



User Communities

- File info simulations
- Task processing
- Data analysis
- Storage



Tape Storage

- Competition to hard drives
- Access speed (seconds)
- Capacity (PB)



Disk Storage

- Search
- Specialty for scientific data
- Searchable on research data
- File targeted for logs
- File targeted for backup
- Typically lots of metadata



Visualization

- Depends on:
 - Schedulers
 - Computers
 - Network



Scheduler

- Plots as Gantt chart
- Queue
- Combinatorics

Externalities

- Computers
- Storage
- Network
- Hardware requirements
- Cost



Computers

- Typically 1:1 individual systems
- 1MB - 1TB cores
- High bandwidth local memory movement
- CO2 footprint (heat)
- Increasingly ultra-low cost calculations (GPU, FPGA, etc.)



Gateway

- Multiple gateway connections
- Network-to-computer connections



Time

- Resolves input data
- External dependencies
- Computers
- Queue

Depends on:

- Computers
- Scheduling from above
- External schedulers

User Problems Related to the Data Community

- Workflows that span multiple systems
- Workflows that span multiple users
- Workflows that span multiple sites
- Workflows that span multiple data formats
- Workflows that span multiple data sources
- Workflows that span multiple data destinations
- Workflows that span multiple data processing steps
- Workflows that span multiple data processing environments
- Workflows that span multiple data processing tools
- Workflows that span multiple data processing languages
- Workflows that span multiple data processing frameworks
- Workflows that span multiple data processing technologies
- Workflows that span multiple data processing standards
- Workflows that span multiple data processing protocols
- Workflows that span multiple data processing interfaces
- Workflows that span multiple data processing APIs
- Workflows that span multiple data processing SDKs
- Workflows that span multiple data processing libraries
- Workflows that span multiple data processing packages
- Workflows that span multiple data processing modules
- Workflows that span multiple data processing components
- Workflows that span multiple data processing services
- Workflows that span multiple data processing providers
- Workflows that span multiple data processing vendors
- Workflows that span multiple data processing partners
- Workflows that span multiple data processing ecosystems
- Workflows that span multiple data processing environments
- Workflows that span multiple data processing ecosystems



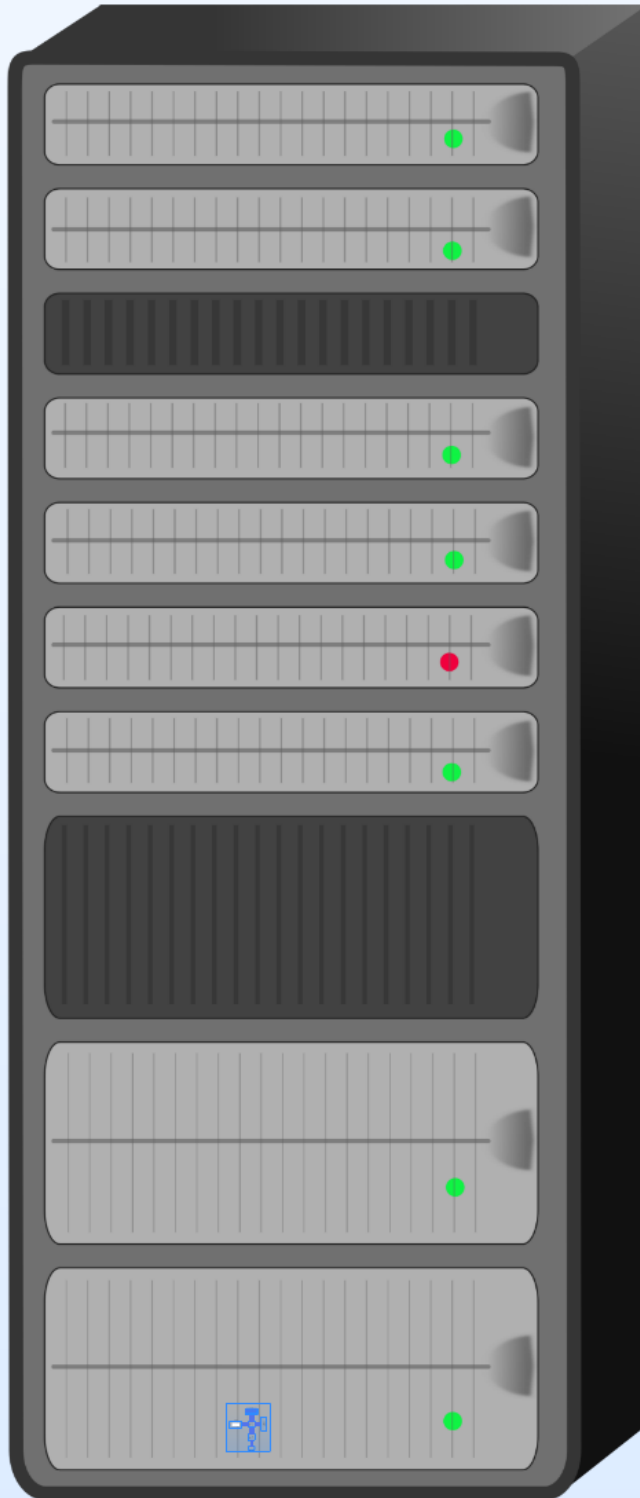
User Communities

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 0.89us)

Increasingly often uses accelerators
(GPGPUs MICs etc)

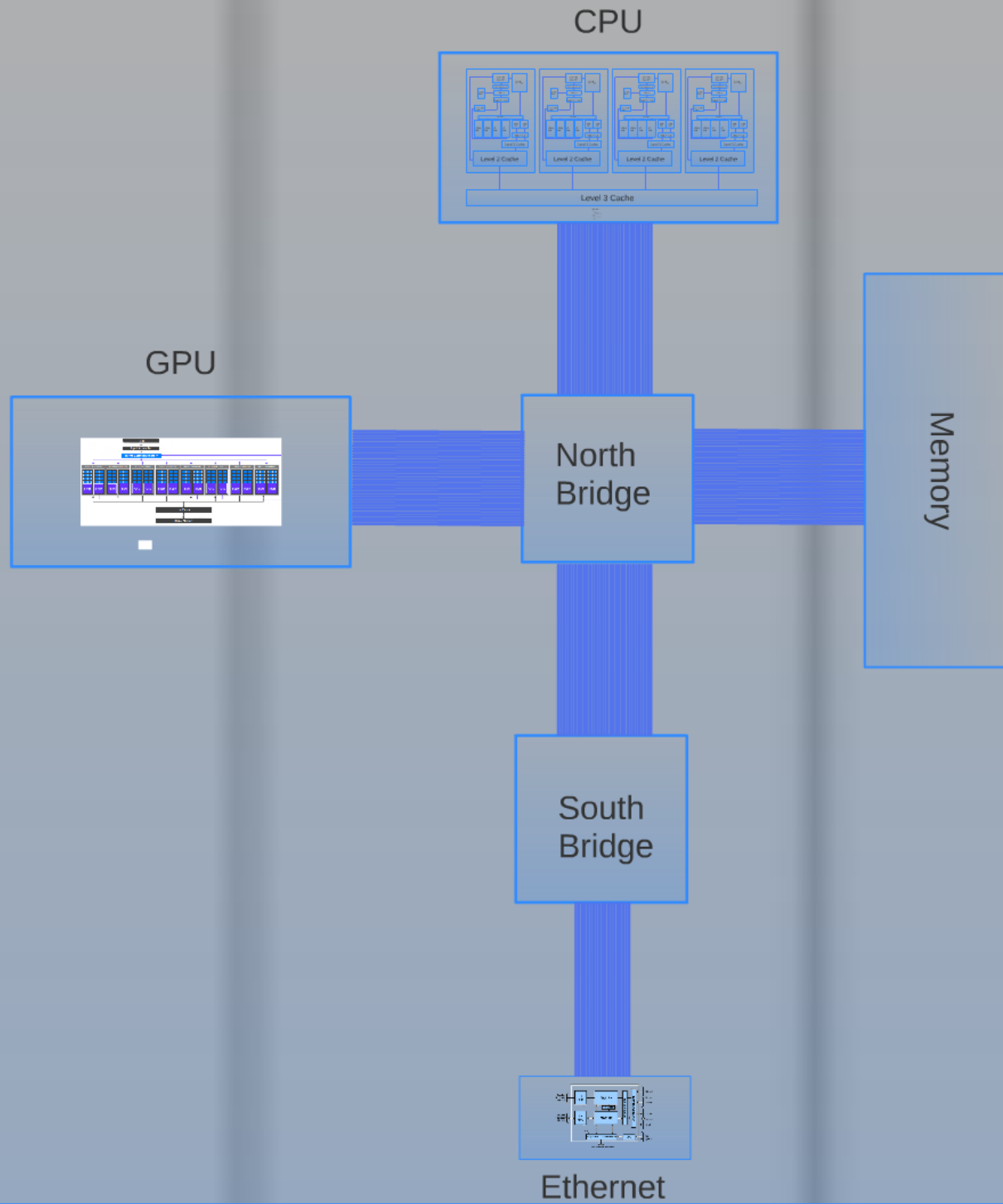
Computers

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)



CPU

GPU

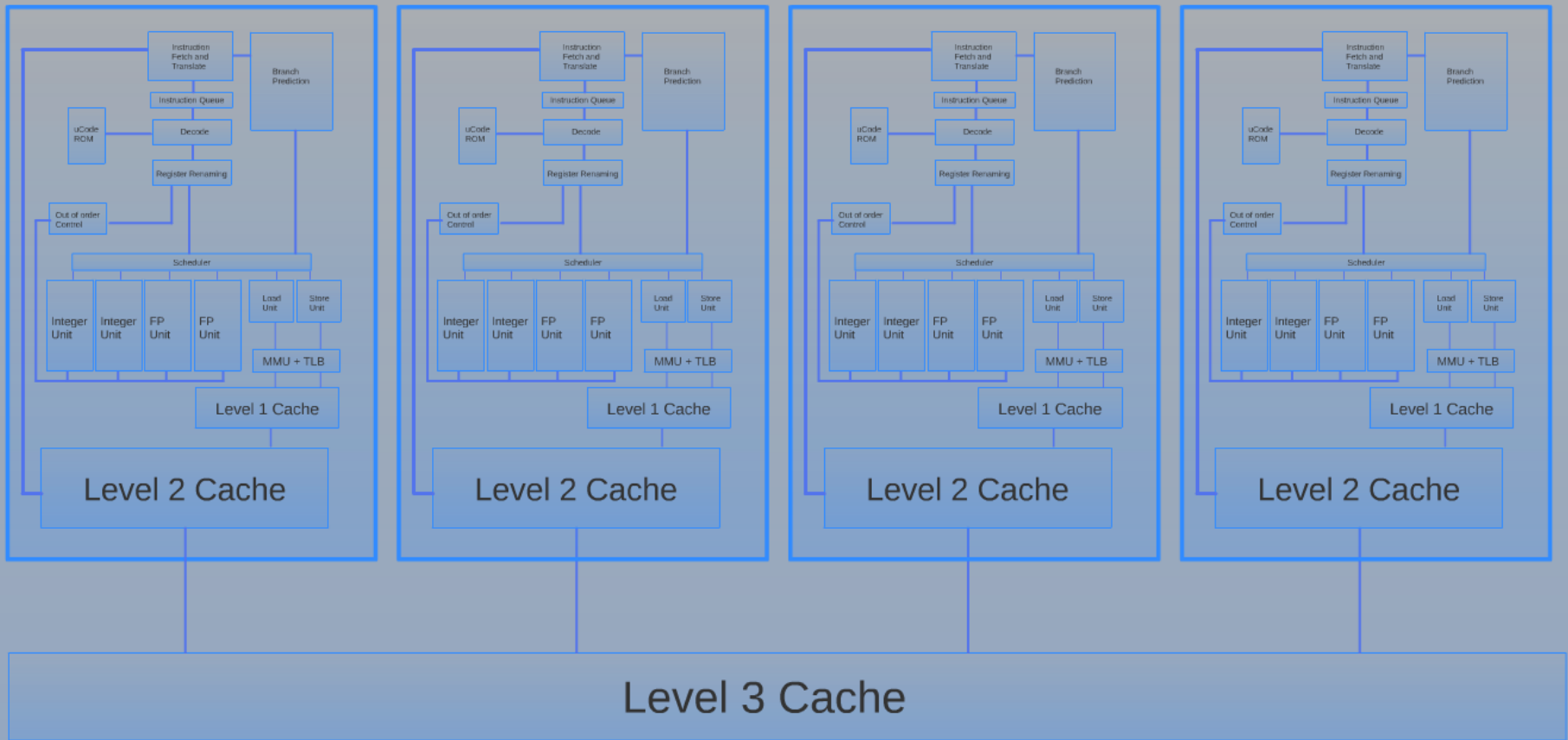
Memory

North Bridge

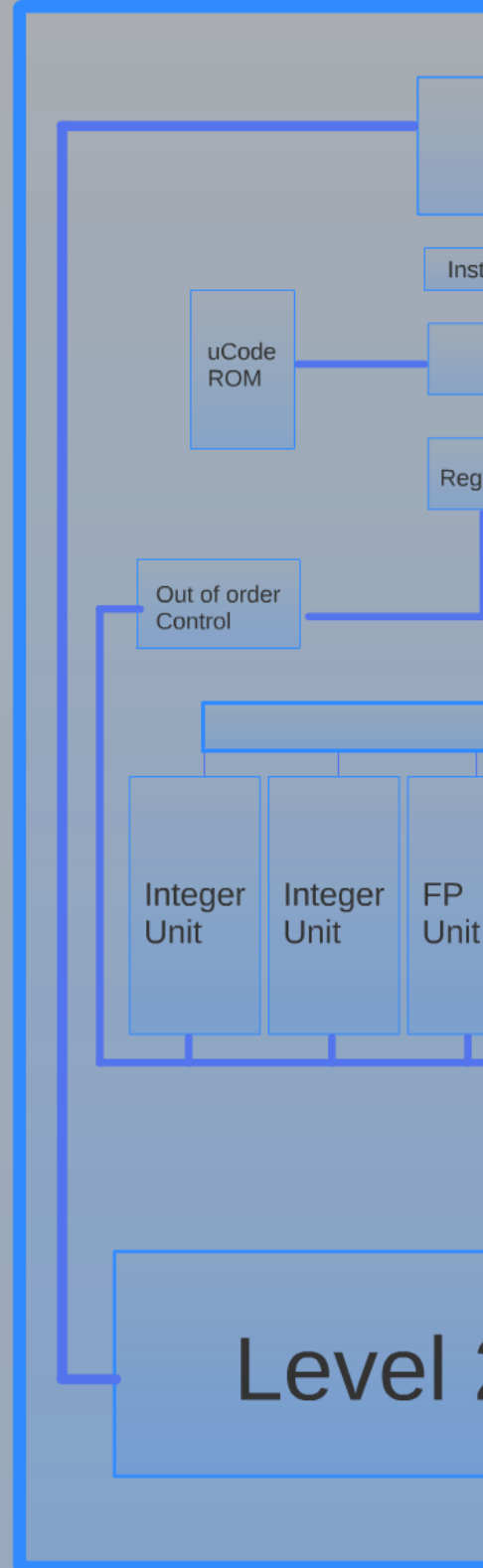
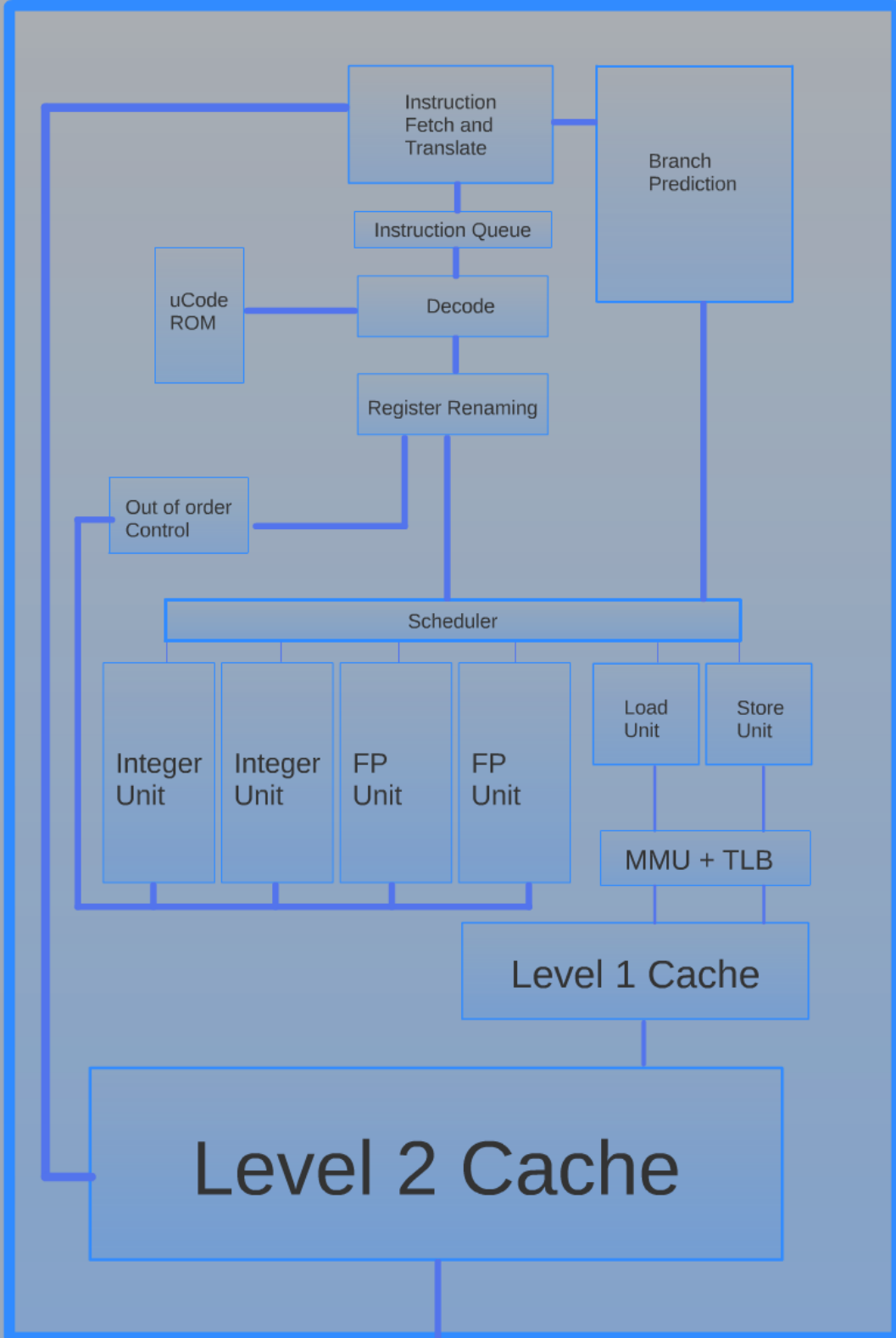
South Bridge

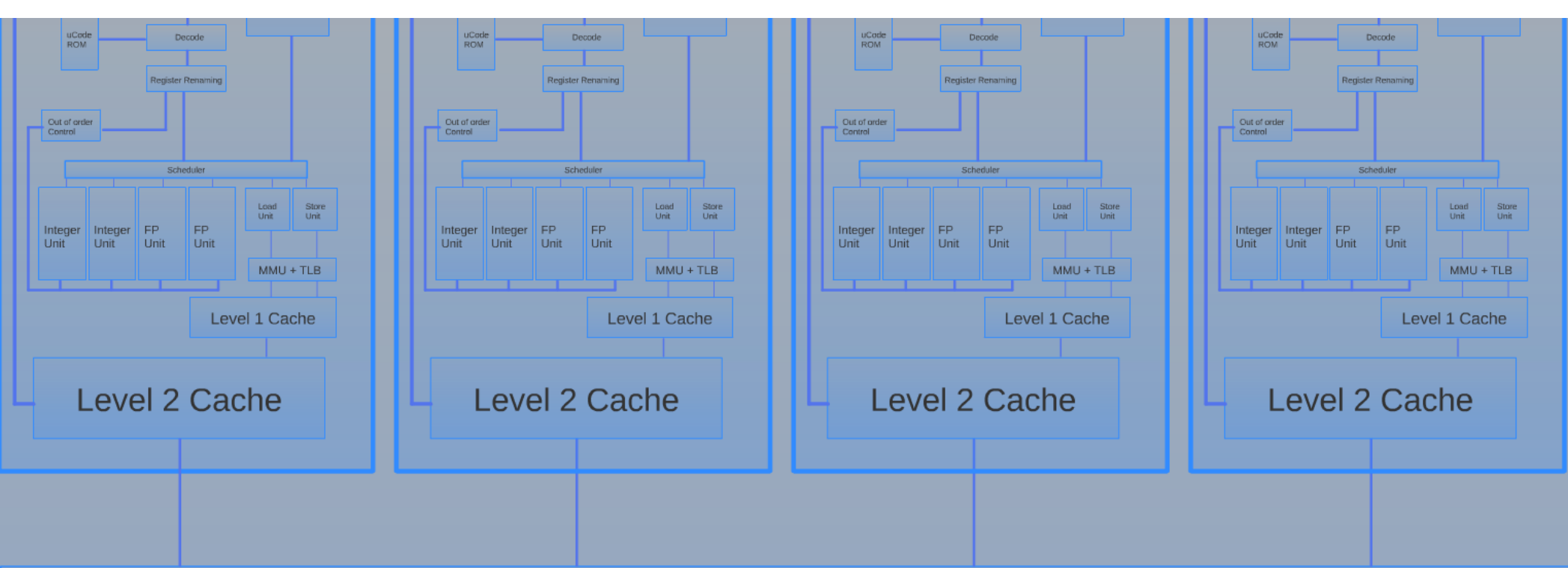
Ethernet

CPU



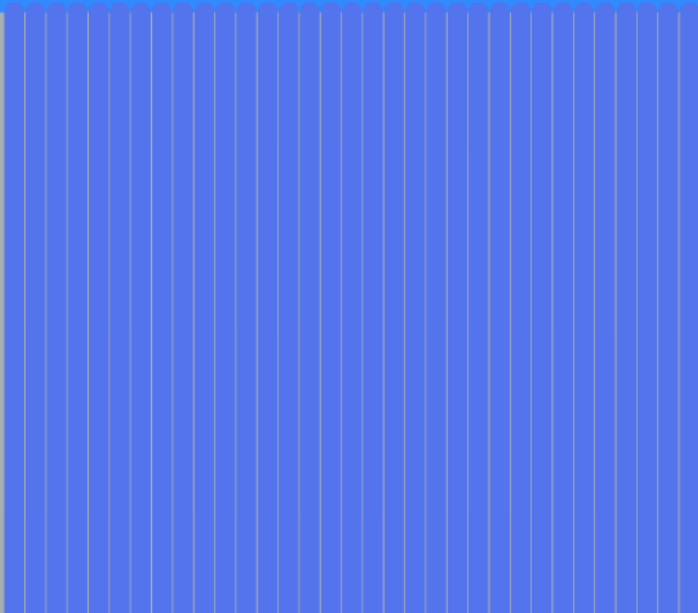
© 2000 Intel Corporation
All rights reserved.
Intel, the Intel logo, and
Celeron are trademarks or
registered trademarks of
Intel Corporation or its
subsidiaries in the United
States and other countries.





Level 3 Cache

The Level 3 Cache is a shared cache that provides a high-speed path to main memory. It is typically implemented as a crossbar or a set of multiplexers that route data between the Level 2 caches and the main memory controller.



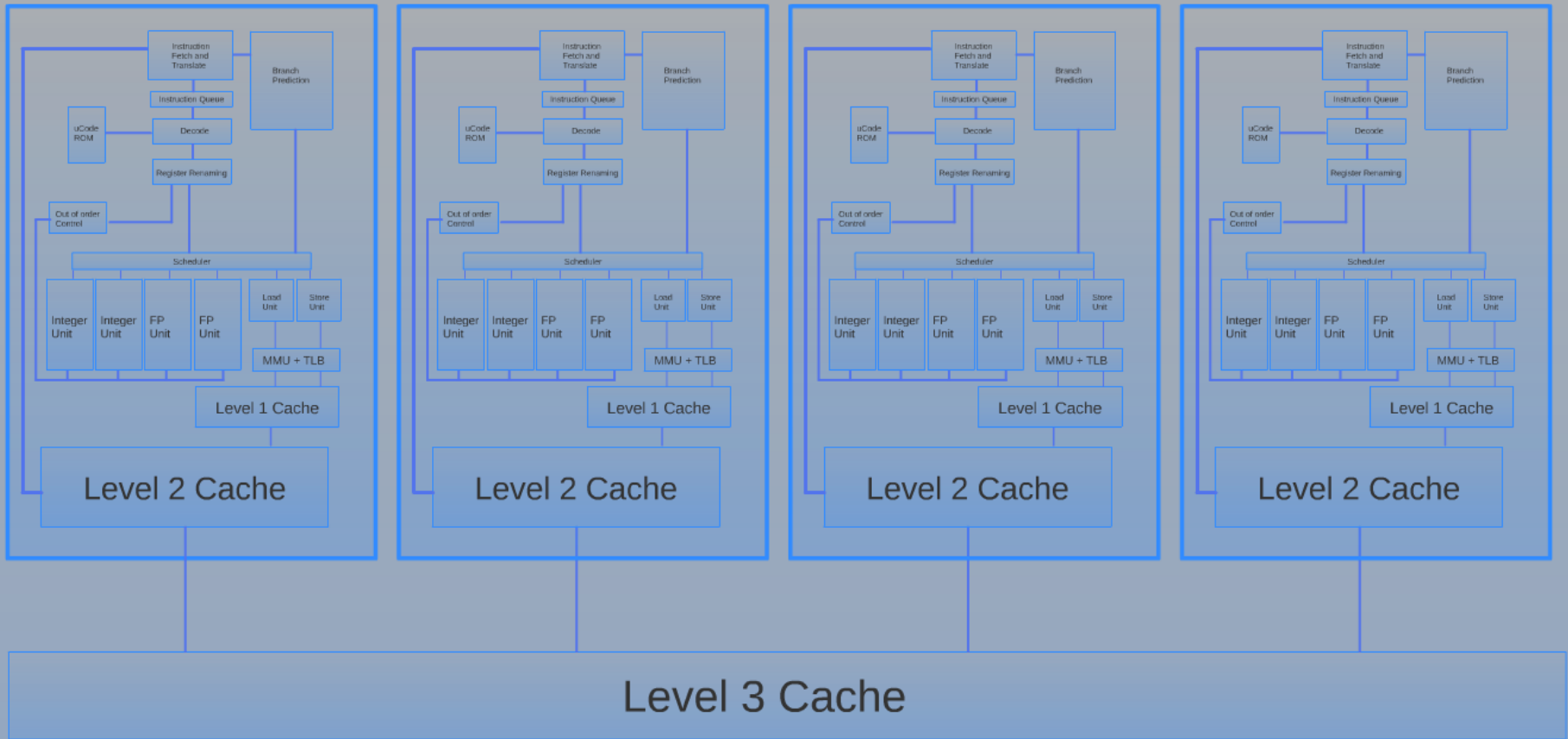
North Bridge



Memory

North Bridge

CPU



© 2010 Intel Corporation. All rights reserved. Intel, the Intel logo, and Intel Inside are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. Other brands and product names are trademarks of their respective owners.

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

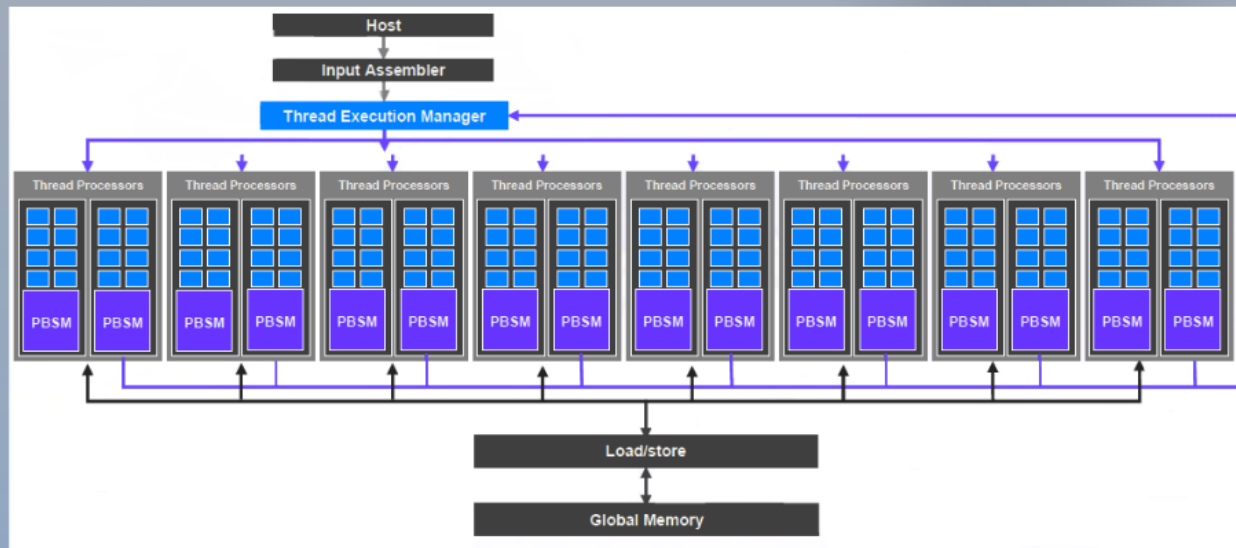
remote L3 CACHE ~100-300 cycles

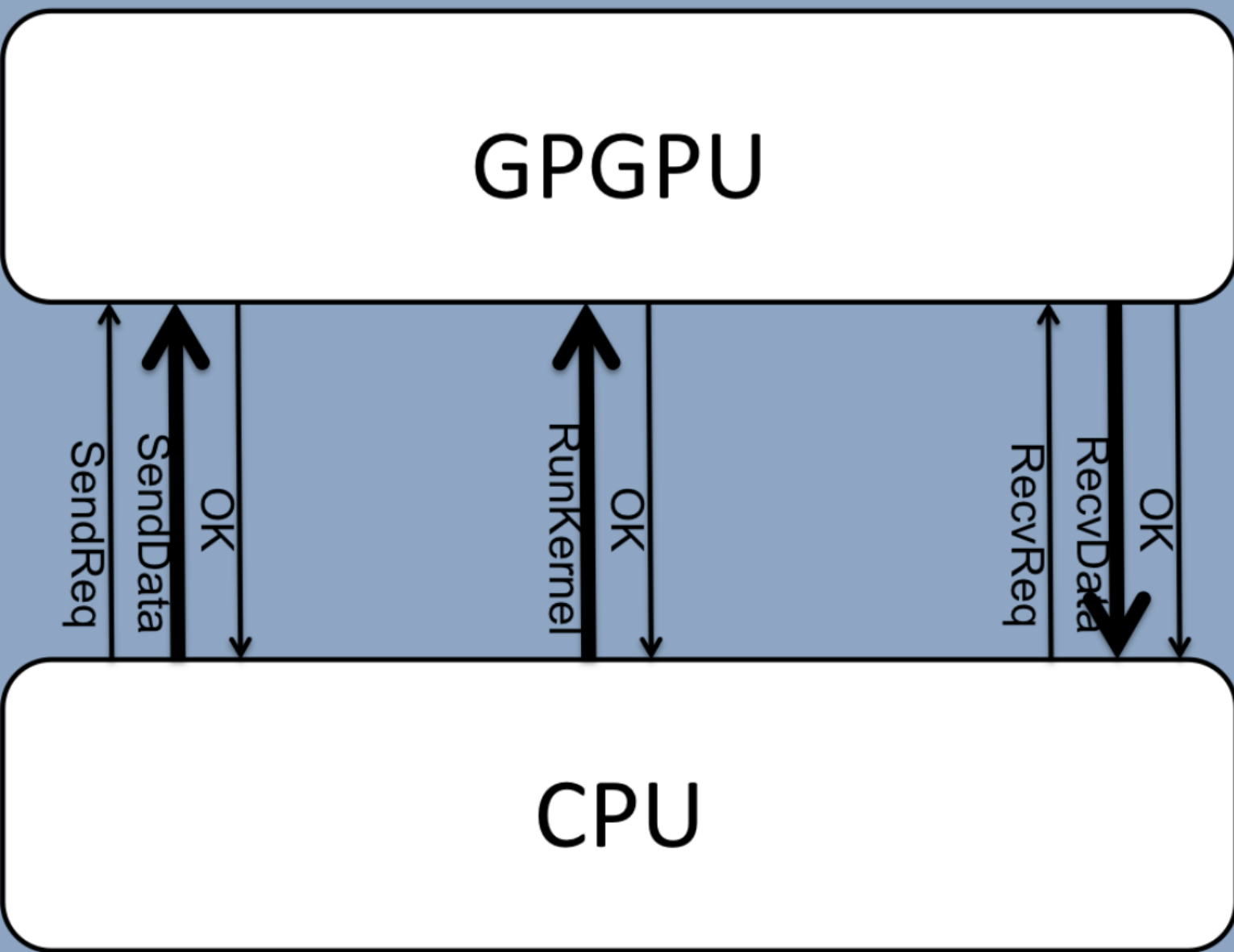
Local Dram ~60 ns

Remote Dram ~100 ns

North Bridge

GPU





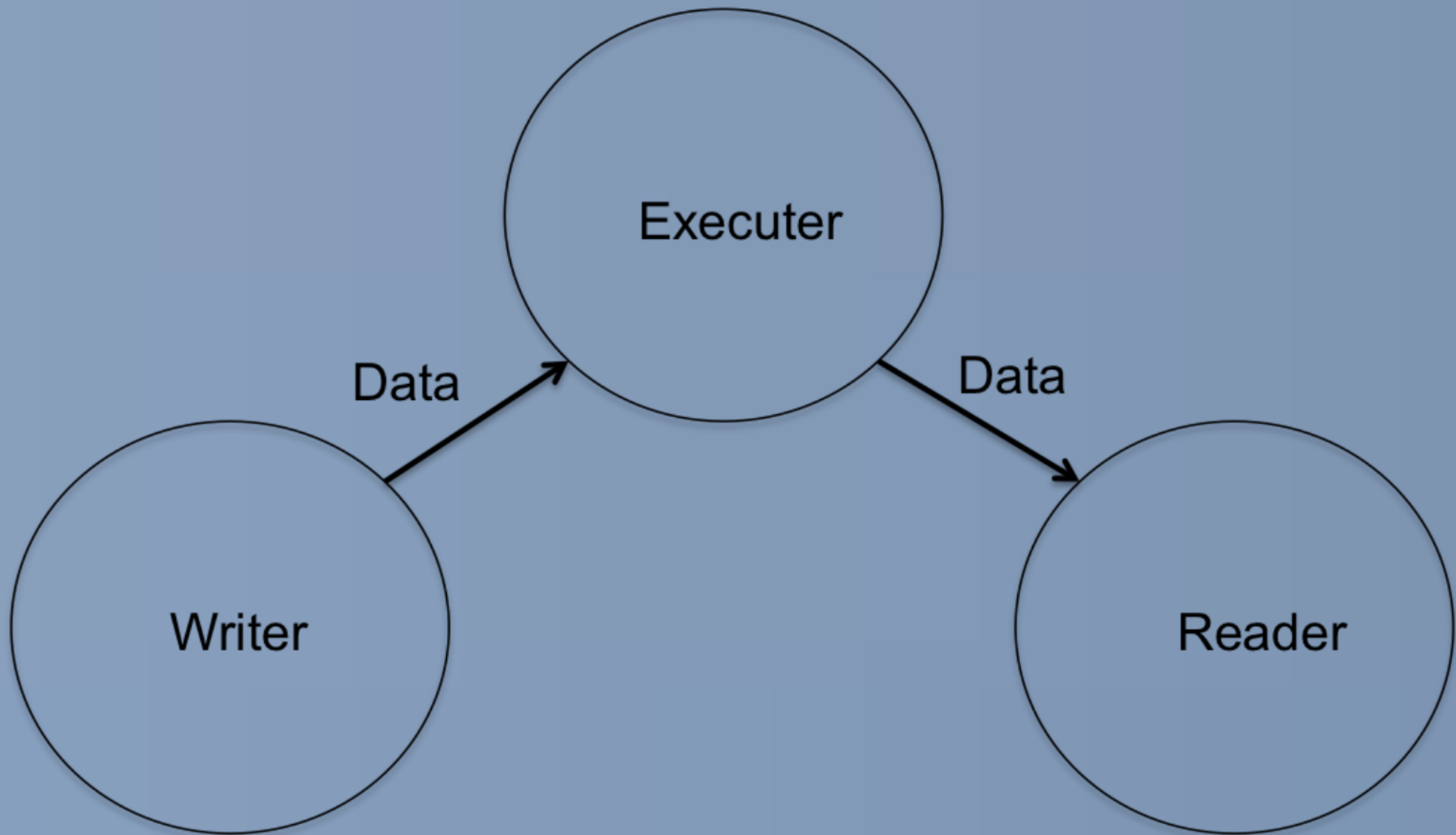
for all buffers:

W: SendReq
 SendData
 OK

E: RunKernel
 OK

R: ReadReq
 RecvData
 OK

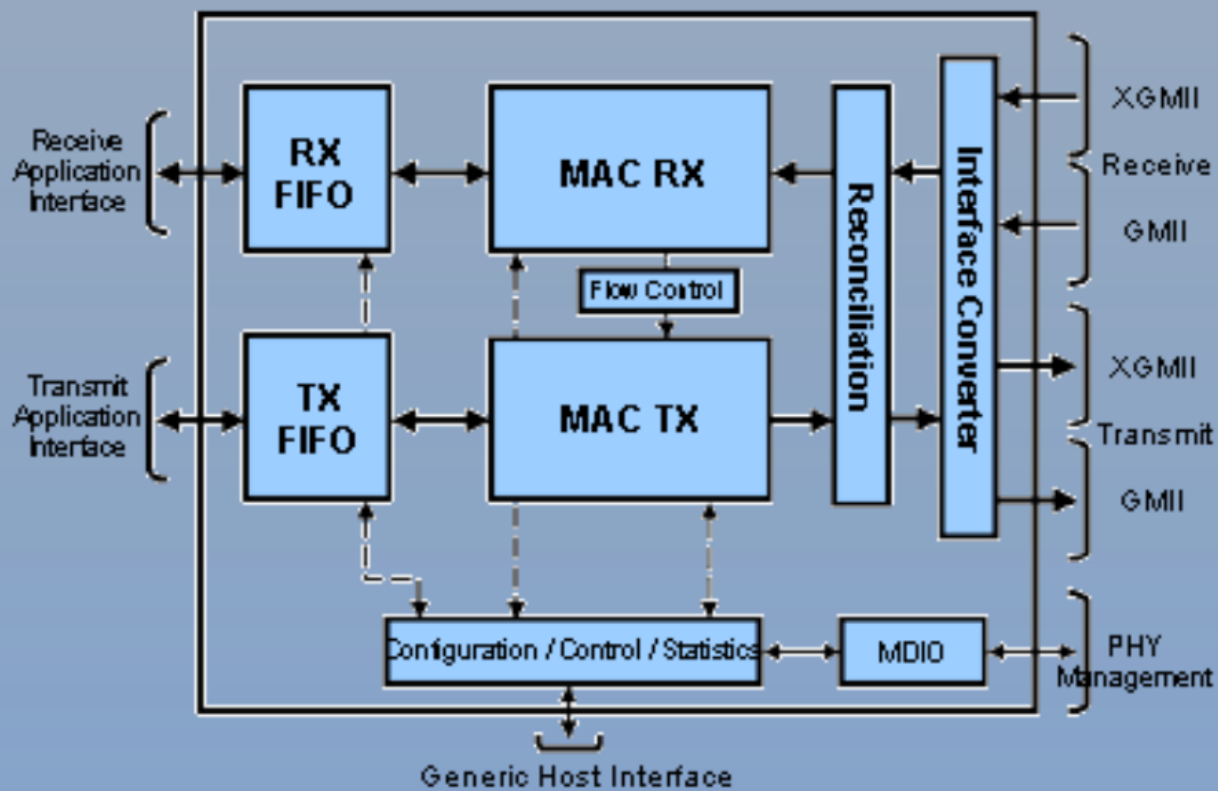
[W];[W||E];[W||E||R];...;[E||R];[R]



$[W]; [W||E]; [W||E||R]; \dots; [E||R]; [R]$

North Bridge

South Bridge

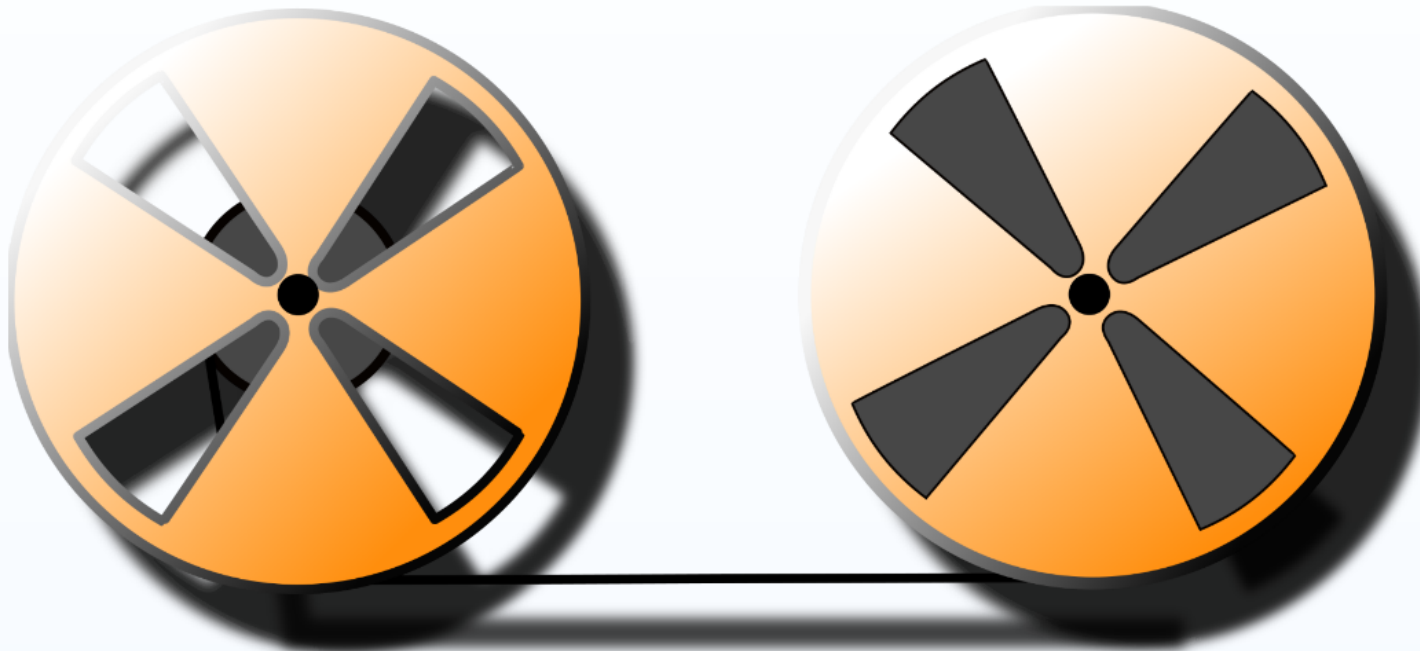




Disk Storage

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

Typically tens of PB disk

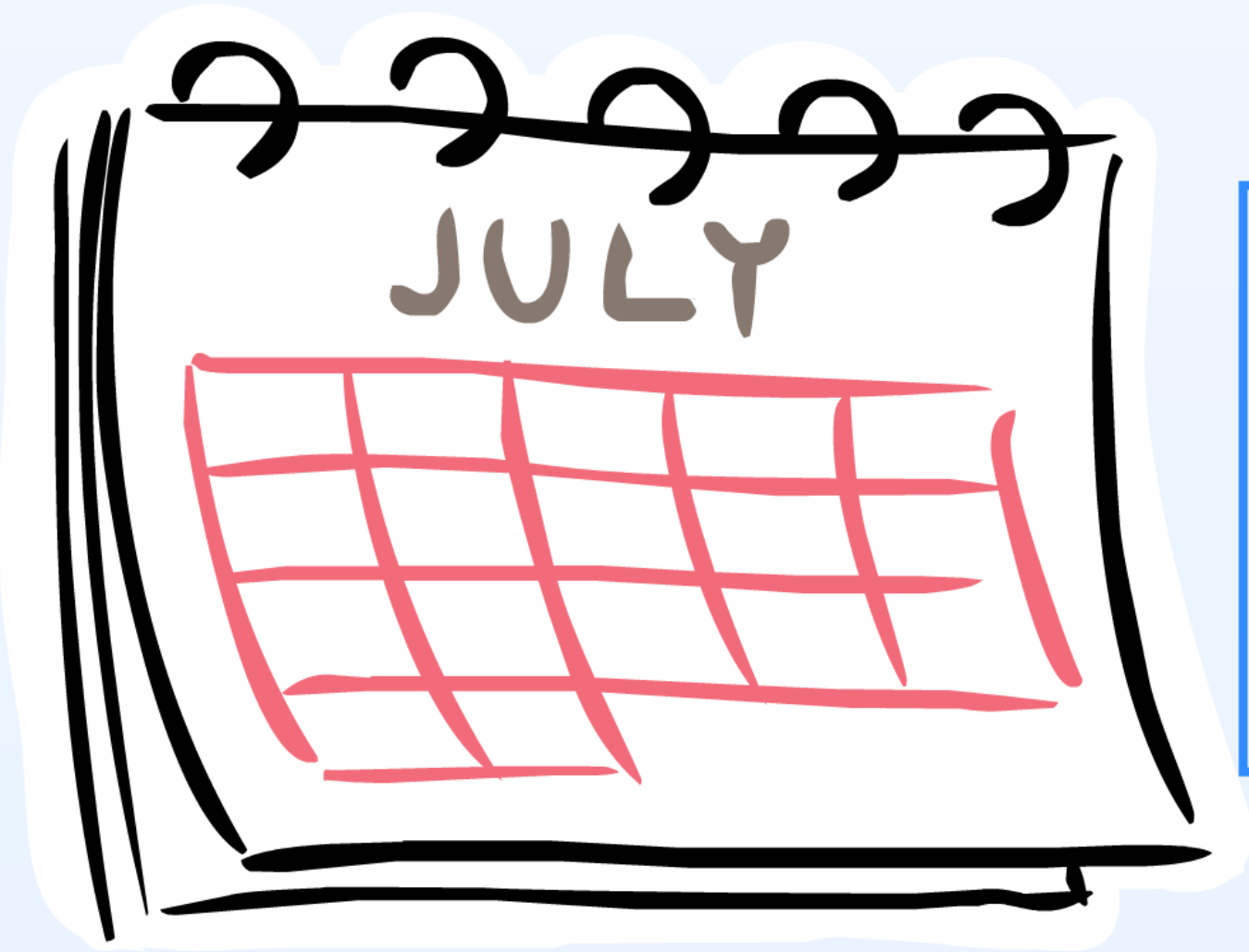


Tape Storage

Cheap mass storage for large datasets

Backup of important data

Typically tens of PB



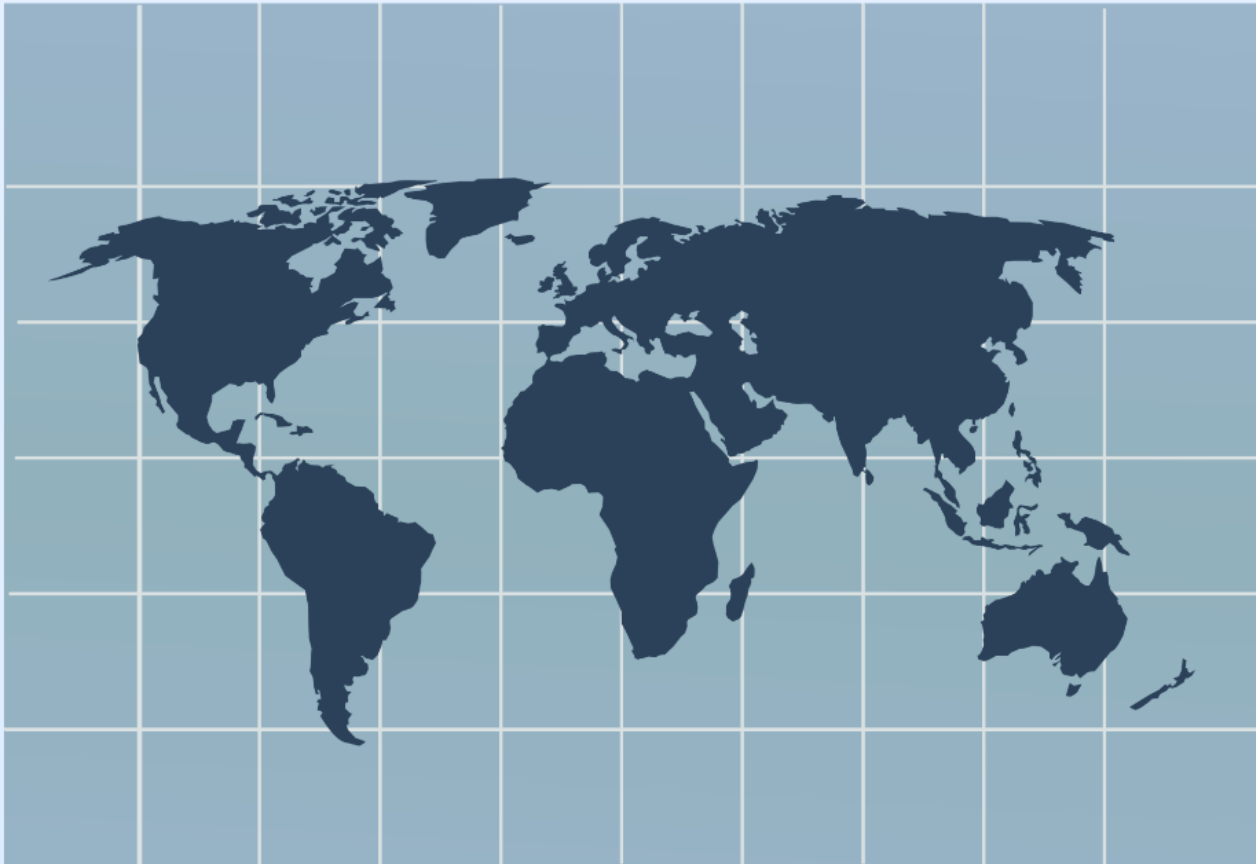
Scheduler

Receives input from

- Users
- Computers

Depends on

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



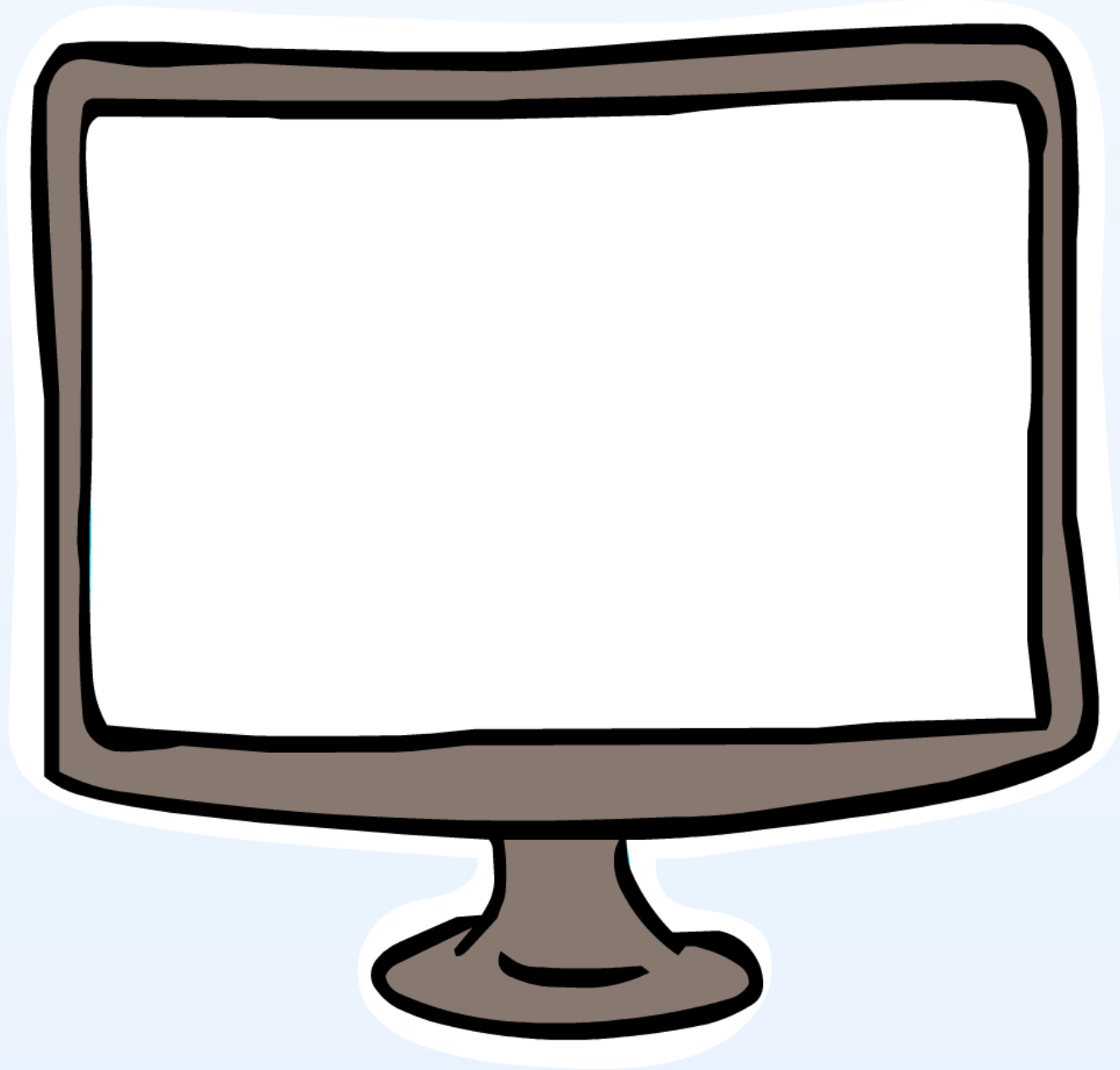
Grid

Receives input from

- External Schedulers
- Computers
- Tape

Depends on

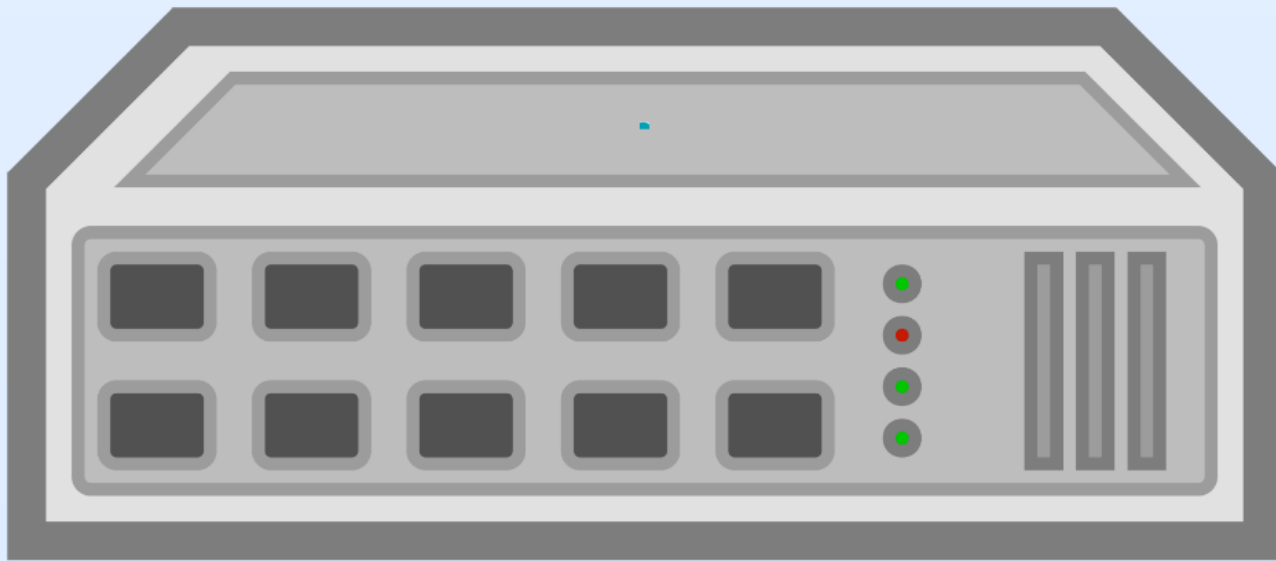
- Computers
- Staging from tape
- External schedulers



Visualization

Depends on

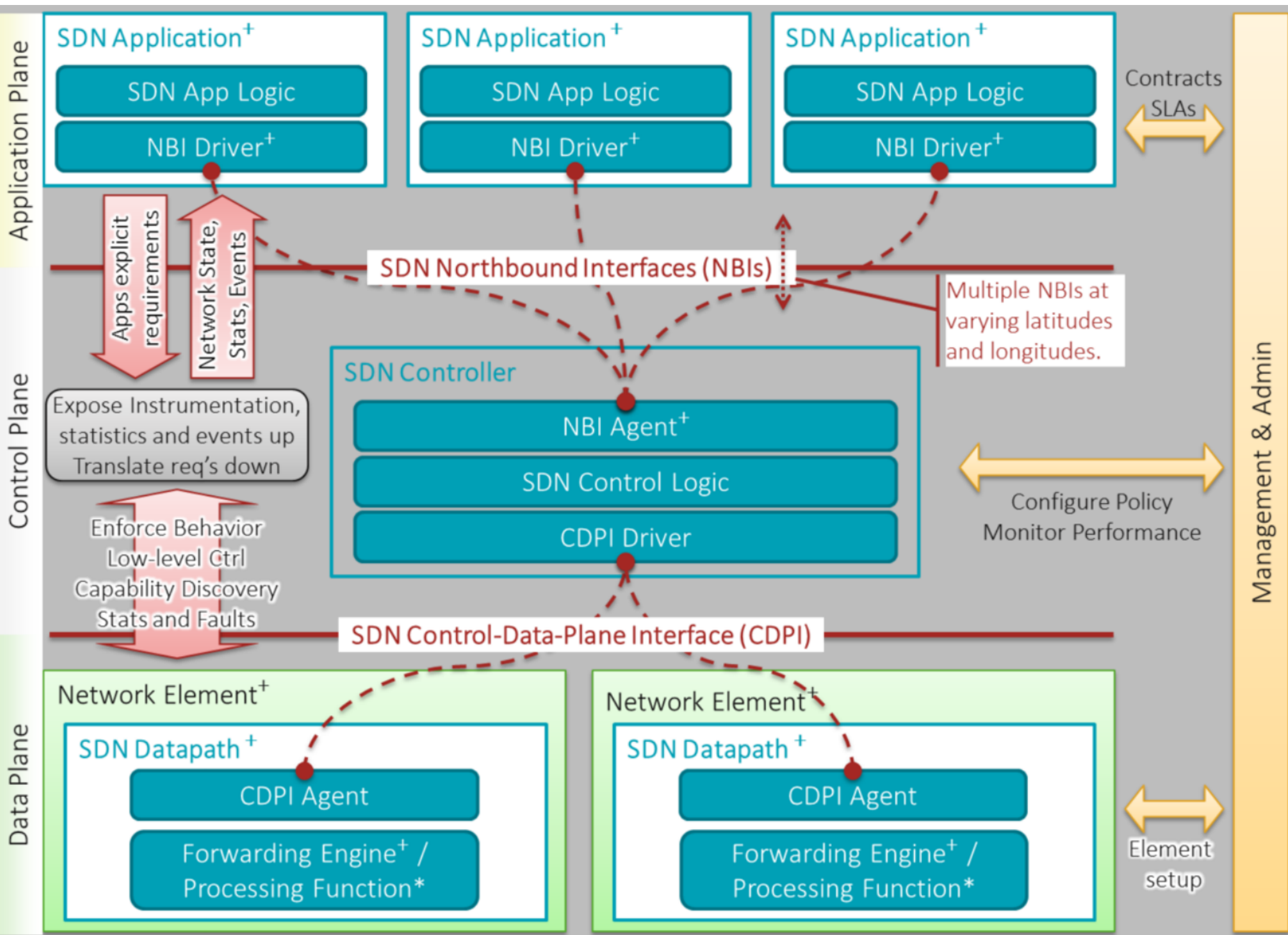
- Scheduler
- Computers
- Network



Gateway

100 Gbps primary connection

Hundreds of concurrent connections



⁺ indicates one or more instances | ^{*} indicates zero or more instances



User Communities

- File info simulation
- Task processing
- Data analysis
- Storage



Tape Storage

- High capacity
- Low cost per GB
- Long retention
- Slow access



Disk Storage

- Highly available for availability
- Available for recovery
- High capacity for large files
- High speed for file access
- Typically low cost per GB



Visualization

- Depends on:
 - Hardware
 - Connectivity
 - Network



Scheduler

- Provides Execution
 - Order
 - Priority
 - Resource
- Interacts with:
 - Computers
 - Storage
 - Network
 - Hardware
 - Software



Computers

- Typically 1:1 individual system
- 1MB - 1TB RAM
- High bandwidth and latency tolerance
- 100-10000 ops/sec
- Increasingly ultra-low cost accelerators (GPU, FPGA, etc.)



Gateway

- High speed network interface
- Network-to-network connectivity



Cloud

- Provides virtualized
 - Storage
 - Connectivity
 - Computers
 - Network
- Depends on:
 - Computers
 - Storage
 - Network
 - Hardware
 - Software

Open Platform Support to the Data Community

- Open standards for data formats
- Open standards for data access
- Open standards for data processing
- Open standards for data storage
- Open standards for data distribution
- Open standards for data security
- Open standards for data management
- Open standards for data integration
- Open standards for data analysis
- Open standards for data visualization
- Open standards for data sharing
- Open standards for data collaboration
- Open standards for data governance
- Open standards for data compliance
- Open standards for data privacy
- Open standards for data ethics
- Open standards for data transparency
- Open standards for data accountability
- Open standards for data responsibility
- Open standards for data stewardship
- Open standards for data leadership
- Open standards for data innovation
- Open standards for data excellence
- Open standards for data success
- Open standards for data achievement
- Open standards for data fulfillment
- Open standards for data realization
- Open standards for data attainment
- Open standards for data accomplishment
- Open standards for data success

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]