User Communities

Ab initio simulations

Post-processing

Data analysis

Storage
Computers

Typically 1-5 individual systems

10K - 1M Cores

High Bandwidth Low Latency Interconnect
(22 GB/sec and 0.80us)

Increasingly often uses accelerators
(GPGPUs MICS etc)
Computers

Typically 1-5 individual systems

10K - 1M Cores

High Bandwidth Low Latency Interconnect (22 GB/sec and 0.89us)

Increasingly often uses accelerators (GPGPUs MICs etc)
North Bridge
Memory
North Bridge
Core i7 Xeon 5500 Series
Data Source Latency (approximate)

L1 CACHE hit, ~4 cycles

L2 CACHE hit, ~10 cycles

L3 CACHE hit, line unshared ~40 cycles

L3 CACHE hit, shared line in another core ~65 cycles

L3 CACHE hit, modified in another core ~75 cycles

remote L3 CACHE ~100-300 cycles

Local Dram ~60 ns

Remote Dram ~100 ns
North Bridge
for all buffers:
W: SendReq
SendData
OK

E: RunKernel
OK

R: ReadReq
RecvData
OK

[W];[W||E];[W||E||R];...;[E||R];[R]
Executor

Writer

Data

Reader

[W];[W||E];[W||E||R];...;[E||R];[R]
North Bridge
South Bridge
Disk Storage

- Hosts
- Input-files for running jobs
- Result-files for running jobs
- Files targeted for tape
- Files staged from tape

Typically tens of PB disk
Tape Storage

- Cheap mass storage for large datasets
- Backup of important data
- Typically tens of PB
Scheduler

Receives input from
- Users
- Computers

Depends on
- Computers
- Staging from tape
- Real-time reservations
- Grid
Grid

Receives input from
- External Schedulers
- Computers
- Tape

Depends on
- Computers
- Staging from tape
- External schedulers
Visualization

Depends on
- Scheduler
- Computers
- Network
Gateway
100 Gbps primary connection
Hundreds of concurrent connections
Open Problems in HPC for the CPA Community
Extensible Software Defined Networking
[BW allocation, user filters for data, interface to IDS]
Flow-based Access Control
[Access to Interactive nodes, and peeking to output graphics]
Resource Availability Coordination
[Ensuring that data is online before jobs are scheduled]
Buffer Migration Management
[Getting data from slow to fast storage, eliminate waiting]
Peer-Scheduler Coordination
[Communicating between Grid schedulers]
Requirement Checking
[Not starting a job that has no storage quota left]
Opportunistic Backup
[If a file that is marked for backup is accessed for other purposes, do backup in parallel]