Texas A&M Supercomputing Facility
A Resource for Research and Discovery

TEXAS A&M UNIVERSITY

The New Hardware and the Role of HPC at Texas A&M

by
Spiros Vellas
Associate Director

The Scene in 2013

- Eos + Lonestar (~48 TFLOPS; ~38m core-hrs/year avail)
  - Commodity x86 & interconnect technologies;
- ~ 700 users; mostly from typical research disciplines: engineering; physics; chemistry; geoscience; …
- 8 + 1 analysts & office admin;
- Teague & Wehner machine rooms … maxed out

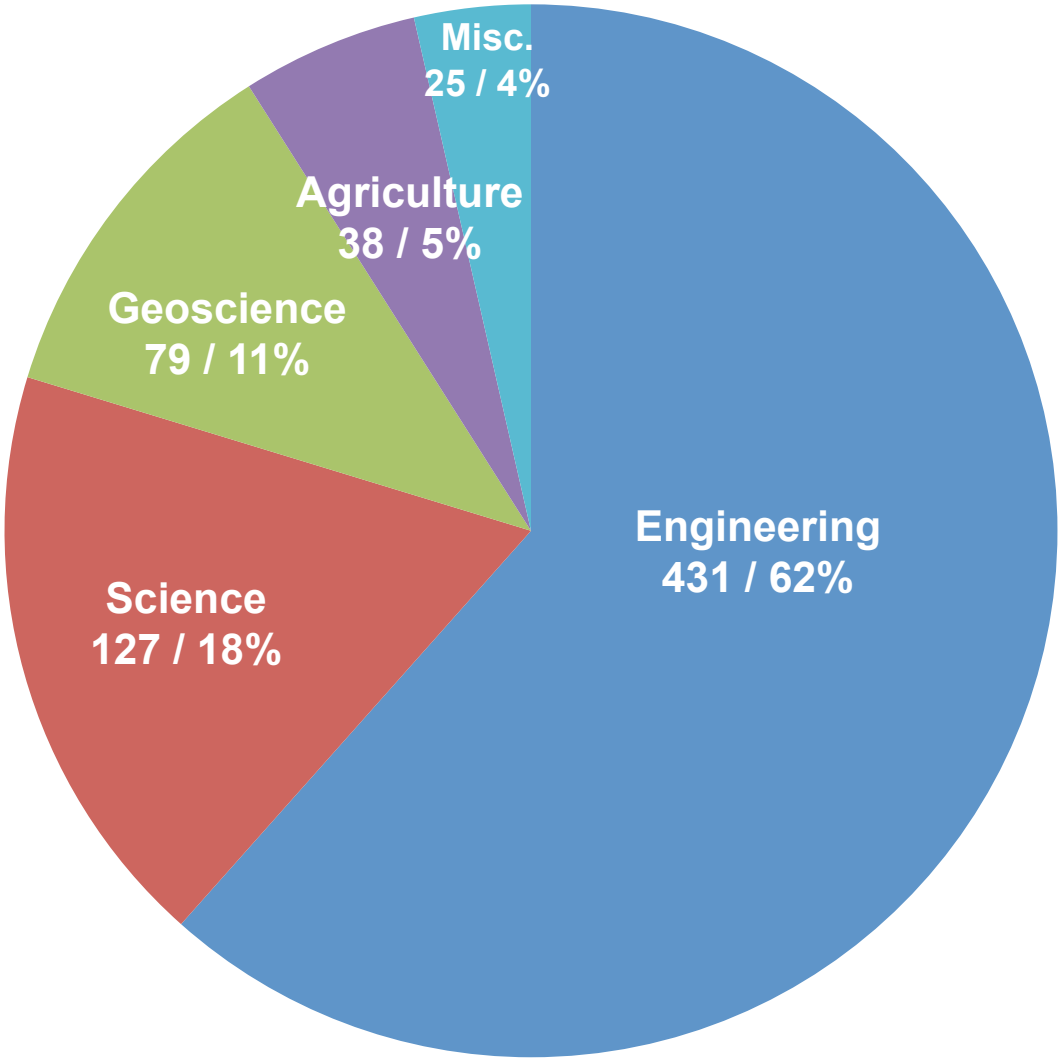
The Scene in 2014

- 4 IBM Systems:
  - 1 2048-node Blue Gene Q cluster; IBM technology;
  - 1 858-node x86 cluster; commodity x86 and interconnect parts;
  - 1 23-node p7+ system; IBM technology;
  - 1 49-node p7+ system; IBM technology;
  - 795 TFLOPS Aggregate Peak Performance … a 16-fold increase
  - 3+ (?) Different programming environments
- ~ 700 users + N (??) users from front-line research areas;
- wider spectrum of computational areas; e.g., genomics
- 8 + 1 analysts & office admin;
- Teague & Wehner machine rooms … maxed out but undergoing power and cooling upgrades

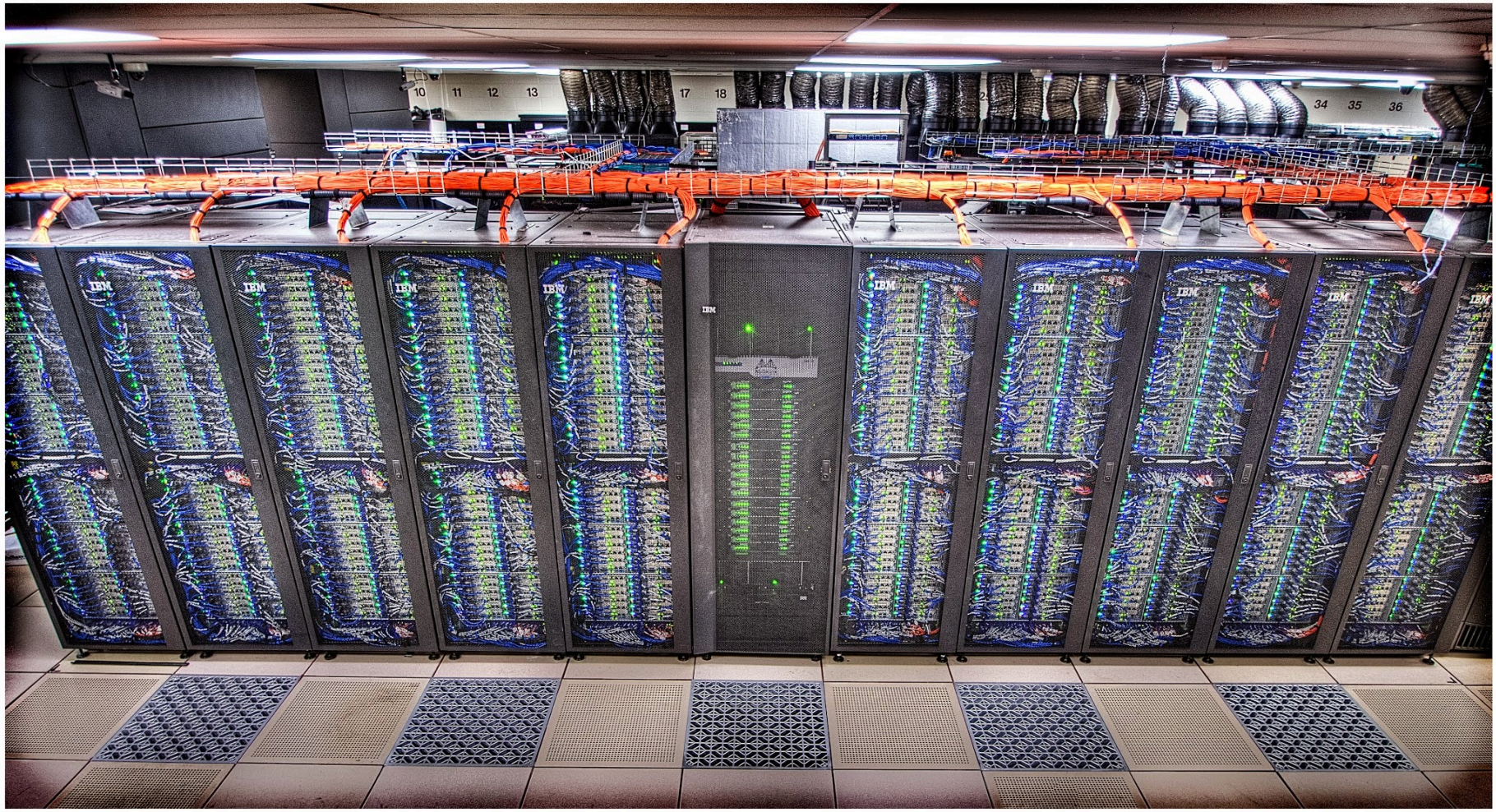
The Scene in 1989 (Supercomputing Facility established)

- 1 Cray YMP2, a 333 MFLOPS system;
  - (OS) UNICOS, Cray Programming Environment
- ~ 60 users;
- Main computational areas: Quantum Chemistry; CFD, ABAQUS;
- 3 + 2 analysts & office admins, respectively;
- 1 Director (1/4-time);
- College of Engineering Initiative; funded mostly by a bond issue

Our Users (~700) as of 4/24/2014



Ada: An IBM x86 Cluster



Ten racks of NeXtScale Compute nodes. The middle rack is a Mellanox SX6536 IB Core Switch, a major part of the InfiniBand fabric connecting all of the nodes. The total number of racks comprising Ada is 16.

Ada: 792 (Regular) Compute Nodes

Node Characteristics		
Component	Technology	Count or size/ Performance
Processor	10-core (IvyBridge) Intel Xeon E5-2670 v2 2.5 GHz; 22 nm fabrication	2 / 400 GFLOPS/node
L3 cache		25 MB per processor
Memory	4 8-byte channels of DDR3-1866 MHz per processor	64 GB / 2 x 59.7 GB/s per node
Local disk	900 GB 2.5” SAS 10K rpm	1 / 6 Gbps
Proc2proc communication	20-bit Intel Quick Path Interconnect (QPI) link	2 paths / 8.0 GT/s p2p 32 GB/s full-duplex
Inter-node communication	Mellanox ConnectX-3 Dual-port FDR14 Host Channel Adapter (HCA)	2 / 56 Gb/s per HCA full-duplex
Global disk	Via GPFS Storage Server (GSS)	4 PB / 9.6 GB/s per node

Ada: 66 Other/Special Compute Nodes

- 20 iDataPlex Nvidia 2 x K20 GPU Nodes
  - 256G Memory, 20 Cores 2.5GHz
- 10 iDataPlex Nvidia 1 x K20 GPU Nodes
  - 64G Memory, 20 Cores 2.5GHz
- 9 iDataPlex Intel 2x Phi Accelerator Nodes
  - 64G Memory, 20 Cores 2.5GHz
- 6 iDataPlex 256GB Medium-Large Memory Nodes
  - 256G Memory, 20 Cores 2.5GHz
- 13 x3850 x5 1TB X-Large Memory Nodes
  - 1024GB Memory, 40 Cores 2.27GHz
- 2 x3850 x5 2TB XX-Large Memory Nodes
  - 2048GB Memory, 80 Cores 2.27GHz
- 6 Login Nodes
  - 256G Memory, 20 Cores 2.5GHz; 1-2 GPUs; 2 Phi; 4 900GB 10K rpm SAS disk drives

Ada: The FDR10 Infiniband Fabric

The Teague InfiniBand fabric connects to all System x compute nodes and storage nodes for MPI and storage traffic.

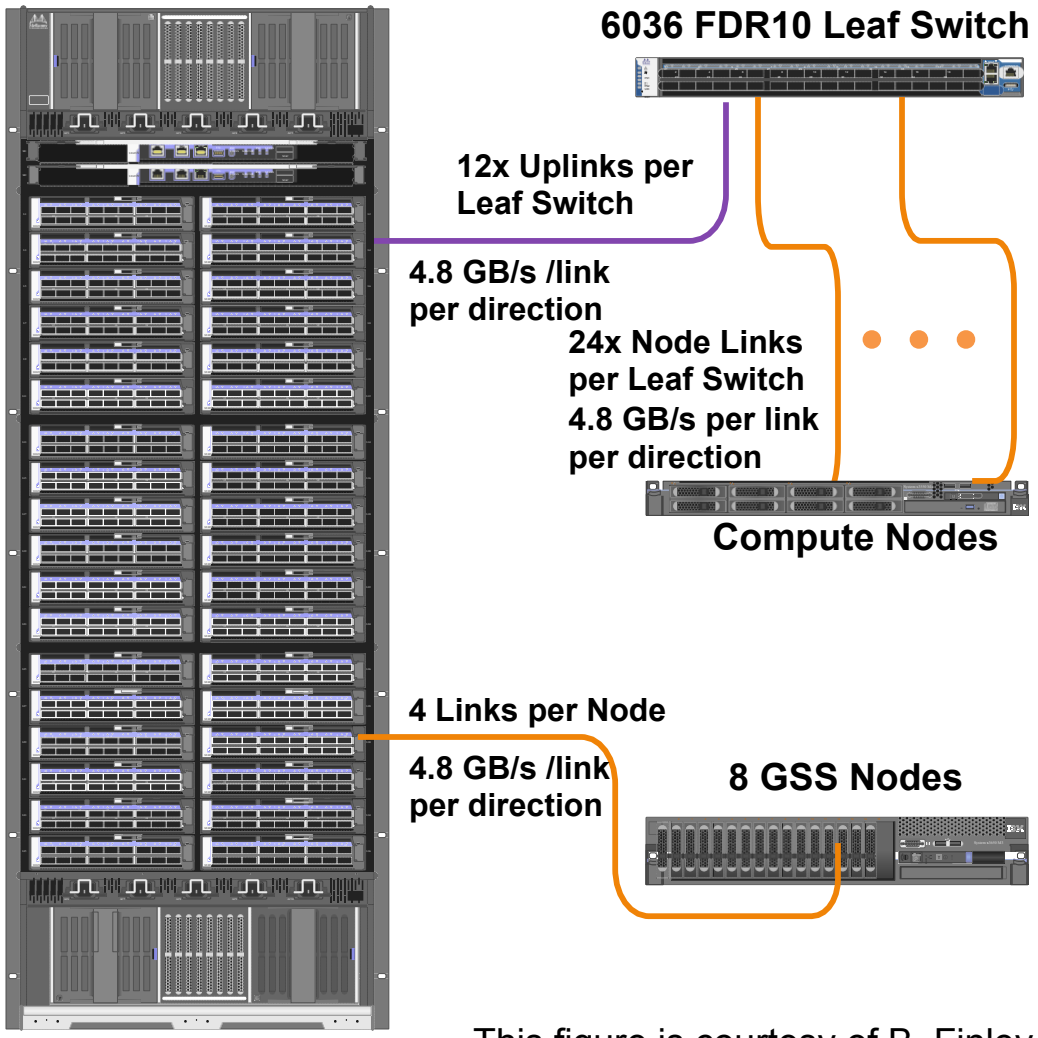
Compute nodes connect to leaf switches, which uplink to the spine at a 2:1 blocking ratio.

GSS nodes (see below) connect directly to the core with 4x node links each, for 160Gb/s aggregate per GSS node.

SUR grant DCS3700 storage (see farther below) nodes also connect directly to the core with 1x node link each.

Node Link

Switch Uplink



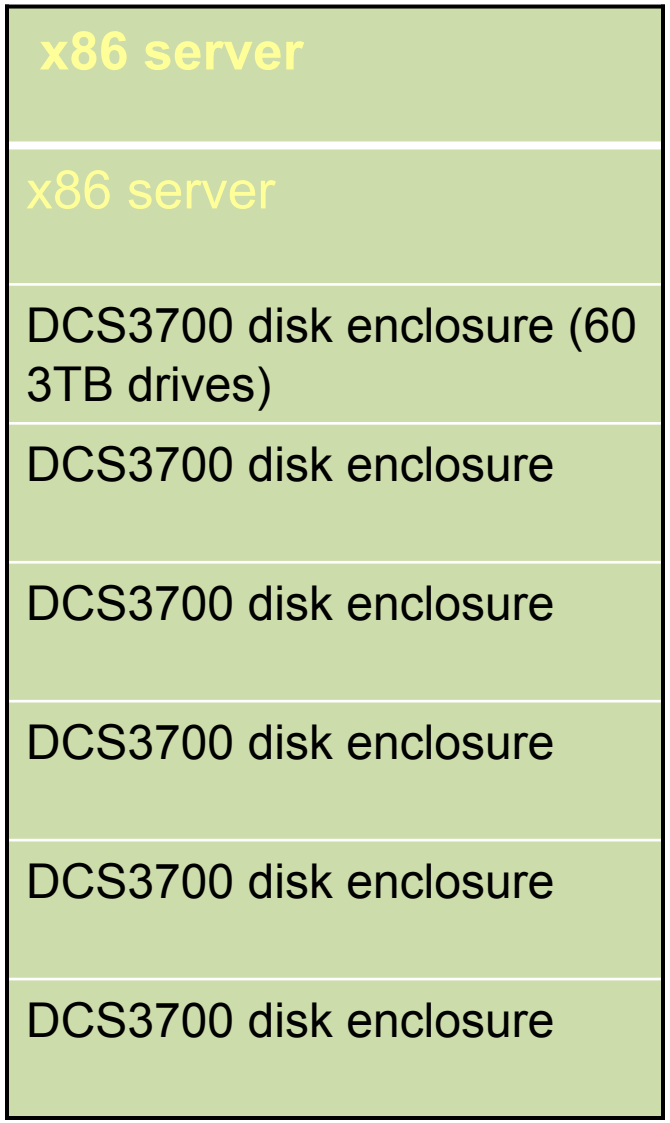
This figure is courtesy of B. Finley

Ada: Shared Fast Mass Storage 1

1 GSS Model 26 Block

4 x GPFS Storage Server Model 26 (GSS26)
4 PB total storage (raw) organized by GPFS
~48 GB/s aggregate throughput
8 x86 storage servers (2/block) to power data transfers between the GSS26 and the Ada & p7+ nodes (see core switch diagram)
32 (4 links per storage server) direct IB FDR10 connections to the InfiniBand Core Switch
16 10GbE connections (2/storage server) to the Ethernet Core Switch in Teague
P7+ nodes (in wehner) access GSS26 via 2 40GbE connections between Wehner and Teague machine rooms

Ada: Shared Slower Mass Storage



IBM System Storage DS3512 Express (Teague machine room)
1.08 PB total storage (raw) organized by GPFS; 6 DCS3700 disk enclosures
~3 GB/s aggregate throughput
2 x86 storage nodes to power data transfers
2 10GbE connections to the Ethernet Core Switch in Teague
Attached to Ada (and the p7+'s) via 10GbE connections
An IBM Shared University Research Grant (SUR Grant) in support of: (1) Plant & Animal Genomics; (2) Health Surveillance; (3) Genomic Signal Processing; (4) Gulf Climate & Environmental Systems Integration; (5) Oil & Gas

Archival & Backup Storage

IBM System Storage TS3500 Tape Library (in Teague machine room)
10 PB total storage (raw) organized by GPFS
~1 GB/s aggregate throughput
2 x86 storage nodes for powering data transfers
1 10GbE connections to the Ethernet Core Switch in Teague
Attached to Ada and the p7+'s via 10GbE
Major user: Geoscience

Crick : 23-node POWER7+ cluster

1 7R2 20-Node Rack

Big Data Analytics Cluster w Storage-rich Nodes IBM 7R2 Node Characteristics
16-core ((2 8-core Power7+ processors) 4.2GHz; 537GFLOPS; 32 nm fabrication
256 GB memory; 34GB/s (??) per processor socket; Shared (on-chip) L3 80MB processor
10GbE port (link to Wehner Core 10G/40G)
4 x 600GB 10K rpm SAS (local disk)
EXP24S + 24 x 600GB (14.4TB) 10K rpm SAS drives
REDHAT ENTERPRISE LNX FOR PWR; GPFS Client
Special Software: IBM InfoSphere BigInsights, Data Explorer Resource; Query Routing for InfoSphere Data Explorer; etc
Target Areas: Big Data Analytics; Genomic Analysis, Breeding Simulation, mining historical data; Map Reduce; etc

Courant : 49-node POWER7+ cluster

1 7R2 20-Node Rack

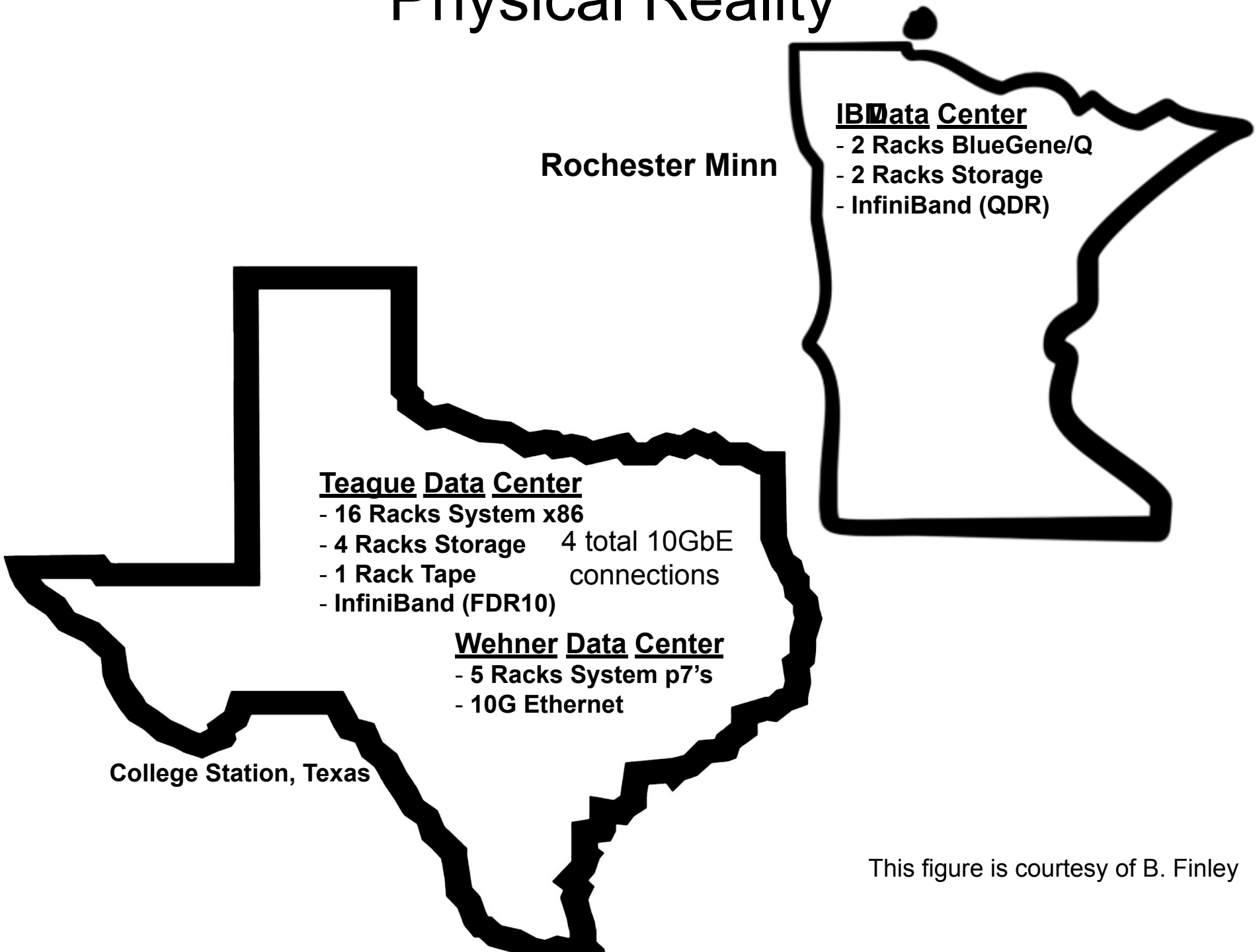
Special HPC Codes & Applications IBM 7R2 Node Characteristics
16-core (2 8-core Power7+ processors) 4.2GHz; 537GFLOPS; 32 nm fabrication
256 GB memory; 34GB/s (??) per processor socket; L3 80MB/processor
10GbE port (link to Wehner Core 10G/40G switch)
4 x 600GB 10K rpm SAS (local disk)
REDHAT ENTERPRISE LINUX FOR PWR ; GPFS Client; IBM Compilers, ESSL, LSF (batch)
Target areas: applications & codes requiring fast memory and fast cpus

Neumann: The 2048-node Blue Gene Q (BG/Q) Cluster

1 (of 2) BG/Q Rack

32768-core, 419 TFLOPS, Massively Parallel, interconnected in 5D Torus topology
Node: 16 compute 1.6 GHz PowerA2 cores; 1 core for system calls; 1 core for spare 16 GB of DDR3 memory 204.8 GFLOPS (1.6 GHz * 8 Flop/s * 16)
1 I/O drawer per rack: 8 I/O nodes with 1 QDR IB link (via the torus interconnect) for every 2 comp-ute nodes. Smallest job is 128-way
Shared Mass Storage 2: 2 PB of GSS26 con-necta to the I/O drawers via 16 IB connections thru 2 QDR/ FDR10 IB switches
4 Front End Nodes: 4-core 3.6 GHz POWER7+ processor, 32 GB RAM, 2 600 GB SAS Drives
OS: Compute Node Kernel (CNK) runs on all compute nodes; a lightweight kernel, similar to Linux, supporting a large subset of its system calls
LoadLeveler (batch); IBM XL and GNU compilers; Engineering & Scientific Library (ESSL)

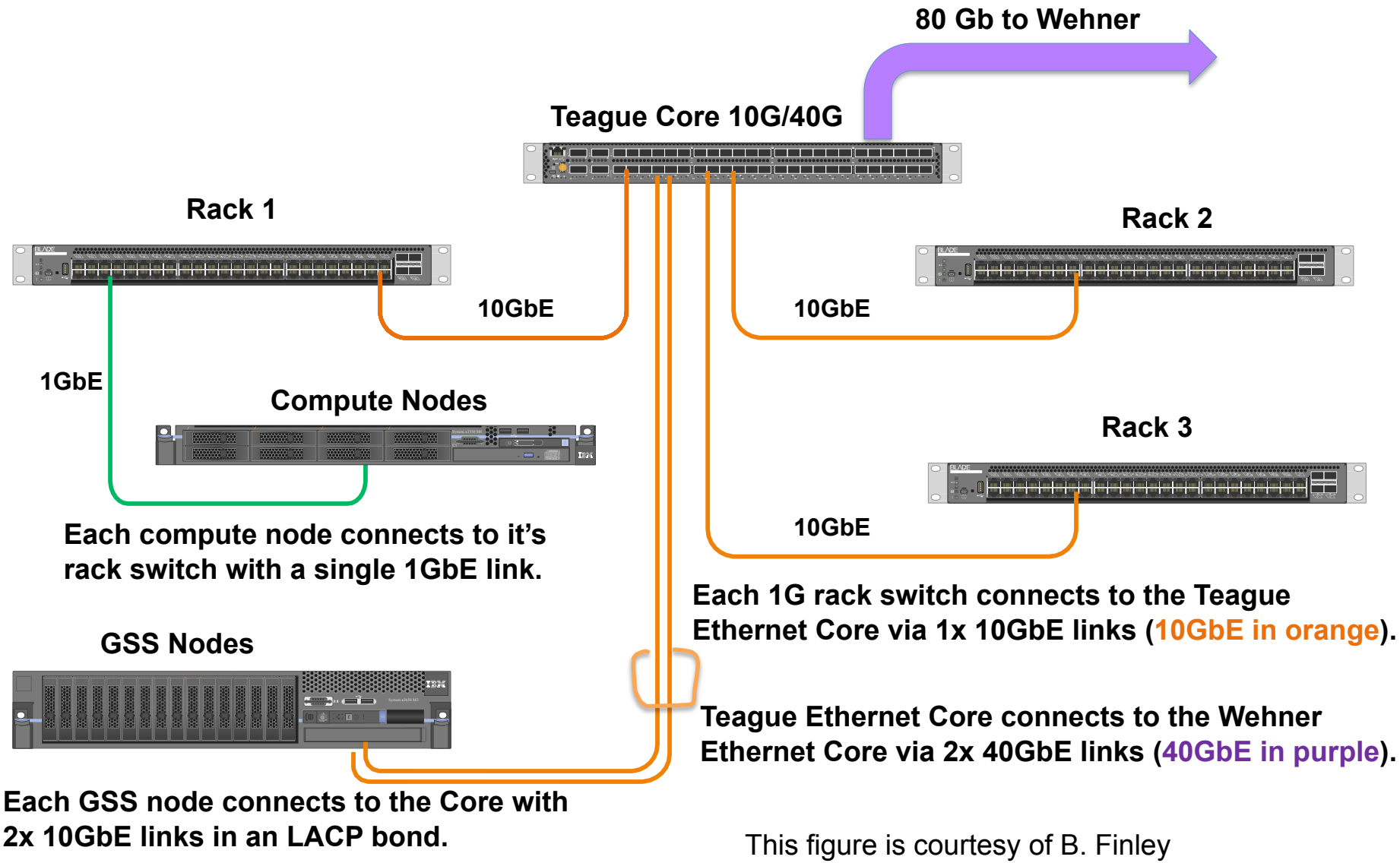
Physical Reality



This figure is courtesy of B. Finley

Teague Ethernet

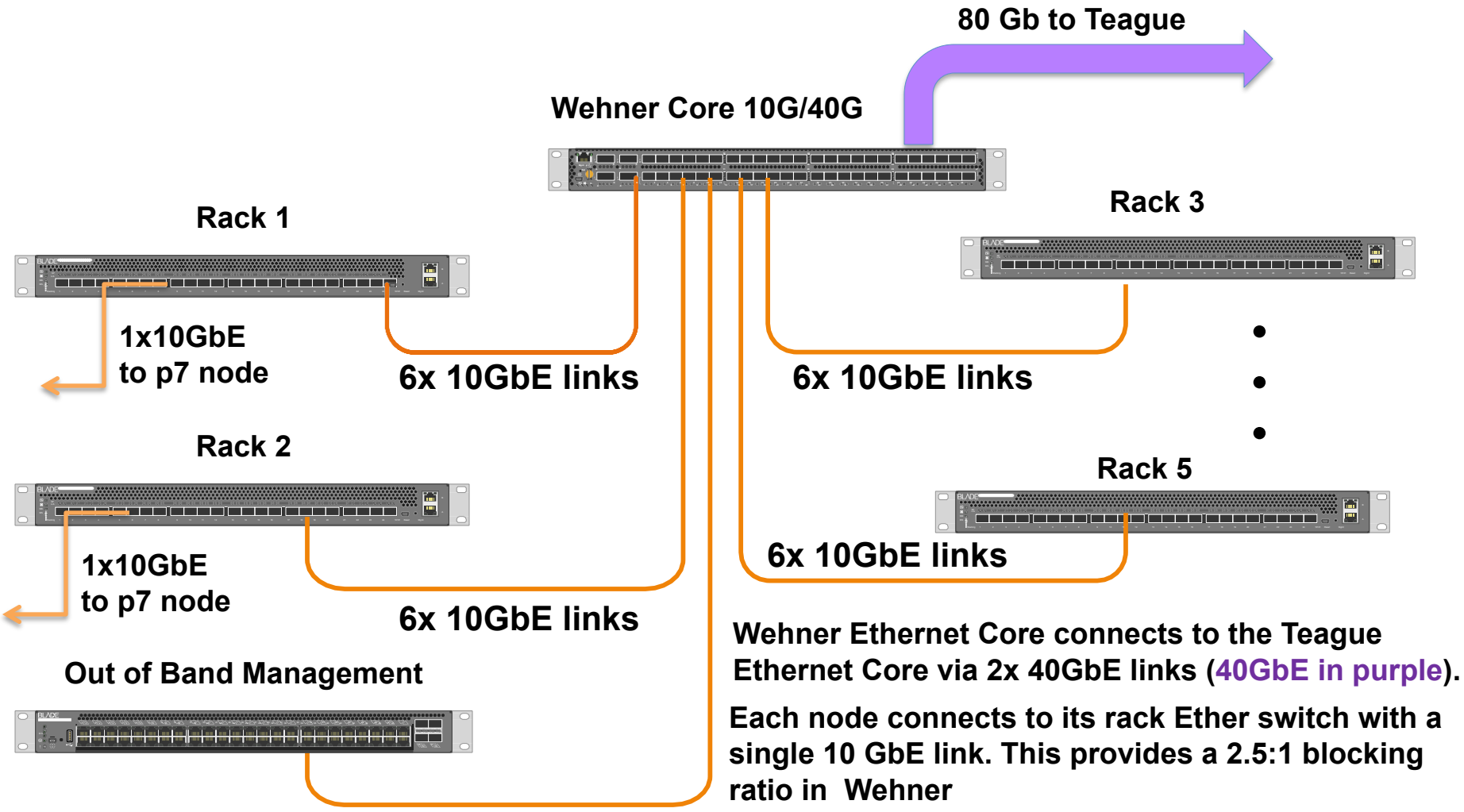
1G Ethernet for the Main OS and Out of Band Management



This figure is courtesy of B. Finley

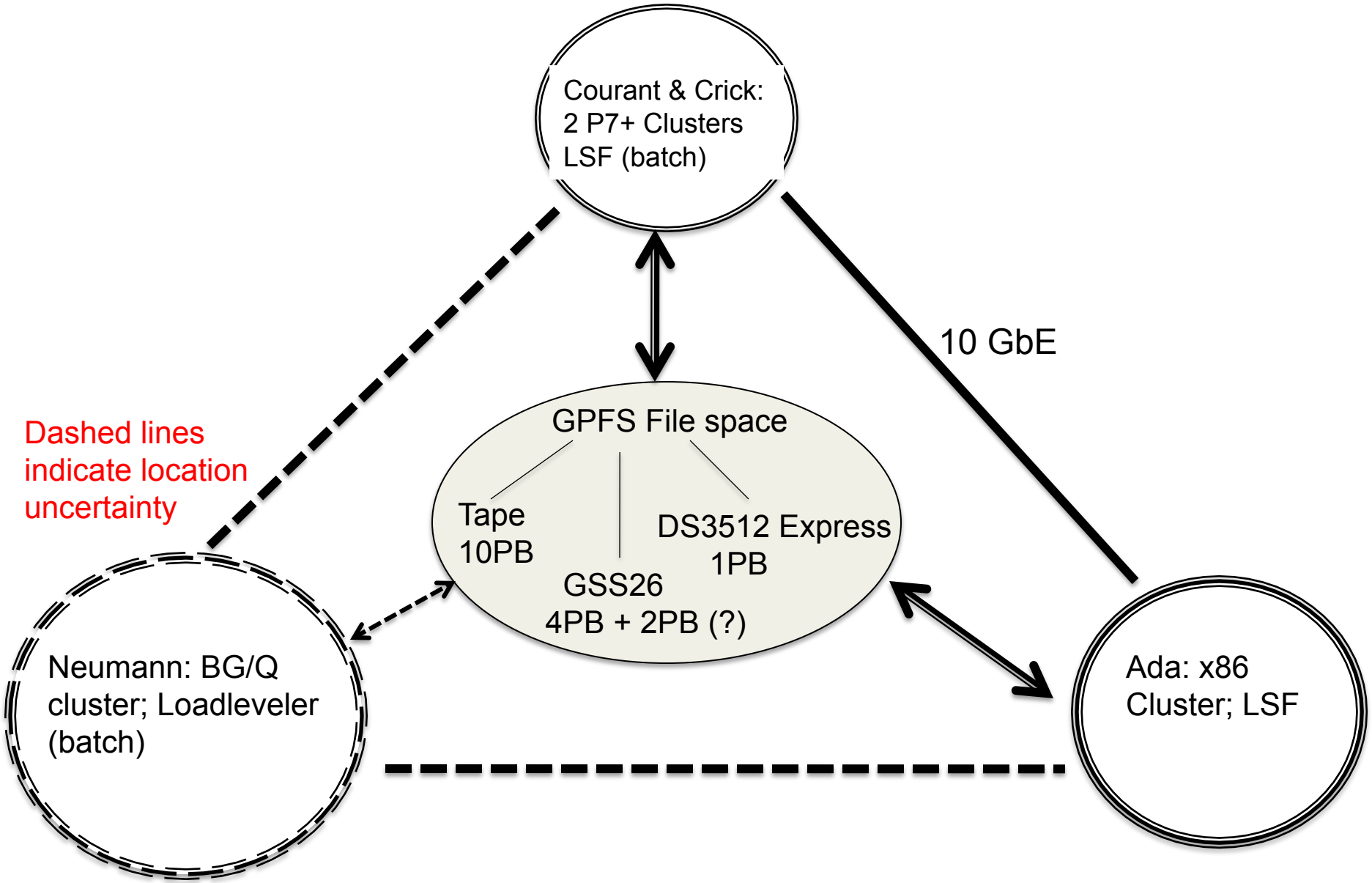
Wehner Ethernet (10GbE)

10G Ethernet for the Main OS and Data Network



This figure is courtesy of B. Finley

2 Unifying Agents: GPFS & LSF



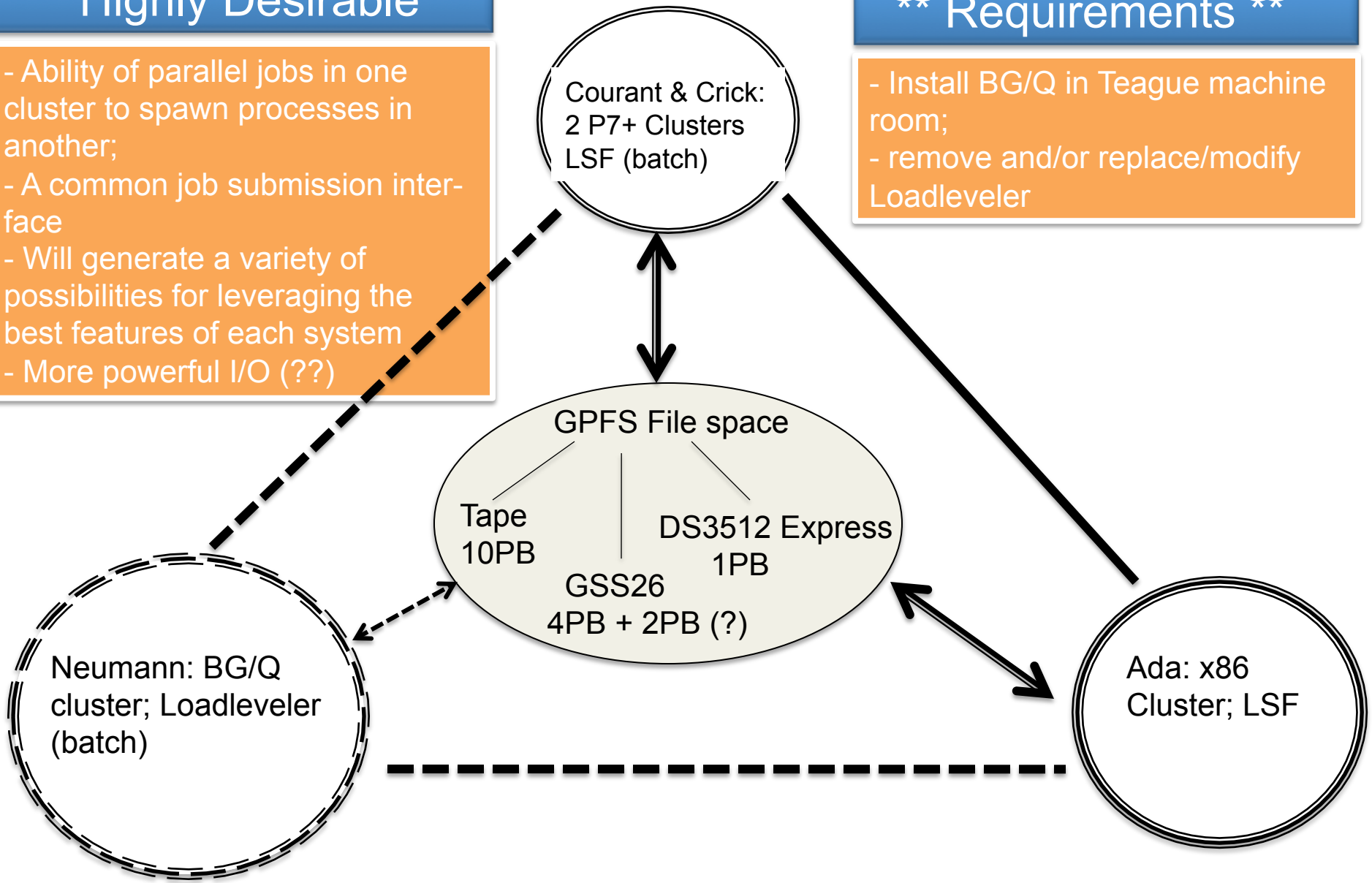
A Challenge to A&M & IBM: Unifying 4 Clusters

** Highly Desirable **

- Ability of parallel jobs in one cluster to spawn processes in another;
- A common job submission inter-face
- Will generate a variety of possibilities for leveraging the best features of each system
- More powerful I/O (??)

** Requirements **

- Install BG/Q in Teague machine room;
- remove and/or replace/modify Loadleveler



Fundamental Objective of the HPC Project

Integration of HPC into Research & Teaching

Parties in the HPC Project

Users ⟺ HPC Infrastructure

Students
Researchers adopting HPC
Misc Others

Skilled Analysts
Hardware Technologies
Software Technologies
Machine Room capacity

Skilled Analysts: Vital to deliver the promise of HPC

We need analysts with strong backgrounds in

- HPC cluster management;
- HPC architectures & related technologies;
- Code optimization & parallelization;
- Scientific (& remote) visualization;
- Big data analytics;
- Web technologies;
- Scientific disciplines: bioinformatics & genomics; weather modeling; molecular dynamics; PDF's; etc

No analysts on the staff with these skills

Facility's 9-member Staff by Activity

- 2 analysts, mostly for system management & hardware technologies;
- 3 analysts mostly for code optimization, parallelization & general user consulting;
- 1 analyst mostly for user accounting & general user consulting;
- 1 analyst for total management & strategic directions, hardware technologies, code optimization & parallelization;
- 1 office admin
- 1 analyst Vacancy …

Needed: More Multi-discipline Analysts to Redeem the Promise of HPC

- Who possess HPC skills as well as at least modest backgrounds in
  - Bioinformatics & Genomics;
  - Big Data Analytics;
  - Scientific (and remote) Visualization;
  - Web Technologies;
  - Applied Mathematics;
  - Physical Science: engineering; physics; chemistry; geoscience

By priority analysts with these skills needed

Production Mode Timelines

- Ada: power & cooling upgrades just completed; installation & configuration ~80% complete; expected on-line by end of May;
- Courant & Crick (p7+'s): power & cooling upgrades underway in Wehner machine room; rack installation in initial phase; expected on-line by end of June;
- Neumann (BG/Q): installation location not determined yet