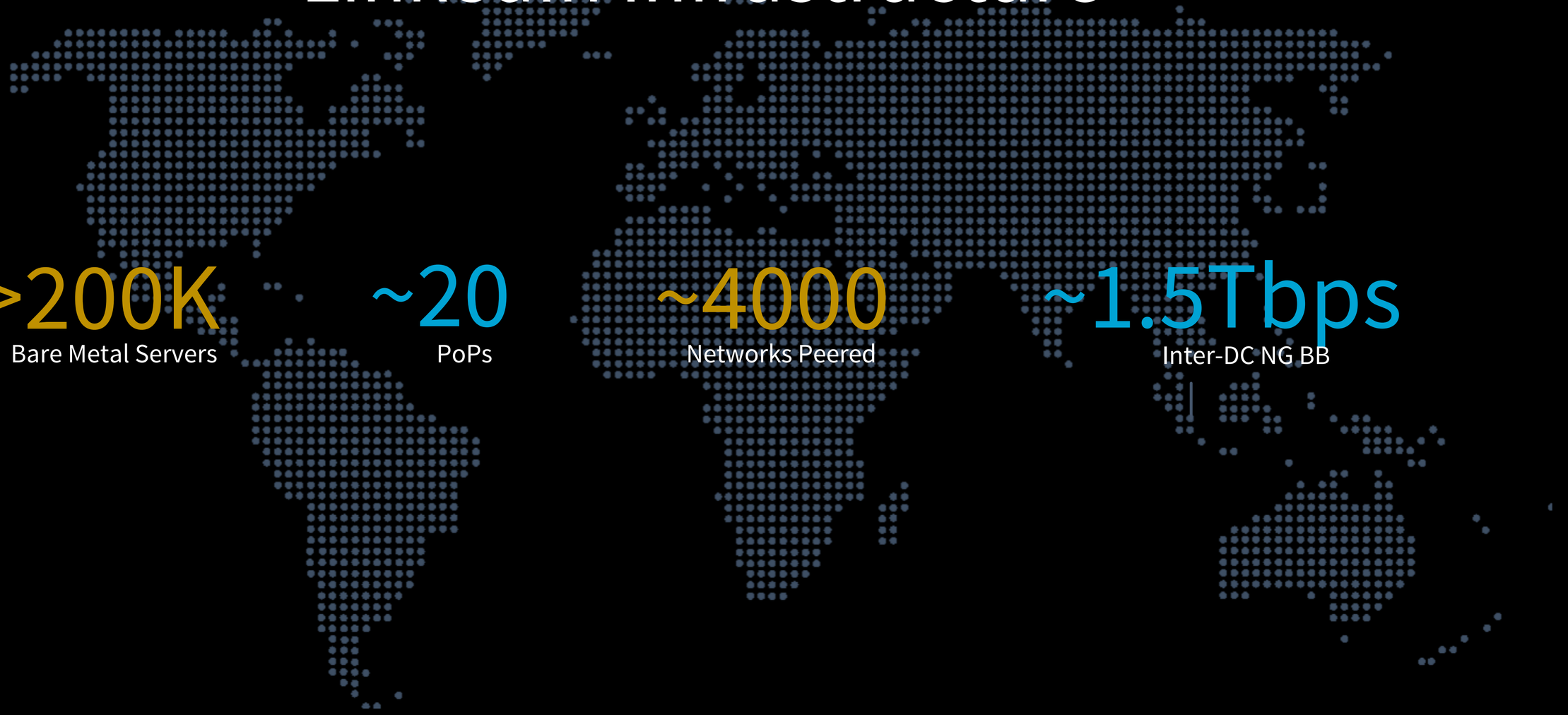
LinkedIn DC Network Architecture

(or how to build a network for 100,000 servers)

Ernesto Ovcharenko
Staff Network Engineer
Infrastructure Engineering

LinkedIn Infrastructure



>200K
Bare Metal Servers

~20
PoPs

~4000
Networks Peered

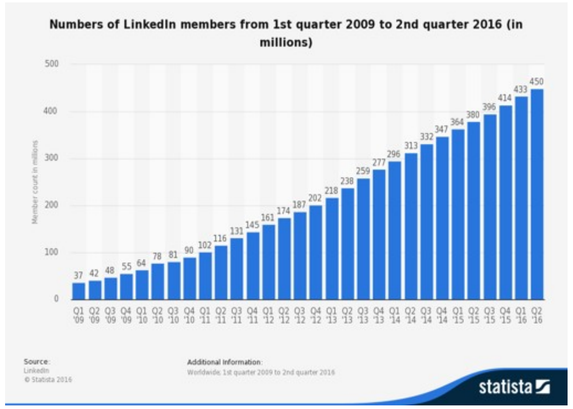
~1.5Tbps
Inter-DC NG BB

Growth

34% infrastructure growth every year…
High bandwidth & compute demand due to the organic growth.

For every single byte, thousands bytes of east-west traffic:

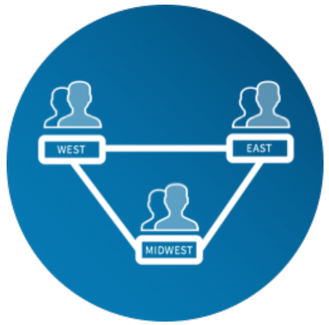
- Application Call Graph
- Kafka (metrics and analytics)
- Hadoop & Offline Compute
- Machine Learning
- Data Replication
- Search and Indexing



Capacity Crisis

Plan for 10x

Scale on Demand



Active Active Datacenters (Multi-colo)

2013-2015
Capacity Uplift

Innovate for Hyperscale

Unlimited Bandwidth

Compute on Demand

Programmable Datacenter

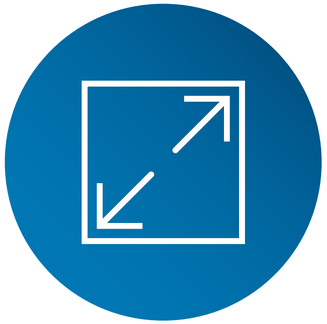
Scale Cost Effectively

2016+
Innovate for hyperscale

Own the code

Freedom and Choice

Flexibility
Customization
Modularity

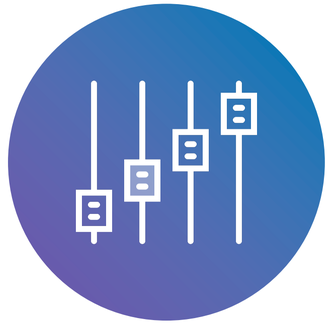


Move Fast

Growth!
Scale
Evolve
Code & Innovate

Independence

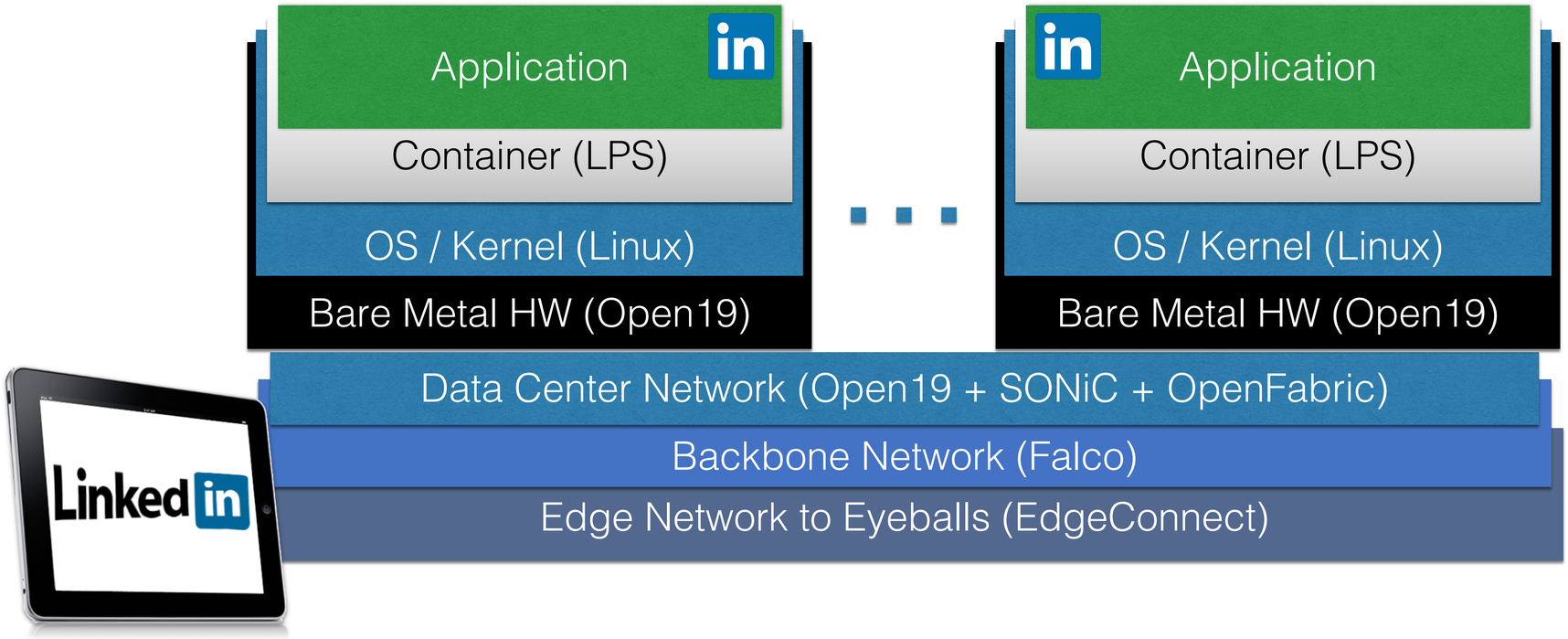
Channel
Procurement
Build Strategy
Ownership



Control

Quality
Maintenance
Risks
Security

Own the code



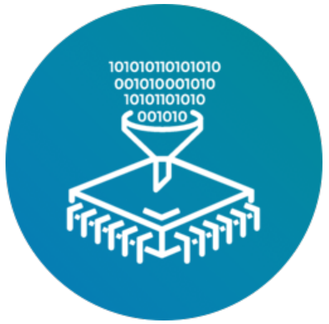
enables us to solve puzzles & complexities in different ways

On solving puzzles in a different way…

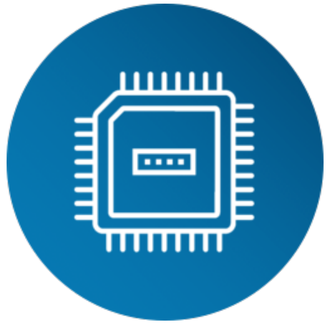
- **Load balancers**: moved to application, x86 server running a BGP daemon.
- **Firewalls**: moved to application/server.
- **NAS filers**: failover complexity moved to servers running BGP daemons, allowed for L2 to L3 network migration.

Project Altair

5-Stage
BGP Clos



Single SKU
Data Center



Single
Chip Architecture

Core Design Principles

Simple

Open

Independent

Programmable

Core Design Principles

- **Simplicity**: “perfection has been reached not when there is nothing left to add, but when there is nothing left to take away.”
- **Openness**: Use community-based tools where possible.
- **Independence**: Refuse to develop a dependence on a single vendor or vendor-driven architecture (and hence avoid the inevitable forklift upgrades)
- **Programmability**: Being able to modify the behavior of the data center fabric in near real time in software…

The Building Block: Hardware

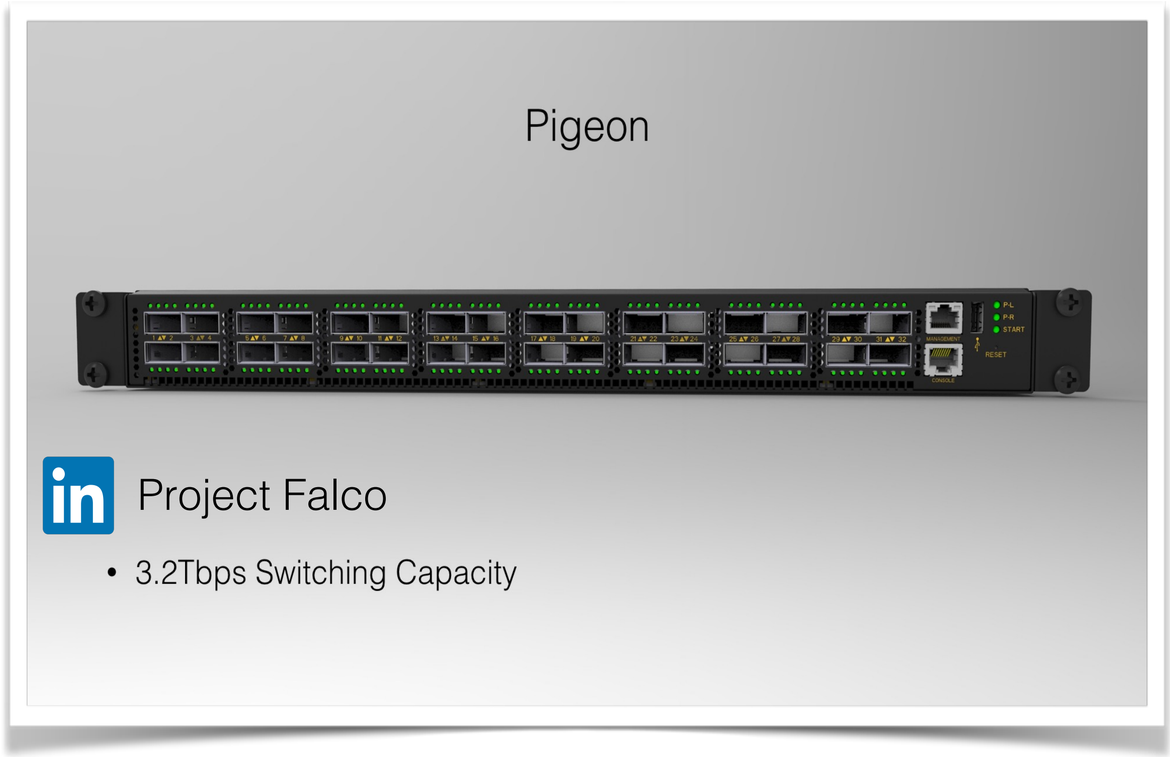
Merchant Silicon Custom Designed Switch (ODM)

No Big Chassis Switches

Designed around robustness (NSR, ISSU, etc.)

Feature-rich but mostly irrelevant to LinkedIn needs

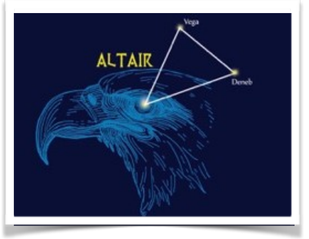
No (FCoE, VXLAN, EVPN, MCLAG, etc.)



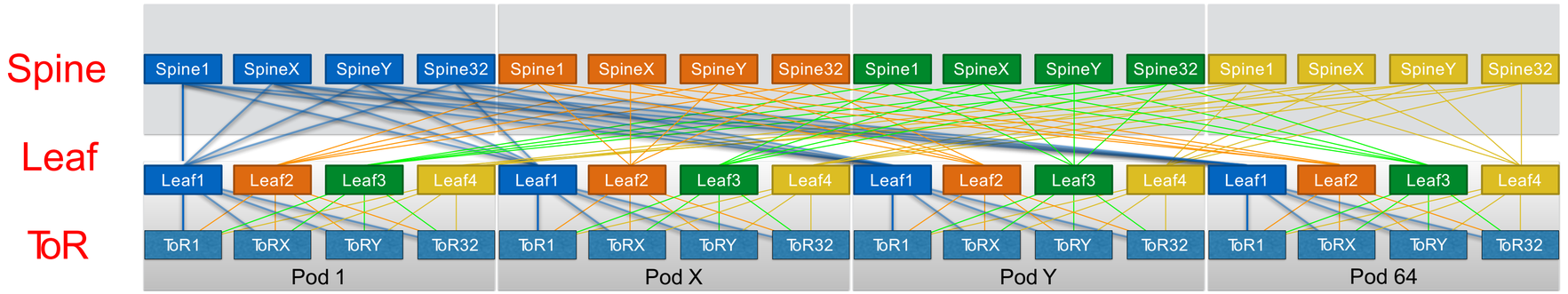
The Building Block: Software

- **Unified Architecture**: Single SKU (hardware and software) for all switches while procuring hardware from multiple ODM channels (multi-homing)
- **Minimum Features**: BGP, BFD, IPv4, IPv6, ECMP, LLDP
- **No Overlay**: For the infrastructure, the application is stateless
- **No Middle-box**: (Firewall, Load-balancer, etc.), moved to application
- Network is only a set of intermediate boxes running linux
- https://github.com/Azure/SONiC

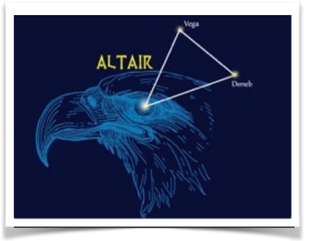
DC Architecture: Altair Design

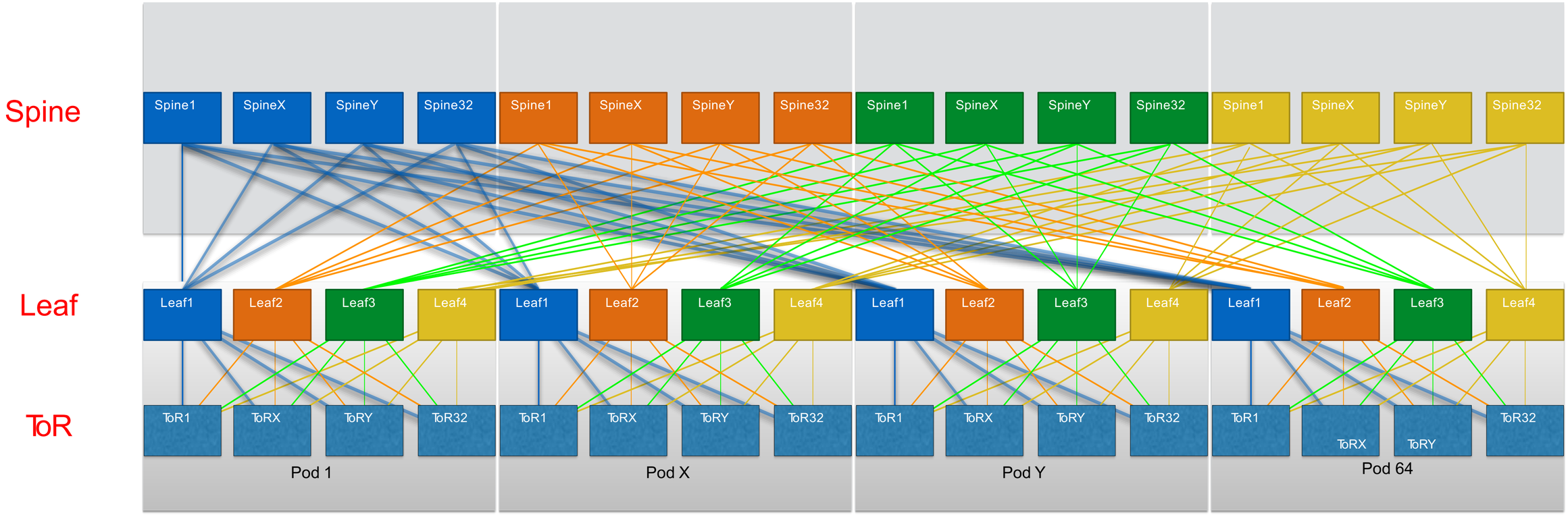


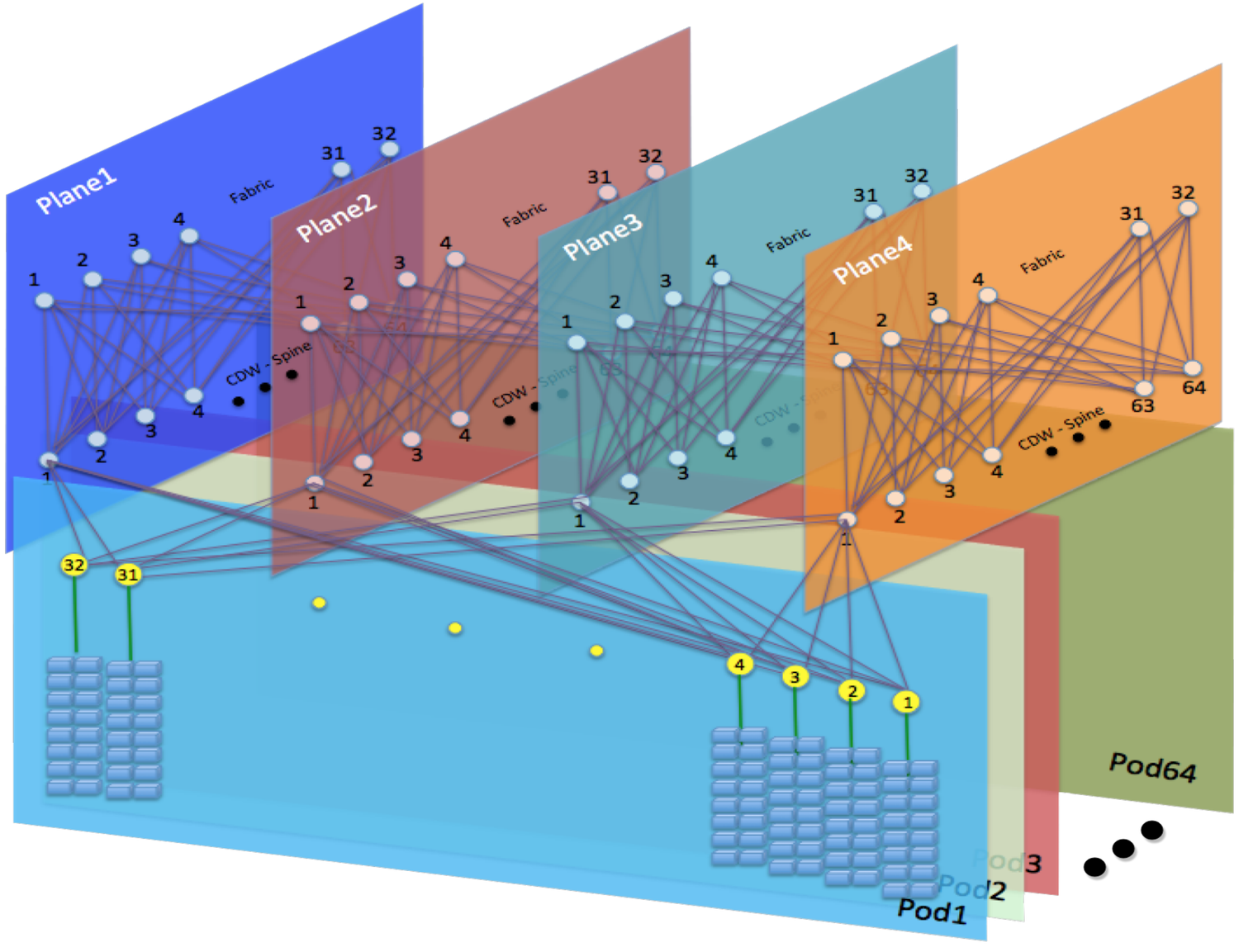
- True 5 Stage Clos Architecture (Maximum Path Length: 5 Chipsets to Minimize Latency)
- Moved complexity from big boxes to our advantage, where we can manage and control!
- Single SKU - Same Chipset - Uniform IO design (Bandwidth, Latency and Buffering)
- Dedicated control plane, OAM and CPU for each ASIC

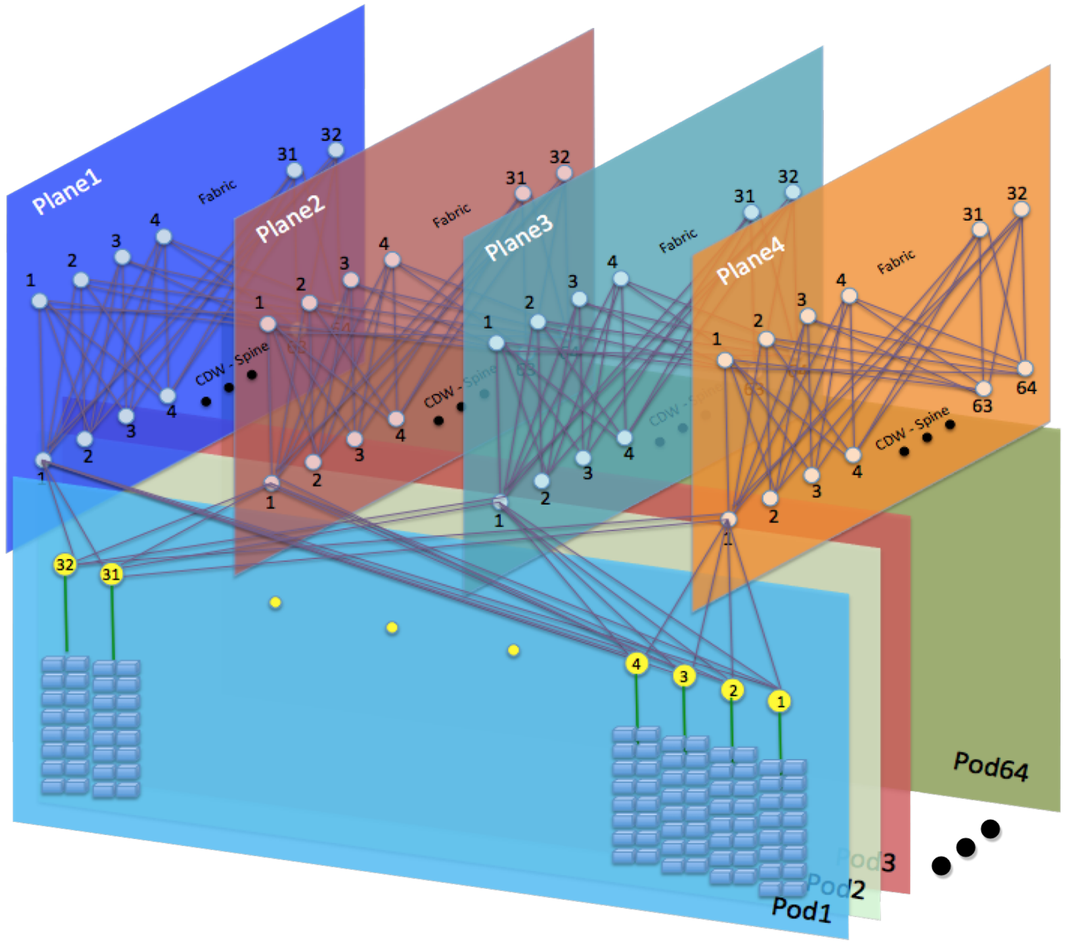


DC Architecture: Altair Design






Plane1
Plane2
Plane3
Plane4
Fabric
CDW - Spine
Pod64
Pod1



Modular and scalable growth

Efficient server deployment

Single protocol

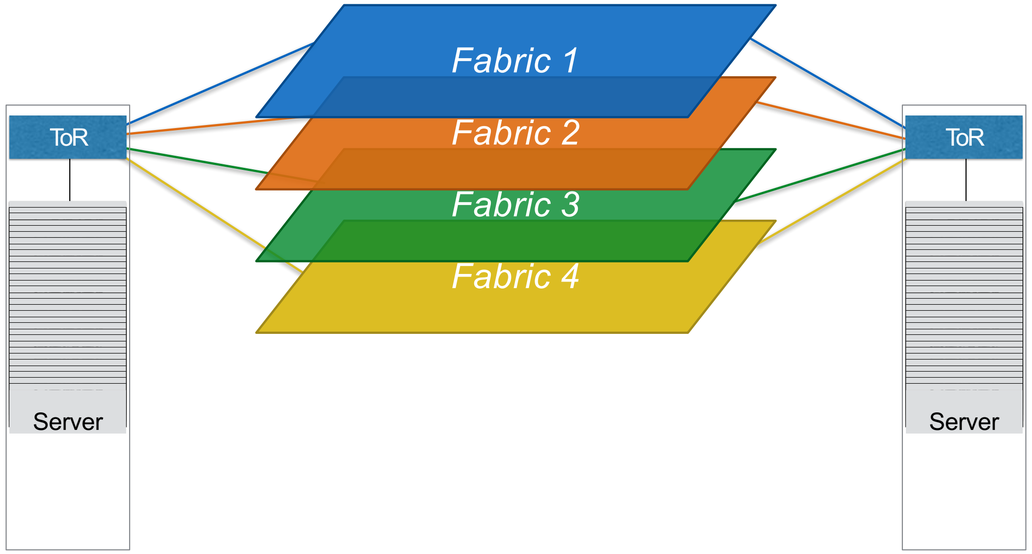
Operations friendly

Predictable performance/failure

Automation friendly

Server-server latency 2.5uS

Non-blocking Parallel Fabric



Tier 1

ToR - Top of the Rack

Broadcom Tomahawk 32x 100G

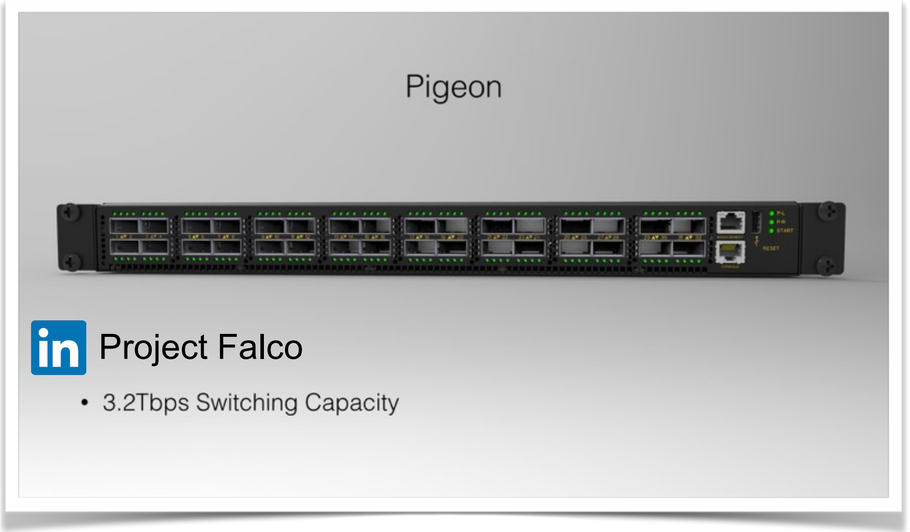
10/25/50/100G Attachement

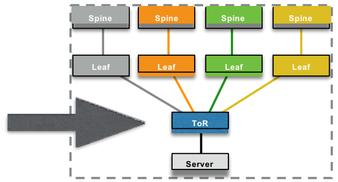
Regular Server Attachement 10G

Each Cabinet: 96 Dense Compute units

Half Cabinet (Leaf-Zone) 48x 10G port for servers + 4 uplinks of 50G

Full Cabinet: 2x Single ToR Zones: 48 + 48 = 96 Servers





Tier 2

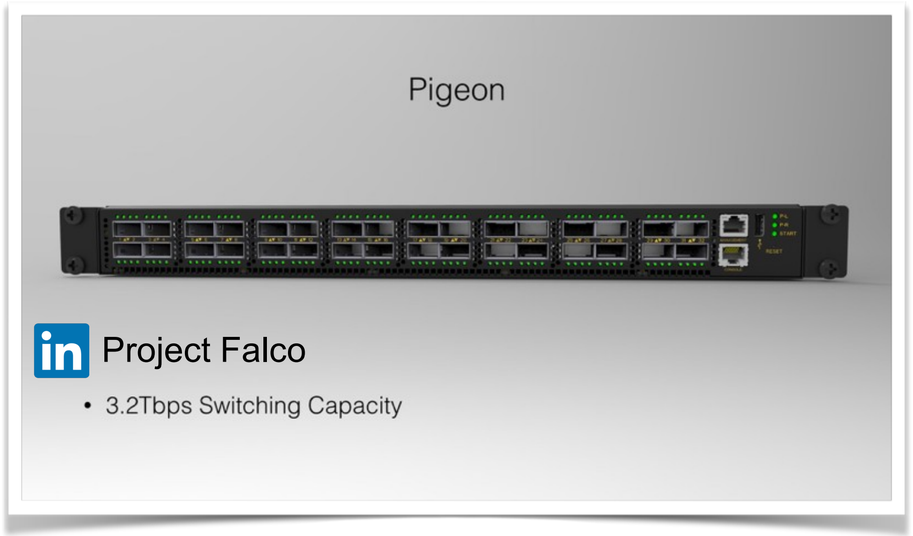
Leaf

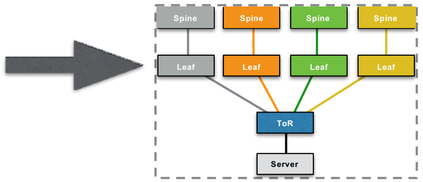
Broadcom Tomahawk 32x 100G

Non-Blocking Topology:

32x downlinks of 50G to serve 32 ToR

32x uplinks of 50G to provide 1:1 Over-subscription





Tier 3

Spine

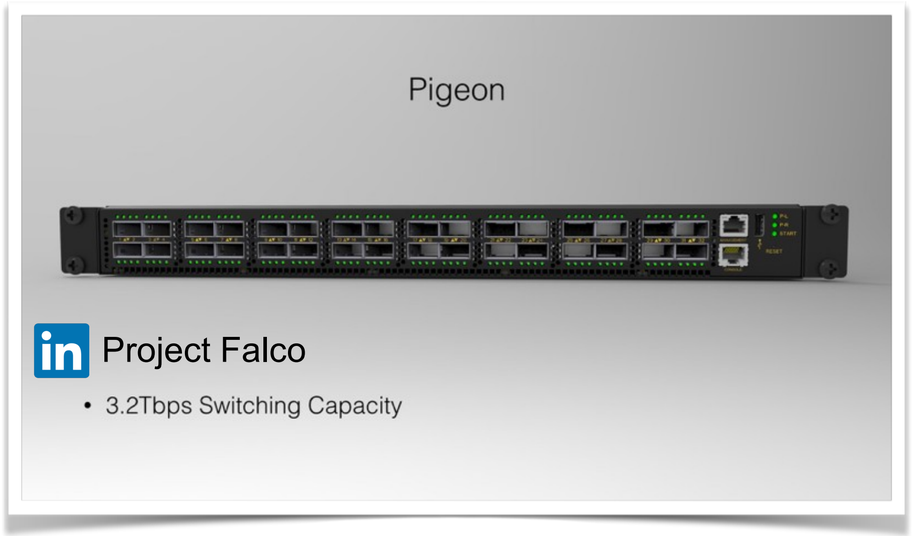
Broadcom Tomahawk 32x 100G

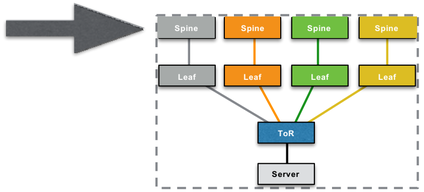
Non-Blocking Topology:

64 downlinks to provide 1:1 Over-subscription

To serve 64 pods (each pod 32 ToR)

100,000 Servers: Each pod (Approximately 1550 Compute)





Challenges

- **Fault isolation.**
- **Fault correlation and remediation.**
- **Build and operations automation.**
- **Physical design.**
- **Logical design.**

Looking Ahead

OPS optimization (Prediction & Remediation Engine)

Open Fabric (New WebScale Protocol)

12.8Tbps chip (Ultra-Low Latency)