cisco

IMAGINE

INTUITIVE

Cisco live!

June 10-14, 2018 · Orlando, FL

#CLUS

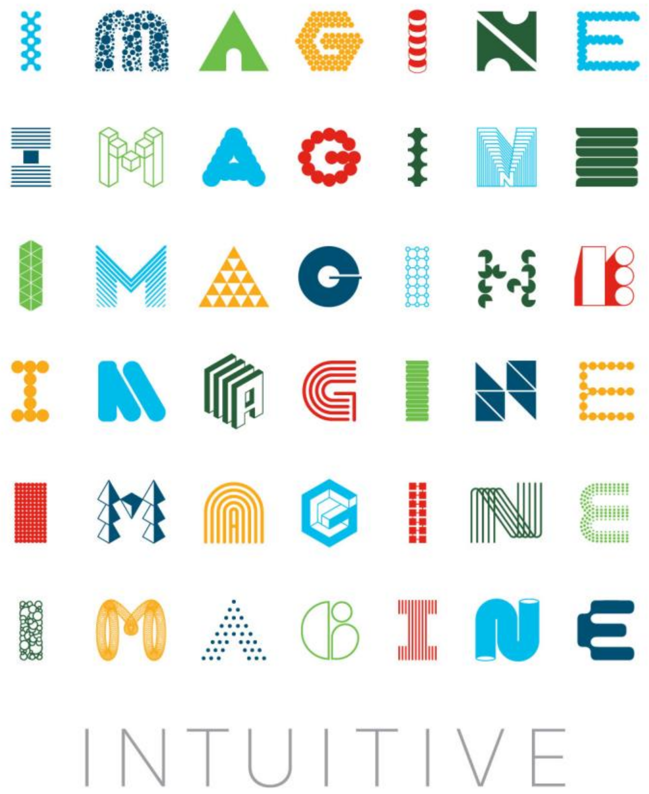
Cisco UCS as an Advanced Computing Platform for the Enterprise

Tae Hwang, Architect, Cisco

Tim Miller, Virtual Systems Engineer, Cisco

Bob Crovella, Solutions Architect, NVIDIA

TECINI-2543



Agenda

- **Introduction**
- Anatomy of Research Computing
- Demo and Hands-on Cluster Operations
- UCS as a Platform for Research Computing
- Hands-on Bonus: UCS Configuration with Ansible
- NVIDIA Deep Learning with GPUs
- Conclusion

Cisco Webex Teams

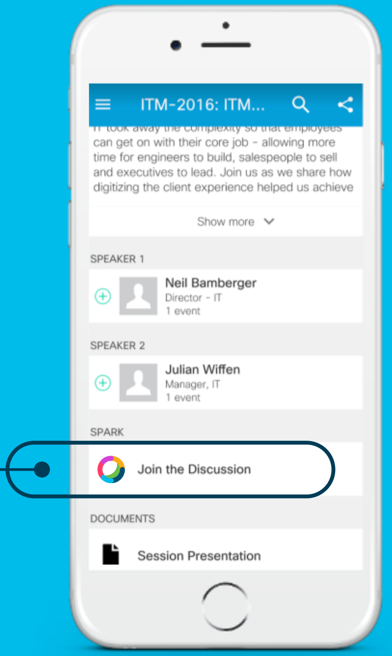
Questions?

Use Cisco Webex Teams (formerly Cisco Spark) to chat with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space

Webex Teams will be moderated by the speaker until June 18, 2018.



cs.co/ciscolivebot#TECINI-2543

Timothy E. Miller, PhD, CCNA, RHCA(*)

- Linux User - 1993
- UNIX/Linux Systems Admin - 1995
- HPC Systems Engineer - 2002
- Network Engineer - 2008
- Network Architect – 2012
- Financial Services Vertical – 2015
- Cisco Systems Engineer - 2017
  - UCS, MDS, VXLAN EVPN, CloudCenter, Programmability, Docker/K8s

Agenda

- Introduction
- **Anatomy of Research Computing**
- Demo and Hands-on Cluster Operations
- UCS as a Platform for Research Computing
- Hands-on Bonus: UCS Configuration with Ansible
- NVIDIA Deep Learning with GPUs
- Conclusion

What is HPC?

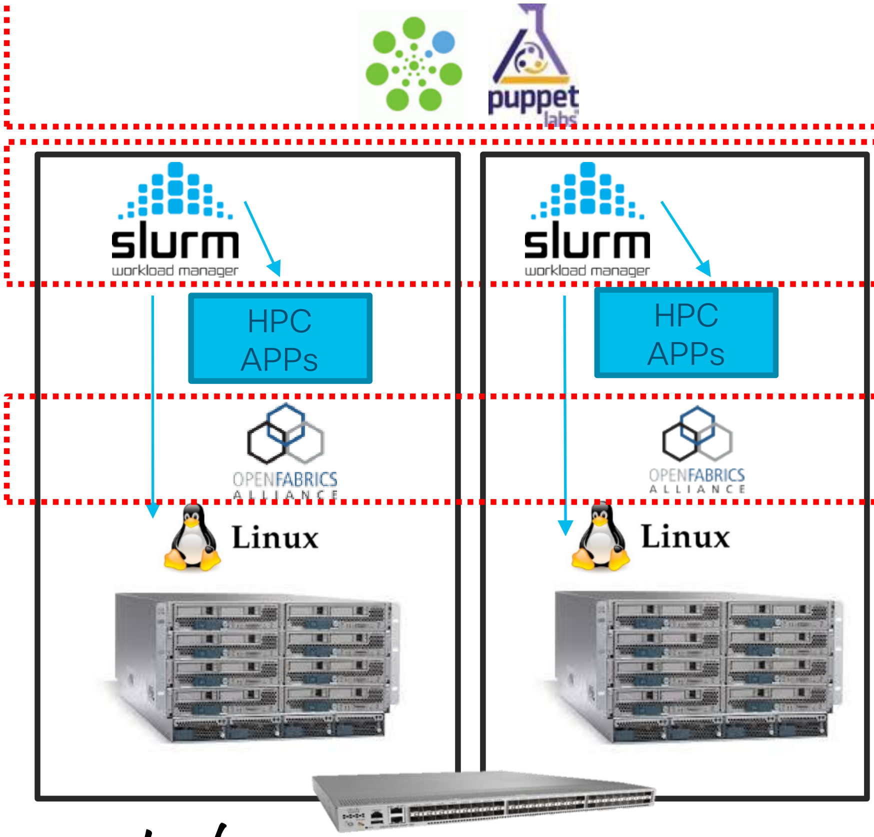
Let's Define High Performance Computing

- **x86 Clustered Linux Baremetal System** that supports high computing needs.
- **No longer FEW fastest Supercomputers in the world**. Several nodes HPC cluster can support hundreds of CPU cores (thousands of GPU cores).
- **Linux is the platform for science/research** – traditional STEM computation, social science, economy/finance, Machine Learning, Big Data, and etc.
- **200 of Top500 use Ethernet as their fabric**
- **System Components:** X86 System + Network/Fabirc (Ethernet, Infiniband, Omnipath, etc) + Storage + HPC Software*

WAKE FOREST
UNIVERSITY

- Has no Engineering School
- 150 Node HPC and growing
- Business school, School of Romantic Language school use HPC

HPC Components



Linux OS Cluster and
Baremetal Management Layer

Resource manager + HPC Apps scheduler

HPC Application Middleware

Generally speaking…

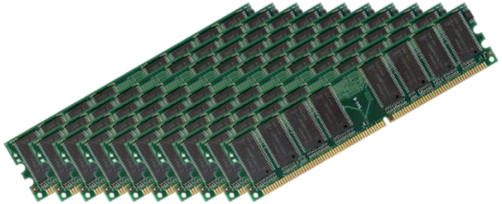
Supercomputer

=

Lots of processors

+

Lots of RAM



+

Lots of disk

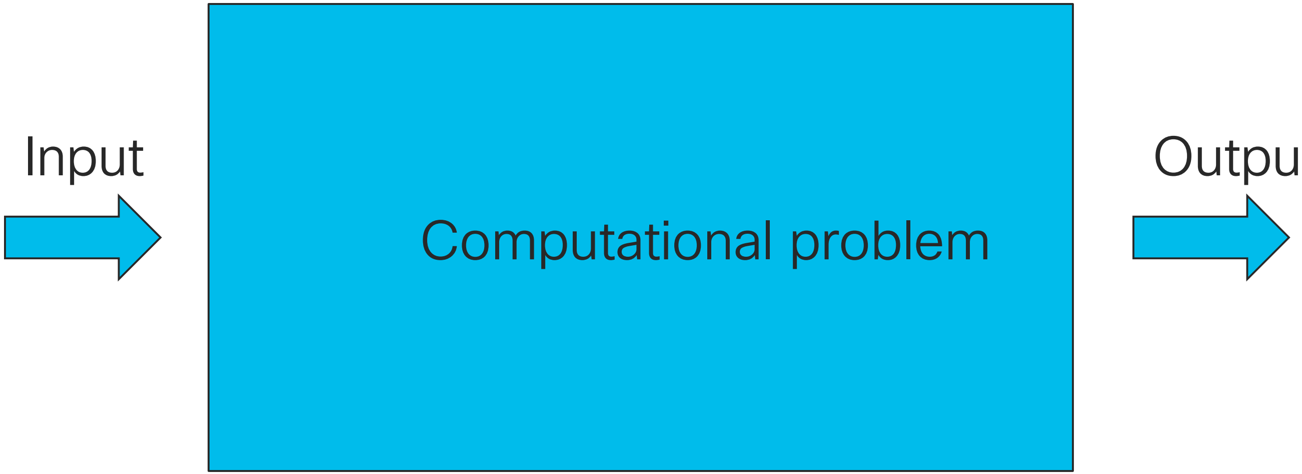
Just a bunch of servers?

The difference between
supercomputers and web farms
and database farms (and …)

All the servers act *together* to
solve a single computational problem

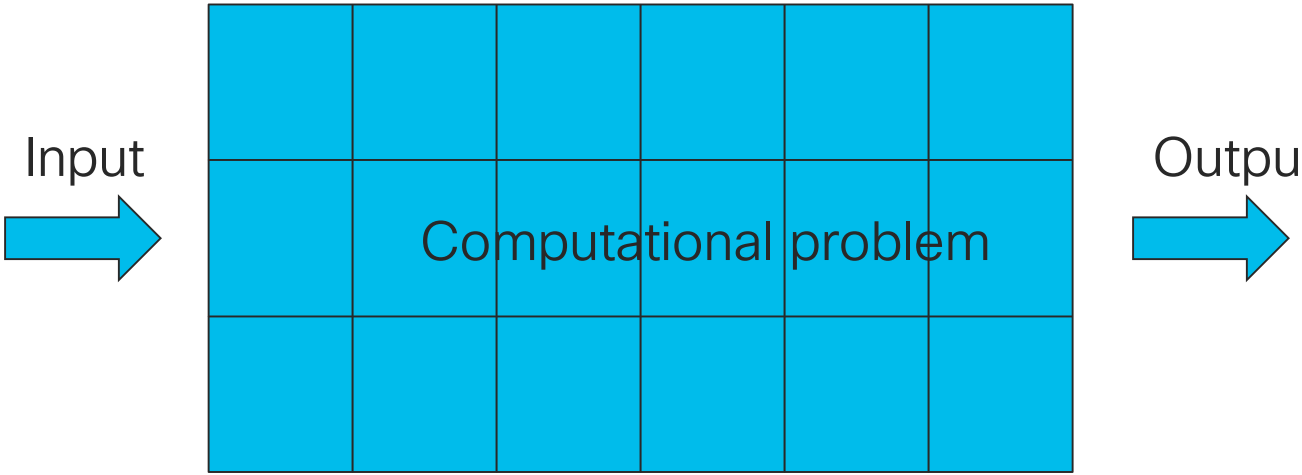
Acting together

Take your computational problem…



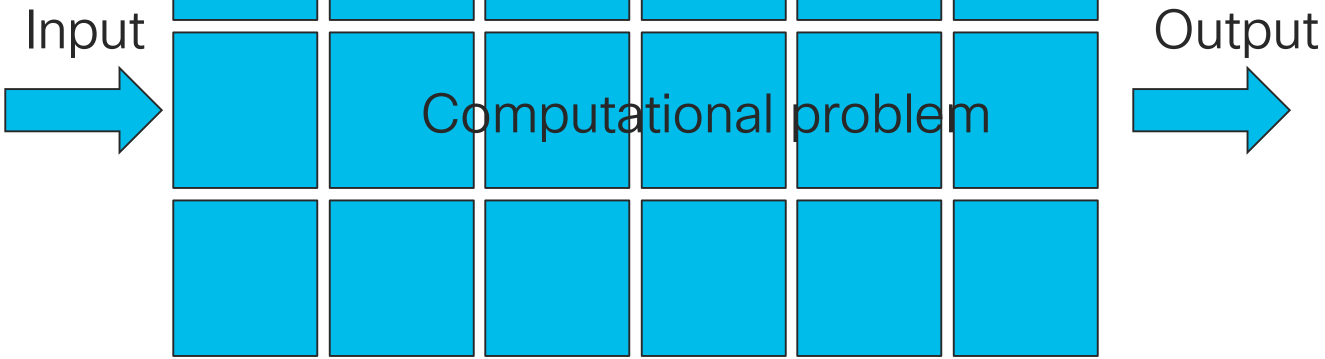
Acting together

...and split it up!



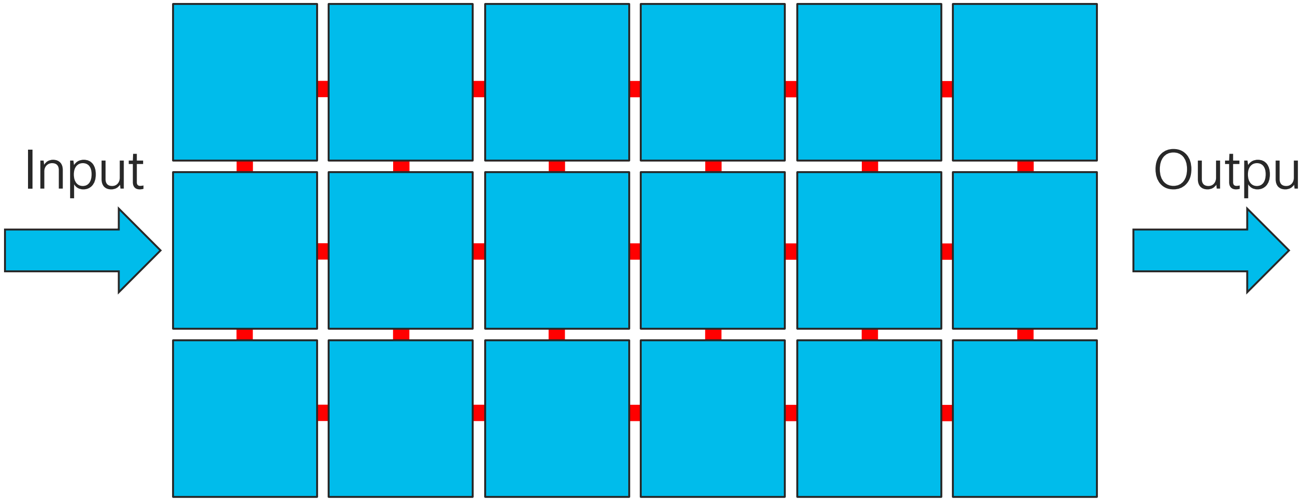
Acting together

Distribute the input data
across a bunch of servers



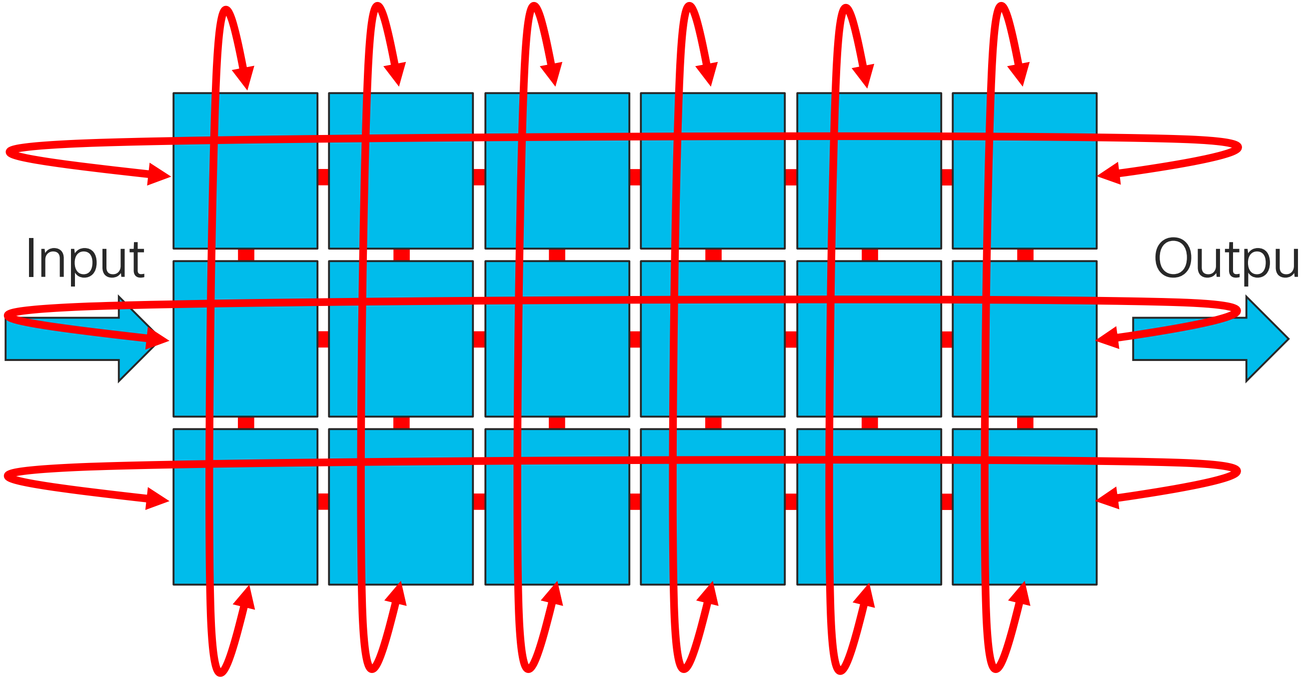
Acting together

Use the network between servers
to communicate / coordinate



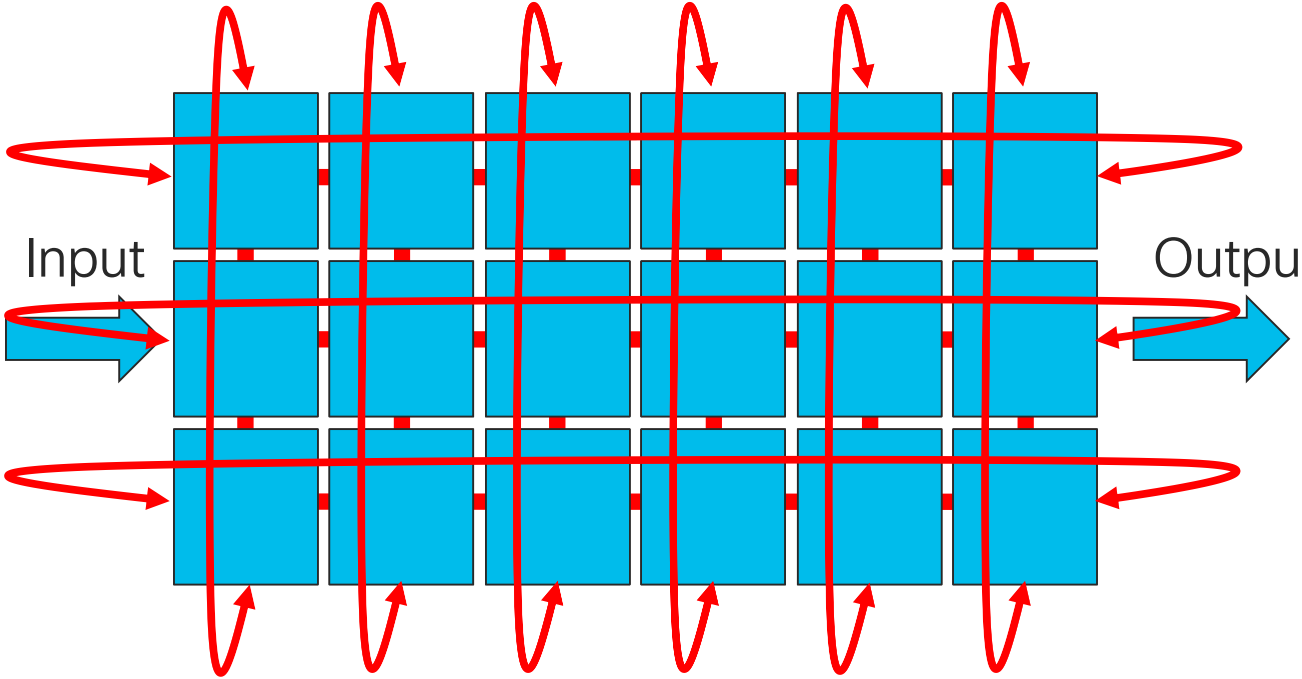
Acting together

Use the network between servers
to communicate / coordinate



Acting together

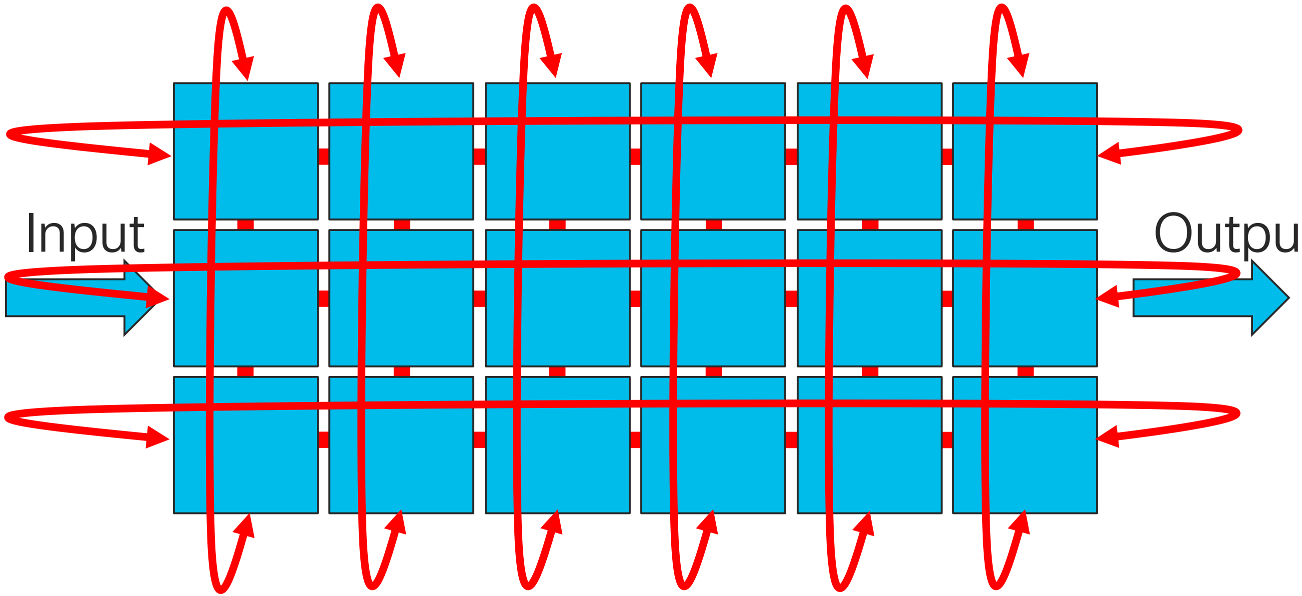
MPI is used for this communication
(we'll get to MPI in a minute)



Acting together

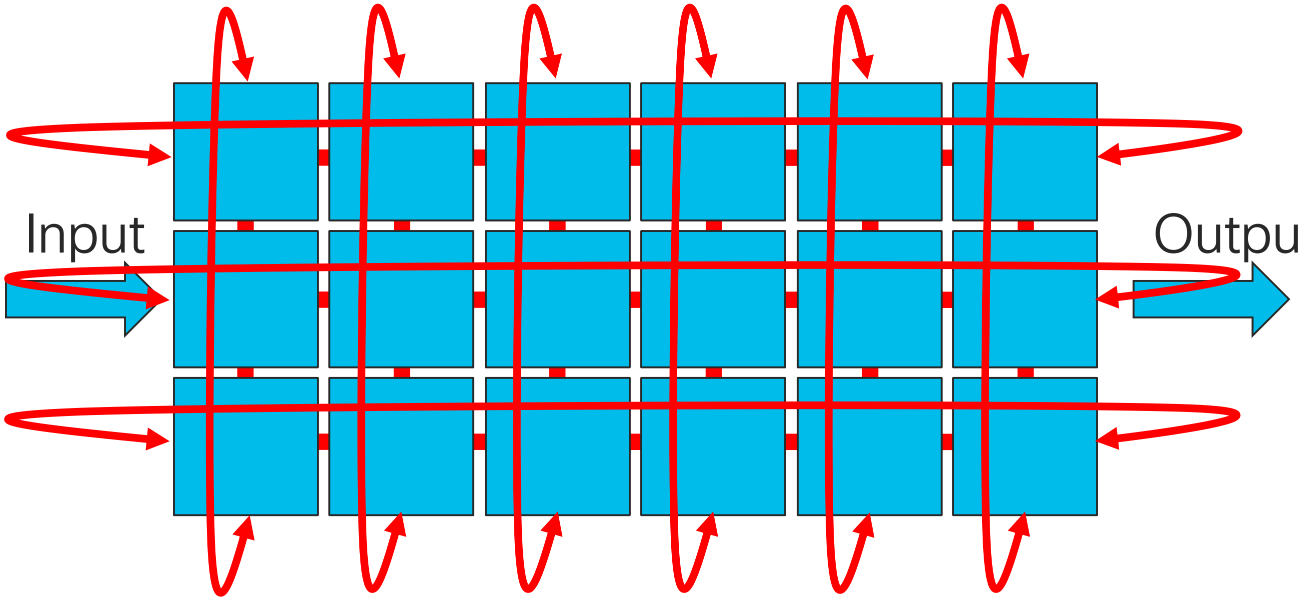
Many small messages are exchanged

Latency of short messages is extremely important

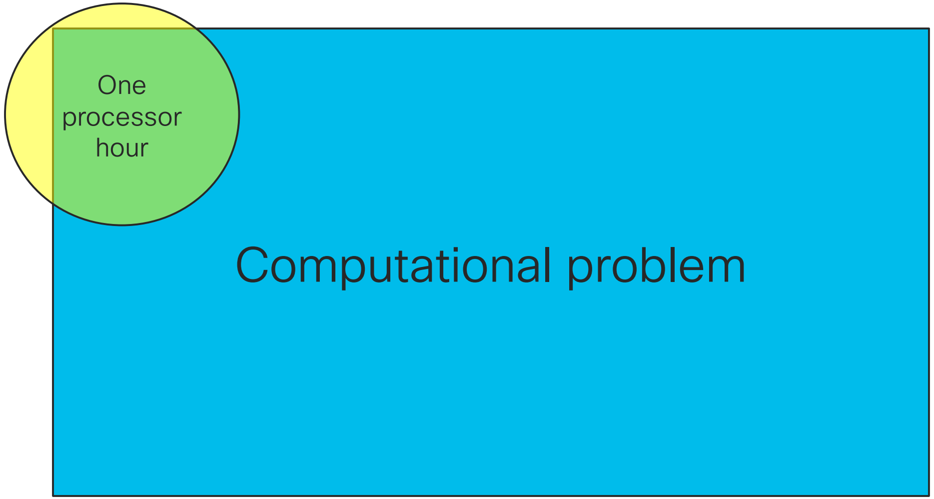


Parallel application

Individual (Linux) processes are running in parallel on each server, coordinating with each other

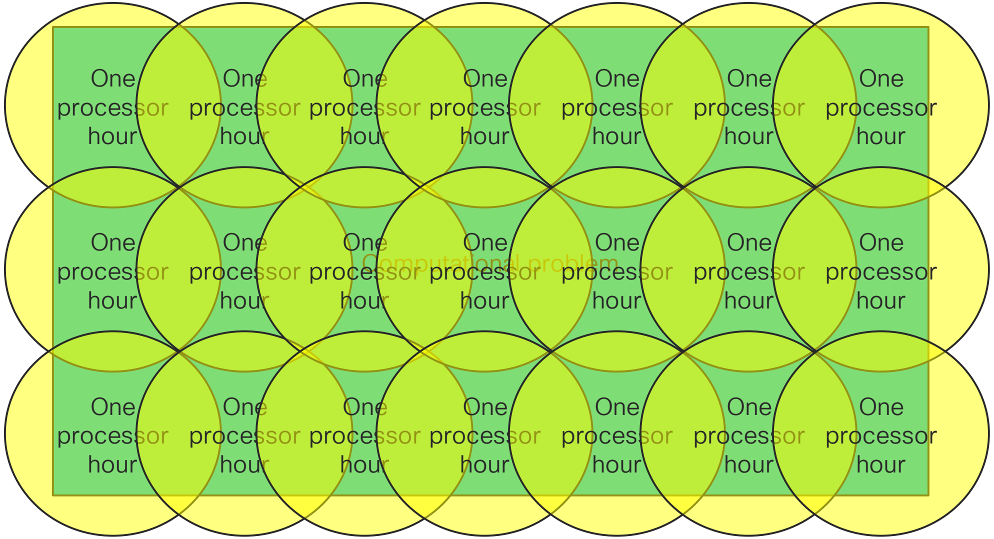


Why go to so much trouble?



1 processor = ...a long time...

Why go to so much trouble?



21 processors = ~1 hour (!)

Disclaimer: scaling is rarely perfect

High Performance Computing

HPC

=

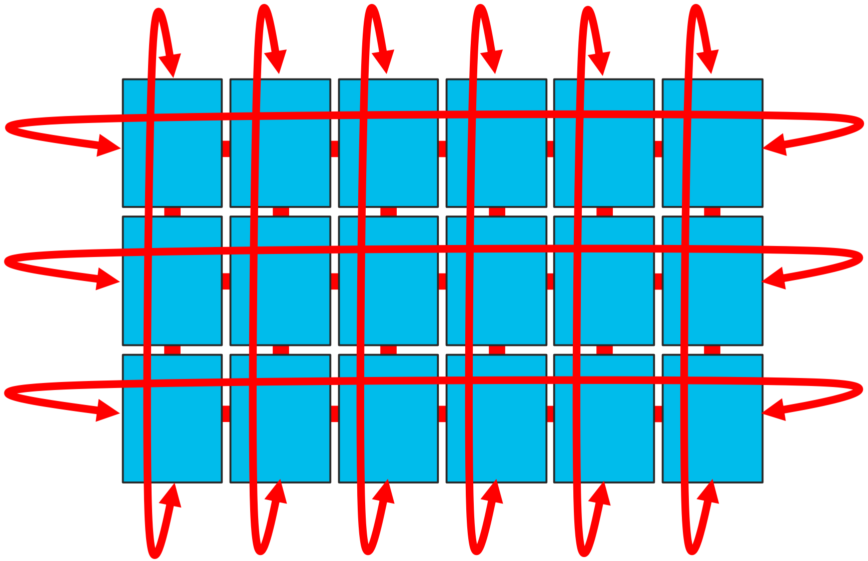
Using “supercomputers” to solve
real world problems that are
TOO BIG
for laptops, desktops,
or individuals servers

Other HPC / TC terms

Tightly coupled application

Each Linux process in the overall parallel application frequently sends short messages, usually to a small number of peer Linux processes.

Network latency -- and congestion management -- of short messages is extremely critical.

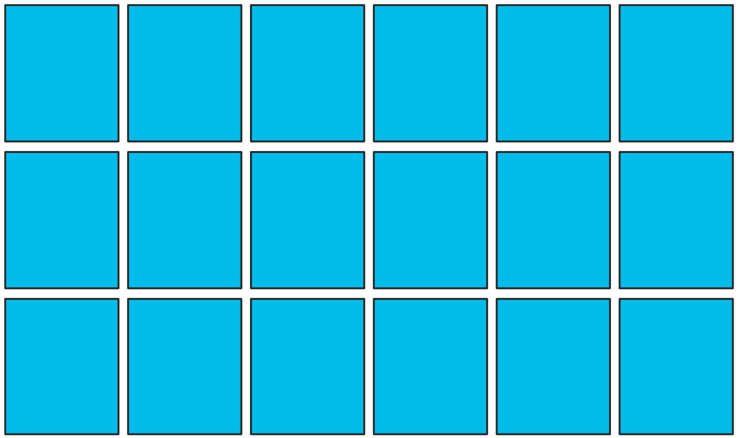


Other HPC / TC terms

Embarrassingly parallel application

The overall job is parallel, but little to no communication or coordination is required to compute the final result.

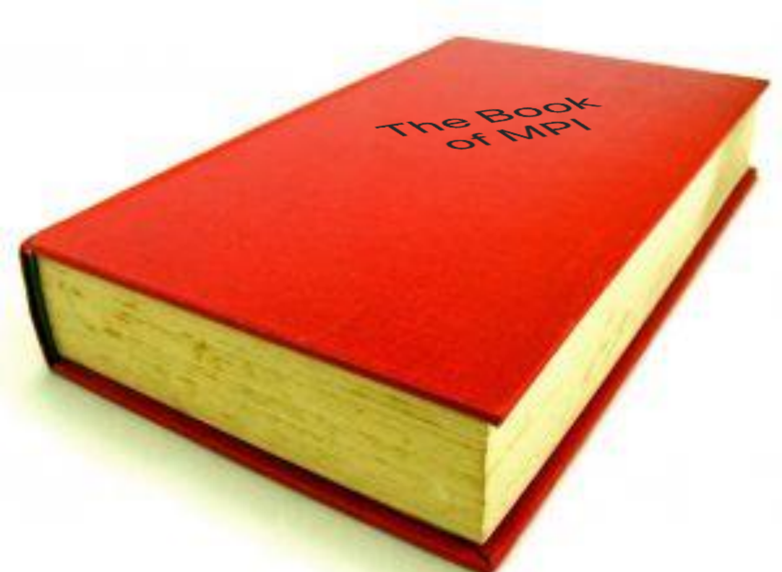
Network latency is likely not a factor. Bandwidth may or may not be, depending on the size of the input / output data sets.



What is MPI?

What is the Message Passing Interface (MPI)?

A standards document



www.mpi-forum.org

There are many implementations of the MPI standard

Some are closed source

Others are open source

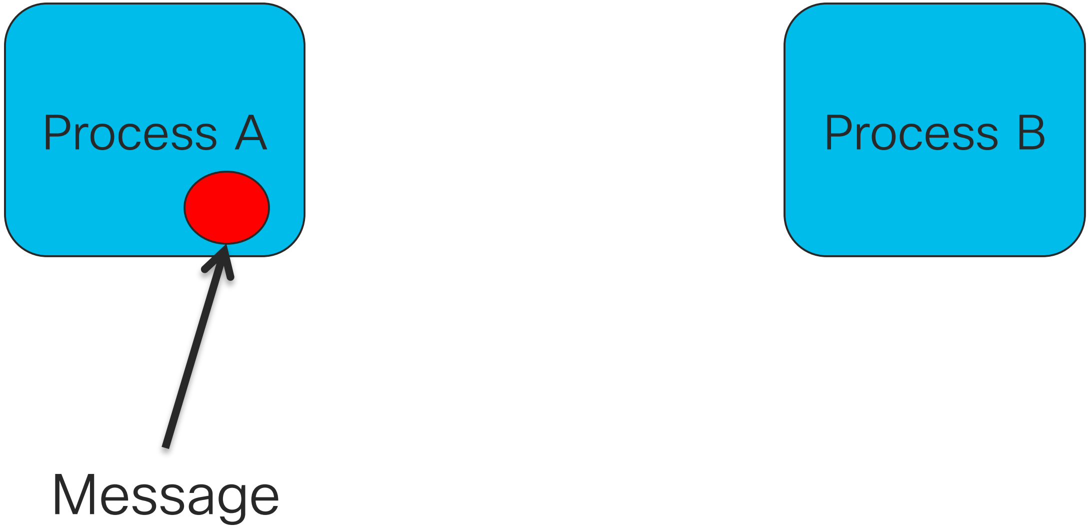
They are usually comprised of middleware and a runtime

So what is MPI *for?*

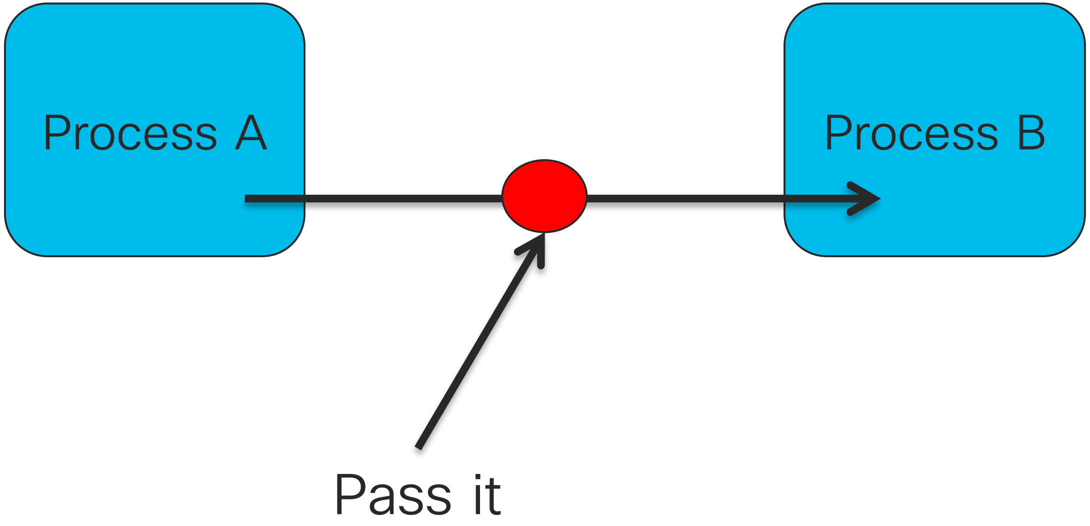
Let's break it down…

Message Passing | Interface

1. Message passing



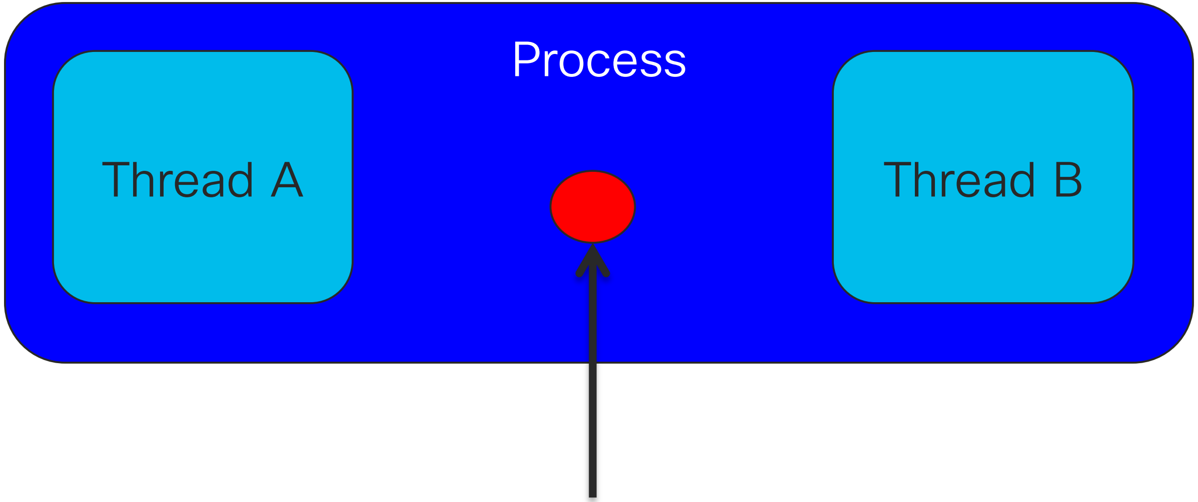
1. Message passing



1. Message passing

Message has been passed

1. Message passing

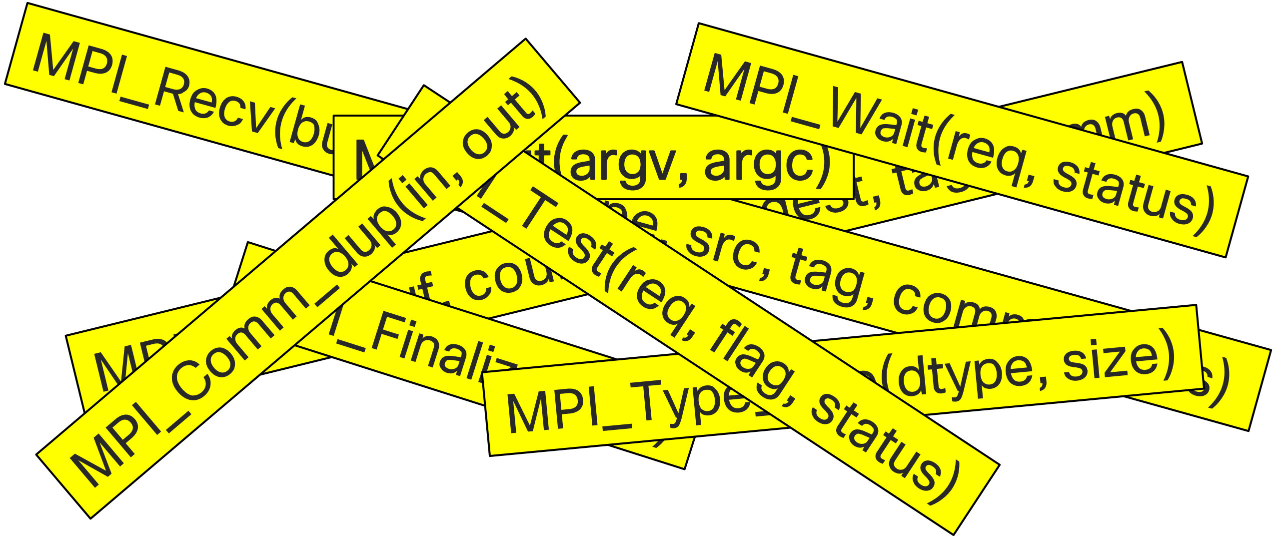


...as opposed to data that is shared

2. Interface

Fortran too!

C programming function calls



Putting it back together

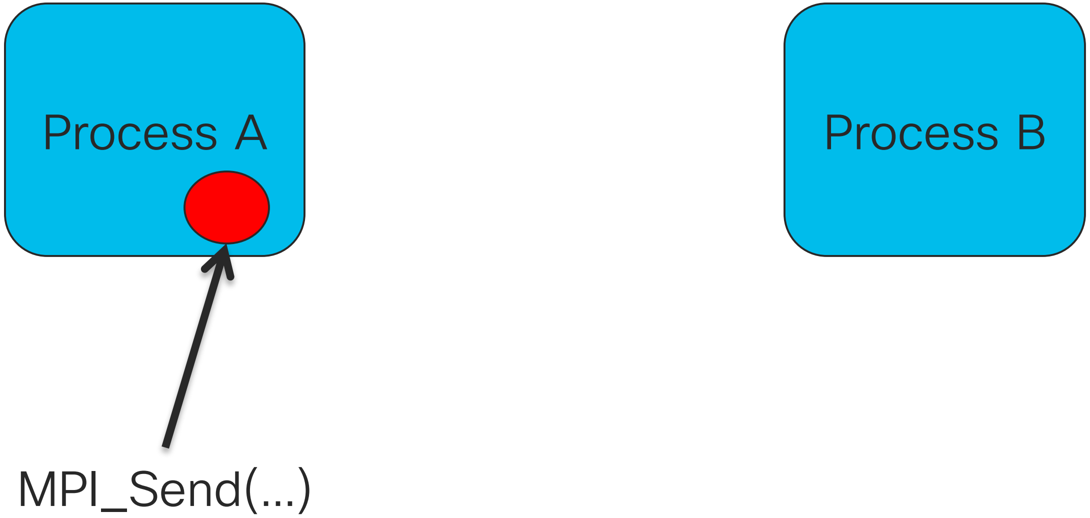
Message Passing | Interface

"An interface for passing messages"

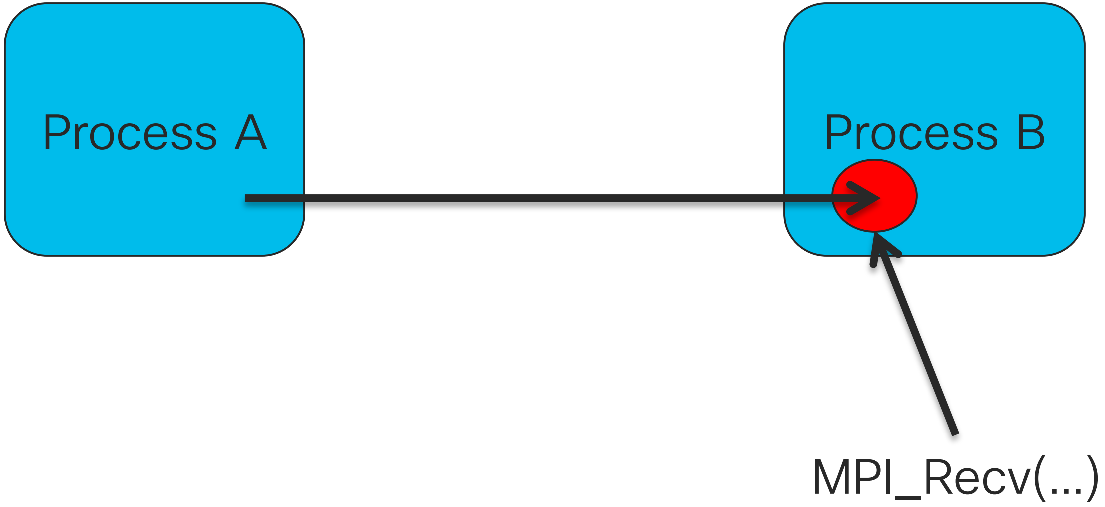
"C functions for passing messages"

Fortran too!

C/Fortran functions for message passing



C/Fortran functions for message passing



Really? Is that all MPI is?

"Can't I just do that with sockets?"

Yes!

(...and no)

Comparison

- (TCP) Sockets

- Connections based on IP addresses and ports
- Point-to-point communication
- Stream-oriented
- Raw data (bytes / octets)
- Network-independent
- "Slow"

- MPI

- Based on peer integer "rank" (e.g., 8)
- Point-to-point and collective and one-sided and ...
- Message oriented
- Typed messages
- Network independent
- Blazing fast

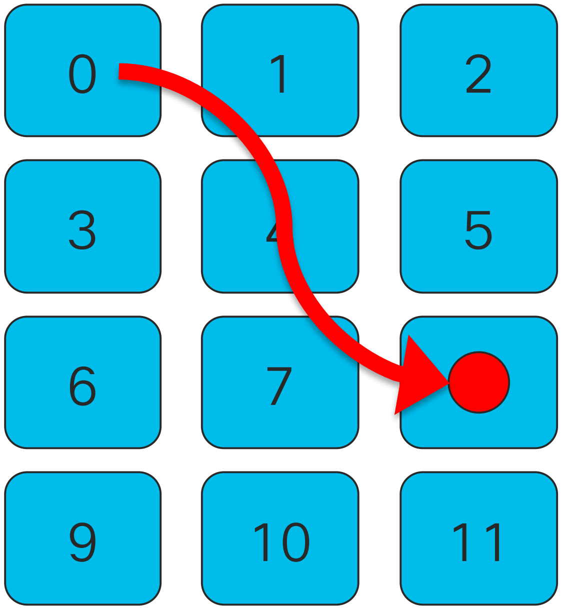
Comparison

- MPI

Whoa!
What are these?

- Based on peer integer "rank" (e.g., 8)
- Point-to-point and collective and one-sided and …
- Message oriented
- Typed messages
- Network independent
- Blazing fast

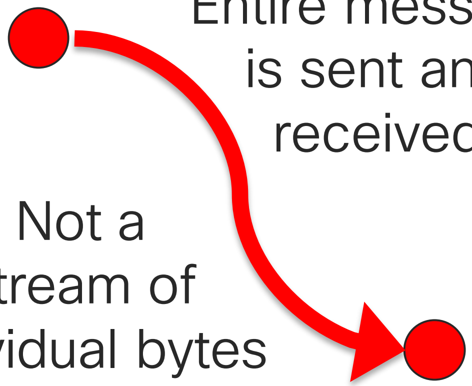
Peer integer “rank”



- MPI

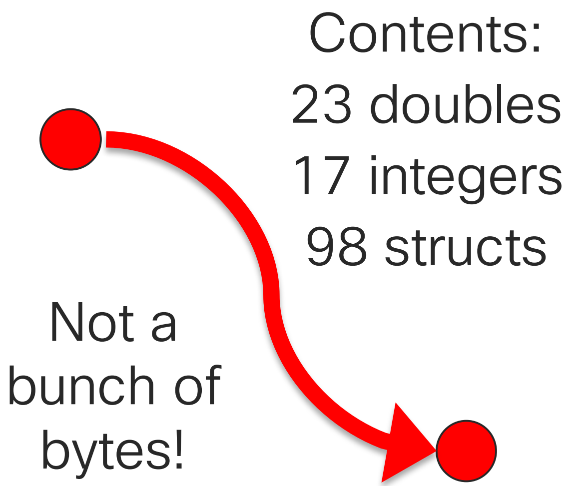
- **Based on peer integer “rank” (e.g., 8)**
- Point-to-point and collective and one-sided and …
- Message oriented
- Typed messages
- Network independent
- Blazing fast

Messages, not bytes



- MPI
  - Based on peer integer “rank” (e.g., 8)
  - Point-to-point and collective and one-sided and …
  - **Message oriented**
  - Typed messages
  - Network independent
  - Blazing fast

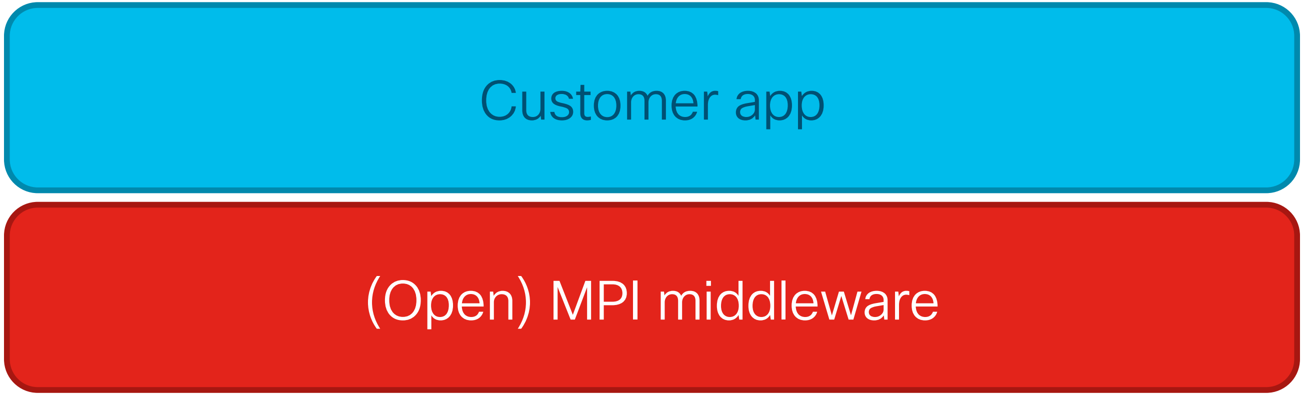
Messages, not bytes



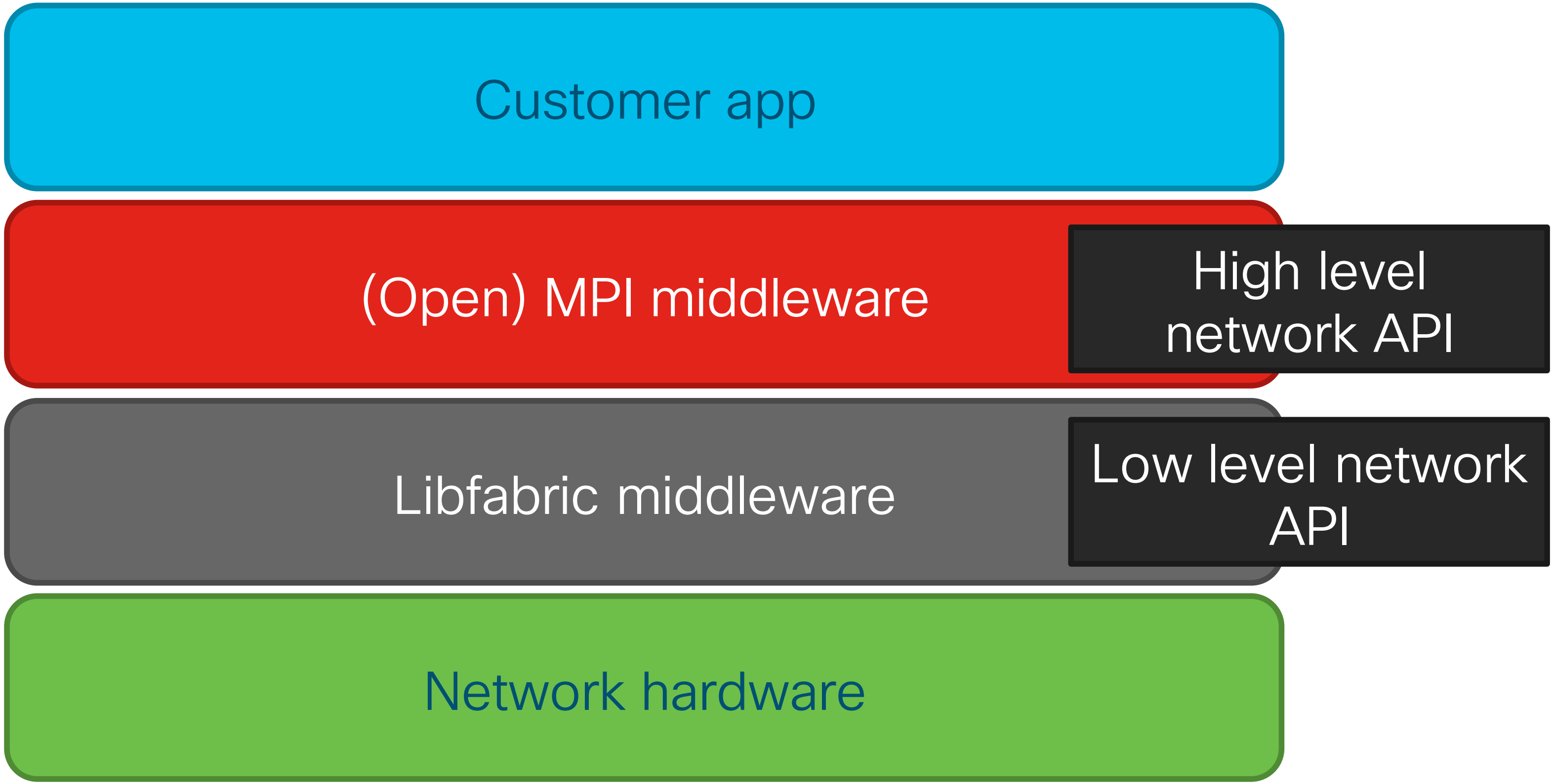
- MPI

- Based on peer integer "rank" (e.g., 8)
- Point-to-point and collective and one-sided and ...
- Message oriented
- **Typed messages**
- Network independent
- Blazing fast

MPI is the top layer of the API stack



But there's more below that



HPC Cluster Components

Traditional HPC cluster

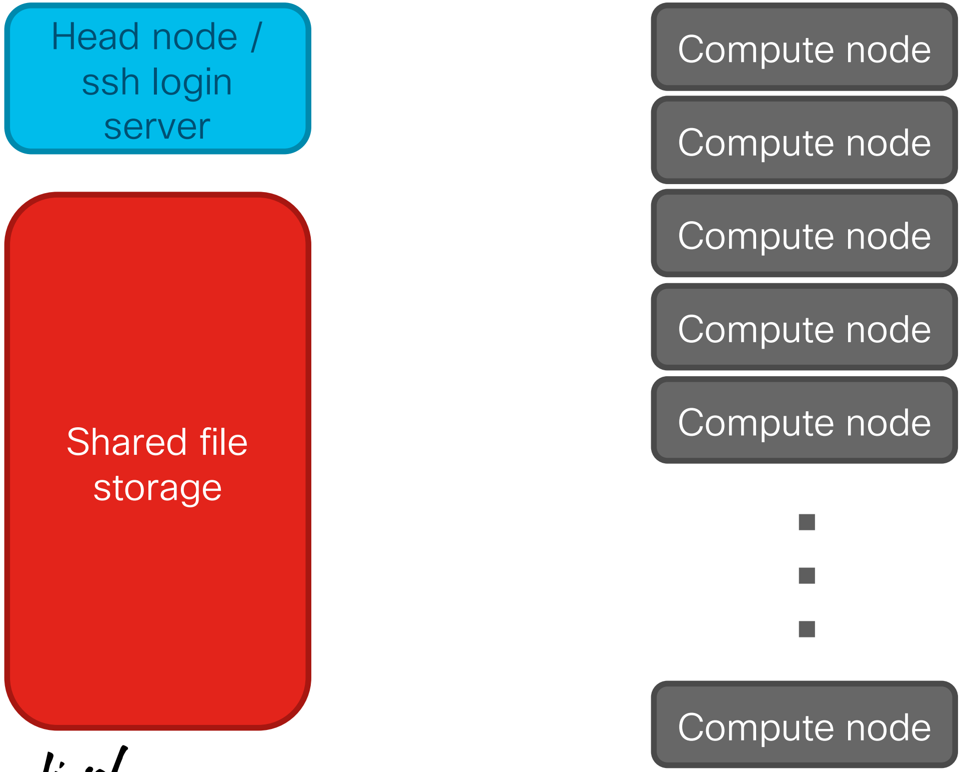
Head node / ssh login server

- Linux on bare metal
- Large RAM
- Used for interactive logins
  - Compile / build applications
  - Submit jobs
- Provides cluster job scheduler, NTP, ...etc.

Traditional HPC cluster

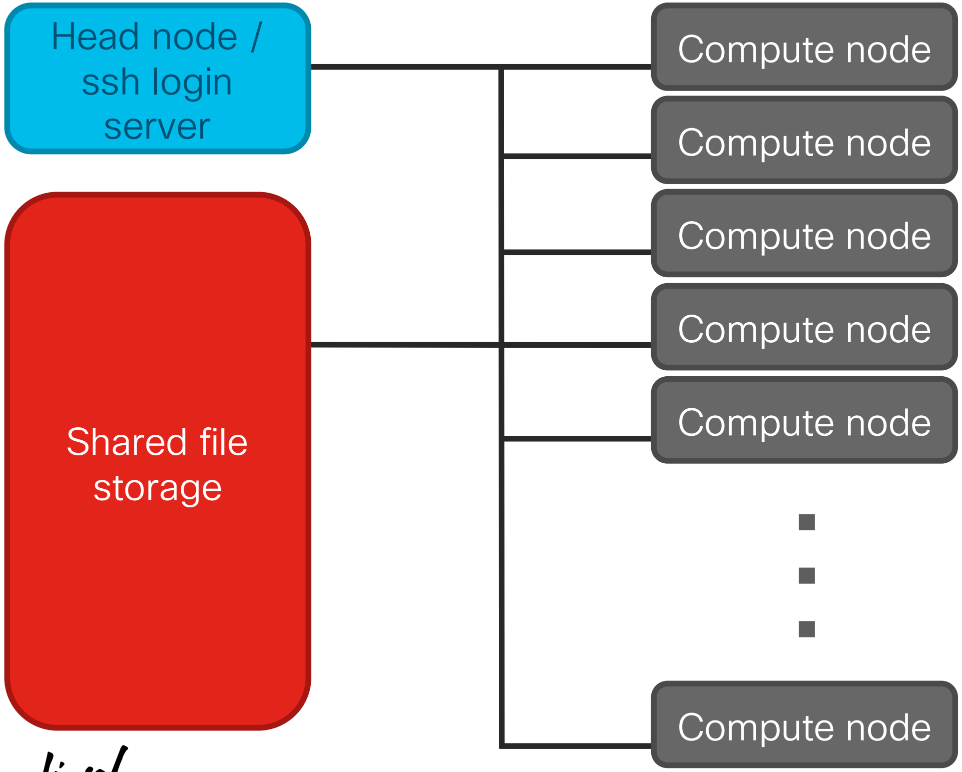
- Network file system
  - Possibly served from head node
  - Possibly have distinct file server(s)
  - May be traditional NFS
  - Or may be a parallel filesystem
- Available across entire cluster

Traditional HPC cluster



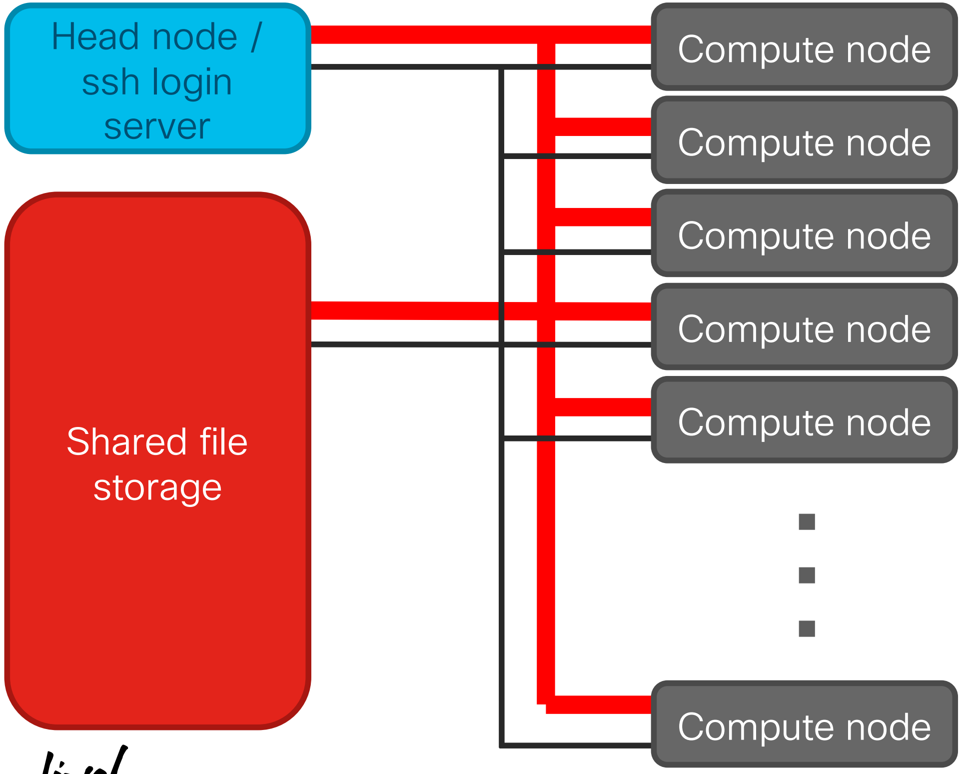
- Large number of compute nodes
- Each compute node:
  - High end Xeon processor
  - 2-8 GB RAM per core
  - Minimal local disk
  - (customers have different religions about the specifics)

Traditional HPC cluster



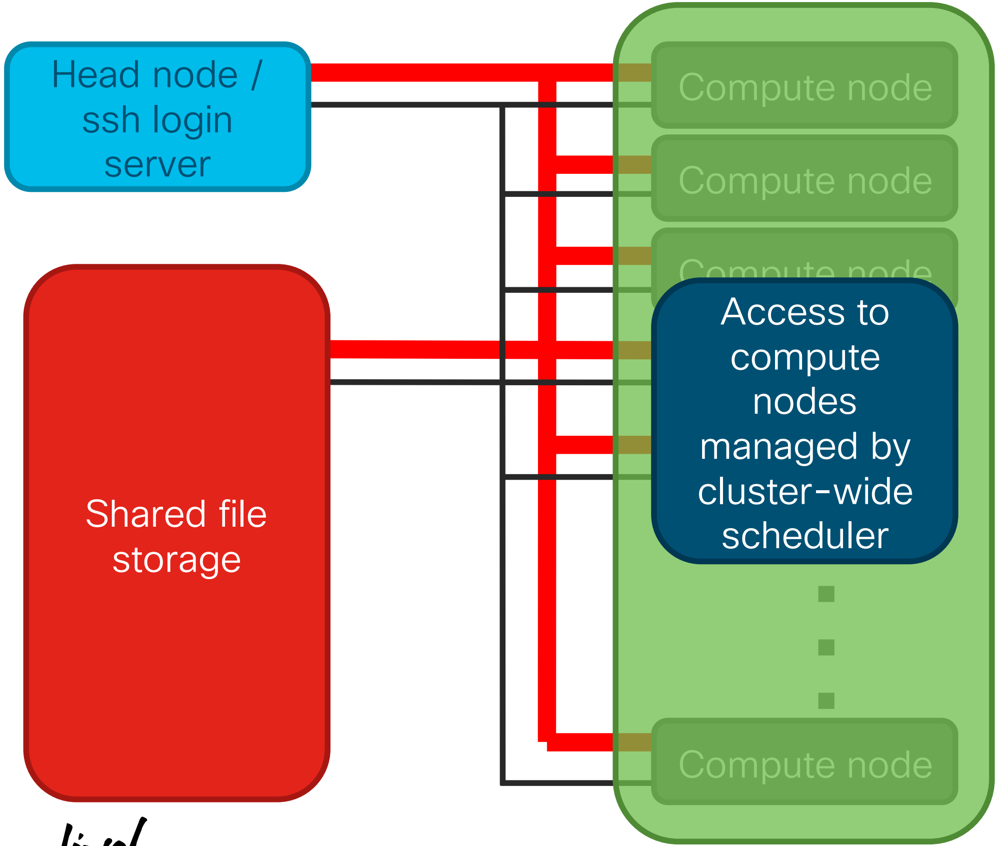
- Low end cluster
  - A single Ethernet/TCP/IP network
  - Possibly even 1 Gbps

Traditional HPC cluster



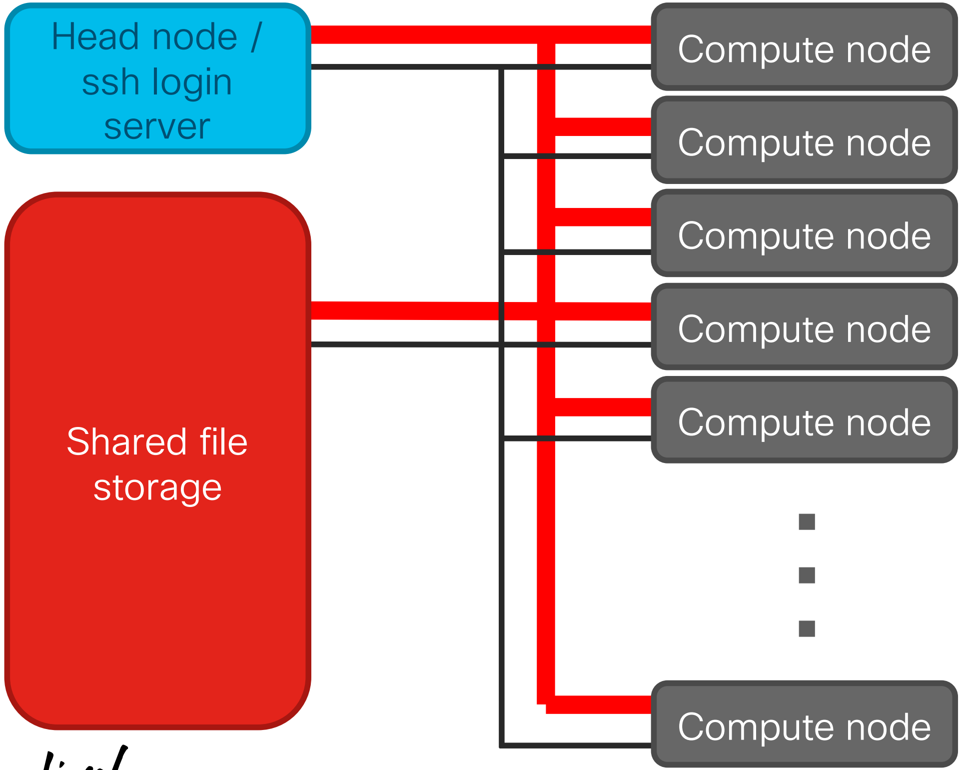
- Mid / high end cluster
  - Cheap Ethernet network for command / control (1/10/40GE)
  - Low latency network for east-west traffic, e.g., InfiniBand.

Traditional HPC cluster



- Scheduler to submit jobs
  - SLURM, LSF, Torque, PBS, ...
- OS/compute node provisioning
  - Bright, XCAT, LSF, ...
  - Some customers have home-grown solutions
  - Will NOT be virtual machines

Traditional HPC cluster

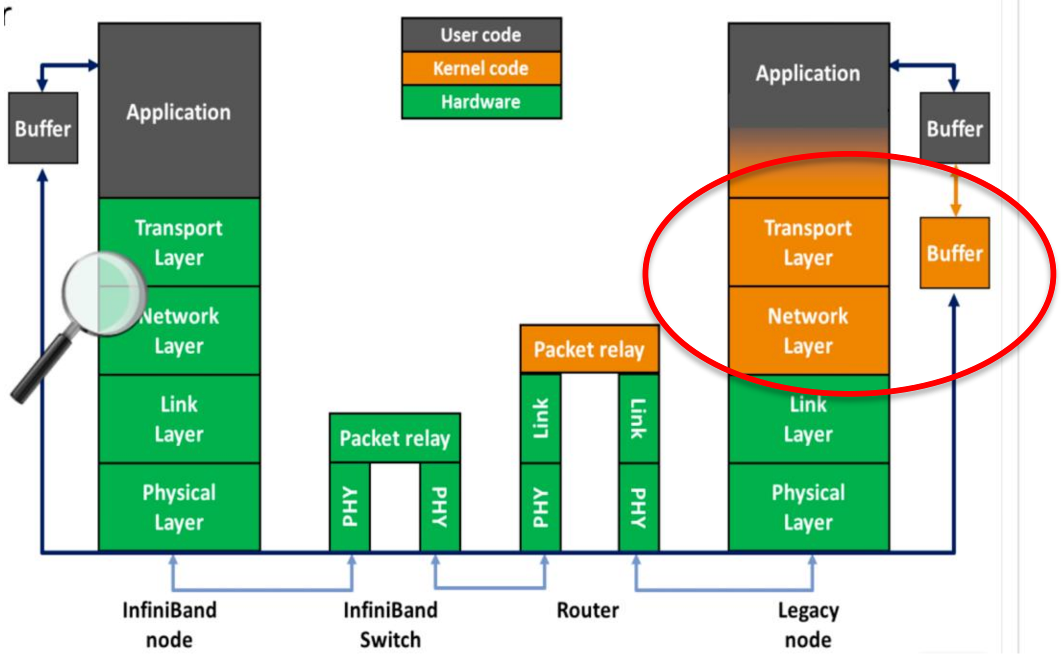


- General purpose clusters
  - Run all kinds of HPC applications
  - Usually for large institutions with many internal customers
- Specific purpose clusters
  - Run only one (or small set of) parallel application(s)
  - Usually a small set of users

HPC Fabric Types

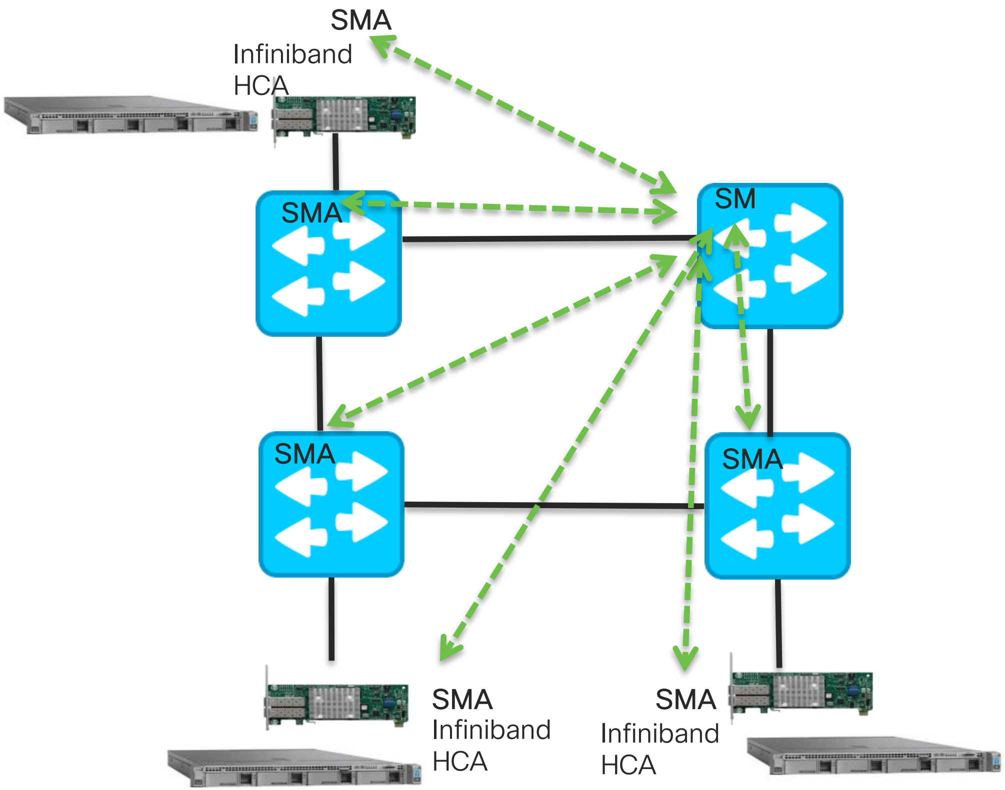
Infiniband

Infiniband Network Stack From a Host Perspective



Infiniband leverages *Kernel Bypass for it does not use OS TCP/IP Stack.*
It reduces latency.

Infiniband – Simple L2 Routing Network



Subnet Manager (SM) – Runs on switch or on server. It communicates with Subnet Manager Agents (SMA). SM can be thought as a master or a controller.

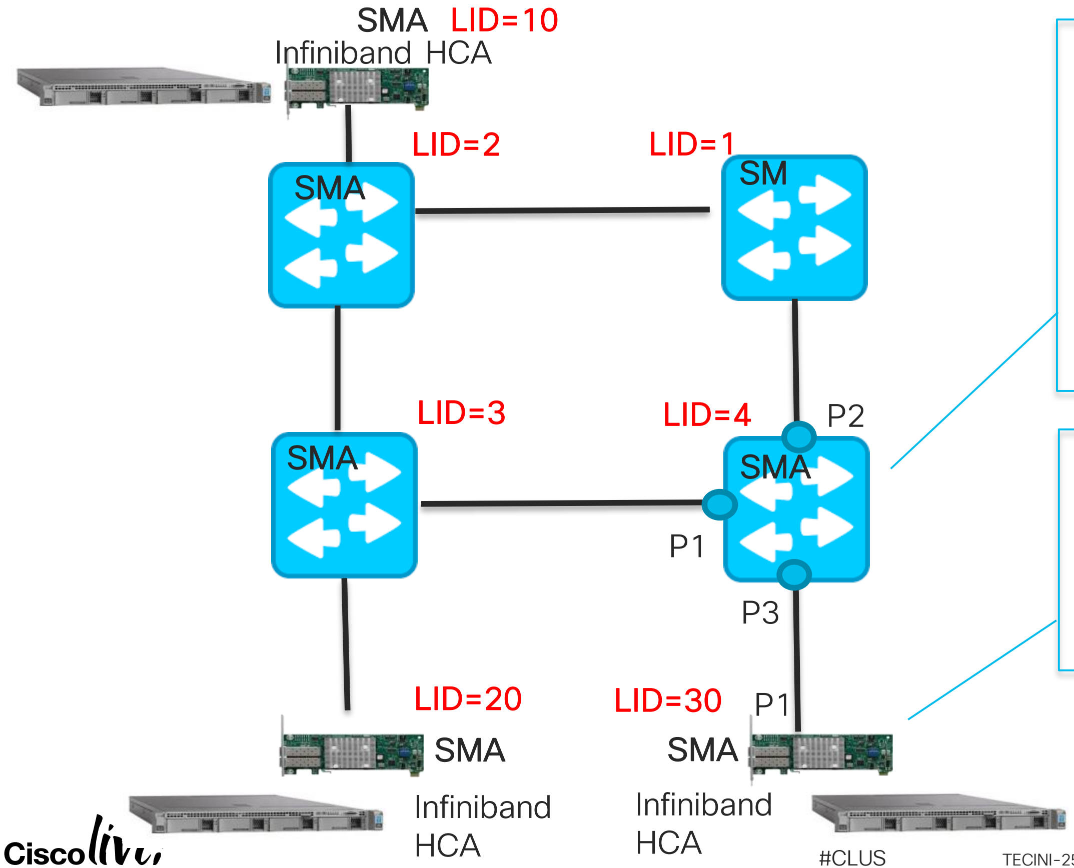
Subnet Manager Agent (SMA) – Runs on switches and HCA. And it communicates with SM. SM can be thought as SMA as slaves.

SM assigns 16 bit Link ID (LID) to all switches and Host Channel Adapters (HCA) – LID is a L2 address (think as MAC Address).

SM calculates L2 routing
Several routing algorithm available. Popular one is shortest-path routing (Dijkstra algorithm).

SM pushes the routing table to SMA.

Infiniband – Sample Routing



Routing Table

Destination LID	Output Port
10	P1, P2
20	P1
30	P3

Routing Table

Destination LID	Output Port
*	P1

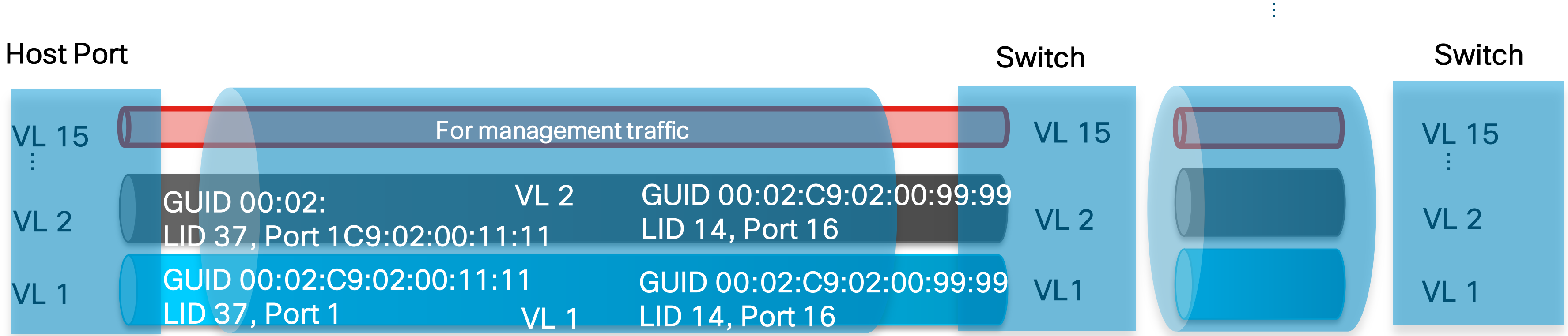
Infiniband Network QoS

Virtual Lanes

- Each Virtual Lane uses different buffers to send its packet toward the other side.
- VL 15 is Subnet Manager (mgmt) traffic only
- VL 0-7 are used for traffic

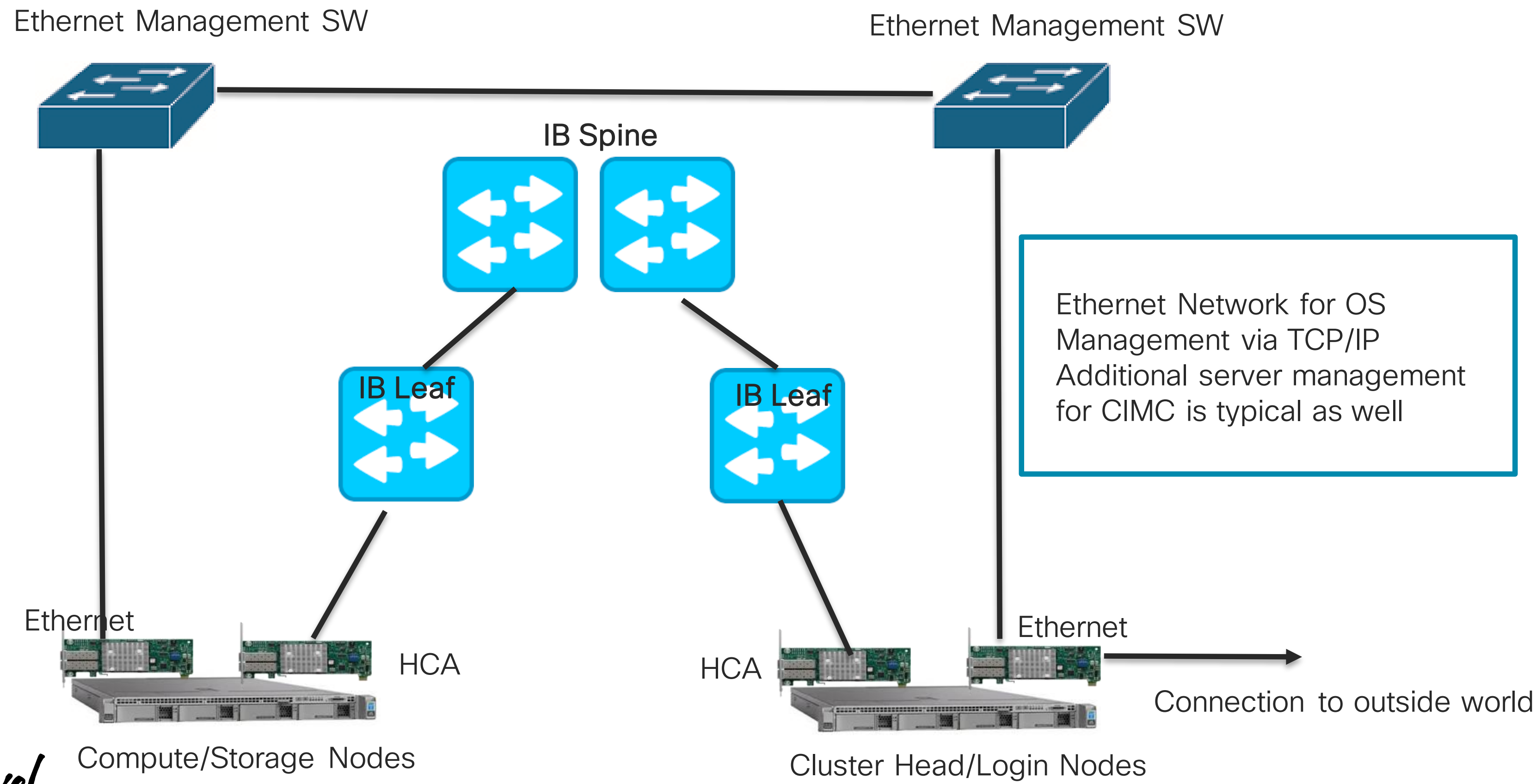
Credit Based Flow Control

- Receiver calculates credit limit sends the info to the Transmitter.
- Transmitter only transmit packets if there are enough credit limits to the packets it can send.



*GUID = 128 bit L3 Address analogous to IP subnet.

Typical Infiniband Network

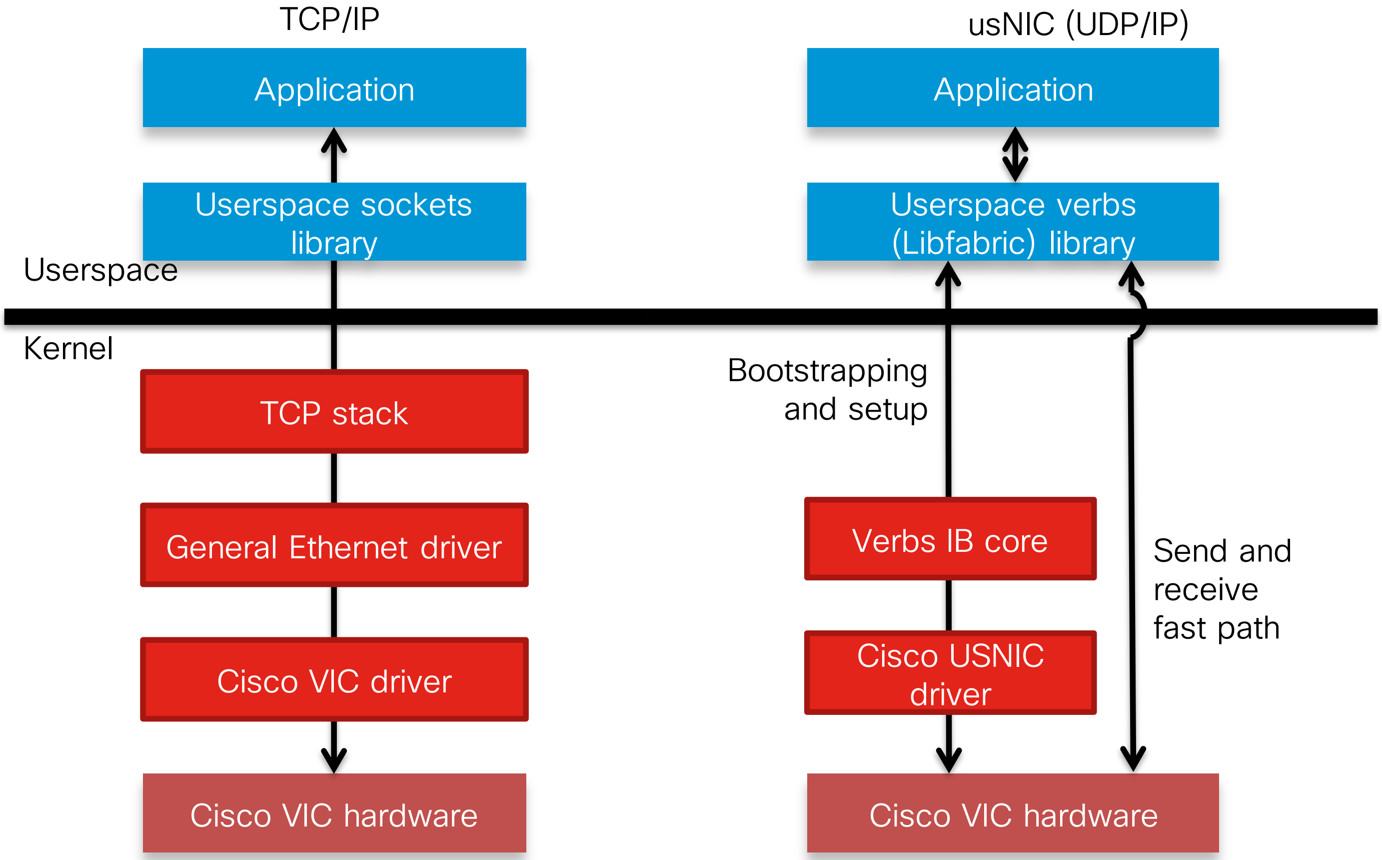


Ethernet Network for OS Management via TCP/IP
Additional server management for CIMC is typical as well

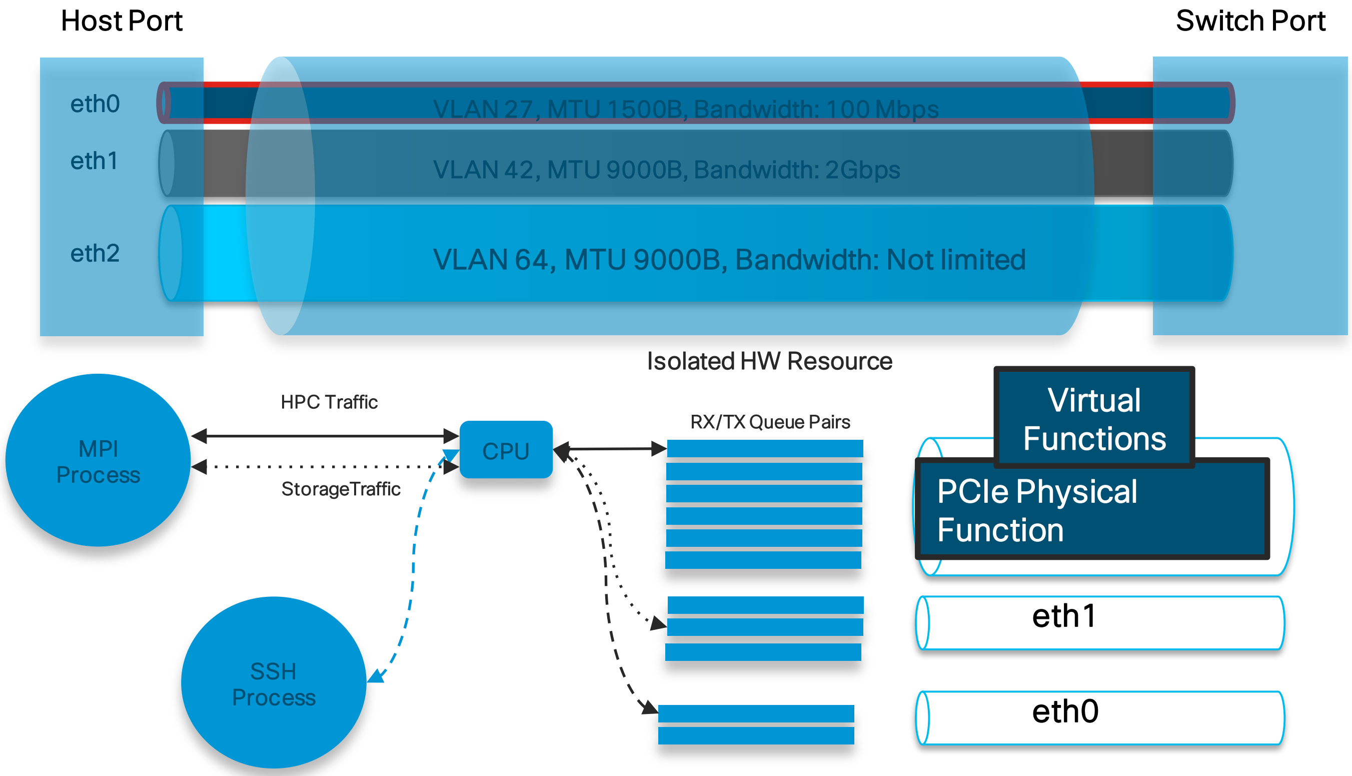
Ethernet is Here

- Ethernet is most widely used L2 Technology - networking can be managed by IT group.
- Ethernet is growing rapidly with many options (1/10/25/40/50/100G) - Infiniband vendors expanding Ethernet solutions
- More and more ultra-low latency Ethernet NICs and switches options are available
- Most Ethernet switches use store and forward operation. For low latency cut-through switching is employed.
- Ethernet in HPC fabric typically means “TCP/IP” thus “slow”.

Cisco Ethernet/usNIC OS bypass

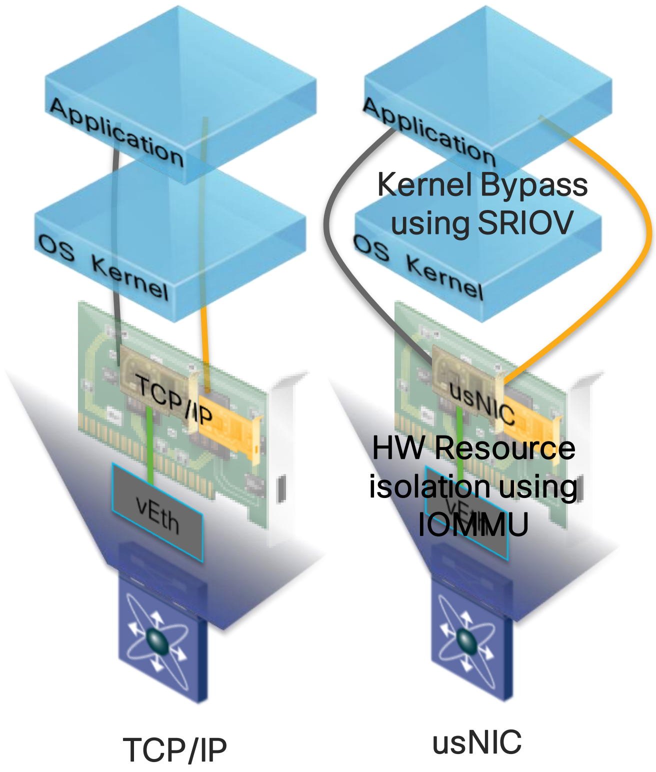


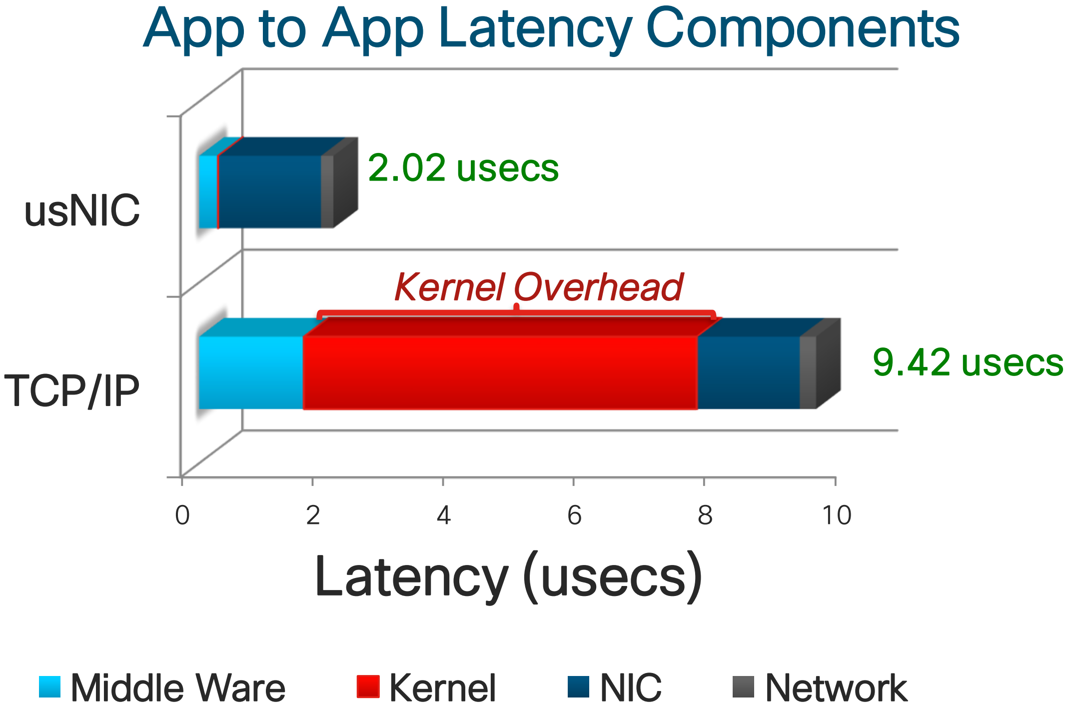
usNIC Network QoS



Up to 256 vNIC can be created.

App to App Latency Factors





Dual Functionality! -> We can use single cable for management and fabric traffic

Networking Latency Comparison

Cisco C-Series vs. B-Series

	C-Series with Nexus 3548	C-Series with Nexus 9k	B-Series with UCS 62xx
User Application Open MPI UD libfabric usNIC	630 ns Each PCIe Gen2 link traverse imposes a minimum of 90 nsecs latency Each Main Memory access imposes a minimum latency of 80 nsecs for local memory and ~ 120 nsecs for remote memory Most of the latency in this layer comes from system delays related to DMA transactions		
VIC Hardware	1200 ns		
Cisco Nexus	190 ns	450 ns	3500 ns

RDMA over Converged Ethernet (RoCE)

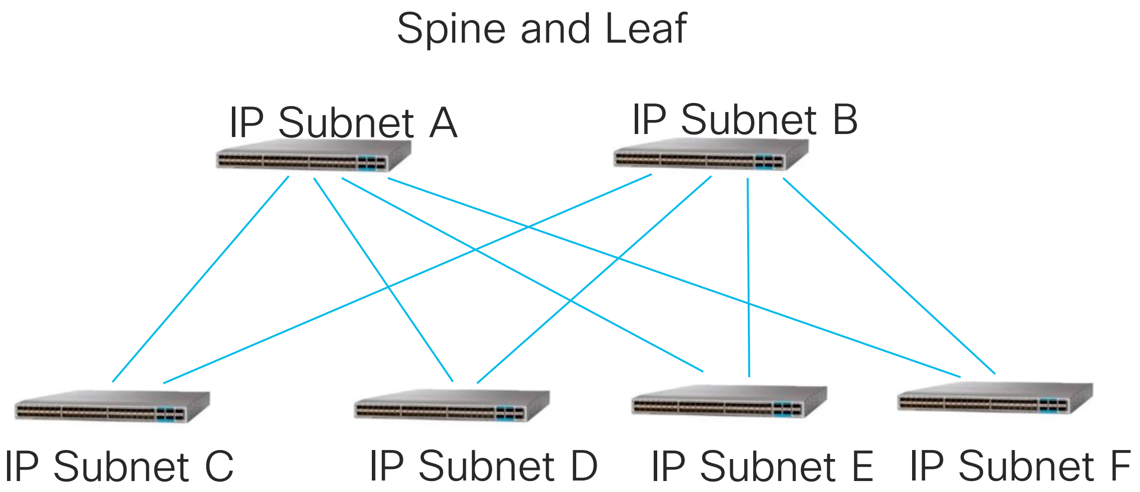
*Existing Infiniband based HPC applications do not need modifications

- Customer app
- MPI middleware
- Open Fabric Enterprise Distribution
- Network hardware/Ethernet

RDMA supports zero-copy networking by enabling the network adapter to transfer data directly to or from application memory, eliminating the need to copy data between application memory and the data buffers in the operating system.

The RoCEv2 protocol exists on top of either the UDP/IPv4 or the UDP/IPv6 protocol.[2] The UDP destination port number 4791 has been reserved for RoCE v2.[10] Since RoCEv2 packets are routable the RoCE v2 protocol is sometimes called Routable RoCE[11]

What is RoCE v2



*Because RoCE v2 is IP routable, the topology can be fully meshed.

Nexus PFC Support for RoCE

```
switch (config)# interface Ethernet1/1
switch (config-if)# flowcontrol receive on
switch (config-if)# flowcontrol send on
switch (config)# interface Ethernet1/1
switch (config-if)# flowcontrol receive on
switch (config-if)# flowcontrol send on
switch (config)# class-map type qos RDMA        <-- RDMA is just the name of the class
switch (config-cmap-qos)# match cos 0           <-- Classify all traffic with CoS 0, in this case untagged traffic.
switch (config-cmap-qos)# exit
```


Nexus PFC Support for RoCE

switch (config)# policy-map type network qos QOS_NETWORK — <-- QOS_NETWORK is just the name of the policy

switch (config-pmap-nqos)# class type network-qos RDMA — <-- Use the RDMA class as the traffic that will be policed in this case.

switch (config-pmap-nqos-c)# pause pfc-cos 0 — <-- Pause all classified traffic that has CoS 0

switch (config-pmap-nqos-c)# exit

switch (config-pmap-nqos)# exit

switch (config)# system qos

switch (config)# service-policy type network-qos QOS_NETWORK — <-- Apply QOS_NETWORK policy to the switch (all ports)

Agenda

- Introduction
- Anatomy of Research Computing
- **Demo and Hands-on Cluster Operations**
- UCS as a Platform for Research Computing
- Hands-on Bonus: UCS Configuration with Ansible
- NVIDIA Deep Learning with GPUs
- Conclusion

Cluster Management and Scheduling Demonstration

Agenda

- Introduction
- Anatomy of Research Computing
- Demo and Hands-on Cluster Operations
- **UCS as a Platform for Research Computing**
- Hands-on Bonus: UCS Configuration with Ansible
- NVIDIA Deep Learning with GPUs
- Conclusion

Agenda

- Introduction
- Anatomy of Research Computing
- Demo and Hands-on Cluster Operations
- **UCS as a Platform for Research Computing**
- Hands-on Bonus: UCS Configuration w/Ansible
- NVIDIA Deep Learning with GPUs
- Conclusion

- Server Overview
- GPU Overview
- UCS Infrastructure
- UCS Advantages
- UCS Designs

Servers

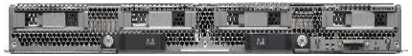
Unified Computing Systems

M5 – Blade Servers

UCS B200 M5

Ideal for Bare Metal Enterprise, VDI, or Dense Virtualization/ Consolidation Workloads

- Up to two Intel Xeon Processor Scalable Family (Max: 56 Cores total)
- 24 DIMMs, up to 3TB
- Up to 80 Gbps I/O
- Up to 2 NVIDIA P6 GPUs
- Up to 2 SD/M.2 cards
- Up to 2 SFF HS drives



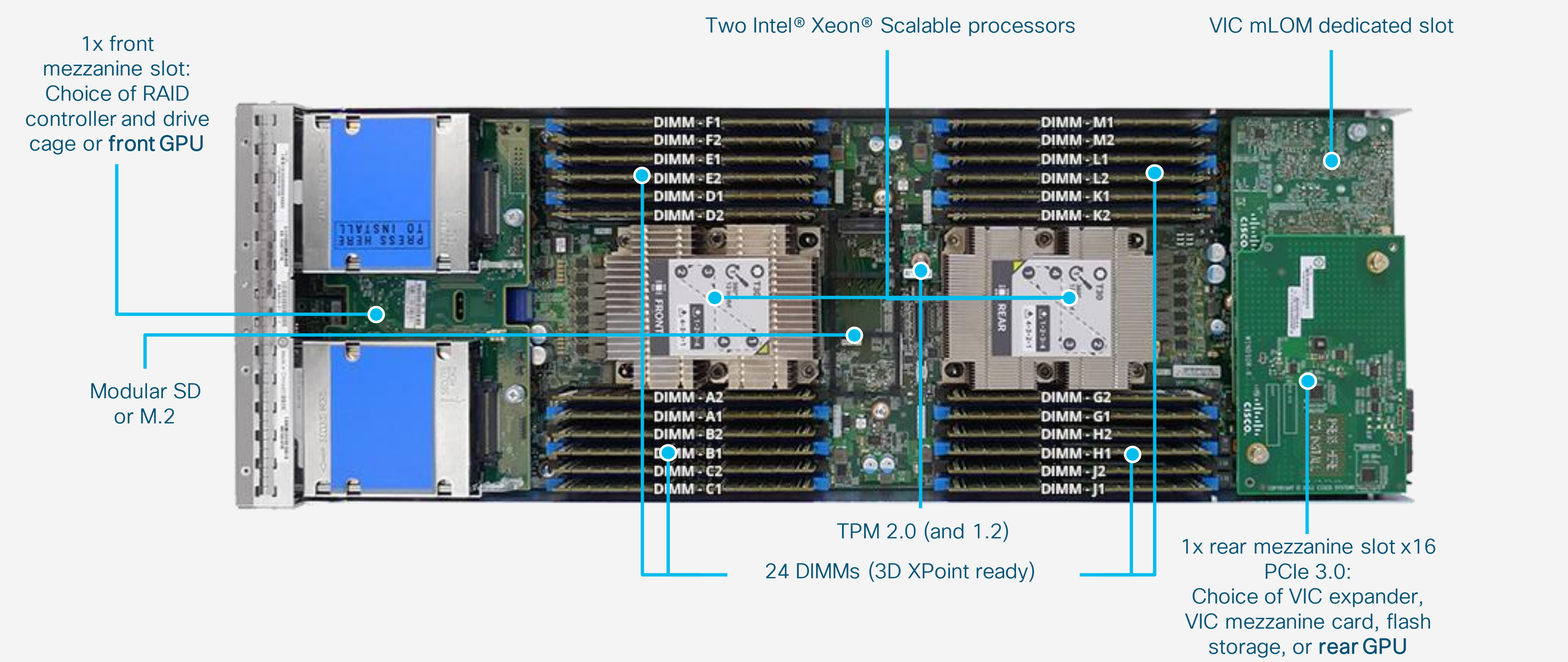
UCS B480 M5

High Performance Platform designed for Compute-Intensive and Memory-Intensive Enterprise Workloads

- Up to four Intel Xeon Processor Scalable Family (Max: 112 Cores total)
- 48 DIMMs, up to 6TB
- Up to 160 Gbps I/O
- Up to 4 NVIDIA P6 GPUs
- Up to 4 SD/M.2 cards
- Up to 4 SFF HS drives

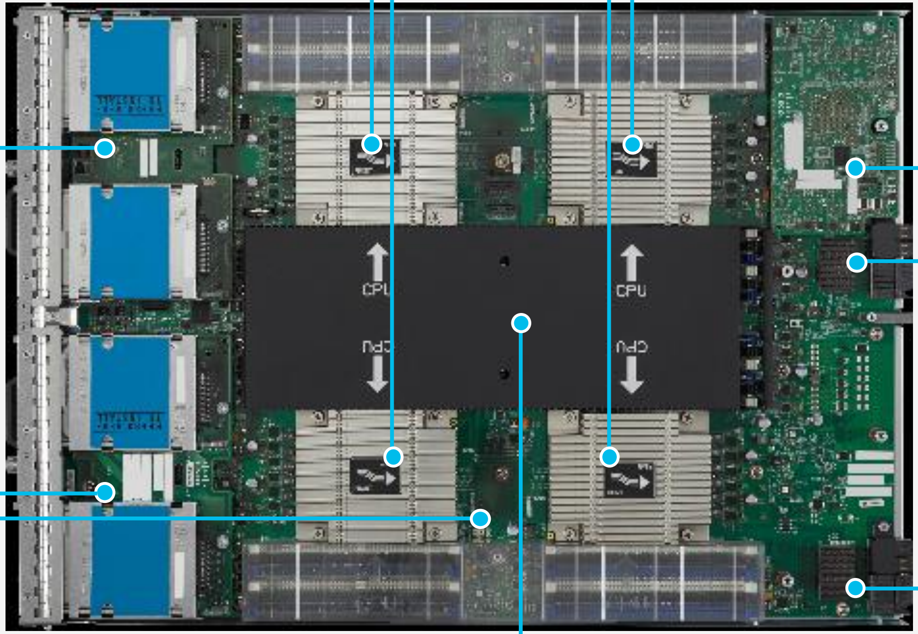
Unified Computing Systems

B200 M5



Unified Computing Systems

B480 M5



UCS Rack Servers

M5 – C-Series

UCS C220 M5

Optimal performance in a dense 1RU form factor for a wide range of workloads including virtualization, web, collaboration, cloud, and bare metal applications

- Up to two Intel Xeon Processor Scalable Family (Max: 56 Cores total)
- 24 DIMMs, up to 3TB
- 10 SFF or 4 LFF HS drives, NVMe PCIe SSD support
- 2 PCIe 3.0 x16 slots Support for 1 NVIDIA P4 GPU

UCS C240 M5

Optimal performance for a wide range of enterprise workloads including big data analytics, collaboration, databases, virtualization, and high performance applications

- Up to two Intel Xeon Processor Scalable Family (Max: 56 Cores total)
- 24 DIMMs, up to 3TB
- 24 SFF or 12 LFF HS drives, NVMe PCIe SSD support, 2 rear SFF HD drives
- 6 PCIe 3.0 slots, Support for 2 NVIDIA P100/V100 GPUs

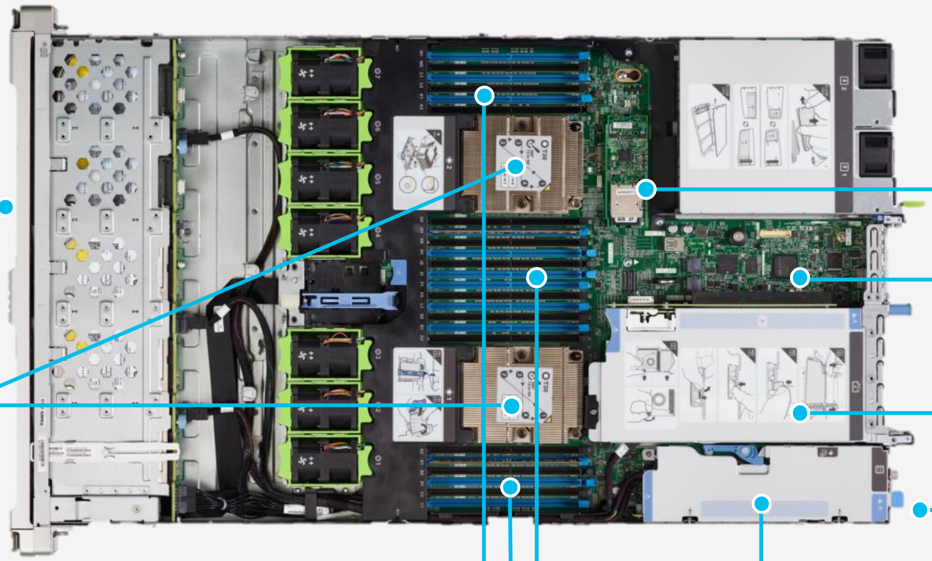
UCS C480 M5

Designed for the most demanding server workloads such as in-memory database, EDA, Machine Learning, Data Analytics, and CPU / GPU rendering

- Up to **FOUR** Intel Xeon Processor Scalable Family (Max: 112 Cores total)
- 48 DIMMs, up to 6TB
- 24 SFF HS drives, NVMe PCIe SSD support, 8 internal top loaded SFF drives
- 12 PCIe 3.0 slots, Support for 6 NVIDIA P100/V100 GPUs

Unified Computing Systems

C220 M5



Unified Computing Systems

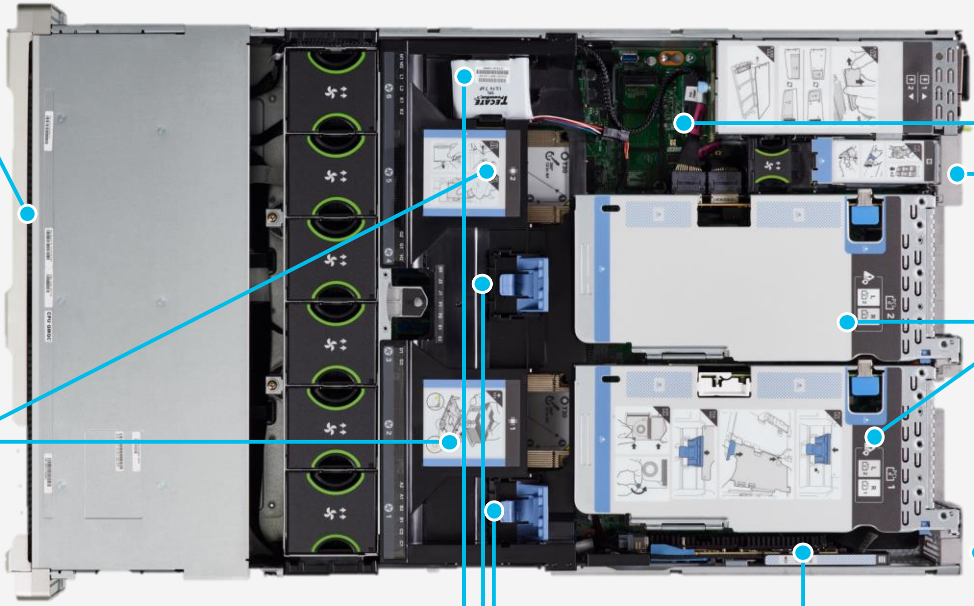
C240 M5

Up to 24 SFF or 12 LFF drives

Support for 4 NVMe PCIe SSDs in mixed configuration, or up to 10 NVMe PCIe SSDs (2 rear) with NVMe-optimized SKU

All SAS and SATA (not NVMe) drives managed by dedicated RAID controller and supported in Cisco UCS® Manager

2 x Intel® Xeon® Scalable processors



Modular SD FlexFlash or Modular M.2 (SATA or NVMe)

Up to 2 optional rear hot-swappable 2.5-inch drives

Can be NVMe PCIe SSDs or SAS, SATA, and SSD in standard SKUs

Up to 6 PCIe 3.0 slots

Up to 4 full-height,full-length x16

Support for 2 GPUs

x16 mLOM slot

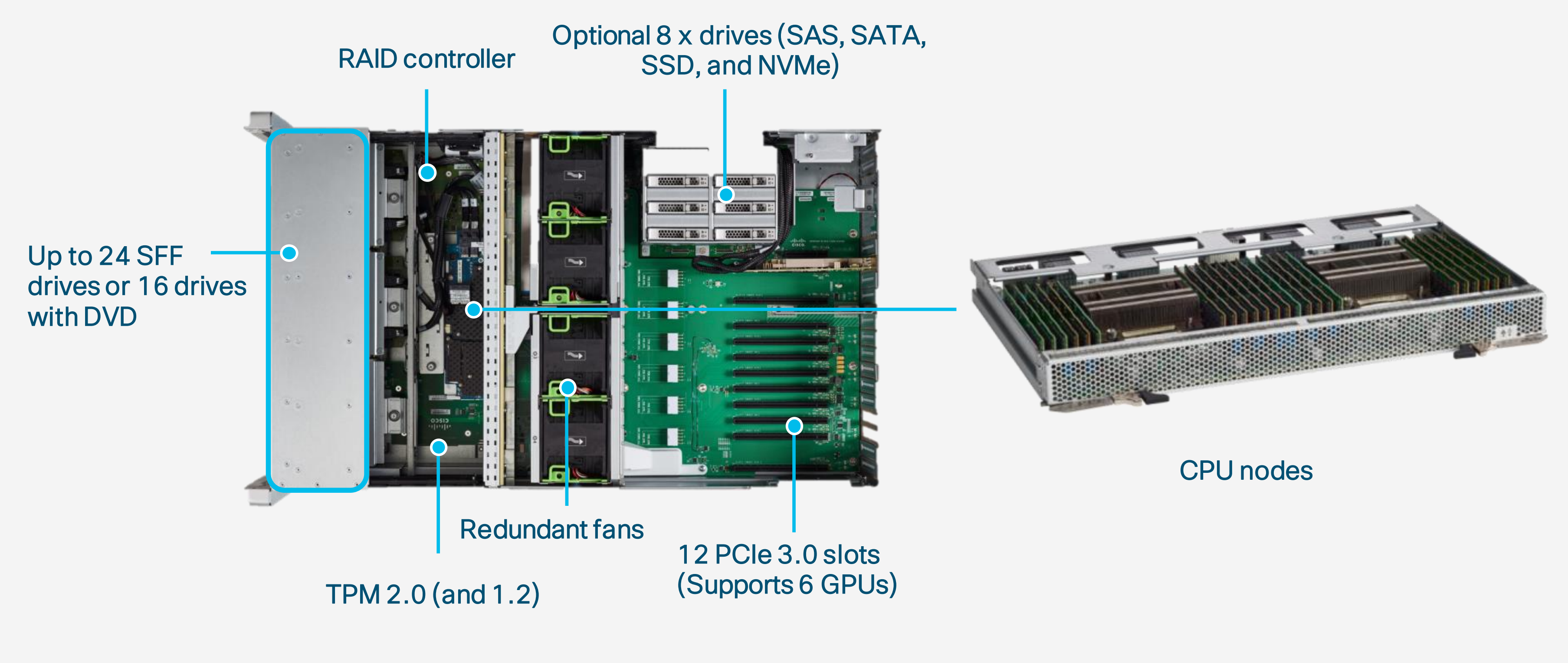
24 DIMM sockets

Up to 2666 MHz speeds

Dedicated RAID slot

Unified Computing Systems

C480 M5

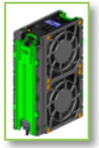


Cisco UCS S3260 Storage Server

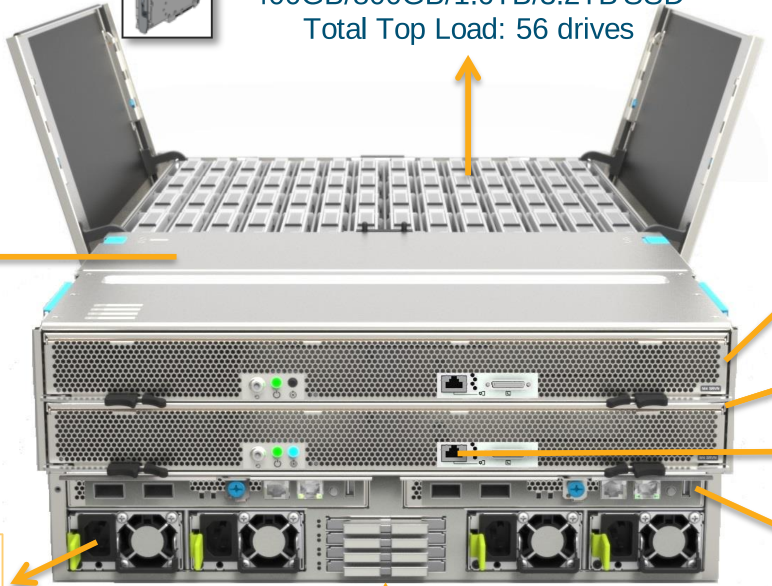
M5 Server Node

Drives
4 Rows of Hot-Swappable HDD
2/4/6/8/10/12TB with up to 2 Rows of
400GB/800GB/1.6TB/3.2TB SSD
Total Top Load: 56 drives

Server Node
Up to (2) Based on 2x Intel Scalable CPUs, LSI 12G Dual-Chip RAID or Pass-through HBA, 14x 16/32/64/128GB DDR4 RAM (2 Slots 3D XPoint Ready), 2x 500GB/1TB/2TB NVMe



FAN
8 Hot-Pluggable Fans



Optional Second Node
Server Node or Drive Expansion or PCIe Expansion

1Gb Host Management Port

Power Supply
4 Hot-Pluggable PSUs

System I/O Controller (SIOC)
Up to (2) Cisco VIC 1300(SIOC2)/ 1400(SIOC3*) or 3rd Party Adapters*
Up to (2) NVMe per SIOC

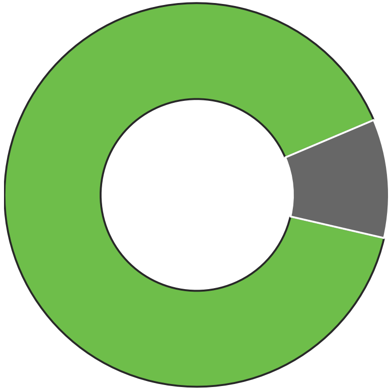
Up to (4) 240GB/480GB/1.6TB SSDs
HW RAID, Hot-Plug, OS/Boot

*Post FCS

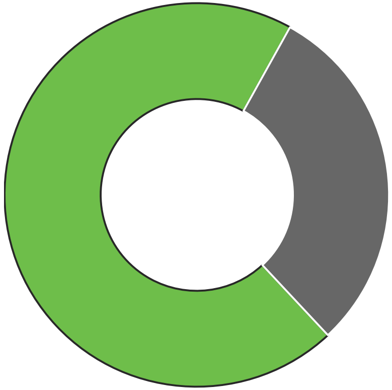
GPUs

70% Of Top HPC Apps Accelerated

INTERSECT360 SURVEY OF TOP APPS



9 of top 10 Apps Accelerated



35 of top 50 Apps Accelerated

Intersect360, Nov 2015
"HPC Application Support for GPU Computing"

TOP 25 APPS IN SURVEY

● (All popular functions accelerated):
- GROMACS
- SIMULIA Abaqus
- NAMD
- AMBER
- ANSYS Mechanical
- Exelis IDL
- MSC NASTRAN

● (Some popular functions accelerated):
- LAMMPS
- NWChem
- LS-DYNA
- Schrodinger

● (In development):
- Gaussian
- GAMESS

● (Some popular functions accelerated):
- ANSYS Fluent
- WRF
- VASP
- OpenFOAM
- CHARMM
- Quantum Espresso

● (Not supported):
- ANSYS CFX
- Star-CD
- CCSM
- COMSOL
- Star-CCM+
- BLAST

● = All popular functions accelerated
● = Some popular functions accelerated
● = In development
● = Not supported

UCS GPU Portfolio

UCS Integrated for accelerated VDI, Deep Learning, and HPC Applications

Virtualization

UCS M6 Blade GPU
For M4 blades only
Enterprise Class NVIDIA GPU for Remote Knowledge Workers, Task Workers and Designers

UCS P6 Blade GPU
For M5 blades only
Doubles user density for Remote Knowledge Workers, Task Workers and Designers

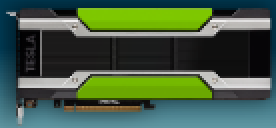
UCS NVIDIA M60
For M4 and M5* server
Ultimate choice for Remote Engineering Workstations and Application Delivery via the Cloud

UCS NVIDIA M10
For M4 and M5 servers
Accelerated Remote Desktop, Maximum User Density per Server

Accelerated Compute

UCS Tesla P100
For M4 and M5
Deep Learning Training and HPC applications

UCS Tesla V100
**For M5 server only
World's most advanced data center GPU ever built to **accelerate AI** and HPC



UCS Tesla P40
For M5 server only
Remote Engineering Workstations and Fast Inferencing for Deep Learning

UCS Tesla P4
For M5 server only
Inferencing Engine for Deep Learning at the edge.

*M5 support for M60 in 3.2(2c) ** Needs UCSM 3.2(3)

Blade Optimized GPU

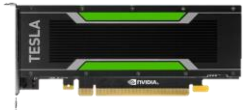
Industry's Only 2-S Blade to offer multiple GPU, densest GPU platform
16 GPU in 6RU Chassis

PID	Description
UCSB-GPU-P6-F	NVIDIA GRID P6 Front Mezzanine
UCSB-GPU-P6-R	NVIDIA GRID P6 Rear Mezzanine

	TESLA M6	TESLA P6
GPUs	**Single** GM204	**Single** GP104
CUDA Cores	1,536	2,048
Memory Size	8 GB GDDR5	16 GB GDDR5
Form Factor	MXM (blade server)	MXM (blade server)
Thermal	bare board	bare board
Power	100W (75W opt)	90W (70W opt)
No. of GPU per B200M5	1	2
Max Concurrent Users	**8 (1GB FB)**	**16 (1GB FB)**
H.264 1080p30 Streams	16	24
3DMark 11	10,558	17,600
SPECviewperf 12	45	75
SGEMM TFLOPS	2.7	4.7
Memory Bandwidth	147 GB/s	192 GB/s

Rack Optimized GPUs

Full UCSM Integration

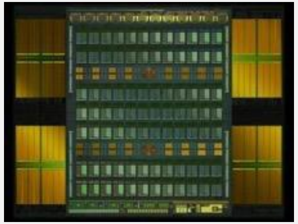


	Tesla P4 for Ultra-Efficient Scale-Out Servers	Tesla P40 for Maximum-Inference Throughput Servers	Tesla P100: The Universal Datacenter GPU
Single-Precision Performance (FP32)	5.5 TeraFLOPS	12 TeraFLOPS	10.6 TeraFLOPS
Half-Precision Performance (FP16)	--	--	21 TeraFLOPS
Integer Operations (INT8)	22 TOPS*	47 TOPS*	--
GPU Memory	8 GB	24 GB	16 GB
Memory Bandwidth	192 GB/s	346 GB/s	732 GB/s
System Interface	Low-Profile PCI Express Form Factor	Dual-Slot, Full-Height PCI Express Form Factor	Dual-Slot, Full-Height PCI Express Form Factor, or SXM2 Form Factor with NVLink
Power	50 W/75 W	250 W	250 W (PCIe) 300W (SXM2)
Hardware-Accelerated Video Engine	1x Decode Engine, 2x Encode Engines	1x Decode Engine, 2x Encode Engines	--

*Tera-Operations per Second with Boost Clock Enabled

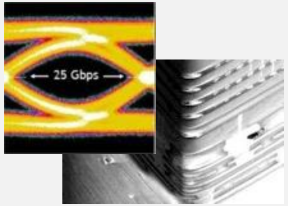
INTRODUCING TESLA V100

Volta Architecture



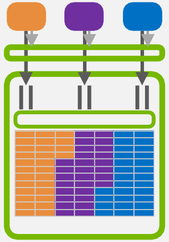
Most Productive GPU

Improved NVLink & HBM2



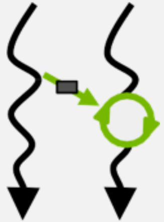
Efficient Bandwidth

Volta MPS



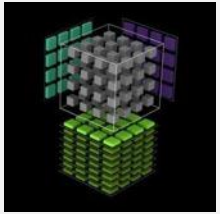
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core



120 Programmable TFLOPS Deep Learning

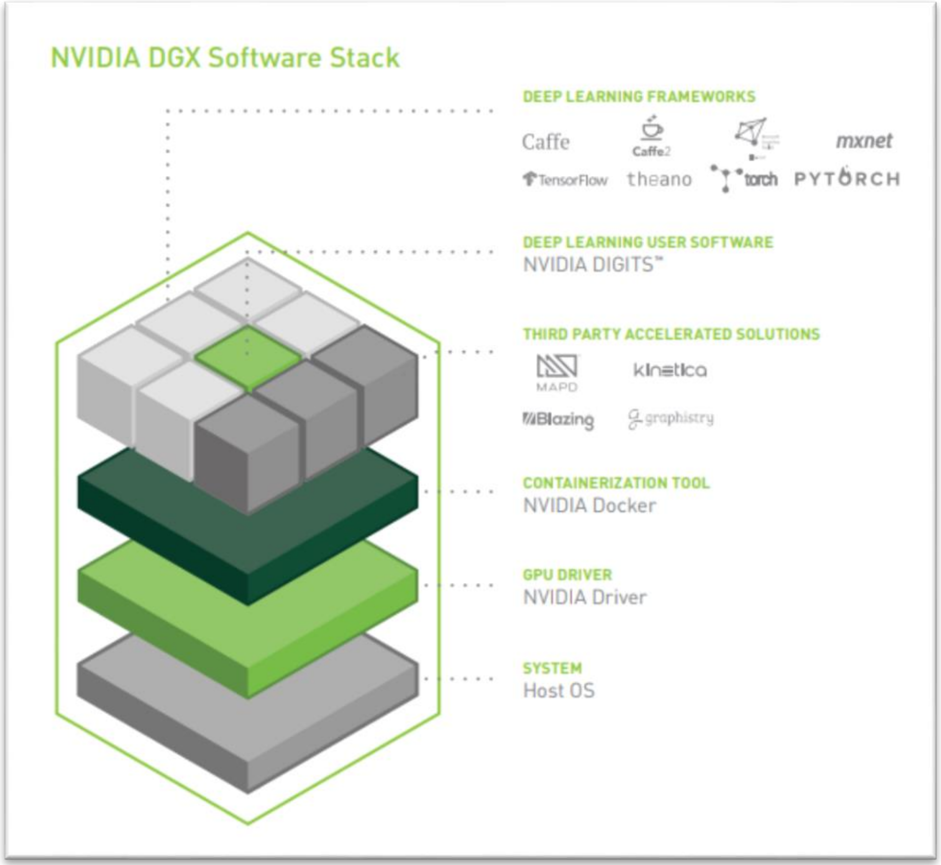
5120 CUDA Cores, 640 Tensor Cores

The Fastest and Most Productive GPU for Deep Learning and HPC

GPU Performance Comparison

	P100	V100	Ratio
Training acceleration	10 TOPS	120 TOPS	12x
Inference acceleration	21 TFLOPS	120 TOPS	6x
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS	1.5x
HBM2 Bandwidth	720 GB/s	900 GB/s	1.2x
NVLink Bandwidth	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

NVIDIA DGX-1



SYSTEM SPECIFICATIONS

GPUs	8X Tesla V100	8X Tesla P100
TFLOPS (GPU FP16)	960	170
GPU Memory	128 GB total system	
CPU	Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz	
NVIDIA CUDA® Cores	40,960	28,672
NVIDIA Tensor Cores (on V100 based systems)	5,120	N/A
Maximum Power Requirements	3,200 W	
System Memory	512 GB 2,133 MHz DDR4 LRDIMM	
Storage	4X 1.92 TB SSD RAID 0	
Network	Dual 10 GbE, 4 IB EDR	
Software	Ubuntu Linux Host OS See Software Stack for Details	

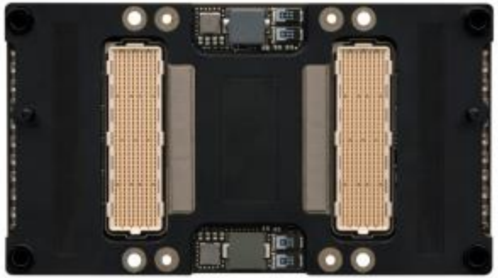
NVIDIA DGX-1

SXM2 and PCIe GPU's Compared

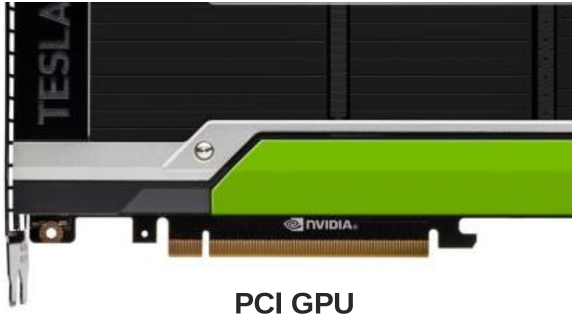
	SXM2	PCIe
Form Factor		
Performance	>6.5 TF DP, >13 TF SP, >26 TF FP16	~6 TF DP, ~12 TF SP, ~24 TF FP16
Memory Size	32 GB HBM2	32 GB HBM2
Memory Bandwidth	~1000 GB/s	~1000 GB/s
GPU Peer to Peer	Scalable, Coherent, NVLink	PCIe Gen3
Power	300 W	250W

NVLINK 2.0 = 300 GBytes/s

PCIe Gen 3 = 32 GBytes/s



SXM GPU back
Connects directly to SXM slot on motherboard



PCI GPU
Connects to one of the available PCI slots

Summary of Cisco UCS AI Solutions

DGX-1 (Resale)
- Starter appliance for research group/ data scientist
- Test multiple Framework easily
- Cloud managed

Training

SXM2/NVLINK

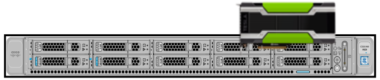
PCIe

C240 M5 + OSS EB3600 (PCIe)
9 x P100/V100 GPUs

UCS C480 M5 (PCIe)
6 x P100/V100 GPUs

Inferencing

C240 M5 + 2xP40
More inferencing GPU with more storage (DC)



C220 M5 + P4
UCS Inferencing appliance for Edge

UCS Infrastructure

Unified Computing System Architecture (Blades)

- UCS Manager
  Embedded– manages entire system

- UCS 6300 Series Fabric Interconnect
  32 x 40GB QSFP+ Port or 24 x 40GB QSFP+ Port & 16 UP Ports

- UCS Fabric Extender – UCS 2x00 Series
  Remote line card

- UCS 5100 Series Blade Server Chassis
  Flexible bay configurations

- UCS B-Series Blade Server
  Industry-standard architecture

- UCS Virtual Adapters
  Choice of multiple adapters

Unified Computing Systems

Fabric Interconnects

FI 6332

- 32 x 40GbE QSFP+ ports
- 2.56Tbps switching performance
- 1RU fixed form factor, two power supplies & four fans

FI 6332-16UP

- 24 x 40GbE QSFP+ & 16 x UP ports (1/10GbE or 4/8/16G FC)
- 2.43Tbps switching performance
- 1RU fixed form factor, two power supplies & four fans

Unified Computing Systems

Fabric Extenders

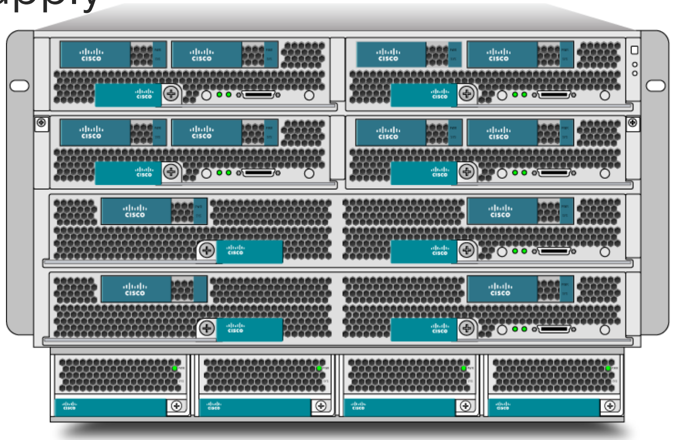
IOM 2304

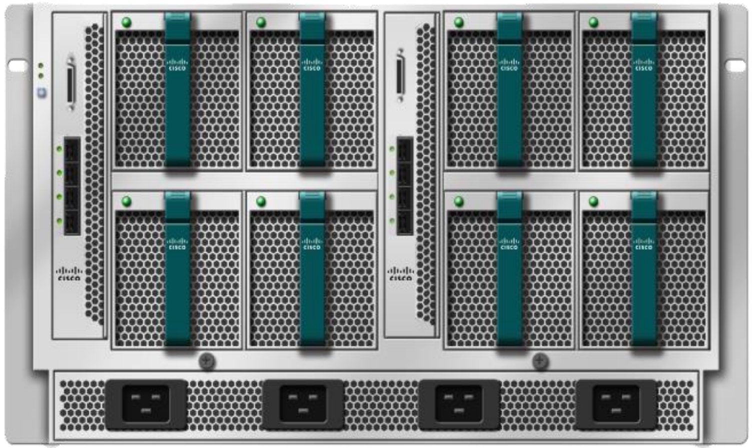
- 8 x 40GbE server links & 4 x 40GbE QSFP+ uplinks
- 960Gbps switching performance
- Modular IOM for UCS 5108

Unified Computing Systems

Blades Chasis

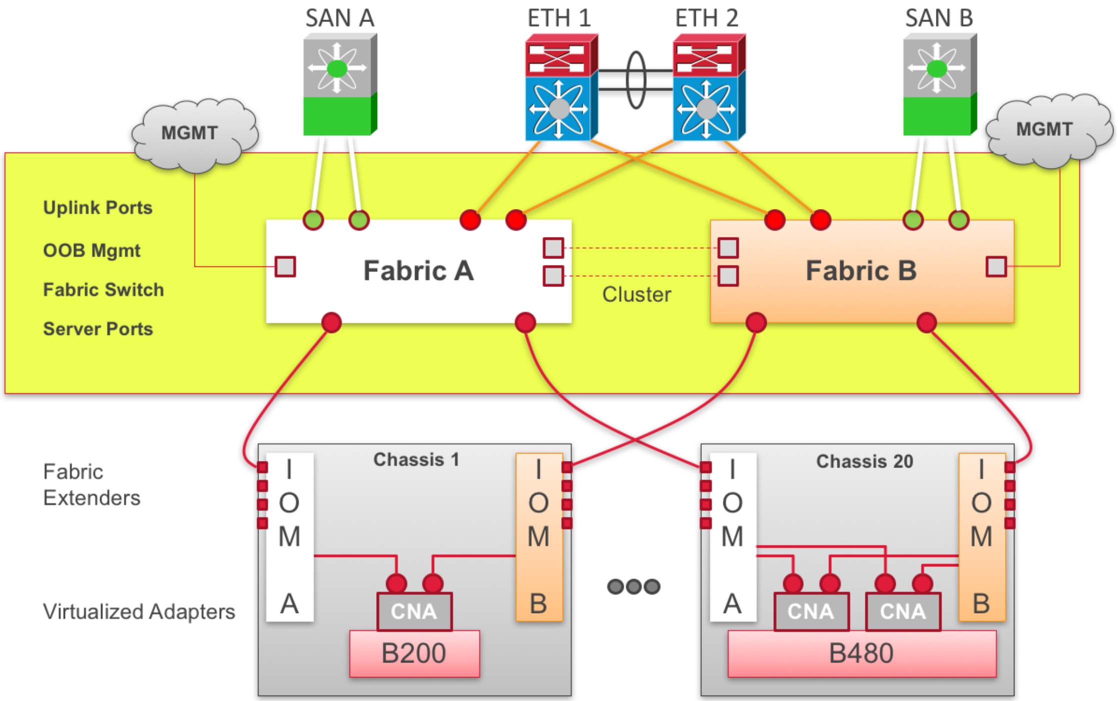
- Up to 8 Half width server blades
- Up to 4 Full width server blades
- Redundant Hot Swap Power Supply
- 6RU Enclosure
- 2 Hot Swap FEX
- 8 Hot Swap Fan Module





Unified Computing Systems

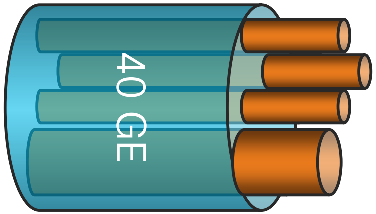
Distributed Architecture, Unified Fabric



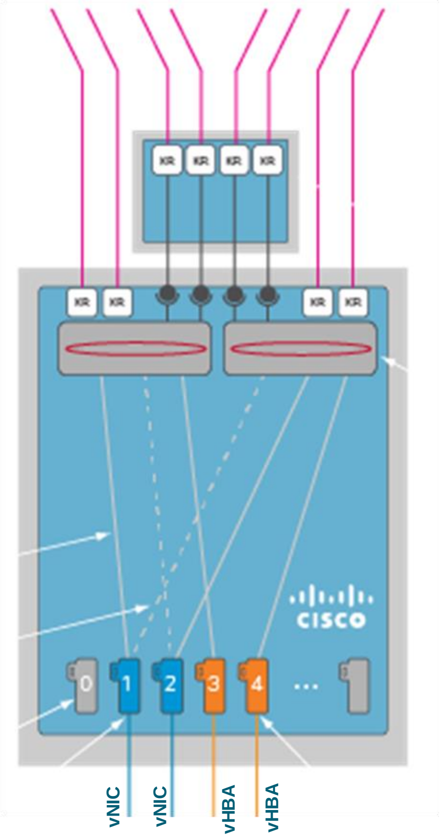
UCS Components

Blade VIC's

- 3rd Generation
  - VIC 1340 mLOM
  - Optional port expander card
  - VIC 1380 Mezz

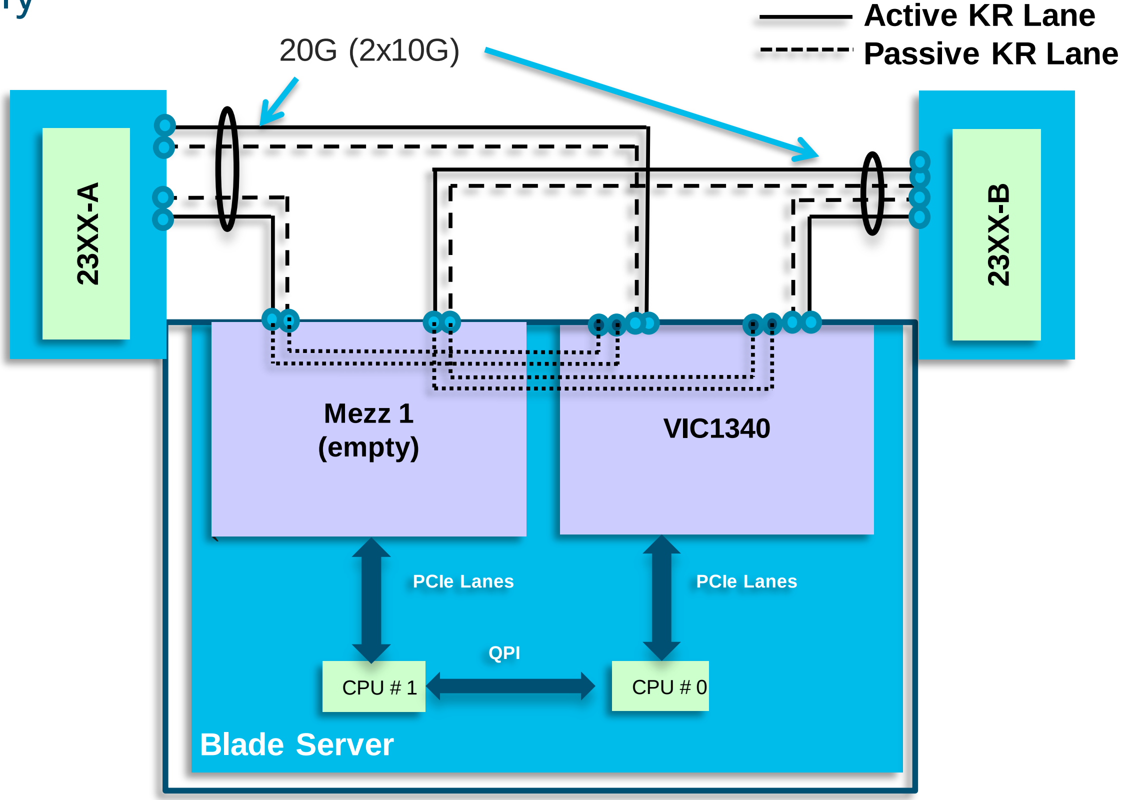


Unified Fabric

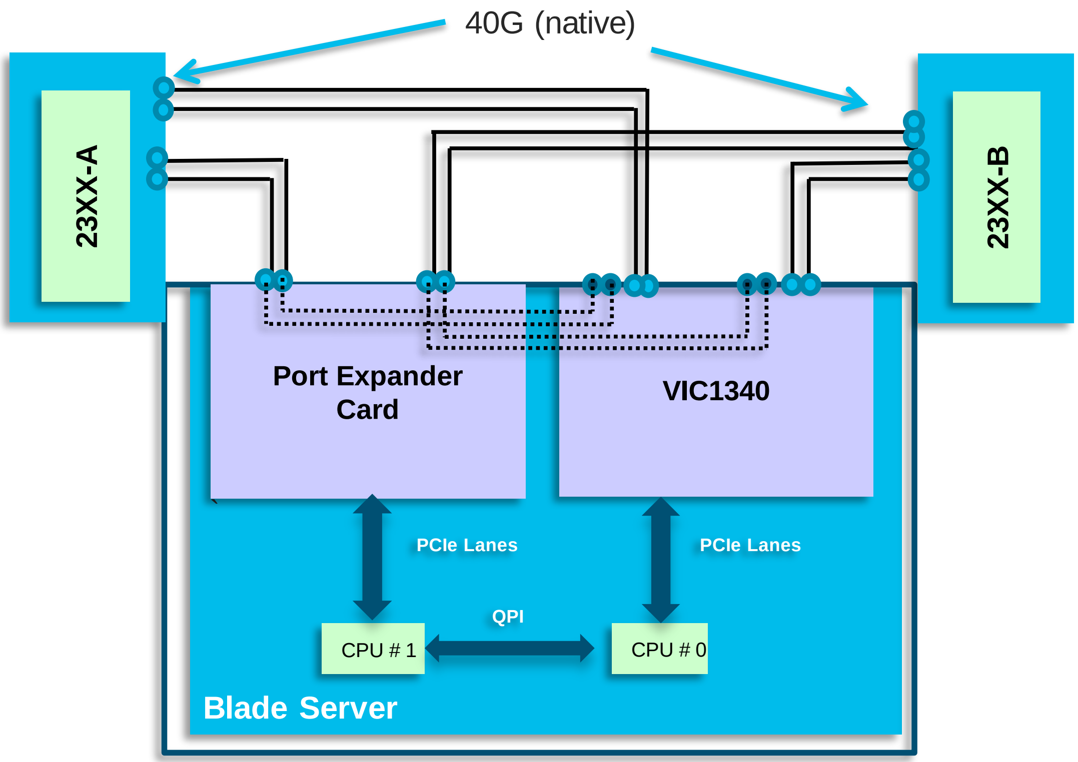


IOM 2304 and Adapter Connection

VIC1340 Only

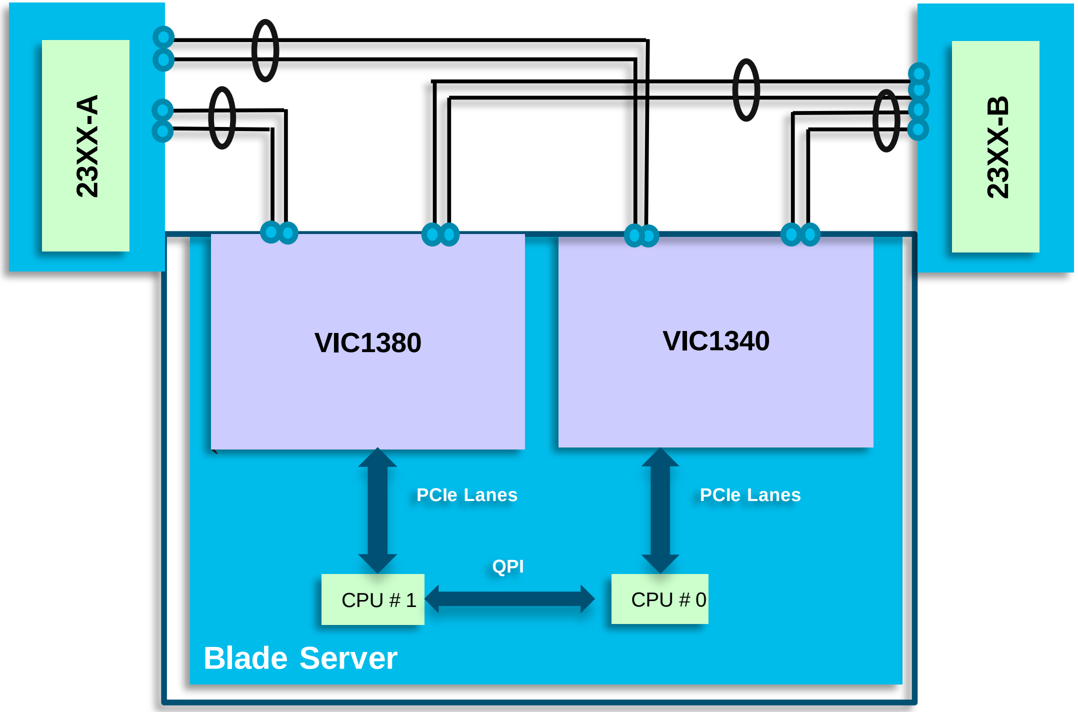


IOM 2304 and Adapter Connection
VIC1340 Plus Port Expander



IOM 2304 and Adapter Connection
VIC1340 Plus VIC1380

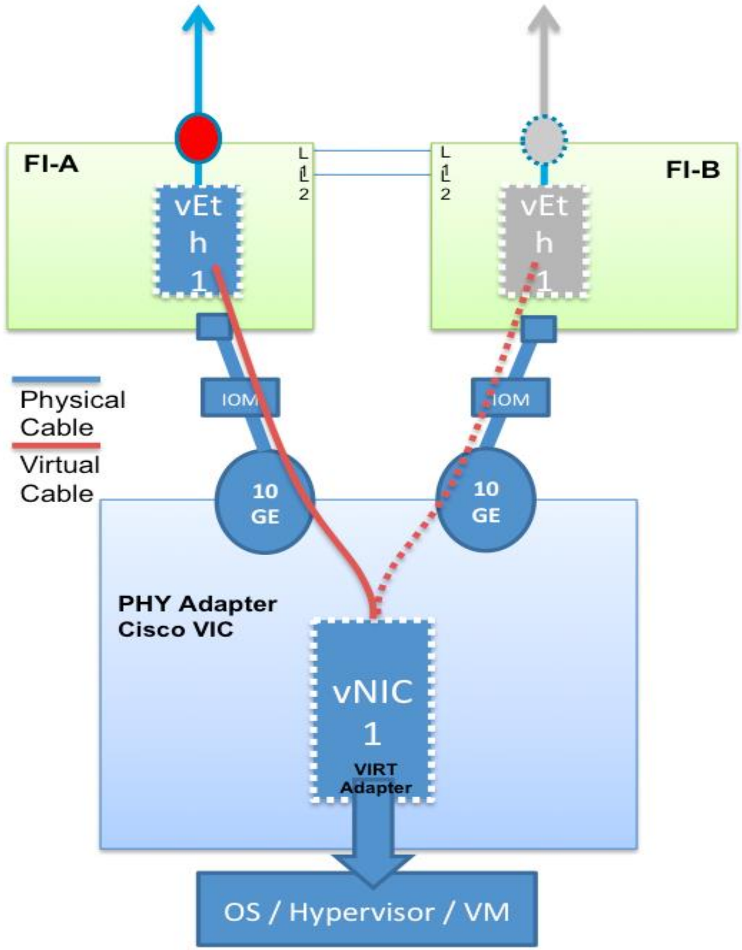
- Adapter Resiliency
  - 2 independent Adapters
  - vCon placement
- 4 20G connections
- 20G are 2x10
  - 7 tuple hashing



Unified Computing Systems

UCS Fabric Failover

- Fabric provides NIC failover capabilities chosen when defining a service profile
- Traditionally done using NIC bonding driver in the OS
- Provides failover for both unicast and multicast traffic
- Works for any OS.



Unified Computing System Architecture (Rack)

- UCS Manager
  Embedded– manages entire system

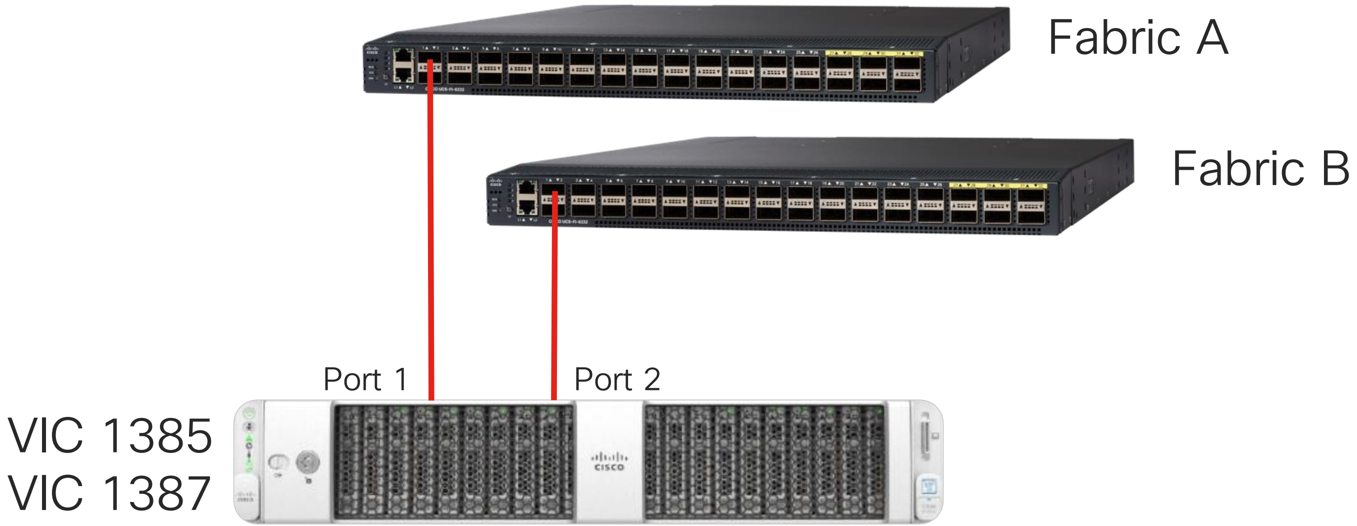
- UCS 6300 Series Fabric Interconnect
  32 x 40GB QSFP+ Port or 24 x 40GB QSFP+ Port & 16 UP Ports

- UCS C-Series Rack Servers (1U, 2U, 4U)
  Industry-standard architecture

- UCS Virtual Adapters
  Choice of multiple adapters

Unified Computing Systems

UCS Managed Rack Server Architecture

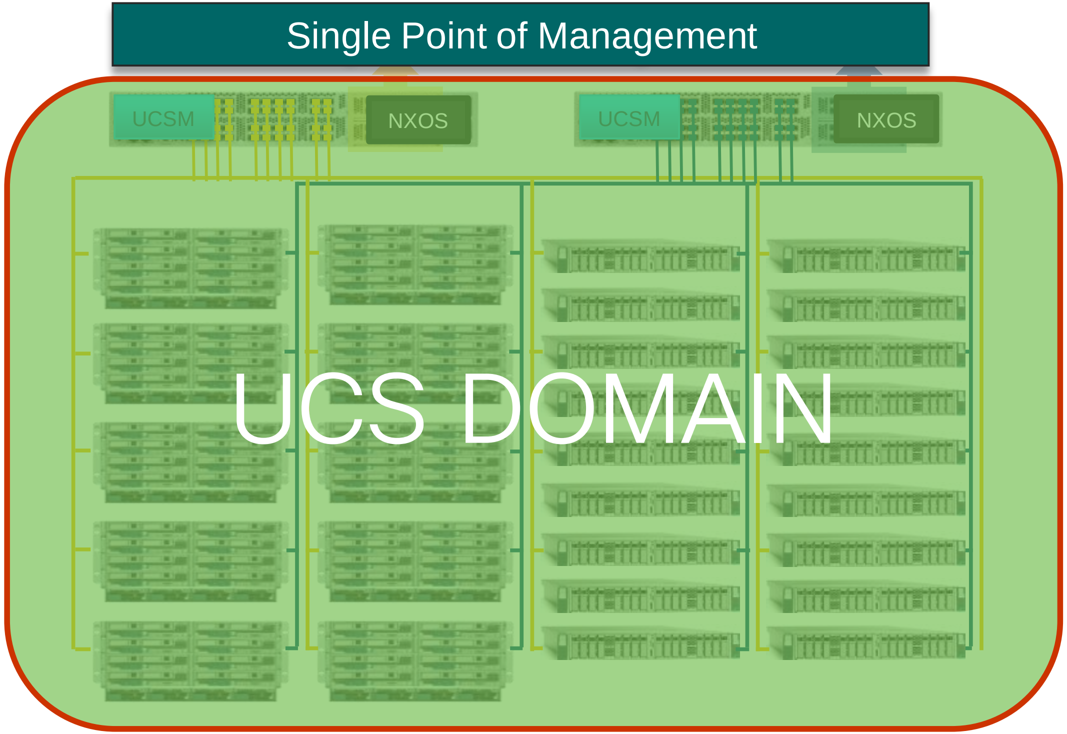


Momentary Lapse for Questions

UCS Advantages

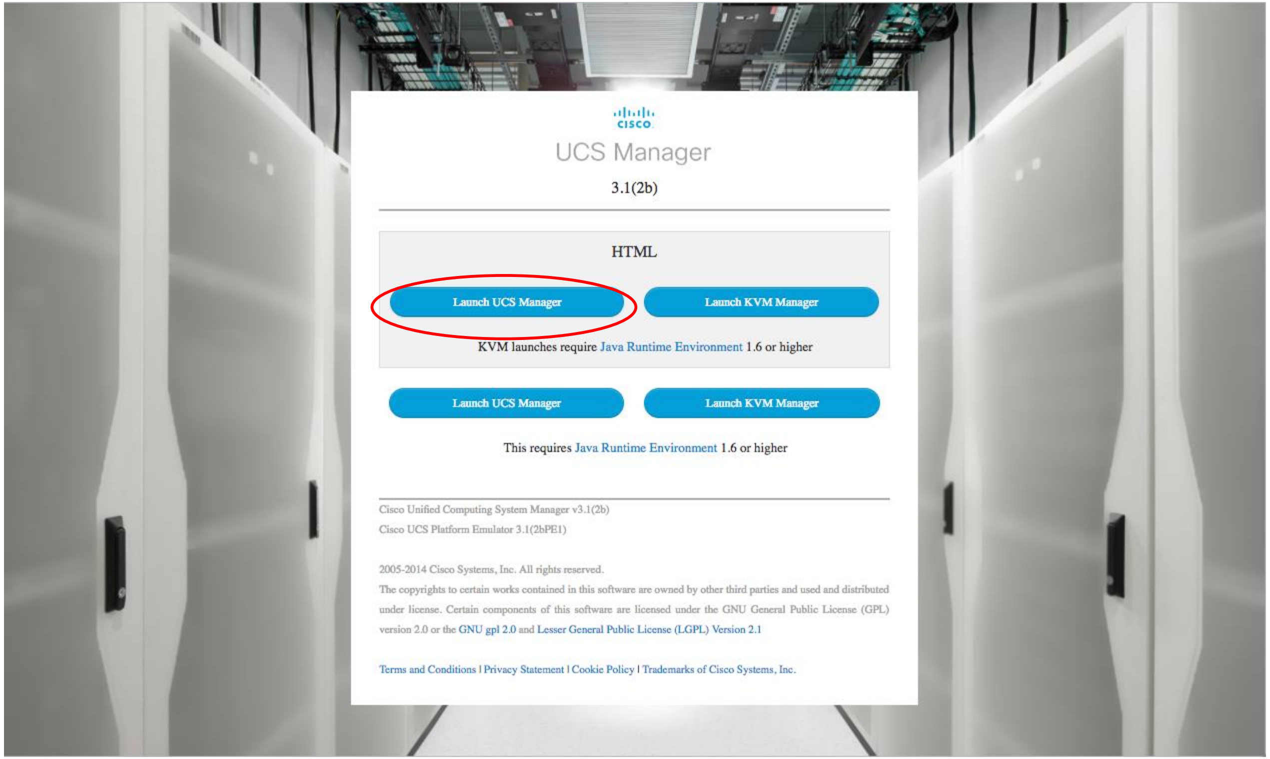
UCS Manager - Advantage #1

Single Point of Management and Scaling – UCS Manager



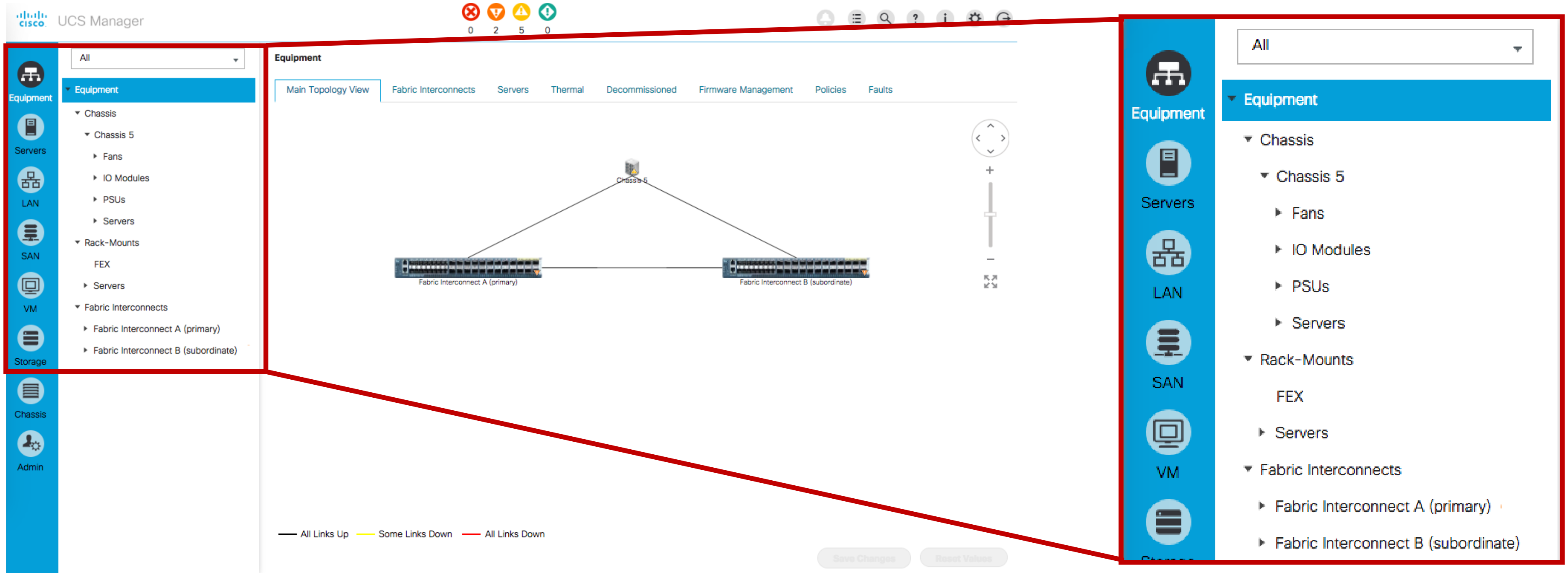
UCS Manager - Advantage #1

HTML5 Browser Based– Management and KVM



UCS Manager - Advantage #1

Single Point of Management – Equipment



UCS Manager - Advantage #1

Single Point of Management – Equipment – Server Inventory

- Equipment
  - Chassis
    - Chassis 5
      - Fans
      - IO Modules
      - PSUs
      - Servers
  - Rack-Mounts
    - FEX
    - Servers
  - Fabric Interconnects
    - Fabric Interconnect A (primary)

Equipment / Chassis / Chassis 5 / Servers

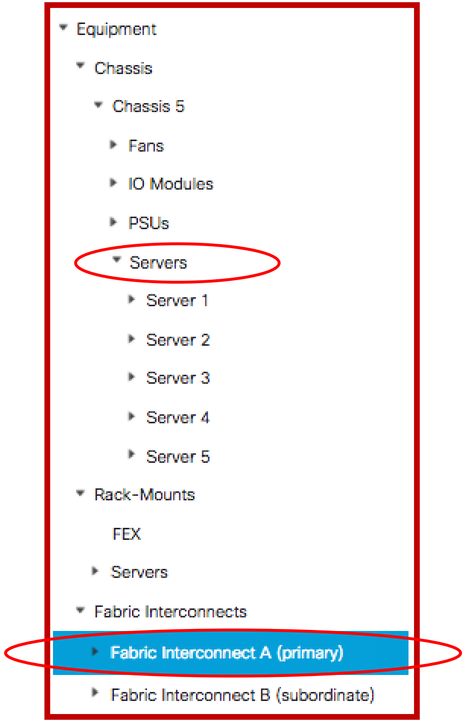
Servers

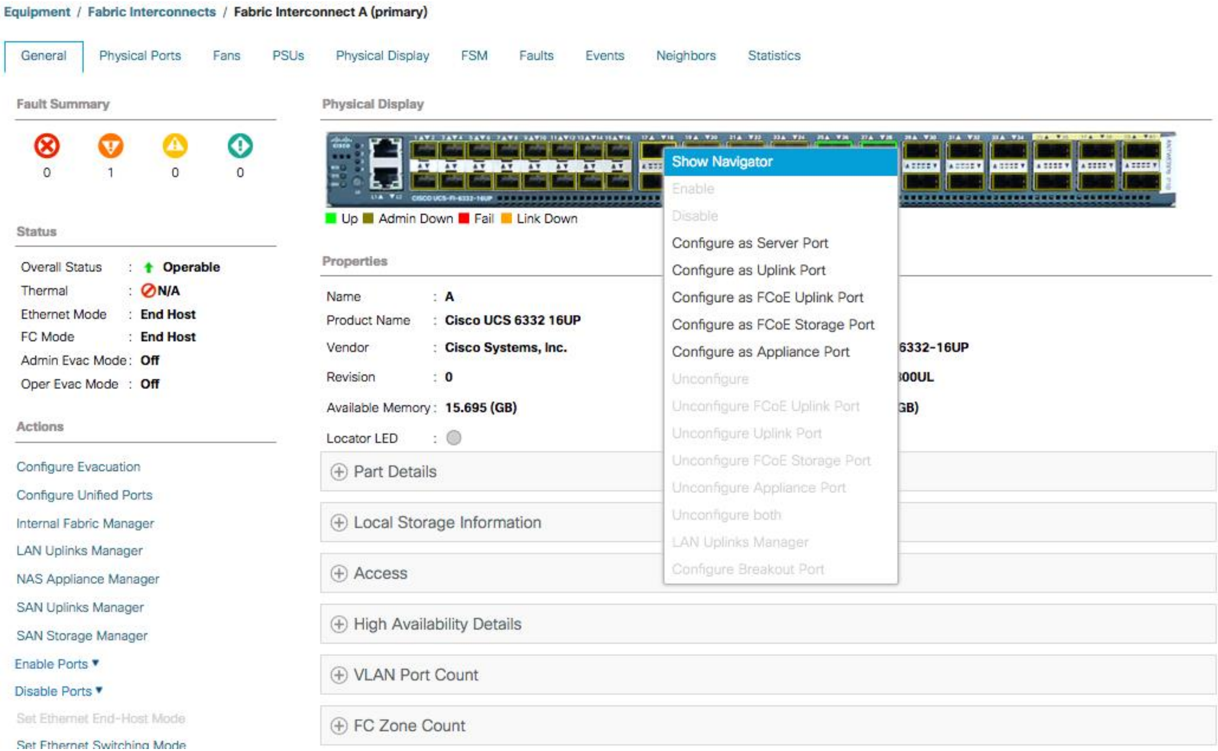
+ − Advanced Filter Export Print

Name	Over...	PID	Model	Serial	Profile	User ...	Cores	Core...	Threa...	Mem...	Adap...	NICs	HBAs	Oper...	Powe...
Se...	Unas...	UCS...	Cisco...	SRV78			8	8	16	49152	1	0	0	Oper...	Off
Se...	Unas...	UCS...	Cisco...	SRV79			8	8	16	49152	1	0	0	Oper...	Off
Se...	Unas...	UCS...	Cisco...	SRV80			8	8	16	49152	1	0	0	Oper...	Off
Se...	Unas...	UCS...	Cisco...	SRV81			8	8	16	49152	1	0	0	Oper...	Off
Se...	Unas...	UCS...	Cisco...	SRV82			16	16	32	49152	1	0	0	Oper...	Off

UCS Manager - Advantage #1

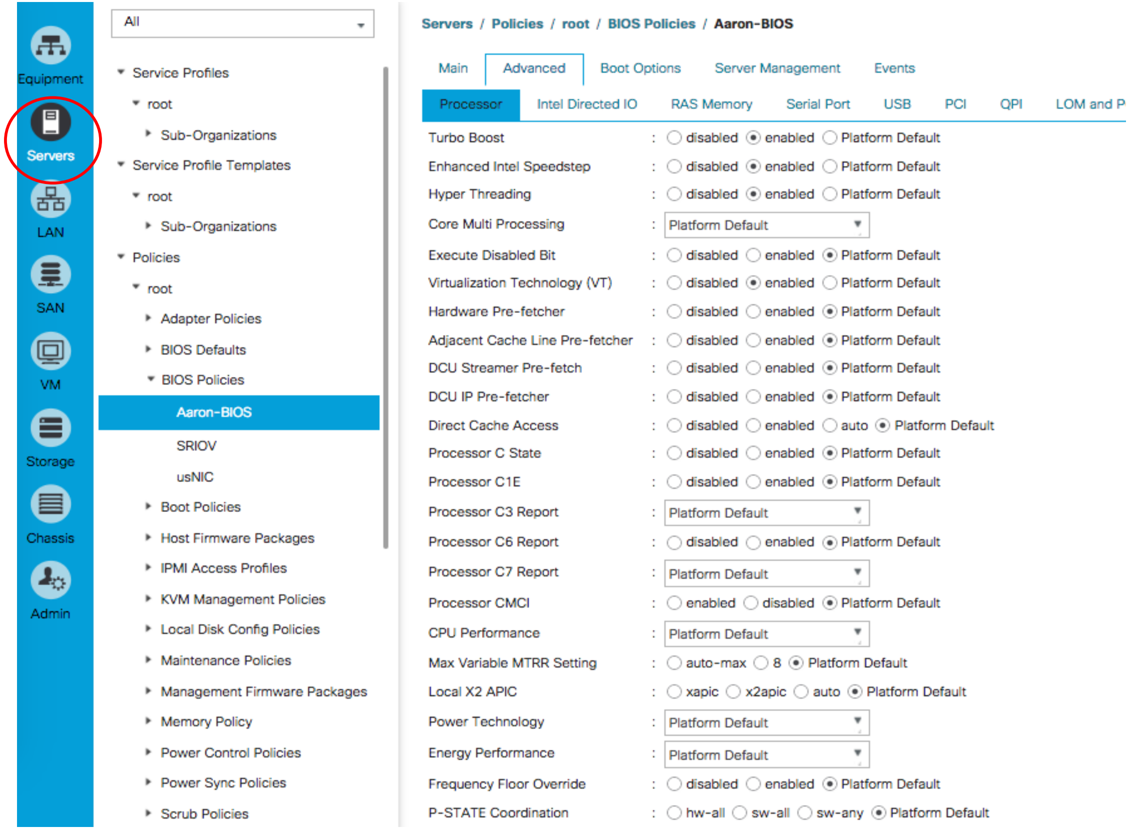
Single Point of Management – Equipment – Fabric Interconnects





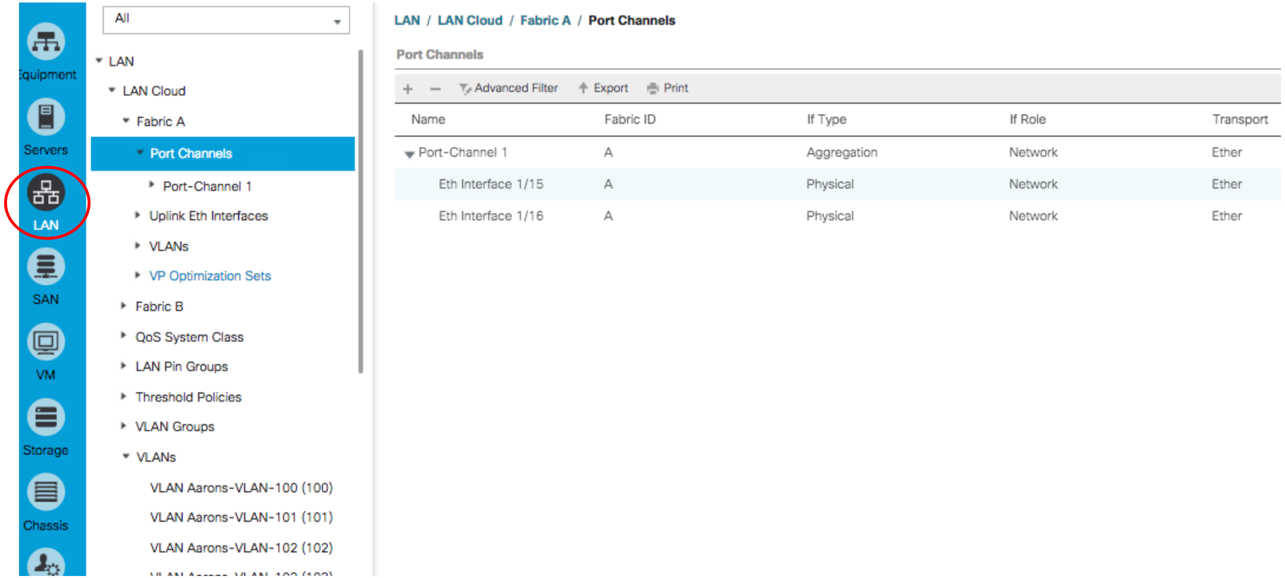
UCS Manager - Advantage #1

Single Point of Management – Servers



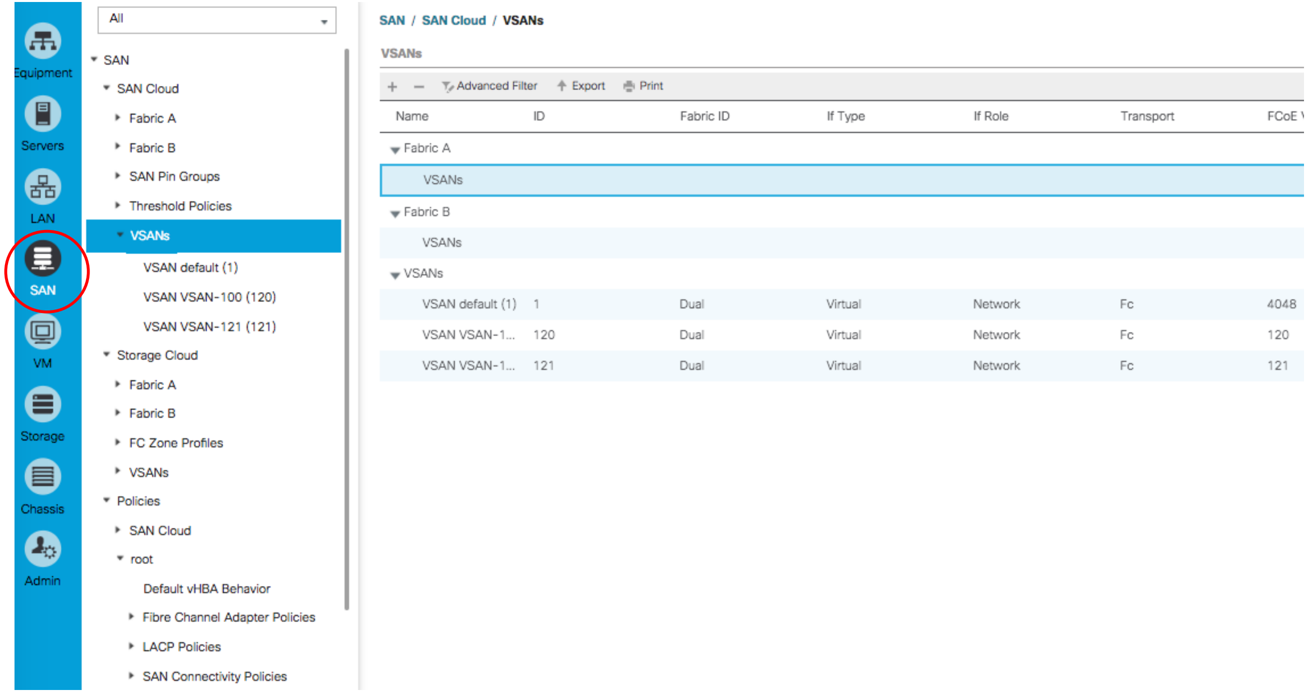
UCS Manager - Advantage #1

Single Point of Management – LAN Connectivity



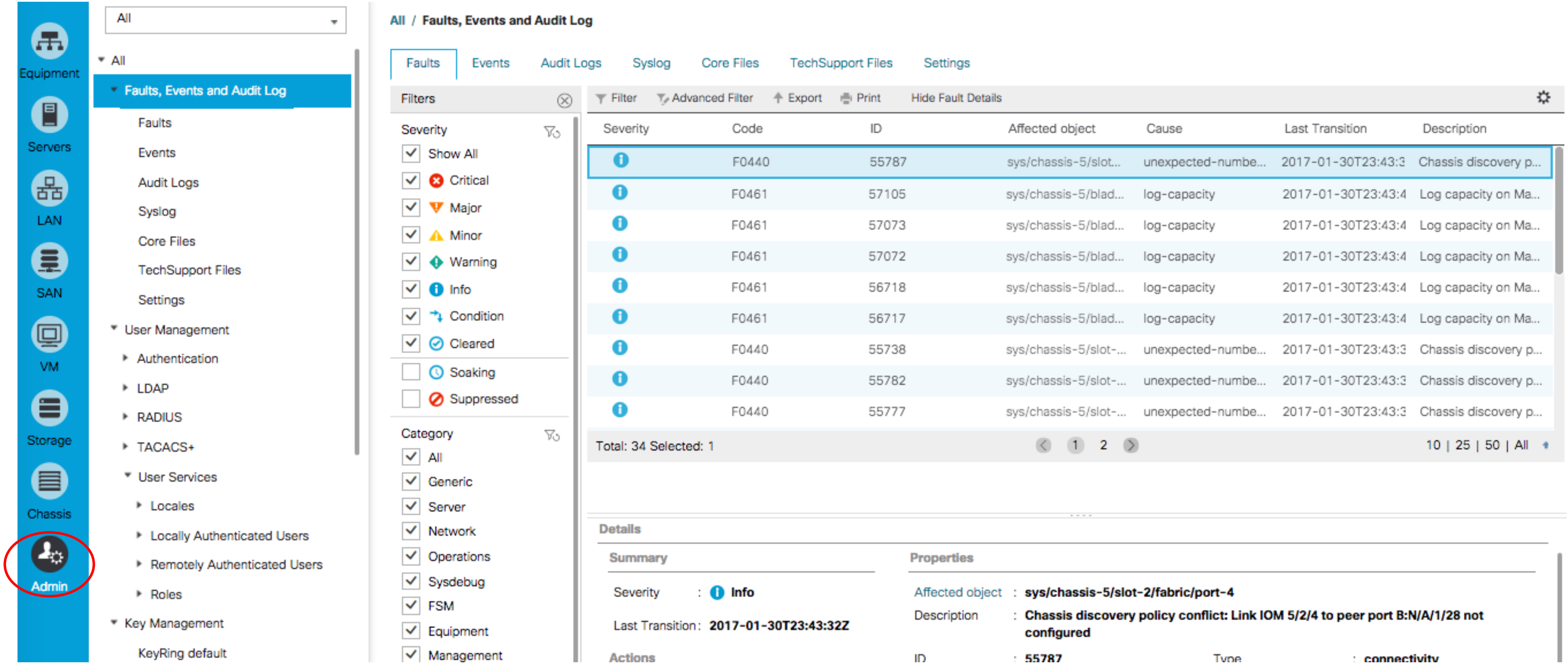
UCS Manager - Advantage #1

Single Point of Management – SAN Connectivity



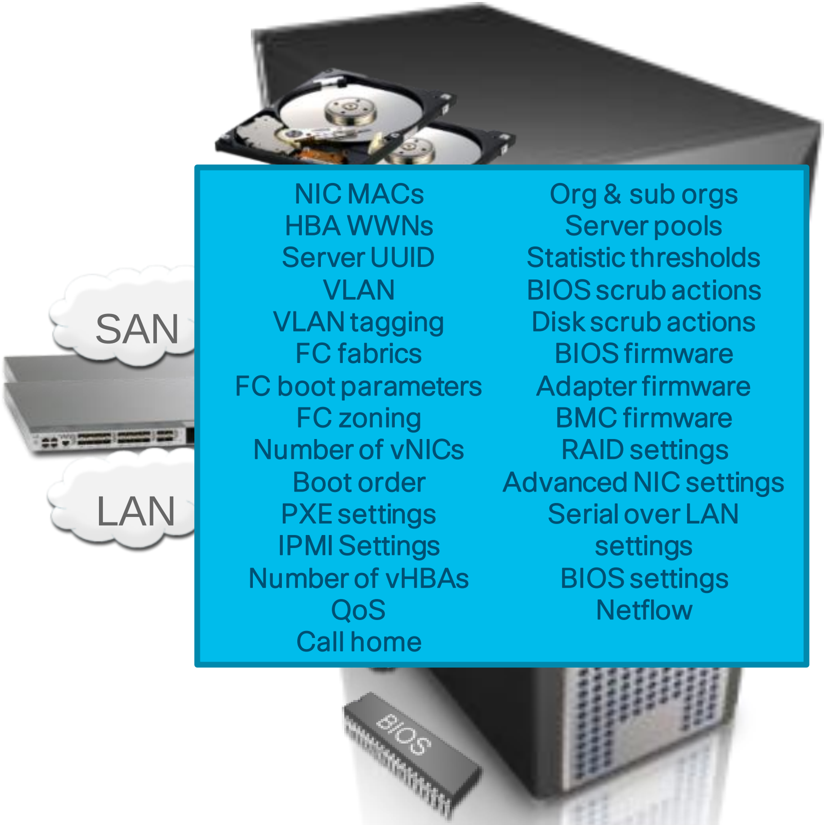
UCS Manager - Advantage #1

Single Point of Management – Central Administration



Stateless Computing – Advantage #2

Service Profiles

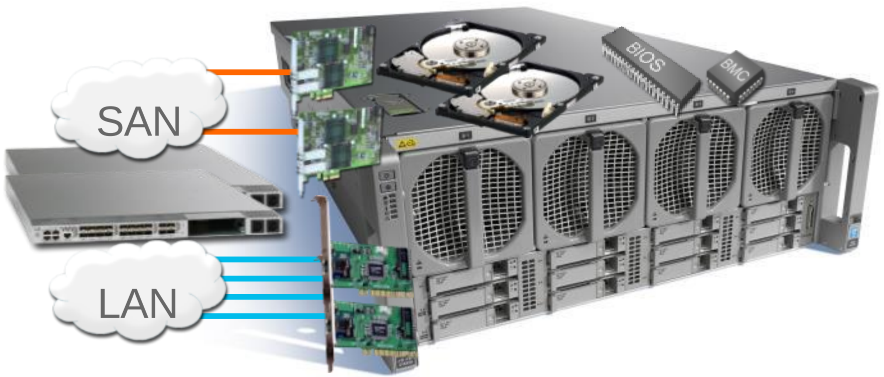


NIC MACs
HBA WWNs
Server UUID
VLAN
VLAN tagging
FC fabrics
FC boot parameters
FC zoning
Number of vNICs
Boot order
PXE settings
IPMI Settings
Number of vHBAs
QoS
Call home

Org & sub orgs
Server pools
Statistic thresholds
BIOS scrub actions
Disk scrub actions
BIOS firmware
Adapter firmware
BMC firmware
RAID settings
Advanced NIC settings
Serial over LAN settings
BIOS settings
Netflow

Stateless Computing – Advantage #2

Service Profiles



NIC MACs	Org & sub orgs
HBA WWNs	Server pools
Server UUID	Statistic thresholds
VLAN	BIOS scrub actions
VLAN tagging	Disk scrub actions
FC fabrics	BIOS firmware
FC boot parameters	Adapter firmware
FC zoning	BMC firmware
Number of vNICs	RAID settings
Boot order	Advanced NIC settings
PXE settings	Serial over LAN settings
IPMI Settings	BIOS settings
Number of vHBAs	Netflow
QoS	
Call home	

Stateless Computing – Advantage #2

Service Profiles

NIC MACs	Org & sub orgs
HBA WWNs	Server pools
Server UUID	Statistic thresholds
VLAN	BIOS scrub actions
VLAN tagging	Disk scrub actions
FC fabrics	BIOS firmware
FC boot parameters	Adapter firmware
FC zoning	BMC firmware
Number of vNICs	RAID settings
Boot order	Advanced NIC settings
PXE settings	Serial over LAN settings
IPMI Settings	BIOS settings
Number of vHBAs	Netflow
QoS	
Call home	

Stateless Computing – Advantage #2

Service Profile Templates - Logical Building Blocks

4.

Service Profile	Service Profile	Service Profile
MAC : 00:25:B5:01:00:05	MAC : 00:25:B5:01:00:06	MAC : 00:25:B5:01:00:07
WWN: 20:01:00:25:B5:01:00:05	WWN: 20:01:00:25:B5:01:00:06	WWN: 20:01:00:25:B5:01:00:07
Firmware 2.2	Firmware 2.2	Firmware 2.2
Boot:SAN	Boot:SAN	Boot:SAN

3.

Service Profile	Service Profile	Service Profile
MAC : 00:25:B5:01:00:05	MAC : 00:25:B5:01:00:06	MAC : 00:25:B5:01:00:07
WWN: 20:01:00:25:B5:01:00:05	WWN: 20:01:00:25:B5:01:00:06	WWN: 20:01:00:25:B5:01:00:07
Firmware 2.1	Firmware 2.1	Firmware 2.1
Boot:SAN	Boot:SAN	Boot:SAN

2.

Templates – Service Profile

MAC : Derived WWN: Derived

Firmware 2.2

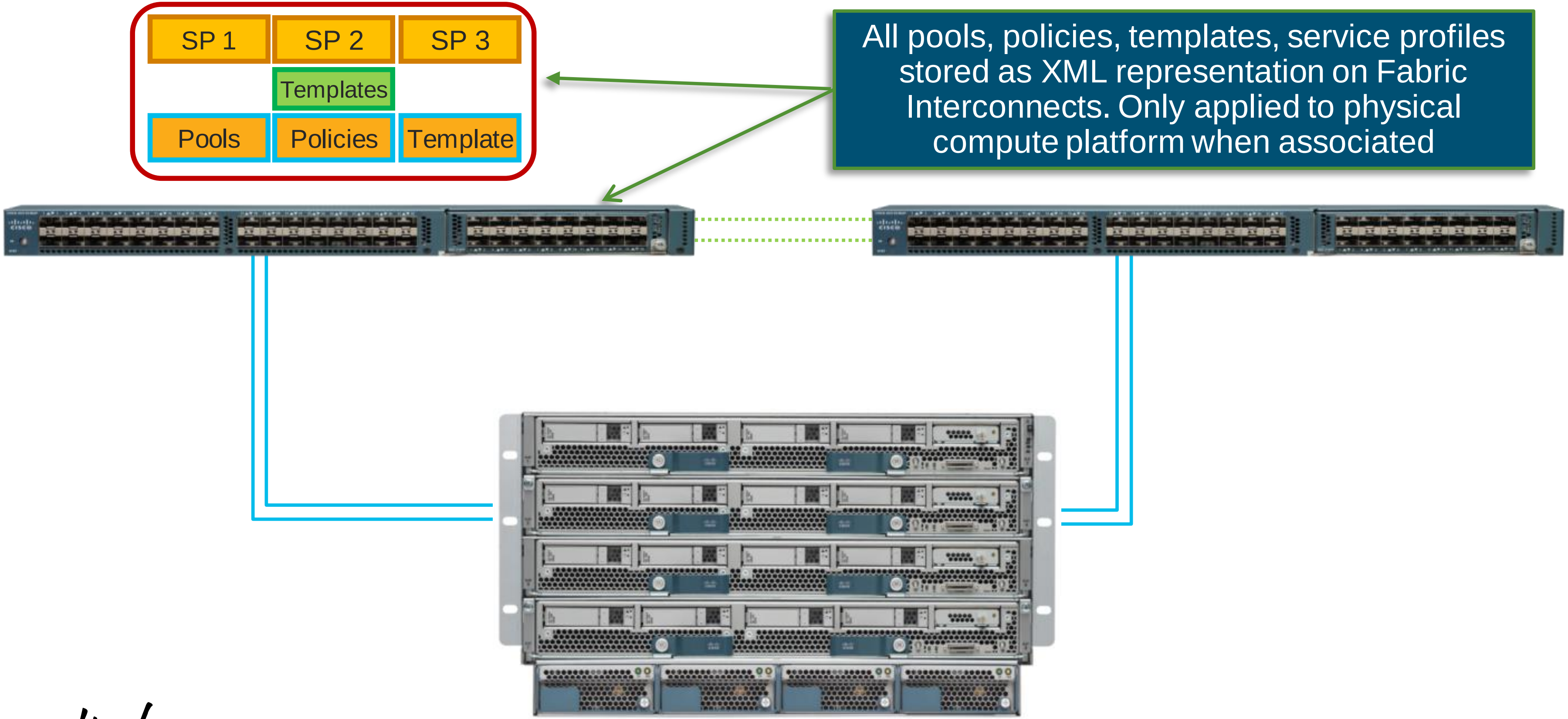
Boot:SAN

1.

Pools	Policies	Templates
UUID	Boot Devices/Order	vNICs
MAC	Host F/W	vHBAs
WWNN / WWPN	QoS	

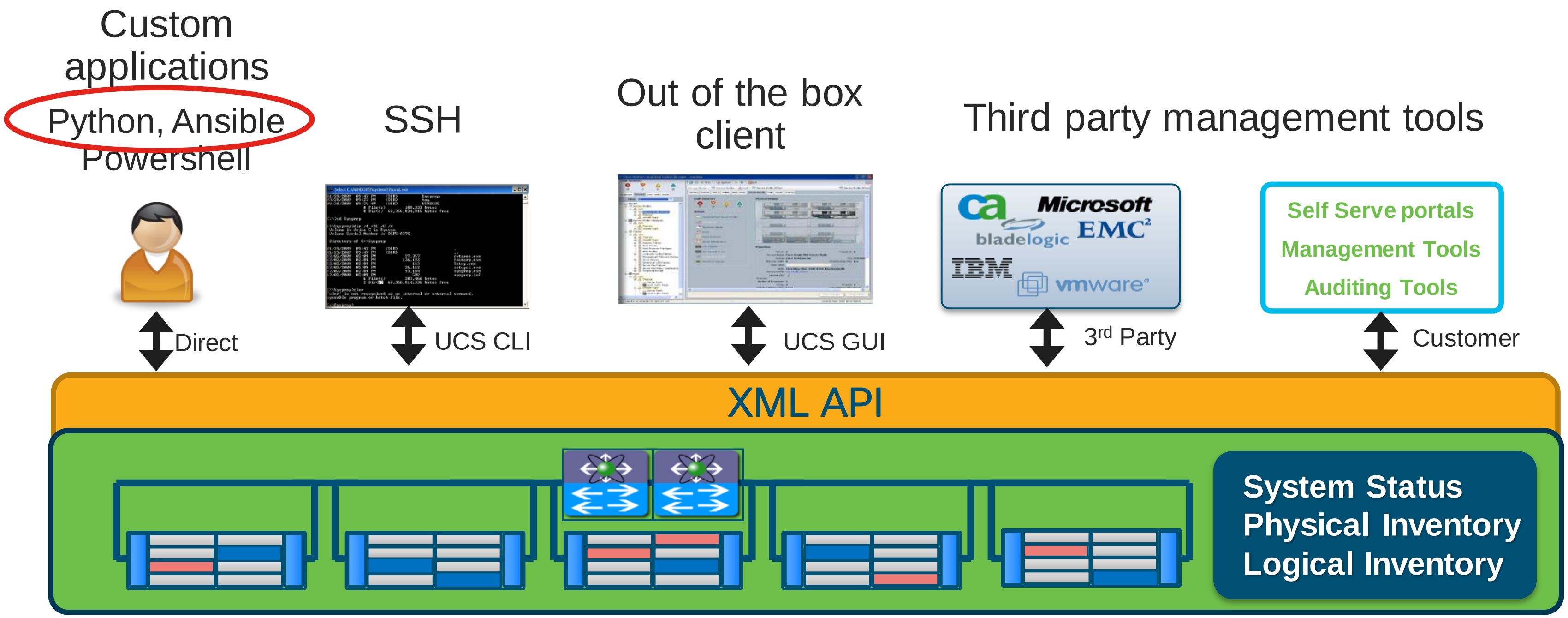
Stateless Computing – Advantage #2

Hardware/Software Abstraction



Programmability – Advantage #3

XML API and Other Hidden Gems - Programmatic interface

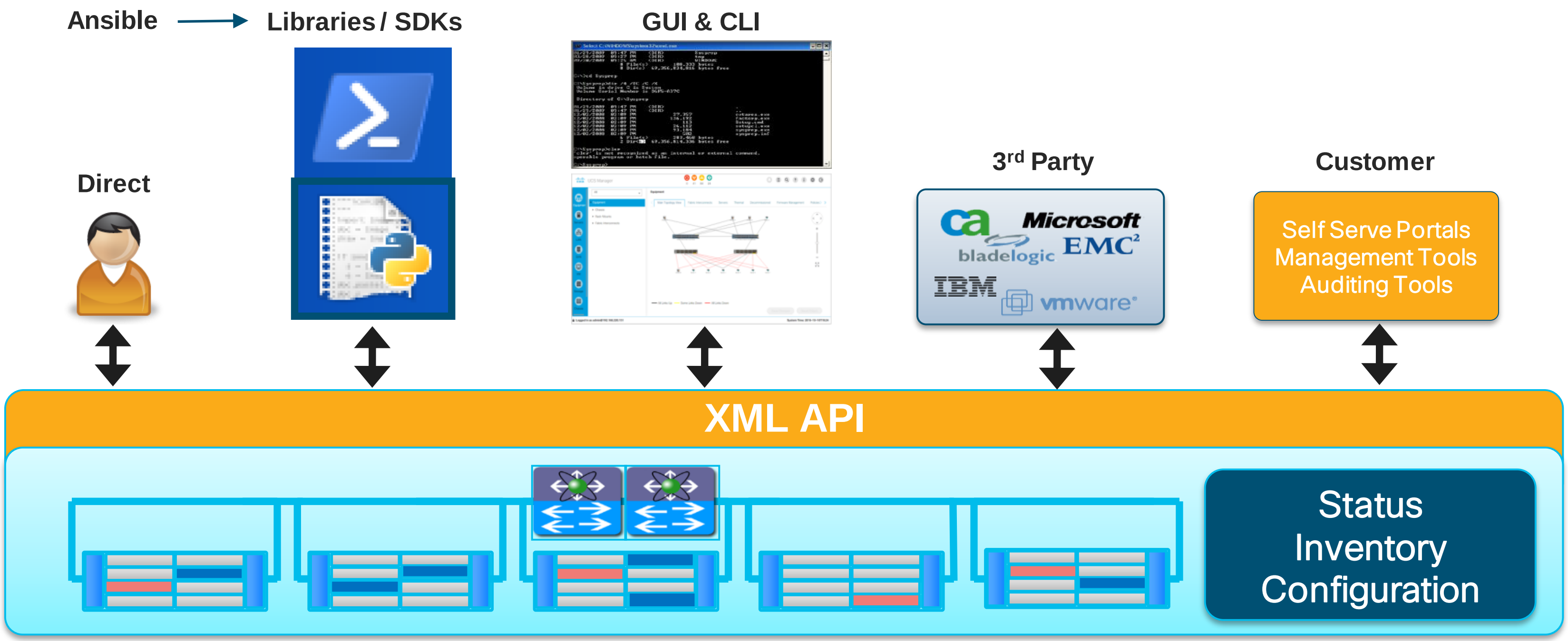


Programmability – Advantage #3

UCS Management Programmatic Infrastructure

DEVNET-1293

DEVNET-2562



DEVNET-1293

DEVNET-2562

Programmability – Advantage #3

Infrastructure as Code (IaC)

- Specific Configuration Management (CM) tools leverage the XML API, Python SDKs, etc. to support the target device
- You define the intent (end state) in plain text files that CM tools consume
- Intent is version controlled with standard SW dev tools, maybe even tested (e.g. Git, Jenkins)

```
#!/usr/bin/env python

from ucsmsdk.ucshandle impo

handle = UcsHandle("10.10.1
handle.login()

blades_and_chassis = handle
print blades_and_chassis
print blades_and_chassis['C

for blade in blades_and_cha
    print blade.dn

for chassis in blades_and_c
print chassis.dn

blade_and_chassis = handle.
print blade_and_chassis
print blade_and_chassis['sy
```

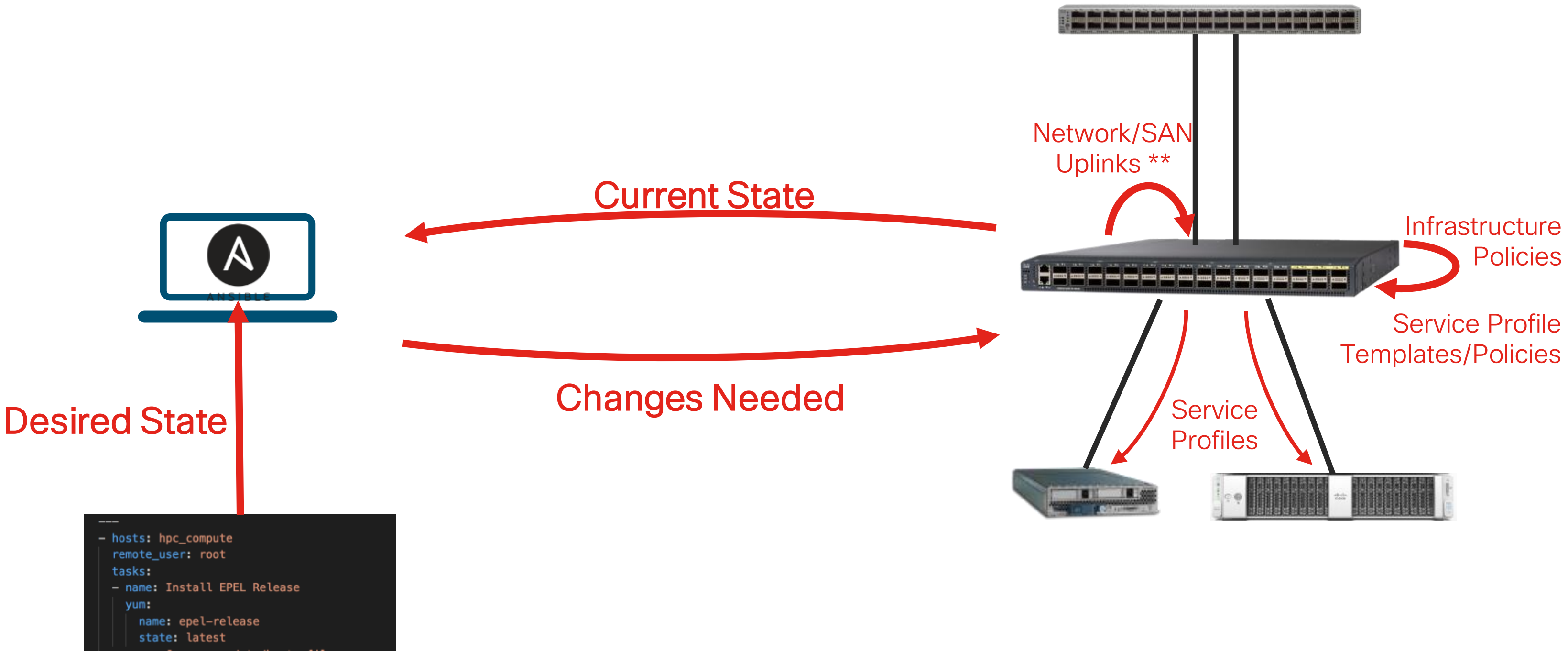
```
---
- hosts: hpc_compute
  remote_user: root
  tasks:
  - name: Install EPEL Release
    yum:
      name: epel-release
      state: latest
  - name: Copy core /etc/hosts file
    copy:
      src: /srv/etc/hosts
      dest: /etc/hosts
      owner: root
      group: root
      mode: 0644
    notify:
    - restart custom service
  handlers:
    - name: restart custom service
      service:
        name: custom
        state: restarted
```

DEVNET-1293

DEVNET-2562

Programmability – Advantage #3

Infrastructure as Code (IaC)

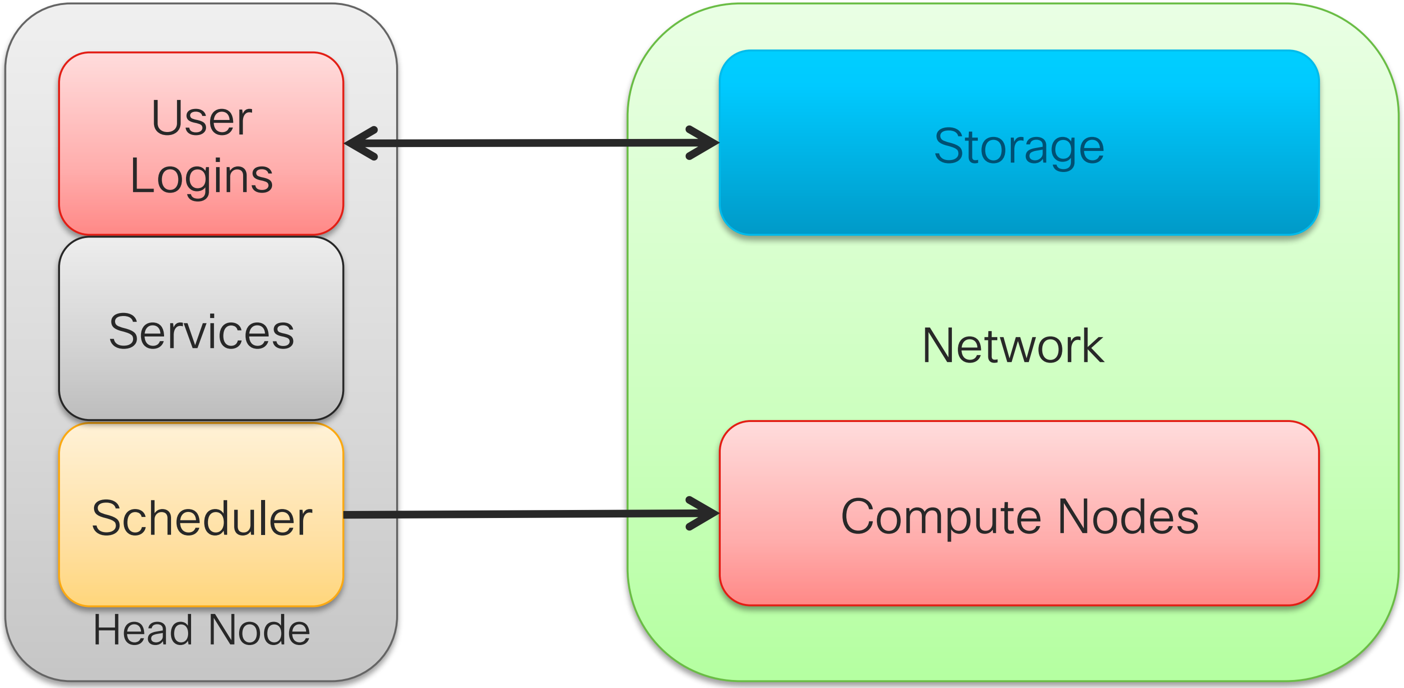


** Future capability

Questionable pause

UCS Designs

General Architecture Review



“Small” Research Computing Environments

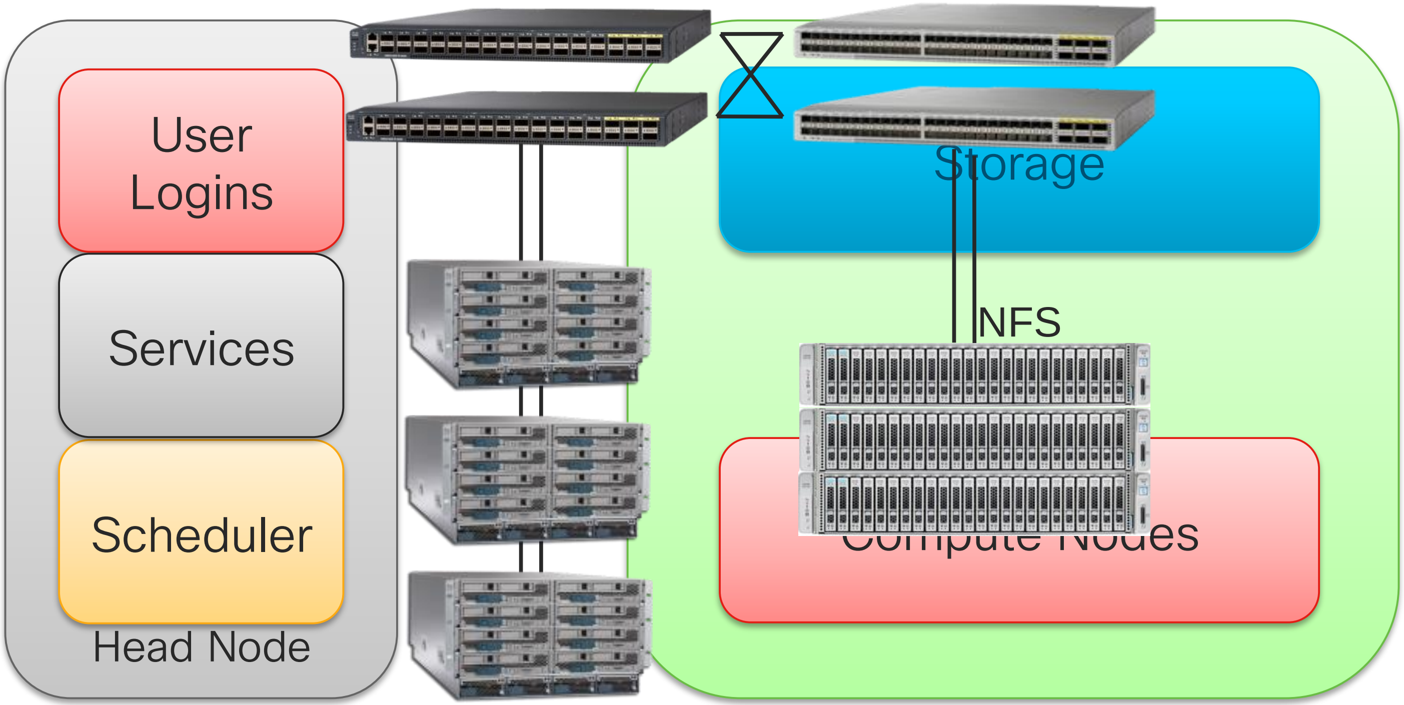
Workload Types

- Compute-bound computations
  - Statistical modeling simulations
  - Small inputs, Large Outputs
- Large scale independent tasks
  - Parameter space search
  - Maybe some multi-core software
- GPU enabled software

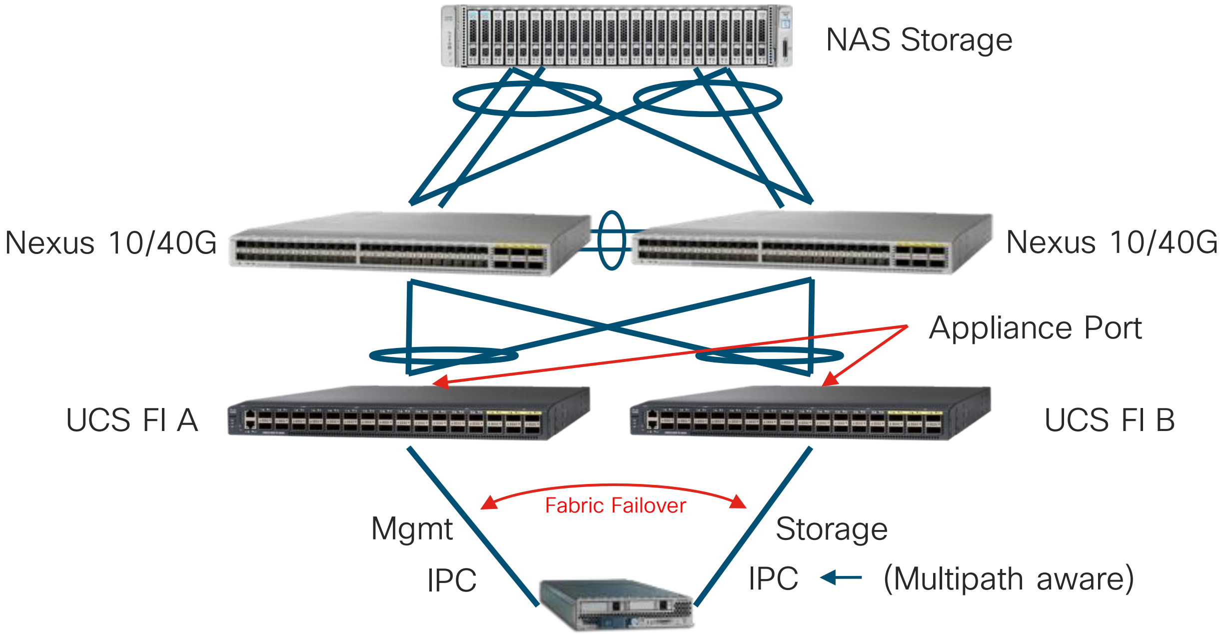
Resource Scale

- Compute
  - 1 UCS domain (<160 nodes)
  - 10GE server
- Networking
  - HPC Local to FI
- Storage
  - Low-modest IOPs
  - GB or smaller streaming r/w

Novel approach to small environments



Implementation Details



“Medium” Research Computing Environments

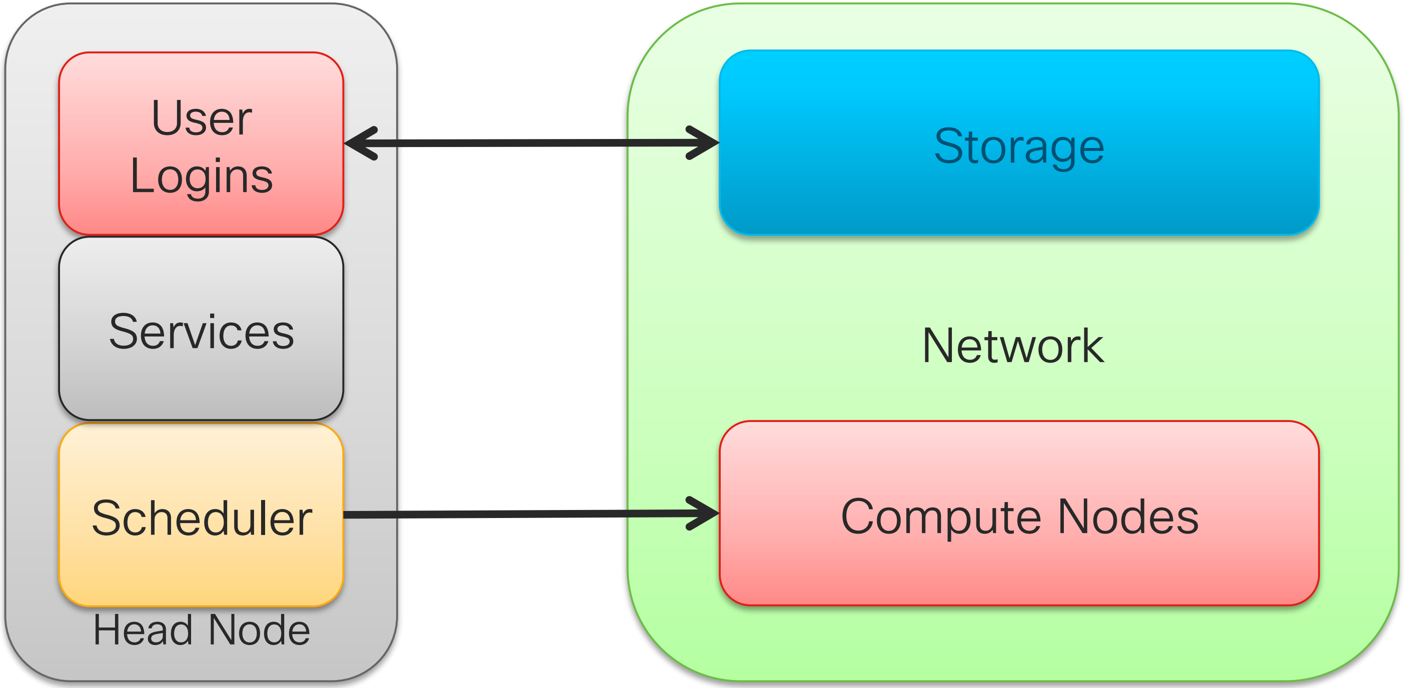
Workload Type

- Multi-node Applications
  - MPI low comm/calc ratio
- TB scale data sets
  - Streaming reads
  - Many nodes, concurrent access
- Balance of I/O and Compute
  - CPU or GPU heavy processing

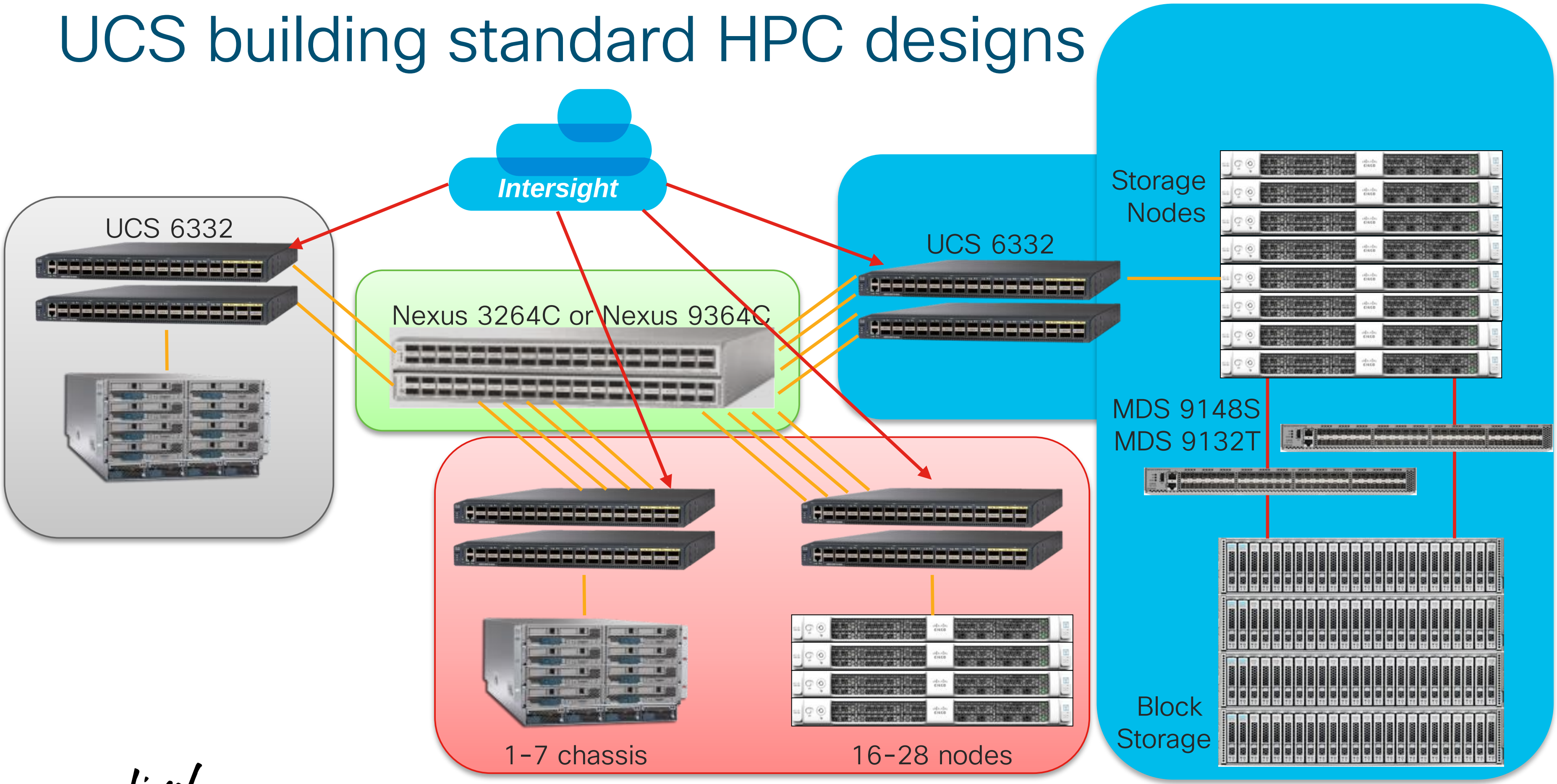
Resource Scale

- Compute
  - Multiple Domains
  - 10GE or 40GE
- Networking
  - Multi-tier UCS/Nexus
- Storage
  - Parallel Filesystems

“Medium” Research Computing Environments



UCS building standard HPC designs



Cisco Validated Designs for Big Data

Cloudera

SAP

Splunk

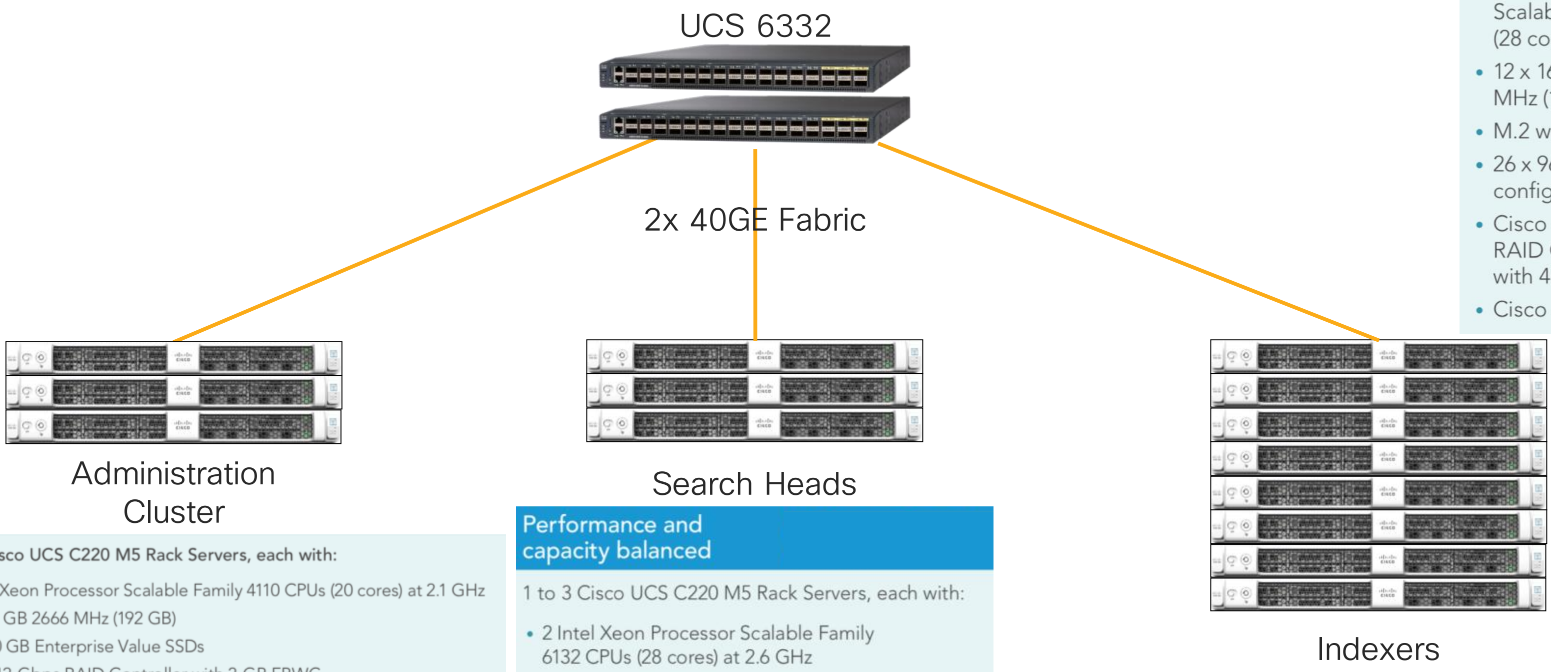
MapR

Hortonworks

IBM

SAS

Splunk Reference Architecture on UCS



8 Cisco UCS C240 M5 Rack Servers[1], each with:

- 2 Intel Xeon Processor Scalable Family 6132 CPUs (28 cores) at 2.6 GHz
- 12 x 16 GB 2666 MHz (192 GB)
- M.2 with 2 x 480-GB SSDs
- 26 x 960-GB SSDs configured as RAID5
- Cisco 12-Gbps RAID Controller with 4-GB FBWC
- Cisco UCS VIC 1387

1 to 3 Cisco UCS C220 M5 Rack Servers, each with:

- 2 Intel Xeon Processor Scalable Family 4110 CPUs (20 cores) at 2.1 GHz
- 12 x 16 GB 2666 MHz (192 GB)
- 2 x 480 GB Enterprise Value SSDs
- Cisco 12-Gbps RAID Controller with 2-GB FBWC
- Cisco UCS VIC 1387

Performance and capacity balanced

1 to 3 Cisco UCS C220 M5 Rack Servers, each with:

- 2 Intel Xeon Processor Scalable Family 6132 CPUs (28 cores) at 2.6 GHz
- 12 x 16 GB 2666 MHz (192 GB)
- 2 x 480-GB Enterprise Value SSDs
- Cisco 12-Gbps RAID Controller with 2-GB FBWC
- Cisco UCS VIC 1387

UCS References Architectures for Custom Needs

Bundle	Blade	High performance	Performance	Capacity	High capacity
Server SKU		UCS-SP-C220M5-A2	UCS-SP-C240M5-A2	UCS-SPC240M5L-S1	UCSS-SP-S3260-BV
Supported platform	Scale out databases such as DataStax Enterprise, Elasticsearch, MongoDB, Oracle NoSQL Database, Pivotal Greenplum DB, MemSQL and Couchbase		Scale out systems such as Cloudera, Hortonworks, MapR, Transwarp, Pivotal Greenplum DB, Pivotal HD, SAS Analytics, Splunk Enterprise, Vertica, Elasticsearch and MarkLogic		Cloudera, Hortonworks, MapR, Splunk Enterprise, MapR-XD, Cisco VSOM, Milestone, Genetec, SDS like Scality, IBM Cloud Object Storage and SwiftStack
Servers	8 x Cisco UCS B200 M5 Blade Servers	8 x Cisco UCS C220 M5 Rack Servers	16 x Cisco UCS C240 M5 Rack Servers with Small-Form-Factor (SFF) drives	16 x Cisco UCS C240 M5 Rack Servers with Large-Form-Factor (LFF) drives	8 x Cisco UCS S3260 Storage Server, each server node with
CPU	2 x Intel Xeon Processor Scalable Family 6132 (2 x 14 cores and 2.6 GHz)	2 Intel Xeon Processor Scalable Family 6132 (2 x 14 cores and 2.6 GHz)	2 Intel Xeon Processor Scalable Family 6132 (2 x 14 cores and 2.6 GHz)	2 Intel Xeon Processor Scalable Family 4110 (2 x 8 cores and 2.1 GHz)	2 Intel Xeon processor E5-
Memory	12 x 16 GB 2666 MHz (192 GB)	12 x 16 GB 2666 MHz (192 GB)	12 x 16 GB 2666 MHz (192 GB)	12 x 16 GB 2666 MHz (192 GB)	8 x 32 GB 2400MHz (256 GB)
Boot	M.2 with 2 x 480-GB SSDs	M.2 with 2 x 480-GB SSDs	M.2 with 2 x 480-GB SSDs	M.2 with 2 x 480-GB SSDs	2 x 480-GB Enterprise Value Boot SSDs
Storage	2 x Cisco 2.5-inch 7.7-TB HGST SN200 NVMe High-Performance Enterprise Value	8 drives of 1.6-TB Enterprise Value SATA SSD SFFs	26 drives of 1.8-TB 10,000-rpm SFF SAS HDDs or 12 x 1.6-TB Enterprise Value SATA SSDs	12 x 8-TB 7200-rpm LFF SAS drives and 2 x 1.6 -TB Enterprise Value SATA SSDs	24 x 6-TB 7200-rpm LFF SAS drives
VIC	40-Gbps Cisco UCS VIC 1340 mLOM	40-Gbps Cisco UCS VIC 1387	40-Gbps Cisco UCS VIC 1387	40-Gbps Cisco UCS VIC 1387	40-Gbps Cisco UCS VIC 1387
Storage controller	Cisco FlexStorage PCIe SSD passthrough module with Hard-Disk-Drive (HDD) cage	Cisco 12-Gbps SAS Modular RAID Controller with 2-GB Flash-Based Write Cache (FBWC) or Cisco 12-Gbps Modular SAS Host Bus Adapter (HBA)	Cisco 12-Gbps SAS Modular RAID Controller with 4-GB FBWC or Cisco 12-Gbps Modular SAS HBA	Cisco 12-Gbps SAS Modular RAID Controller with 2-GB FBWC or Cisco 12-Gbps Modular SAS HBA	Cisco 12-Gbps SAS Modular RAID Controller with 4-GB FBWC
Network connectivity	Cisco UCS 6332 Fabric Interconnect	Cisco UCS 6332 Fabric Interconnect	Cisco UCS 6332 Fabric Interconnect	Cisco UCS 6332 Fabric Interconnect	Cisco UCS 6332 Fabric Interconnect

"Large" Research Computing Environments

Workload Types

- Any and all of the above
  - At scales beyond normal DC design
- Tightly coupled analyses
  - Communication heavy
  - Large multi-node applications
  - Messages are small (4-16K)
- Large data sets feeding GPUs
- GPU-to-GPU Direct RDMA Copy

Resource Scale

- Compute
  - 100s to 1000s
- Network
  - Low latency, large bandwidth
  - Separate network infrastructure
- Storage
  - Dedicated parallel FS appliance(s)
  - Multiple clusters of storage servers

Nexus 100G Portfolio

32p 100G QSFP28
Nexus 3232C

64p 100G QSFP28
Nexus 3264C

450ns latency

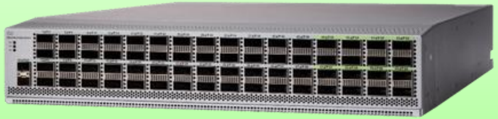
Merchant Silicon

Line Rate Telemetry

MACSEC Encryption

36p 100G QSFP28
Nexus 9336C-FX2

64p 40/100G QSFP28
Nexus 9364C



100G/40G Line Rate Performance (>250B)

Nexus 100G Portfolio

NX-OS Based Switches

32p 100G QSFP
Nexus 3232C

3.2Tbps switching with single chip

16MB Shared Buffer, 450ns switching latency

64p 100G QSFP28
Nexus 3264C

6.4Tbps switching with single chip

42MB Shared Buffer, 450ns switching latency

36p 100G QSFP
Nexus 9336C-FX2

3.6 Tbps switching with single chip

40MB Shared Buffer

100G Line Rate MACSEC, SSX Telemetry

64p 40/100G QSFP
Nexus 9364C

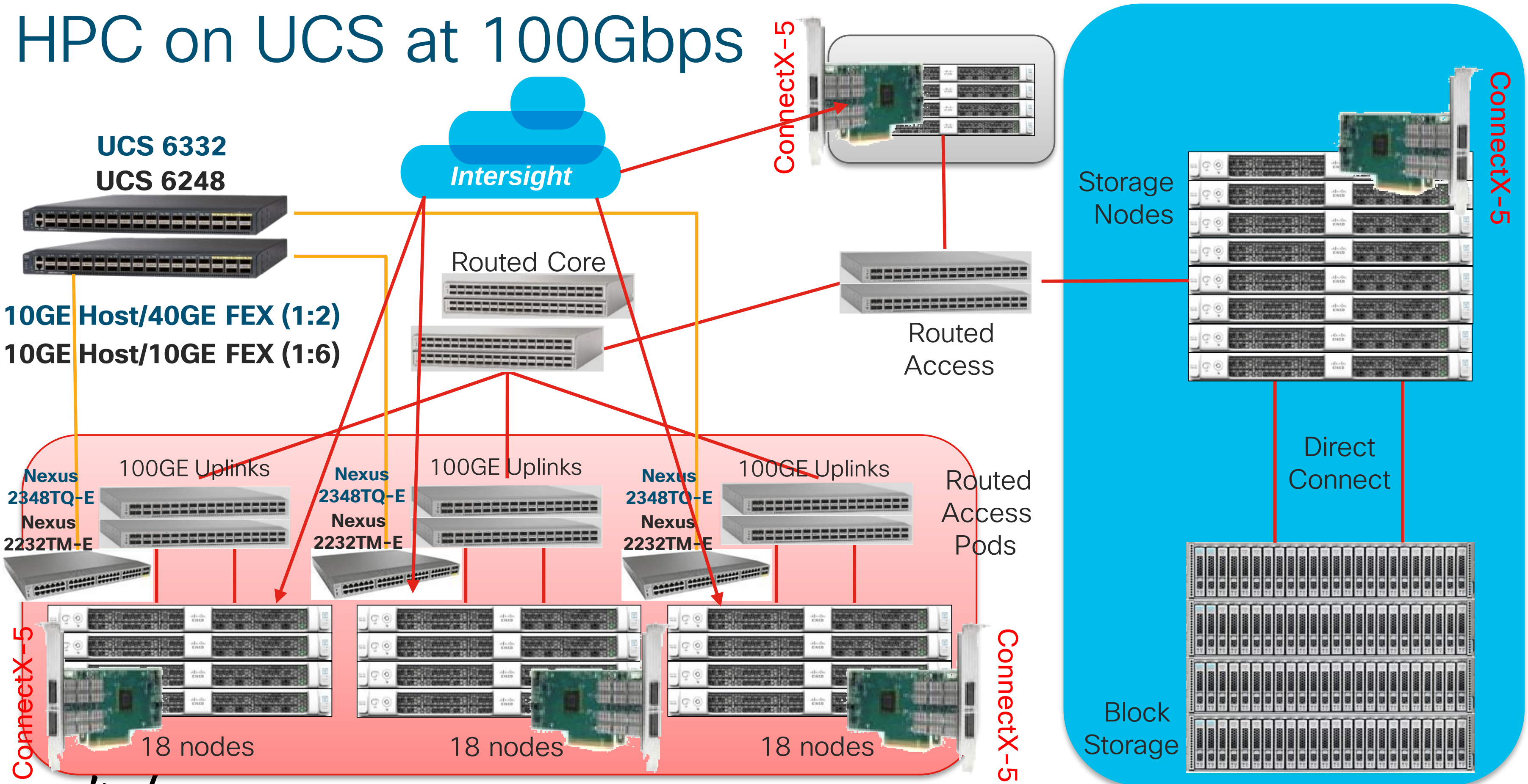
6.4Tbps switching with single chip

40MB Shared Buffer

100G Line Rate MACSEC, SSX Telemetry

100G/40G Line Rate Performance (>250B)

HPC on UCS at 100Gbps



UCS Drills into Oil and Gas

10x pair of Racks = 8 SU

A pair of Rack Computing:

72 ports 100 or 40G
- 12 ports 100G uplinks
- 60 ports 40G downlinks
- Oversubscription 2:1

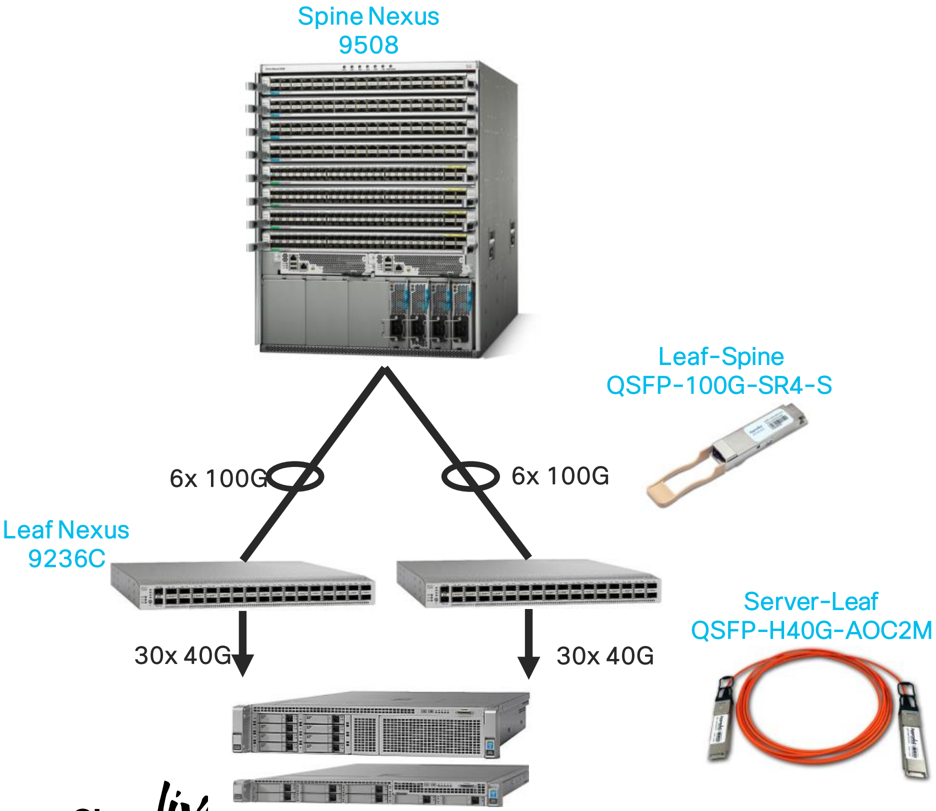
16 UCS C240-M5 servers
Total of **23.1 Tera FLOPs** (CPUs)

8 Magma Chassis
- 64x NVIDIA P100 or
- **595.2 Tera FLOPs** (GPUs)

14 UCS C220-M5 servers

Total 629.9 Tera FLOPs

Oil and Gas Network Topology



Spine ports

32-ports 100G module

Leaf ports

6 100G uplink ports (600G)

30 40G downlink ports (1200G)

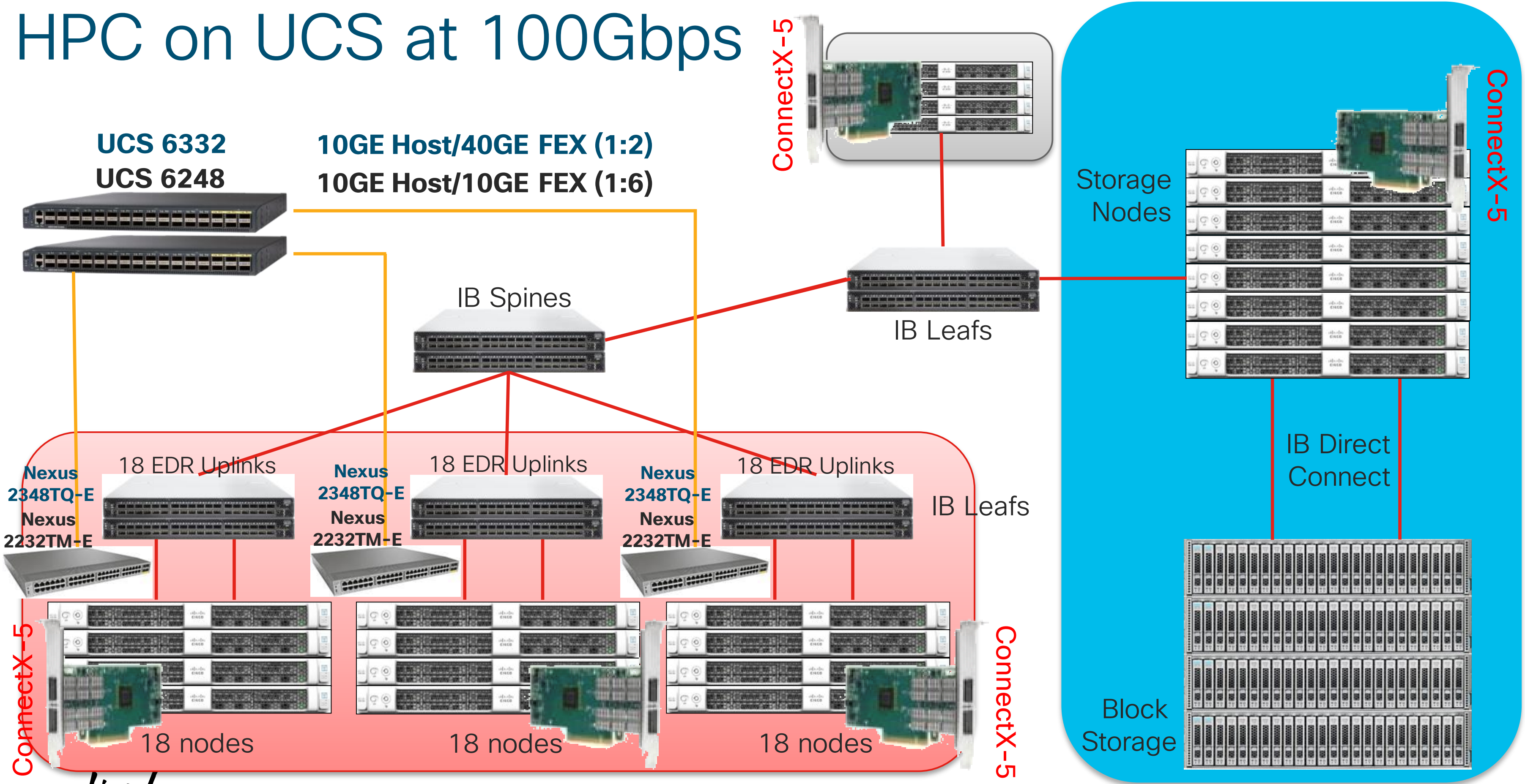
Oversubscription 2 : 1

Fabric Scalability

3 Nexus 9508 Spine supports up to 256 Leafs

3 Nexus 9516 Spine supports up to 512 Leafs

HPC on UCS at 100Gbps



It’s been a long time since I asked for questions

UCS Solution Summary

- UCS has a reliable, low risk portfolio
- UCS offers a proven ease-of-use operational benefits
- UCS enables automation of hardware provisioning that can integrate with established OS configuration management tools
- UCS provides competitive options for small and mid-size research computing environments
- UCS extends competitive computing solutions through support of 3rd party components for advanced communication needs

UCS Sweet Spots

- Better TCO for Enterprises with sustained usage (>30%)
- Leverage HPC environment as part of DR plan
- No dedicated HPC specialists to run infrastructure
- Need platform with automation to integrate into your existing toolsets (Ansible, Puppet, etc)

Agenda

- Introduction
- Anatomy of Research Computing
- Demo and Hands-on Cluster Operations
- UCS as a Platform for Research Computing
- **Hands-on Bonus: UCS Configuration with Ansible**
- NVIDIA Deep Learning with GPUs
- Conclusion

Demo

Agenda

- Introduction
- Anatomy of Research Computing
- Demo and Hands-on Cluster Operations
- UCS as a Platform for Research Computing
- Hands-on Bonus: UCS Configuration with Ansible
- **NVIDIA Deep Learning with GPUs**
- Conclusion

Deep Learning with GPUs

CiscoLive, June, 2018

AGENDA

What is Deep Learning?

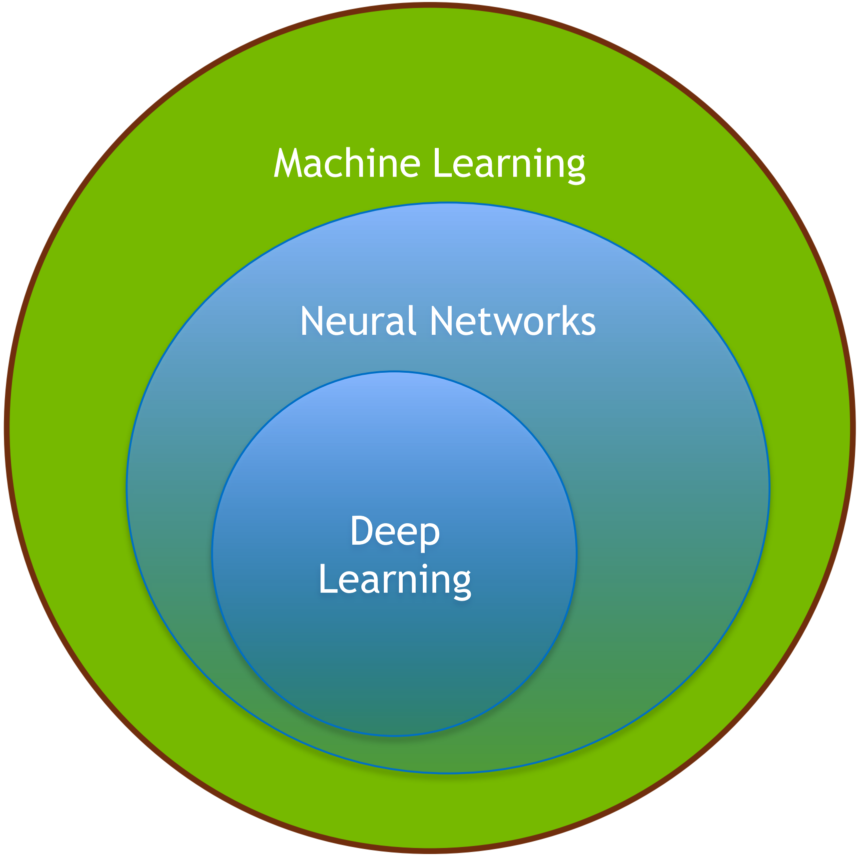
Example Use Cases (Healthcare emphasis)

GPUs and DL

DL in practice

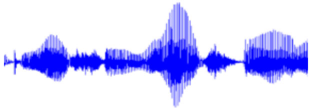
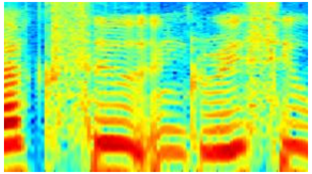
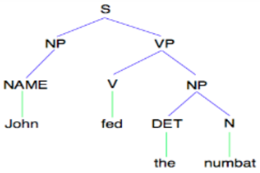
DIGITS Demo (time permitting)

What is Deep Learning?


Machine Learning
Neural Networks
Deep
Learning

Traditional machine perception

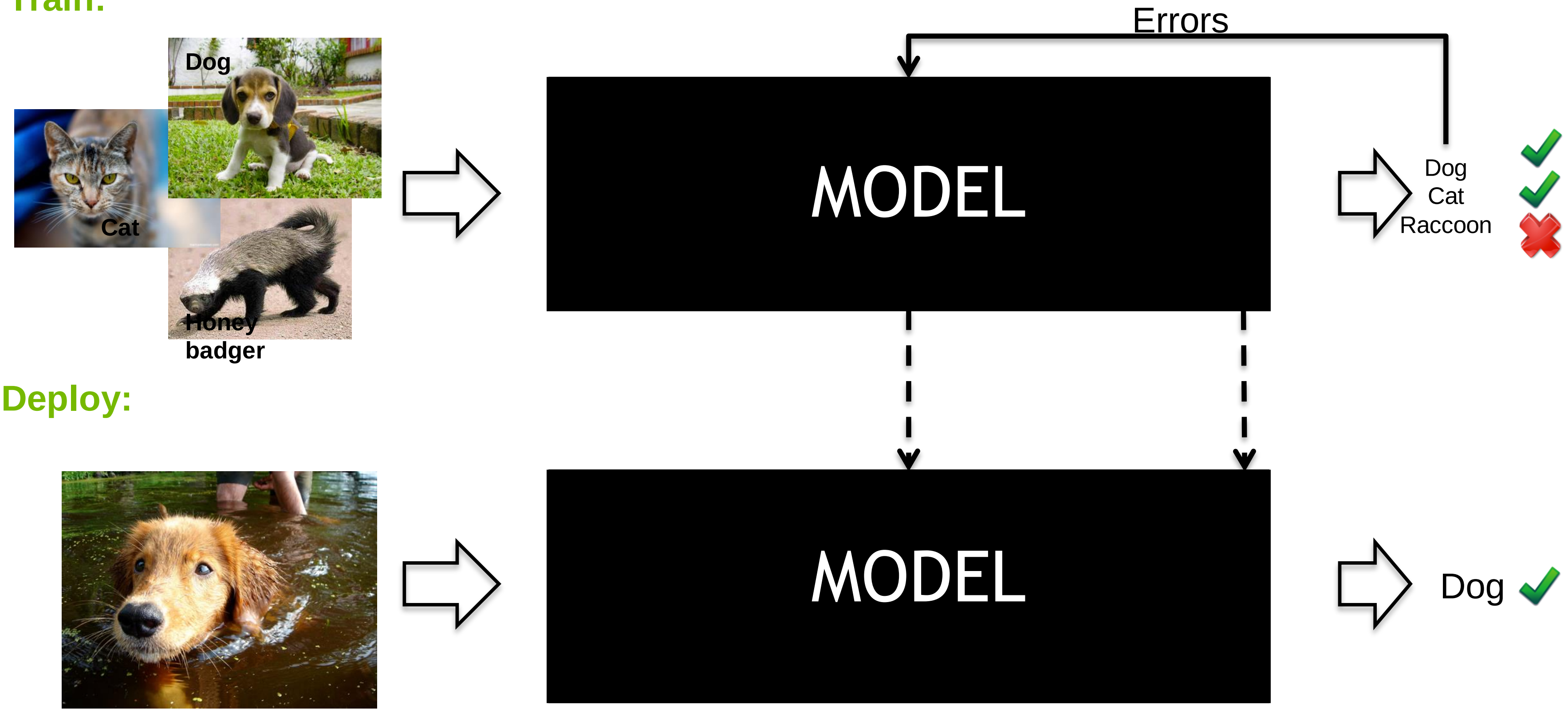
Hand crafted feature extractors

Raw data	Feature extraction	Classifier/ detector	Result
		SVM, shallow neural net, …	
		HMM, shallow neural net, …	Speaker ID, speech transcription, …
		Clustering, HMM, LDA, LSA …	Topic classification, machine translation, sentiment analysis…

Machine learning approach

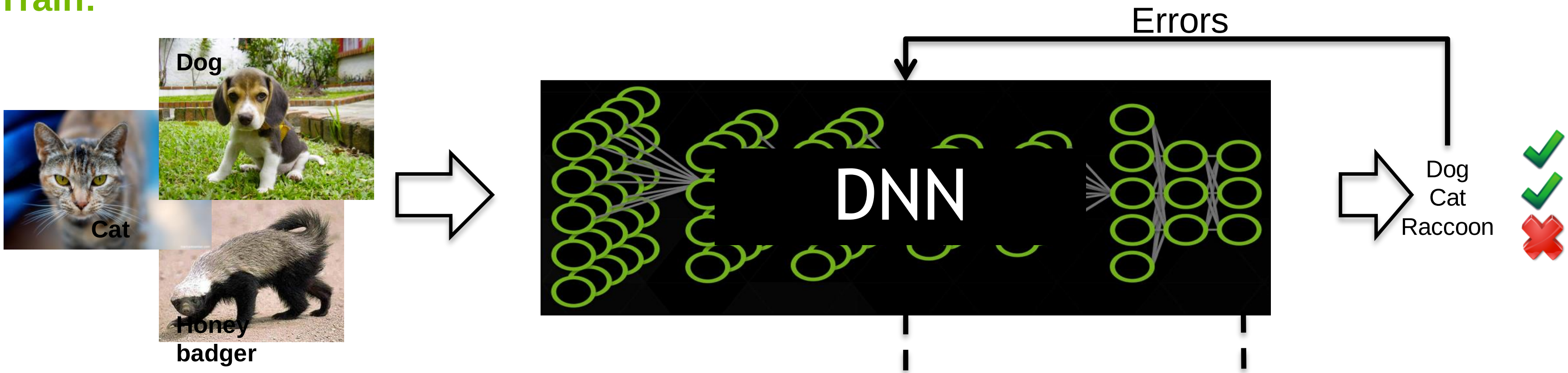
Train:

Deploy:

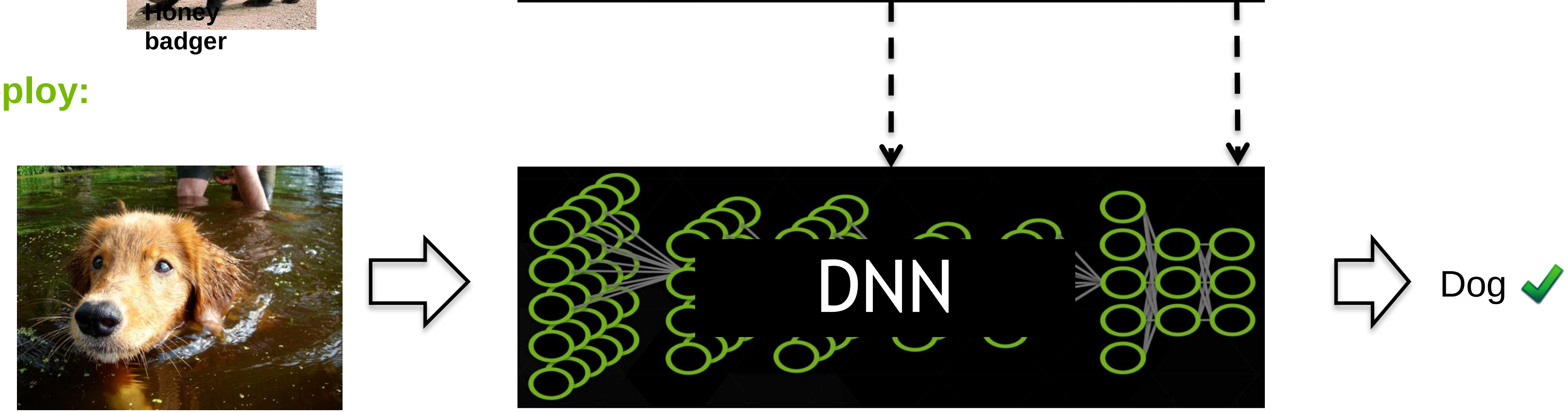


Deep learning approach

Train:



Deploy:

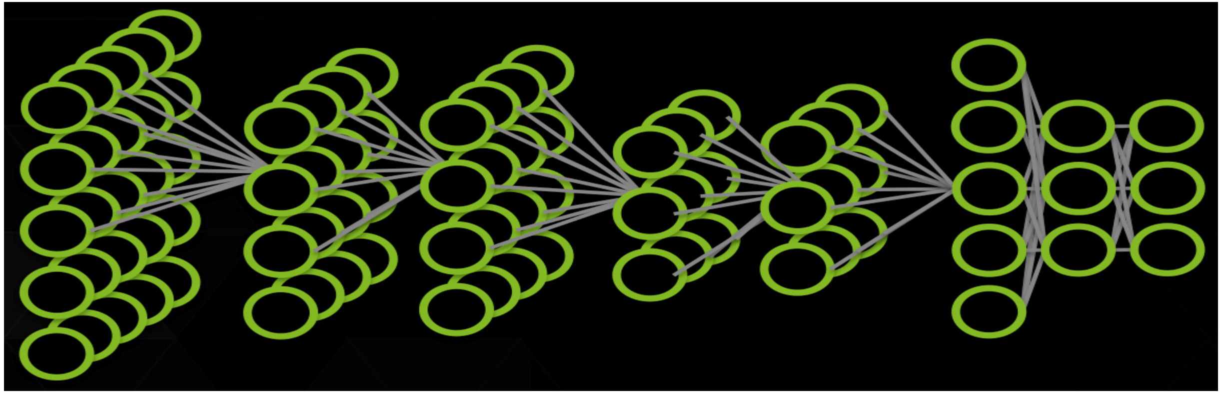


Artificial neural network

A collection of simple, trainable mathematical units that collectively learn complex functions

Hidden layers

Input layer

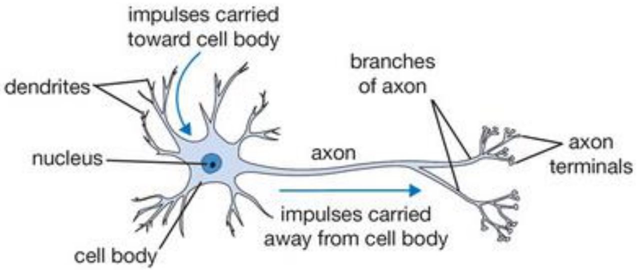


Output layer

Given sufficient training data an artificial neural network can approximate very complex functions mapping raw data to output decisions

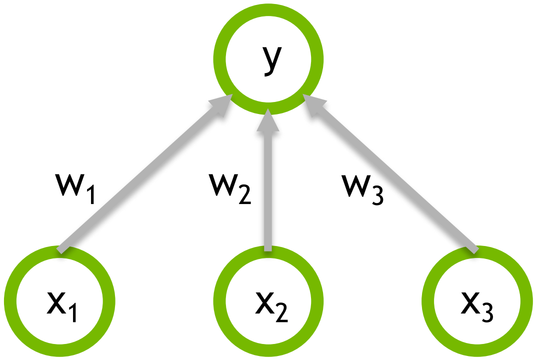
Artificial neurons

Biological neuron



From Stanford cs231n lecture notes

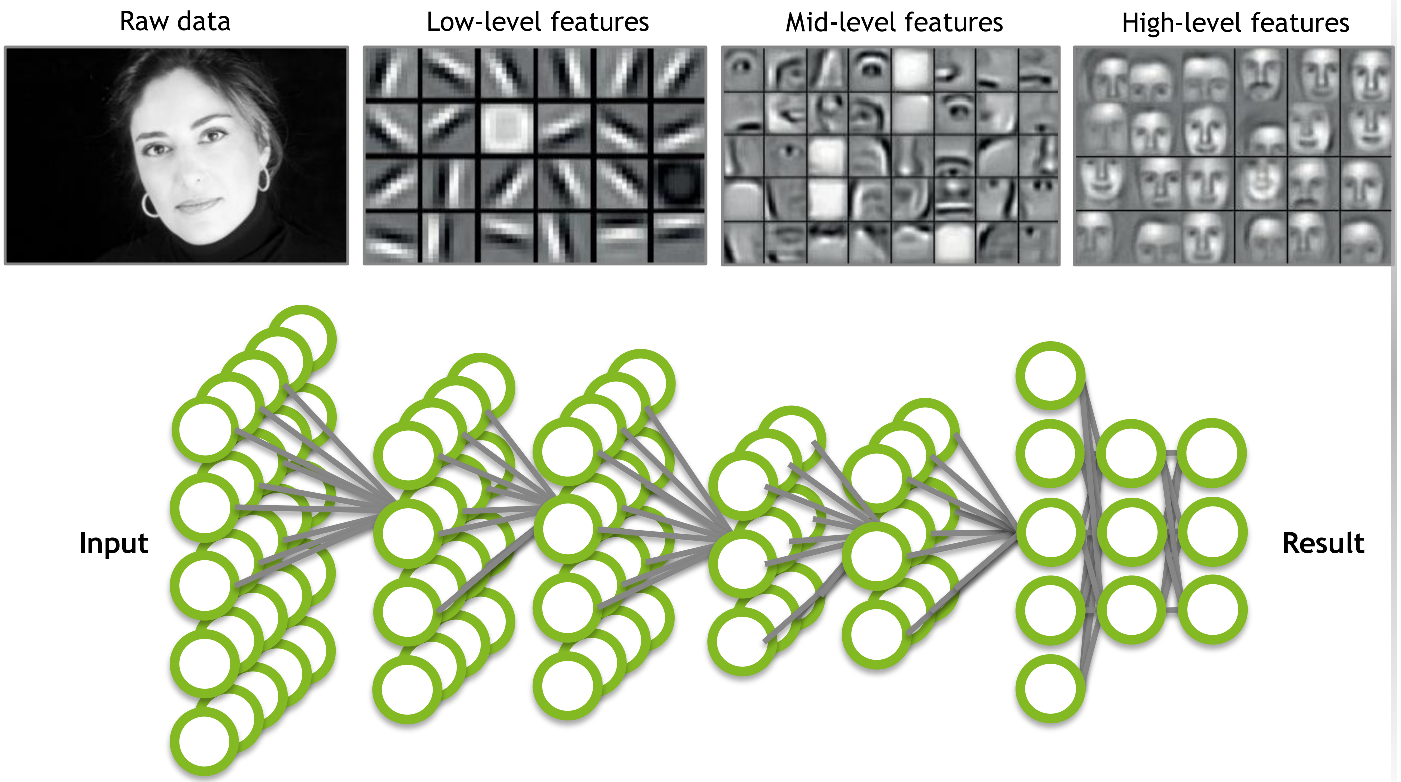
Artificial neuron



$$y=F(w_1x_1+w_2x_2+w_3x_3)$$

$$F(x)=max(0,x)$$

Deep neural network (DNN)



Application components:

Task objective
e.g. Identify face

Training data
10-100M images

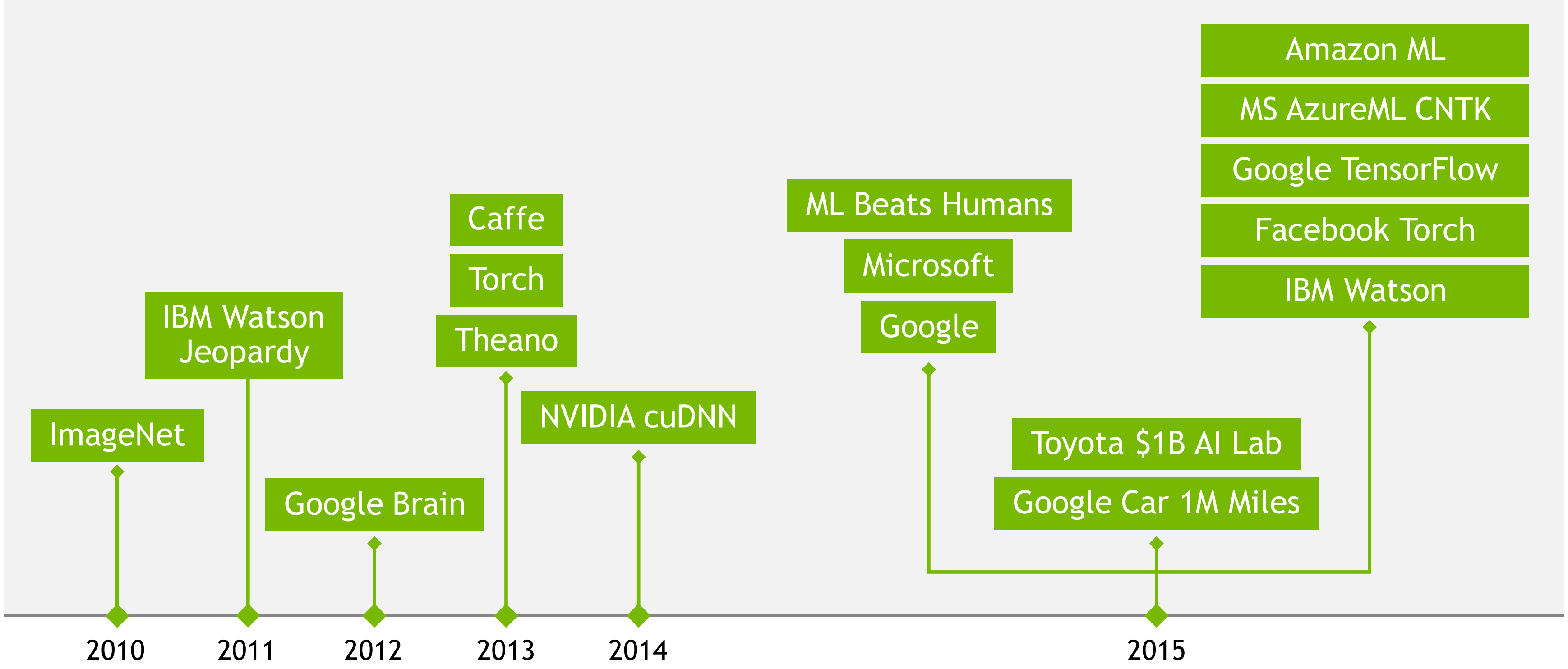
Network architecture
~10s-100s of layers
1B parameters

Learning algorithm
~30 Exaflops
1-30 GPU days

Deep learning benefits

- Robust
  - No need to design the features ahead of time – features are automatically learned to be optimal for the task at hand
  - Robustness to natural variations in the data is automatically learned
- Generalizable
  - The same neural net approach can be used for many different applications and data types
- Scalable
  - Performance improves with more data, method is massively parallelizable

The AI race is on



AlphaGo

First Computer Program to Beat a Human Go Professional

Training DNNs: 3 weeks, 340 million training steps on 50 GPUs

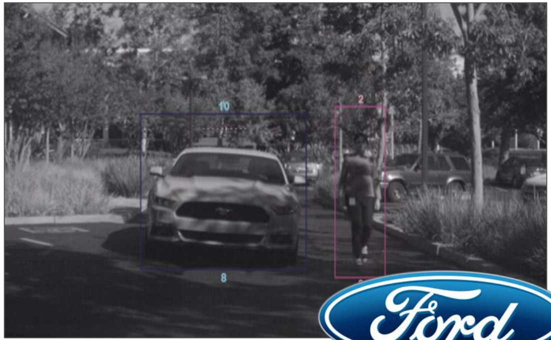
Play: Asynchronous multi-threaded search

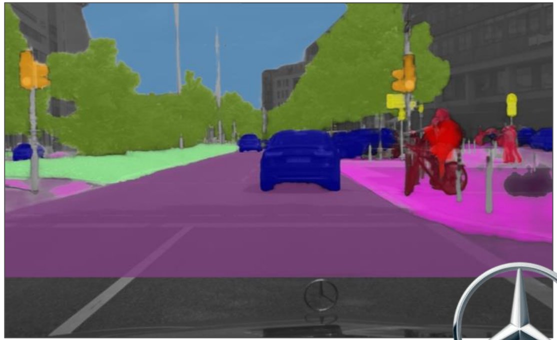
- Simulations on CPUs, policy and value DNNs in parallel on GPUs
- Single machine: 40 search threads, 48 CPUs, and 8 GPUs
- Distributed version: 40 search threads, 1202 CPUs and 176 GPUs

Outcome: Beat both European and World Go champions in best of 5 matches

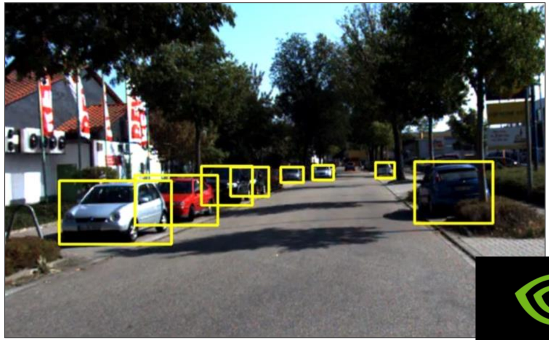
http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
http://deepmind.com/alpha-go.html

Deep Learning for Autonomous vehicles





Audi

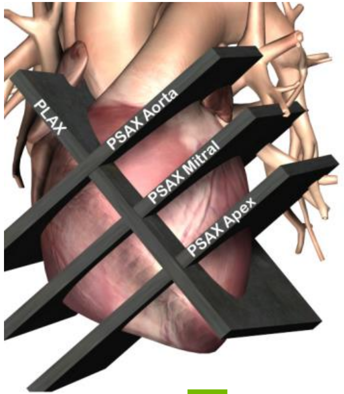


NVIDIA

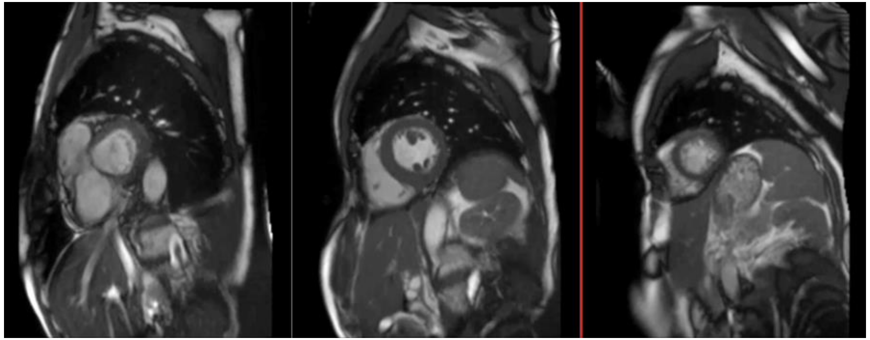
Audi

Automating Cardiac MRI analysis

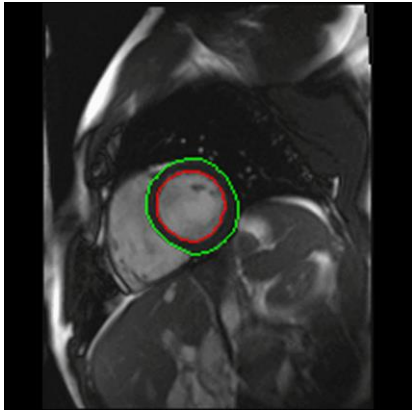
DL performance matches expert cardiologist at computing ejection fraction – a key indicator of heart disease



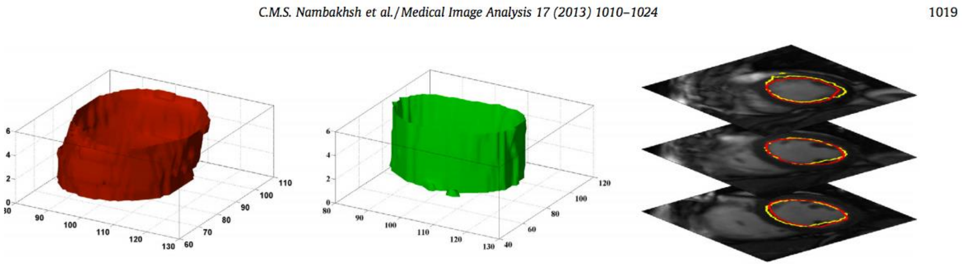
MRI imaging



Manual annotation



Software volume estimate



Safeguarding patients’ health through enhanced preventative medicine

- ‘Deep Patient’ analyzes electronic health records to predict 78 diseases, up to one year prior to onset
- Neural network trained on 100,000’s records using NVIDIA® Tesla® K80 GPU and CUDA® programming model.

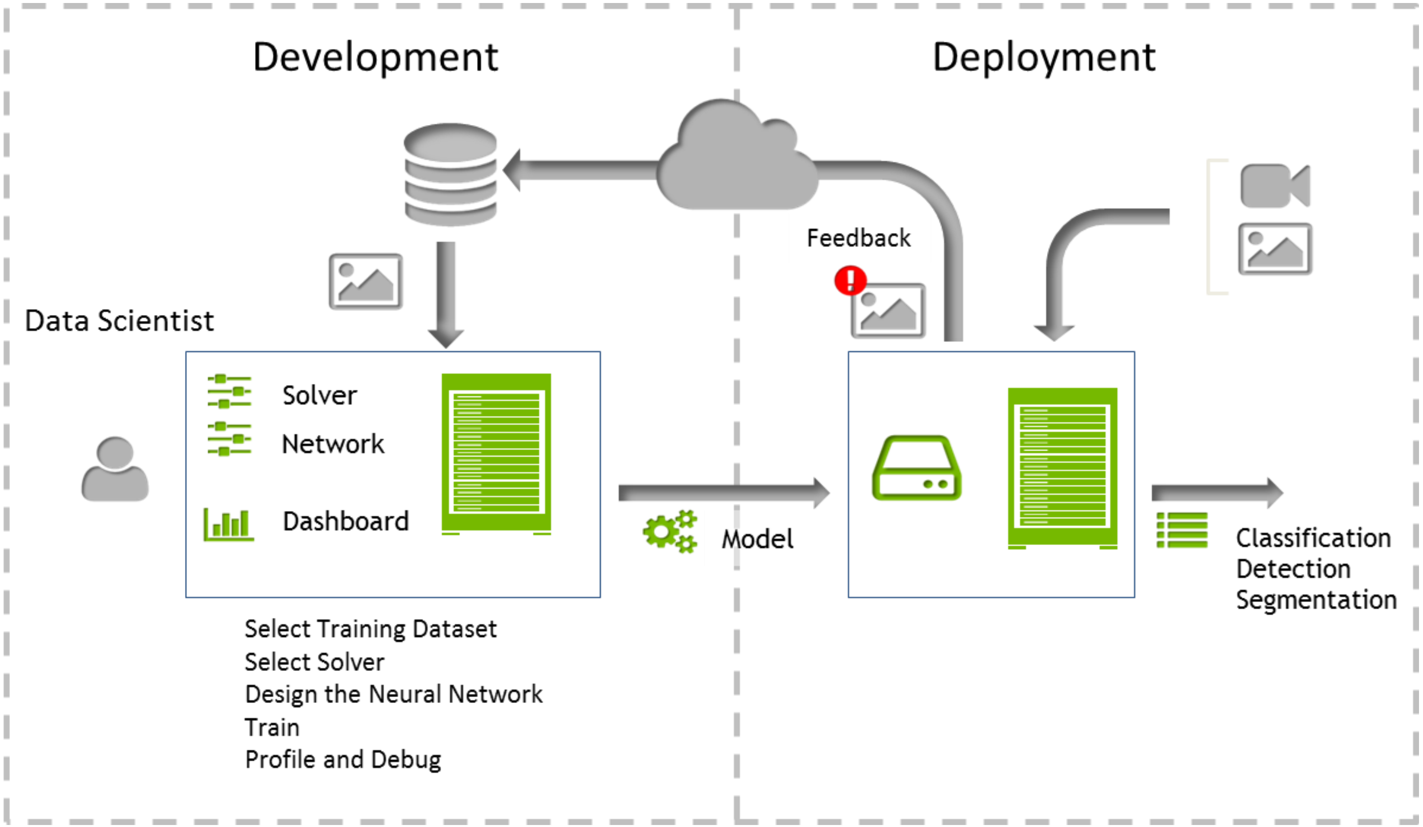
“For most diseases, prevention is easier than reversal. Deep Patient could have a huge impact on people’s health.”

-Joel T. Dudley, Assistant Professor of Genetics, Genomic Sciences Director of Biomedical Informatics

GPUs and DL

USE MORE PROCESSORS TO GO FASTER

Deep learning development cycle



TESLA PLATFORM FOR AI

NEURAL NETWORK COMPLEXITY IS EXPLODING

Bigger and More Compute Intensive

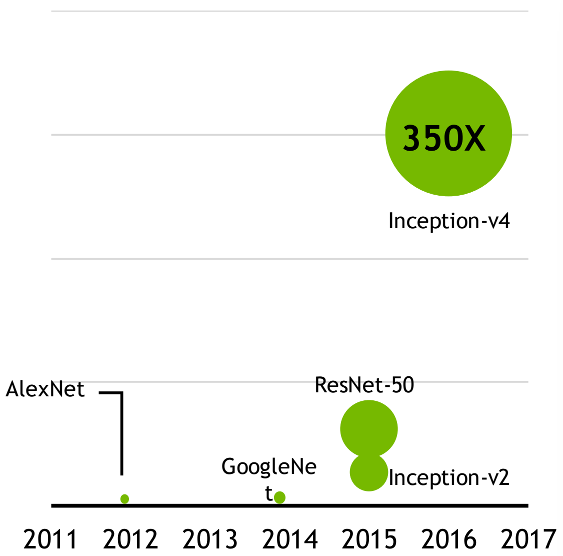
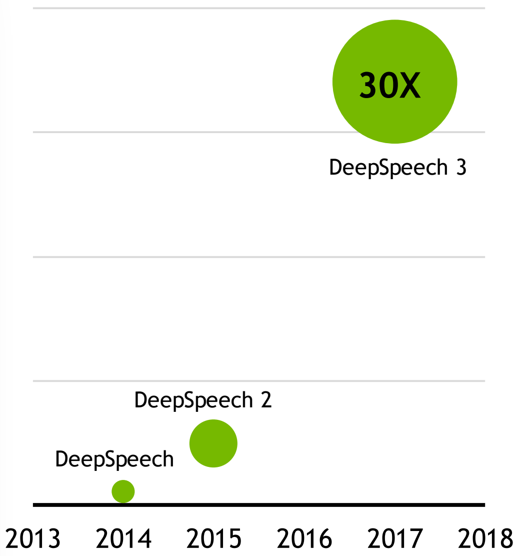
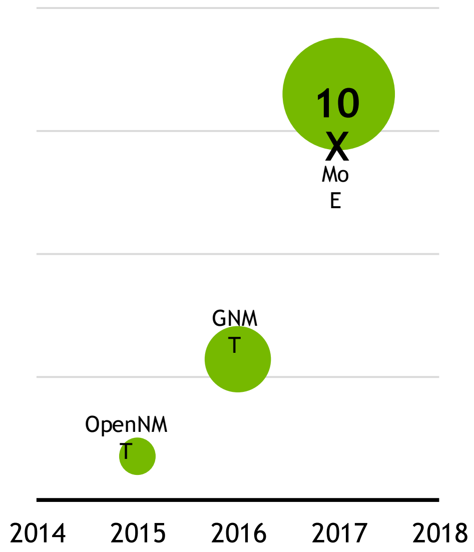


Image
(GOP * Bandwidth)



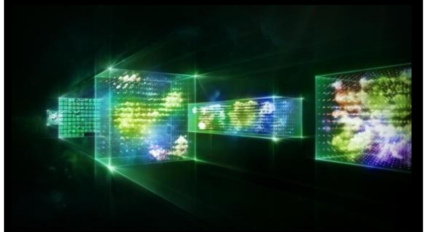
Speech
(GOP * Bandwidth)



Translation
(GOP * Bandwidth)

PLATFORM BUILT FOR AI

Delivering 125 TFLOPS of DL Performance with Volta



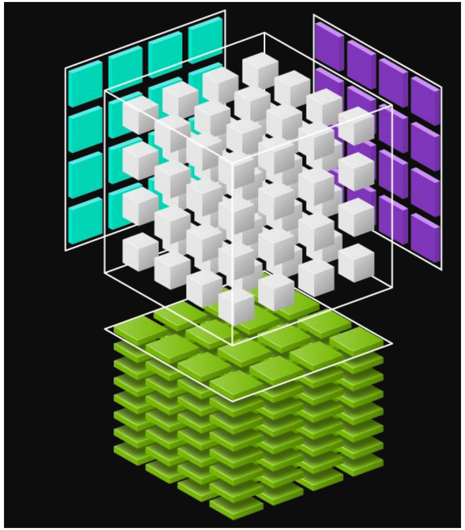
MATRIX DATA OPTIMIZATION:

Dense Matrix of Tensor Compute

TENSOR-OP CONVERSION:

FP32 to Tensor Op Data for Frameworks

VOLTA-OPTIMIZED cuDNN



VOLTA TENSOR CORE

4x4 matrix processing array

D[FP32] = A[FP16] * B[FP16] + C[FP32]

Optimized For Deep Learning

Caffe2

Microsoft Cognitive Toolkit

mxnet

PYTORCH

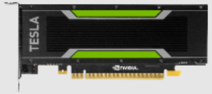
TensorFlow

ALL MAJOR FRAMEWORKS

END-TO-END PRODUCT FAMILY

HYPERSCALE HPC

Training & Inference - Tesla V100



Most Efficient Inference & Transcoding - Tesla P4

Deep learning training & inference

STRONG-SCALE HPC

Tesla V100 with NVLink

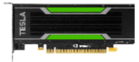
HPC and DL workloads scaling to multiple GPUs

MIXED-APPS HPC

Tesla V100 with PCI-E

HPC workloads with mix of CPU and GPU workloads

TESLA PRODUCTS RECOMMENDATION

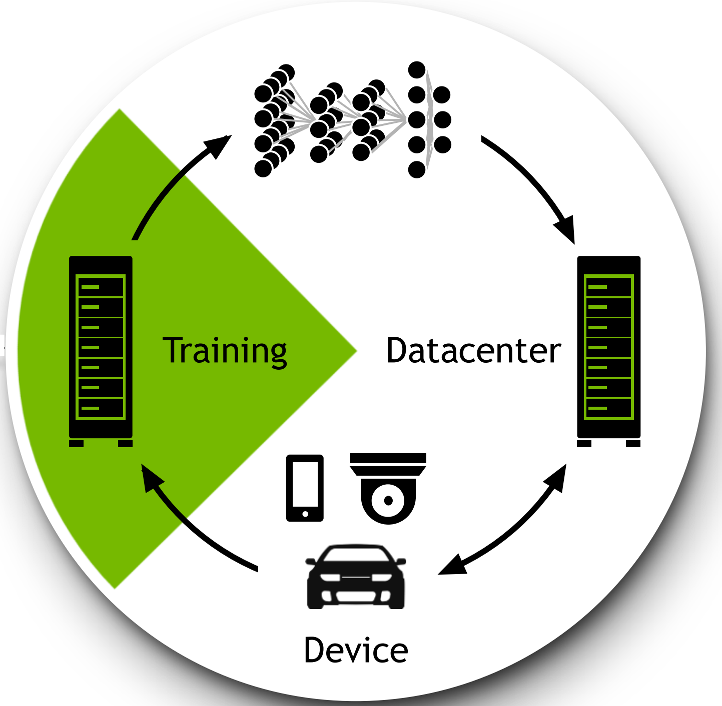
PRODUCT	V100			P4
	 V100 for PCIe	 V100 for NVLink	 V100 for Hyperscale	 P4 for PCIe
Target Use Cases	• Universal GPU for accelerating HPC and AI Workloads • Available in 3 form factors			• Low power, low profile optimized for scale out DL inference deployment • Most efficient inference and video processing
Form Factors	• Tesla V100 for NVLink: Ultimate performance for DL • Tesla V100 for PCIe: Highest versatility for all workloads • Tesla V100 for Hyperscale: Maximum efficiency for scale-out hyperscale data centers			• PCIe
Best Configs.	• PCIe: 2-4 GPU/node • NVLink: 8 way Hybrid Cube Mesh • Hyperscale: 2-4 GPU/node			• 1-2 GPU/node
1st Server Ship	• Available Now			• Available Now

GPU DEEP LEARNING IS A NEW COMPUTING MODEL

Billions of Trillions of Operations

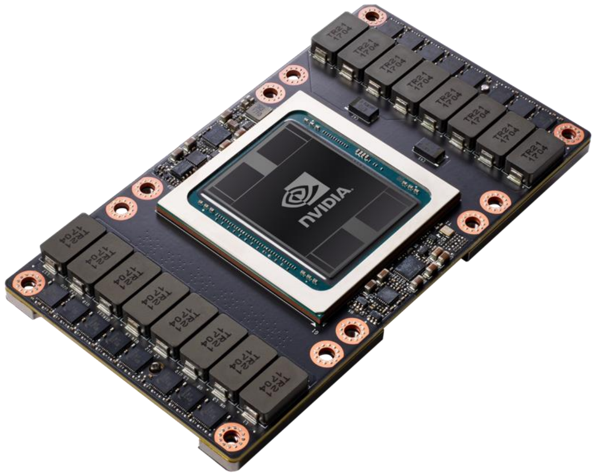
GPU train larger models, accelerate time to market

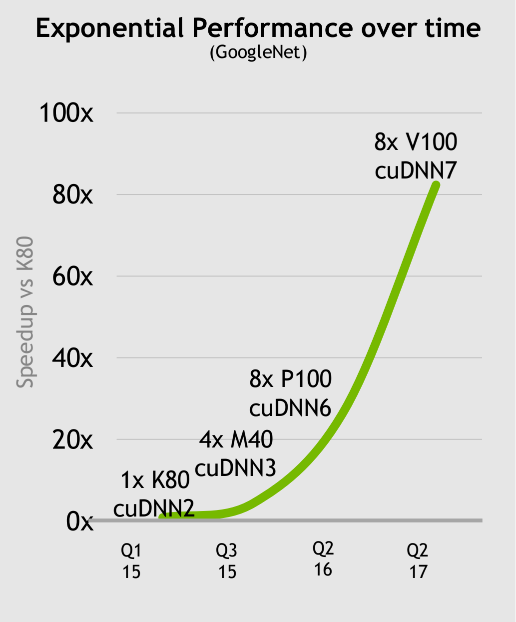
TRAINING



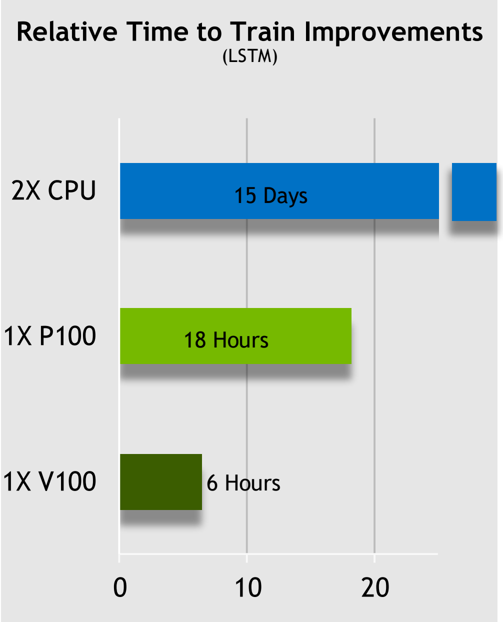
REVOLUTIONARY AI PERFORMANCE

3X Faster DL Training Performance





Over 80X DL Training Performance in 3 Years



3X Reduction in Time to Train Over P100

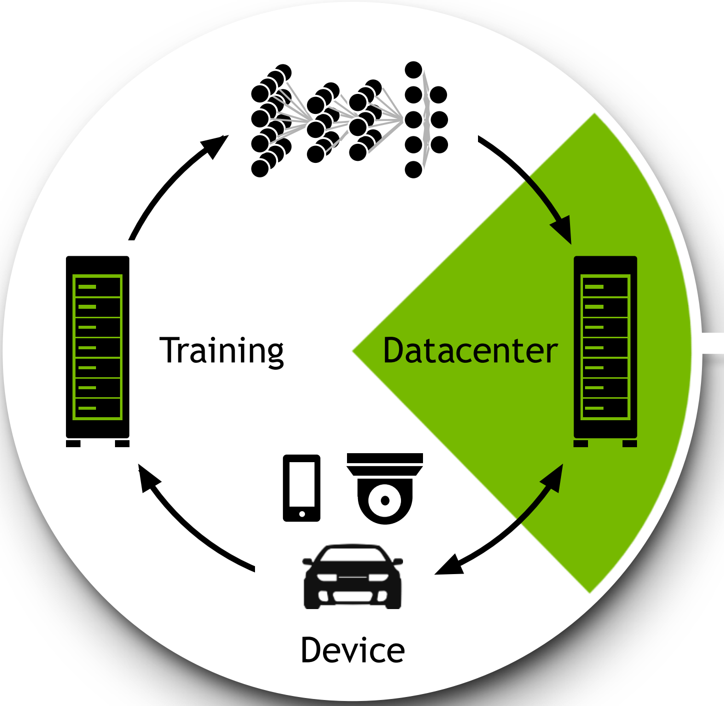
GoogleNet Training Performance on versions of cuDNN Vs 1x K80 cuDNN2

Neural Machine Translation Training for 13 Epochs |German ->English, WMT15 subset | CPU = 2x Xeon E5 2699 V4

TESLA V100

	 For NVLink Servers	 For PCIe Servers
Core	5120 CUDA cores, 640 Tensor cores	5120 CUDA cores, 640 Tensor cores
Compute	7.8 TF DP · 15.7 TF SP · 125 TF DL	7 TF DP · 14 TF SP · 112 TF DL
Memory	HBM2: 900 GB/s · 16 GB	HBM2: 900 GB/s · 16 GB
Interconnect	NVLink (up to 300 GB/s) + PCIe Gen3 (up to 32 GB/s)	PCIe Gen3 (up to 32 GB/s)
Power	300W	250W
Available	Now	Now

GPU DEEP LEARNING IS A NEW COMPUTING MODEL

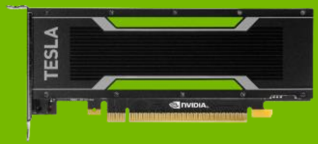


10s of billions of image, voice, video queries per day

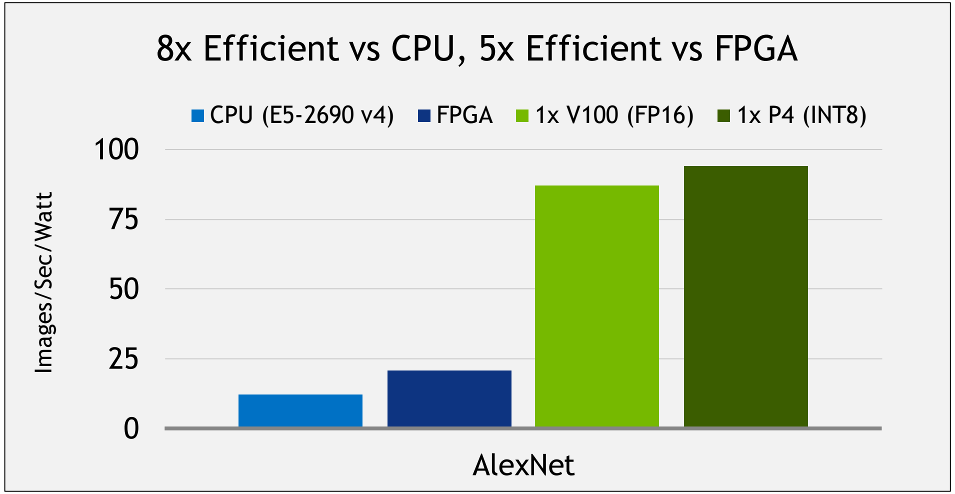
GPU inference for fast response, maximize datacenter throughput

DATACENTER INFERENCING

TESLA P4



Maximum Efficiency for Scale-out Servers



P4	
# of CUDA Cores	2560
Peak Single Precision	5.5 Teraflops
Peak INT8	22 TOPS
Low Precision	4x 8-bit vector dot product with 32-bit accumulate
Video Engines	1x decode engine, 2x encode engine
GDDR5 Memory	8 GB @ 192 GB/s
Power	50W & 75 W

AlexNet, batch size = 128, CPU: Intel E5-2690v4 using Intel MKL 2017, FPGA is Arria10-115
P4 board power: 64W, V100 board power: 218W, CPU power: 196W includes CPU and RAM power, FPGA performance/watt taken from Intel whitepaper titled “An OpenCLTM Deep Learning Accelerator on Arria 10”

DL in practice (DL SDK)

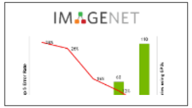
POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK accelerates every major framework

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION



SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

DEEP LEARNING FRAMEWORKS

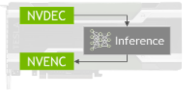
Caffe, Caffe2, Chainer, Microsoft Cognitive Toolkit, mxnet, PYTORCH, theano, torch, TensorFlow

NVIDIA DEEP LEARNING SDK

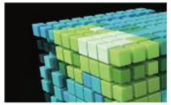
cuDNN

TensorRT

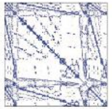
DeepStream SDK



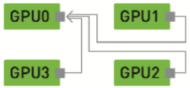
cuBLAS



cuSPARSE

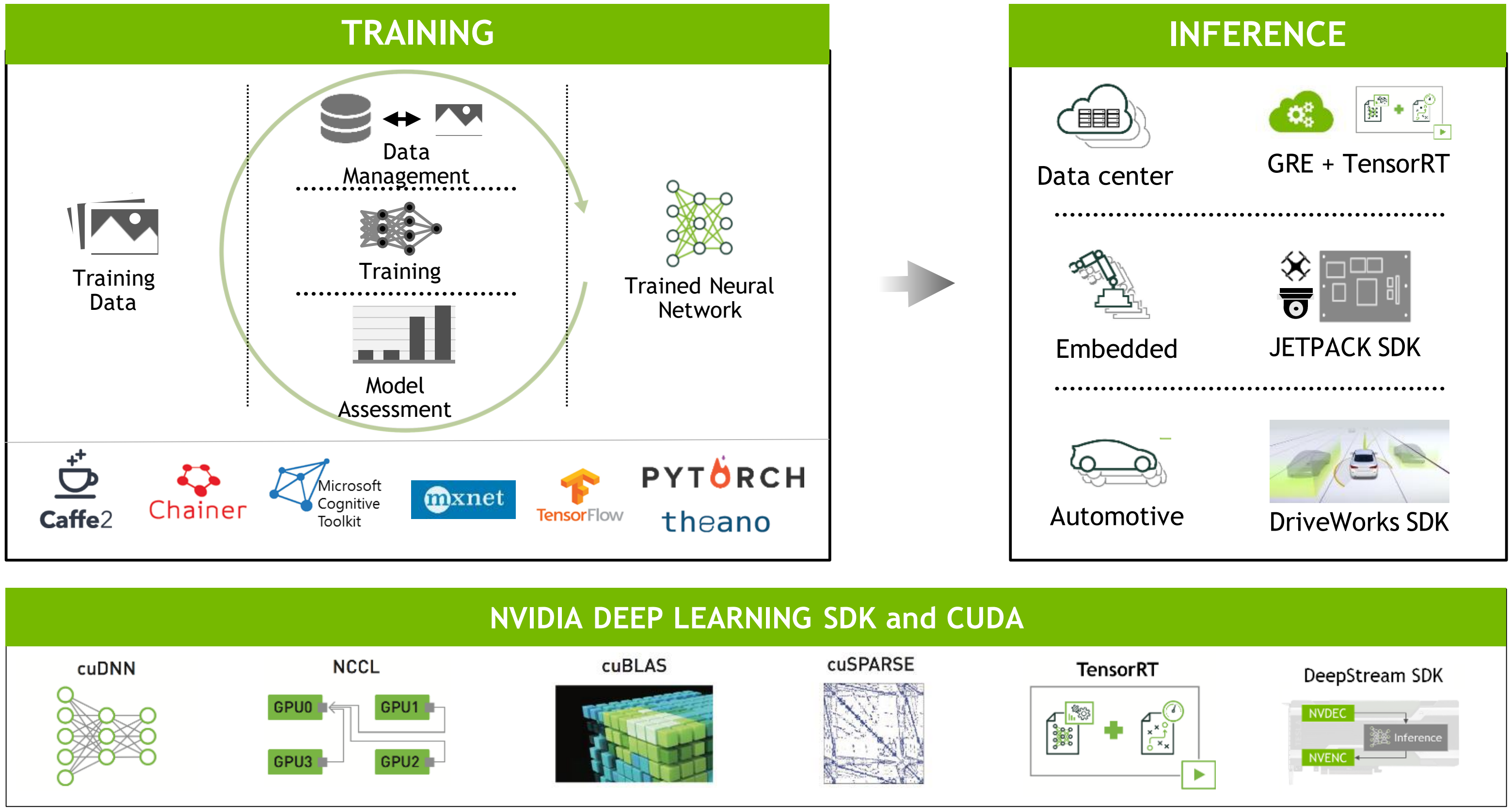


NCCL



developer.nvidia.com/deep-learning-software

NVIDIA DEEP LEARNING SOFTWARE PLATFORM



developer.nvidia.com/deep-learning-software

NVIDIA DEEP LEARNING SDK

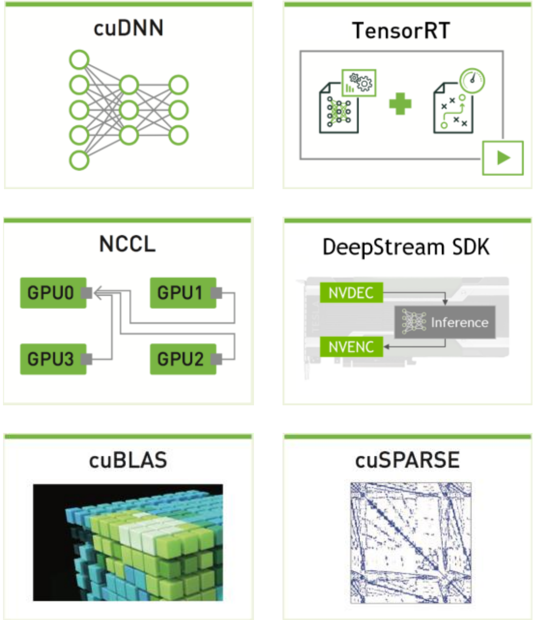
High performance GPU-acceleration for deep learning

Powerful tools and libraries for designing and deploying GPU-accelerated deep learning applications

High performance building blocks for training and deploying deep neural networks on NVIDIA GPUs

Industry vetted deep learning algorithms and linear algebra subroutines for developing novel deep neural networks

Multi-GPU and multi-node scaling that accelerates training on up to eight GPU



“ We are amazed by the steady stream of improvements made to the NVIDIA Deep Learning SDK and the speedups that they deliver.”

— Frédéric Bastien, Team Lead (Theano) MILA

developer.nvidia.com/deep-learning-software

NVIDIA cuDNN

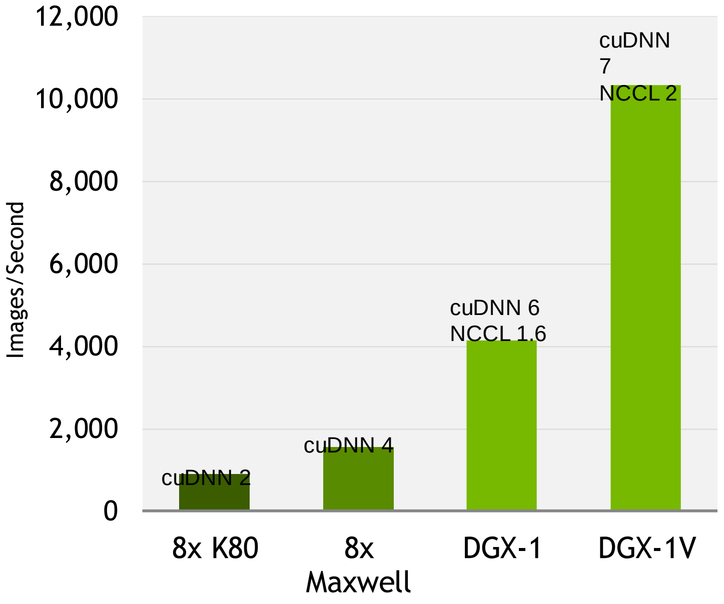
Deep Learning Primitives

High performance building blocks for deep learning frameworks

- Drop-in acceleration for widely used deep learning frameworks such as Caffe2, Microsoft Cognitive Toolkit, PyTorch, Tensorflow, Theano and others
- Accelerates industry vetted deep learning algorithms, such as convolutions, LSTM RNNs, fully connected, and pooling layers
- Fast deep learning training performance tuned for NVIDIA GPUs

developer.nvidia.com/cudnn

Deep Learning Training Performance



"NVIDIA has improved the speed of cuDNN with each release while extending the interface to more operations and devices at the same time."

— Evan Shelhamer, Lead Caffe Developer, UC Berkeley

NVIDIA TensorRT 3

Programmable Inference Accelerator

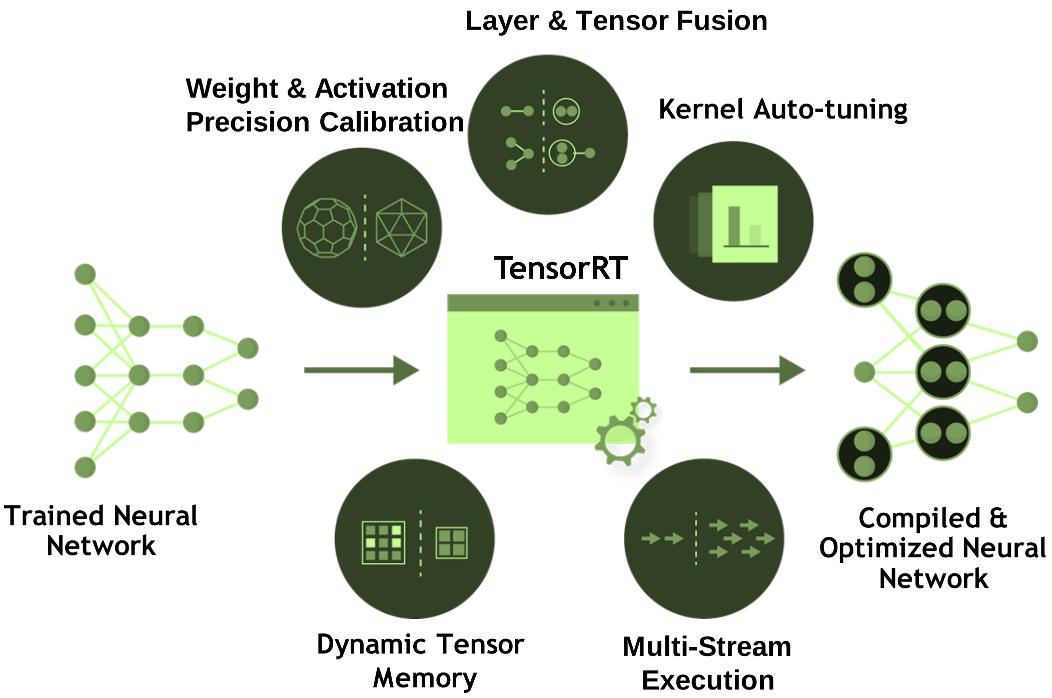
Compiler for Optimized Neural Networks

Weight & Activation Precision Calibration

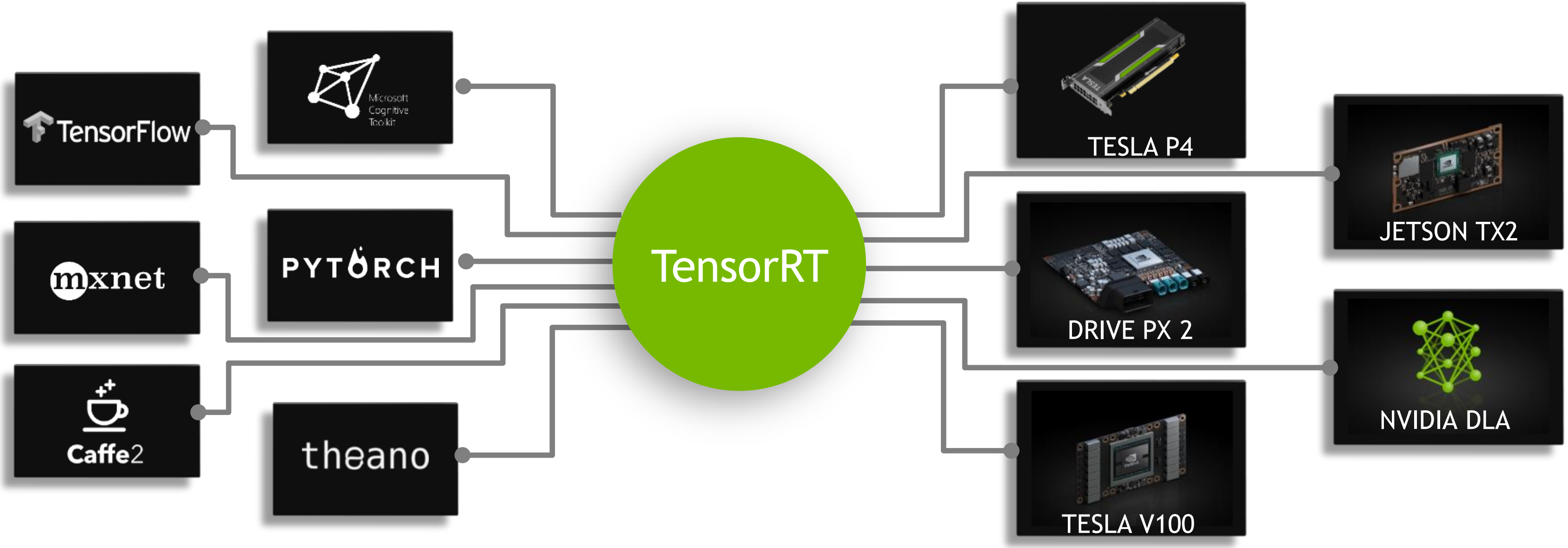
Layer & Tensor Fusion

Kernel Auto-Tuning

Multi-Stream Execution

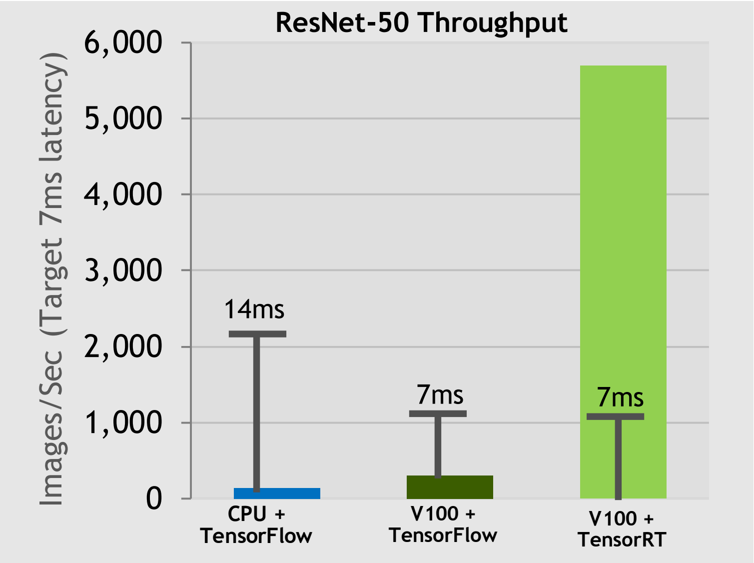


NVIDIA TENSORRT PROGRAMMABLE INFERENCE ACCELERATOR

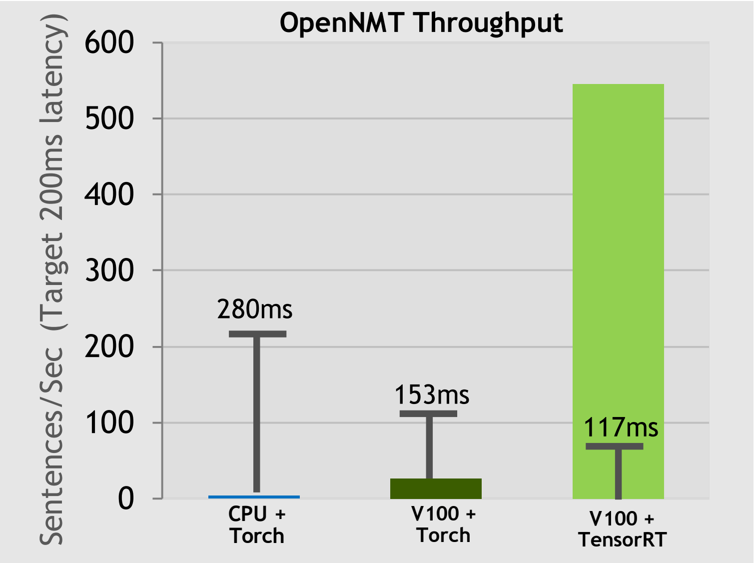


NVIDIA TENSORRT 3

World's Fastest Inference Platform



IMAGES



TRANSLATION

NVIDIA Collective Communications Library (NCCL)

Multi-GPU and multi-node collective communication primitives

High-performance multi-GPU and multi-node collective communication primitives optimized for NVIDIA GPUs

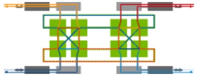
Fast routines for multi-GPU multi-node acceleration that maximizes inter-GPU bandwidth utilization

Easy to integrate and MPI compatible. Uses automatic topology detection to scale HPC and deep learning applications over PCIe and NVLink

Accelerates leading deep learning frameworks such as Caffe2, Microsoft Cognitive Toolkit, MXNet, PyTorch and more

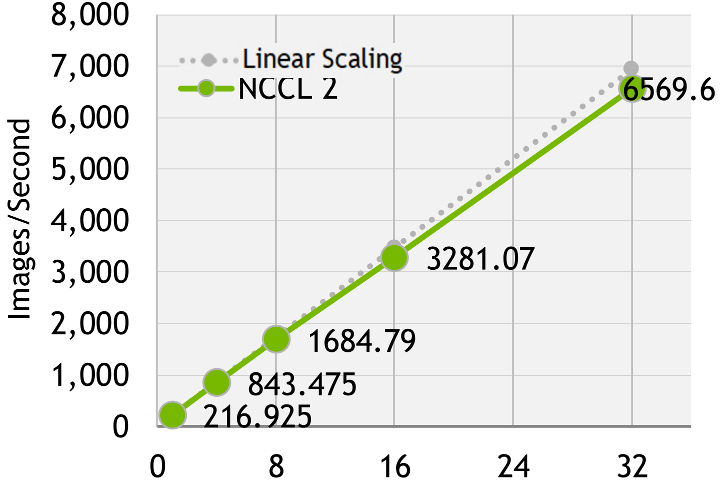
Multi-GPU: NVLink, PCIe

Multi-Node: InfiniBand verbs, IP Sockets



Automatic Topology Detection

Near-Linear Multi-Node Scaling



Microsoft Cognitive Toolkit multi-node scaling performance (images/sec), NVIDIA DGX-1 + cuDNN 6 (FP32), ResNet50, Batch size: 64

developer.nvidia.com/nccl

NVIDIA DIGITS

Interactive Deep Learning GPU Training System

Interactive deep learning training application for engineers and data scientists

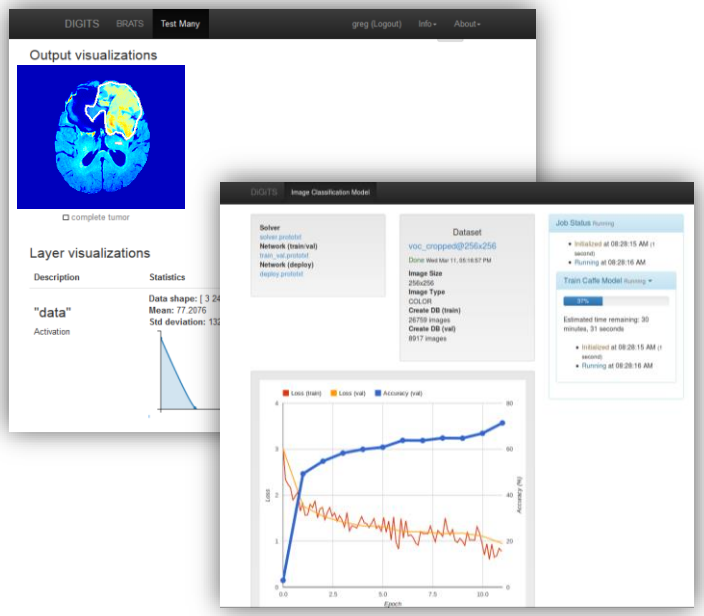
Simplify deep neural network training with an interactive interface to train and validate, and visualize results

Built-in workflows for image classification, object detection and image segmentation

Improve model accuracy with pre-trained models from the DIGITS Model Store

Faster time to solution with multi-GPU acceleration

developer.nvidia.com/digits



DIGITS Demo

Takeaways

TAKEAWAYS

DL, GPUs and Big Data are combining to create the "**Big Bang**" in Artificial Intelligence

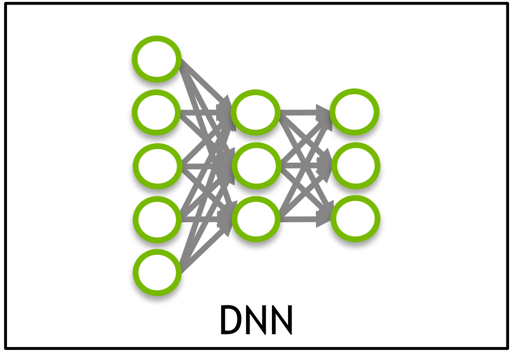
Deep Learning is solving problems that ML can't

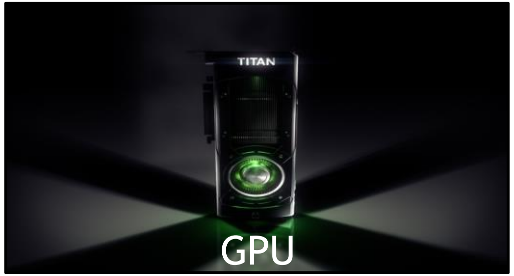
NVIDIA provides both HW and SW building blocks to accelerate DL workflows: **"You choose the DL framework, we'll make it run fast"**

Tesla V100 for most HPC and AI/DL workflows

Tesla P4 for datacenter inferencing

developer.nvidia.com/deep-learning





Agenda

- Introduction
- Anatomy of Research Computing
- Demo and Hands-on Cluster Operations
- UCS as a Platform for Research Computing
- Hands-on Bonus: UCS Configuration with Ansible
- NVIDIA Deep Learning with GPUs
- **Conclusion**

Continue your education

Demos in the Cisco campus

Walk-in self-paced labs

Meet the engineer 1:1 meetings

Related sessions

Continue your education

- BRKINI-2012 UCS Operational Agility and Best Practices
  - Monday, Jun 11, 04:00 p.m. - 05:30 p.m. | W311B
- BRKINI-2205 UCS Central Advanced Principles
  - Tuesday, Jun 12, 08:30 a.m. - 10:00 a.m. | W108A
- BRKINI-2348 Demystify AI/ML with Cisco UCS
  - Thursday, Jun 14, 01:00 p.m. - 02:30 p.m. | W208A
- DEVNET-1293 UCS Automation and Orchestration with Ansible
  - Monday, Jun 11, 10:30 a.m. - 11:15 a.m. | WoS, DevNet Classroom 1
- DEVNET-2562 UCS PowerTool – Deploy at Scale
  - Tuesday, Jun 12, 10:30 a.m. - 11:15 a.m. | WoS, DevNet Classroom 1

Related sessions

Complete your online session evaluation

Give us your feedback to be entered into a Daily Survey Drawing.

Complete your session surveys through the Cisco Live mobile app or on www.CiscoLive.com/us.

Don’t forget: Cisco Live sessions will be available for viewing on demand after the event at www.CiscoLive.com/Online.

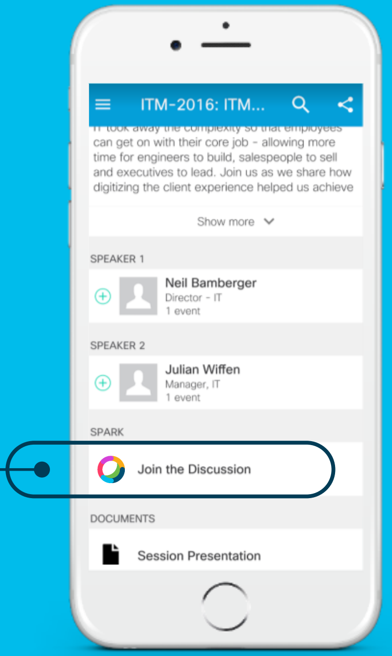
Cisco Webex Teams

Questions?

Use Cisco Webex Teams (formerly Cisco Spark) to chat with the speaker after the session

How

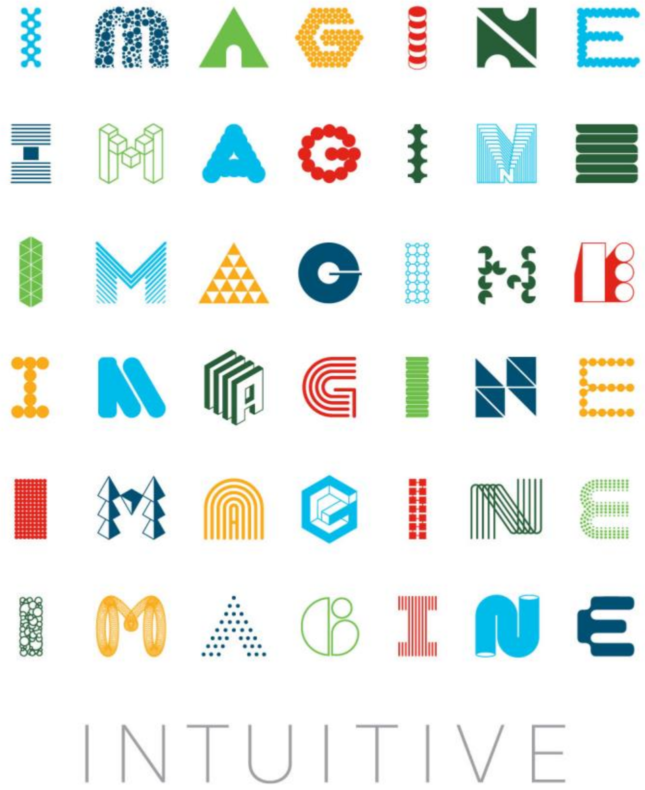
1. Find this session in the Cisco Live Mobile App
2. Click "Join the Discussion"
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space



Webex Teams will be moderated by the speaker until June 18, 2018.

cs.co/ciscolivebot#TECINI-2543

Thank you

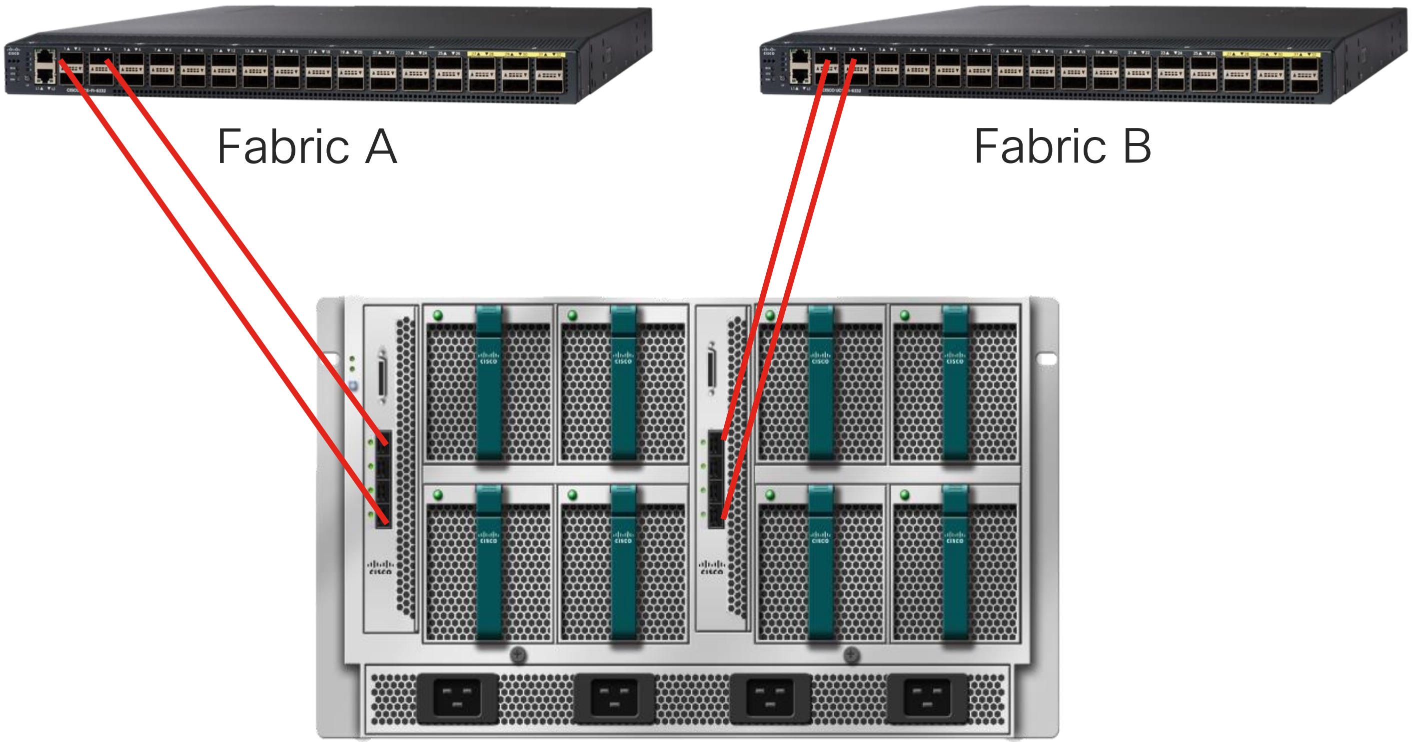


IMAGINE

INTUITIVE

Unified Computing Systems

UCS Blade Server Architecture



Nvidia GPUs for Compute on Cisco UCS M5 server

With full UCS-Manager integration

	P100 (UCSC-GPU-P100-12G UCSC-GPU-P100-16G)	P40 (UCSC-GPU-P40)	P4 (UCSC-GPU-P4)
GPU	GP100	GP102	GP104
PEAK FP64 (TFLOPs)	4.7	NA	NA
PEAK FP32 (TFLOPs)	9.3	12	5.5
PEAK FP16 (TFLOPs)	18.7	NA	NA
PEAK TIOPs	NA	47	22
Memory Size	16/12 GB HBM2	24 GB GDDR5	8 GB GDDR5
Memory BW	732/549 GB/s	346 GB/s	192 GB/s
Interconnect	PCIe Gen3	PCIe Gen3	PCIe Gen3
ECC	Internal + HBM2	GDDR5	GDDR5
Form Factor	PCIE Dual Slot	PCIE Dual Slot	PCIE LP
Power	250 W	250 W	50-75 W