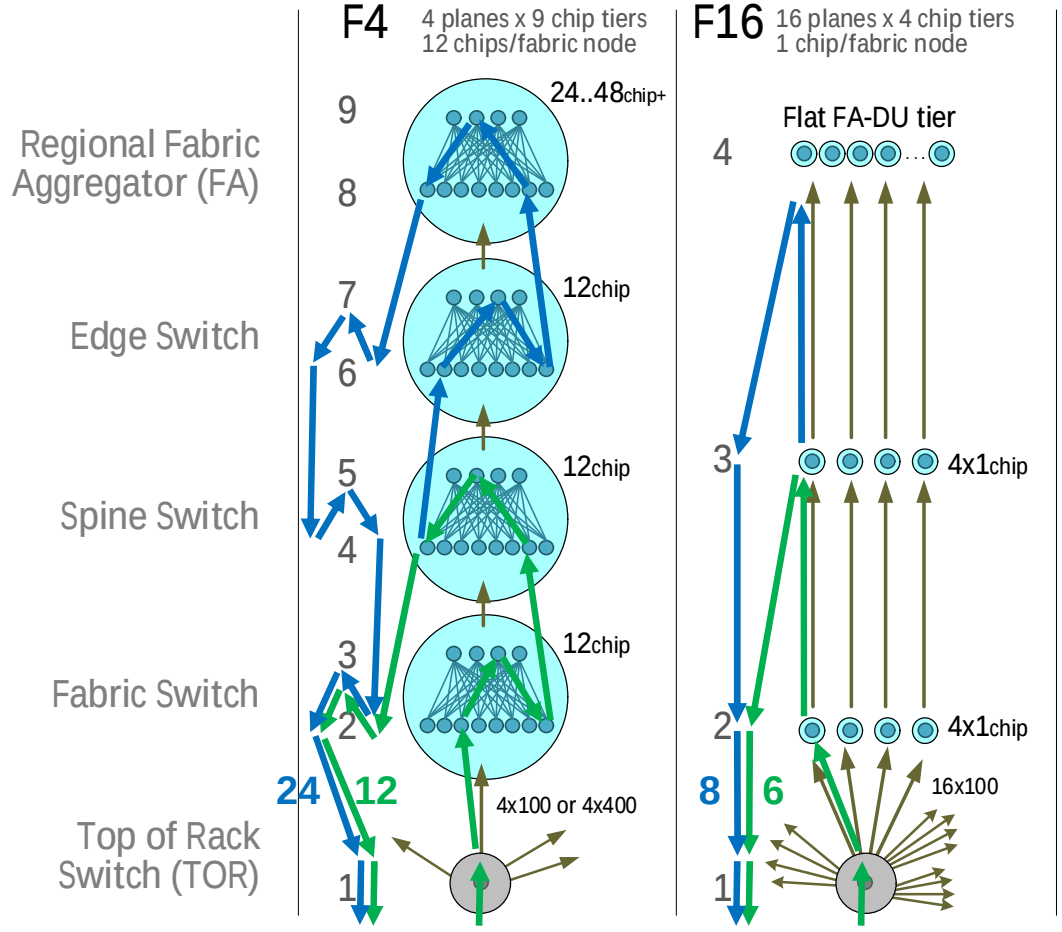
- Over 3X less switch ASICs and control planes in fabric
- 2.25X less tiers of chips in the topology



NETWORKING

Shorter paths

- Up to 2X less host-to-host network hops intra-fabric
- Up to 3X less host-to-host network hops intra-region
- More consistency, less queuing points



NETWORKING

Building blocks

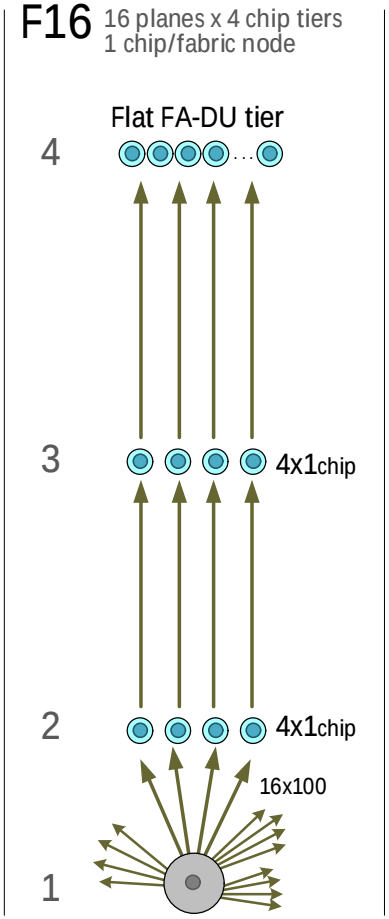
- Minipack
  128 x 100G, 4RU,
  Tomahawk-3, ~1.3kW

Single-chip,
uniform building block

Regional Fabric Aggregator (FA)

Spine Switch

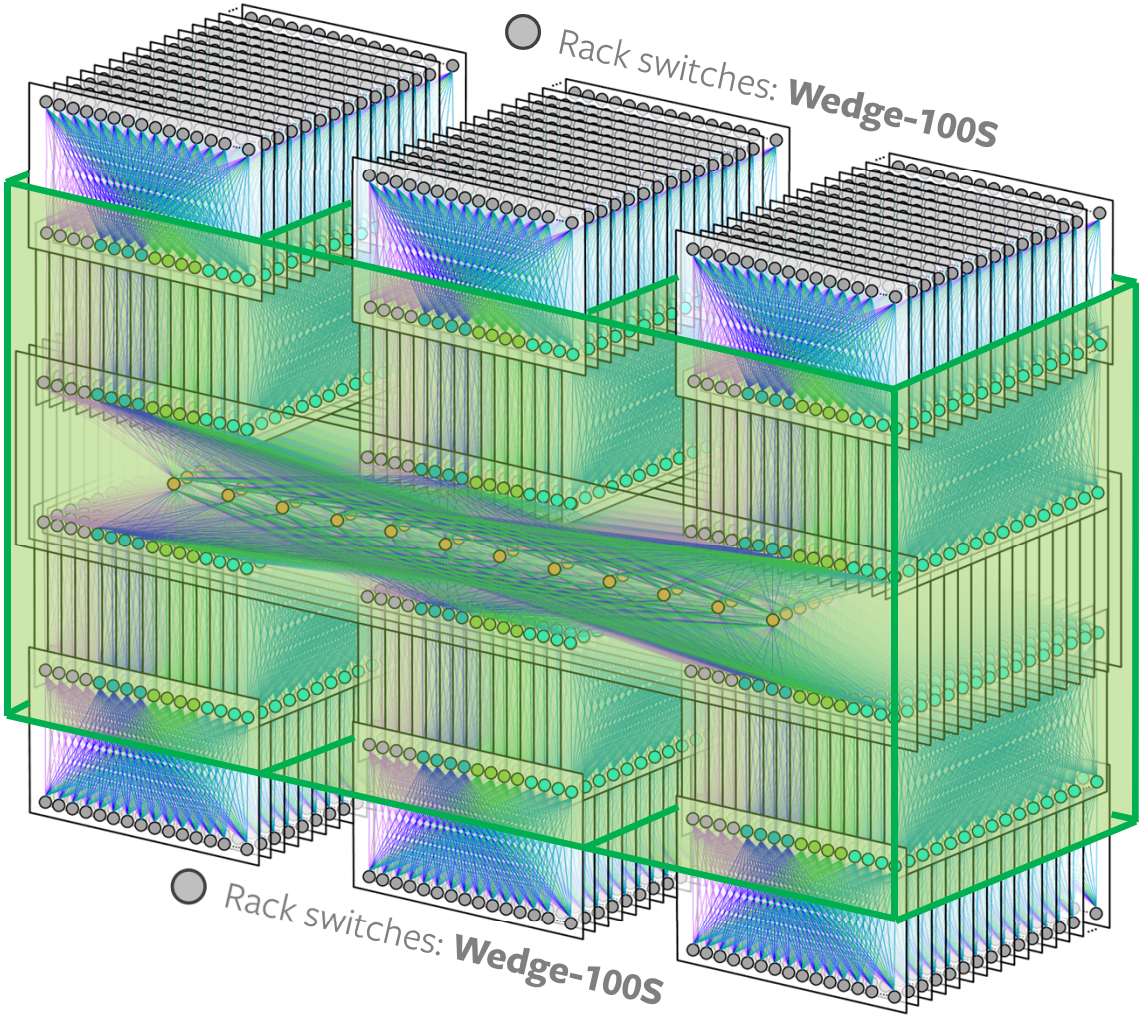
Fabric Switch



Rack switches:
Wedge-100S

NETWORKING

Building blocks



NETWORKING

- Minipack
  128 x 100G, 4RU,
  Tomahawk-3, ~1.3kW

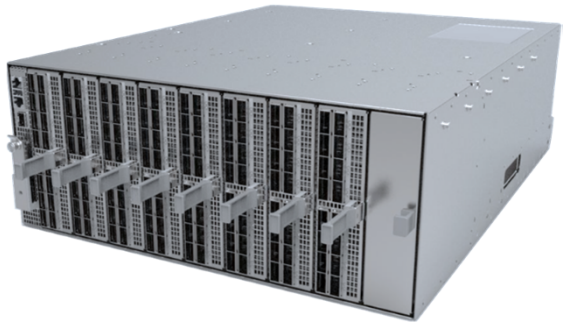
All fabric tiers and roles

Building blocks

NETWORKING

- Facebook Minipack
  FBOSS

- Arista 7368X4
  FBOSS or EOS



Single-chip,
uniform building block

modular PIMs = interface flexibility

To summarize

NETWORKING

- F16 fabric: achieving 4X bandwidth at scale, without 4X faster links
- 8 planes, 16 planes: new dimension of scaling
- 100G links: not forced to adapt next-gen optics from early day1
- Power savings: both now and in the future iterations
- Next steps: clear path to the next 2-4X – on specific tiers or all-around

To summarize

NETWORKING

- Simpler: single-chip large-radix systems improve efficiency
- Flattened: 3X+ less ASICs, 2.25+X less tiers, 2-3X less hops between servers
- Minipack: one flexible and efficient building block for all roles in fabric
- HGRID: disaggregated aggregation – scaling the multi-fabric regions in both bandwidth and size


Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT