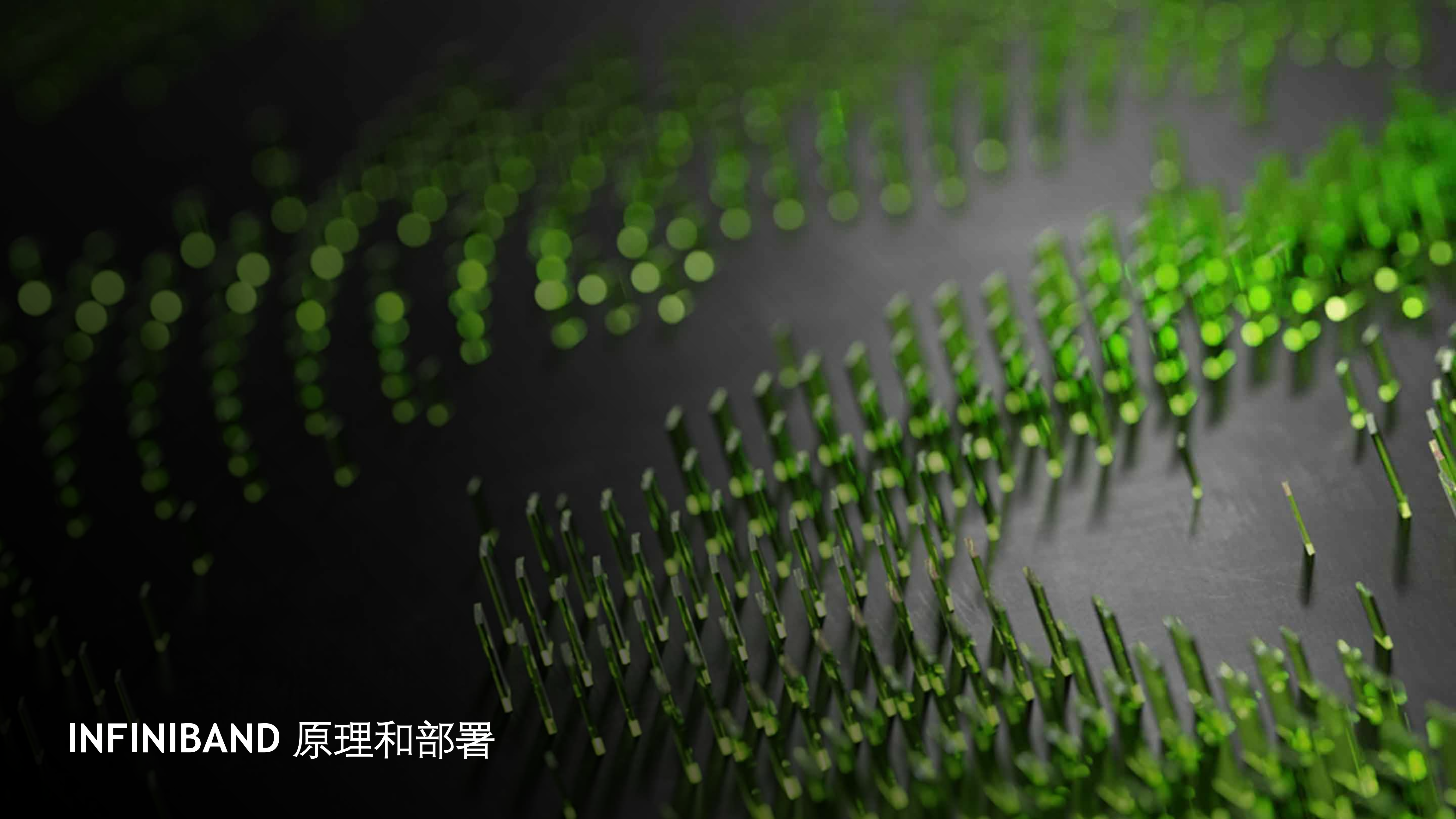
INFINIBAND HPC CLOUD SOLUTION

Aganda：

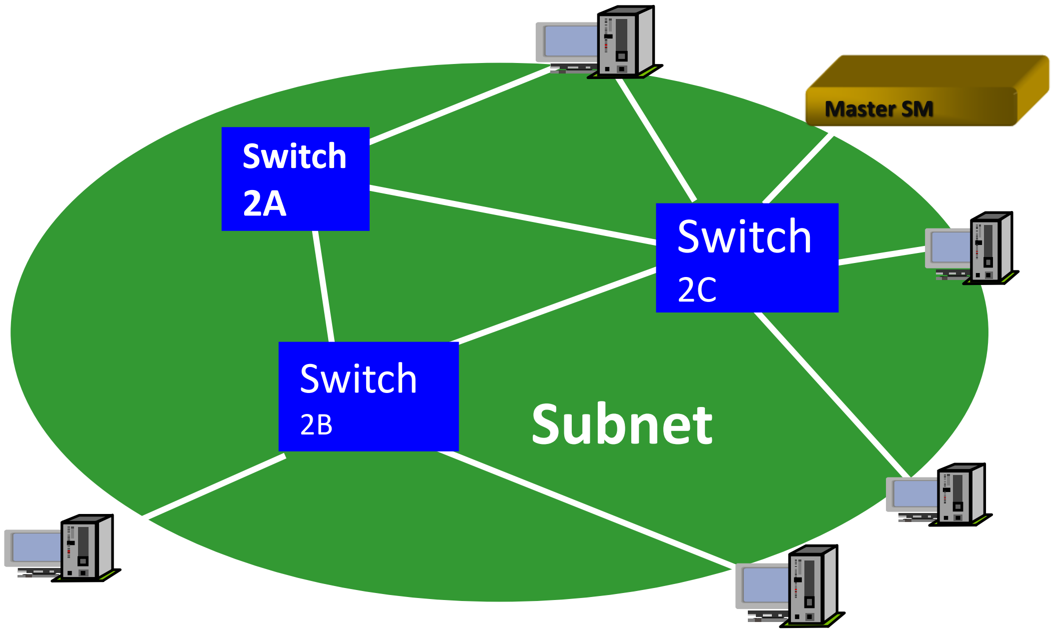
1. Infiniband如何在公有云建设，如何部署
2. Infiniband建设/部署的拓扑，最小启动建设单元
3. Infiniband如何监控运维
4. IB部署方案分享



INFINIBAND 原理和部署

SUBNET MODEL

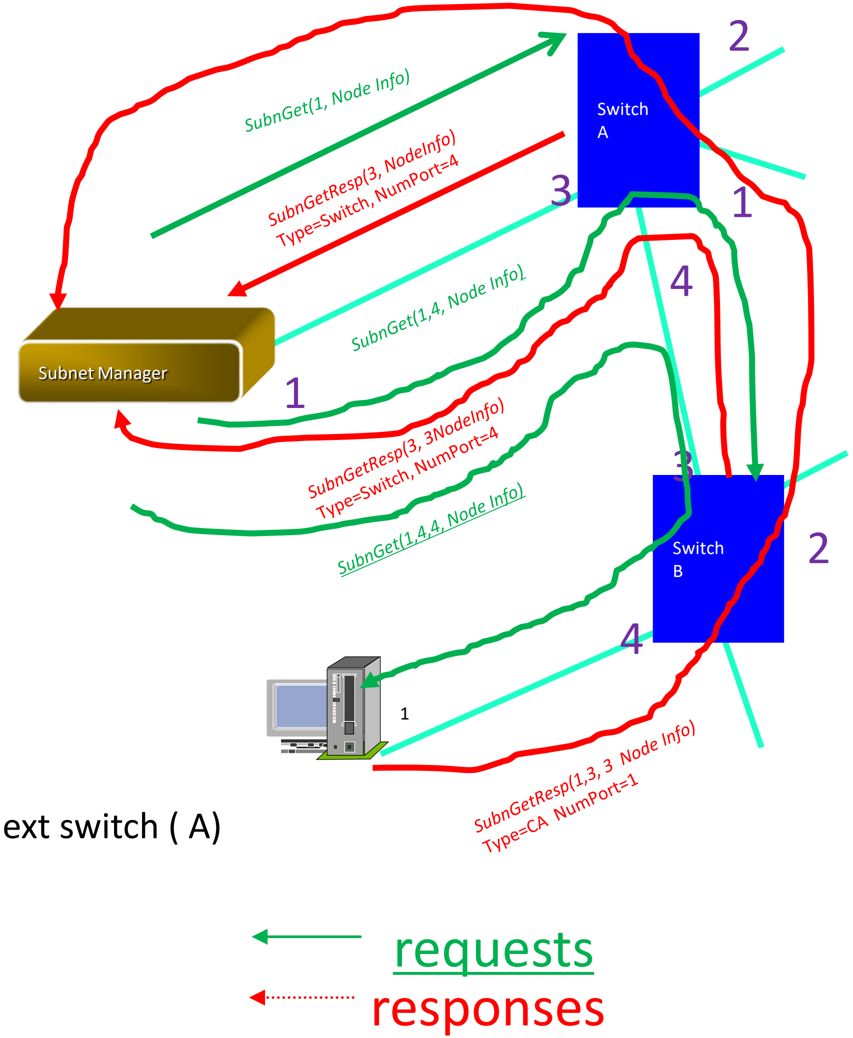
- Subnet = HCAs and interconnected through switches
- Each subnet has its own LID space
- Each subnet has at least one SM and exactly one (logical) Master SM



INFINIBAND SUBNET MANAGER

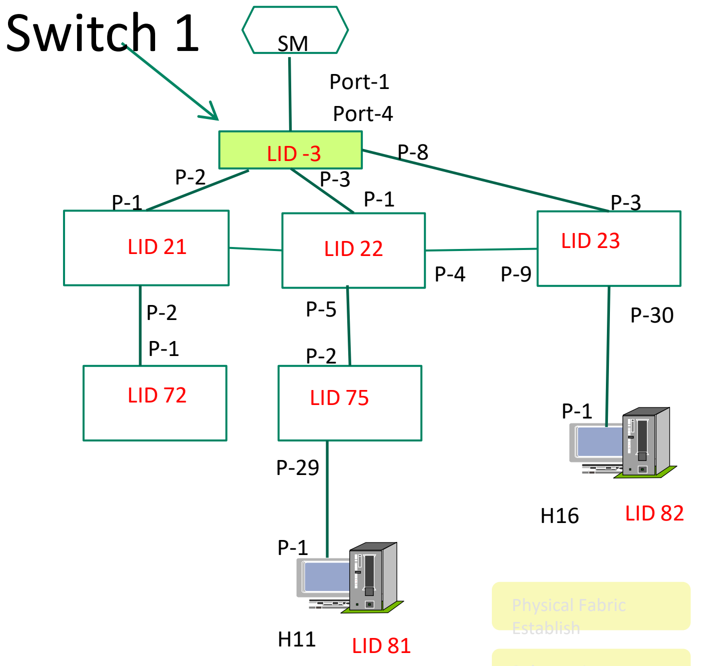
- SM requests like devices responses will include :
  - ❑ Node Info
  - ❑ Ports info

1. SM – I am requesting info via port number 1
2. Switch A – I am responding via port number 3 ,
   I have an Active port Number 4
3. SM – I am requesting info via :
   - my port 1- next switch (A) port 4
4. SWITCH B – I am a switch responding via
   my port 3 via next switch (A) port 3
   - I have a live port port 4
5. SM – I am requesting info via :
   - my port 1- next switch (A) port 4 ,next switch (B) port 4
6. Host – I am a CA , responding via my port 1 , next switch (B) port 3 , next switch (A) port 3



IB FORWARDING TABLE

- After the SM finished gathering all Fabric information , including direct route tables , it assigns a LID to each one of the NODES
- At this stage the LMX table will be populated with the relevant routes option to each one of the nodes
- The output of the LMX will provide the Best Route to Reach a DLID .
  That Result Will be based on Shortest Path First (SPF)



LMX Switch_1

PORT D-LID	2	3	8	Min Hops
21	1	2	3	1
22	2	1	2	1
23	3	2	1	1
75	3	2	3	2
81	4	3	4	3
82	4	3	2	2

LFT   Switch_1

The Dest. LID	Best Route / exit port
21	2
22	3
23	8
75	3
81	3
82	8

Physical Fabric Establish

Subnet Discovery

Information Gathering

Lid Assignment

Path Establishment

Port Configuration

Switch Configuration

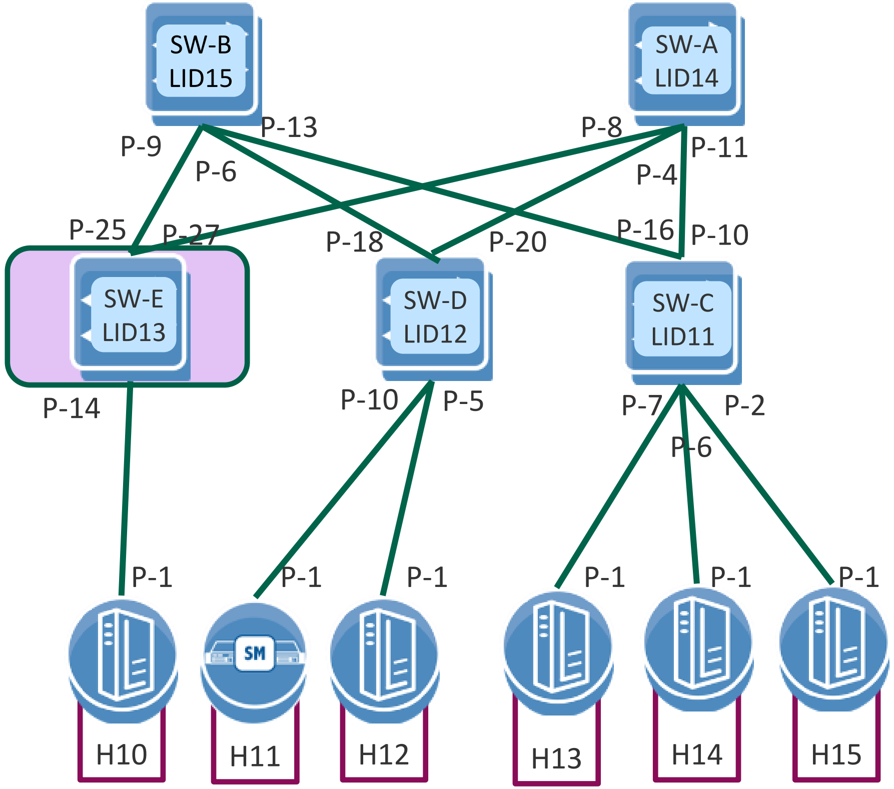
Subnet Activation

ADAPTIVE ROUTING (DYNAMIC)

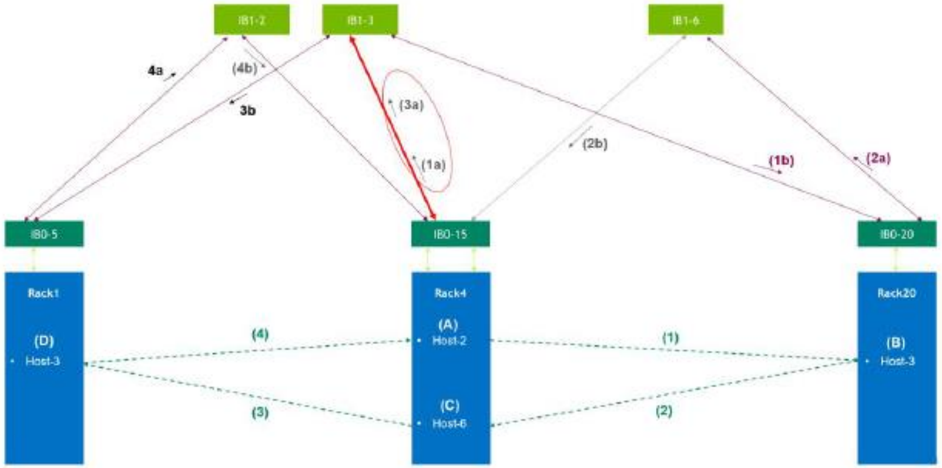
- AR PORT GROUP – AR will automatically group switch multiple ports , Having the same cost (minhop) towards any specific destination lid
- Allows a switch ,to move the data connection between exit ports, selecting the least congested port of that AR port group
- The best exit port to be used , is analyzed and selected by a port transmit “virtual que manager”
- AR functionality is managed by a new SM component called Adaptive Routing Manager
(AR plugin)

AR Group	Ports
1	{25,27}

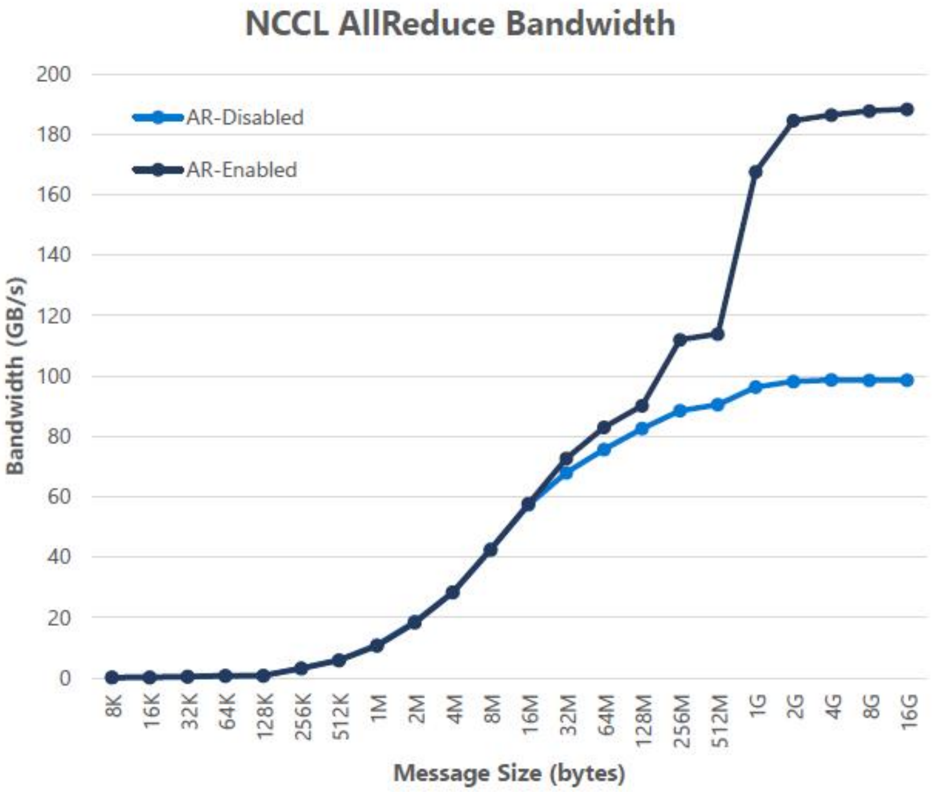
Dest. LID	AR Group
5	1
6	1
7	1
8	1
9	1



Adaptive Routing



Communication paths during NCCL AllReduce



Impact of Adaptive Routing

- Congestion can happen with static routing if a single link is being used by two or more communicating pairs
- AR avoids congestion and offers stable performance

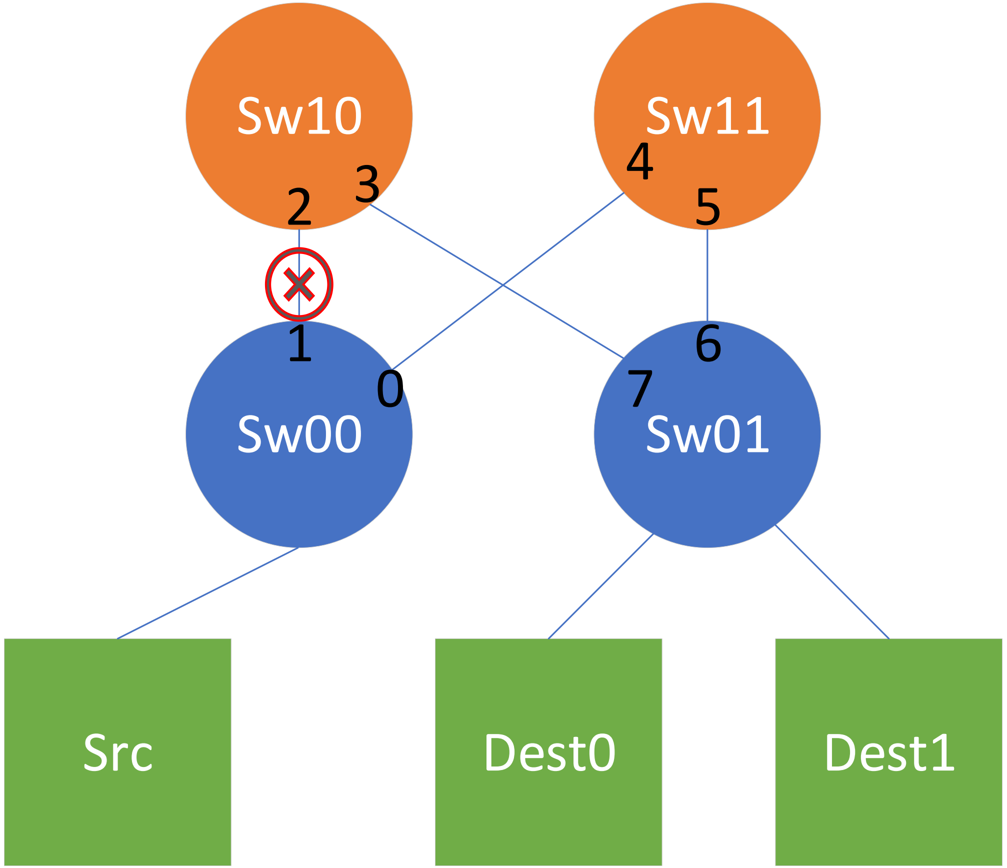
SHIELD

- Regular forwarding decision is a single output port per DLID.
  - Example Forwarding Data Base (FDB) for Sw00:

Dlid	Output Port	Port Status
Dest0	0	Up
Dest1	1	Down

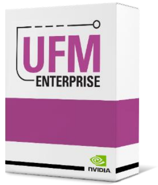
  - When a link fails, traffic sent over is discarded.
- With SHIELD and FRNs, other link options are made available and used if needed.
  - Example FDB with AR for Sw00:

Dlid	Output Port	Port Status
Dest0	0,1	Up, Down
Dest1	0,1	Up, Down



OPENSM

	Switch SM	UFM SM	OFED SM
Cost	Free	Not Free	Free
HA	Opensm HA	UFM HA	None HA
Failover Speed	Medium	Fast	Normal
Version	Less	More	Normal
Feature	Old	Newest	Normal
Configuration	Very Difficult	Easy	Difficult

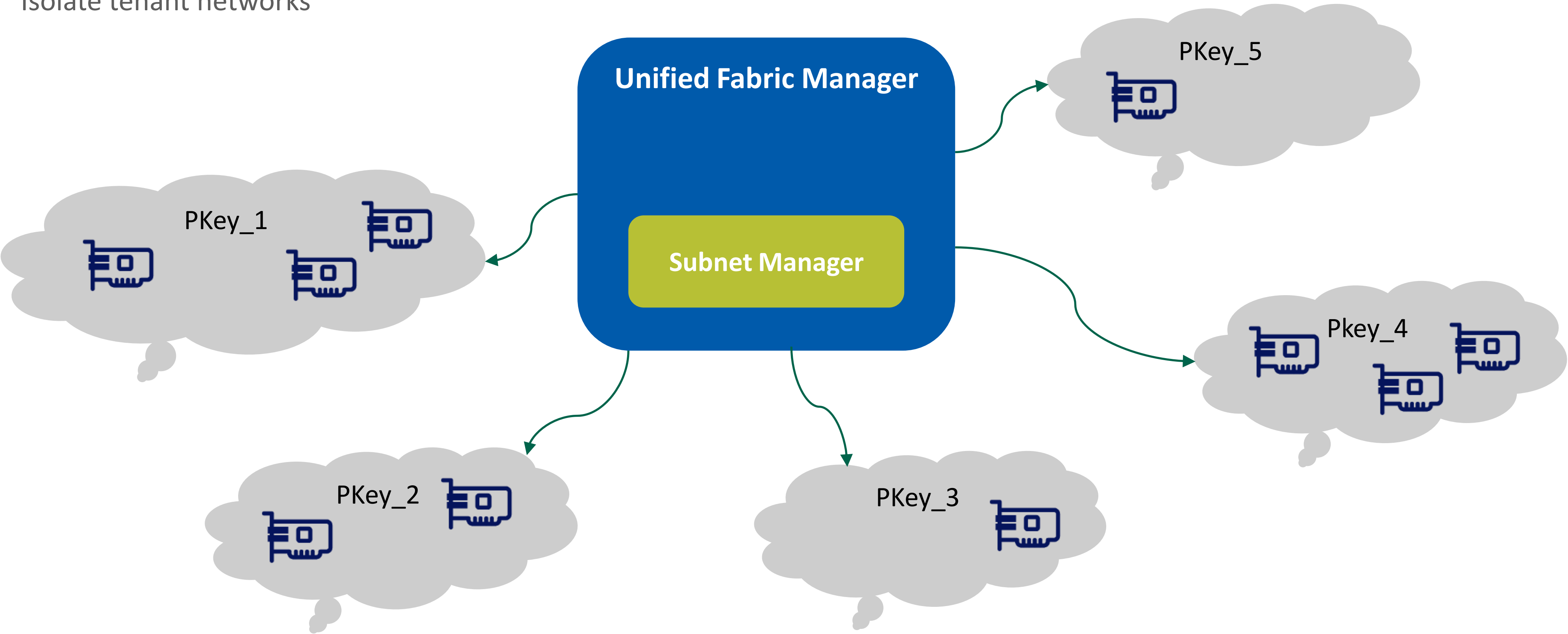


CLOUD NETWORKING API

Allows operators to create/remove/update tenants

Manages PKey GUIDs by getting, adding, and removing GUIDs from PKeys

Isolate tenant networks

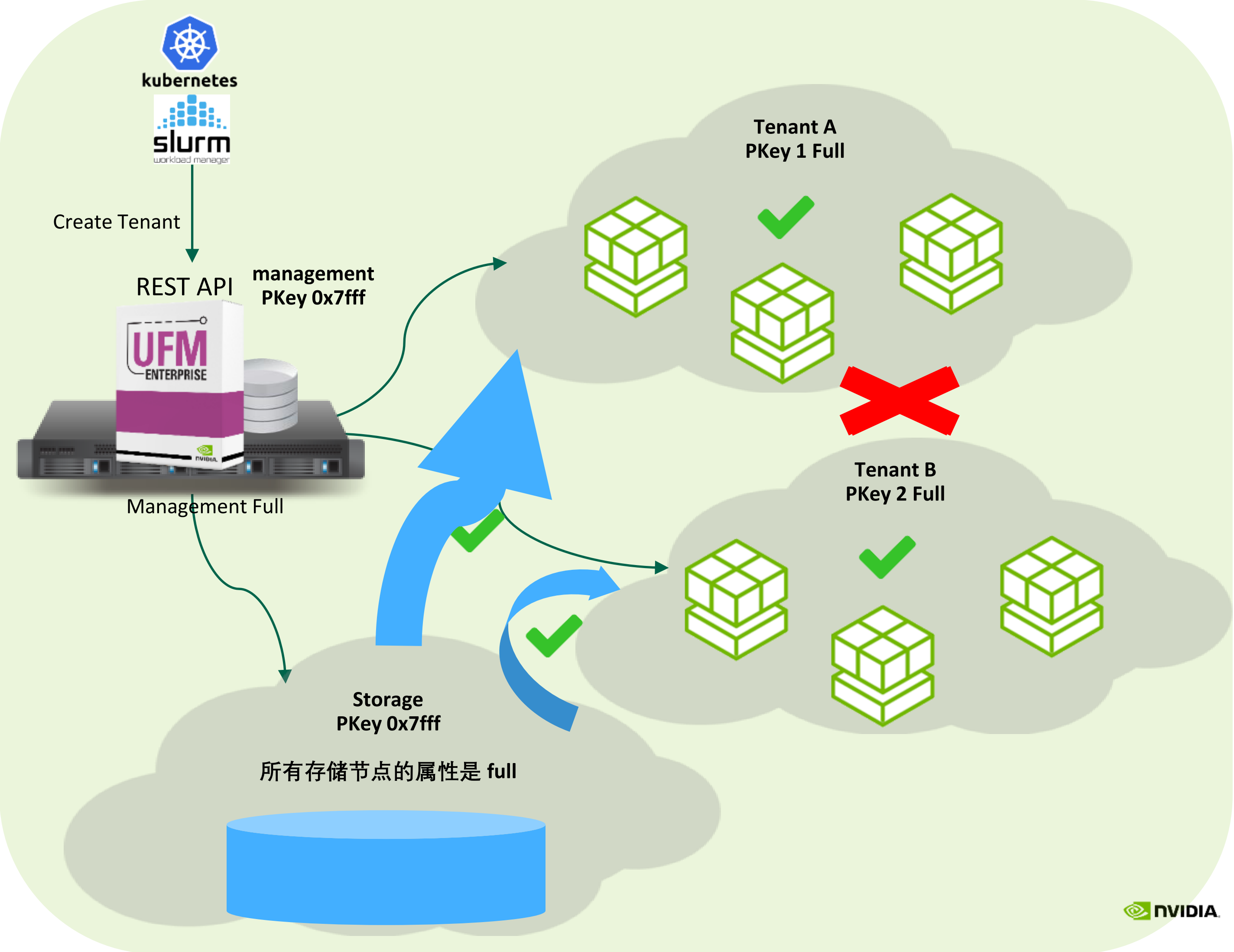


IB NETWORK AUTOMATION & PROVISIONING

- ✓ Default Network
  - ✓ FULL - Only SM and Storage
  - ✓ Limited - All nodes
- ✓ Tenant Network
  - ✓ Full - All nodes

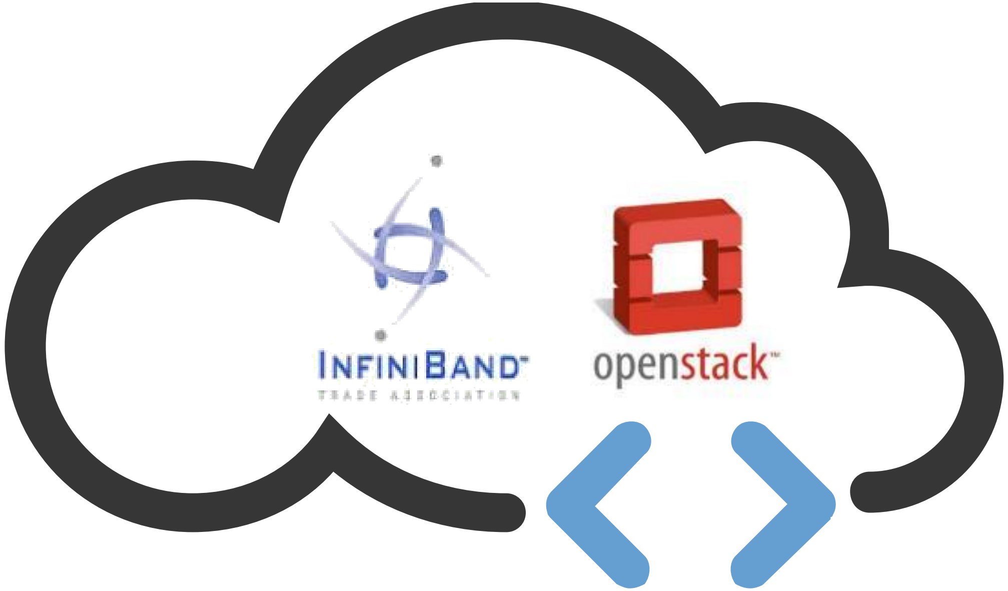
File	Required configuration in UFM

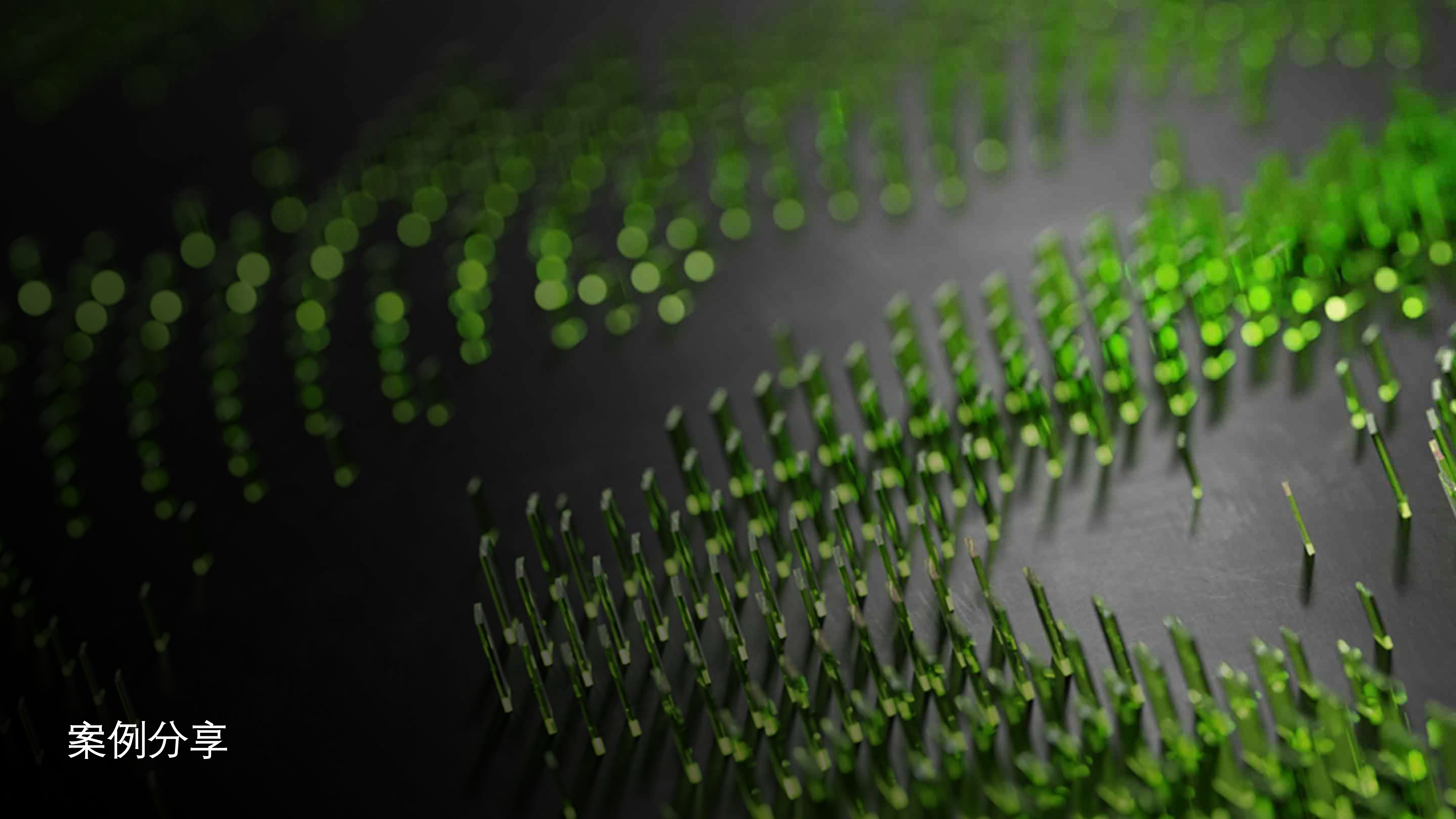
partition.conf	Default=0x7fff, all=limited, self=FULL, IO-Nodes-GUID=FULL
	TenantA=0x8001, all=full
	TenantB=0x8002, all=full
	…



INFINIBAND OPENSTACK – ORCHESTRATING HPC CLOUDS

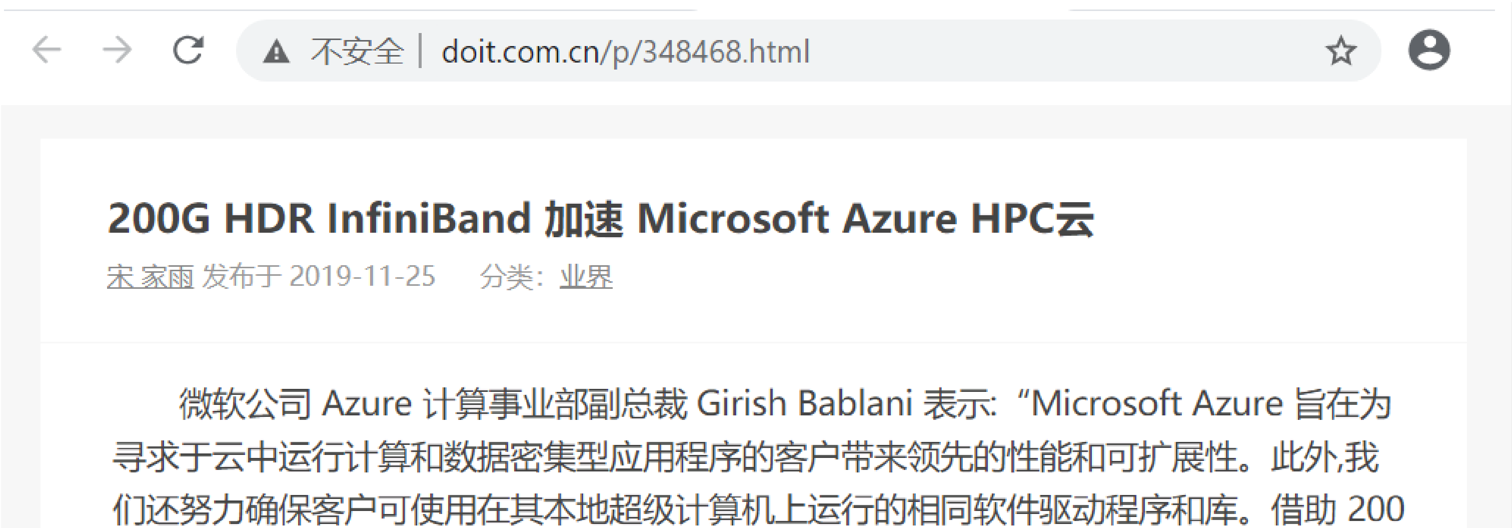
- Native InfiniBand integration into OpenStack
- RDMA-enabled virtual machines
- Network isolation and partition
- Cluster management with UFM appliance
- InfiniBand In-Network Computing
- Accelerate cloud storage NVMe over Fabrics





案例分享

INFINIBAND CLOUD



▲ 不安全 | doit.com.cn/p/348468.html

200G HDR InfiniBand 加速 Microsoft Azure HPC云

宋家雨 发布于 2019-11-25 分类：业界

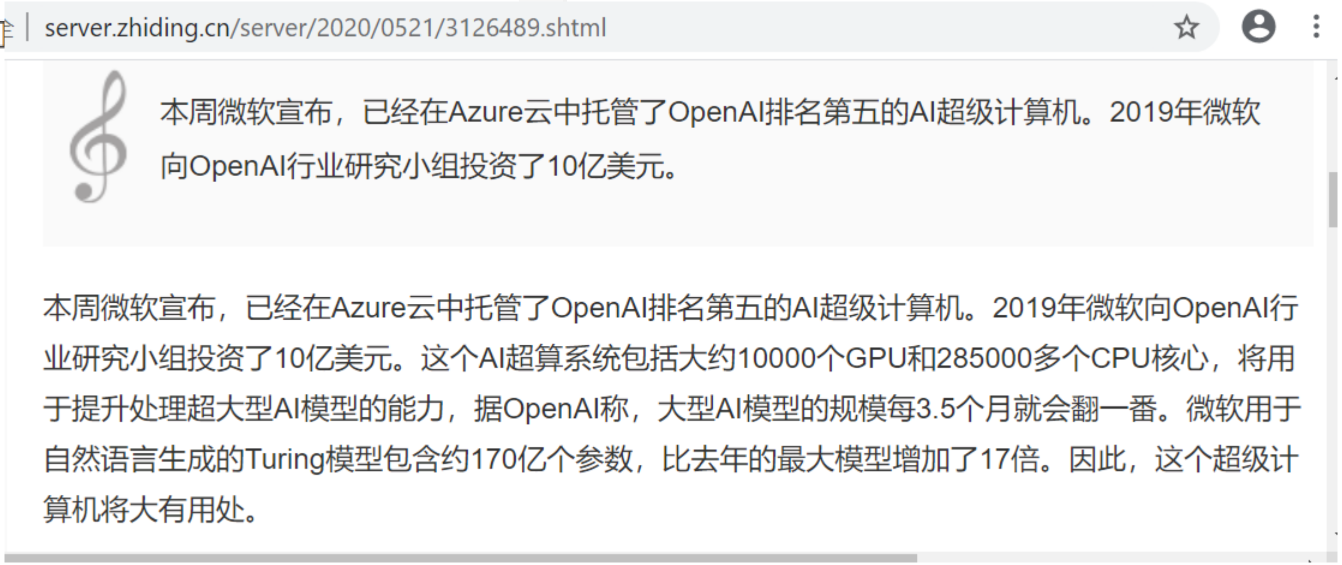
微软公司 Azure 计算事业部副总裁 Girish Bablani 表示：“Microsoft Azure 旨在为寻求于云中运行计算和数据密集型应用程序的客户带来领先的性能和可扩展性。此外,我们还努力确保客户可使用在其本地超级计算机上运行的相同软件驱动程序和库。借助 200 GB HDR InfiniBand,我们能够为真实的 HPC 和人工智能工
metal 超级计算机相媲美的可扩展性与性能。”

Mellanox Quantum

SHARP 2

OpenAI

Microsoft Azure



server.zhiding.cn/server/2020/0521/3126489.shtml

> 本周微软宣布，已经在Azure云中托管了OpenAI排名第五的AI超级计算机。2019年微软向OpenAI行业研究小组投资了10亿美元。

本周微软宣布，已经在Azure云中托管了OpenAI排名第五的AI超级计算机。2019年微软向OpenAI行业研究小组投资了10亿美元。这个AI超算系统包括大约10000个GPU和285000多个CPU核心，将用于提升处理超大型AI模型的能力，据OpenAI称，大型AI模型的规模每3.5个月就会翻一番。微软用于自然语言生成的Turing模型包含约170亿个参数，比去年的最大模型增加了17倍。因此，这个超级计算机将大有用处。

Achieve more with Azure HPC

Purpose-built HPC	A full range of CPU and GPU capabilities that help applications scale to 80K+ cores
Fast, secure networking	Fast InfiniBand inter-connects as well as edge-to-cloud connectivity
High performing storage	A range of storage capabilities to support simple-to-complex storage needs
Workload orchestration	End-to-end workflow agility using known, familiar tools and processes
Intelligence services	AI, machine learning, and deep learning at supercomputer scale

ANSYS FLUENT ON HBV2

- **App:** ANSYS Fluent
- **Version:** 14.06.004
- **Model:** External Flow over a Formula-1 Race Car (f1_racecar_140m)
- **Configuration Details:** 60 MPI ranks were run (2 out of 4 cores per NUMA) in each HBv2 VM in order to leave nominal resources to run Linux background processes and give ~6 GB/s of memory bandwidth per core. In addition, Adaptive Routing was enabled and DCT (Dynamic Connected Transport) was used as the transport layer, while **HPC-X version 2.50 (UCX v1.6) was used for MPI**. Azure CentOS HPC 7.6 image was used from https://github.com/Azure/azhpc-images
- **Summary:** HBv2 VMs scale super linearly (112%) up to the top end measured number of VMs (128). The Fluent Solver Rating measured at this top-end level of scale is 83% more performance than the current leader submission on ANSYS public database for this model (https://bit.ly/2OdAExM).

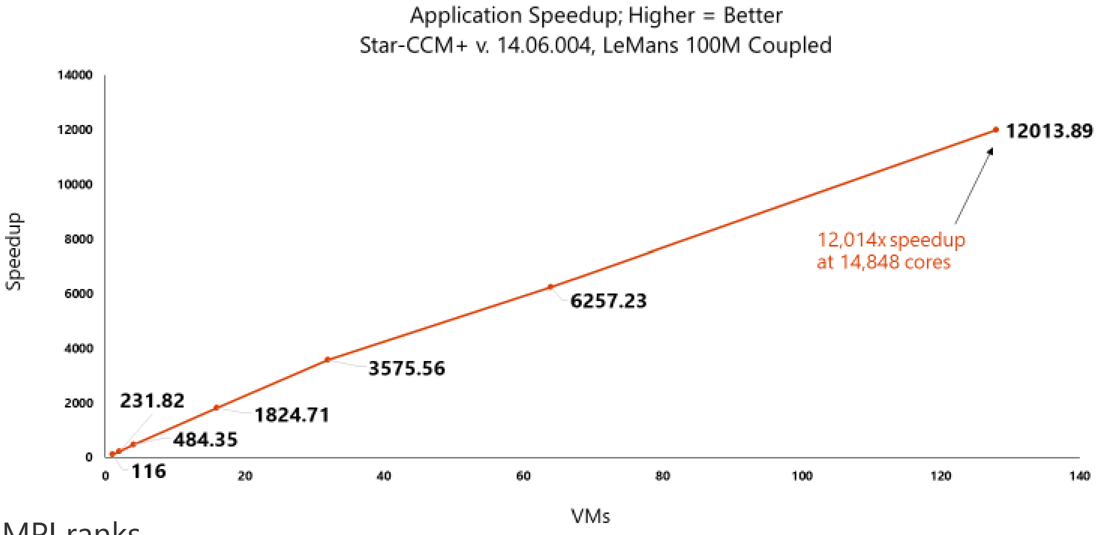
Microsoft Azure

STAR-CCM + ON HBV2

App: Siemens Star-CCM+
Version: 14.06.004
Model: LeMans 100M Coupled Solver
Configuration Details: 116 MPI ranks were run (4 ranks from each of 29 NUMA) in each HBv2 VM in order to leave nominal resources to run Linux background processes. In addition, Adaptive Routing was enabled and DCT (Dynamic Connected Transport) was used as the transport layer, while HPC-X version 2.50 (UCX v1.6) was used for MPI. Azure CentOS HPC 7.6 image was used from https://github.com/Azure/azhpc-images

Microsoft

Microsoft Azure

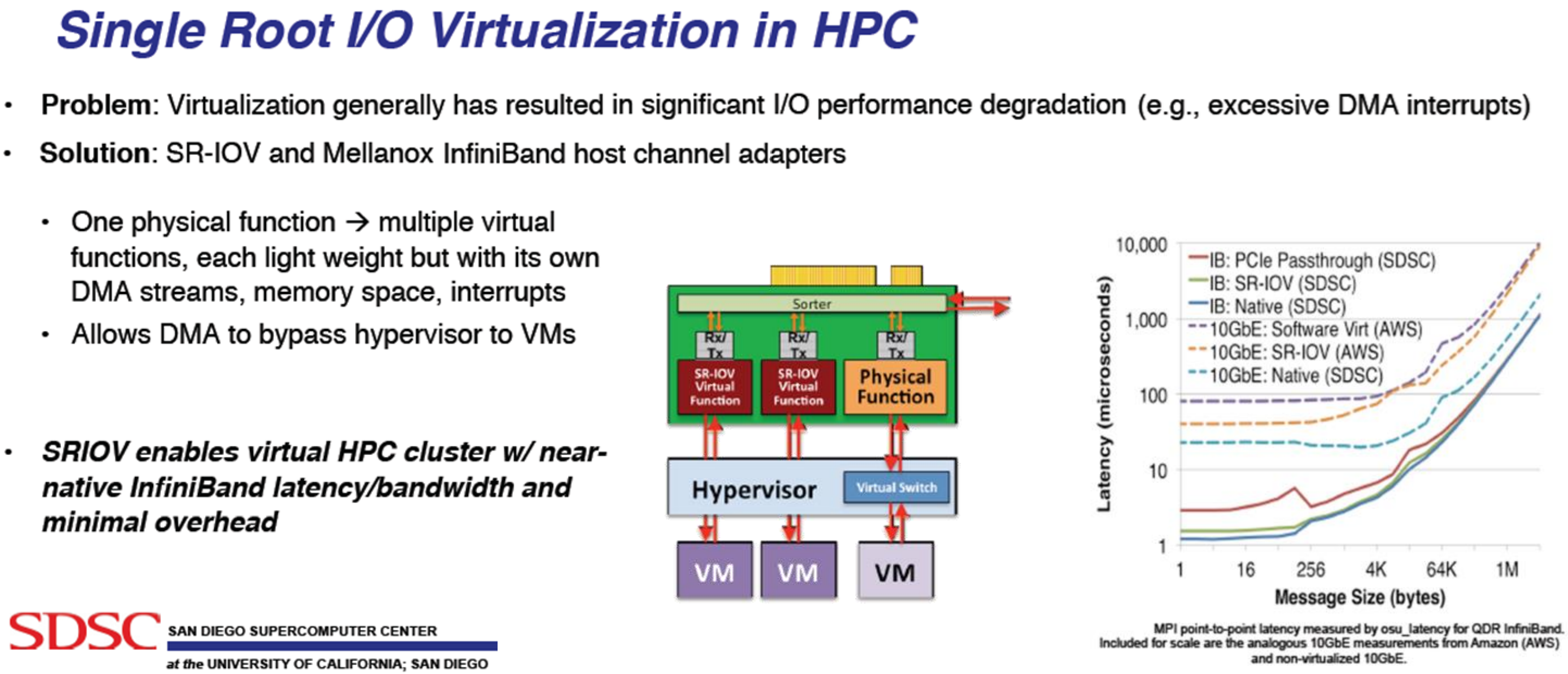

Application Speedup; Higher = Better
Star-CCM+ v. 14.06.004, LeMans 100M Coupled

12,014x speedup
at 14,848 cores

Summary: Star-CCM+ was scaled at 81% efficiency to nearly 15,000 MPI ranks delivering an application speedup of more than 12,000x. This compares favorably to Azure’s previous best of more than 11,500 MPI ranks, which itself was a world-record for MPI scalability on the public cloud.

NVIDIA

VIRTUALIZATION FOR HPC: INFINIBAND SRIOV



Single Root I/O Virtualization in HPC

- **Problem**: Virtualization generally has resulted in significant I/O performance degradation (e.g., excessive DMA interrupts)
- **Solution**: SR-IOV and Mellanox InfiniBand host channel adapters
  - One physical function → multiple virtual functions, each light weight but with its own DMA streams, memory space, interrupts
  - Allows DMA to bypass hypervisor to VMs
- ***SRIOV enables virtual HPC cluster w/ near-native InfiniBand latency/bandwidth and minimal overhead***

MPI point-to-point latency measured by osu_latency for QDR InfiniBand. Included for scale are the analogous 10GbE measurements from Amazon (AWS) and non-virtualized 10GbE.

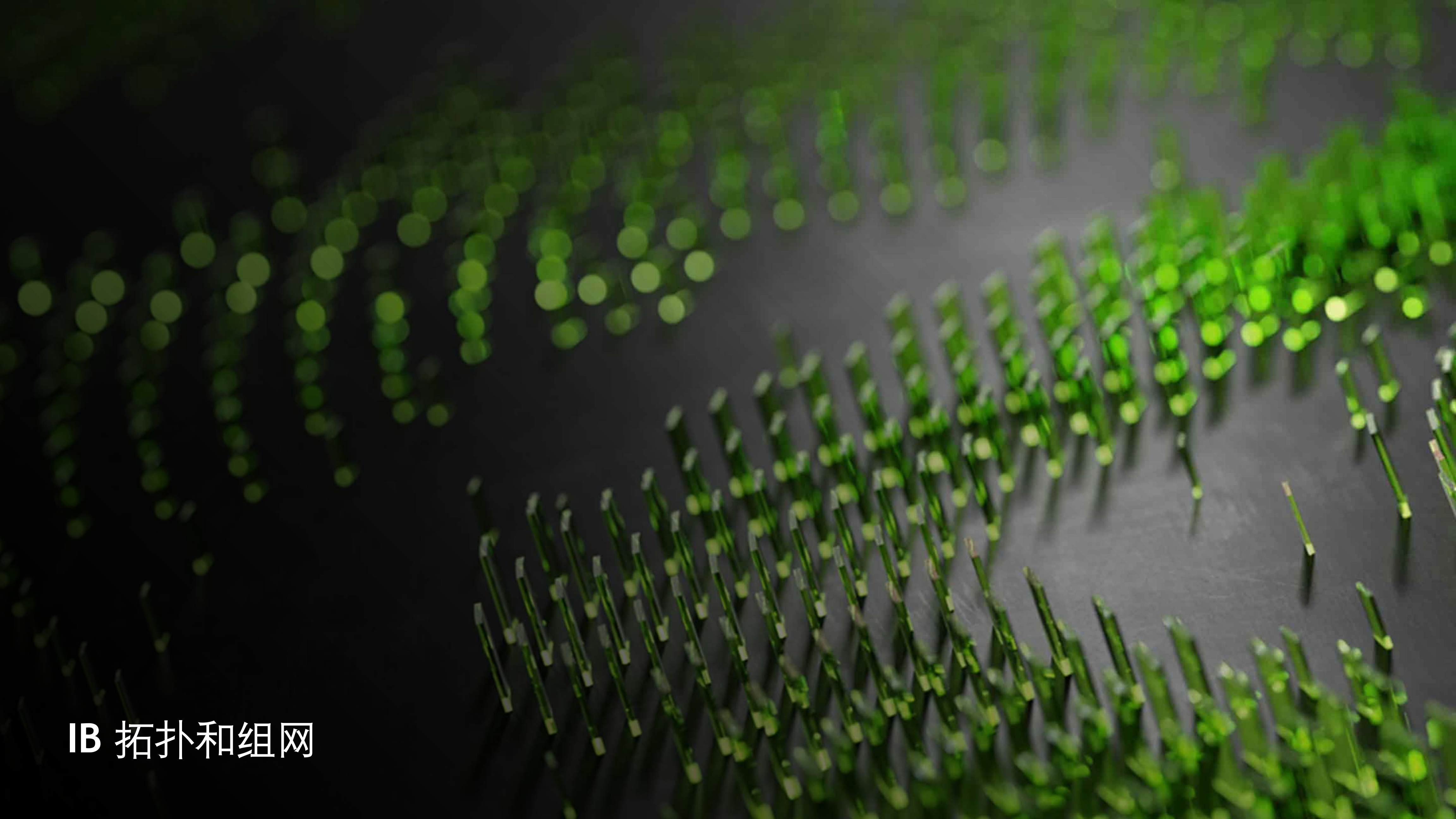
SDSC SAN DIEGO SUPERCOMPUTER CENTER
at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

UNIVERSITY OF CAMBRIDGE: HPC CLOUD CONVERGENCE

- Motivation behind OpenStack Research Cloud
  - Make computing, data, applications and workflows more accessible, flexible and secure
  - Allow a wide range of services to be delivered from a single framework
  - Make research computing easier to use, easier to share, decrease the time to science and increasing innovation and research

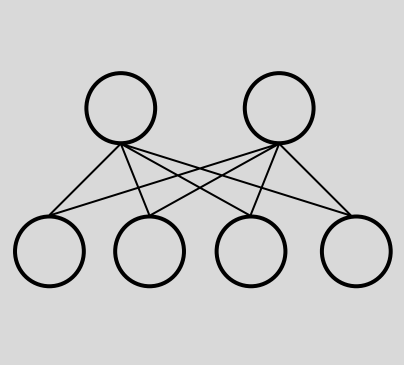
- Use cases
  - Research computing as a Service
  - HPC as a Service
  - HPDA as a Service

- Research Cloud Network Requirement
  - SRIOV and RDMA essential for HPC
  - Network virtualization with no compromise
  - High-performance data I/O

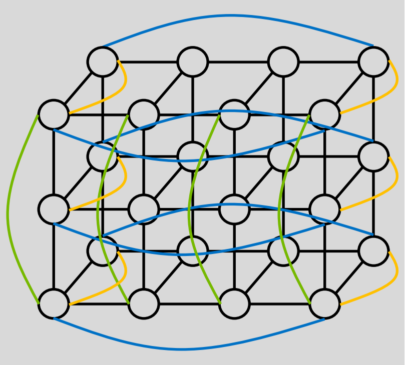


IB 拓扑和组网

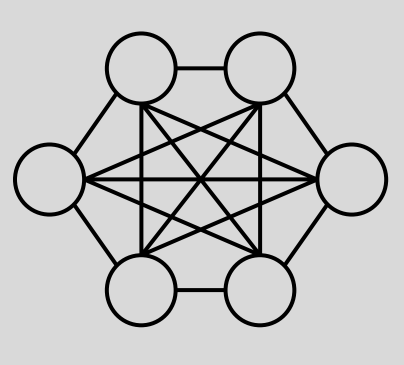
SUPPORTING VARIETY OF TOPOLOGIES



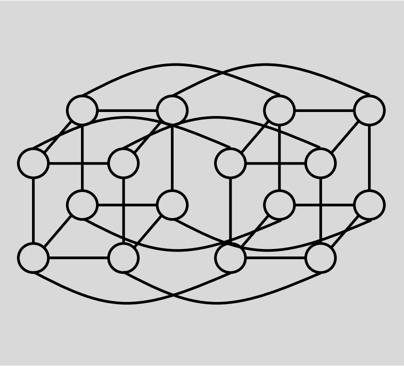
Fat Tree



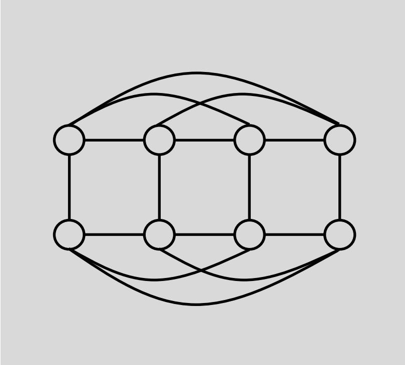
Torus



Dragonfly

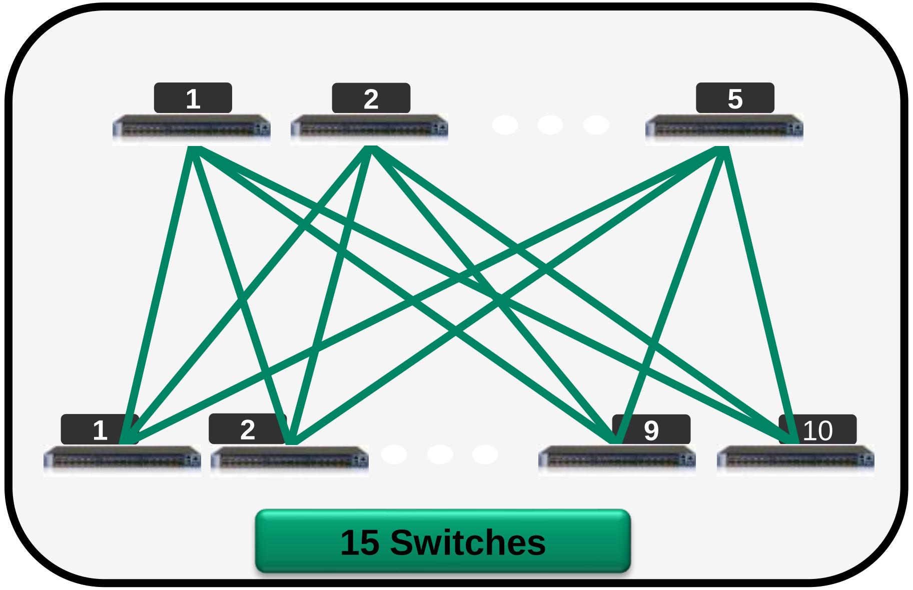


Hypercube

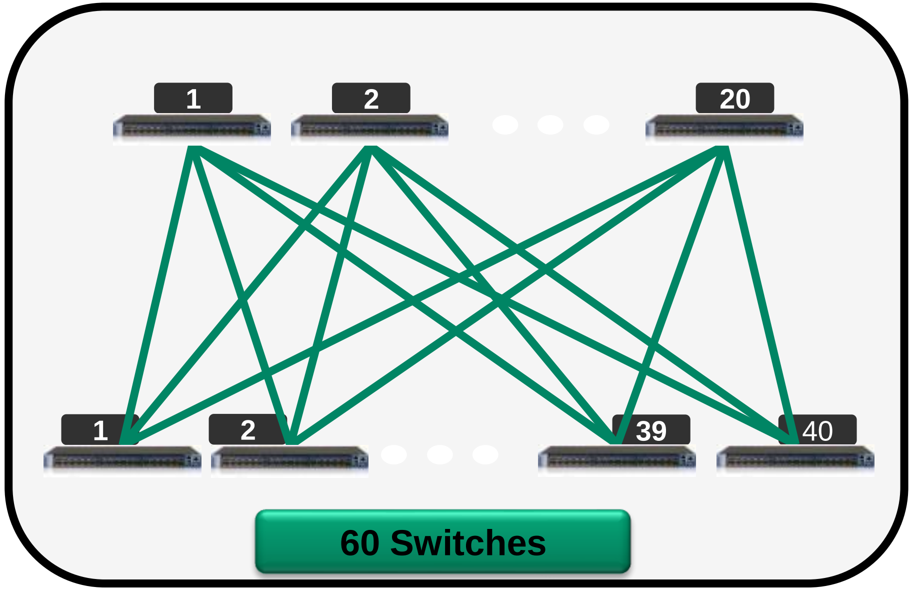


HyperX

200-Node 200G InfiniBand Platform



800-Node 200G InfiniBand Platform

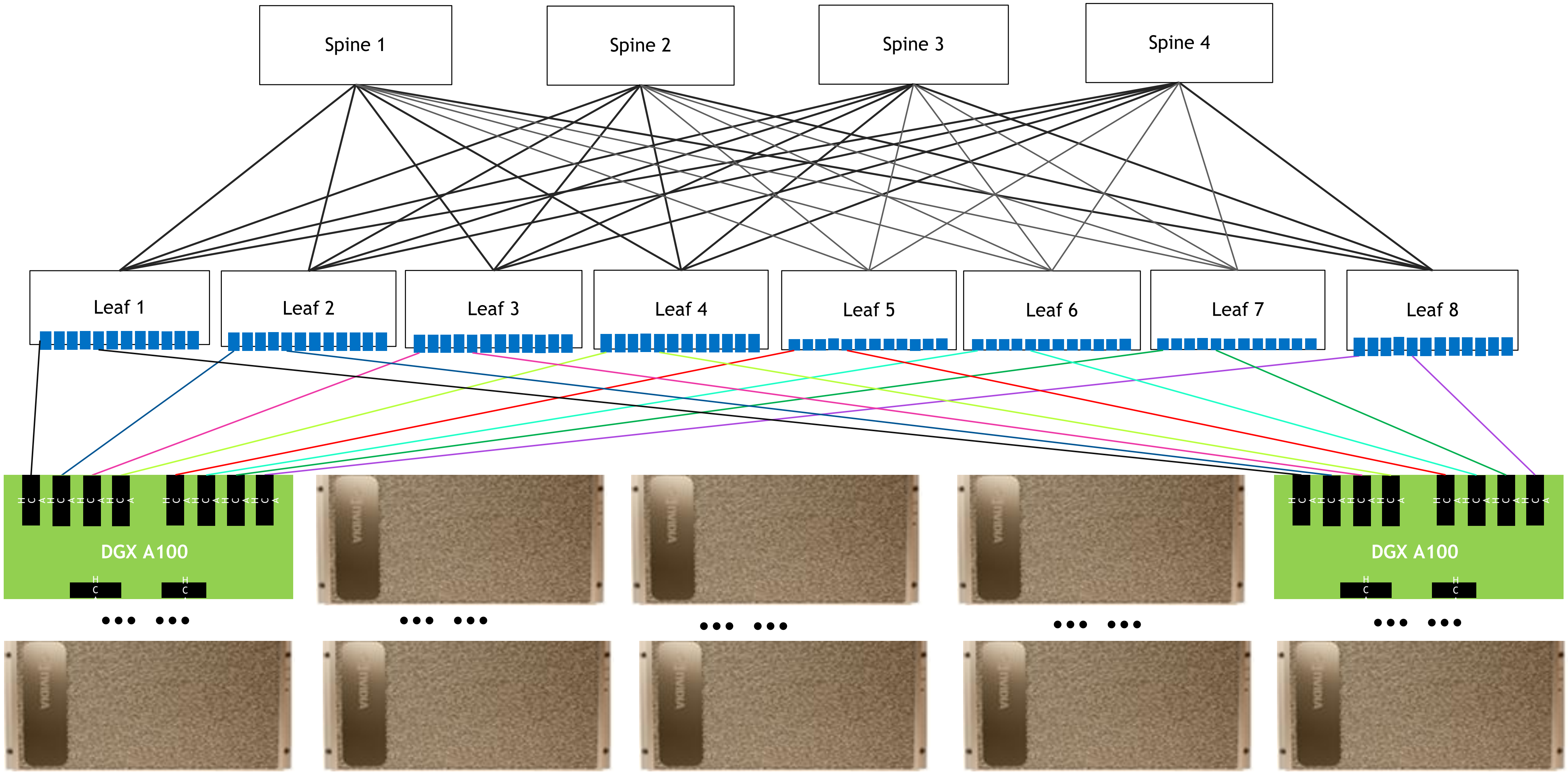


Quantum
The smartest switch, became smarter

ConnectX-6

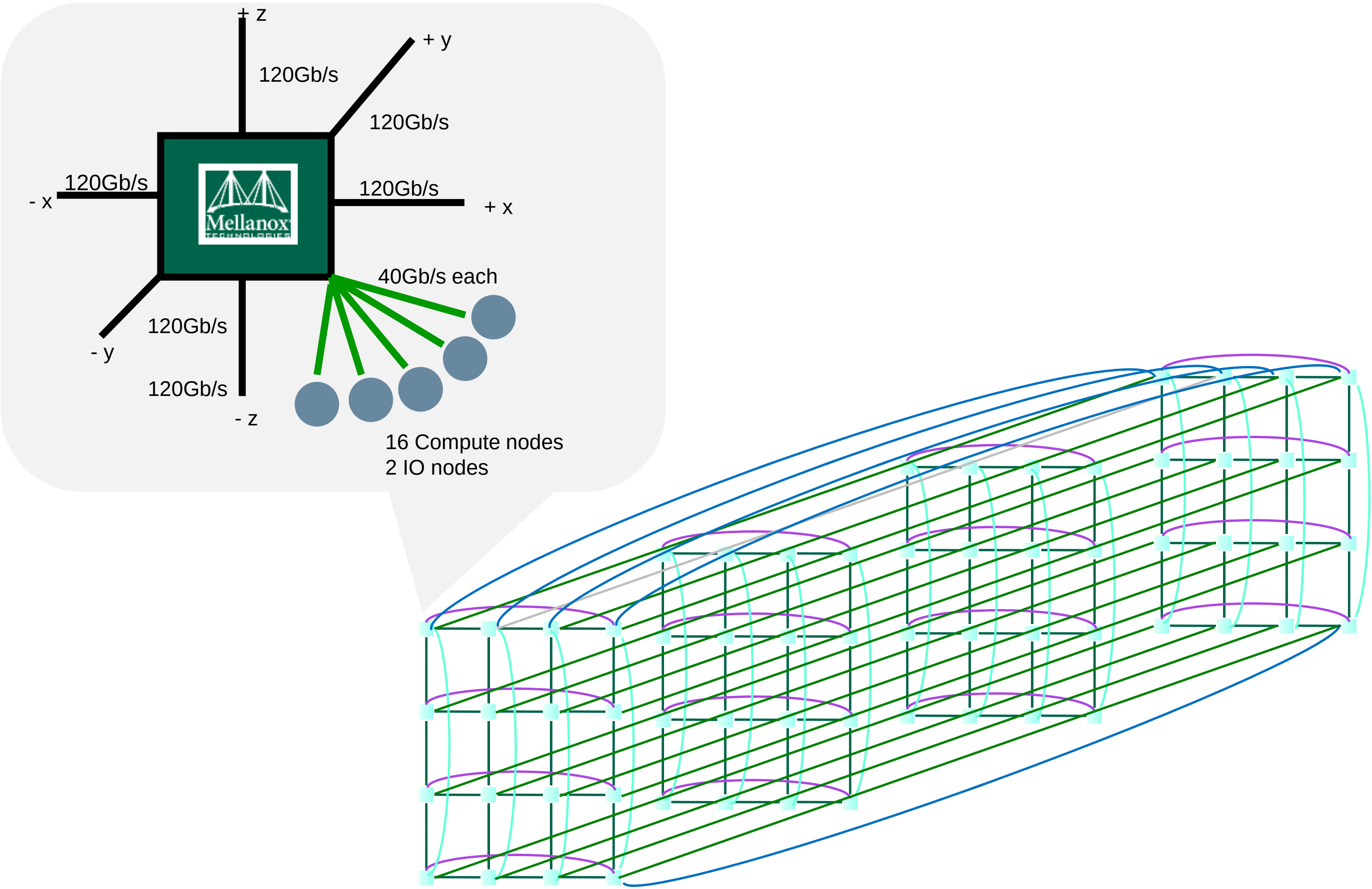
NVIDIA

SUPERPOD SCALABLE UNIT(SU)

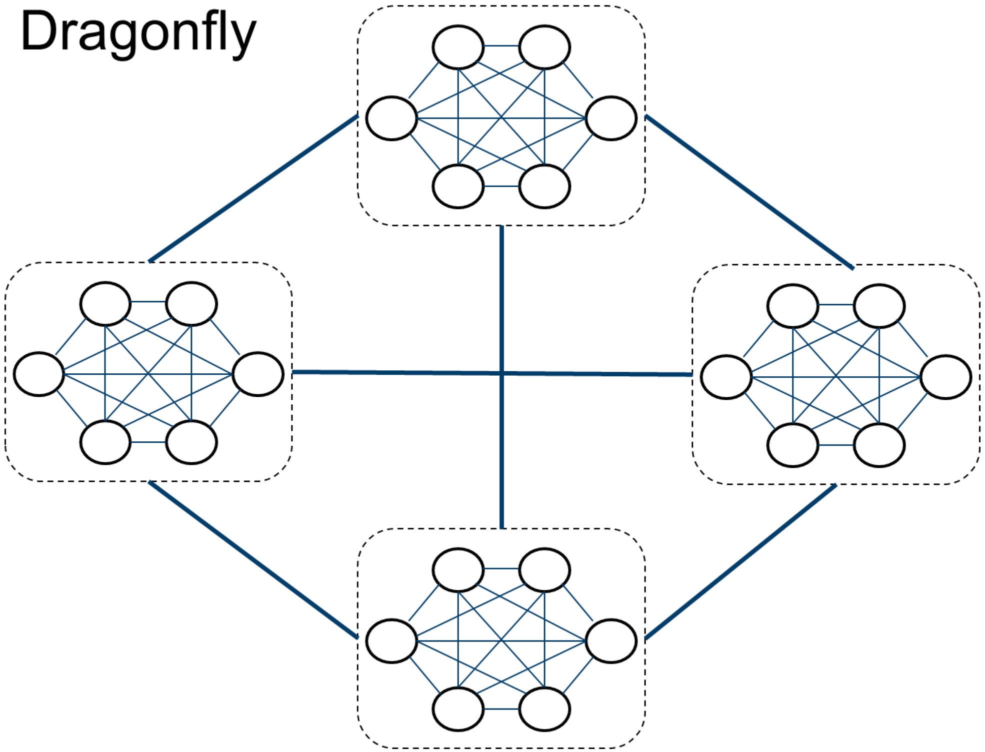


20 nodes per Scalable Unit

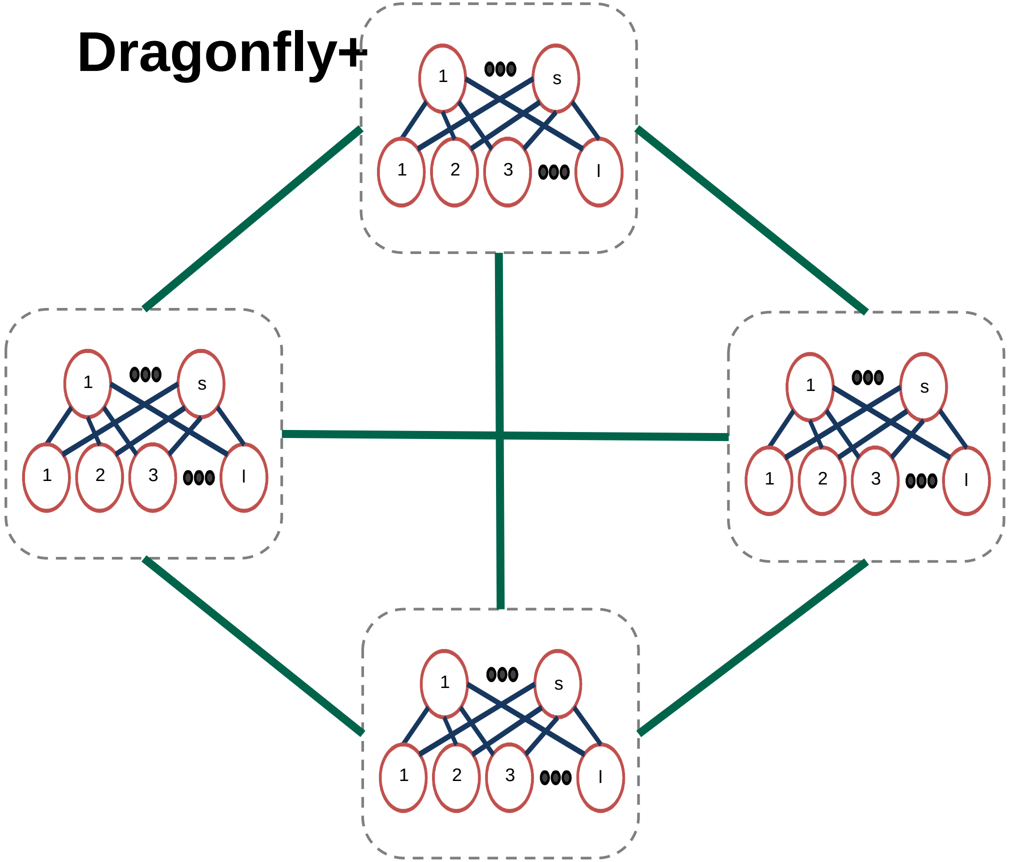
Example 3D Torus – SDSC Gordon



TRADITIONAL DRAGONFLY VS DRAGONFLY+



Dragonfly



Dragonfly+

DRAGONFLY+ TOPOLOGY

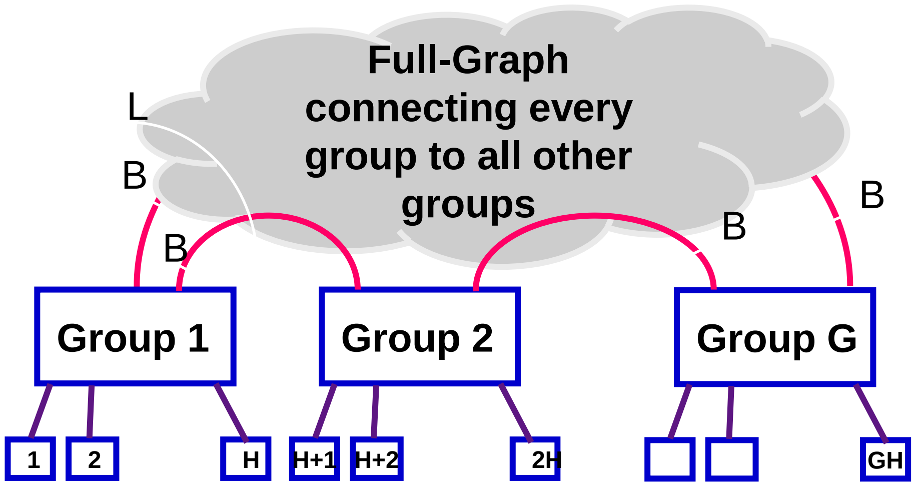
Several “groups”, connected using all to all links

The topology inside each group can be any topology

Reduce total cost of network (fewer long cables)

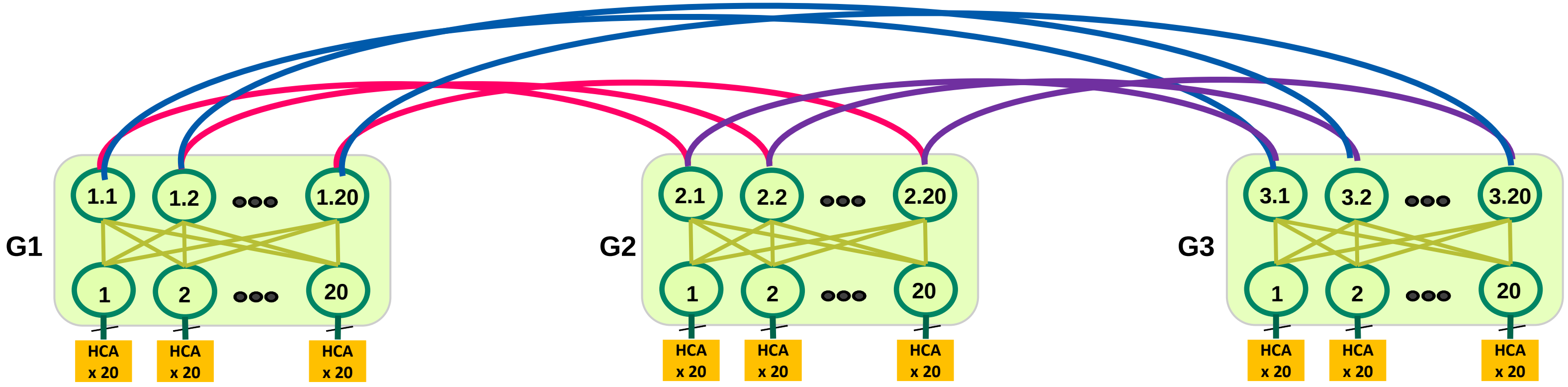
Utilizes Adaptive Routing to for efficient operations

Simplifies future system expansion



Full-Graph connecting every group to all other groups

1200-Nodes Dragonfly+ Systems Example

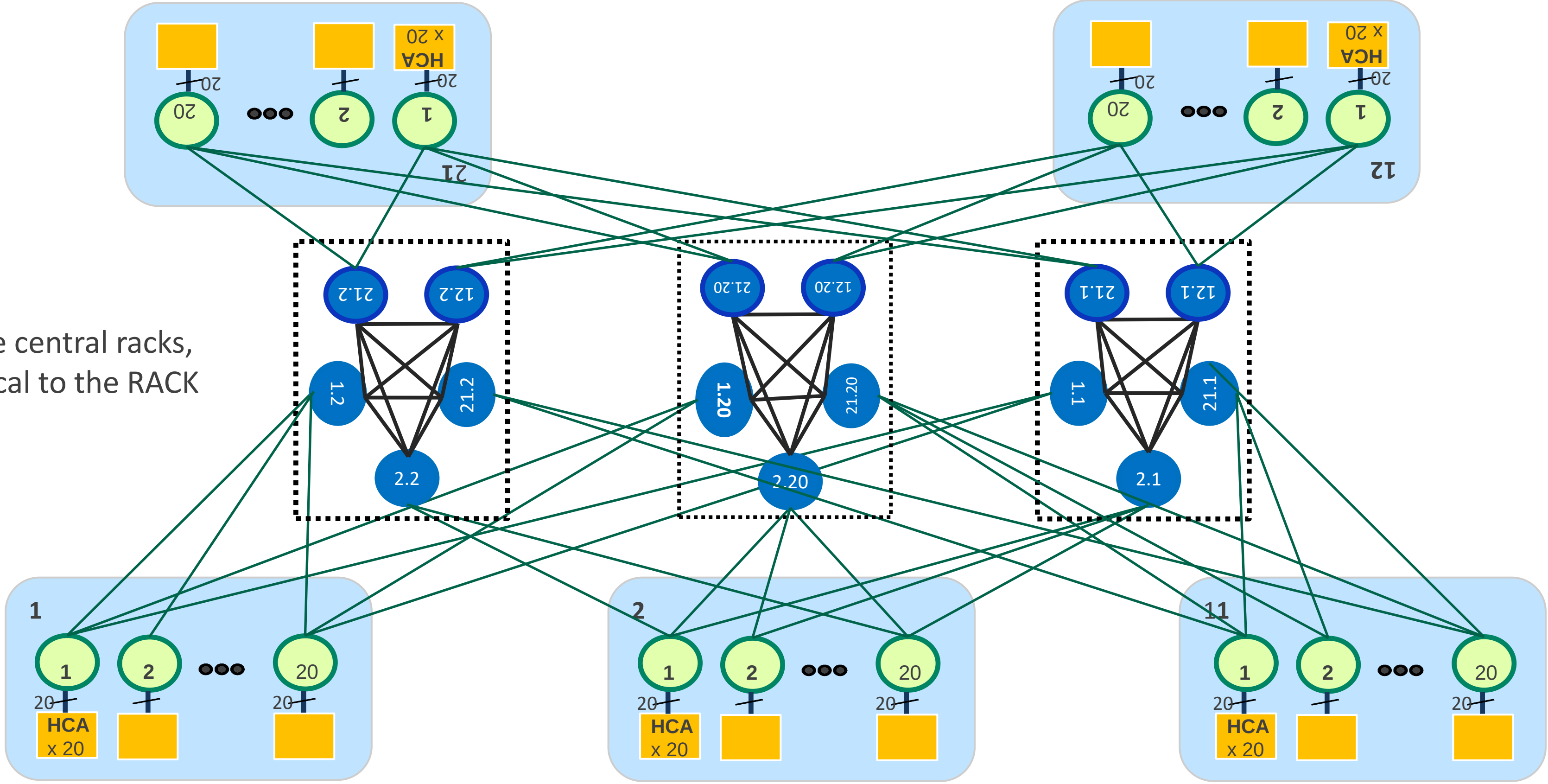


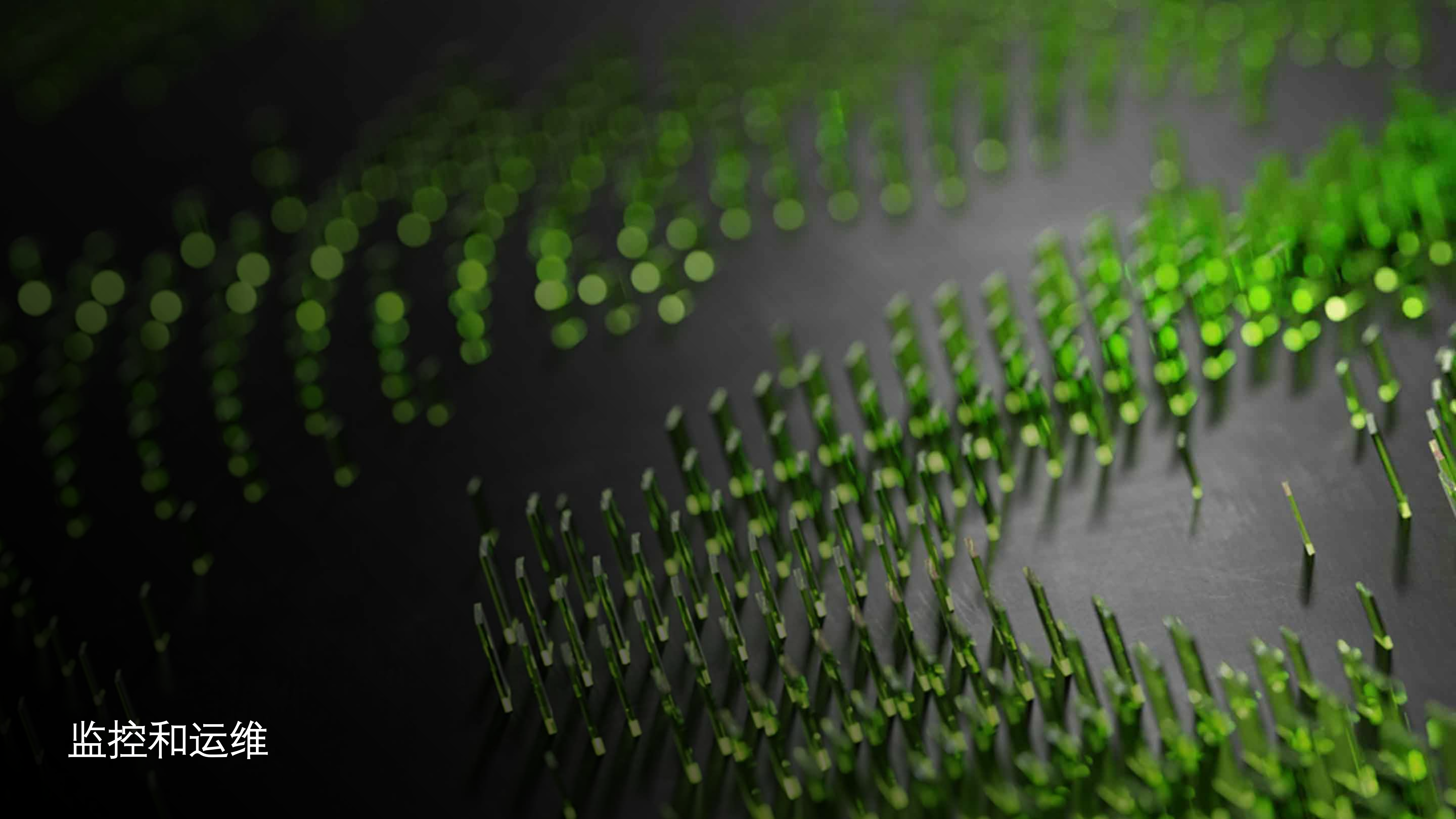
FUTURE EXPANSION OF DRAGONFLY+ BASED SYSTEM

Phase 2:
+10x400 =
8400 hosts

Re-cable the central racks,
a change local to the RACK

Phase 1:
11x400 =
4400 hosts





监控和运维

UFM PLATFORMS PORTFOLIO

UFM Telemetry

Real-Time Monitoring

UFM Enterprise

Management, Monitoring & Orchestration

(UFM Enterprise includes UFM Telemetry)

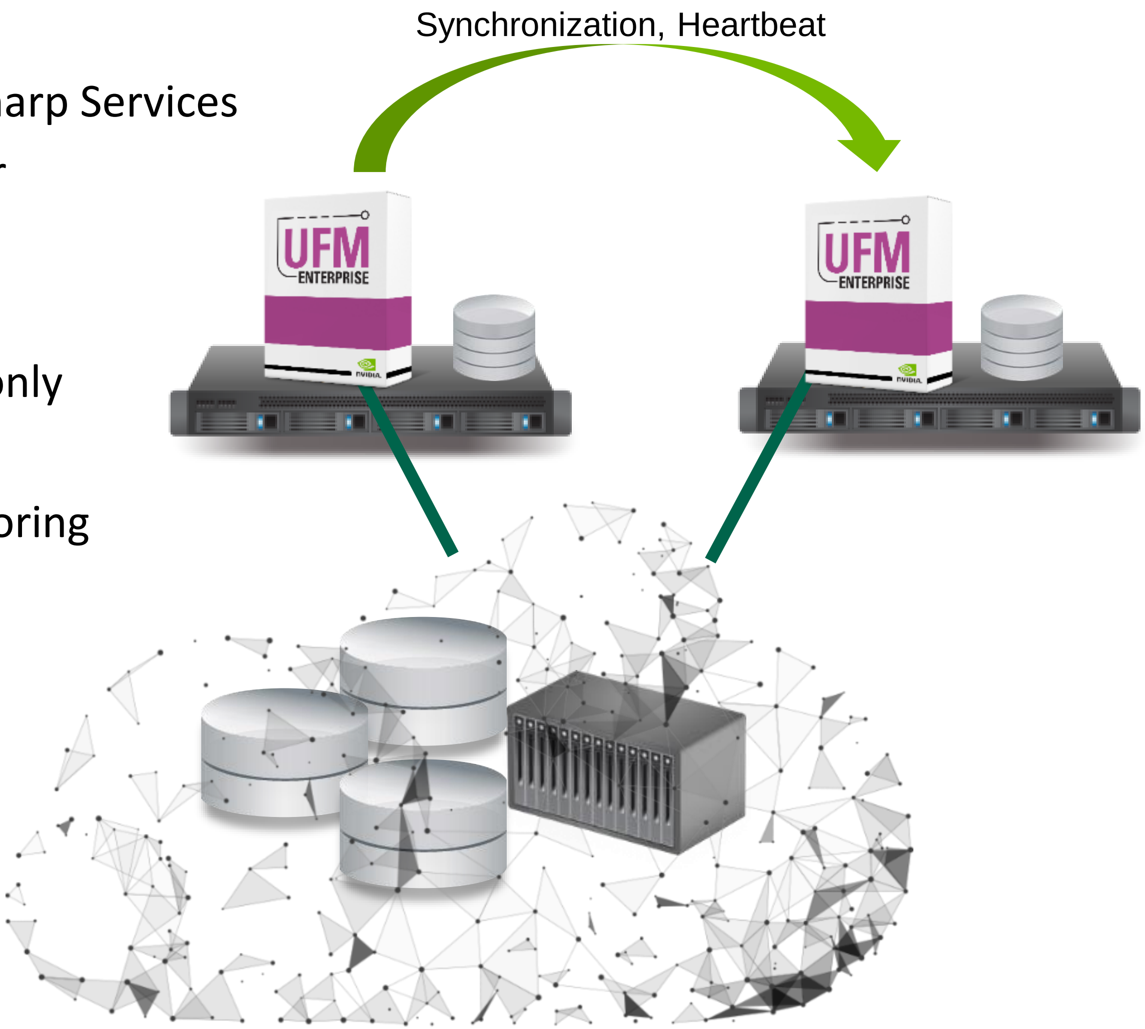
UFM Cyber-AI

Cyber Intelligence and Analytics

(UFM Cyber-AI includes UFM Enterprise)

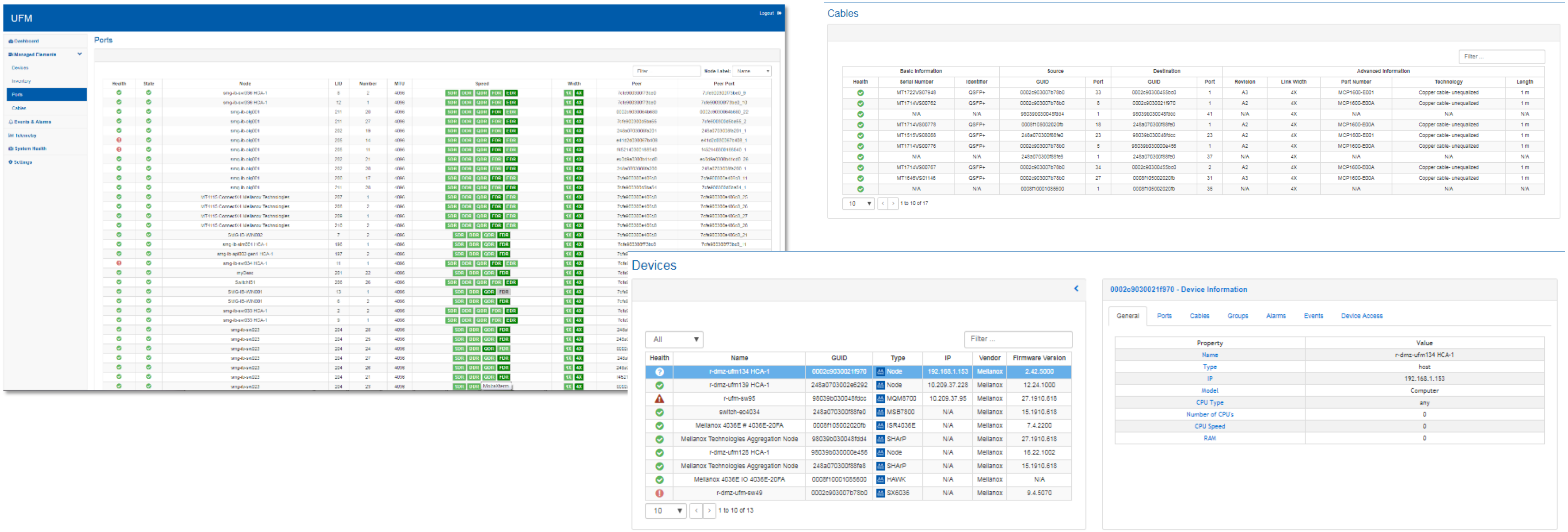
UFM IN THE FABRIC

- Manages Subnet Manager and Sharp Services
- Software or appliance form factor
- High availability - 2 or more
- Switch and adapter management
- Full management or monitoring only
- Layer 2 level monitoring
- REST API for configuration/monitoring
- Single Interface for all network



CENTRALIZE DEVICE MANAGEMENT

- Manage an inventory of assets, switches and nodes
- Centrally upgrade firmware and software across all managed and unmanaged systems



Cables

Filter ...

Basic Information			Source		Destination		Advanced Information				
Health	Serial Number	Identifier	GUID	Port	GUID	Port	Revision	Link Width	Part Number	Technology	Length
✓	MT1722VS07948	QSFP+	0002c903007b78b0	33	0002c90300455bc0	1	A3	4X	MCP1600-E001	Copper cable- unequalized	1 m
✓	MT1714VS00762	QSFP+	0002c903007b78b0	8	0002c9030021f970	1	A2	4X	MCP1600-E00A	Copper cable- unequalized	1 m
✓	N/A	N/A	98039b030048fdd4	1	98039b030048fdcc	41	N/A	4X	N/A	N/A	N/A
✓	MT1714VS00778	QSFP+	0008f105002020fb	18	248a070300f88fe0	1	A2	4X	MCP1600-E00A	Copper cable- unequalized	1 m
✓	MT1515VS08068	QSFP+	248a070300f88fe0	23	98039b030048fdcc	23	A2	4X	MCP1600-E001	Copper cable- unequalized	1 m
✓	MT1714VS00776	QSFP+	0002c903007b78b0	5	98039b030000e456	1	A2	4X	MCP1600-E00A	Copper cable- unequalized	1 m
✓	N/A	N/A	248a070300f88fe8	1	248a070300f88fe0	37	N/A	4X	N/A	N/A	N/A
✓	MT1714VS00767	QSFP+	0002c903007b78b0	34	0002c90300455bc0	2	A2	4X	MCP1600-E00A	Copper cable- unequalized	1 m
✓	MT1646VS01146	QSFP+	0002c903007b78b0	27	0008f105002020fb	31	A3	4X	MCP1600-E00A	Copper cable- unequalized	1 m
✓	N/A	N/A	0008f10001085600	1	0008f105002020fb	35	N/A	4X	N/A	N/A	N/A

10 ▼ ‹ › 1 to 10 of 17

Devices

All ▼ Filter ...

Health	Name	GUID	Type	IP	Vendor	Firmware Version
?	r-dmz-ufm134 HCA-1	0002c9030021f970	Node	192.168.1.153	Mellanox	2.42.5000
✓	r-dmz-ufm139 HCA-1	248a0703002e6292	Node	10.209.37.228	Mellanox	12.24.1000
⚠	r-ufm-sw95	98039b030048fdcc	MQM8700	10.209.37.95	Mellanox	27.1910.618
✓	switch-ec4034	248a070300f88fe0	MSB7800	N/A	Mellanox	15.1910.618
✓	Mellanox 4036E # 4036E-20FA	0008f105002020fb	ISR4036E	N/A	Mellanox	7.4.2200
✓	Mellanox Technologies Aggregation Node	98039b030048fdd4	SHArP	N/A	Mellanox	27.1910.618
✓	r-dmz-ufm128 HCA-1	98039b030000e456	Node	N/A	Mellanox	16.22.1002
✓	Mellanox Technologies Aggregation Node	248a070300f88fe8	SHArP	N/A	Mellanox	15.1910.618
✓	Mellanox 4036E IO 4036E-20FA	0008f10001085600	HAWK	N/A	Mellanox	N/A
!	r-dmz-ufm-sw49	0002c903007b78b0	SX6036	N/A	Mellanox	9.4.5070

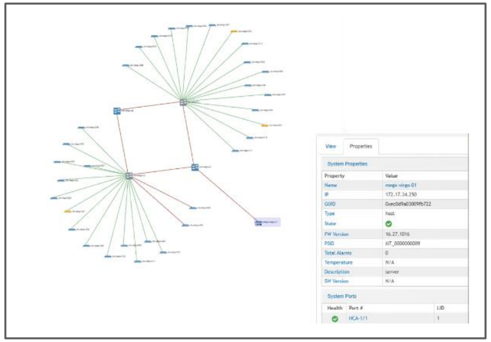
10 ▼ ‹ › 1 to 10 of 13

0002c9030021f970 - Device Information

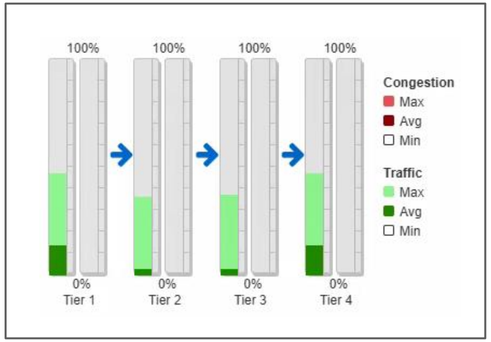
General | Ports | Cables | Groups | Alarms | Events | Device Access

Property	Value
Name	r-dmz-ufm134 HCA-1
Type	host
IP	192.168.1.153
Model	Computer
CPU Type	any
Number of CPU's	0
CPU Speed	0
RAM	0

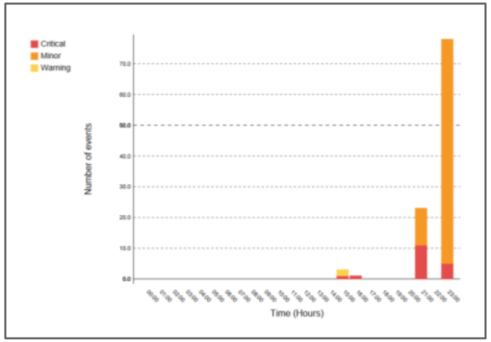
UFM DASHBOARD



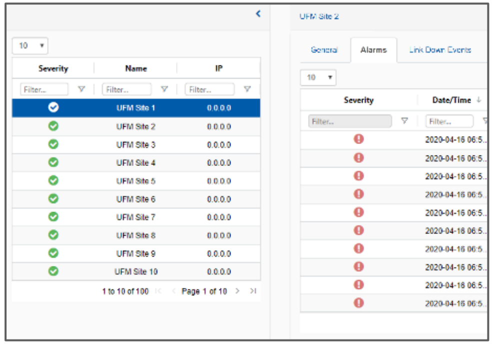
Network Validation



Congestion Mapping



Health Reports



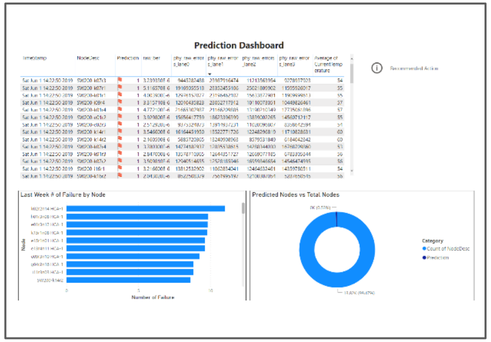
Severity	Name	IP
✓	UFM Site 1	0.0.0.0
✓	UFM Site 2	0.0.0.0
✓	UFM Site 3	0.0.0.0
✓	UFM Site 4	0.0.0.0
✓	UFM Site 5	0.0.0.0
✓	UFM Site 6	0.0.0.0
✓	UFM Site 7	0.0.0.0
✓	UFM Site 8	0.0.0.0
✓	UFM Site 9	0.0.0.0
✓	UFM Site 10	0.0.0.0

1 to 10 of 100 Page 1 of 10

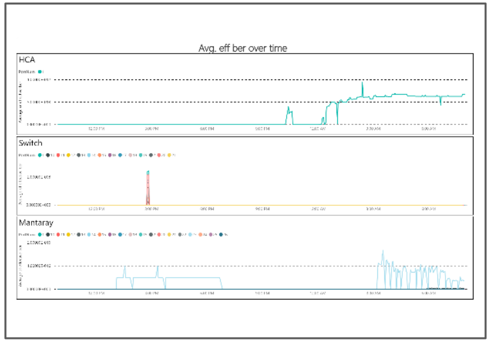
UFM Site 2

Severity	Date/Time
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...
!	2020-04-16 06:5...

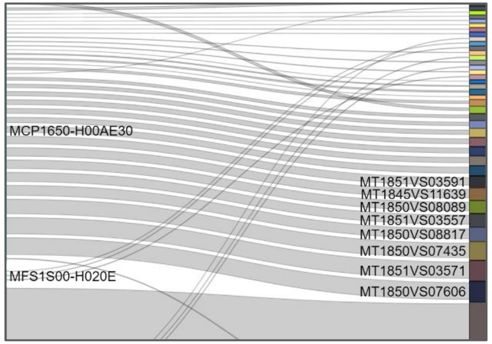
Inventory Mapping



Prediction Dashboard



Real-Time Analysis

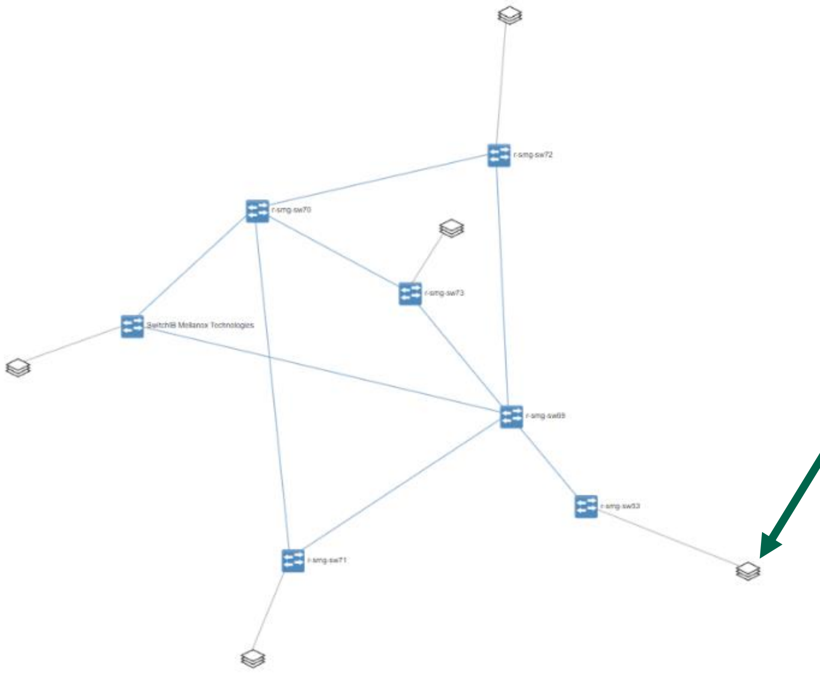


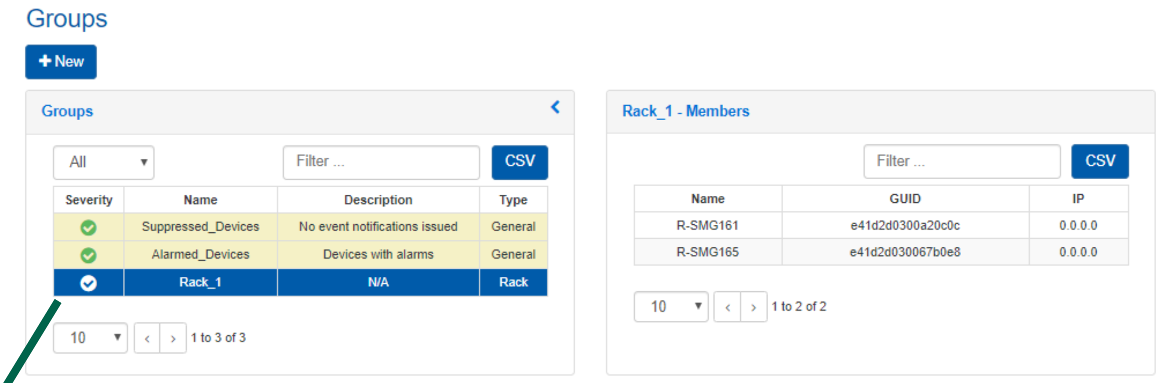
Performance Monitoring

Secure Cable Management

CENTRALIZE DEVICE MANAGEMENT

- Manage an inventory of assets, switches and nodes
- Centrally upgrade firmware and software across all managed and unmanaged systems



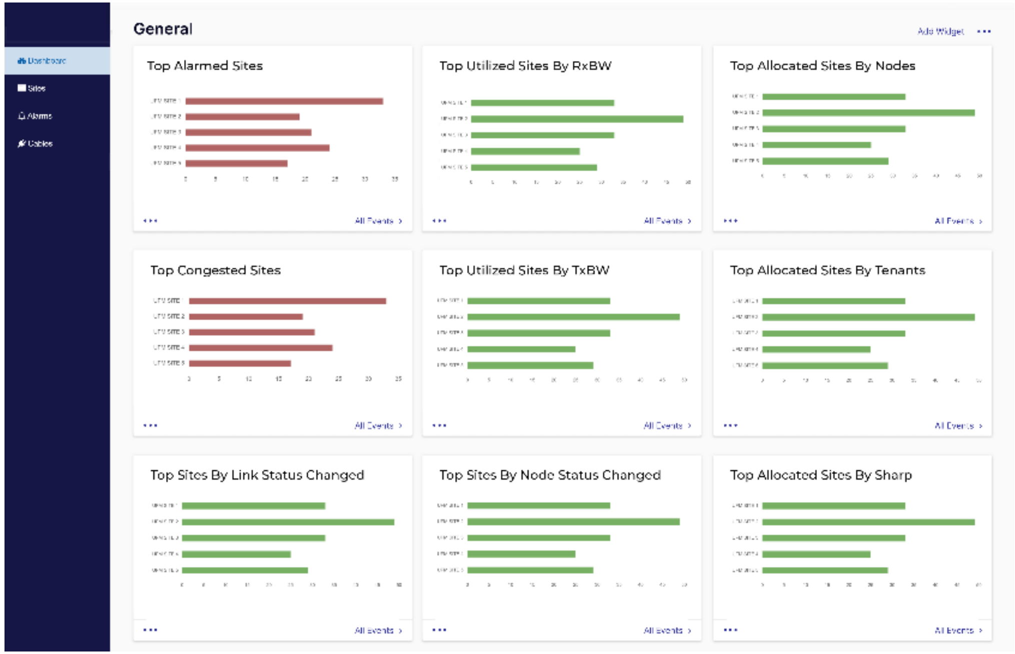


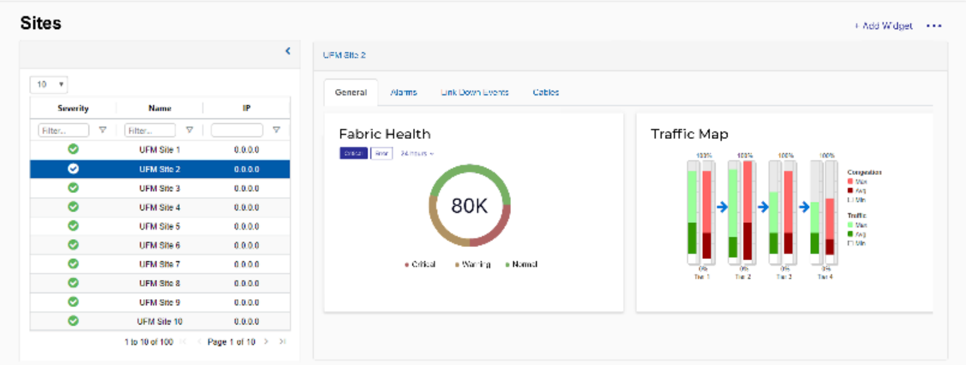
Severity	Name	Description	Type
✓	Suppressed_Devices	No event notifications issued	General
✓	Alarmed_Devices	Devices with alarms	General
✓	Rack_1	N/A	Rack

Name	GUID	IP
R-SMG161	e41d2d0300a20c0c	0.0.0.0
R-SMG165	e41d2d030067b0e8	0.0.0.0

MULTI CLOUD SOLUTION

- Correlation between Pkey creation/GUID assignment and traffic utilization/congestion
- Single main dashboard for all managed cloud/clusters
- Alerts, Traffic utilization, Congestion, Cable Info, Health

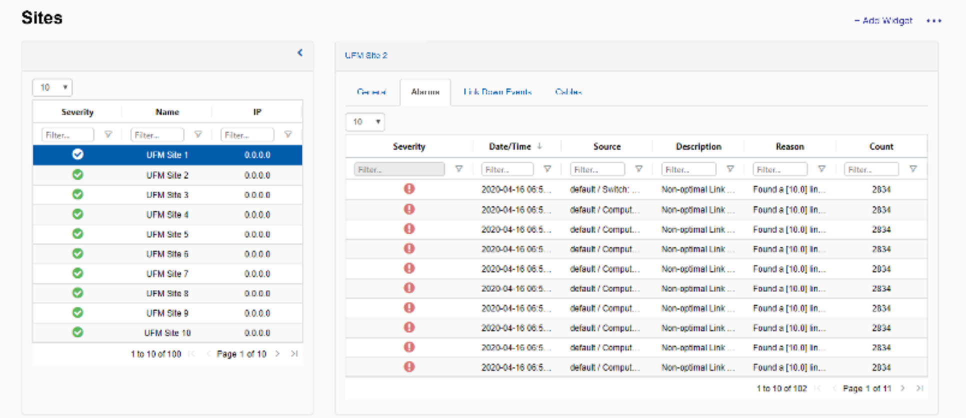

Dashboard
Sites
Alarms
Cables
General
Add Widget
Top Alarmed Sites
Top Utilized Sites By RxBW
Top Allocated Sites By Nodes
Top Congested Sites
Top Utilized Sites By TxBW
Top Allocated Sites By Tenants
Top Sites By Link Status Changed
Top Sites By Node Status Changed
Top Allocated Sites By Sharp
All Events ›



10

Severity	Name	IP
✓	UFM Site 1	0.0.0.0
✓	UFM Site 2	0.0.0.0
✓	UFM Site 3	0.0.0.0
✓	UFM Site 4	0.0.0.0
✓	UFM Site 5	0.0.0.0
✓	UFM Site 6	0.0.0.0
✓	UFM Site 7	0.0.0.0
✓	UFM Site 8	0.0.0.0
✓	UFM Site 9	0.0.0.0
✓	UFM Site 10	0.0.0.0

1 to 10 of 100 Page 1 of 10



10

Severity	Name	IP
✓	UFM Site 1	0.0.0.0
✓	UFM Site 2	0.0.0.0
✓	UFM Site 3	0.0.0.0
✓	UFM Site 4	0.0.0.0
✓	UFM Site 5	0.0.0.0
✓	UFM Site 6	0.0.0.0
✓	UFM Site 7	0.0.0.0
✓	UFM Site 8	0.0.0.0
✓	UFM Site 9	0.0.0.0
✓	UFM Site 10	0.0.0.0

1 to 10 of 100 Page 1 of 10

10

Severity	Date/Time ↓	Source	Description	Reason	Count
!	2020-04-16 06:5...	default / Switch: ...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834
!	2020-04-16 06:5...	default / Comput...	Non-optimal Link ...	Found a [10.0] lin...	2834

1 to 10 of 102 Page 1 of 11

UFM IN CLOUD

Day 1 operations	Day 2 Operations
Fabric Bring up validation	Network Auto Provisioning
Cable check	Tenant Security and Isolation
Link check	Chassis Fault Detection
Connectivity Check	Network Congestion
BW Check	Network Issues
Latency Check	Network Analysis and Monitoring
Chassis Check	HA service for the network
Inventory Discovery + Health	Events and Alarms

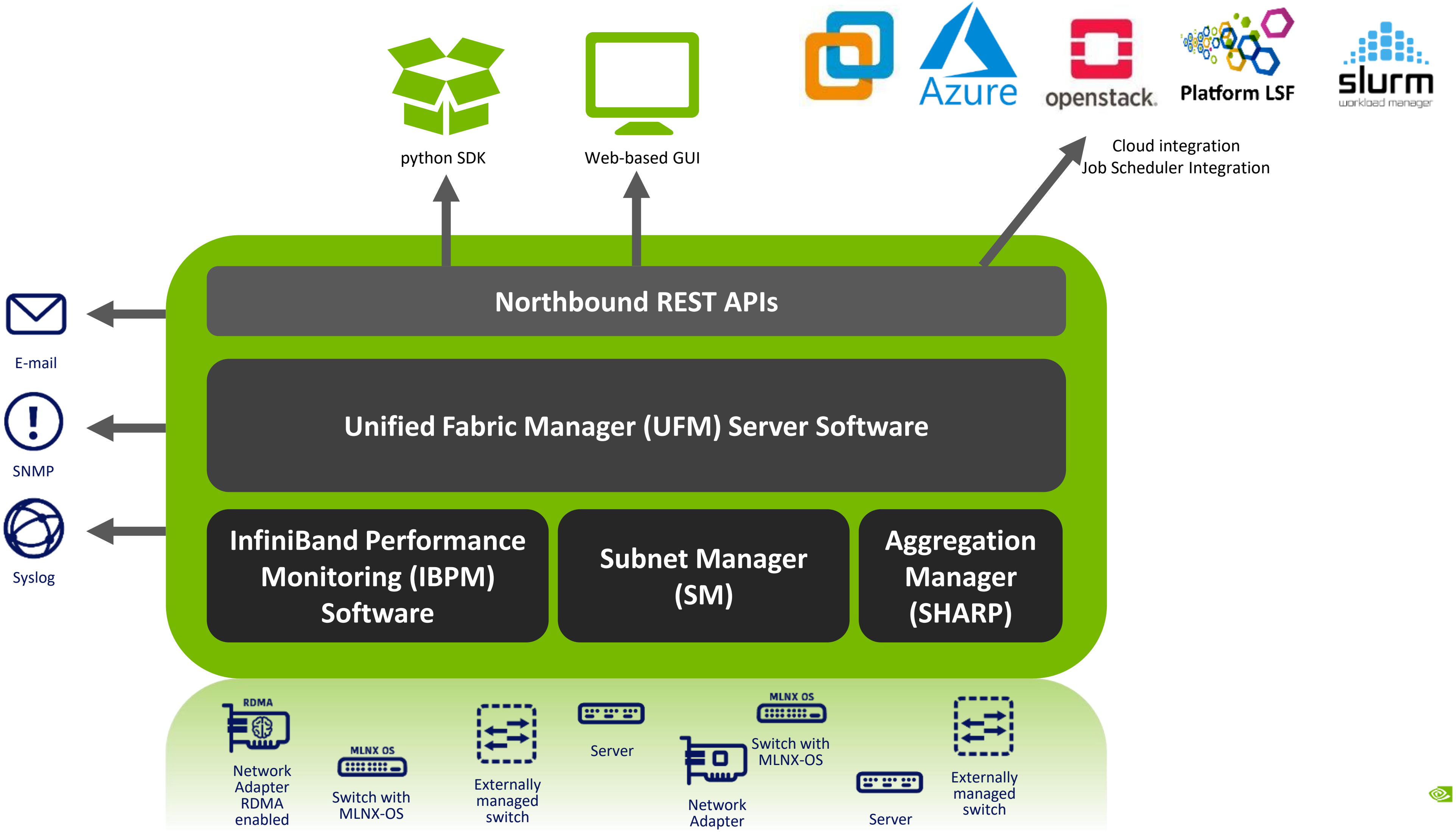
UFM SECURITY FEATURES FOR CLOUD

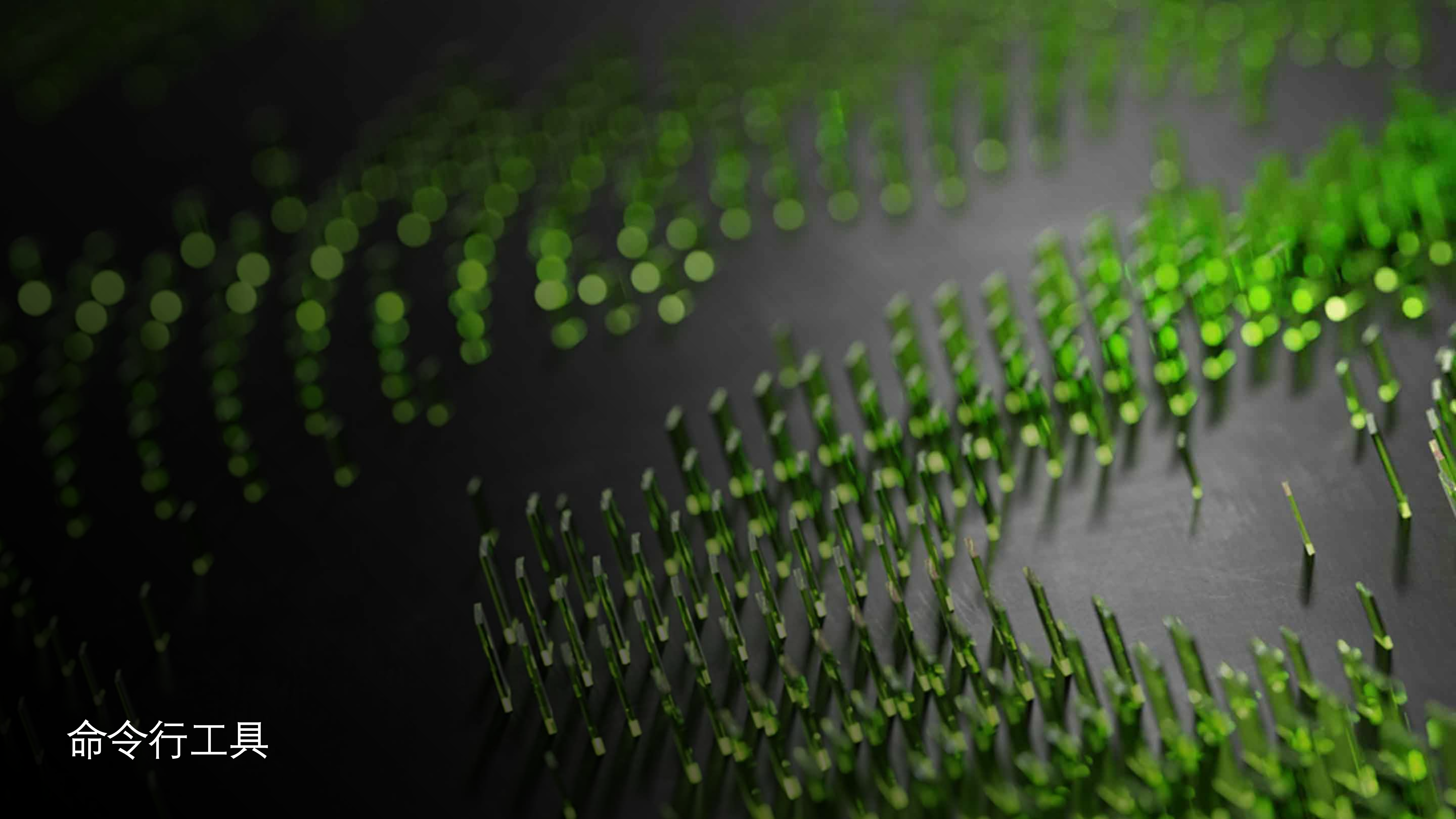
Subject
Alert and action on SA_Key violation (detecting malicious queries and reporting)
Alert and action on SA DoSc (detecting and reporting attack)
Randomize SA_Key on SM start (in order to protect the SM from item 1)
Support for M_Key per port in SM and tools (protect fabric from malicious configuration)
Flows to isolate violator of security alerts (fabric operator action)
ConnectX-5 And 6 SLID anti-masquerading feature in steering logic
Prevent dDOS by malicious registration in SM
Secured Cable Management (Detecting cable changes)
Switch Port Bad Pkey Alert

UFM SECURITY EVENTS

ID	Subject	
256	Bad M_Key	Found bad Management key. Check your HCA driver or partition settings. Management Key: Enforces the control of a master subnet manager
257	Bad P_Key	Found a bad Partition key. Check your partitioning settings. Partition Key: Enforces membership. Administered through the subnet manager by the partition manager (PM).
258	Bad Q_Key	Found bad Queue key. Security error. Queue Key: Enforces access rights for reliable and unreliable datagram service (RAW datagram service type not included)
259	Bad P_Key Switch External Port	Found a bad Partition key. Check your partitioning settings. Partition Key: Enforces membership. Administered through the subnet manager by the partition manager (PM)
560	User Connected	User Connected
561	User Disconnected	User Disconnected
1300	SA Key violation	SA Key Volation Committed
1301	SGID Spoofed	SGID spoofed by VPort/port
1302	SA High Rate detected	Rate Limit Exceeded

UFM SOFTWARE ARCHITECTURE





命令行工具

IBDIAGNET

Ibutils2

Scans the fabric using directed route packets and extracts all the available information regarding its connectivity and devices. An ibdiagnet run performs the following stages:

- Fabric discovery
- Duplicated GUIDs detection
- Links in INIT state and unresponsive links detection
- Counters fetch
- Error counters check
- Routing checks
- Link width and speed checks
- Alias GUIDs check
- Subnet Manager check
- Partition keys check

IBDIAGNET COMMAND

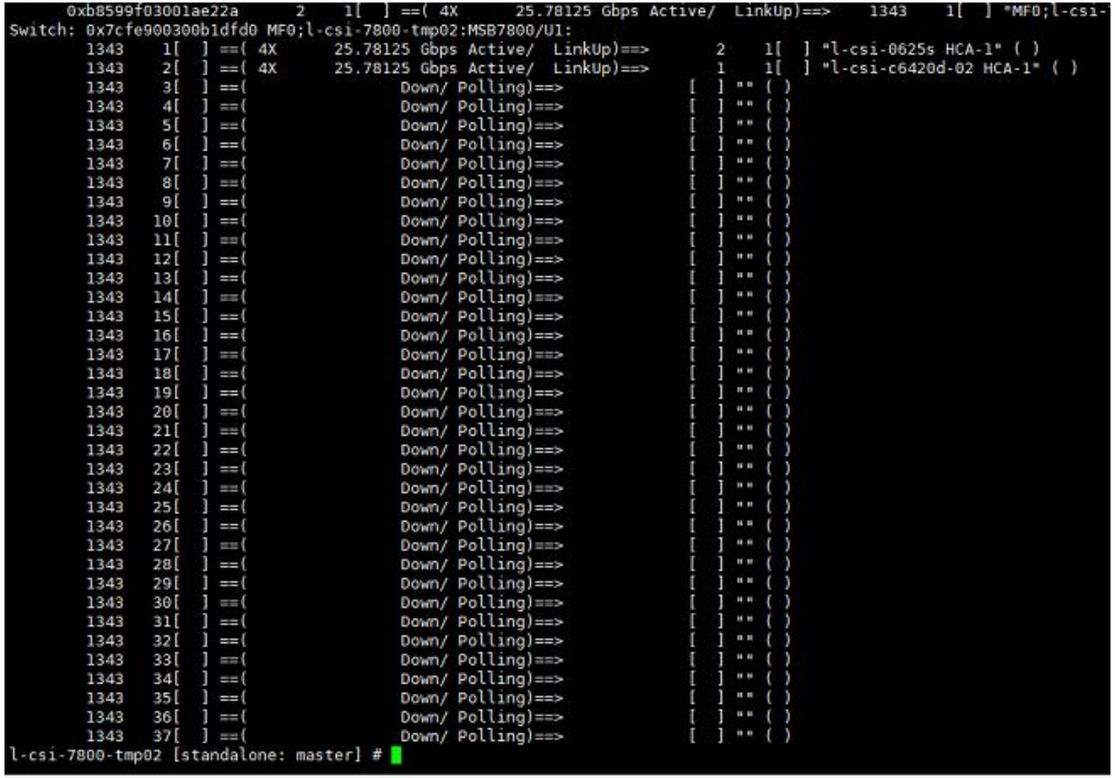
INSTALLATION

```
------------
mkdir /tmp/mlnx
cd /tmp/mlnx
cp  <path>/ibdiagnet_monitor_4.5.tgz .
tar zxf ibdiagnet_monitor_4.5.tgz
export IBDIAGNET_PLUGINS_PATH=/tmp/mlnx/usr/share/ibdiagnet2.1.1/plugins
export LD_LIBRARY_PATH=/tmp/mlnx/usr/lib
/tmp/mlnx/usr/bin/ibdiagnet <relevent flags>
/tmp/mlnx/usr/bin/ibdiagnet -pc --pm_pause_time 300 -P all=1 --get_cable_info --get_phy_info
```

Logs will be at the same place in /var/tmp/ibdiagnet2/*

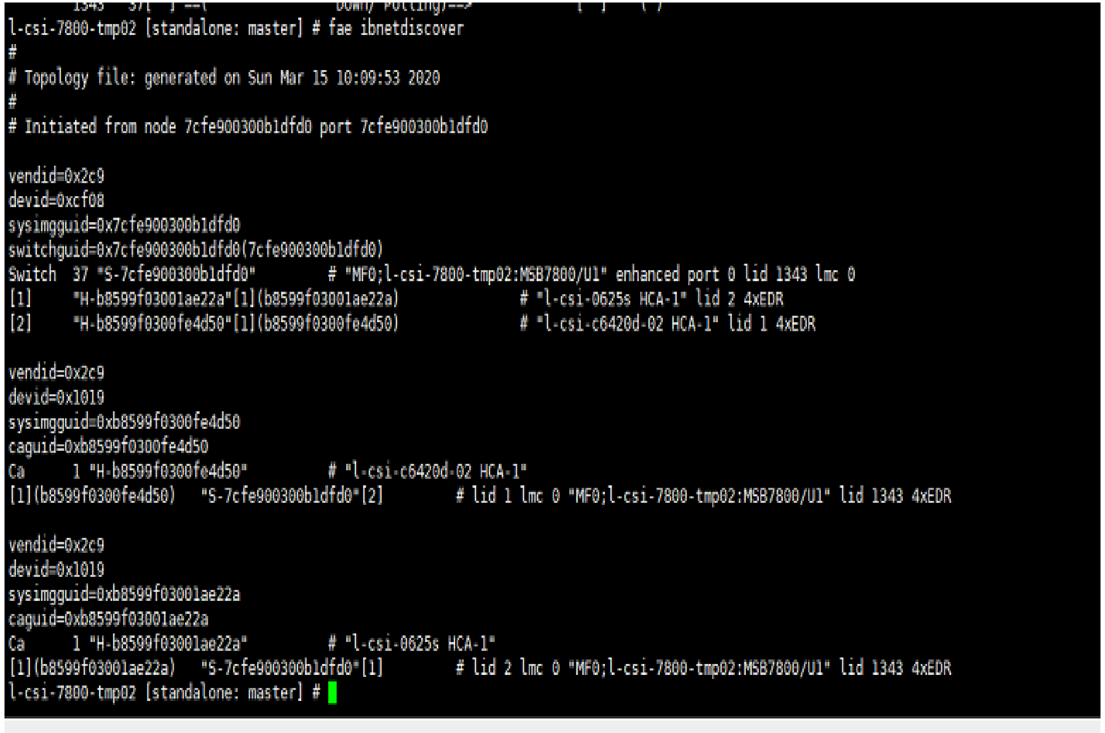
ibdiagnet_monitor_4.5.tgz

#iblinkinfo



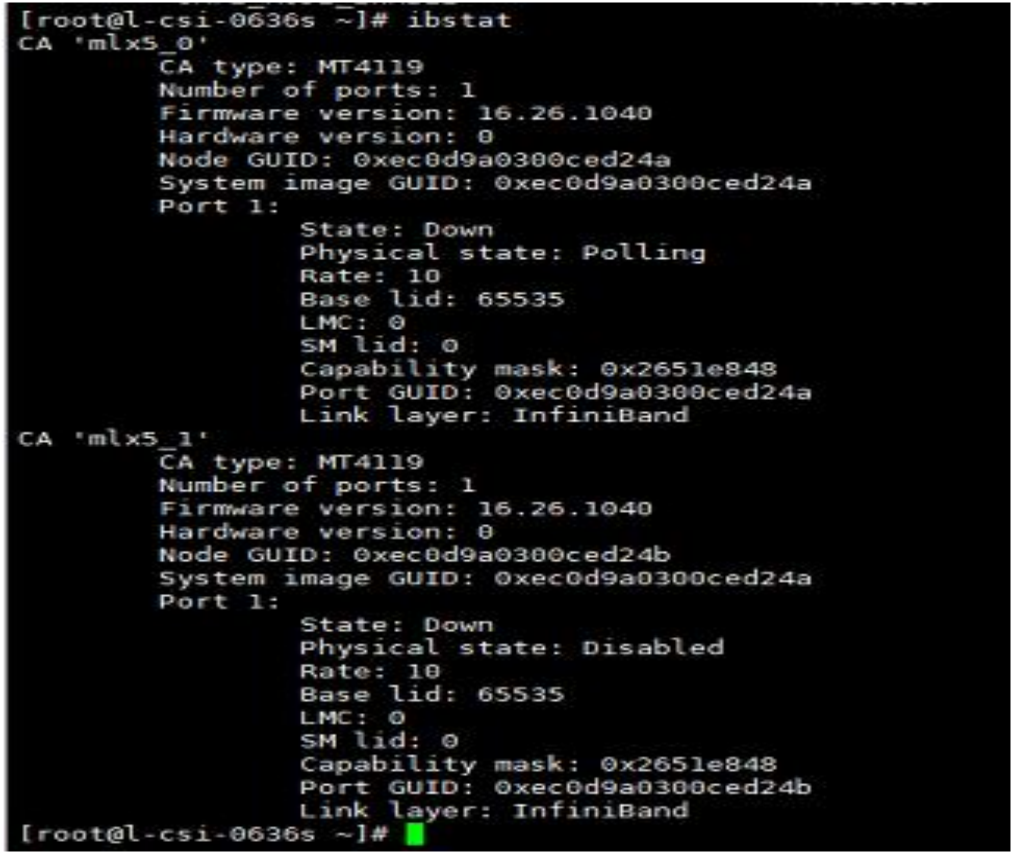
```
      0xb8599f03001ae22a      2    1[  ] ==( 4X      25.78125 Gbps Active/  LinkUp)==>    1343    1[  ] "MF0;l-csi-
Switch: 0x7cfe900300b1dfd0 MF0;l-csi-7800-tmp02:MSB7800/U1:
        1343     1[  ] ==( 4X       25.78125 Gbps Active/  LinkUp)==>       2     1[  ] "l-csi-0625s HCA-1" ( )
        1343     2[  ] ==( 4X       25.78125 Gbps Active/  LinkUp)==>       1     1[  ] "l-csi-c6420d-02 HCA-1" ( )
        1343     3[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343     4[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343     5[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343     6[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343     7[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343     8[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343     9[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    10[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    11[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    12[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    13[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    14[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    15[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    16[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    17[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    18[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    19[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    20[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    21[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    22[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    23[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    24[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    25[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    26[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    27[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    28[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    29[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    30[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    31[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    32[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    33[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    34[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    35[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    36[  ] ==(                Down/ Polling)==>             [  ] "" ( )
        1343    37[  ] ==(                Down/ Polling)==>             [  ] "" ( )
l-csi-7800-tmp02 [standalone: master] #
```

#ibnetdiscover



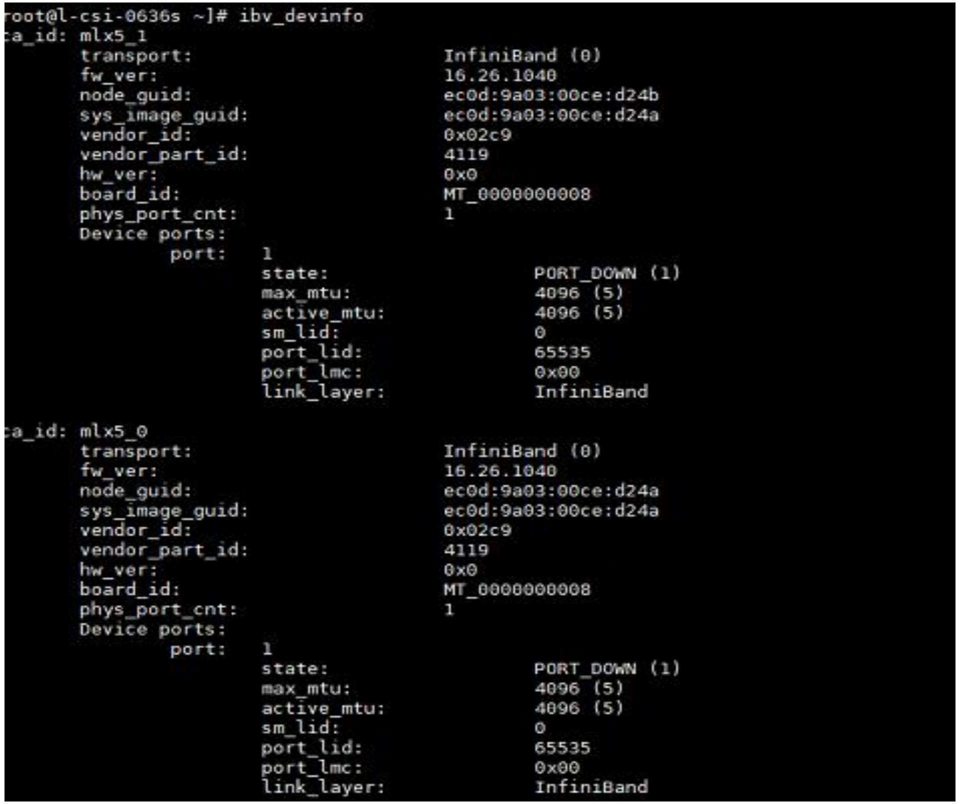
```
       1343   37[  ] ==(                Down/ Polling)==>            [  ]    ( )
l-csi-7800-tmp02 [standalone: master] # fae ibnetdiscover
#
# Topology file: generated on Sun Mar 15 10:09:53 2020
#
# Initiated from node 7cfe900300b1dfd0 port 7cfe900300b1dfd0

vendid=0x2c9
devid=0xcf08
sysimgguid=0x7cfe900300b1dfd0
switchguid=0x7cfe900300b1dfd0(7cfe900300b1dfd0)
Switch  37 "S-7cfe900300b1dfd0"         # "MF0;l-csi-7800-tmp02:MSB7800/U1" enhanced port 0 lid 1343 lmc 0
[1]     "H-b8599f03001ae22a"[1](b8599f03001ae22a)               # "l-csi-0625s HCA-1" lid 2 4xEDR
[2]     "H-b8599f0300fe4d50"[1](b8599f0300fe4d50)               # "l-csi-c6420d-02 HCA-1" lid 1 4xEDR

vendid=0x2c9
devid=0x1019
sysimgguid=0xb8599f0300fe4d50
caguid=0xb8599f0300fe4d50
Ca      1 "H-b8599f0300fe4d50"          # "l-csi-c6420d-02 HCA-1"
[1](b8599f0300fe4d50)   "S-7cfe900300b1dfd0"[2]         # lid 1 lmc 0 "MF0;l-csi-7800-tmp02:MSB7800/U1" lid 1343 4xEDR

vendid=0x2c9
devid=0x1019
sysimgguid=0xb8599f03001ae22a
caguid=0xb8599f03001ae22a
Ca      1 "H-b8599f03001ae22a"          # "l-csi-0625s HCA-1"
[1](b8599f03001ae22a)   "S-7cfe900300b1dfd0"[1]         # lid 2 lmc 0 "MF0;l-csi-7800-tmp02:MSB7800/U1" lid 1343 4xEDR
l-csi-7800-tmp02 [standalone: master] #
```

#ibstat



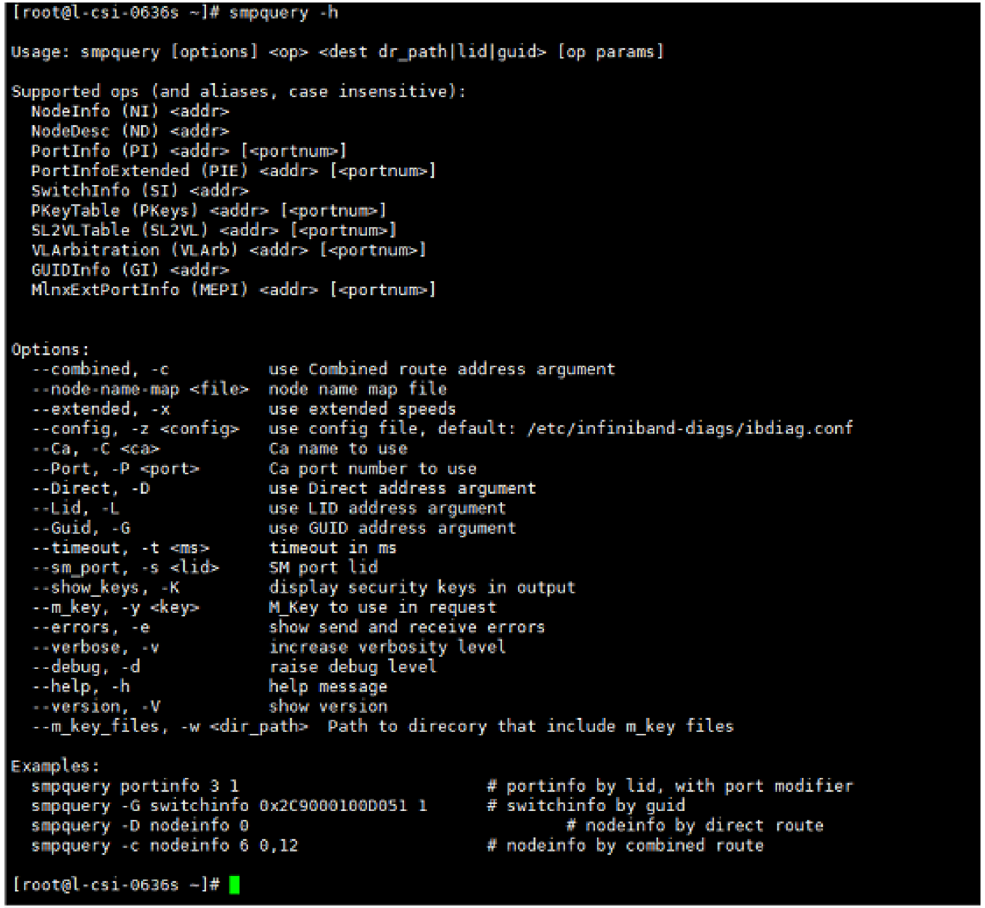
```
[root@l-csi-0636s ~]# ibstat
CA 'mlx5_0'
        CA type: MT4119
        Number of ports: 1
        Firmware version: 16.26.1040
        Hardware version: 0
        Node GUID: 0xec0d9a0300ced24a
        System image GUID: 0xec0d9a0300ced24a
        Port 1:
                State: Down
                Physical state: Polling
                Rate: 10
                Base lid: 65535
                LMC: 0
                SM lid: 0
                Capability mask: 0x2651e848
                Port GUID: 0xec0d9a0300ced24a
                Link layer: InfiniBand
CA 'mlx5_1'
        CA type: MT4119
        Number of ports: 1
        Firmware version: 16.26.1040
        Hardware version: 0
        Node GUID: 0xec0d9a0300ced24b
        System image GUID: 0xec0d9a0300ced24a
        Port 1:
                State: Down
                Physical state: Disabled
                Rate: 10
                Base lid: 65535
                LMC: 0
                SM lid: 0
                Capability mask: 0x2651e848
                Port GUID: 0xec0d9a0300ced24b
                Link layer: InfiniBand
[root@l-csi-0636s ~]#
```

#ibv_devinfo



```
root@l-csi-0636s ~]# ibv_devinfo
ca_id: mlx5_1
       transport:                       InfiniBand (0)
       fw_ver:                          16.26.1040
       node_guid:                       ec0d:9a03:00ce:d24b
       sys_image_guid:                  ec0d:9a03:00ce:d24a
       vendor_id:                       0x02c9
       vendor_part_id:                  4119
       hw_ver:                          0x0
       board_id:                        MT_0000000008
       phys_port_cnt:                   1
       Device ports:
               port:   1
                       state:                   PORT_DOWN (1)
                       max_mtu:                 4096 (5)
                       active_mtu:              4096 (5)
                       sm_lid:                  0
                       port_lid:                65535
                       port_lmc:                0x00
                       link_layer:              InfiniBand

ca_id: mlx5_0
       transport:                       InfiniBand (0)
       fw_ver:                          16.26.1040
       node_guid:                       ec0d:9a03:00ce:d24a
       sys_image_guid:                  ec0d:9a03:00ce:d24a
       vendor_id:                       0x02c9
       vendor_part_id:                  4119
       hw_ver:                          0x0
       board_id:                        MT_0000000008
       phys_port_cnt:                   1
       Device ports:
               port:   1
                       state:                   PORT_DOWN (1)
                       max_mtu:                 4096 (5)
                       active_mtu:              4096 (5)
                       sm_lid:                  0
                       port_lid:                65535
                       port_lmc:                0x00
                       link_layer:              InfiniBand
```

#smpquery



```
[root@l-csi-0636s ~]# smpquery -h

Usage: smpquery [options] <op> <dest dr_path|lid|guid> [op params]

Supported ops (and aliases, case insensitive):
  NodeInfo (NI) <addr>
  NodeDesc (ND) <addr>
  PortInfo (PI) <addr> [<portnum>]
  PortInfoExtended (PIE) <addr> [<portnum>]
  SwitchInfo (SI) <addr>
  PKeyTable (PKeys) <addr> [<portnum>]
  SL2VLTable (SL2VL) <addr> [<portnum>]
  VLArbitration (VLArb) <addr> [<portnum>]
  GUIDInfo (GI) <addr>
  MlnxExtPortInfo (MEPI) <addr> [<portnum>]


Options:
  --combined, -c          use Combined route address argument
  --node-name-map <file>  node name map file
  --extended, -x          use extended speeds
  --config, -z <config>   use config file, default: /etc/infiniband-diags/ibdiag.conf
  --Ca, -C <ca>           Ca name to use
  --Port, -P <port>       Ca port number to use
  --Direct, -D            use Direct address argument
  --Lid, -L               use LID address argument
  --Guid, -G              use GUID address argument
  --timeout, -t <ms>      timeout in ms
  --sm_port, -s <lid>     SM port lid
  --show_keys, -K         display security keys in output
  --m_key, -y <key>       M_Key to use in request
  --errors, -e            show send and receive errors
  --verbose, -v           increase verbosity level
  --debug, -d             raise debug level
  --help, -h              help message
  --version, -V           show version
  --m_key_files, -w <dir_path>  Path to direcory that include m_key files

Examples:
  smpquery portinfo 3 1                         # portinfo by lid, with port modifier
  smpquery -G switchinfo 0x2C9000100D051 1      # switchinfo by guid
  smpquery -D nodeinfo 0                                # nodeinfo by direct route
  smpquery -c nodeinfo 6 0,12                   # nodeinfo by combined route

[root@l-csi-0636s ~]#
```

OTHER COMMAND

- #ibswitches
- #ibnodes
- #ibhosts
- #ofed_info –S
- #mlxlink –d lid-<lid#> -p <port#>
- #perfquery
- #ibportstate
- #flint
- #mlxvpd

- #mlxup
- #mlxconfig
- #mlxfwmanage
- #ib_write_bw
- #ib_write_lat
- #ibping
- #ib_read_bw
- #ib_read_lat

HOST TOOLS

Sysinfo-snapshot

Windows

```
<installation_directory>\ManagementTools\MLNX_System_Snapshot.exe
```

Linux

```
#./sysinfo-snapshot.py
```

ESXi

```
# esxi-sysinfo-snapshot.py
```

NVIDIA®