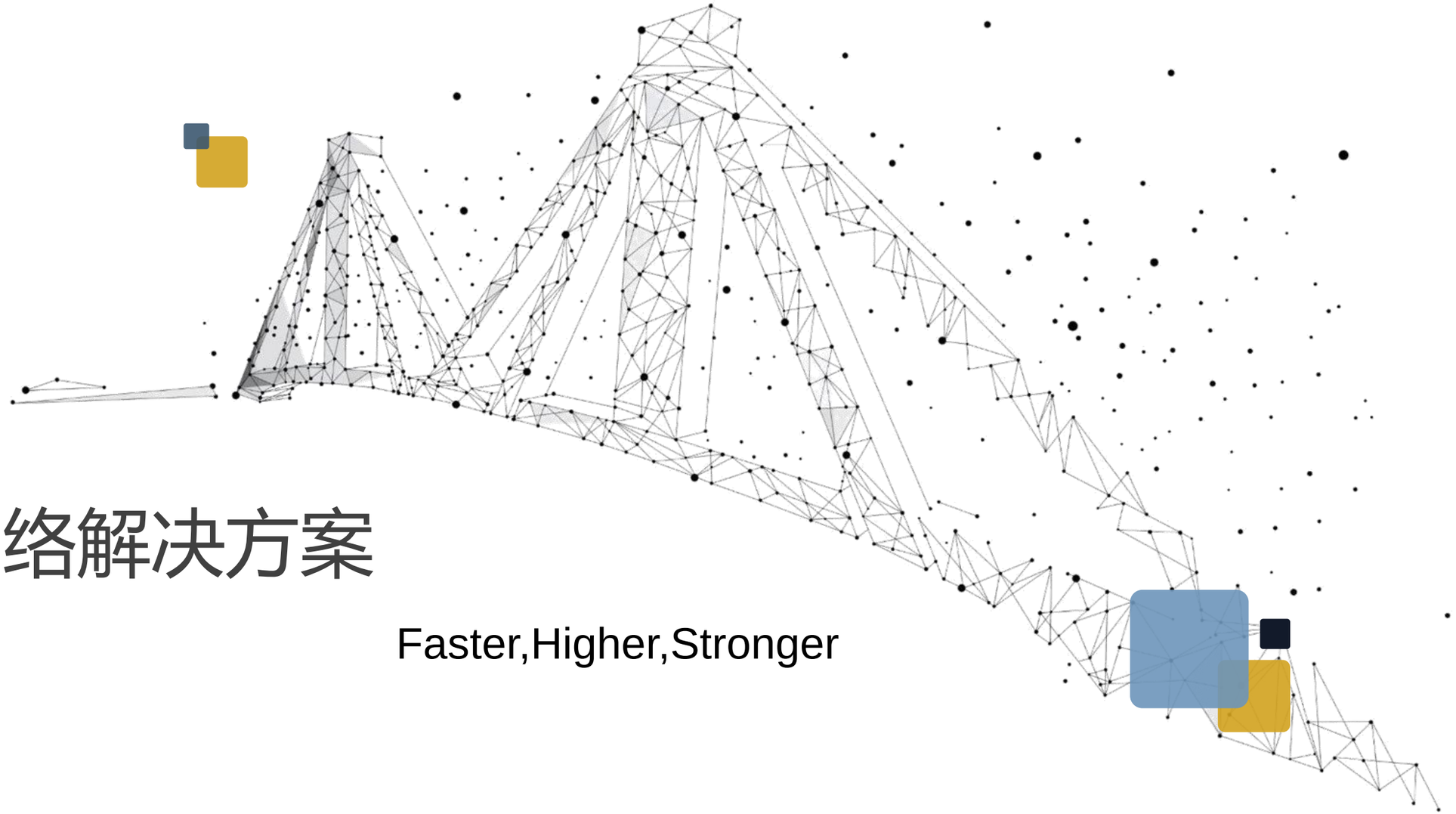


新时代网络解决方案

Faster,Higher,Stronger

Mellanox 概述

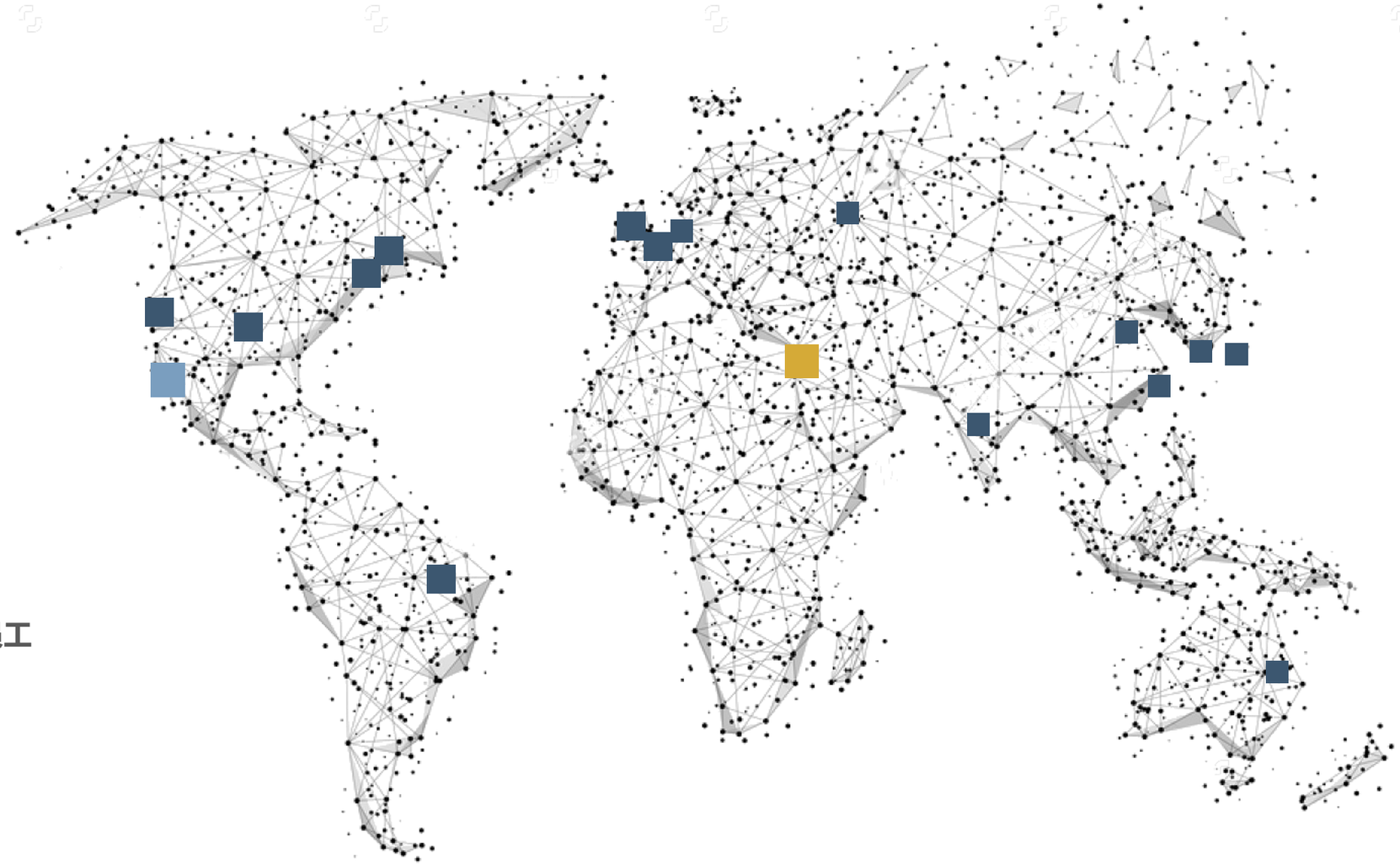
公司总部

- 以色列约克奈姆
- 加州桑尼维尔
- 全球办事处

全球**约** 2,900 **名员工**

NASDAQ®

股票代码：
MLNX



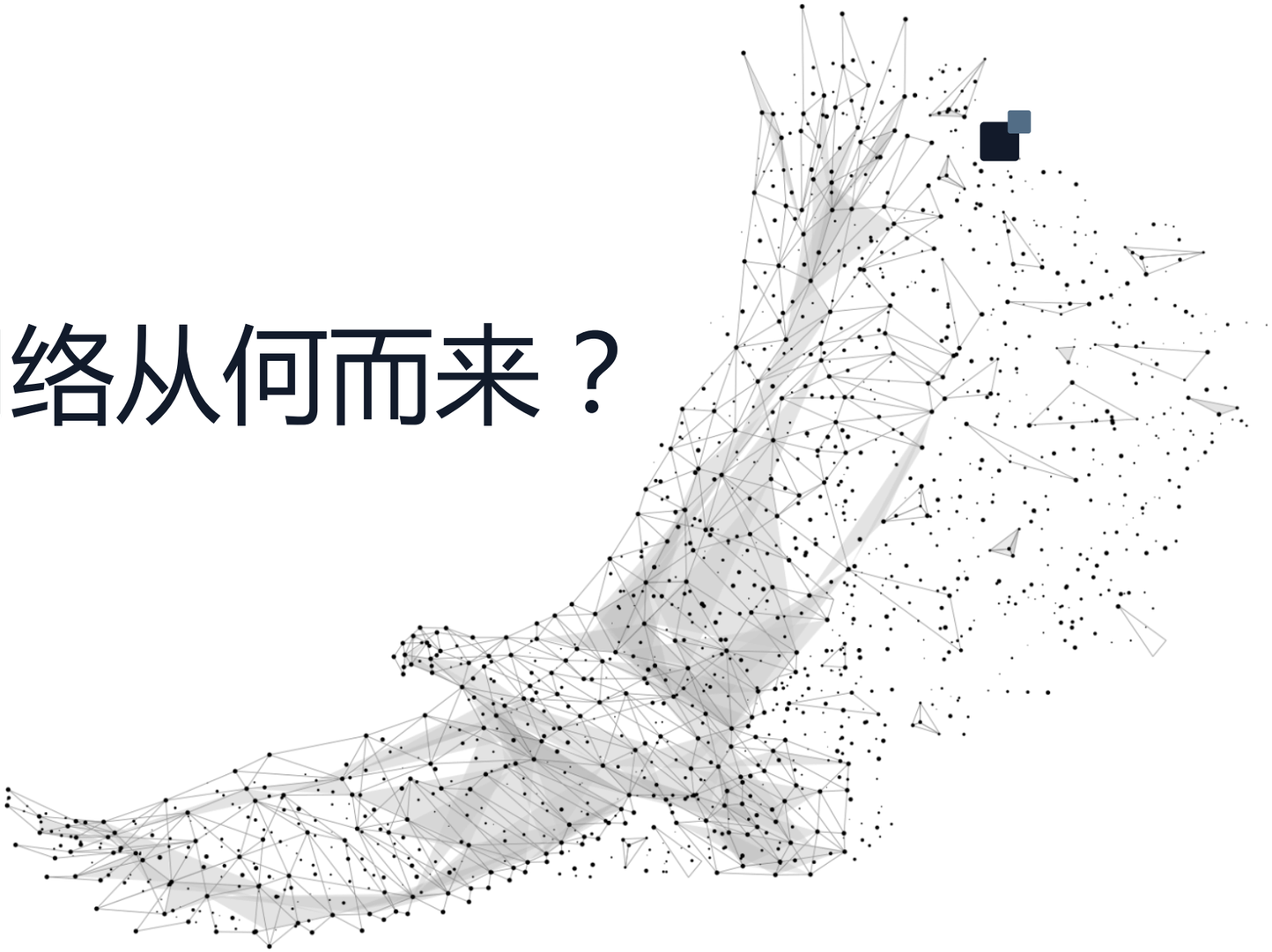
全面的端到端产品组合

InfiniBand 和以太网

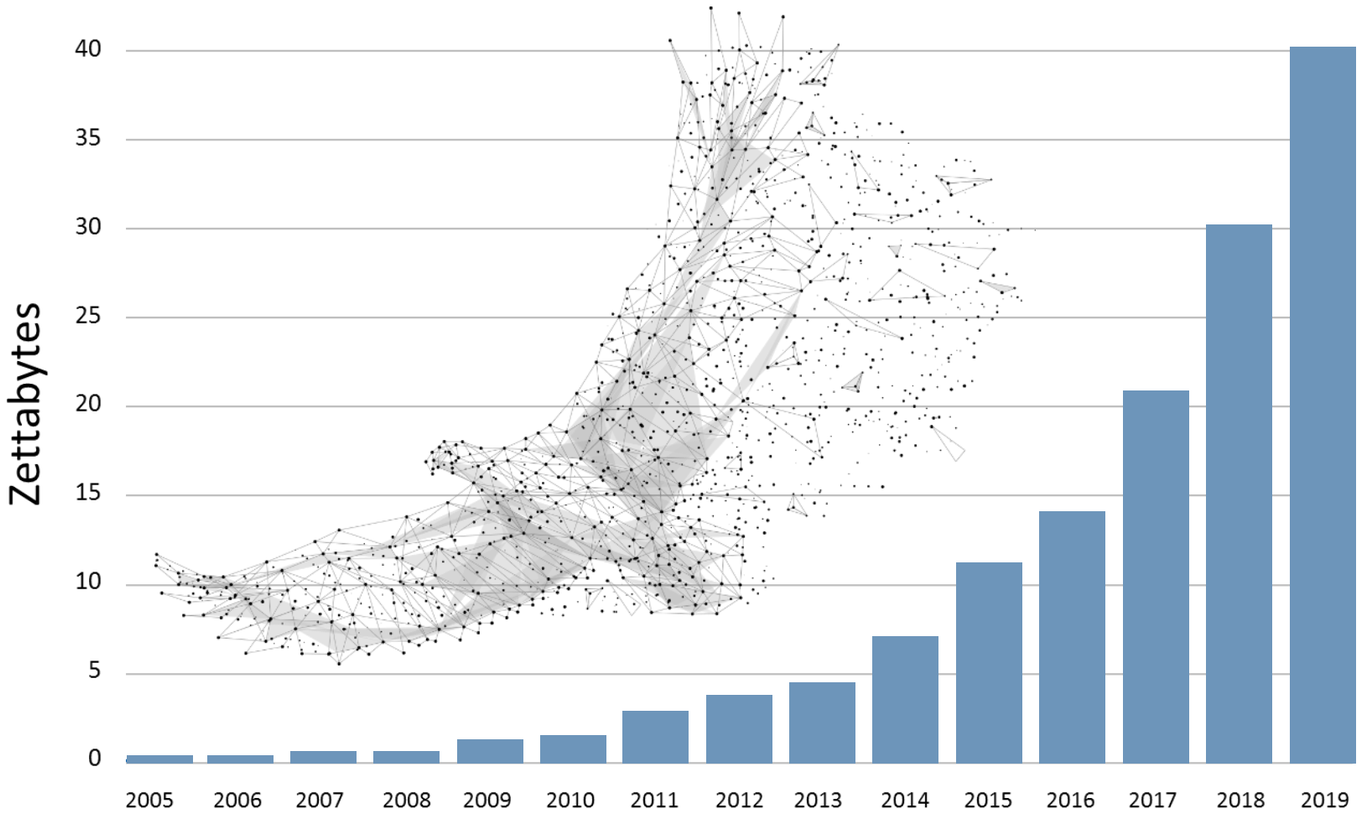
- 软件
- NPU和多核
- 城域/广域网
- 交换机/网关
- 网卡适配器
- 芯片
- 线缆/模块

领先的端到端互连解决方案提供商

25G网络从何而来？



指数级数据增长无处不在



更高的 数据速度

更快的 数据处理

更好的 数据安全

云计算

高性能计算

大数据

安全

物联网

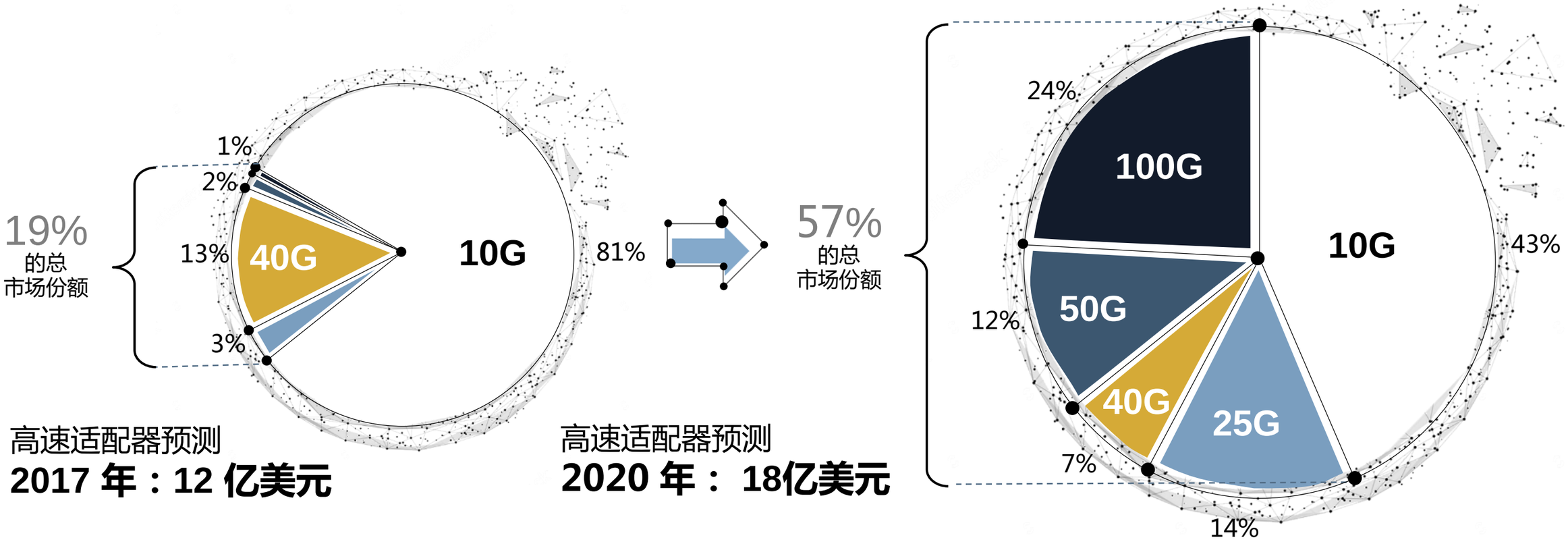
企业
商业智能

存储

机器学习

25Gb/s 及以上以太网市场的大幅增长

2017 年上半年：Mellanox 的 25、40、50、100G 网卡市场份额达 86%



Crehan Research：2016 年第二季度服务器级适配器市场份额；Crehan Research：长期预测 – 服务器级适配器和 LOM（2016 年 7 月）；Crehan Research：长期预测 – 数据中心交换机（2016 年 7 月）

为什么是25G？

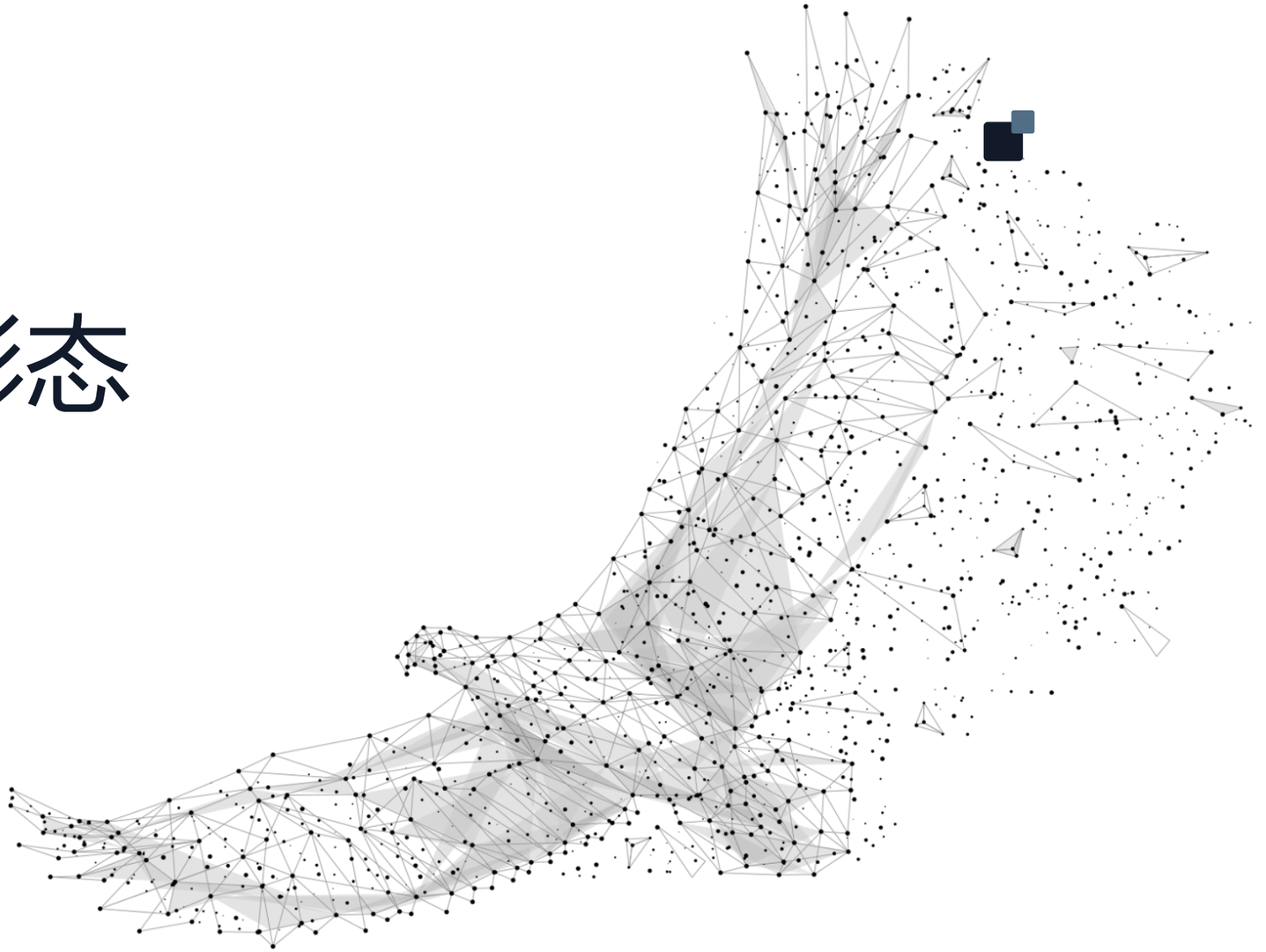
- 25G比10G快，效率比40G高，成本又比40G低
- 25G与SerDes的速率更匹配！

主流Serdes的速率正好就是25Gbps，从25G网卡出来到对端的25G网卡，端到端的所有连接全只需要一条25Gbps速率的SerDes连接通道，40GE需要四个10G SerDes连接通道才能实现，两个40GE网卡之间的通信，需要四对的光纤

- 25G网卡对PCIe通道的使用效率更高
- 数据中心采用25GE布线成本更低

25GE网卡和交换机上使用的是SFP28模块，因为仅采用单通道连接，所以兼容过去10GE时代的LC光纤。如果是从10GE升级到25GE的话，则无需重新布线

产品形态



主流25G网卡ConnectX-4

ConnectX-4: Highest Performance Adapter in the Market

InfiniBand: SDR / DDR / QDR / FDR / EDR

Ethernet: 10 / 25 / 40 / 50 / 56 / 100GbE

100Gb/s, <0.7us latency

150 million messages per second

OpenPOWER CAPI technology

CORE-Direct technology

GPUDirect RDMA

Dynamically Connected Transport (DCT)

Ethernet offloads (HDS, RSS, TSS, LRO, LSOv2)

Multi-Host Technology

开放以太网10/25/40/50/100 交换机产品

SN2700 – 32x100GbE (64x50GbE)
Ideal 100GbE ToR / Aggregation

SN2410 – 8x100GbE + 48x25GbE
25GbE ➔ 100GbE ToR

SN2100 – 16x100GbE ports
Ideal storage/Database Switch
Highest 25GbE Density per rack unit

Spectrum™

- Predictable Performance
- Fair Traffic Distribution for Cloud
- Best-in-Class Throughput, Latency, Power Consumption
- Zero Packet Loss

cumulus®

LinX互联线缆模块

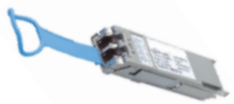
- Passive Copper Cables
  - SFP28: 0.5, 1, 1.5, 2, 2.5, 3 and 5m
  - QSFP28: 0.5, 1, 1.5, 2, 2.5, 3 and 5m


- Active Optical Cables (AOC)
  - SFP28: 5, 10, 15 and 30m
  - QSFP28: 3, 5, 10, 15, 20, 30, 50 and 100m

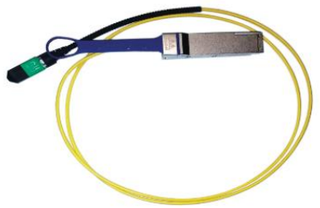


- Breakout Cables
  - QSFP28 to 4xSFP28: 1, 1.5, 2, 2.5, 3 and 5m
  - QSFP28 to 2xQSFP28: 1, 1.5, 2, 2.5, 3 and 5m


- Short reach Transceiver modules
  - SFP28: SR, 100m on OM4 MMF, LC/LC
  - QSFP28: SR4, 100m on OM4 MMF, MPO

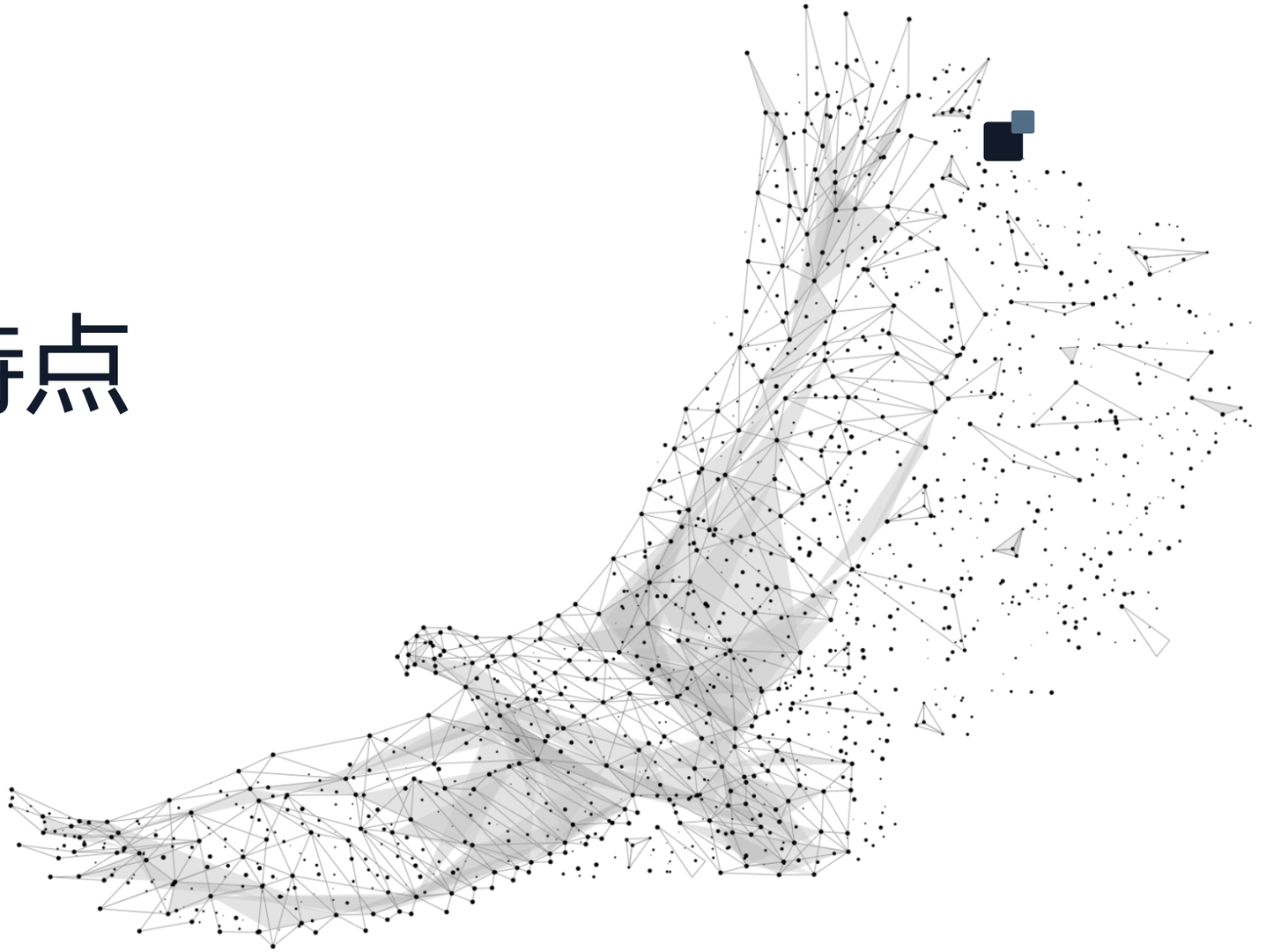



- 2km reach Transceiver modules
  - QSFP28: PSM4, 2km on SMF, MPO
  - QSFP28 Pigtail: PSM4, 2km on SMF, MPO
  - QSFP28: WDM, 2km on SMF, LC/LC

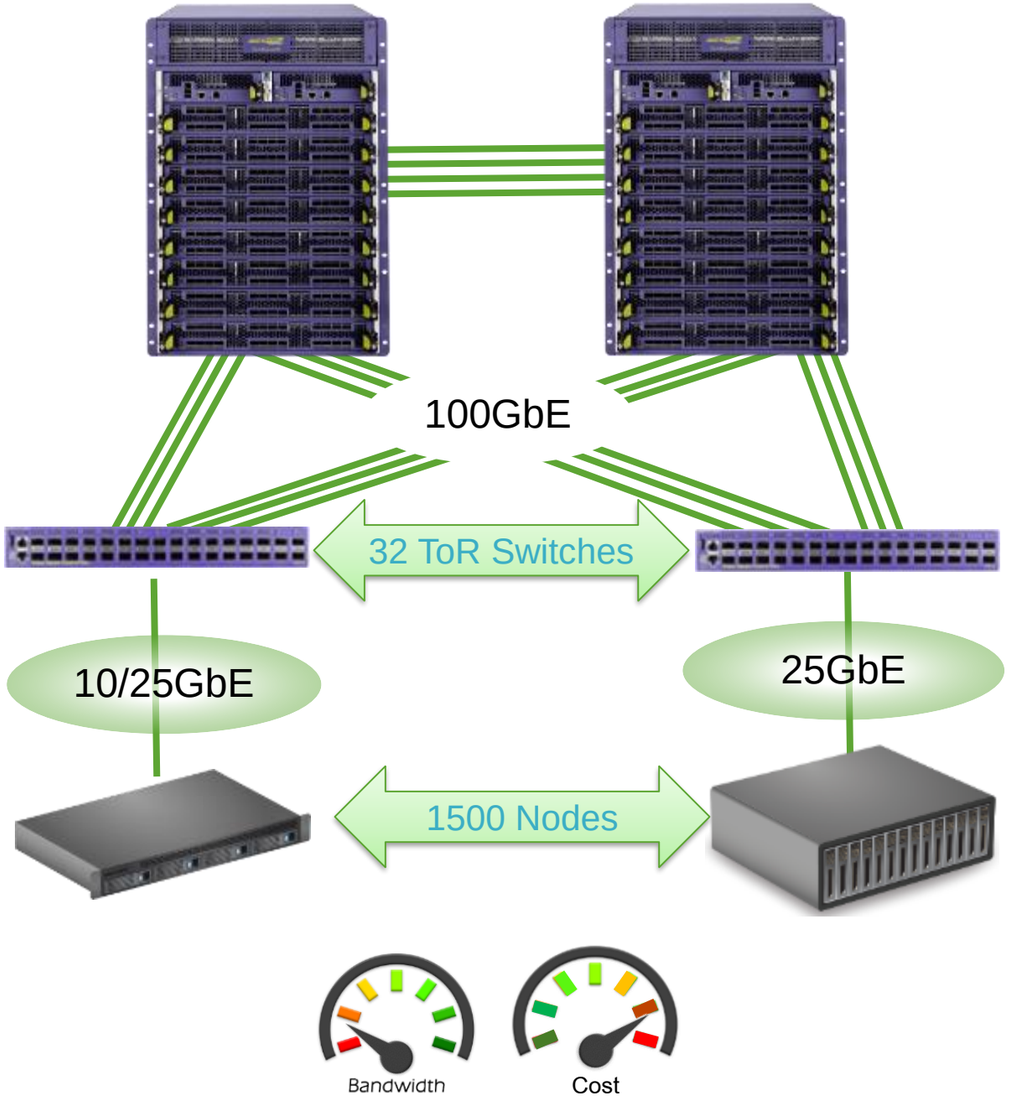



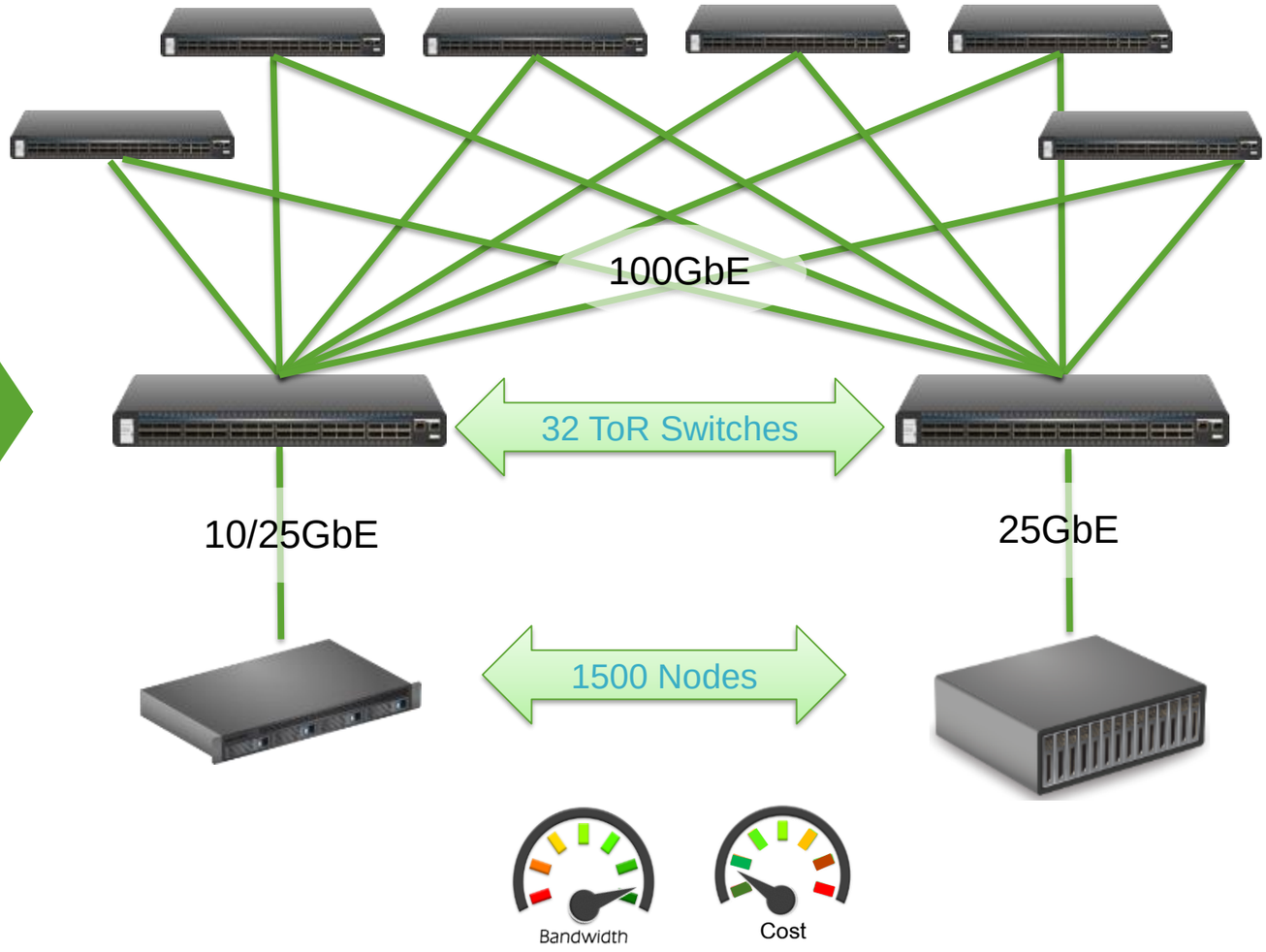
- 10km reach Transceiver modules
  - QSFP28: LR4, 10km on SMF, LC/LC

架构特点



现代数据中心架构演进

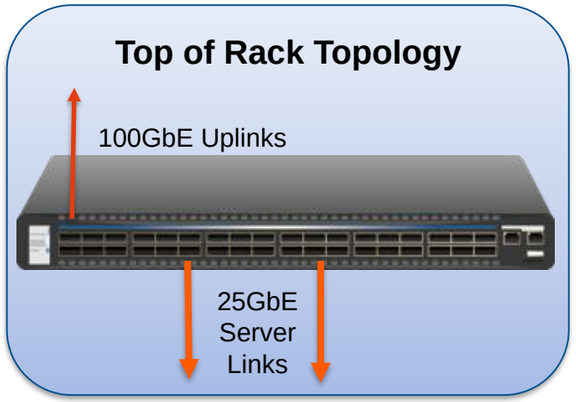




VMS架构优势

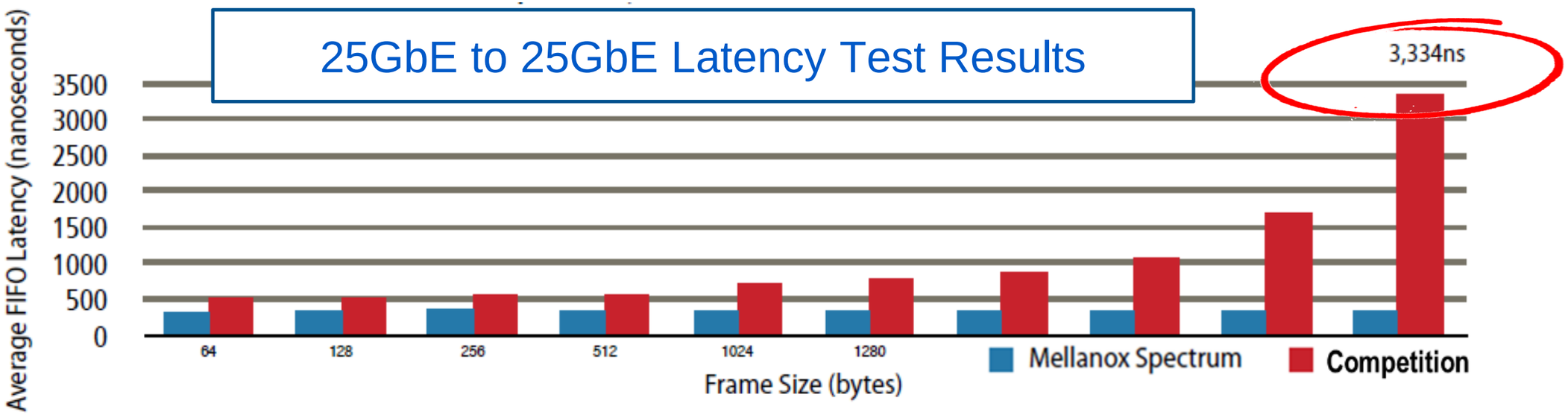
	Chassis	Mellanox VMS
Scale-Out	Expansions are painful and inflexible	Pay-as-you-grow with leaf incremental
Traffic Pattern	North-South, traversing three levels with poor performance	Dynamic movement between for hyper-converged and virtualized networks
Over-Subscription	Chassis as spine is non-blocking with full spine CAPEX	Simple over-subscription, can utilize less switches and interconnect to lower CAPEX
Visibility & Control	Little visibility or control (if any) on the traffic behavior inside the chassis	All traffic flows and known and debug-able
Buffers	Large buffers are expensive and not needed Can create unnecessary high latency and hurt network predictability	Single shared buffer, 10x microburst capacity
Resiliency	Chassis failure can take entire fabric down	Switch failure is negligible by re-routing the fabric
Freedom of NOS	Vendor lock-in	HW<>SW disaggregation, chose the preferred NOS
Power Consumption	High power consumption	Up-to 75% lower power consumption
Price	$$$$	$$

极低时延网络



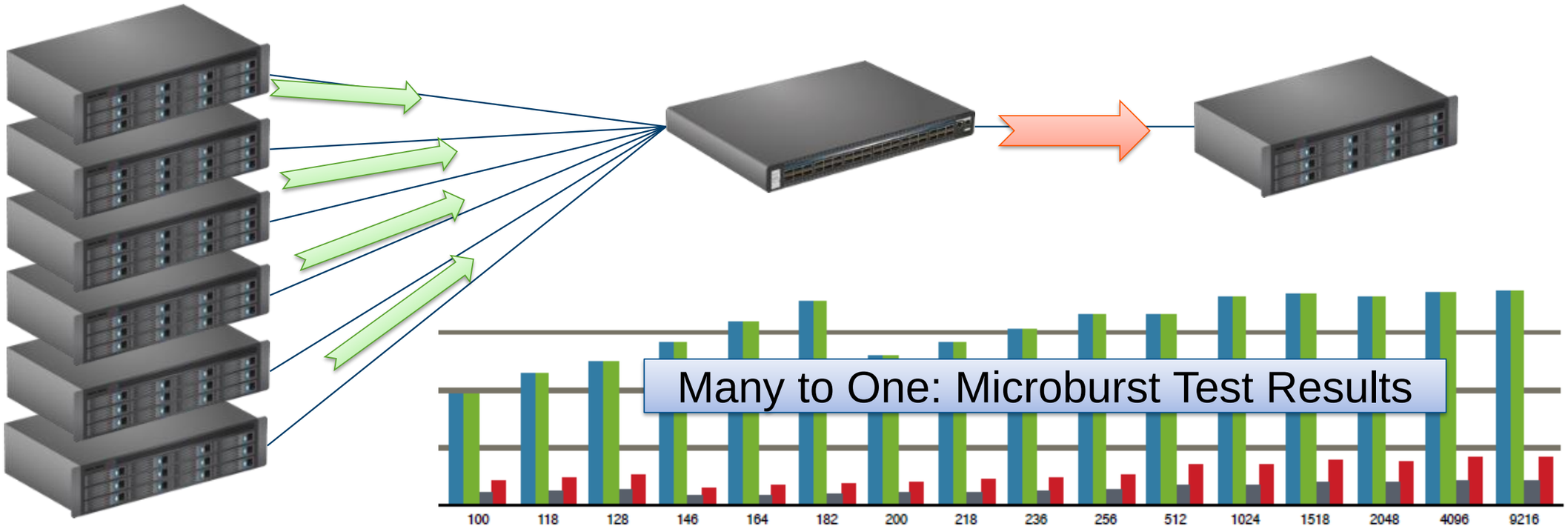
- **Top of rack topology**
  - 25/50G servers talking to each other
- **Spectrum: 300ns latency**
  - Consistent cut-through latency

- **Who cares about latency?**
  - Machine Learning Neural Networks
  - ***Spectrum is the only low latency switch >10G***

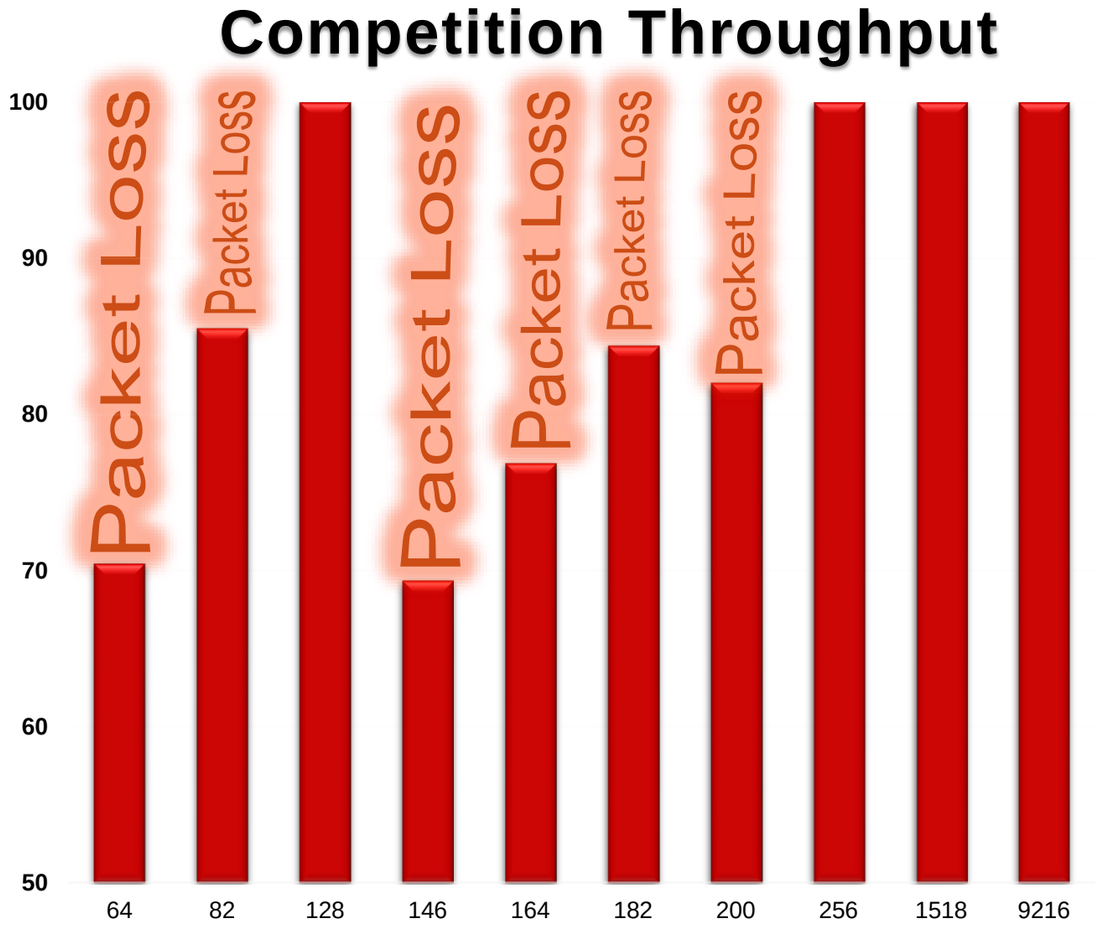


10倍业界的Microburst 性能

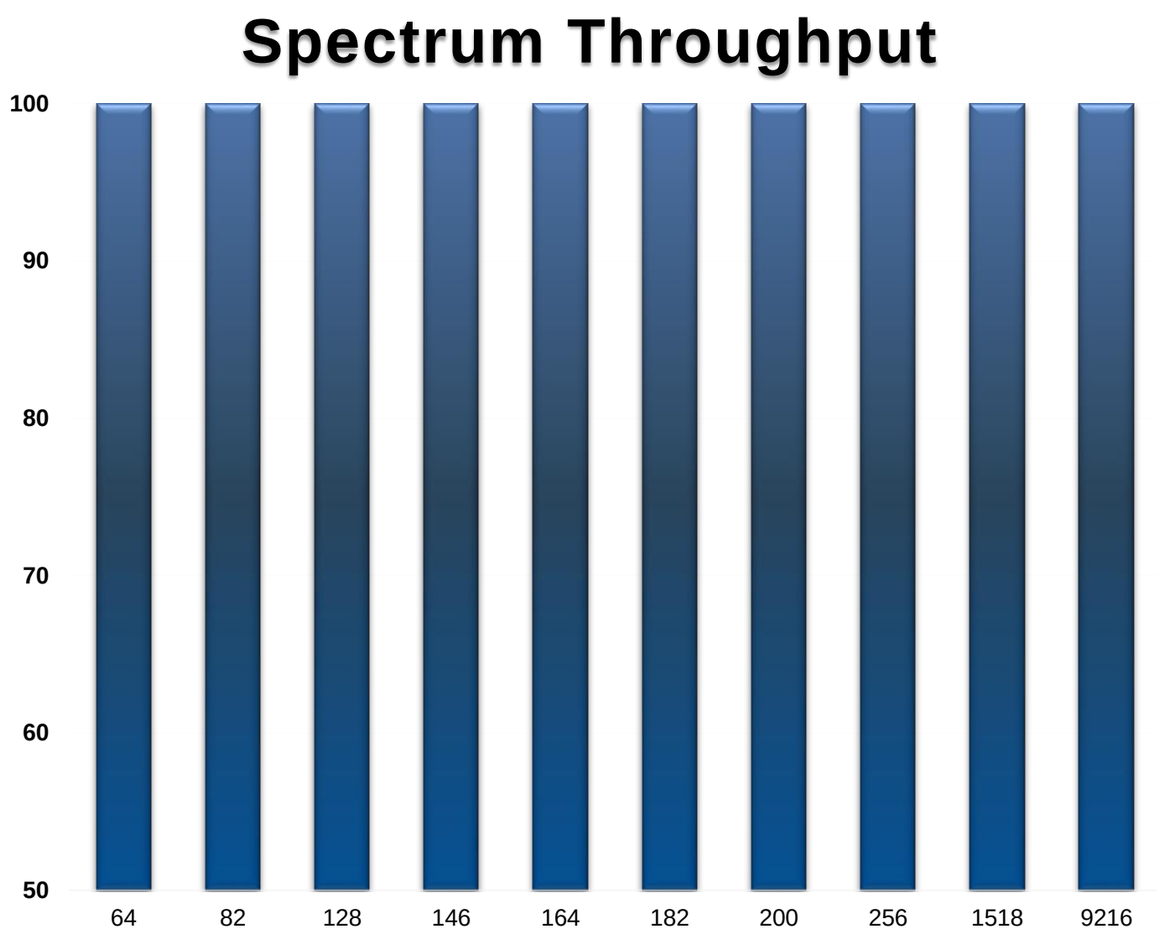
- Spectrum performs better:
  - 10x better microburst absorption
  - Fully shared buffer
- Who cares about Microburst?
  - Spark & Big Data Applications
  - Machine Learning Applications



真正的零丢包

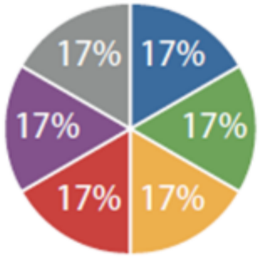
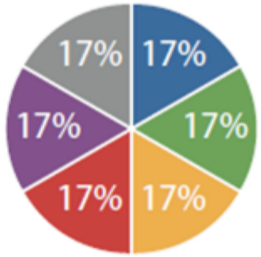
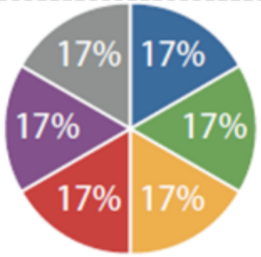
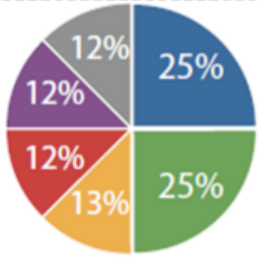
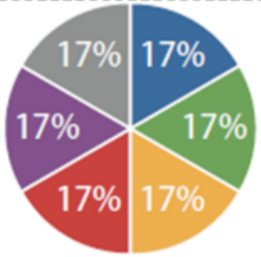
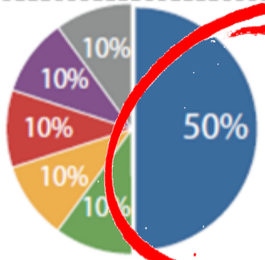


VS



- Packet loss creates problems at any speed
- 32 Ports @ 100Gb/s requires a packet rate of 4.76Bpps
  - Anything less will lose data when small packet microbursts occur

公平流量负载

	Mellanox Spectrum	Competition
	Always Fair bandwidth distribution for each stream	Unfair bandwidth distribution in most test cases
Test 1		
Test 2		
Test 3		

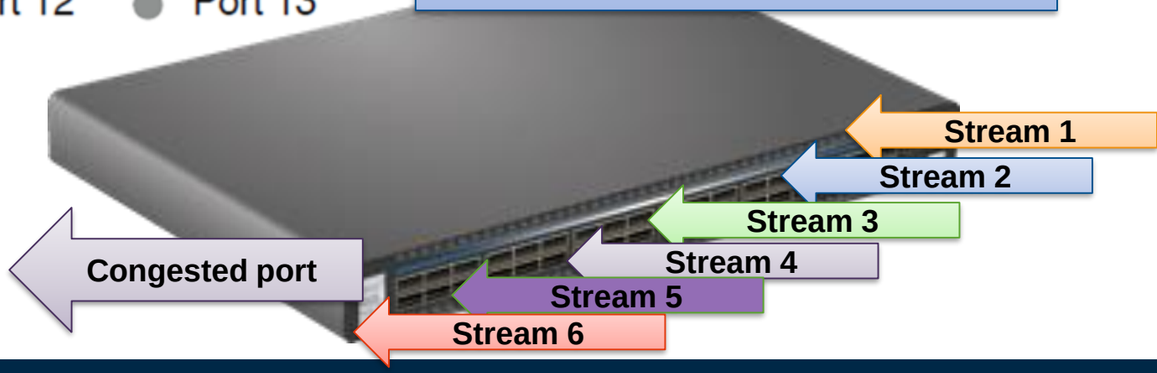
Destination port is Port 31 for all streams
Following is the source port of each stream

- Test 1: Port 9, Port 10, Port 11, Port 12, Port 13, Port 14
- Test 2: Port 7, Port 8, Port 9, Port 10, Port 11, Port 12
- Test 3: Port 8, Port 9, Port 10, Port 11, Port 12, Port 13

One server unfairly gets 5x bandwidth than others

Spectrum Delivers Fair File Transfer

Oversubscribed Scenario
6 streams sent to 1 port
Common in all datacenters



- Bandwidth must be fairly allocated
- Competition unpredictably allocates bandwidth

开放以太网: 白盒交换机模型，OEM品质

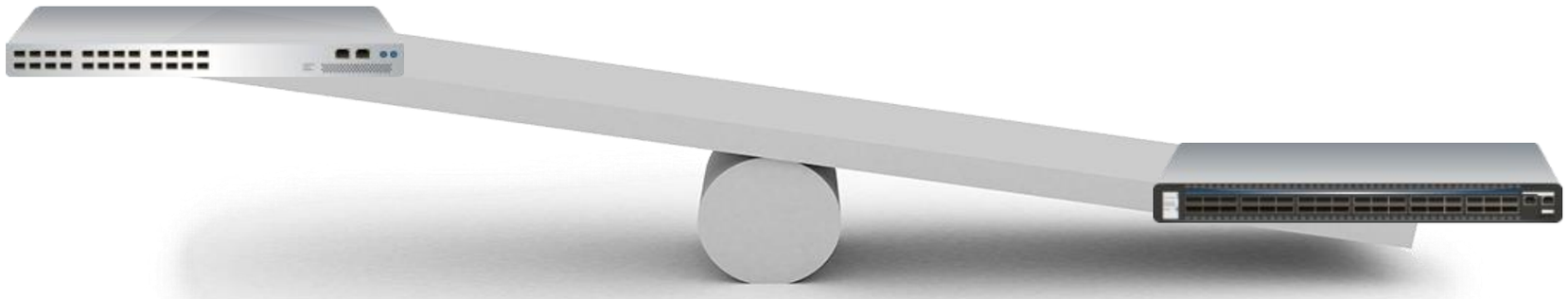
- ❑ Server Expertise
- ❑ White Box Service Level
- ❑ White Box Quality
- ❑ HW-only support

White Box

VS.

Mellanox TECHNOLOGIES

- ☑ Network Expertise
- ☑ OEM Service Level
- ☑ OEM Quality
- ☑ Most Reliable (MTBF)
- ☑ One Throat to Choke



统一网管平台：NEO™ 令管理简单，智能，高效

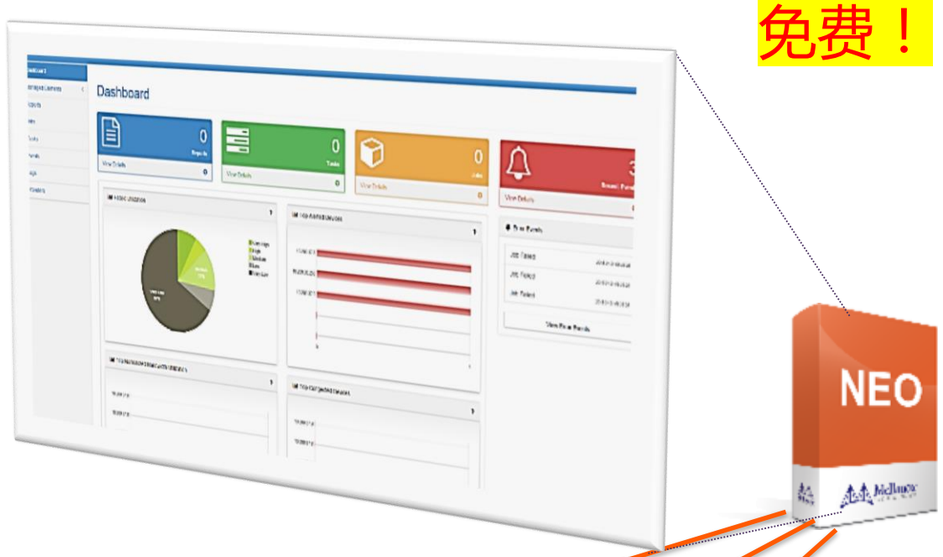
Fabric Bring-up Made Easy

Efficient Network Monitoring

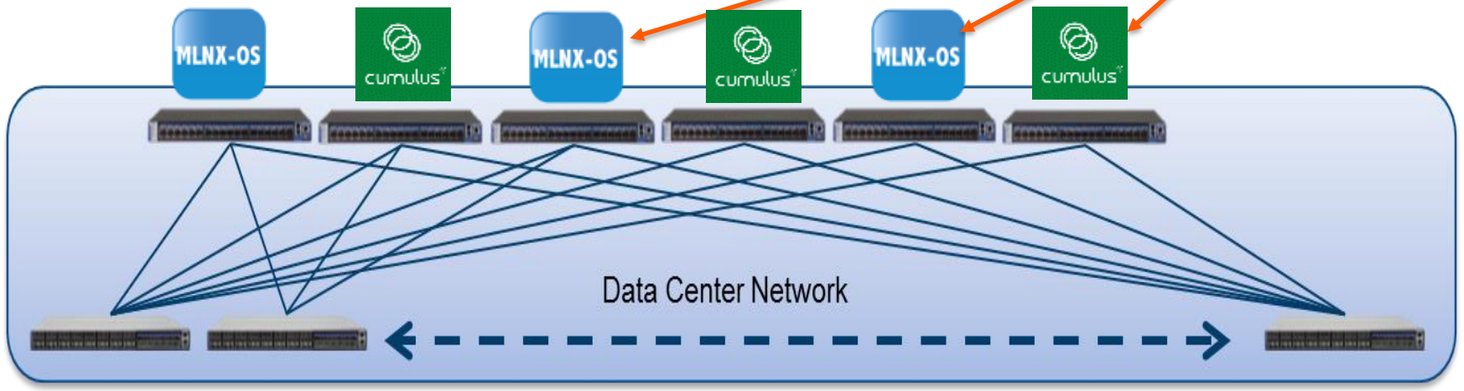
Save Time on Day-to-Day Operations

Network API, Built for Integration

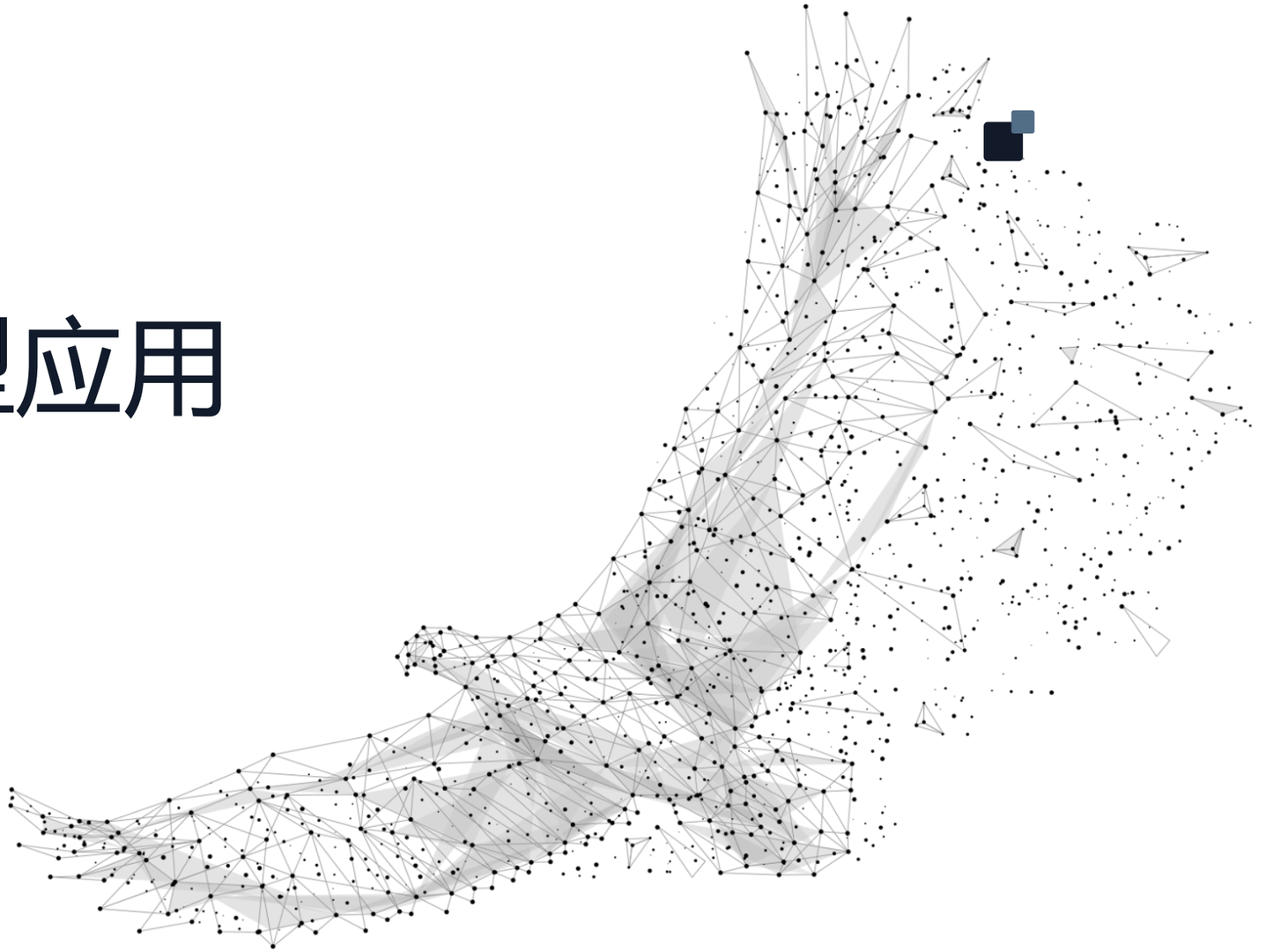
免费！



- ✓ MLNX-OS & Cumulus Linux
- ✓ End-to-End RoCE Automation
- ✓ Auto-Provisioning with Nutanix and OpenStack



典型应用



主要市场

云

存储

大数据
机器学习

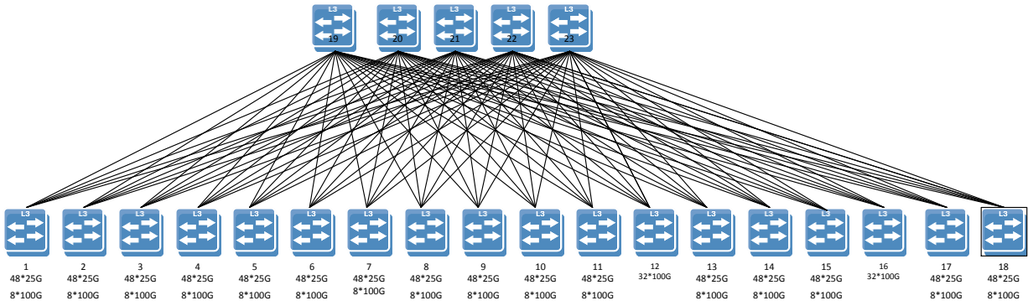
广电/传媒

高频交易

Defense

案例------科大讯飞

- Customer: iFLYTEK – 600 Nodes 25G Cluster
- Application: Hadoop/MPI/Lustre/GlusterFS
- *Needs for networking*:
  - High efficiency 25G to host and 100G for uplink
  - Convergence Network for Hadoop/MPI/Storage
  - Easy to scale



Solution → Mellanox Ethernet end to end with SN2410/SN2700 switch, NEO, ConnectX-4 Lx NIC and LinkX cable

- High efficiency 25G network
  - 3:2 over subscription, 1us end to end low latency.
  - Automatically networking provision & management.
- Convergence network with QoS
  - TCP/IP and RoCEv2 support in converged network.
  - PFC and ECN deploy for different application.
- Scalability
  - Hyper scale topology with full layer 3 BGP as routing
  - 8/16 ECMP

"Mellanox help iFLYTEK to build next generation data center network, easy to adopt different high performance application. Easy to scale more servers without network services disruption.

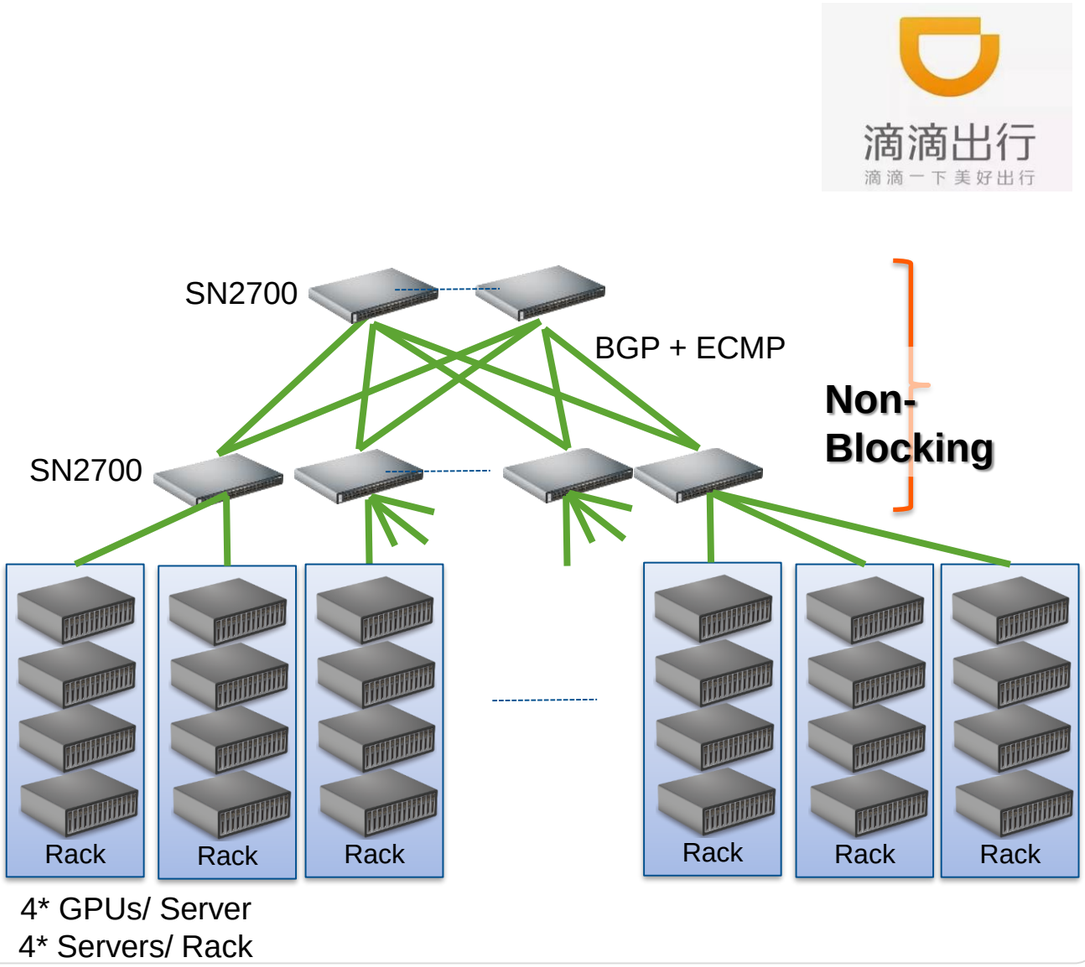
iFLYTEK

案例------滴滴

- E2E: LinkX, ConnectX-4 Lx SN2000
- Vertical: ride-hailing, transportsation
  +400M users across +400 cities in China
  in Oct'16 Didi clocked +20M/day, 4x the entire US
- Application/Use Case: Deep learning cluster to provide image and video recognition for self-driving vehicles, safety and enhanced customer experience.

1. Low latency: with Spectrum Leaf & Spine measured 1usec latency between application to application
2. Mellanox is the RoCE company expertise in deploying lossless (PFC+ECN) fabric with mature RDMA feature support E2E
3. Easy to scale: Leaf & Spine arch running BGP/ECMP

Solution Architecture

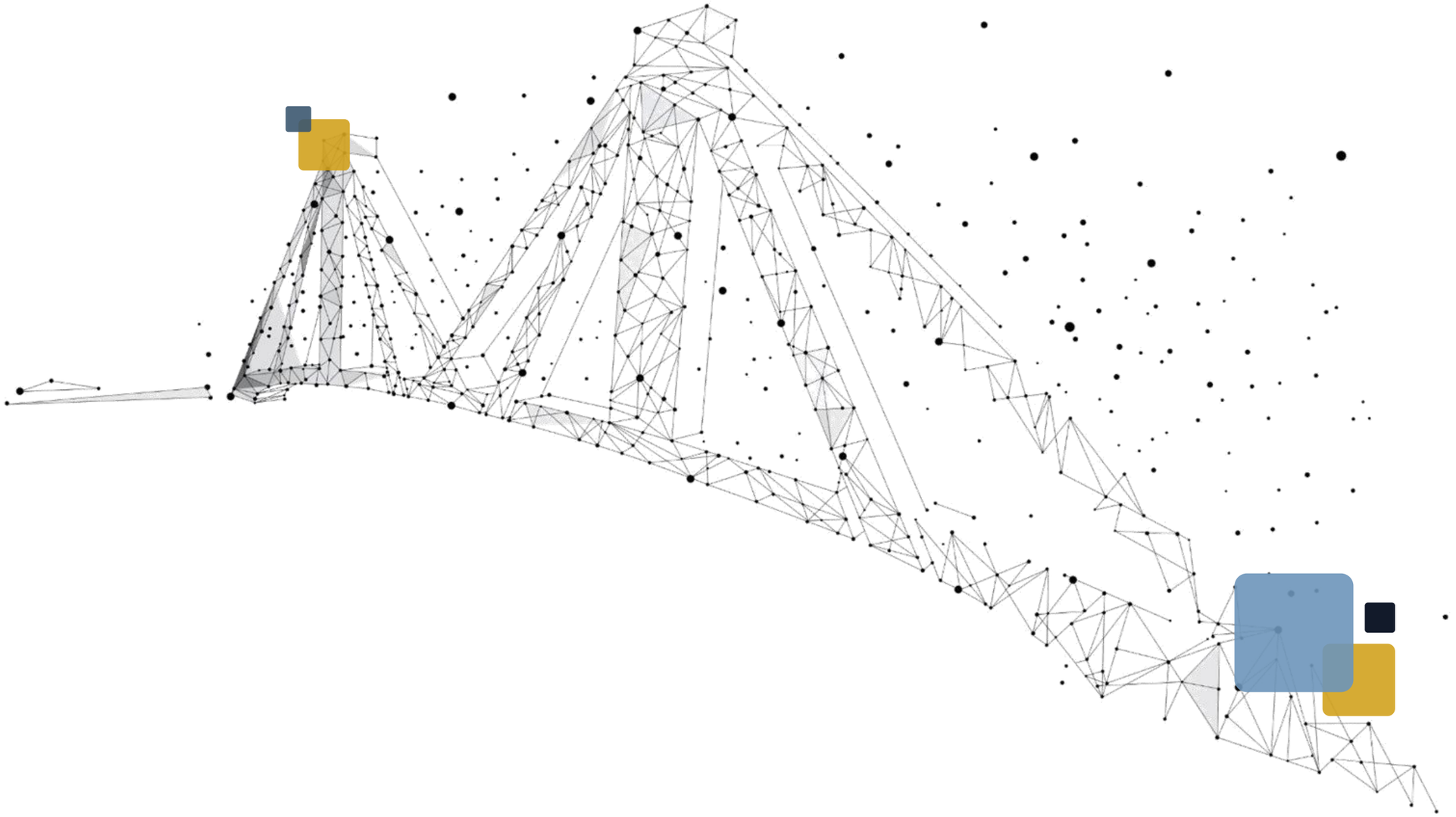


MLNX-OS™

总结

面向智能计算和存储平台

多核 CPU 和智能网络	机器学习的未来	BlueField Quantum ConnectX
以太网卡	25 是新的 10 版！50G、100G 是现在的主流产品！	ConnectX Cloud
以太网交换机	开放式以太网交换机，用于可编程网络	Spectrum Vendor NOS: MLNX-OS, cumulus Open Source NOS: Open Switch, SONiC Linux: switchdev
线缆和转发器	行业集成和硅光技术	LinkX
多核和网络处理器	智能网络解决方案支持无所不在的安全性！	Indigo BlueField Innova



谢谢