ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

OPERATING SYSTEMS & PARALLEL COMPUTING

High Performance Computing (HPC)

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

HPC Overview

HPC Definition

"High-Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business."

--Inside HPC

Forms of HPC

- Dedicated supercomputer.
- Commodity HPC cluster.
- Grid computing.
- HPC in cloud.

What does HPC include ?

- **High-performance computing is fast computing**
  - Computations in parallel over lots of compute elements (CPU, GPU)
  - Very fast network to connect between the compute elements

- **Hardware**
  - Computer Architecture
    - Vector Computers, MPP, SMP, Distributed Systems, Clusters
  - Network Connections
    - InfiniBand, Ethernet, Proprietary (Myrinet, Quadrics, Cray-SeaStar etc.)

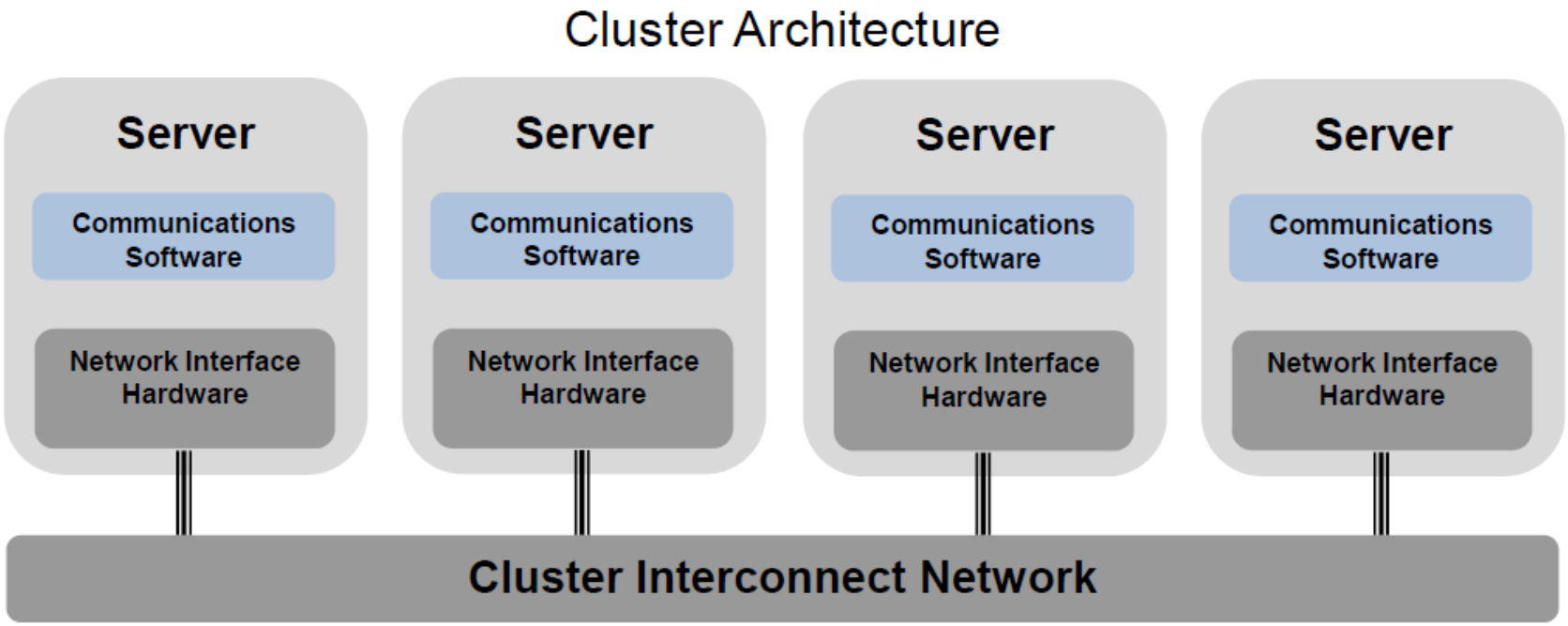
- **Software**
  - Programming models
    - MPI (Message Passing Interface), SHMEM (Shared Memory), PGAS, etc.
  - Applications
    - Open source, commercial

Rise and Fall of HPC Computer Architecture

- **Vector Computers (VC) - proprietary system**
  - Provided the breakthrough needed for the emergence of computational science, but they were only a partial answer
- **Massively Parallel Processors (MPP) - proprietary systems**
  - High cost and a low performance/price ratio.
- **Symmetric Multiprocessors (SMP)**
  - Suffers from scalability
- **Distributed Systems**
  - Difficult to use and hard to extract parallel performance
- **Clusters – commodity and highly popular**
  - High Performance Computing - Commodity Supercomputing
  - High Availability Computing - Mission Critical Applications

HPC Clusters: Affordable, Efficient, Scalable

- **Since the 1990s, there has been an increasing trend to move away from expensive /specialized proprietary parallel supercomputers to clusters of computers**
  - From specialized supercomputers to cost effective, general purpose systems
- **So What's So Different about Clusters?**
  - Commodity, standard, affordable, cost effective, scalable and reliable architecture



HPC Workloads

- HPC workloads aggregate computing power to solve large problems in science, engineering, or business

Fin-tech
Risk analysis
High-frequency trading

Manufacturing
Product Lifecycle Management,
CAE, Fluid dynamics

Retail
Recommendation engines
Customer analysis

Genomics
DNA sequencing
Protein analysis

Media and entertainment
3D-rendering
Transcoding

Oil and gas
Reservoir modelling
Seismic analysis

HPC Workloads Challenge

Cost of capacity and “right-sizing” purchases

“We have high capital costs and long purchasing cycles for HPC infrastructure.”

“Each provisioning cycle we either buy too little and run short of capacity after a few months, or buy to meet peak needs and end up underutilizing it most of the time. Either way, we live with this choice for years.”

“Still, we are always looking for as much compute capacity as possible, at the lowest cost.”

On-premises level of control and performance

“I don’t want to lose the performance and control we have on-premises.”

“Some of our applications demand large storage capacity, high bandwidth, and low latency that we haven’t seen in public clouds.”

Cloud flexibility in cost and scheduling

“Long budget cycles don’t match short-term projects well.”

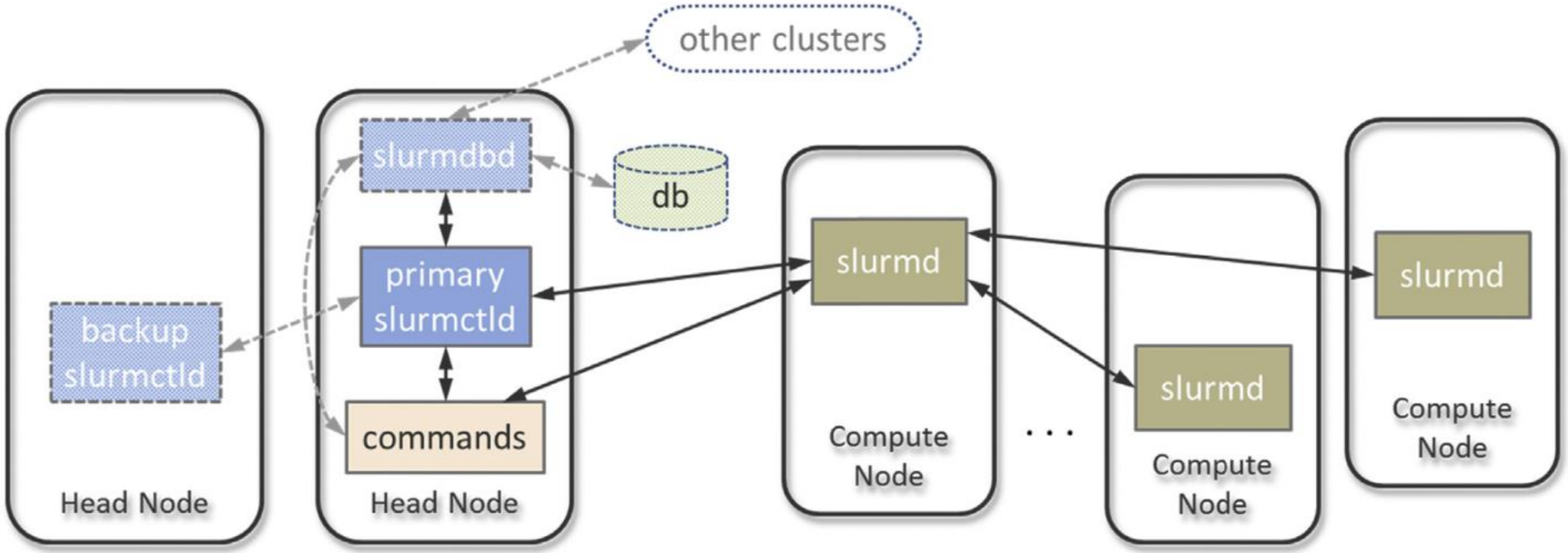
“New projects must conform to architectural decisions made for past projects, often resulting in sub-optimal solutions.”

“We need to be able to try out the newest technologies without the high up-front costs and long budget cycles required to bring them on site at scale.

What is SLURM?

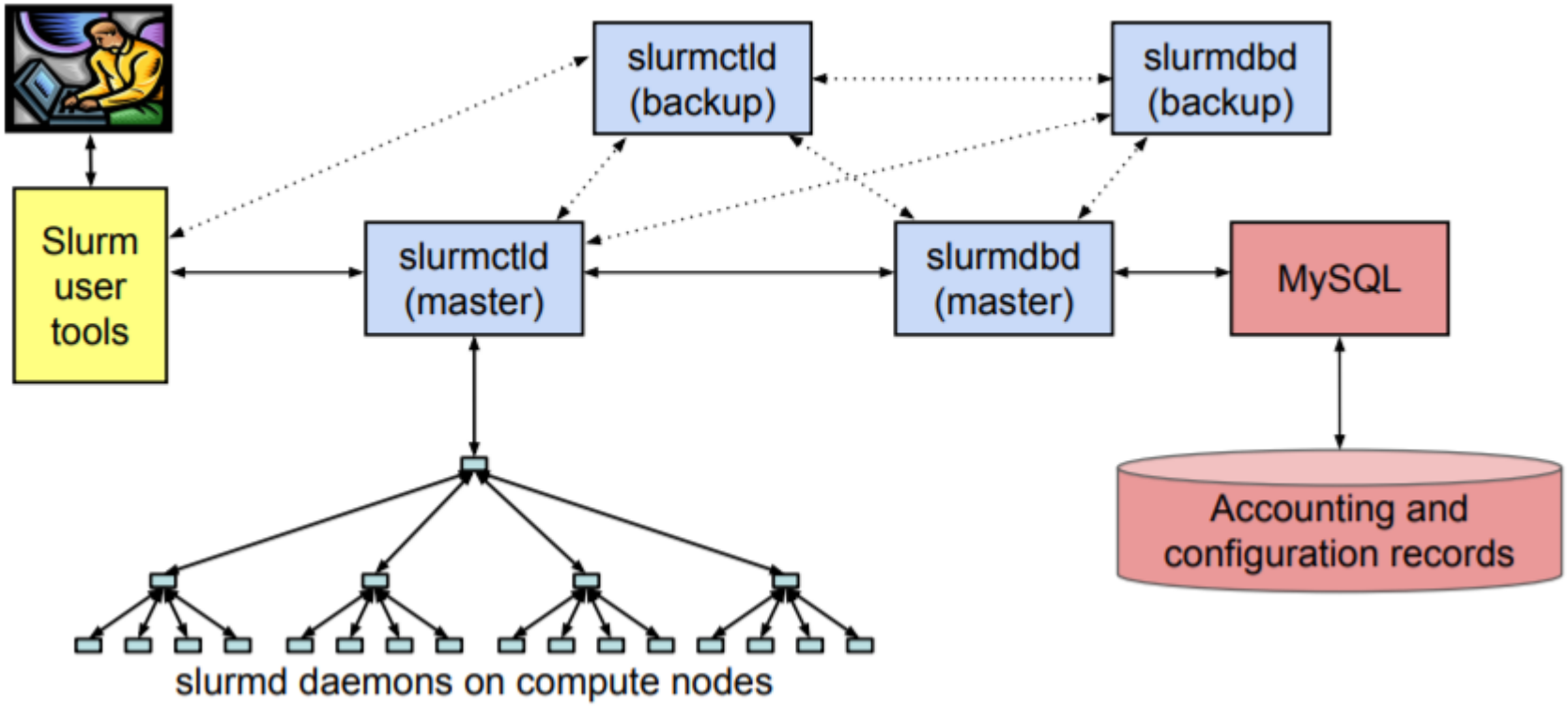
- Historically Slurm was an acronym standing for
  - **S**imple **L**inux **U**tility for **R**esource **M**anagement
- Development started in 2002 at Lawrence Livermore National Laboratory as a resource manager for Linux clusters
- Sophisticated scheduling plugins added in 2008
- About 500,000 lines of C code today (plus test suite and doc)
- Used on many of the world's largest computers
- Active global development community

Simplified architecture of SLURM.



Components framed by dashed lines are optional.

Cluster Architecture



Role of Resource Manager

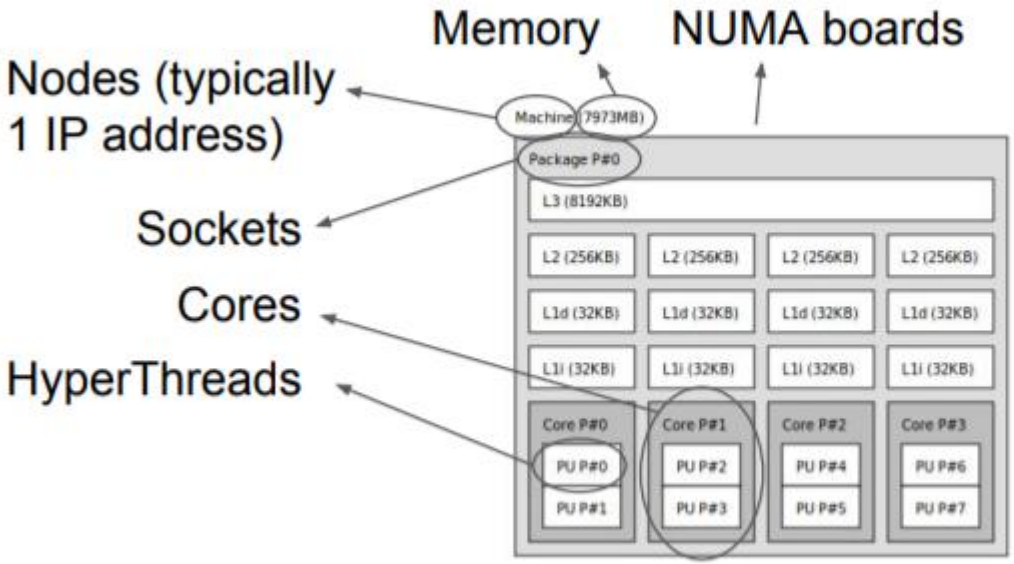
- The “glue” for a parallel computer to execute parallel jobs
- It should make a parallel computer as almost easy to use as a PC

On a PC. Execute program “a.out” `a.out`	On a cluster. Execute 8 copies of “a.out” `srun -n8 a.out`

- MPI would typically be used to manage communications within the parallel program

Roles of Resource Manager

- Allocate resources within a cluster

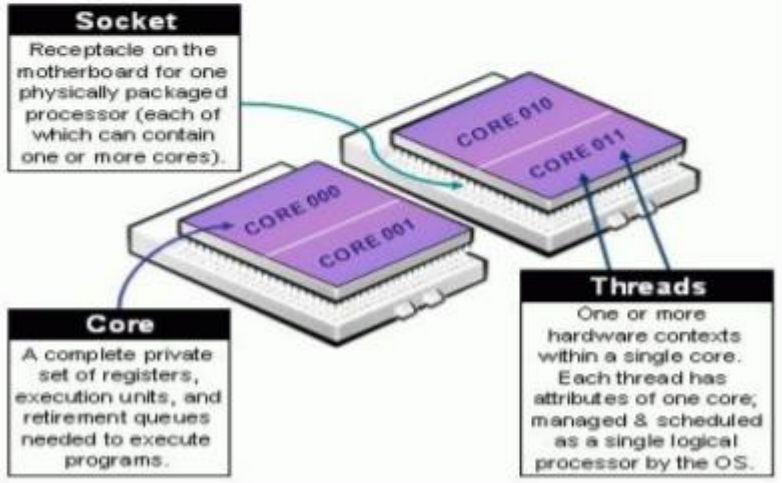


Interconnect/Switch resources

Licenses

Generic Resources (e.g. GPUs)

Can require extensive knowledge about the hardware and system software (e.g. to alter network routing or manage switch window)



- Launch and otherwise manage jobs

Role of Job Scheduler

- When there is more work than resources, the job scheduler manages queue(s) of work
  - Supports complex scheduling algorithms
    - Optimized for network topology, fair-share scheduling, advanced reservations, preemption, gang scheduling (time-slicing jobs), backfill scheduling, etc.
    - Job can be prioritized using highly configurable parameters such as job age, job partition, job size, job QOS, etc.
  - Supports resource limits (by queue, user, group, etc.)

ie
SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

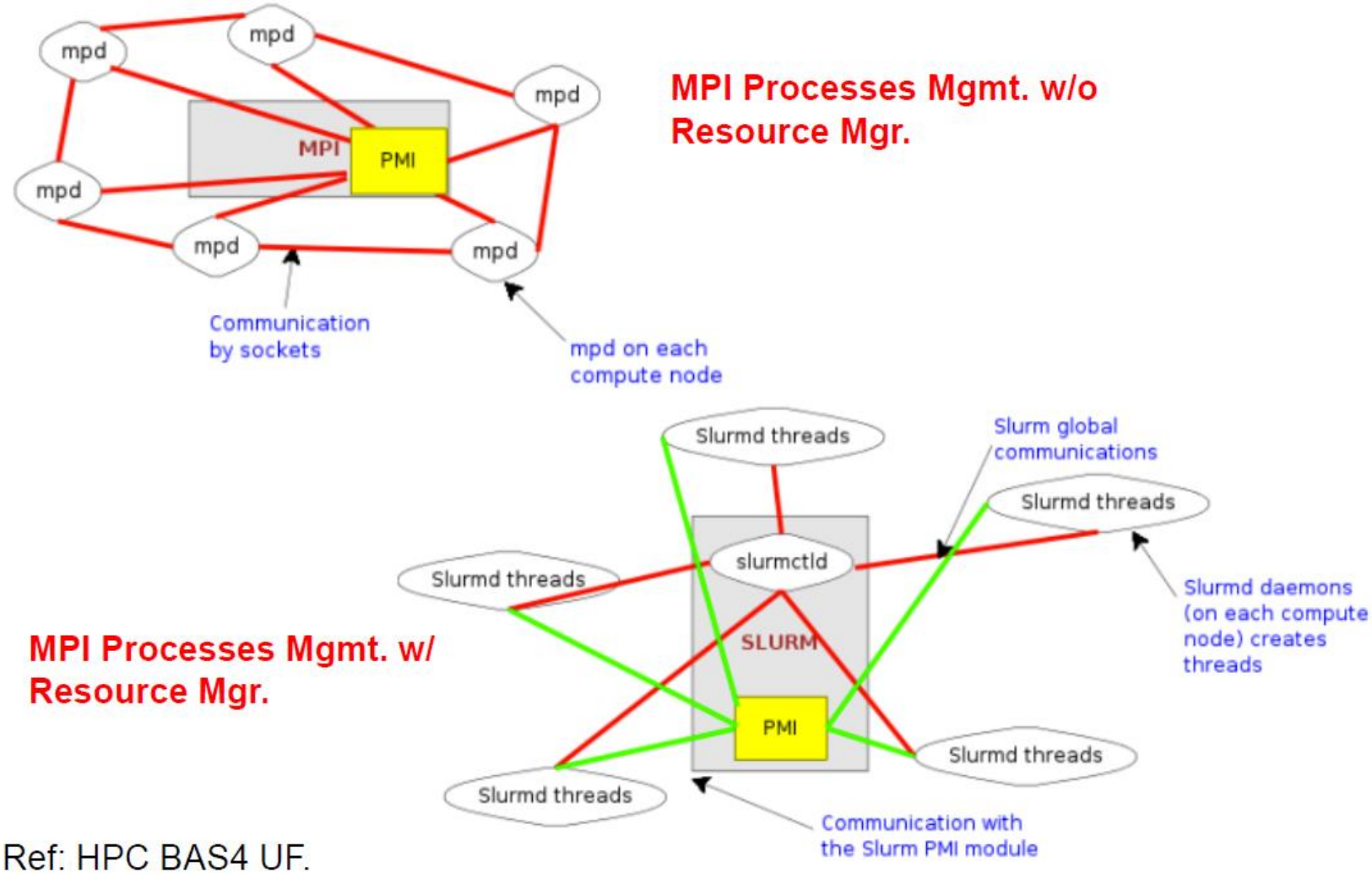
Daemons

- **slurmctld** – Central controller (typically one per cluster)
  - Monitors state of resources
  - Manages job queues
  - Allocates resources
- **slurmdbd** – Database daemon (typically one per enterprise)
  - Collects accounting information
  - Uploads configuration information (limits, fair-share, etc.) to slurmctld

Daemons

- **slurmd** – Compute node daemon (typically one per compute node)
  - Launches and manages slurmstepd (see below)
  - Small and very light-weight
  - Quiescent after launch except for optional accounting
  - Supports hierarchical communications with configurable fanout
- **slurmstepd** – Job step shepherd
  - Launched for batch job and each job step
  - Launches user application tasks
  - Manages application I/O, signals, etc.

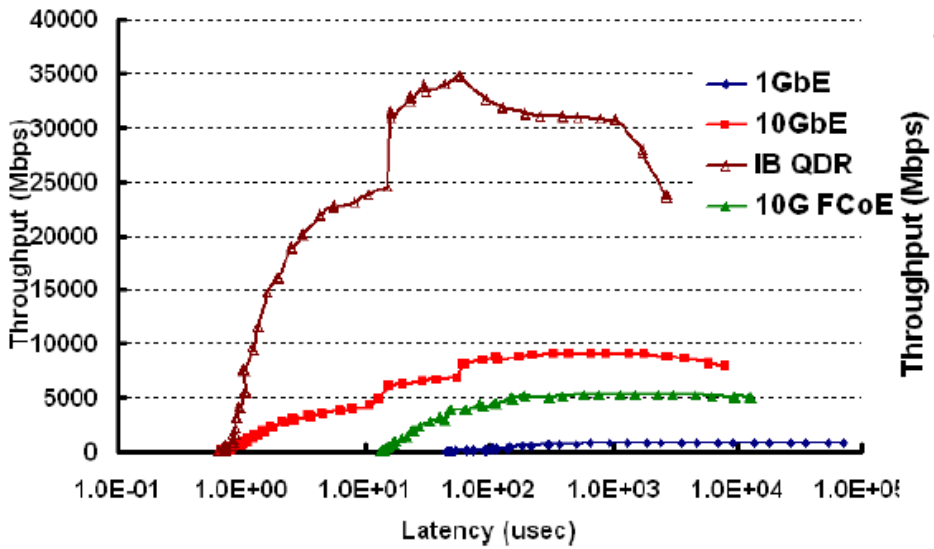
Cluster MPI & Resource Manager

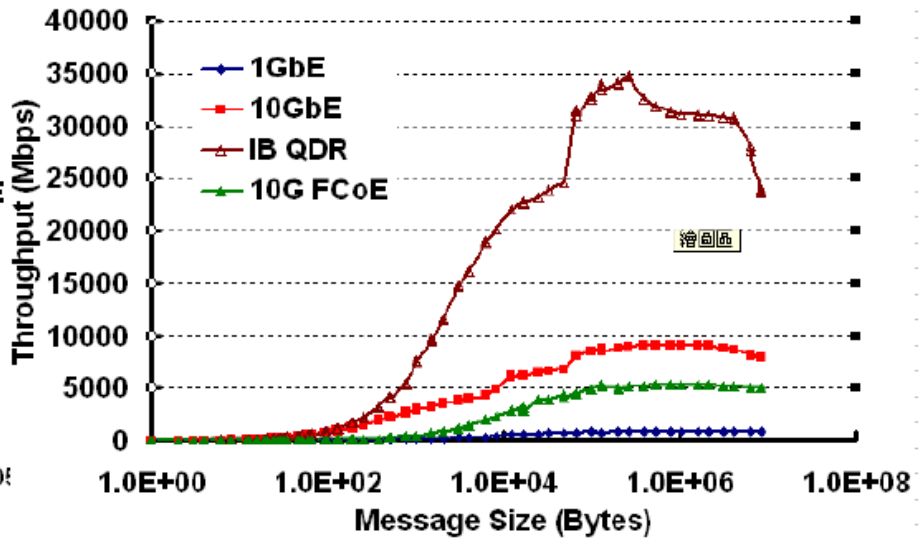


Ref: HPC BAS4 UF.

Network Performance: Thoughput vs Latency

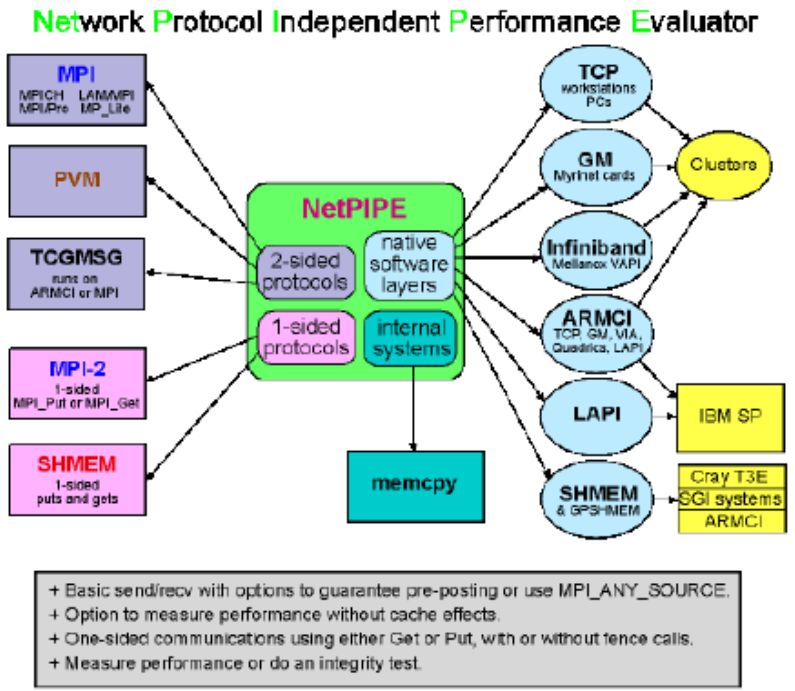
- Peak 10G 9.1Gbps ~ 877 usec (Msg Size: 1MB)
- IB QDR reach 31.1Gbps with same msg size
  - Only 29% of 10G Latency (~256 usec)
  - Peak IB QDR 34.8Gbps ~ 57 usec (Msg Size: 262KB)

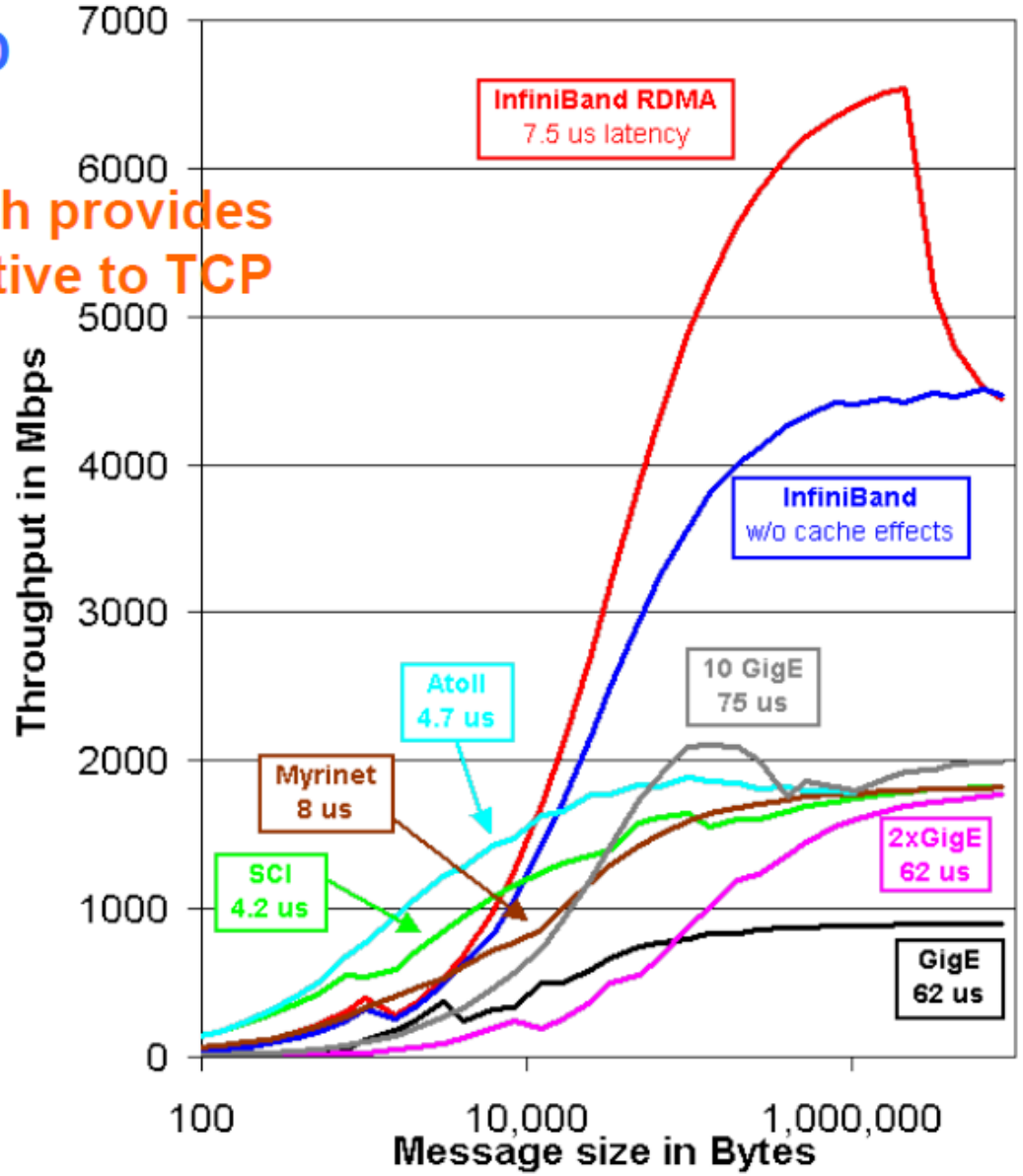




Cluster: File Server Performance

- Preload SDP provided by OFED
- Sockets Direct Protocol (SDP)
  - Note: Network protocol which provides an RDMA accelerated alternative to TCP over InfiniBand





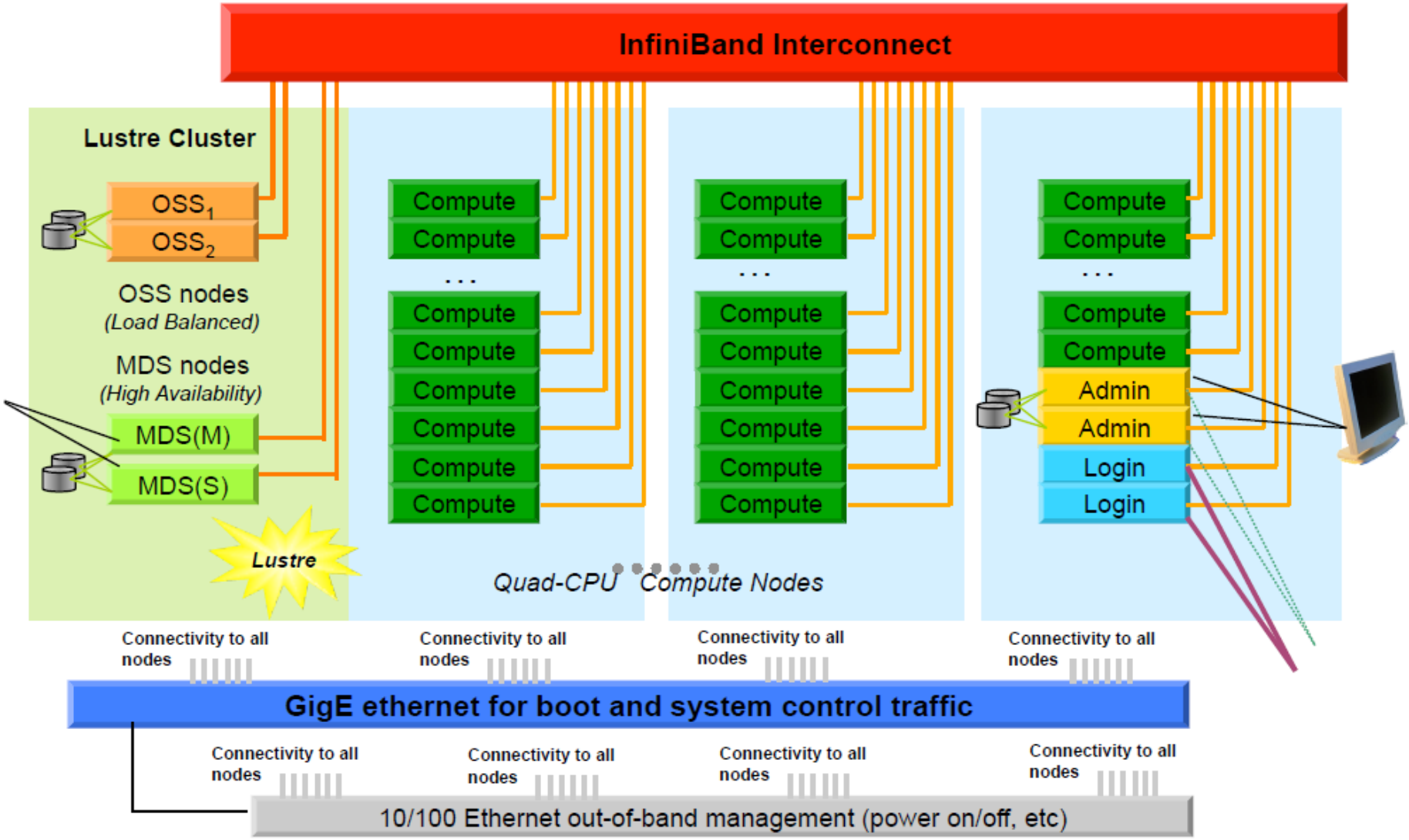
Cluster I/O: Cluster FS

- OCFS2 (Oracle Cluster File System)
  - Once proprietary, now GPL
  - Available in Linux vanilla kernel
  - not widely used outside the database world
- PVFS (Parallel Virtual File System)
  - Open source & easy to install
  - Userspace-only server
  - kernel module required only on clients
  - Optimized for MPI-IO
  - POSIX compatibility layer performance is sub-optimal
- pNFS (Parallel NFS)
  - Extension of NFSv4
  - Proprietary solutions available: “Panasas”
  - Put together benefits of parallel IO using standard solution (NFS)

Cluster I/O: Cluster FS (2)

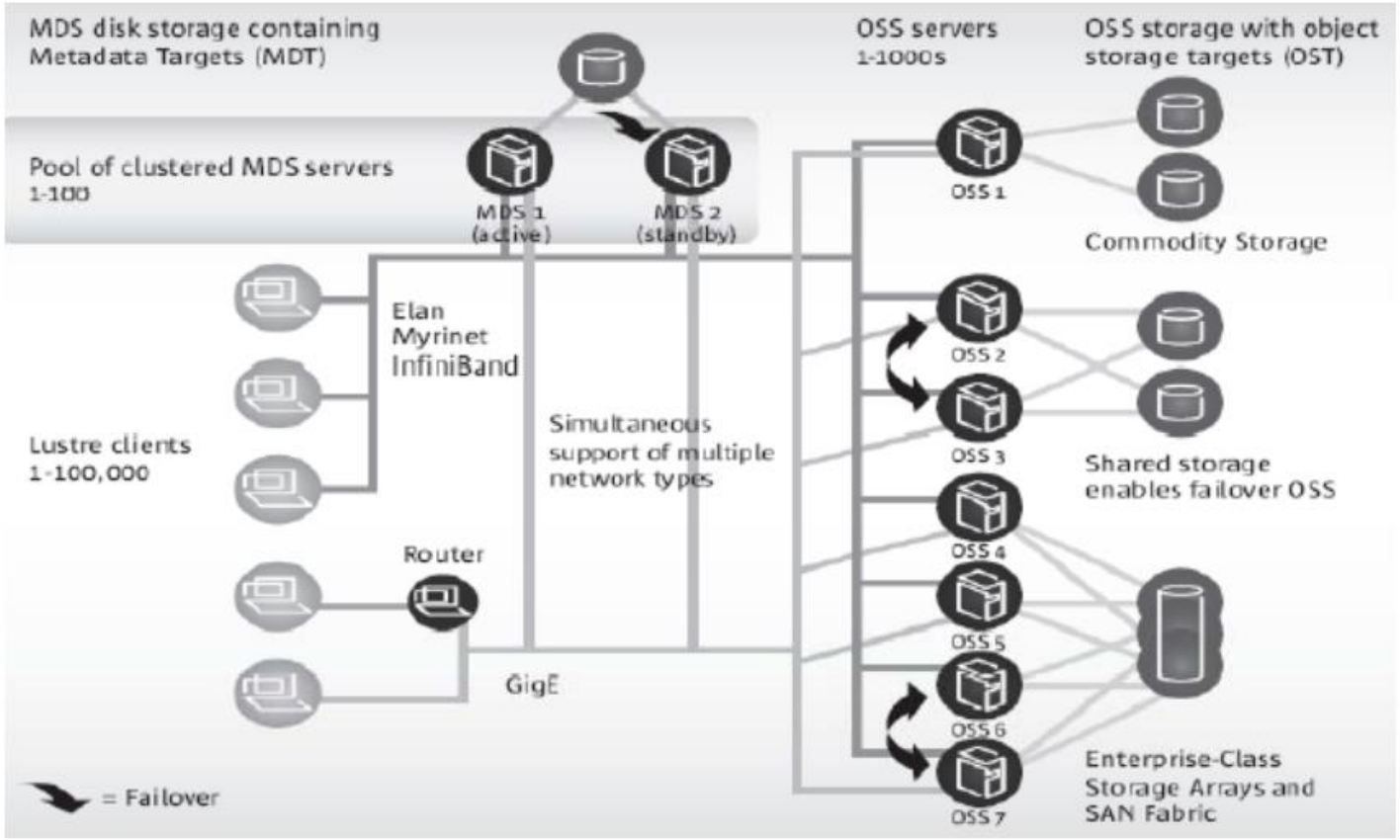
- **GPFS (General Parallel File System)**
  - Rock-solid w/ 10-years history
  - Available for AIX, Linux & Windows Server 2003
  - Proprietary license
  - Tightly integrated with IBM cluster management tools
- **Lustre**
  - HA & LB implementation
  - highly scalable parallel filesystem: ~ 100K clients
  - Performance:
    - Client: ~1 GB/s & 1K Metadata Op/s
    - MDS: 3K ~ 15K Metadata Op/s
    - OSS: 500 ~ 2.5 GB/s
  - POSIX compatibility
  - Components:
    - single or dual Metadata Server (MDS) w/ attached Metadata Target (MDT) (if consider scalability & load balance)
    - multiple “up to ~O(3)” Object Storage Server (OSS) w/ attached Object Storage Targets (OST)

Lustre Cluster I/O

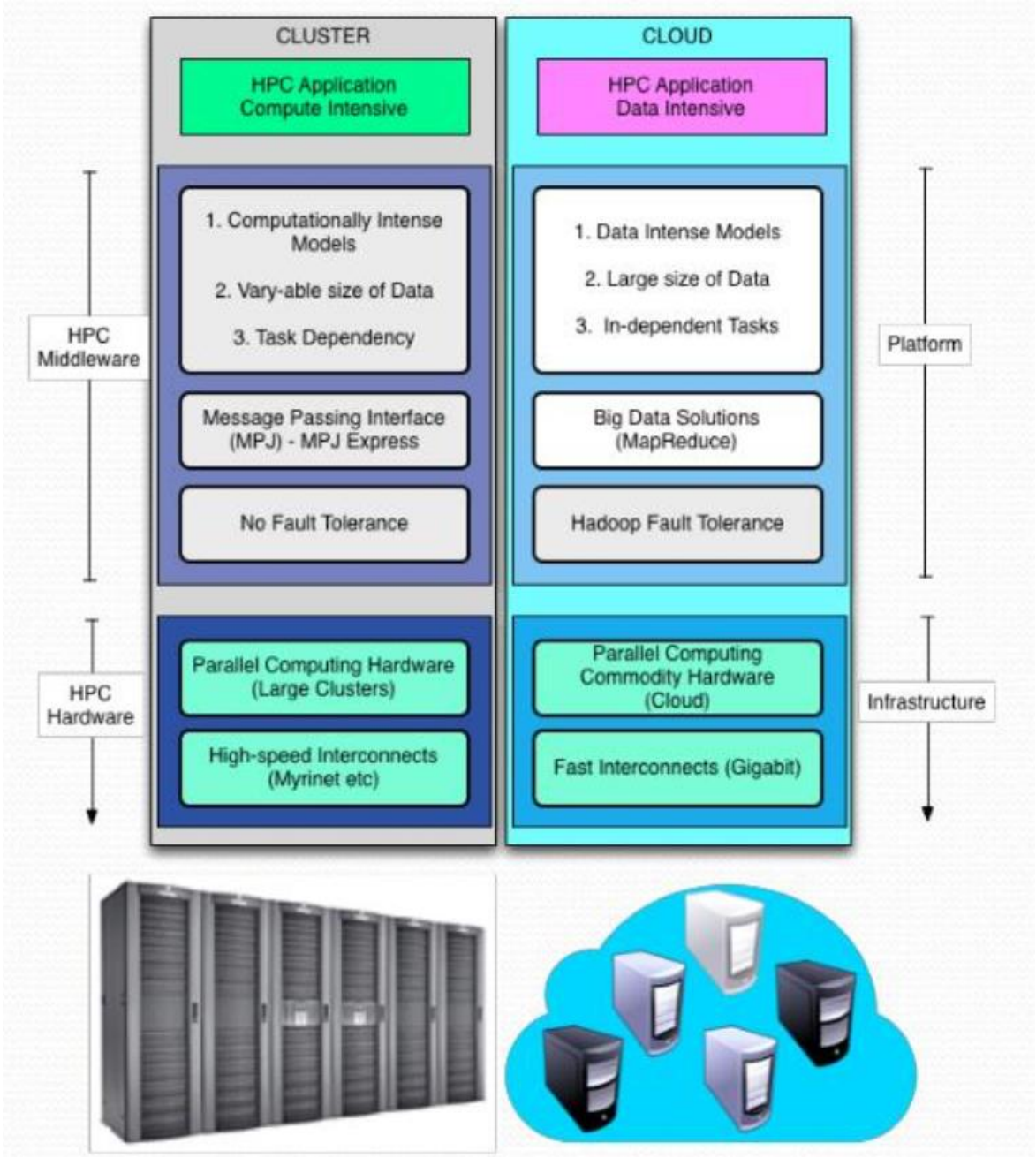


Cluster I/O: Parallel Filesystem using Lustre

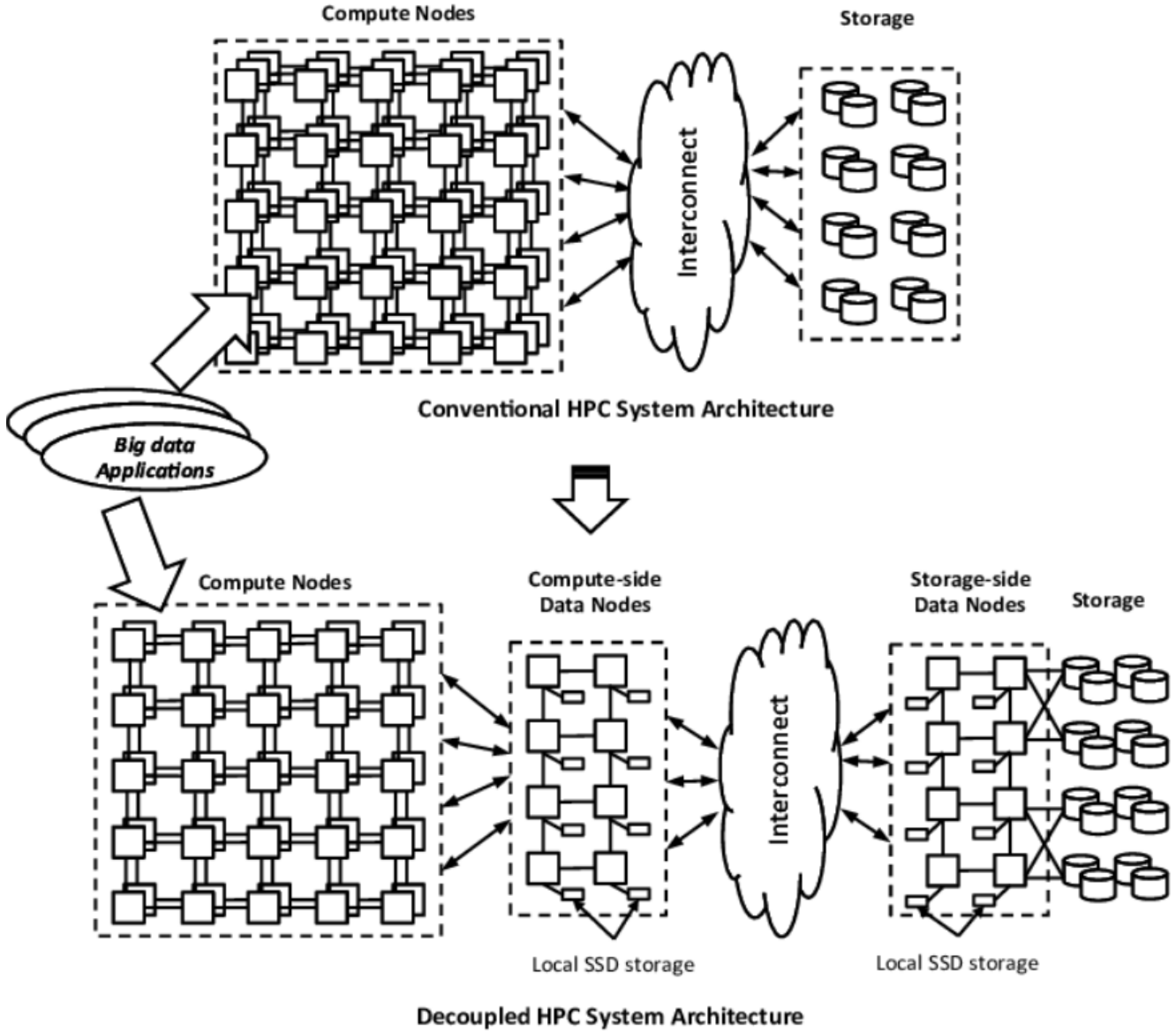
- **Typical Setup**
  - **MDS: ~ O(1) servers with good CPU and RAM, high seek rate**
  - **OSS: ~ O(3) server req. good bus bandwidth, storage**



Motivation for HPC over Cloud



Decoupled High Performance Computing System Architecture



ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

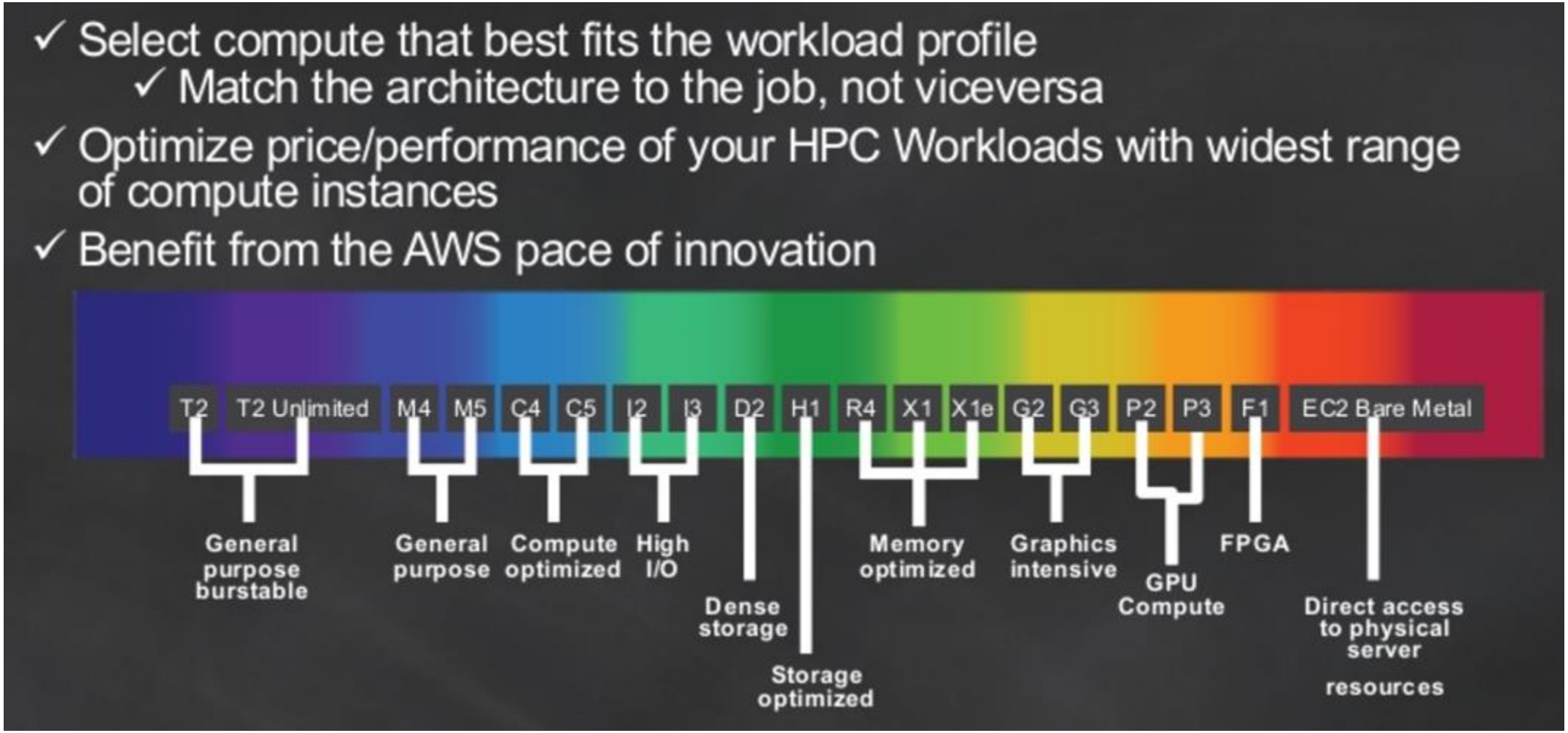
HPC on AWS

https://aws.amazon.com/hpc/

High Performance Computing on AWS

AWS Advantages for Multiple HPC Workload Types

Amazon EC2 Instances



Cost Advantages

AWS Networks

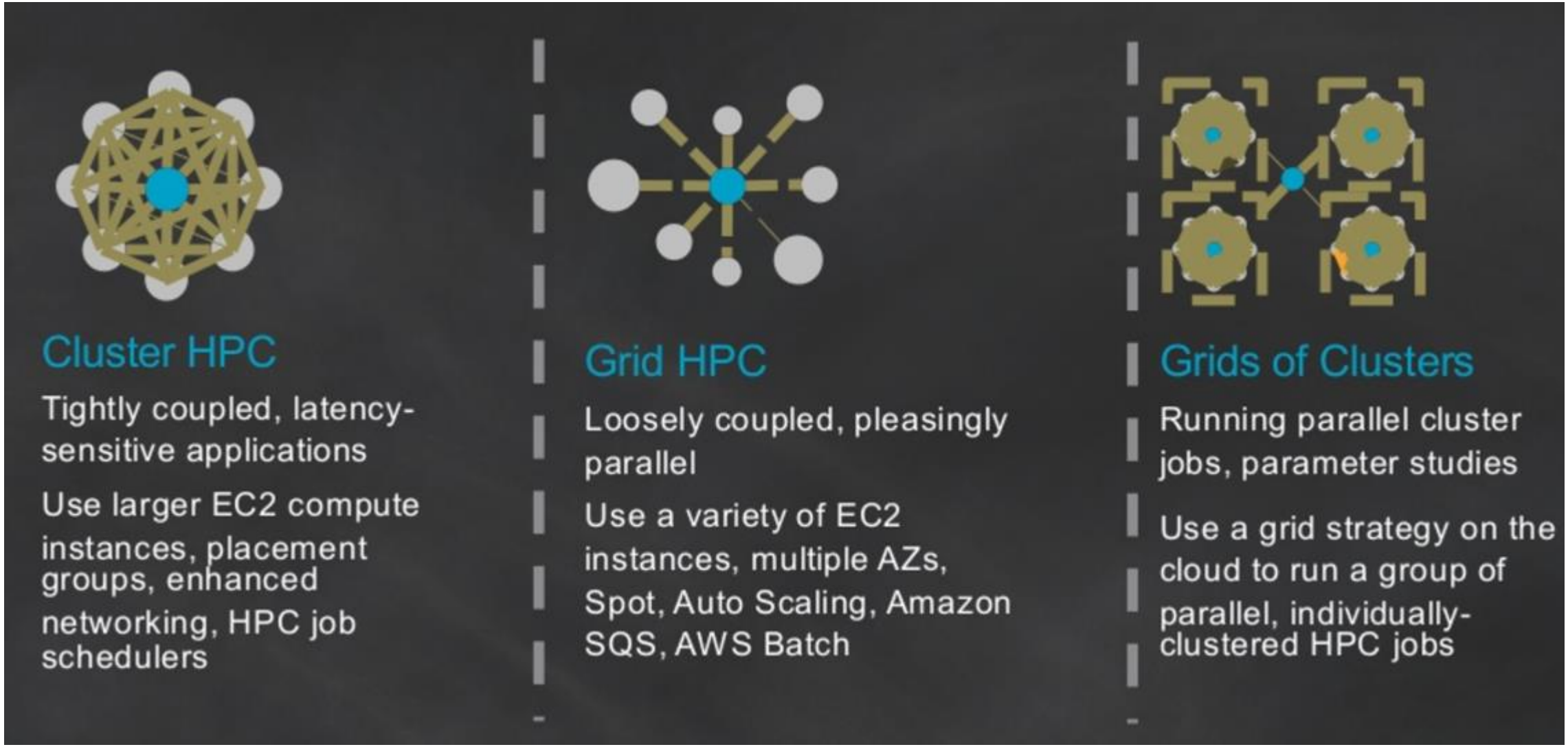
- AWS proprietary networking
  - Full bi-section bandwidth in placement groups
- Elastic Network Adapter (ENA)
  - Supports network speeds of up to 25 Gbps in placement groups
  - Multi-queue support
- VPC (Network segregation)
- Direct Connect (1/10GigE)

Important enablers for HPC on Cloud

- Compute performance – CPUs, GPUs, FPGAs
- Memory performance – high RAM requirements in many applications
- Network performance – throughput, latency, and consistency
- Storage performance – including shared filesystems
- Automation and cluster/job management
- Remote graphics for interactive applications

…and SCALE

Cluster and Grid HPC in the Cloud



HPC in Design and Manufacturing

Applications for engineering:

- Molecular dynamics, CAD, CAE, EDA
- Collaboration tools for engineering
- Big data for manufacturing

HGST
a Western Digital brand

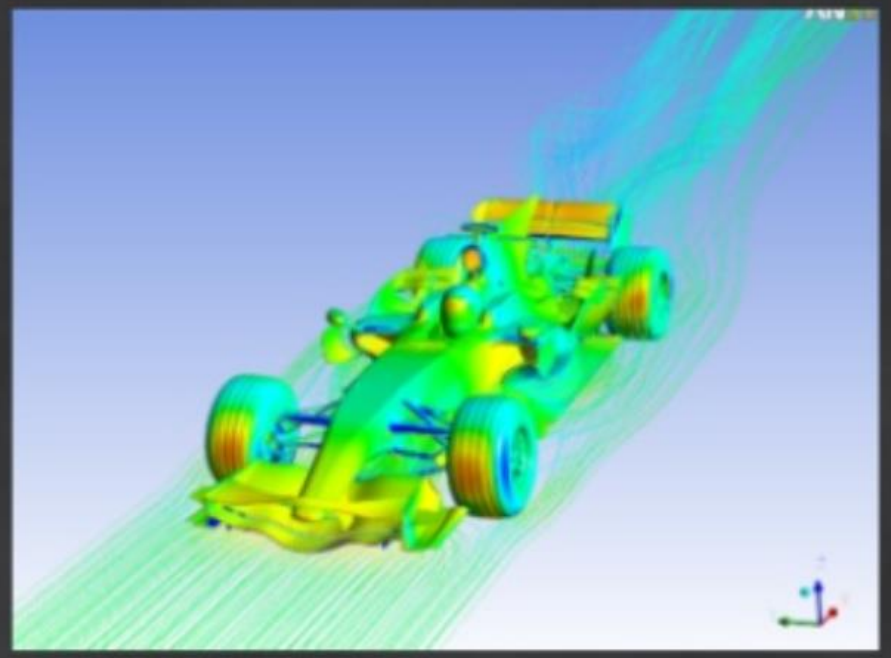
Running drive-head simulations at scale

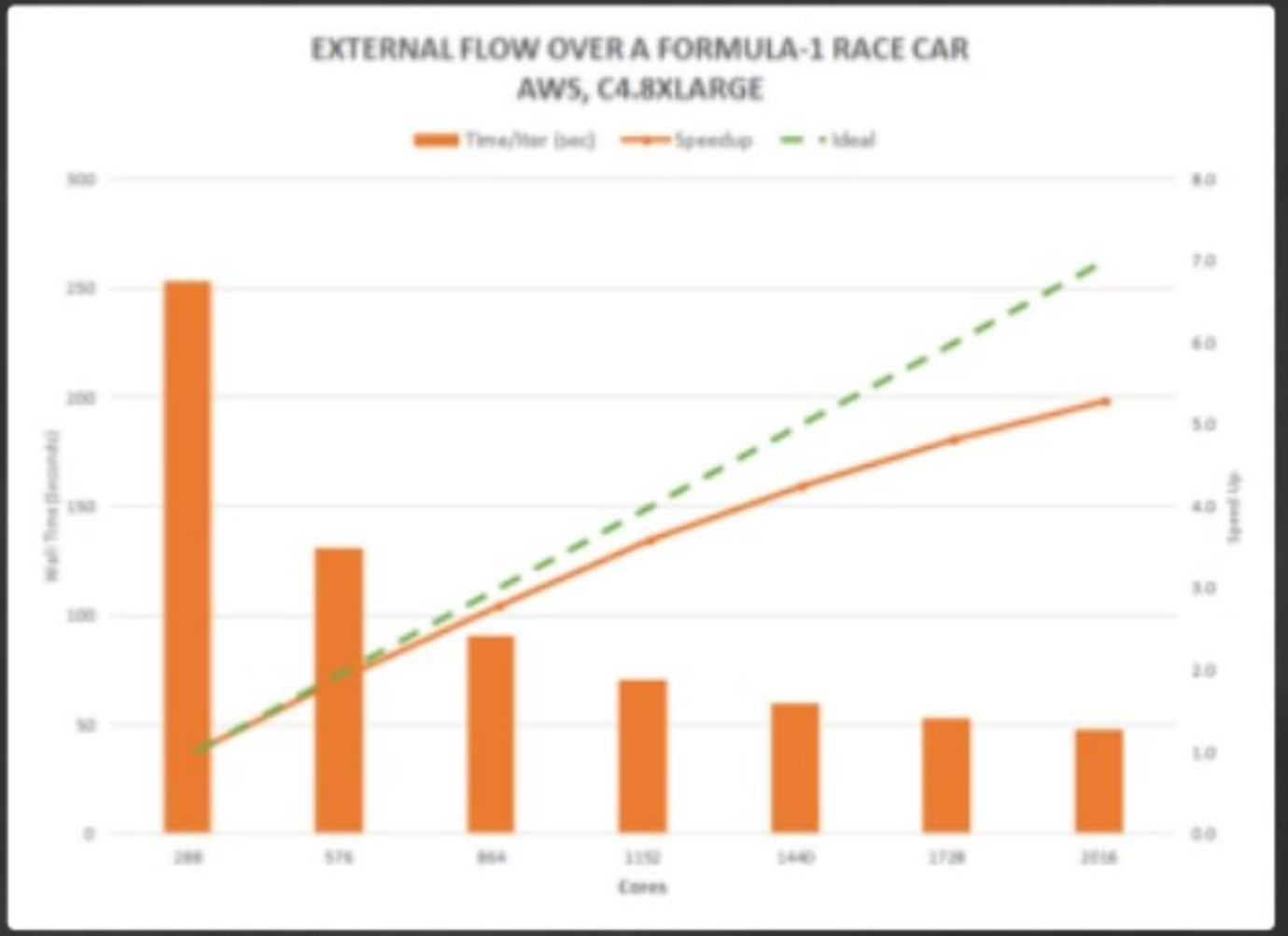
Millions of parallel parameter sweeps, running months of simulations in just hours

Over 85,000 Intel cores running at peak, using Spot Instance

Fluid Dynamics – Ansys Fluent

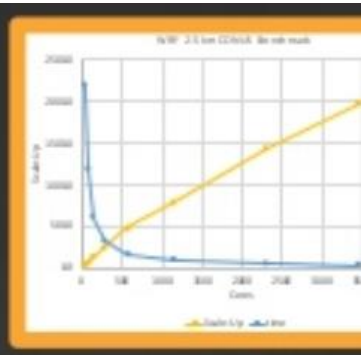
- C4.8xlarge instance type
- 140M cell model
- F1 car CFD benchmark





Performance considerations

For tightly-coupled cluster workloads



Network

- Use a placement group
- Enable enhanced networking

MPI libraries

- Test with Intel MPI and OpenMPI 3.0, and make use of available tunings

Domain decomposition

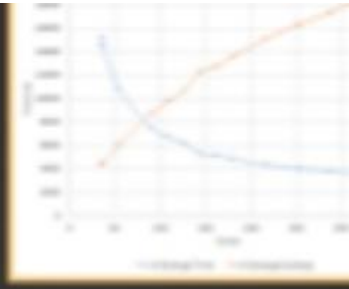
- Choose number of cells per core for either pre-core efficiency or for faster results

Test using real-world examples

- Use large cases for testing: do not benchmark scalability using only small examples

Performance considerations

For all HPC workloads



OS version

- Use Amazon Linux or an updated 3.10+ kernel – 4.0+ if using NVME on F1 or I3

Instance types

- C5, C4, M5, R4 are the best choices today – but always test with the latest EC2 instances

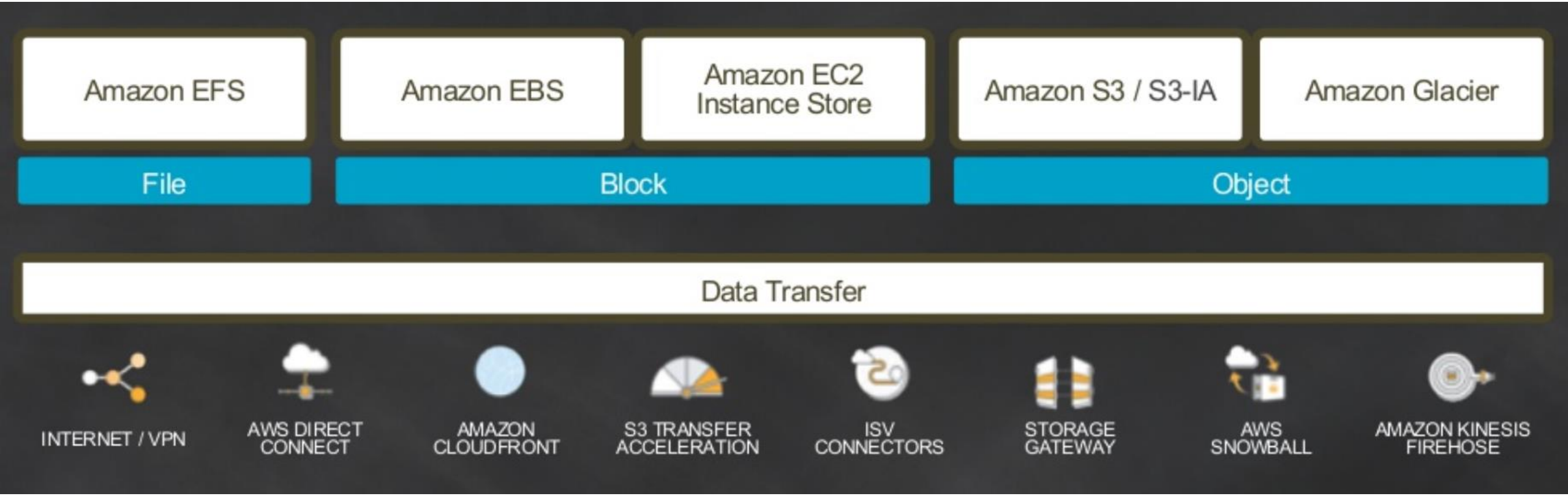
Processor states

- Use P-states to reduce processor variability

Hyper-threading and affinity

- Test with Hyper-threading (HT) on and off – usually off is best, but not always
- Use CPU affinity to pin threads to CPU cores when HT is off

AWS Storage is a platform



Optimize AWS HPC Storage

EFS

Highly available, multi-AZ, fully managed network-attached elastic file system.

For near-line, highly-available storage of files in a traditional NFS format (NFSv4).

Use for read-often, temporary working storage

EBS + EC2

Create a single-AZ shared file system using EC2 and EBS, with third-party or open source software (ZFS, Weka.io, Intel Lustre, etc).

For near-line storage of files optimized for high IOPS.

Use for high-IOPS, temporary working storage

Amazon S3

Secure, durable, highly-scalable object storage. Fast access, low cost.

For long-term durable storage of data, in a readily accessible get/put access format.

Primary durable and scalable storage for critical data

Amazon Glacier

Secure, durable, long term, highly cost-effective object storage.

For long-term storage and archival of data that is infrequently accessed.

Use for long-term, lower-cost archival of critical data

Deploy multiple HPC clusters

Running at the same time, and tuned for each workload

HPC automation with CfnCluster

- **CfnCluster** simplifies deployment of HPC in the cloud, including integrating with popular HPC schedulers
- Built on AWS CloudFormation, easy to modify to meet specific application or project requirements

AWS Batch for HPC workloads

Fully Managed

No software to install or servers to manage

AWS Batch provisions, manages and scales your infrastructure

Integrated with AWS

Natively integrated with the AWS platform

AWS Batch jobs can easily and security interact with services such as Amazon S3, DynamoDB, and Rekognition

Cost-optimized Resource Provisioning

AWS Batch automatically provisions compute resources tailored to the needs of your jobs using EC2 and EC2 Spot

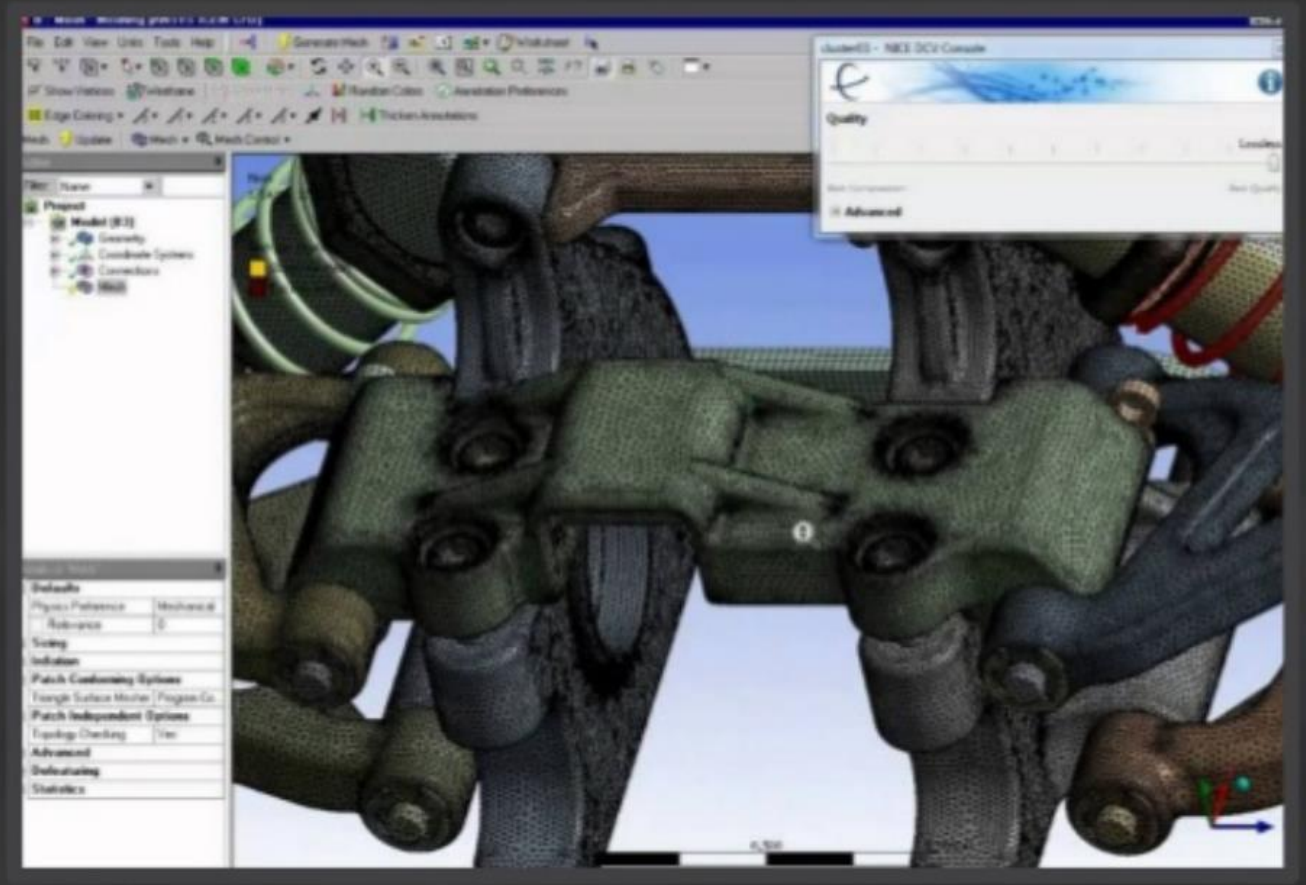
Graphics and Collaboration with DCV and AppStream

Cloud can be used for pre-and post processing as well as HPC

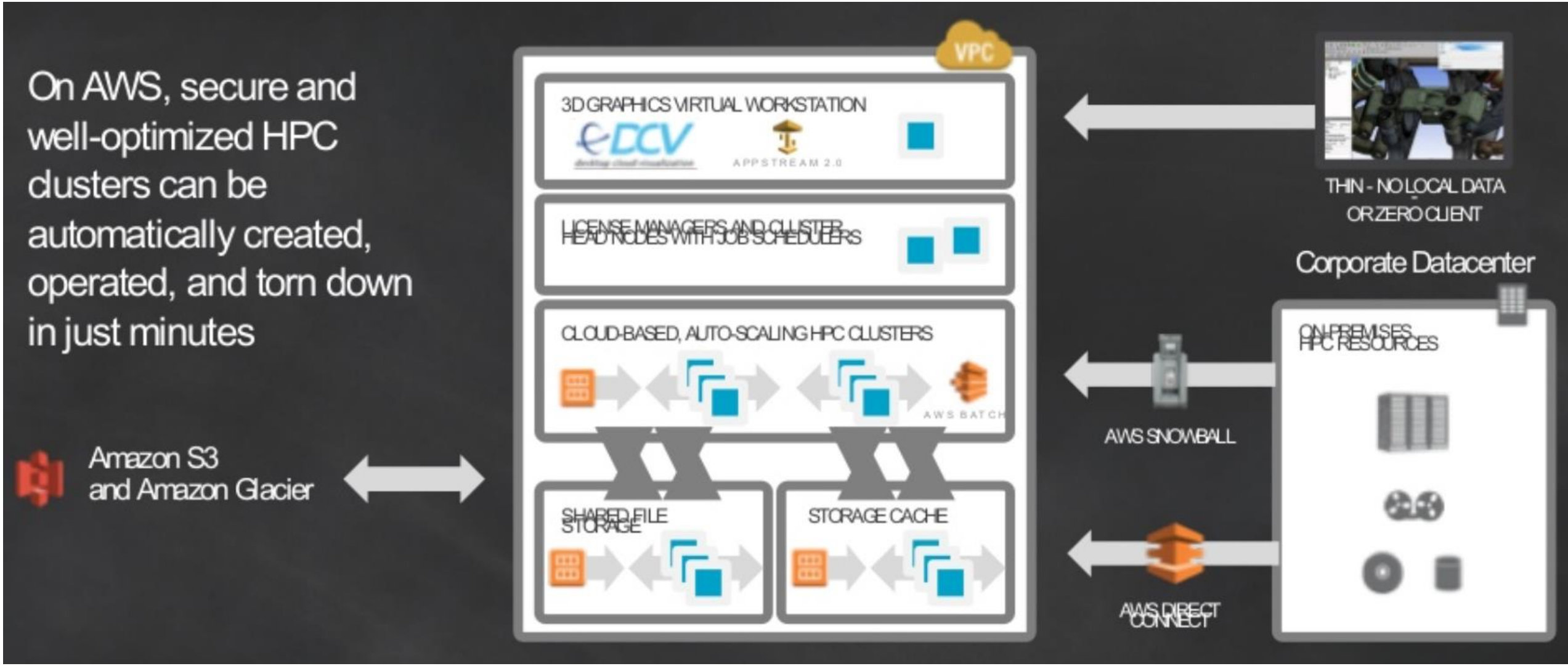
- Use GPUs in the cloud for remote rendering and remote desktops

Cloud is more secure for collaboration

- Encrypt the data in flight and at rest
- Manage your own keys and credentials
- Deliver pixels to your collaborators, not the actual data



Deploying HPC on AWS



Financial modeling has grown more onerous

Diverse risk analysis models

- Market risk
- Credit risk
- Liquidity risk

Broad regulatory requirements

- Comprehensive Capital Analysis and Review (Banking/Dodd Frank)
- Solvency Capital Requirements (Insurance/Basel II)
- Fundamental Review of the Trading Book (Insurance/Basel III)

Compute-intensive calculations

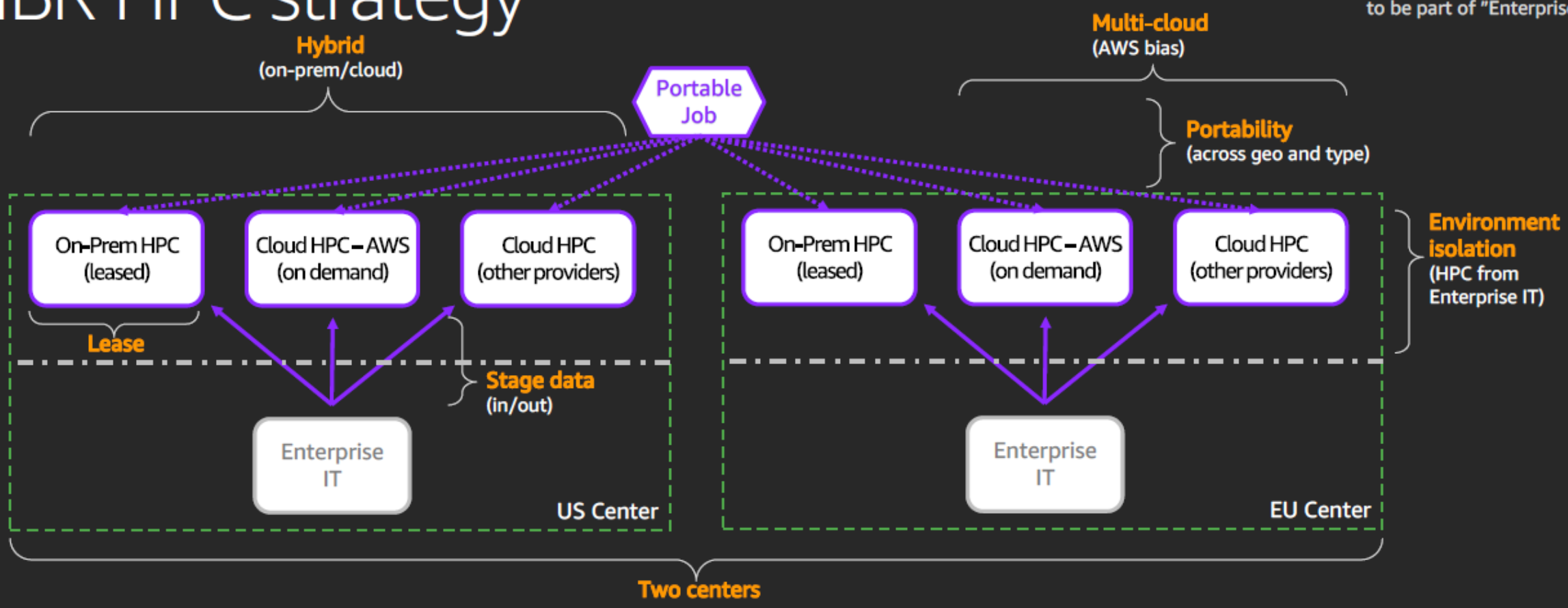
- More granular risk factors
- Wider range of scenarios
- More historical data

Large amounts of compute resources needed to run simulations

ie
SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

NIBR HPC strategy

- Strategic principles annotated with **this style**
- Instruments not shown, assumed to be part of "Enterprise IT"



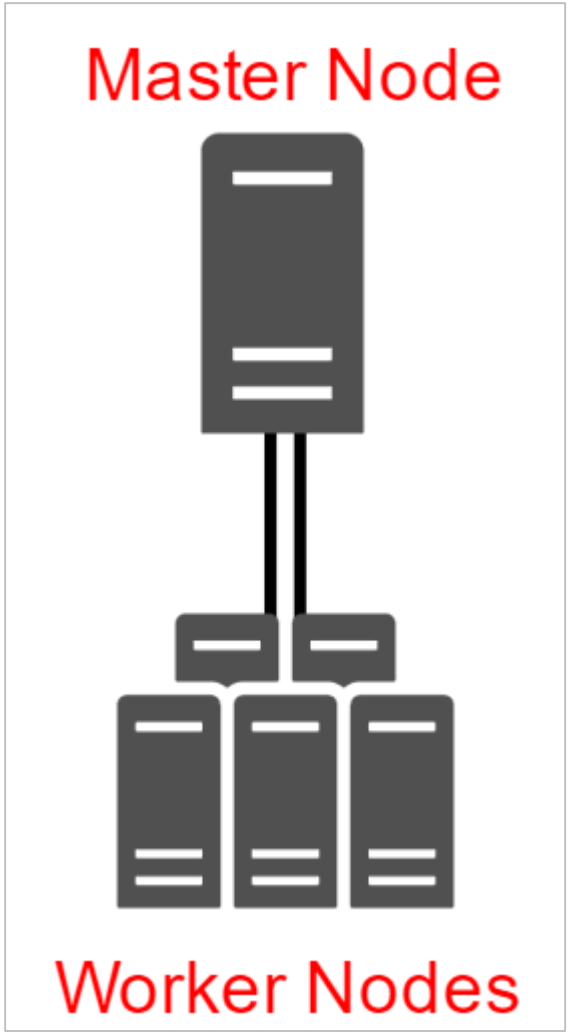
ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

HPC on Azure

https://azure.microsoft.com/en-us/solutions/high-performance-computing/#applications

HPC Clusters

- HPC typically involves **clusters** of computers interconnected by a high-speed network
- A single computer in the cluster is called a **node**
- **Workloads** are managed by **master** nodes that distribute workloads across **worker** nodes



ie
SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

HPC in Azure

- Run massively parallel compute jobs in the cloud
  - Photorealistic 3D rendering
  - Brute force cryptographical analysis
  - Financial risk modeling, genomics research, and more
- Deploy an HPC cluster in minutes and scale as needed
- Automate deployments with deployment templates
- Combine with Azure Batch for batch scheduling and compute management (http://bit.ly/a4r-batch)
- Linux or Windows

Virtual-Machine Sizes

A-Series

Up to 8 cores and 56 GB RAM

A0	A1	A2	A3
A4	A5	A6	A7

Compute-intensive

A8	A9	A10	A11

D-Series

Also available in DS sizes

D1	D2	D3	D4
D11	D12	D13	D14

Also available in DS sizes

D1v2	D2v2	D3v2	D4v2
D11v2	D12v2	D13v2	D14v2

20 cores, 140 GB Ram

D15v2	DS15v2

F/G/H-Series

Also available in Fs sizes

F1	F2	F4	F8	F16

Also available in GS sizes

G1	G2	G3	G4	G5

Molecular modeling etc.

H8	H16	H8m	H16m	
H16r		H16mr		

N-Series

NVIDIA M60 x 1/2/4

NV6	NV12	NV24

NVIDIA K80 x 1/2/4

NC6	NC12	NC24

** Currently in preview*

Choosing a VM Size

CPU core = Memory	A0 - A7	D1v2 - D5v2	D1 - D4
CPU core > Memory	F1, F2, F4, F8, F16		
CPU core < Memory	D11v2 - D15v2	D11 - D14	G
GPU	N		
CPU core++	A8 - A11	G(S)5	D(S)15v2
Memory++	G(S)4, G(S)5	D(s)15v2	
Networking++	A10 - A11		

Power vs. Cost

A8	D1	G5
8 cores 56 GB RAM 382 GB SSD drives 32 Gbit/sec InfiniBand RDMA	1 core 3.5 GB RAM 50 GB SSD drives	32 cores 448 GB RAM 6,144 GB SSD drives Latest Xeon E5 v3 processors
$ 1.47/hr. or $1,091/mo.	$0.14/hr. or $104/mo.	$ 9.65/hr. or $7,180/mo.
$0.98/hr. or $725/mo.	$0.077/hr. or $57/mo.	$ 8.69/hr. or $6,465/mo.

See bit.ly/a4r-vm-pricing for up-to-date pricing information

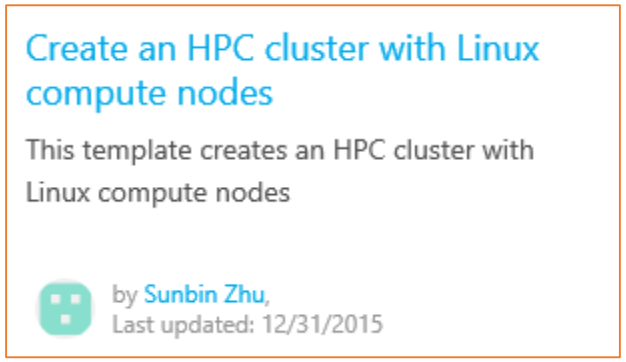
ie
SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

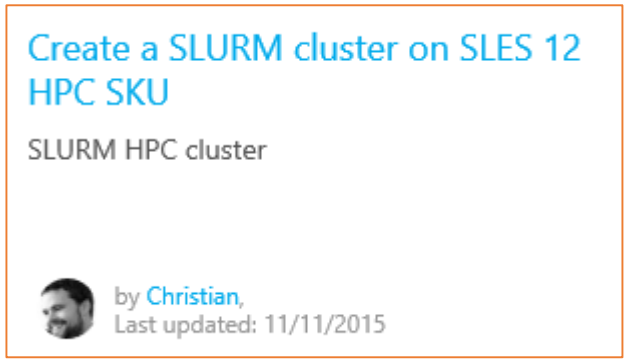
Azure Resource Manager

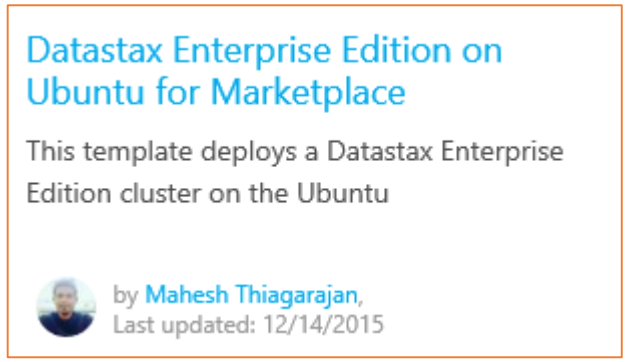
- Allows resources to be collated into resource groups
  - Deploy, manage, monitor, and delete all resources at once rather than one resource at a time
- Allows complex deployments to be performed declaratively via deployment templates
  - Deployment templates specify all the resources — VMs, switches, storage accounts, etc. — to be provisioned using JSON syntax
  - Templates can include parameters that are filled in at runtime
  - Learn more at http://bit.ly/a4r-arm

Azure Quickstart Templates

- Free, open-source deployment templates



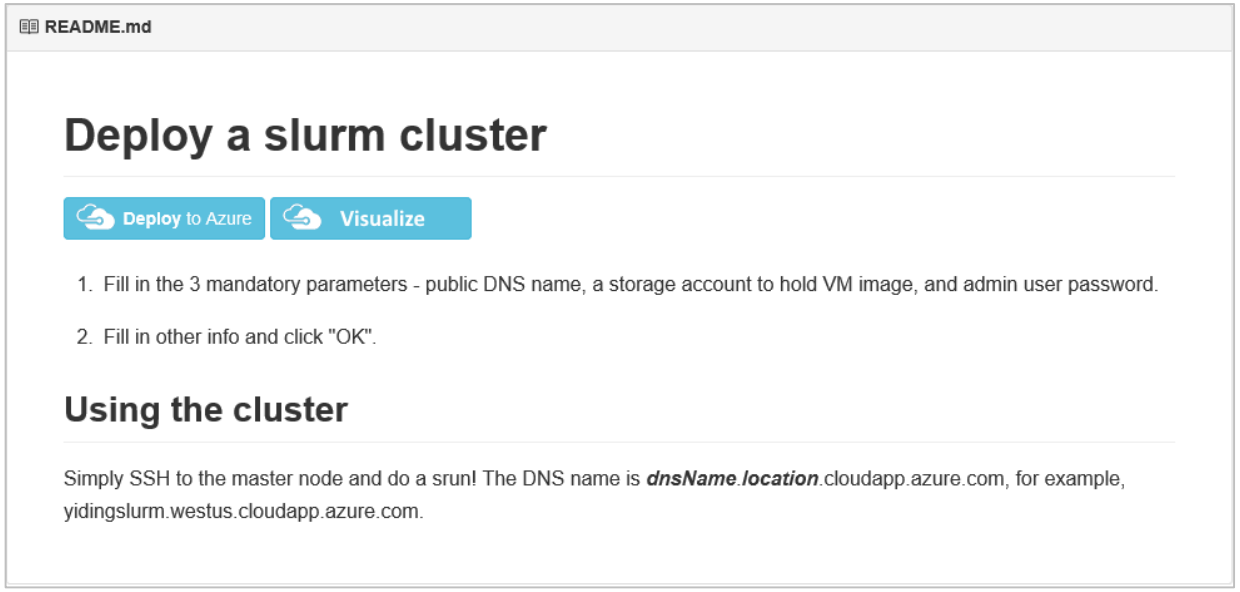




- Find them on the Azure site (http://bit.ly/a4r-quickstart)
- Or browse them on GitHub (http://bit.ly/a4r-github)

SLURM Clusters

- Simple Linux Utility for Resource Management (SLURM)
- Quickstart template at http://bit.ly/a4r-slurm enables easy deployment of SLURM clusters of user-specified sizes



README.md

Deploy a slurm cluster

Deploy to Azure | Visualize

1. Fill in the 3 mandatory parameters - public DNS name, a storage account to hold VM image, and admin user password.
2. Fill in other info and click "OK".

Using the cluster

Simply SSH to the master node and do a srun! The DNS name is ***dnsName.location***.cloudapp.azure.com, for example, yidingslurm.westus.cloudapp.azure.com.

Other Azure resources

- https://www.youtube.com/user/OfficeGarageSeries/search?query=HPC

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

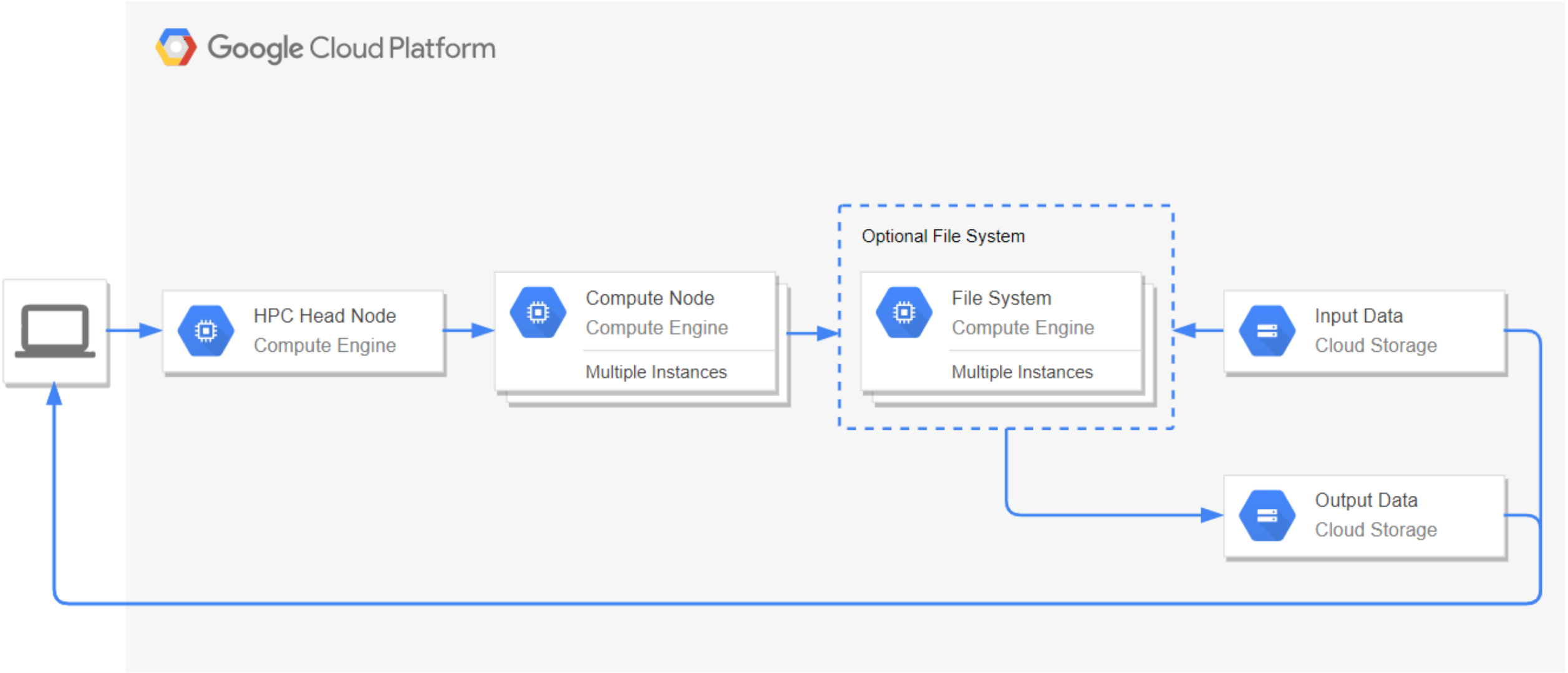
See Azure Presentation

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

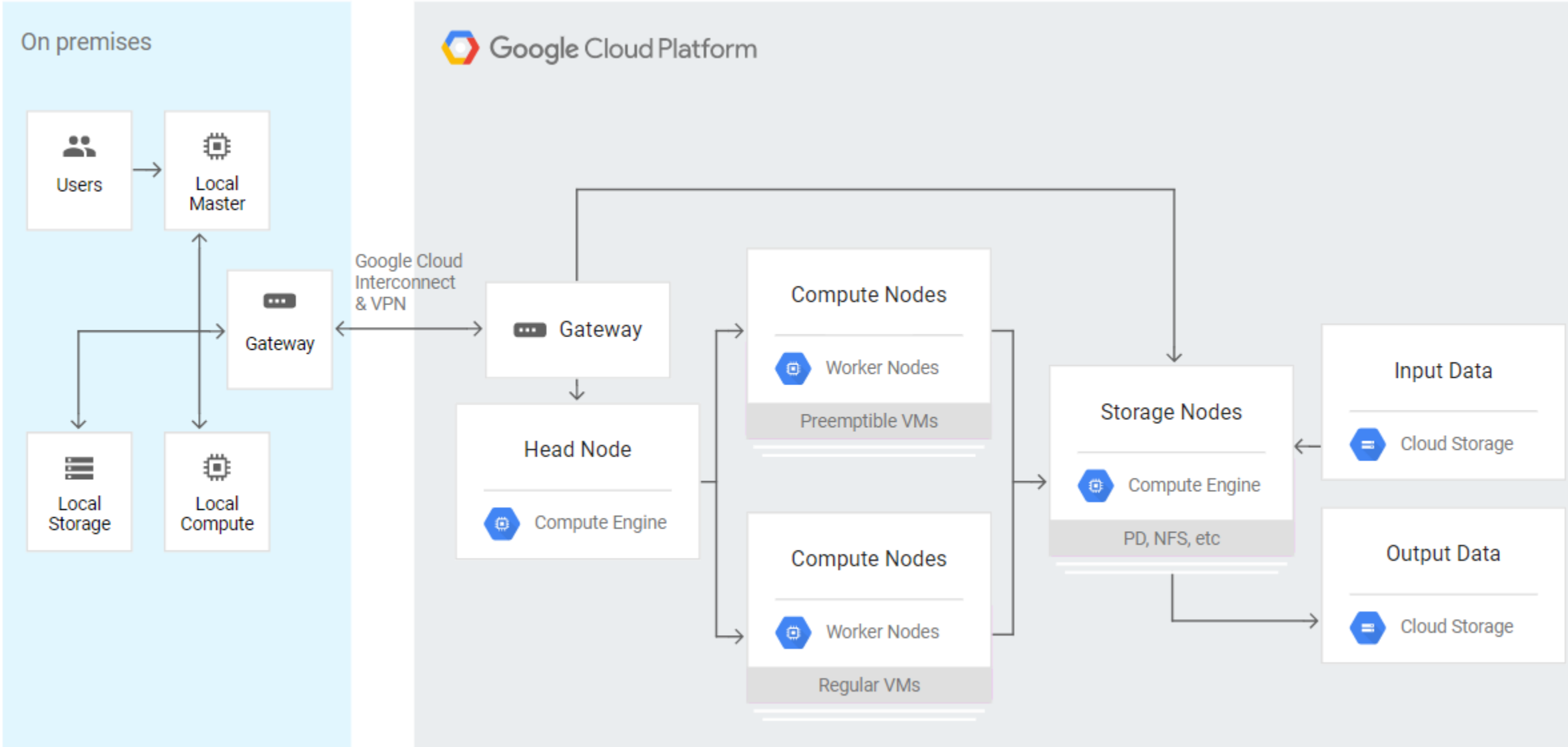
HPC on Google Cloud

https://cloud.google.com/solutions/hpc

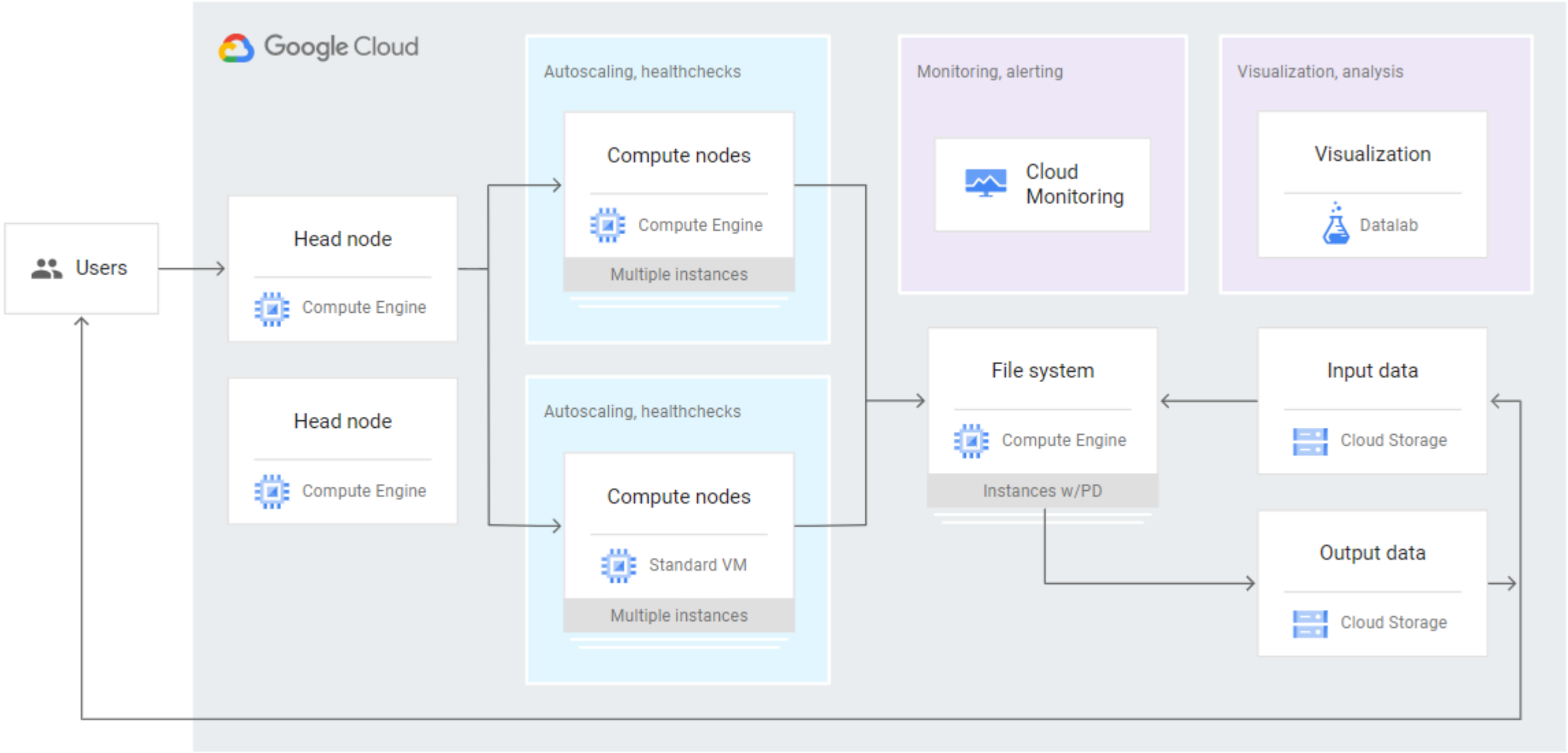
HPC on Google Cloud



HPC Comparison on Google Cloud



HPC Scaling on Google Cloud



Other resources

- https://www.schedmd.com/

- https://codelabs.developers.google.com/codelabs/hpc-slurm-federated-on-gcp/?_ga=2.58697710.534408772.1585177692-1203274973.1579591965&_gac=1.195443934.1584603641.Cj0KCQjwjcfzBRCHARIsAO-1_OroJDaxMbSKTV3UyMeJ-WlywpDqgAiTgBXNAGriwZOROelTmIvvPm4aAuylEALw_wcB#0

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

HPC on Oracle Cloud

https://www.oracle.com/es/cloud/solutions/hpc.html

OCI Hardware

Bare Metal Standard
52 Cores, 768 GB RAM,
up to 1 PB Block Storage

Bare Metal DenseIO
51.2 TB of local NVMe SSD
2x 25Gbe Network Interfaces

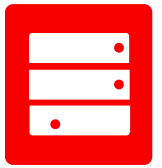
Bare Metal GPU
28 Cores, 192 GB RAM,
2x Tesla P100 GPUs
Pre-Configured Images

Bare Metal GPU V2
52 Cores, 768 GB RAM,
8x Tesla V100 GPUs
NVLINK Interconnect

File Storage Service
Managed distributed file service
POSIX, NFSv3 mount point

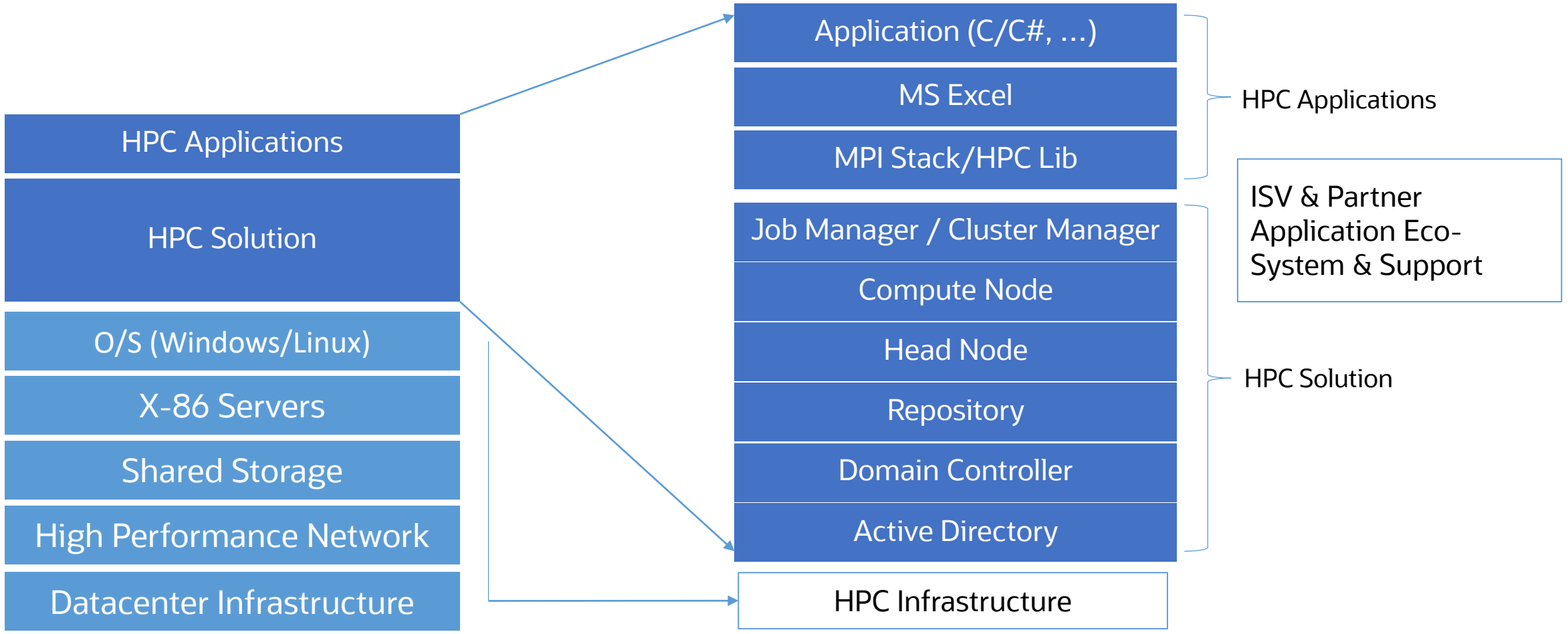
Block Storage
50 GB-2 TB volumes
Up to 400K IOPS per host

Bare-Metal HPC
36 cores, 3.7 GHz
384GB RAM
6.7 TB NVME, 1PB block
RDMA



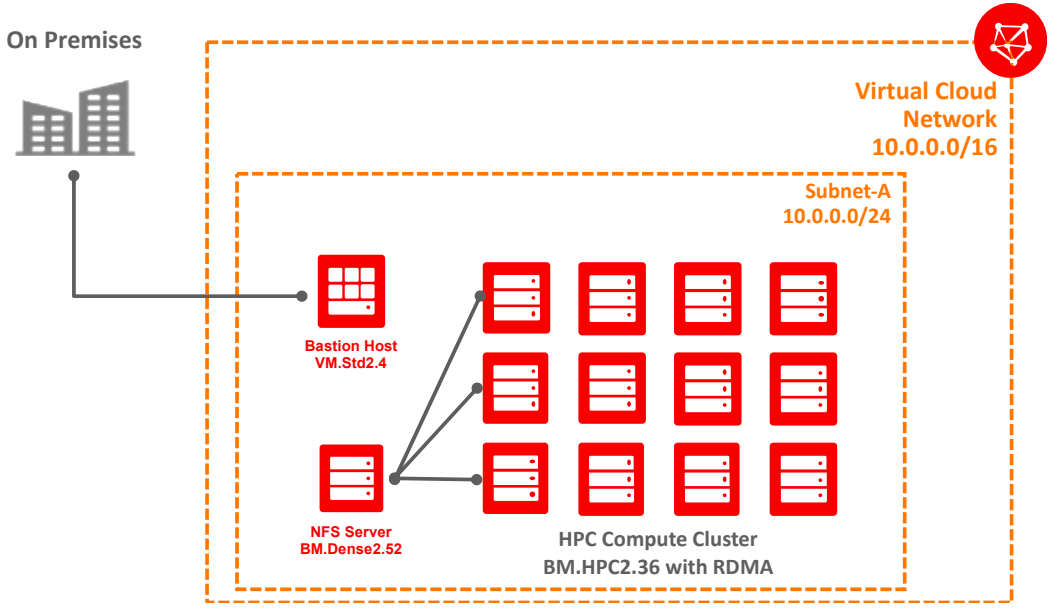
GPU Visualization
P100 GPU
NVIDIA Quadro Enabled
Teradici & Citrix Support

Typical HPC Technology Stack



RDMA Architecture

- RDMA
  - Bastion host IP will be provided
  - Each compute node will have access to external internet
  - Firewalls will be enabled, disable if your testing requires
  - DenseIO node will provide NFS share across all nodes
  - RDMA will be enabled and configured
  - FSS can be included if desired



Required capabilities for HPC Infrastructure

Capability	Consideration	Oracle offering
Instances	No Hypervisor overhead	✓ Bare-Metal Cloud Instances
CPU	Cutting edge features	✓ Intel Skylake Processor
GPU	3D visualization, computational fluid dynamics, deep learning	✓ NVIDIA p100GPU, NVIDIA Tesla Volta GPUs
Storage	High-throughput, low-latency	✓ up to 512TB of NVMe Block ✓ Managed Distributed File System Service
Network	Predictable Performance with low Latency and high throughput	✓ Truly Non-Over Subscribed Flat Networking, up to 2 x 25 Gbps Ethernet ✓ 100 Gbps, ulta-low latency RoCE

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

See Oracle HPC Presentation

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

HPC on IBM Cloud

https://www.ibm.com/es-es/cloud/hpc

IBM Cloud Machine

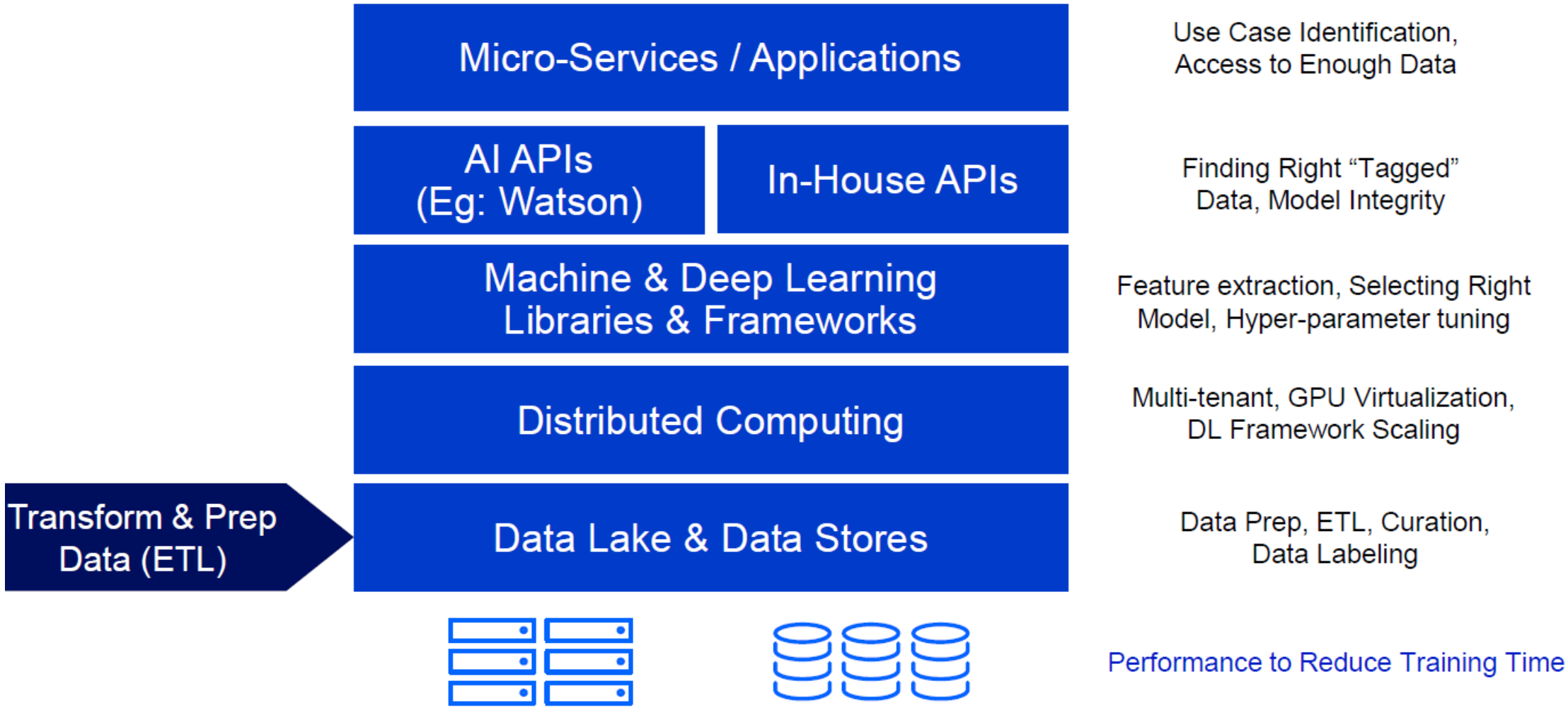
IBM Cloud Bare Metal Servers:

- High level of customization to fit any workload requirement
- Superior security
- No hypervisor overhead
- All resources dedicated to a single user; no "noisy neighbors"

IBM Cloud Virtual Servers:

- Low provisioning time; pre-configured virtual servers
- Flexible instance options per storage type, memory, and number of GPUs
- Efficient and scalable for burst workloads with hourly and monthly options

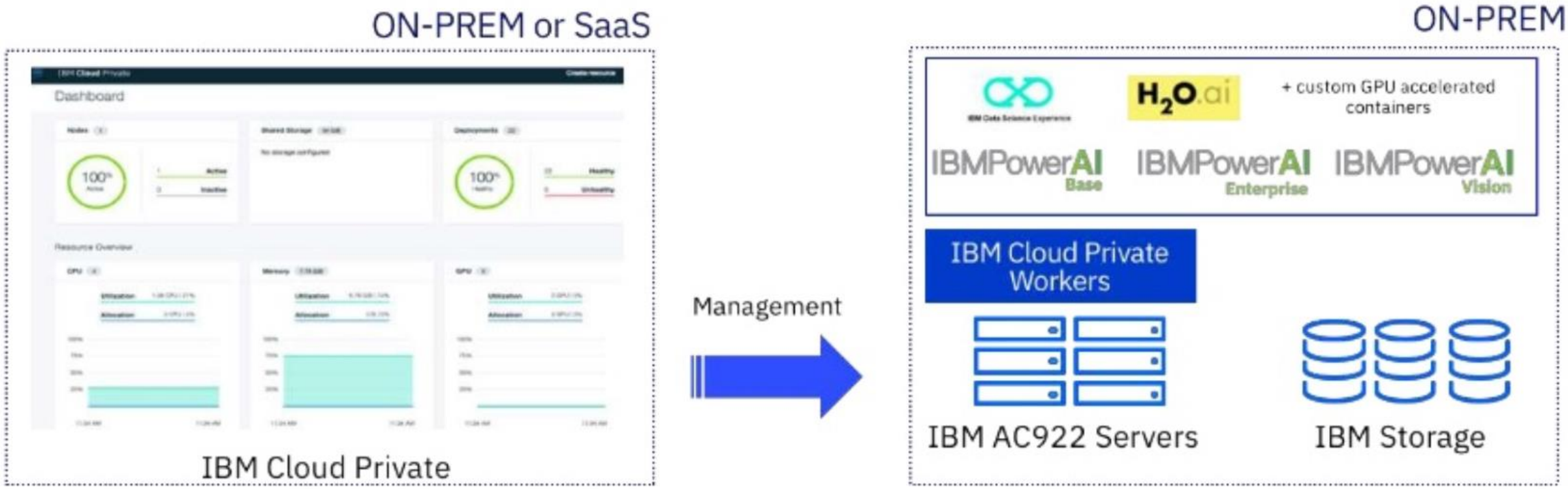
AI Infrastructure Stack Challenges



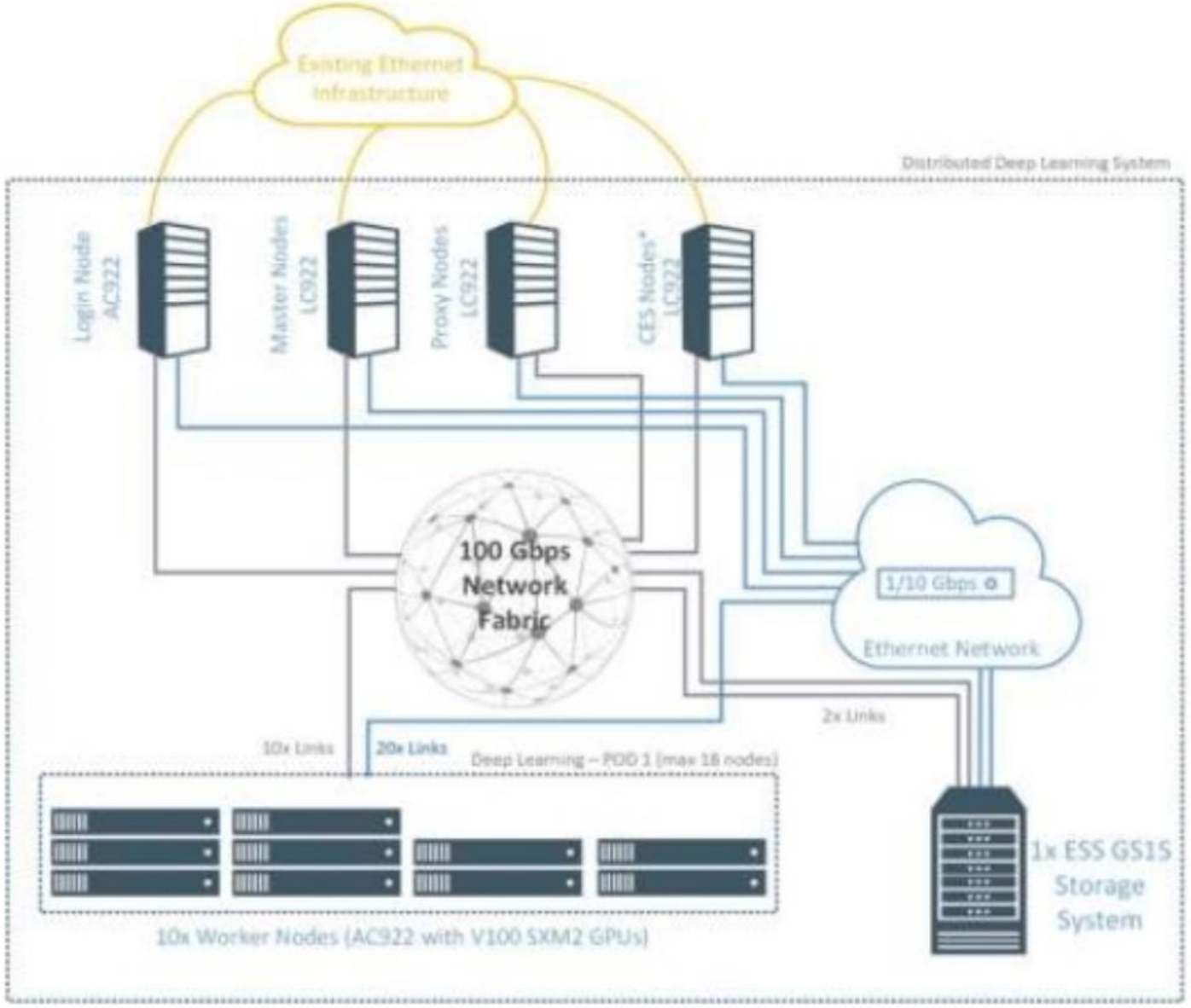
GPU Solutions

Selection	NVIDIA Tesla M60	NVIDIA Tesla K80	NVIDIA Tesla P100	NVIDIA Tesla V100
Ideal environment	Fundamental enterprise performance for virtualization and professional graphics	Reliable enterprise performance for introductory AI computing	Essential performance for growing advanced AI and HPC capabilities	Maximum performance for progressive deep learning workloads
Availability	Monthly bare metal servers	Monthly bare metal servers	Monthly and hourly bare metal servers Monthly and hourly virtual servers	Monthly bare metal servers Monthly and hourly virtual servers
Enabled data centers	North Amerca, Europe, Asia, Australia and South America data centers	North America, Europe, Asia, Australia and South America data centers	(Bare Metal) Dallas, TX San Jose, CA Washington D.C. Amsterdam Seoul Tokyo (Virtual) Dallas, TX Washington D.C.	Dallas, TX Washington D.C.

Enterprise AI



IBM AC92 Deep Learning Architecture



SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY