OCTOBER 2023

NETWORK DESIGN EXERCISE

Barbara Krašovec

The objective of the exercise

The objective is to think about a network design. We will take HPC infrastructure as an example.

Security and network design

- Integrating security early on prevents multiple security gaps later on.
- Always plan for growth.
- Maintain network documentation and update it, when the topology changes.
- Network redundancy - how much does it cost if services are down?
- Which servers are accessible to external users?
- Where is sensitive data?
- Plan remote access to your internal network.
- Do not segregate too much, the complexity of the system will also make debugging more complex.

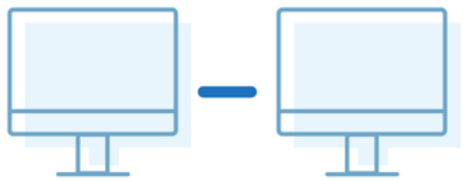
Traditional HPC network

- Traditionally HPC network’s topology is either star or tree
- vendors concentrate more on performance than on security,
- usually only two network fabrics: management network (administration) and servers network (compute),
- top of rack switches (leaf switches) connected to core switches (top level switches - TLS),
- compute network in usually on private IPs, local network .

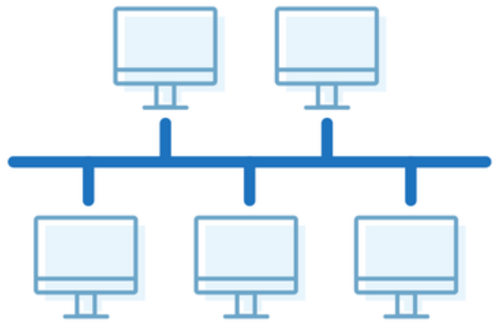
Network topology

Network Topology Types

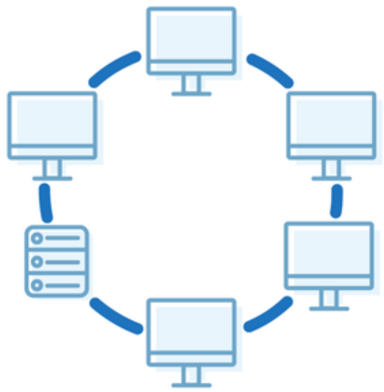
1 Point to point



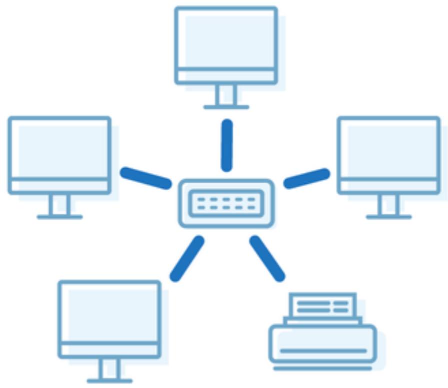
2 Bus



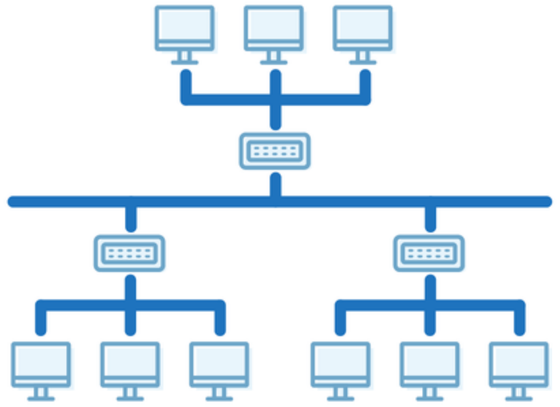
3 Ring



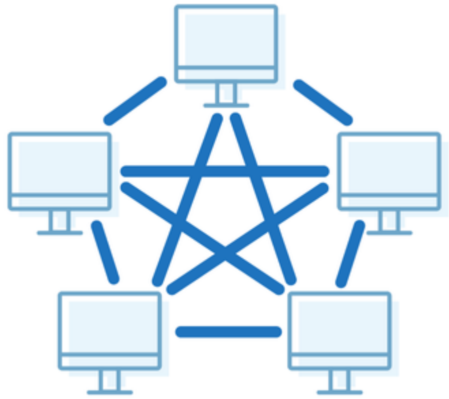
4 Star



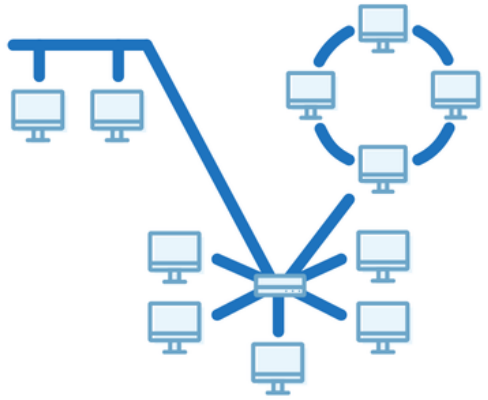
5 Tree



6 Mesh



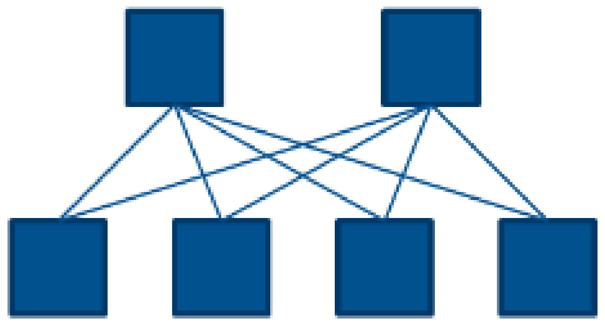
7 Hybrid



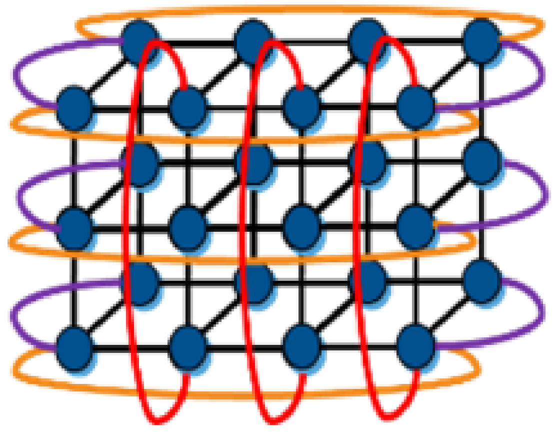
Cluster interconnect

- Multiple options: ethernet, infiniband, slingshot, omnipath etc.
- Low latency and high throughput required (OpenMPI requires less than 1 µs of latency)
- two dominant topologies: fat-tree and dragonfly (also torus, 2D/3D mesh, hypercube)

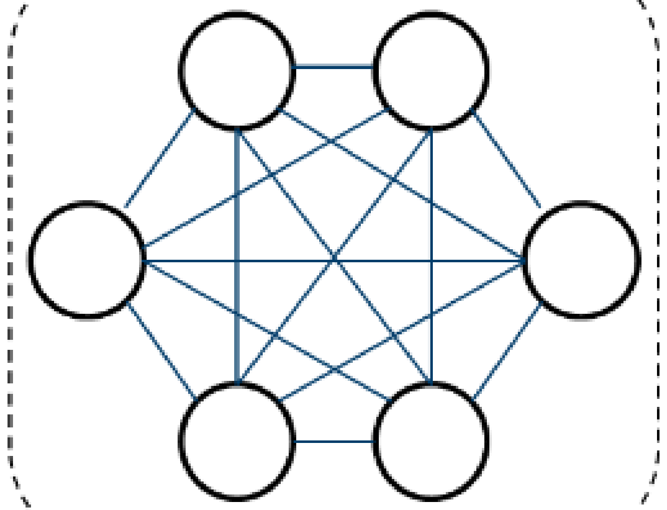
Interconnect topology



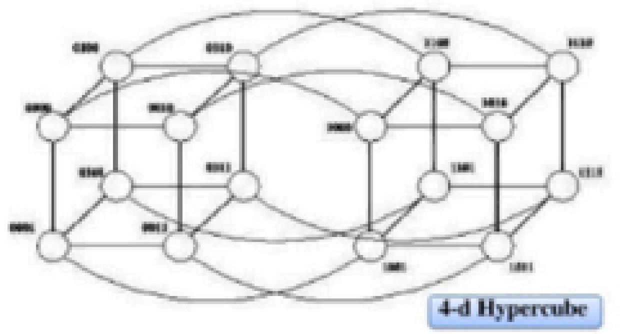
Fat Tree



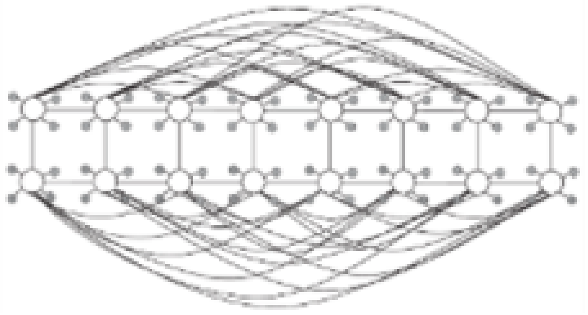
Torus



Dragonfly



Hypercube



HyperX

Complexity grows

- International collaboration and access to external storage systems requires changes.
- Federated access to resources requires changes.
- Where to put login nodes?
- Adding storage to the cluster - scratch storage vs permanent storage.
- How to transfer files? Access to the permanent storage via login node or directly? Authn/authz?
- Some users need VMs for their UIs (cloud), access to storage?

IPv4 and IPv6

- Problems of IPv6 network design? Some devices don’t support IPv6.
- DHCP IPv4 vs IPv6 auto configuration?
- Dual stack, public IPv6 only, private IPv4/IPv6?
- Network security: never forget that IPv6 is often enabled by default, even if you don’t configure it, network ACLs should be set

Exercise: our infrastructure

- **10 Login nodes** with a two-port Mellanox Connectx6 card (can be configured as ethernet or Infiniband)
- **2 management nodes** for central HPC services (central rsyslog, slurmctld, cluster shell, ipmi access)
- IPMI devices, monitoring tools, cooling management, power management services
- **30 S3 storage nodes** + disks (eg. Ceph)
- **30 storage nodes for permanent storage** + disks (eg. Ceph, EOS, dCache)
- **60 storage nodes for high-speed scratch** (eg. Ceph, Lustre, GPFS or other)
- **1000 compute nodes**
- **30 virtualization nodes** for cloud services
  - **VMs for backend services** (DNS, squid, websites, monitoring tools, configuration management etc)
  - **VMs for end users** (private cloud)

Prepare network design

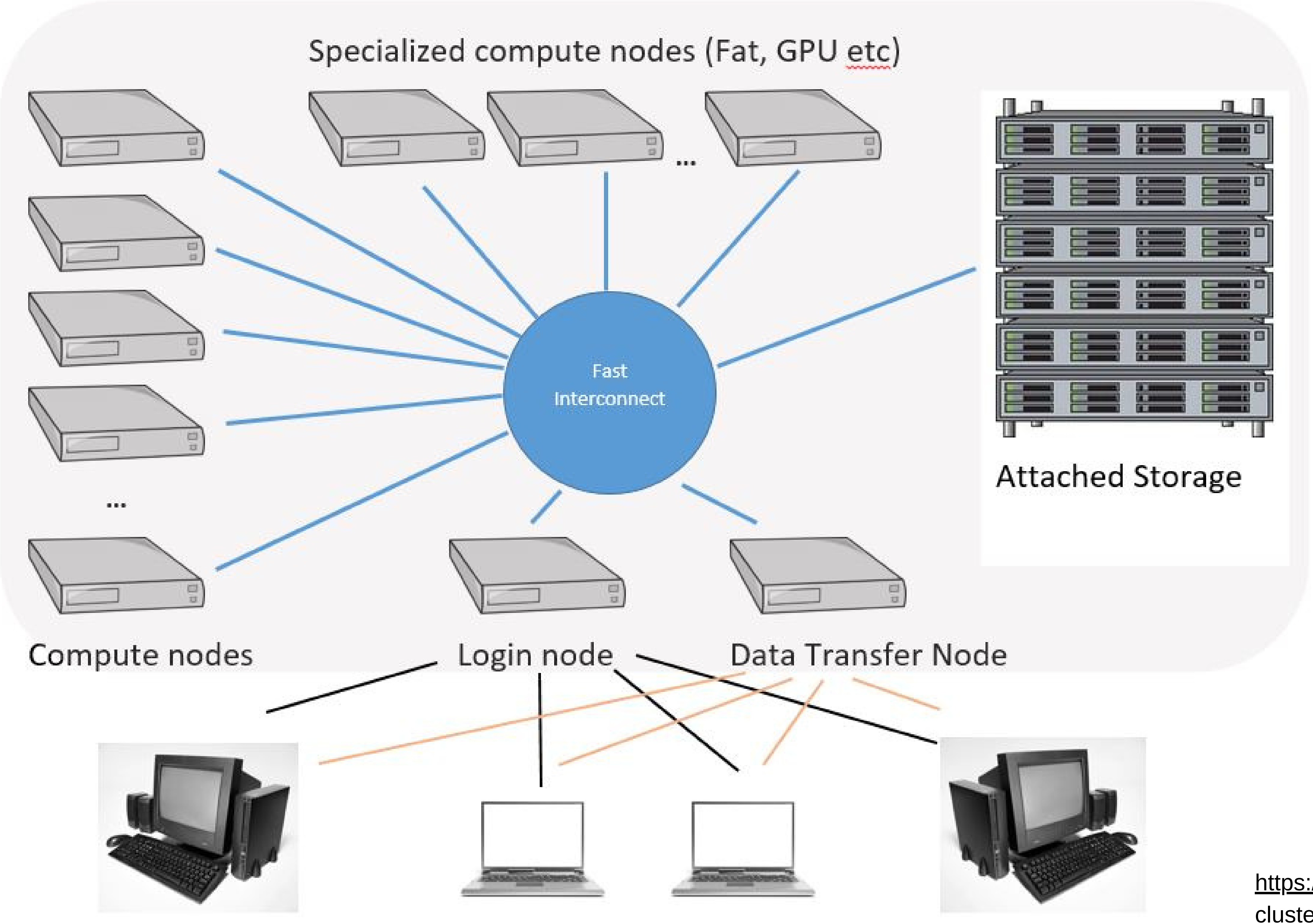
In groups of 3-4 discuss the possible network design:

- How many VLANs/subnets and argument why?
- How to provide access to users?
- Will you use NAT, if yes for which services and why?
- Which services will run on private, and which on public networks?
- Where would you, if even, put the firewall?
- Consider IPv6/IPv4 network, dual stack or exclusive?
- Estimate the number of public IPv4 addresses.
- How will the storage be accessed? Directly, via login node, another jump host?
- How will you segment remote networks?
- How to plan your cloud infrastructure? What are the considerations?

Examples

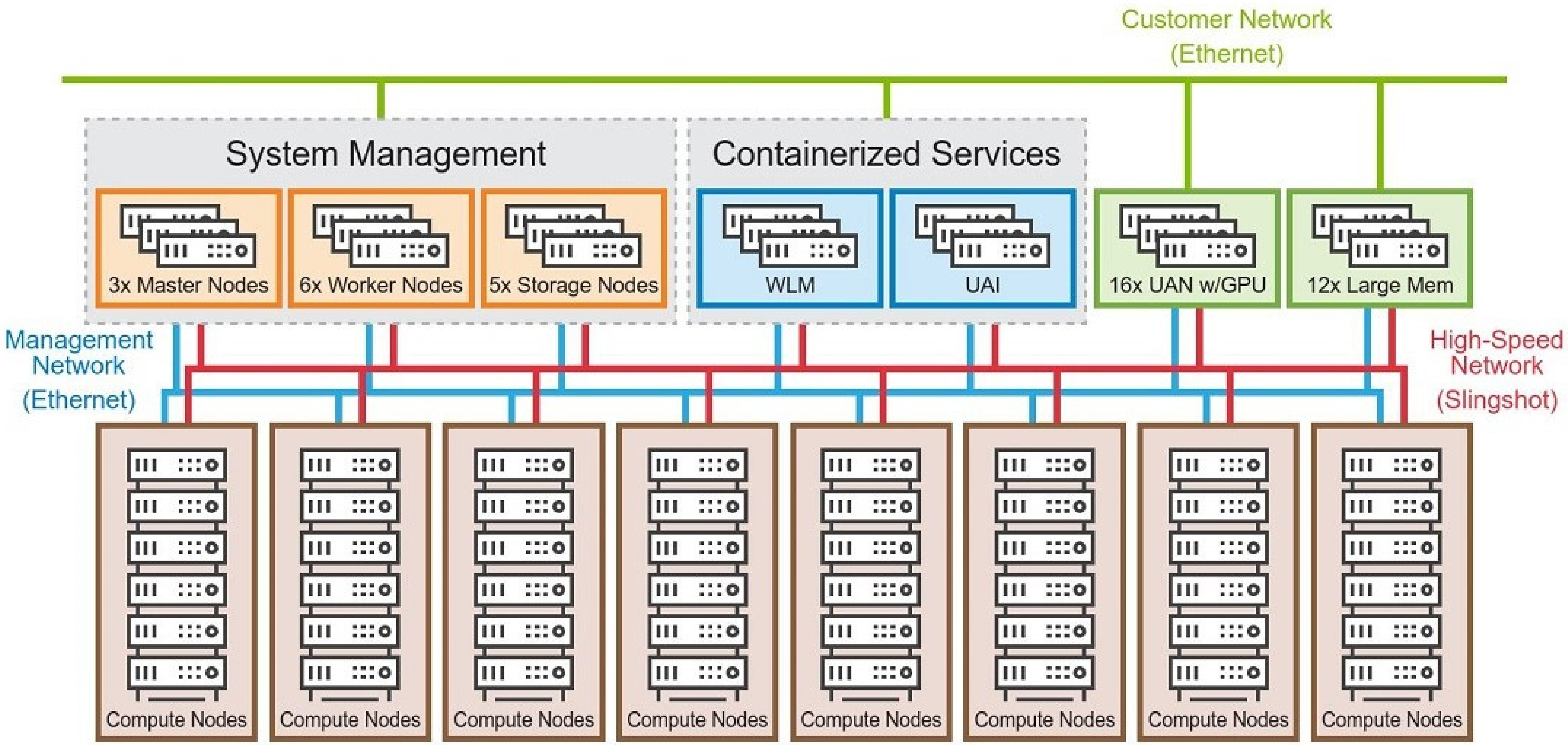
Take a look at some examples, comment on them and make your own design

Example 1



https://www.hpc.iastate.edu/guides/introduction-to-hpc-clusters/what-is-an-hpc-cluster

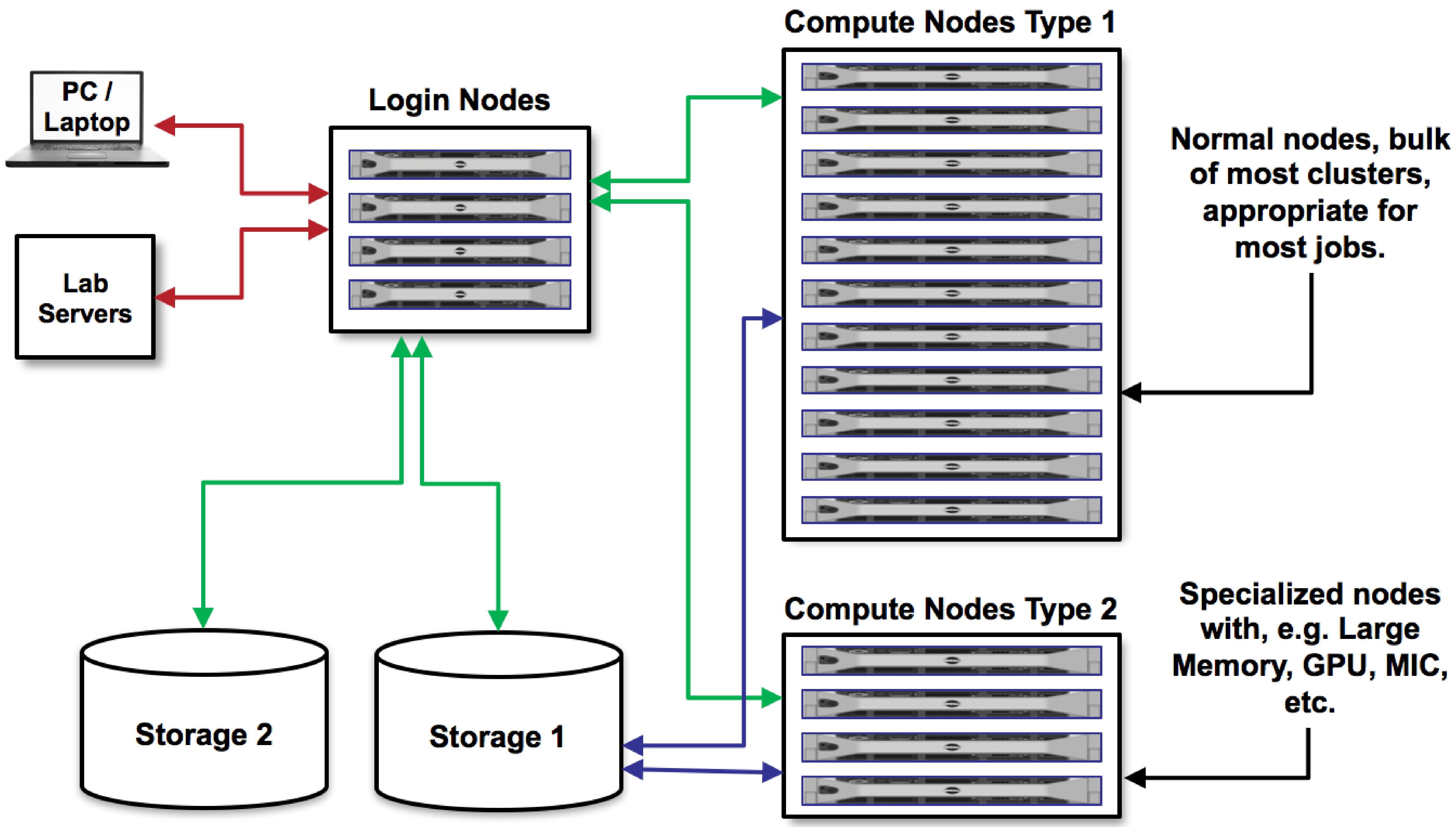
Example 2



HPC Architecture Connection Diagram

https://www.gigabyte.com/Article/high-performance-computing-cluster

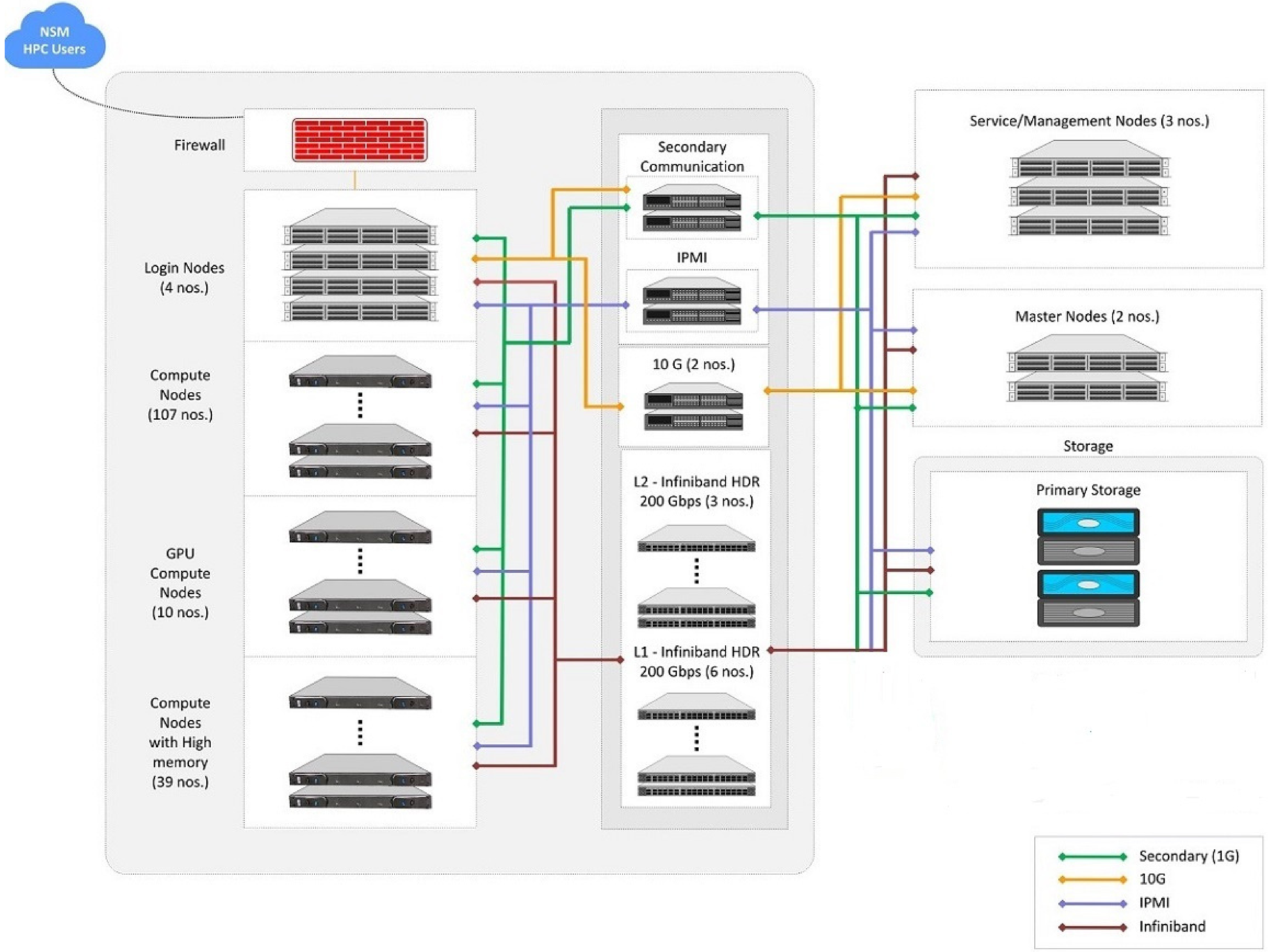
Example 3



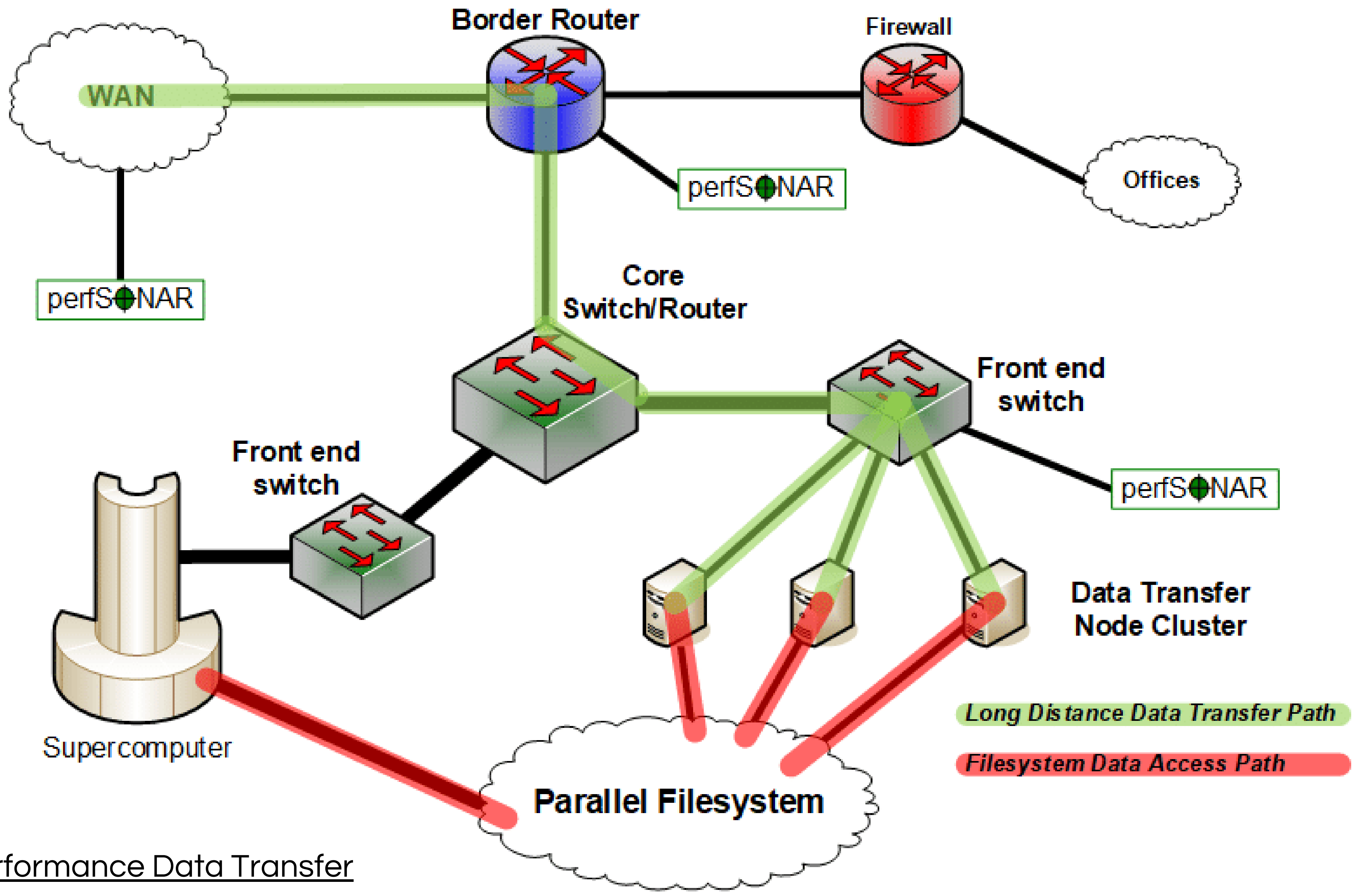
PC / Laptop
Lab Servers
Login Nodes
Compute Nodes Type 1
Normal nodes, bulk of most clusters, appropriate for most jobs.
Compute Nodes Type 2
Specialized nodes with, e.g. Large Memory, GPU, MIC, etc.
Storage 2
Storage 1

https://tacc.github.io/ctls2017/docs/intro_to_hpc/intro_to_hpc_01.html

Example 4

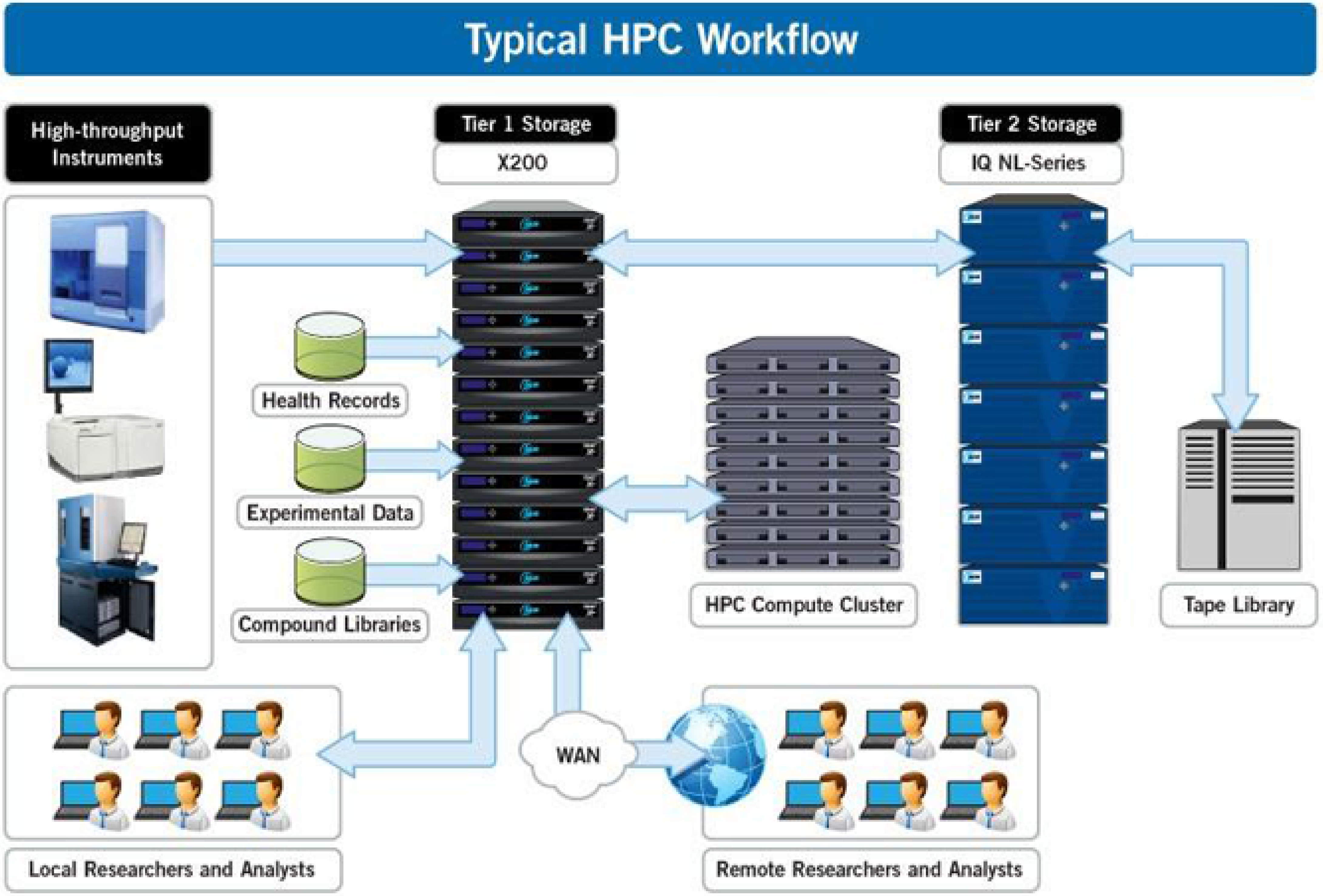

NSM
HPC Users
Firewall
Login Nodes
(4 nos.)
Compute
Nodes
(107 nos.)
GPU
Compute
Nodes
(10 nos.)
Compute
Nodes
with High
memory
(39 nos.)
Secondary
Communication
IPMI
10 G (2 nos.)
L2 - Infiniband HDR
200 Gbps (3 nos.)
L1 - Infiniband HDR
200 Gbps (6 nos.)
Service/Management Nodes (3 nos.)
Master Nodes (2 nos.)
Storage
Primary Storage
Secondary (1G)
10G
IPMI
Infiniband

Example 5



The Petascale DTN Project: High Performance Data Transfer for HPC Facilities, 2021

Example 6



http://hpcmicrosystems.net/?page_id=79