User Communities

All-hills simulations

Post-processing

Data analysis

Storage

Disk Storage

Ready

Input files for running jobs

Results files for viewing jobs

Files targeted for tape

Files staged from tape

Typically on one of the slow

Visualization

Depends on

- Scheduler
- Computers
- Filesystem

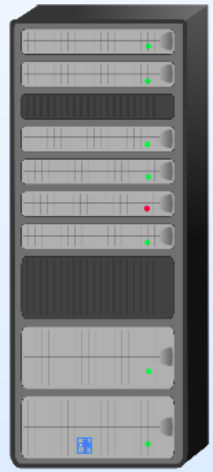
Scheduler

Receives input from

- Users
- Computers

Depends on

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



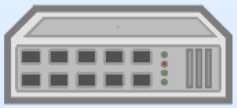
Computers

Typically 1-5 exaflop/s systems

10K - 1M Cores

High Bandwidth, Low Latency Interconnect (2-5 TB/sec and 0.5 us)

Increasingly often uses accelerators (GPGPUs, FPGAs, etc)



Gateway

100 Gbps primary connection

Hundreds of concurrent connections

Grid

Receives input from

- External Schedulers
- Computers
- Tape

Depends on

- Computers
- Staging from tape
- External schedulers

User Communities

All-hills simulations

Post-processing

Data analysis

Storage

Disk Storage

Ready

Input files for running jobs

Results files for viewing jobs

Files targeted for tape

Files staged from tape

Typically on one of the slow

Visualization

Depends on

- Scheduler
- Computers
- Filesystem

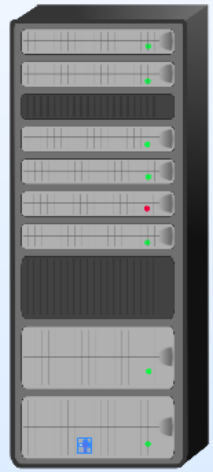
Scheduler

Receives input from

- Users
- Computers

Depends on

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



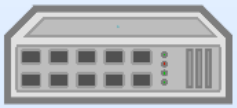
Computers

Typically 1-5 exaflop/s systems

10K - 1M Cores

High Bandwidth, Low Latency Interconnect (2-5 TB/sec and 0.5 us)

Increasingly often uses accelerators (GPGPUs, FPGAs, etc)



Gateway

100 Gbps primary connection

Hundreds of concurrent connections

Grid

Receives input from

- External Schedulers
- Computers
- Tape

Depends on

- Computers
- Staging from tape
- External schedulers

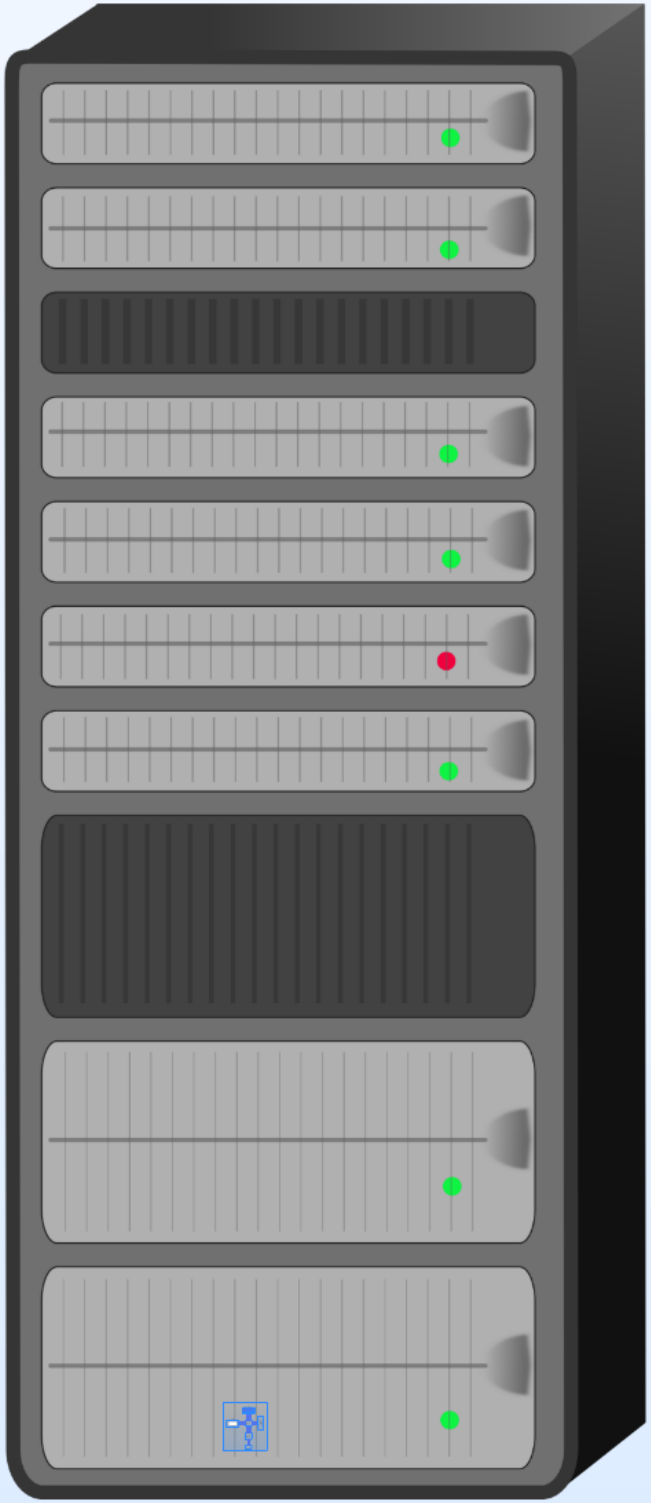
User Communities

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)

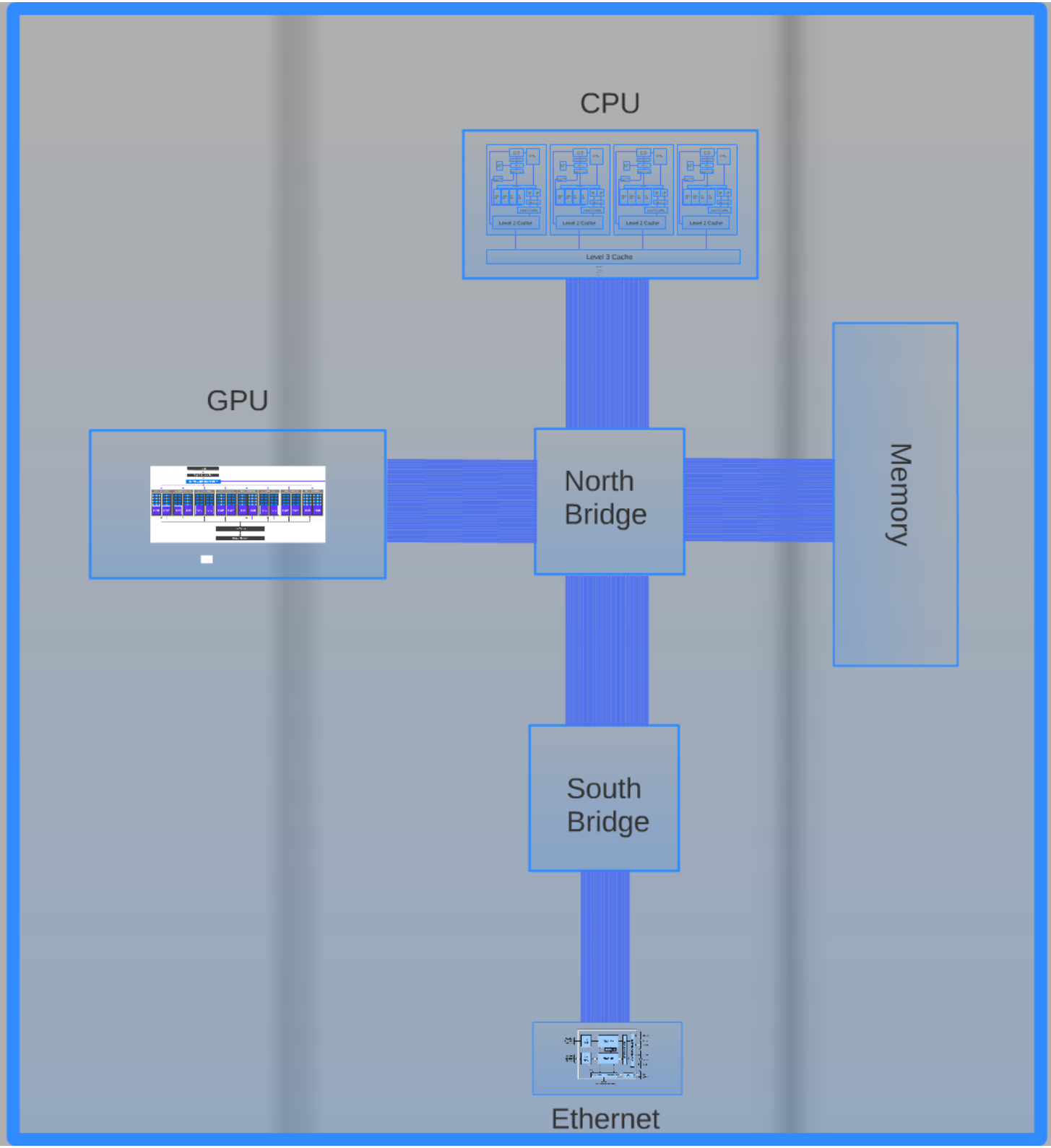
Computers

Typically 1-5 individual systems

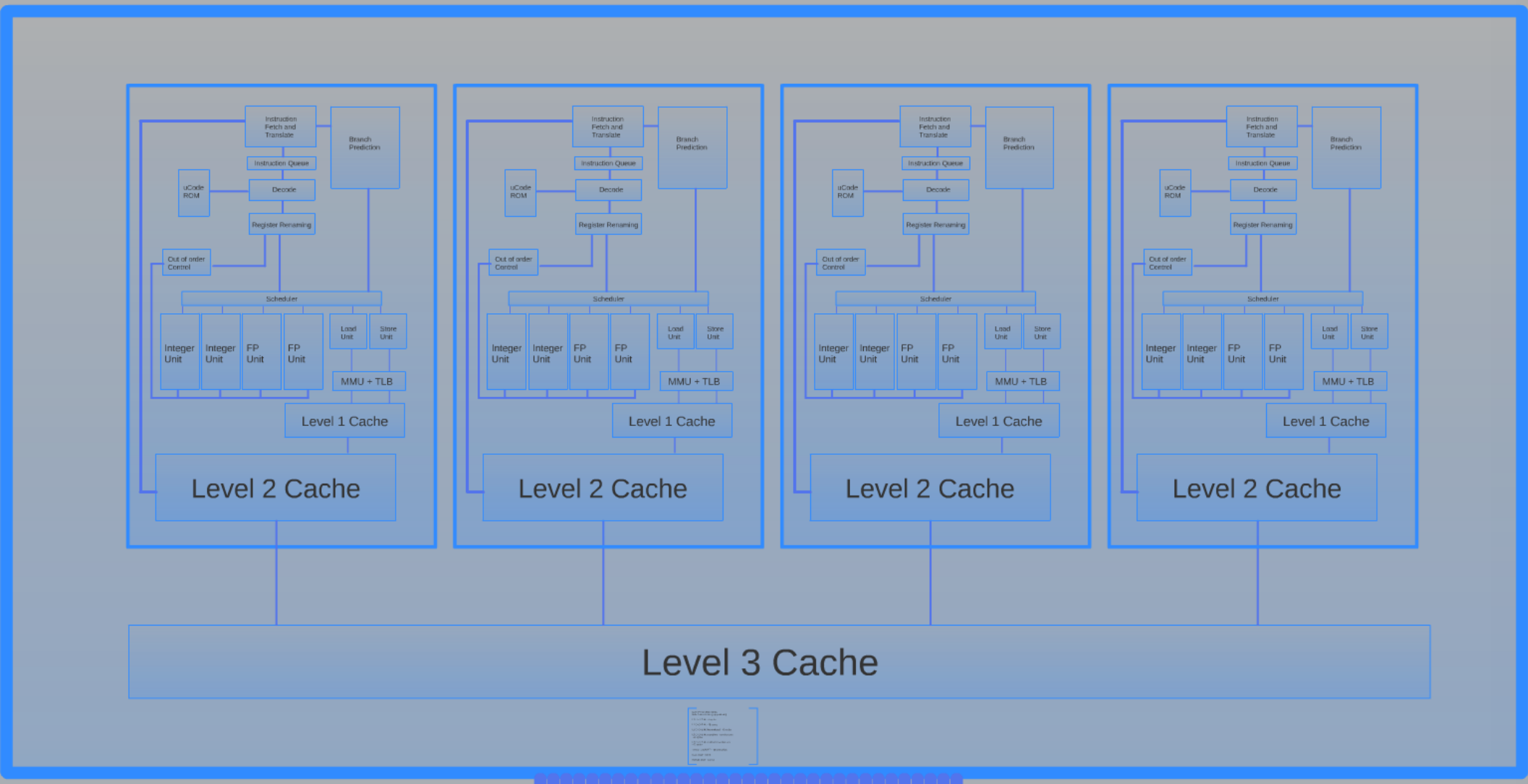
10K - 1M Cores

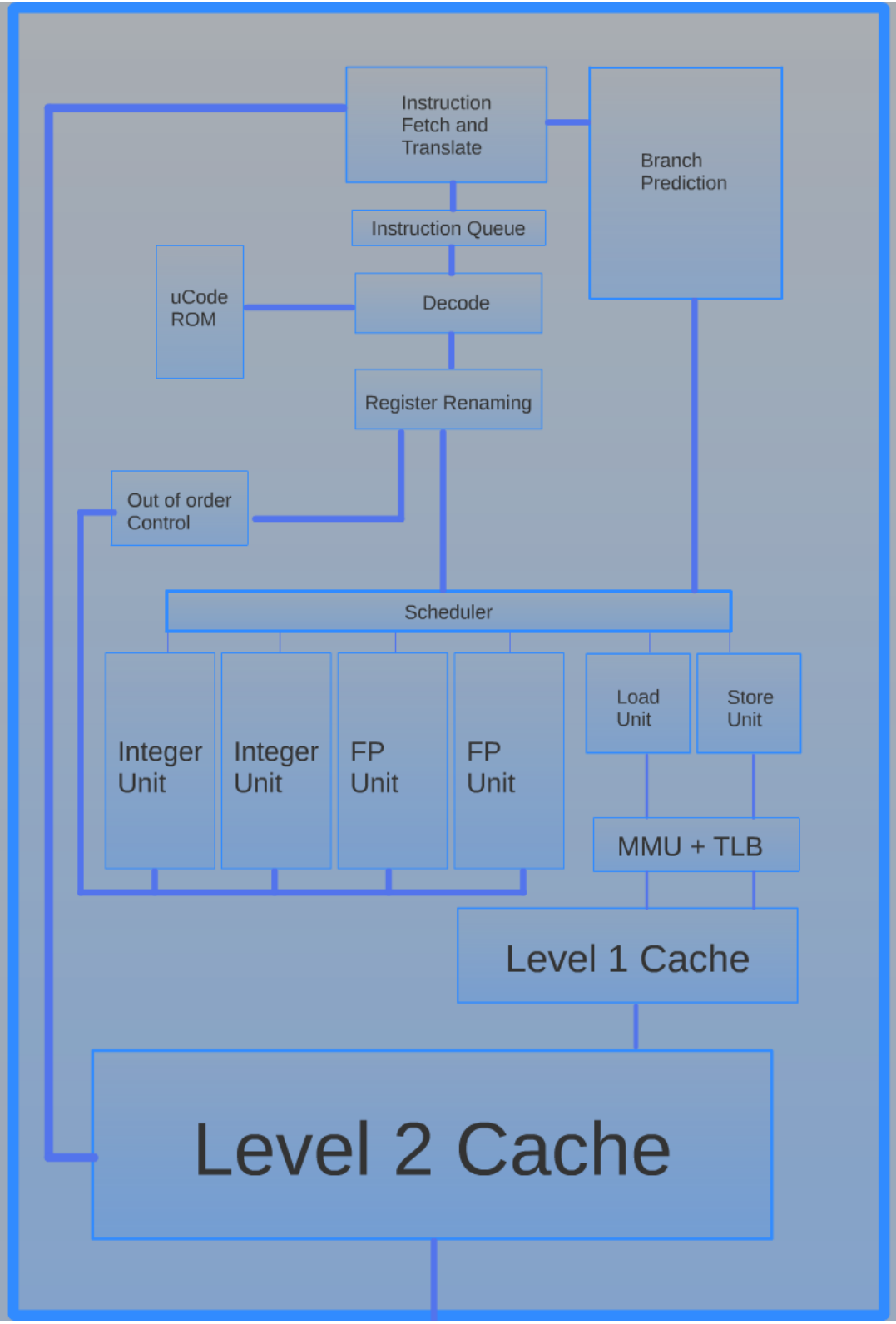
High Bandwidth Low Latency Interconnect (22 GB/sec and 0.89us)

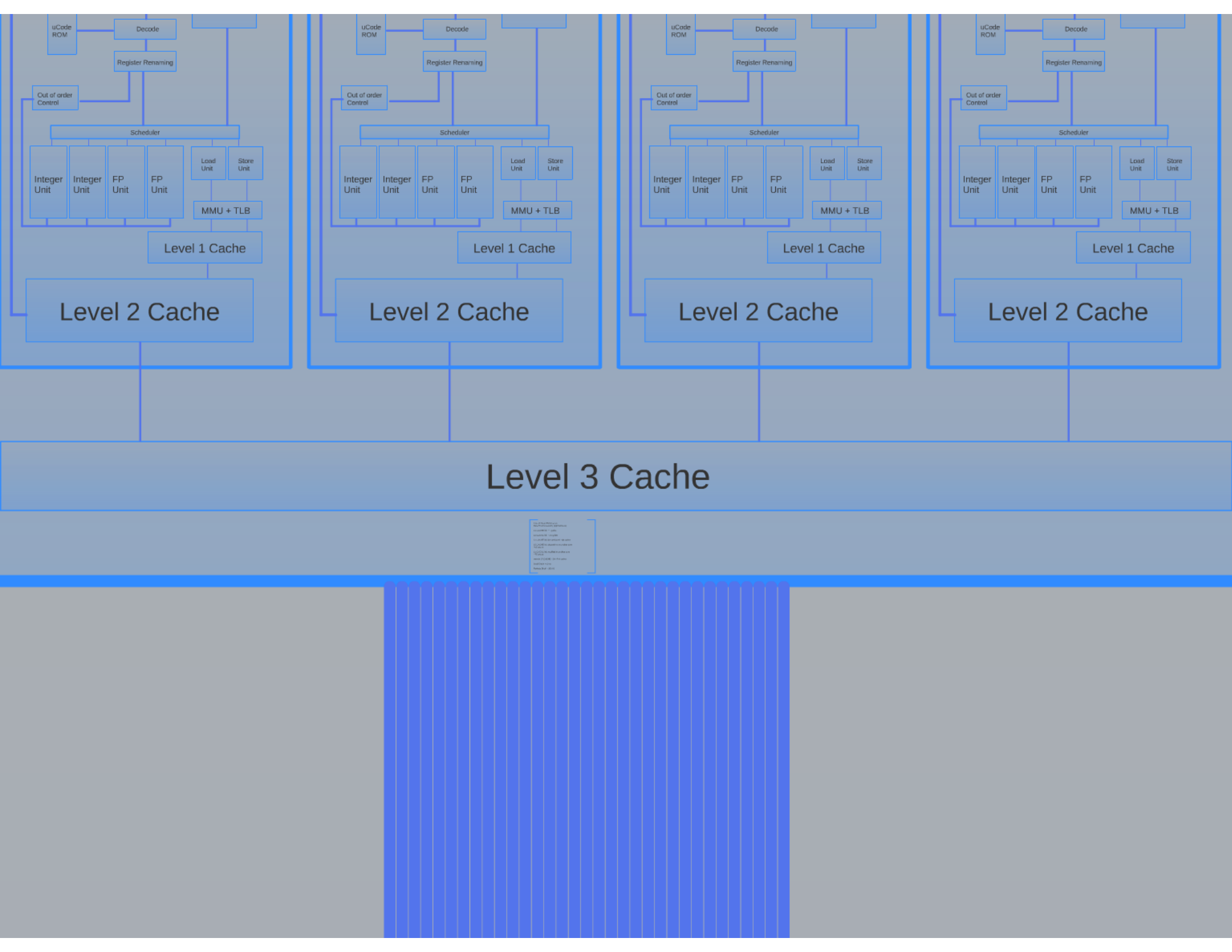
Increasingly often uses accelerators (GPGPUs MICs etc)


CPU
Level 2 Cache
Level 2 Cache
Level 2 Cache
Level 2 Cache
Level 3 Cache
GPU
North Bridge
Memory
South Bridge
Ethernet

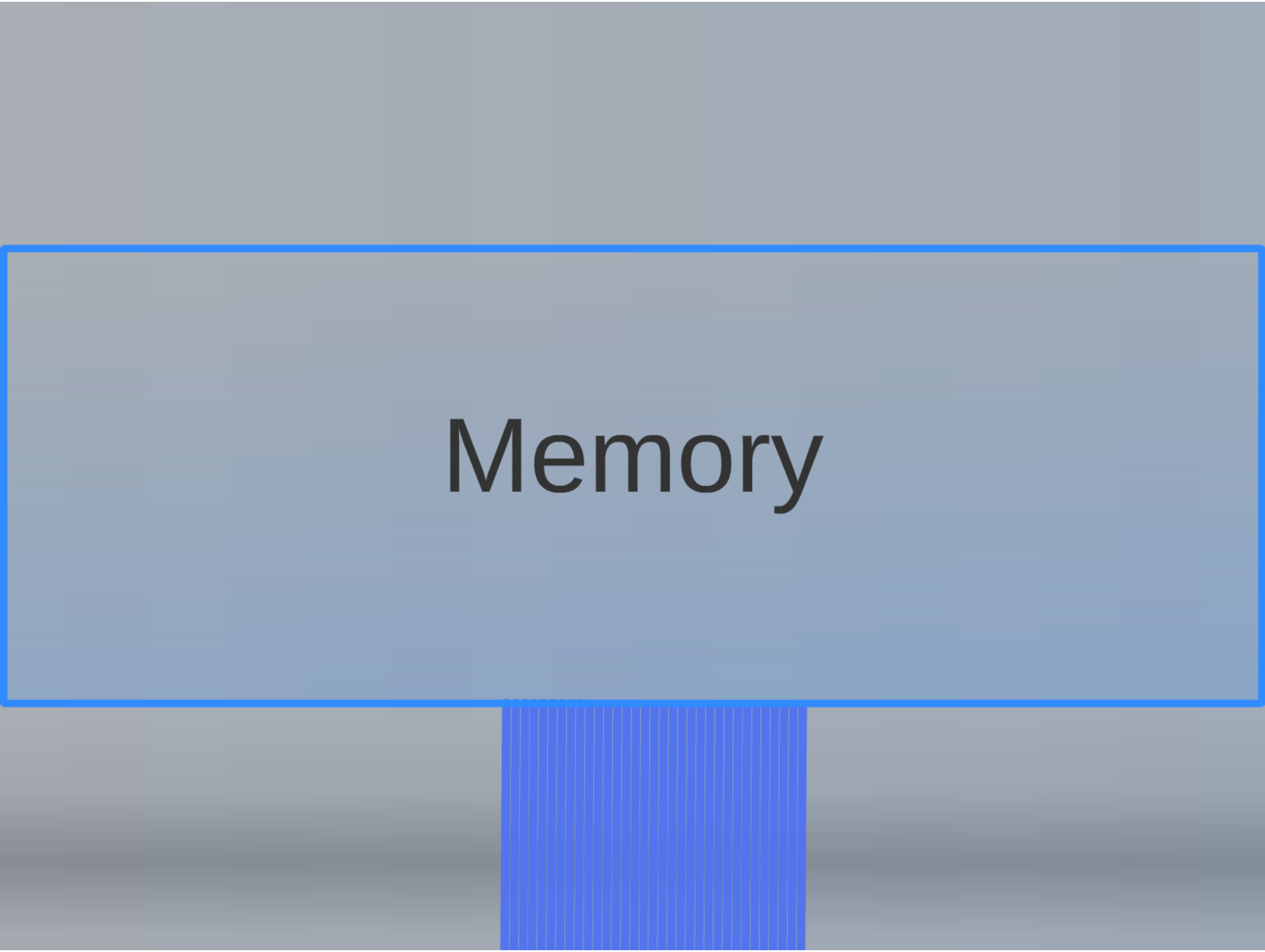
CPU


Instruction Fetch and Translate
Branch Prediction
Instruction Queue
uCode ROM
Decode
Register Renaming
Out of order Control
Scheduler
Integer Unit
Integer Unit
FP Unit
FP Unit
Load Unit
Store Unit
MMU + TLB
Level 1 Cache
Level 2 Cache
Level 3 Cache


Instruction Fetch and Translate
Branch Prediction
Instruction Queue
uCode ROM
Decode
Register Renaming
Out of order Control
Scheduler
Integer Unit
Integer Unit
FP Unit
FP Unit
Load Unit
Store Unit
MMU + TLB
Level 1 Cache
Level 2 Cache

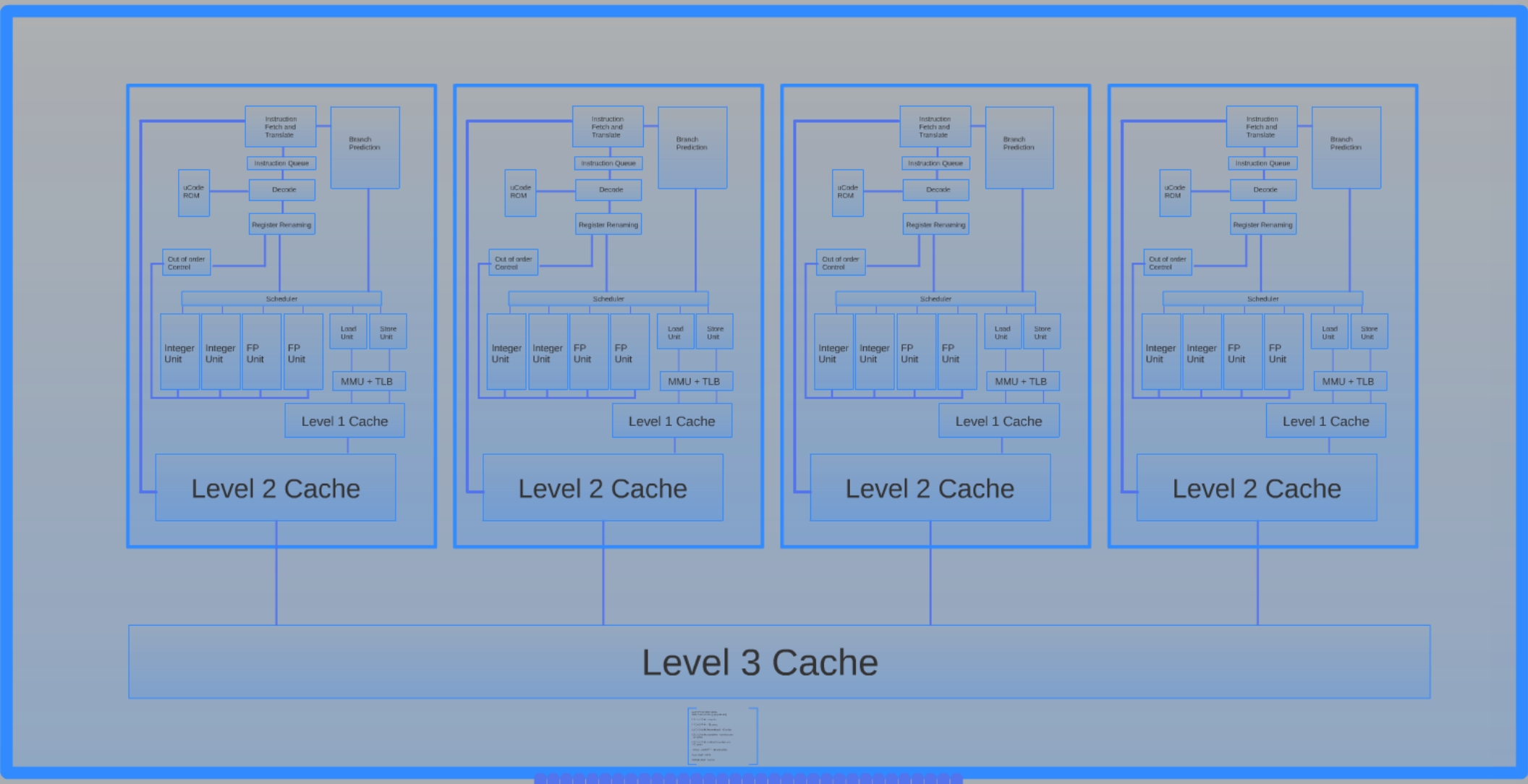

uCode ROM
Decode
Register Renaming
Out of order Control
Scheduler
Integer Unit
Integer Unit
FP Unit
FP Unit
Load Unit
Store Unit
MMU + TLB
Level 1 Cache
Level 2 Cache
Level 3 Cache


North
Bridge


Memory


North
Bridge

CPU


Instruction Fetch and Translate
Branch Prediction
Instruction Queue
uCode ROM
Decode
Register Renaming
Out of order Control
Scheduler
Integer Unit
Integer Unit
FP Unit
FP Unit
Load Unit
Store Unit
MMU + TLB
Level 1 Cache
Level 2 Cache
Level 3 Cache

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

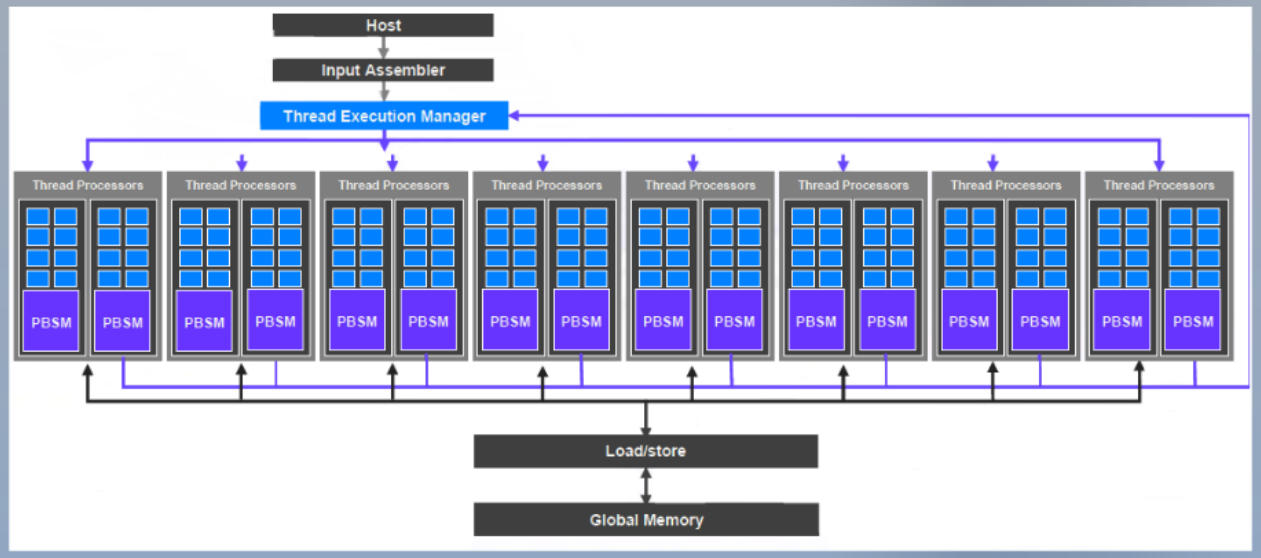
remote L3 CACHE ~100-300 cycles

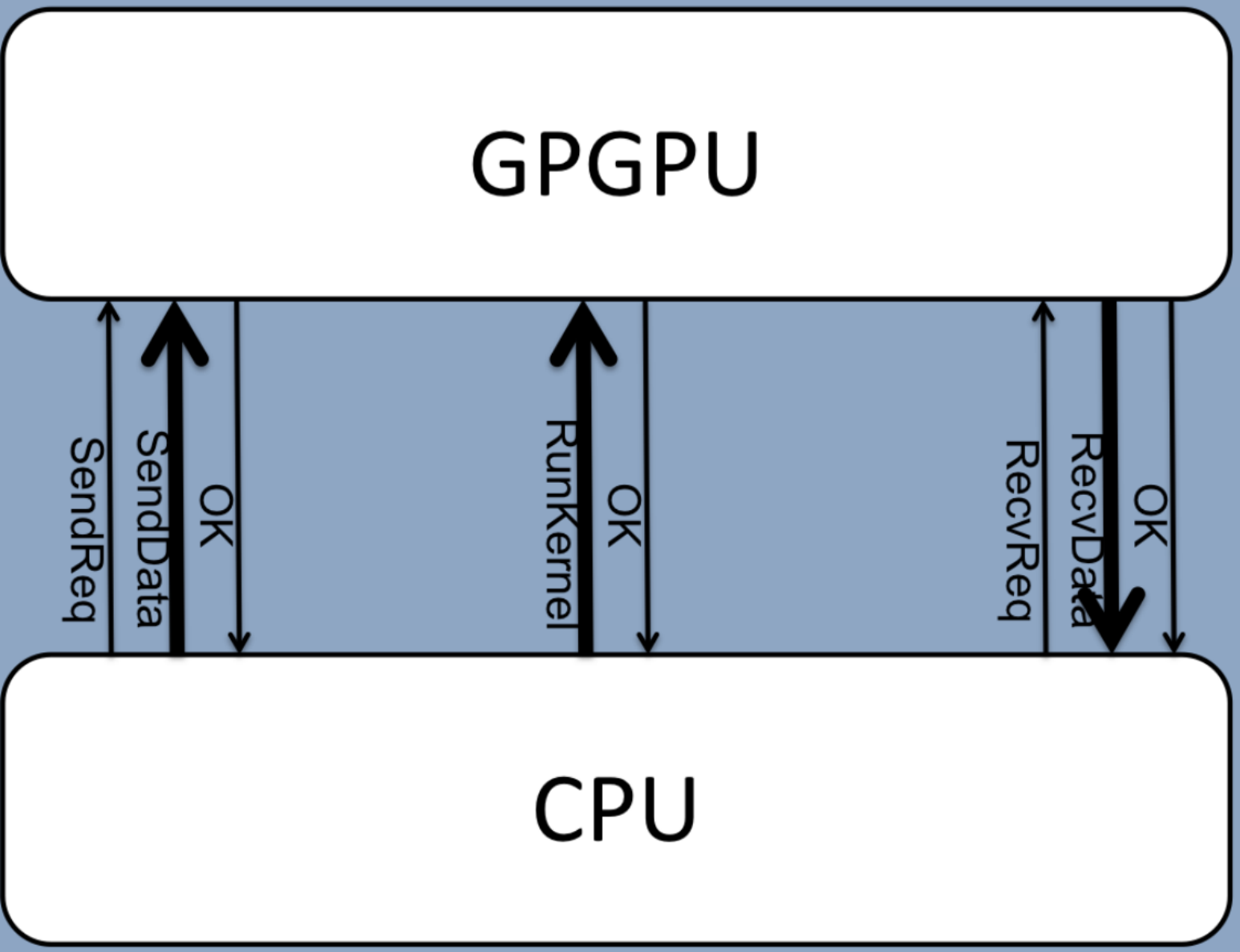
Local Dram ~60 ns

Remote Dram ~100 ns


North
Bridge

GPU


Host
Input Assembler
Thread Execution Manager
Thread Processors
PBSM
Load/store
Global Memory



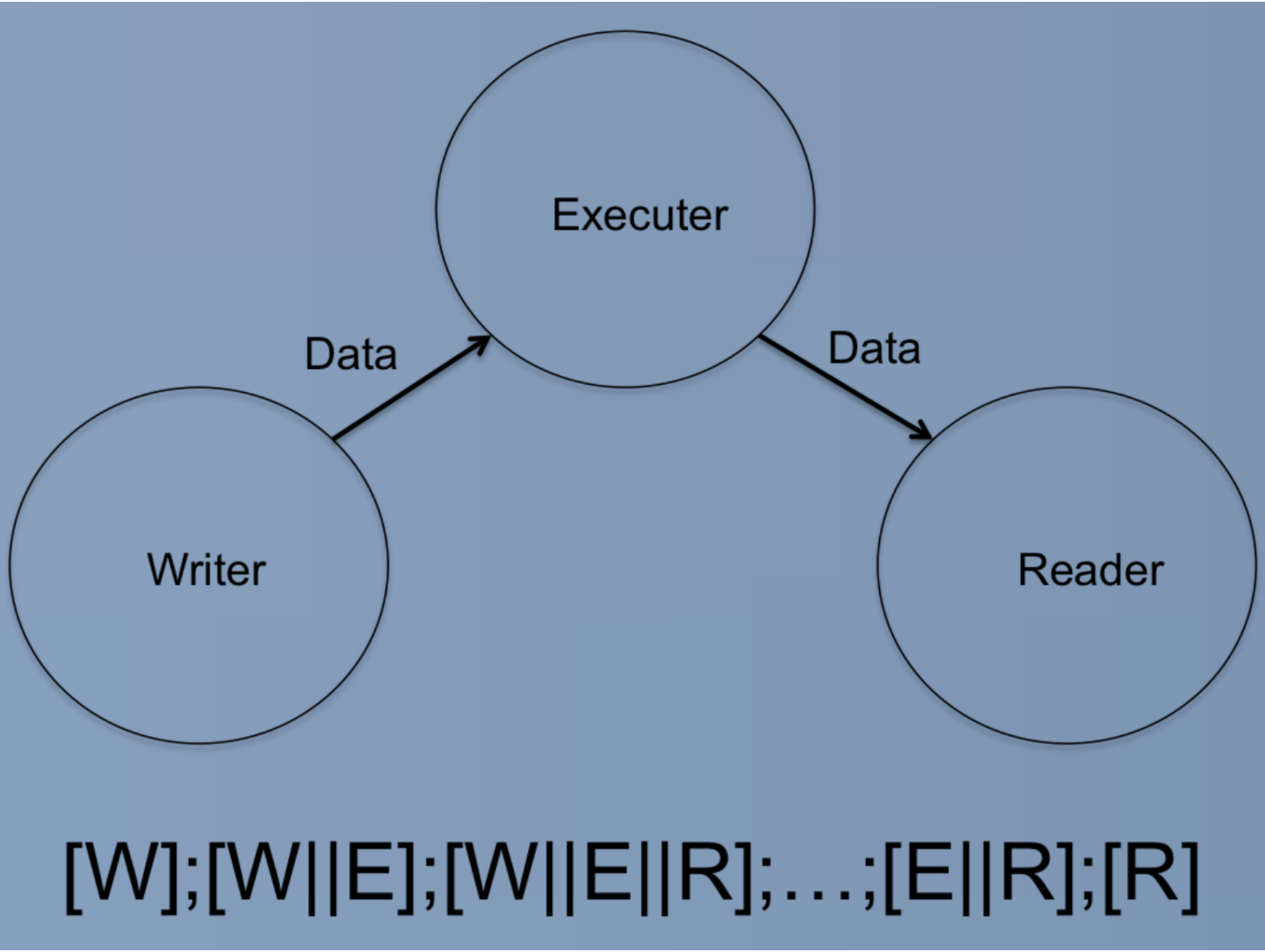
for all buffers:

W: SendReq
SendData
OK

E: RunKernel
OK

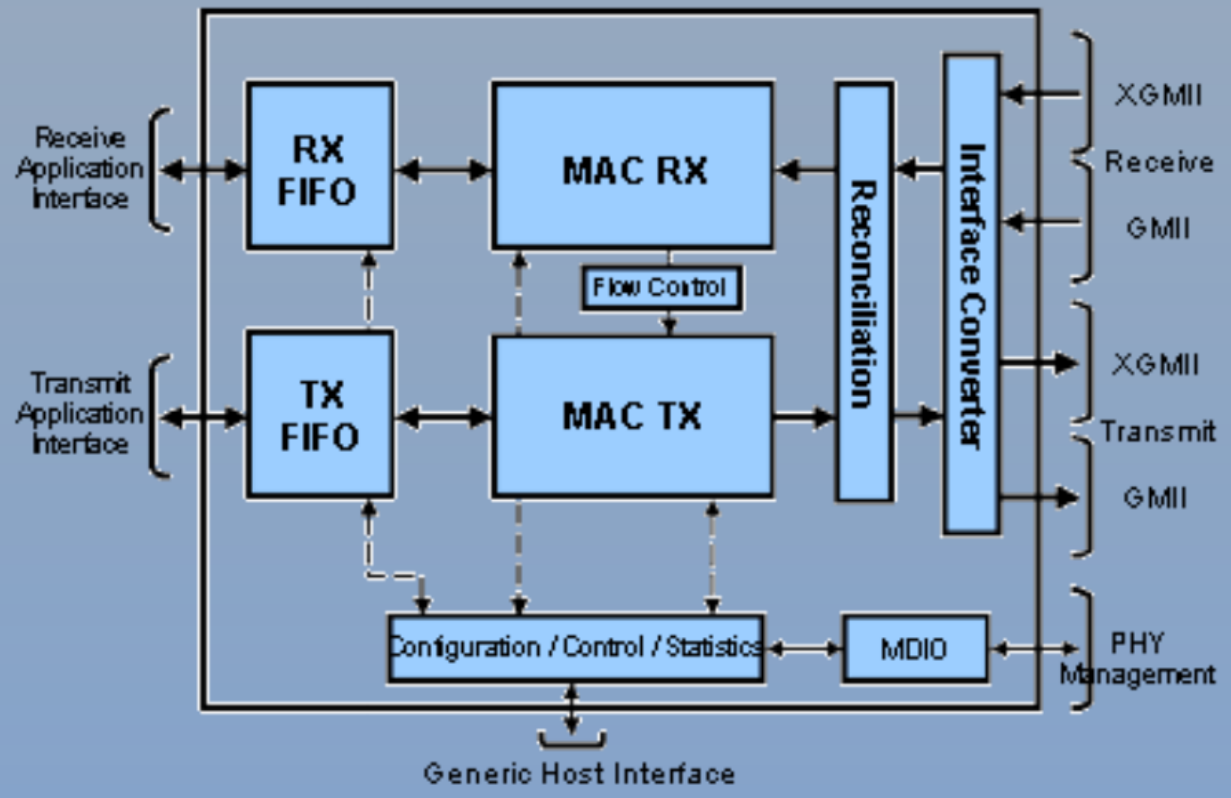
R: ReadReq
RecvData
OK

[W];[W||E];[W||E||R];…;[E||R];[R]


Executer
Data
Data
Writer
Reader
[W];[W||E];[W||E||R];…;[E||R];[R]


North
Bridge


South
Bridge

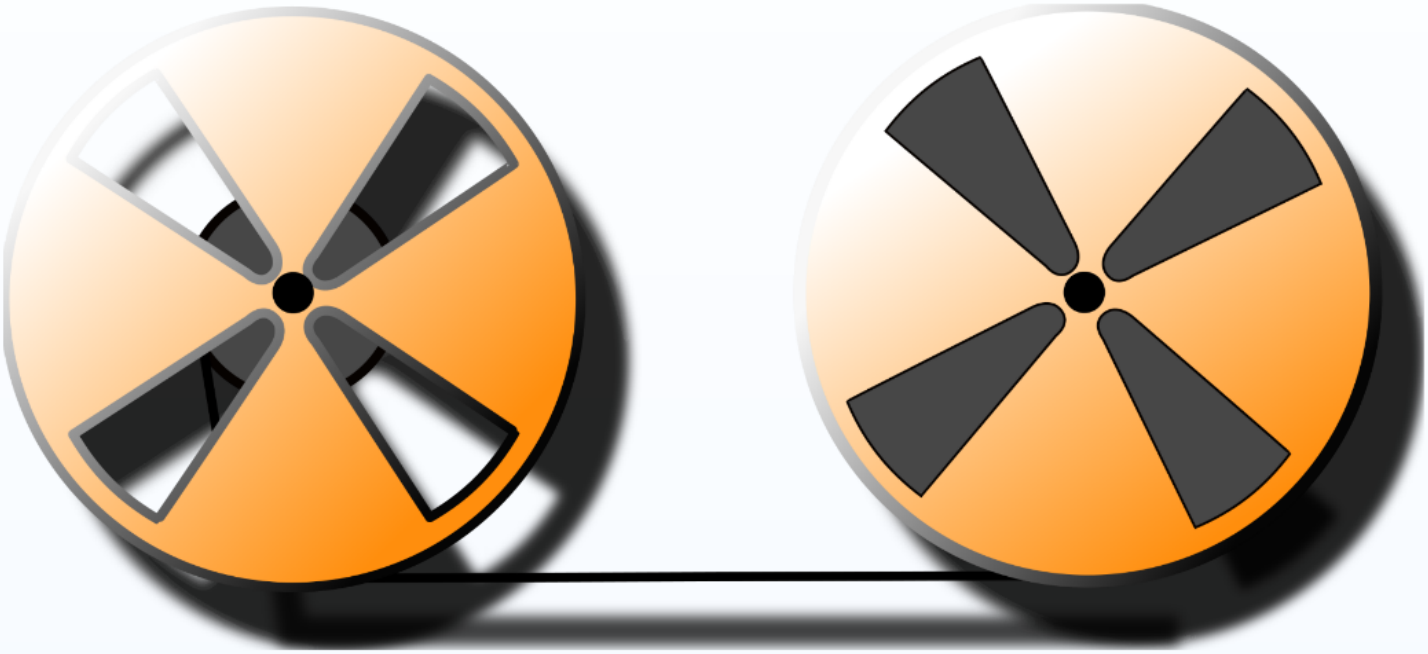

Receive Application Interface
RX FIFO
MAC RX
Reconciliation
Interface Converter
XGMII
Receive
GMII
Flow Control
Transmit Application Interface
TX FIFO
MAC TX
XGMII
Transmit
GMII
Configuration / Control / Statistics
MDIO
PHY Management
Generic Host Interface

Ethernet

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

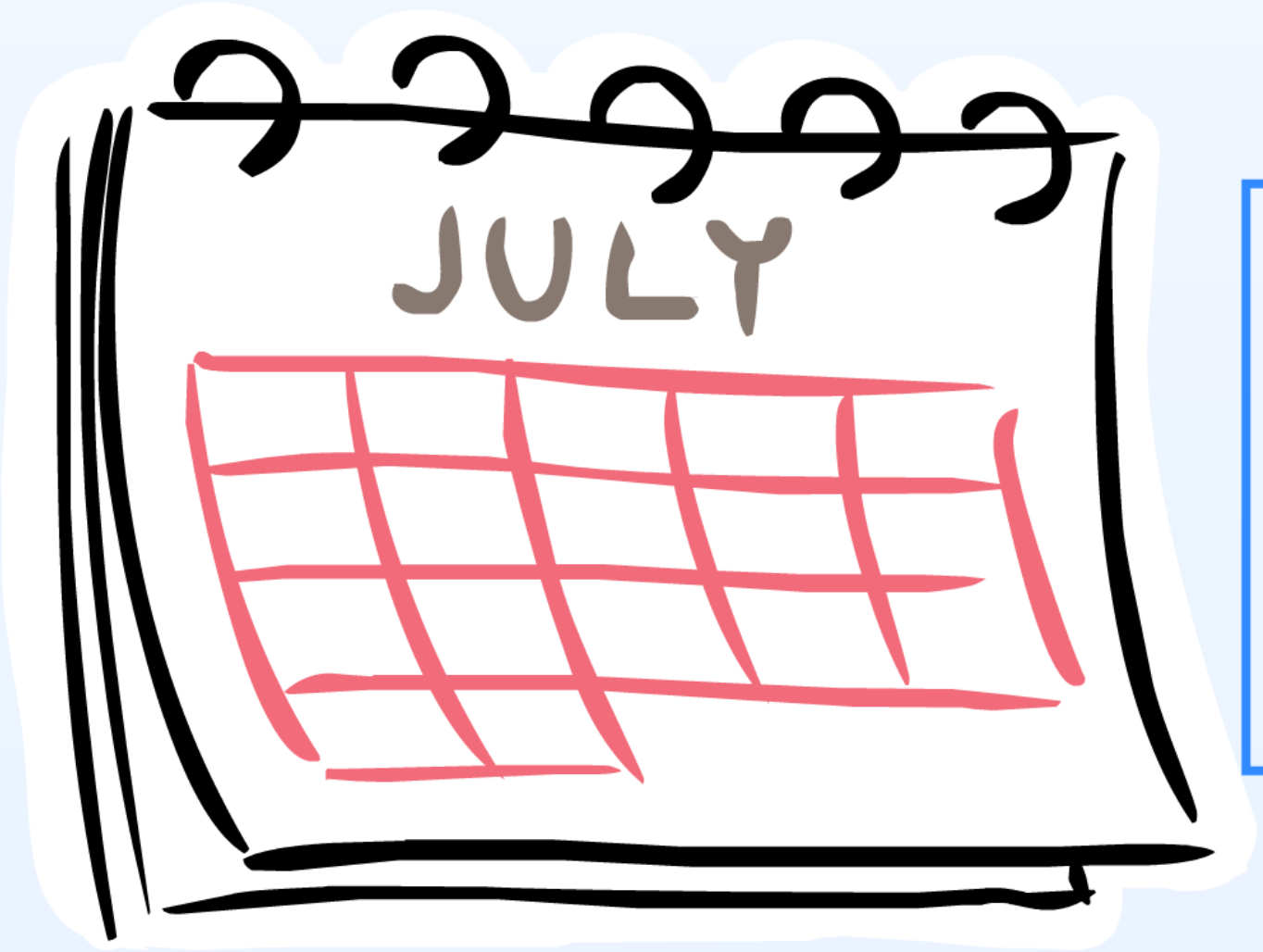


Tape Storage

Cheap mass storage for large datasets

Backup of important data

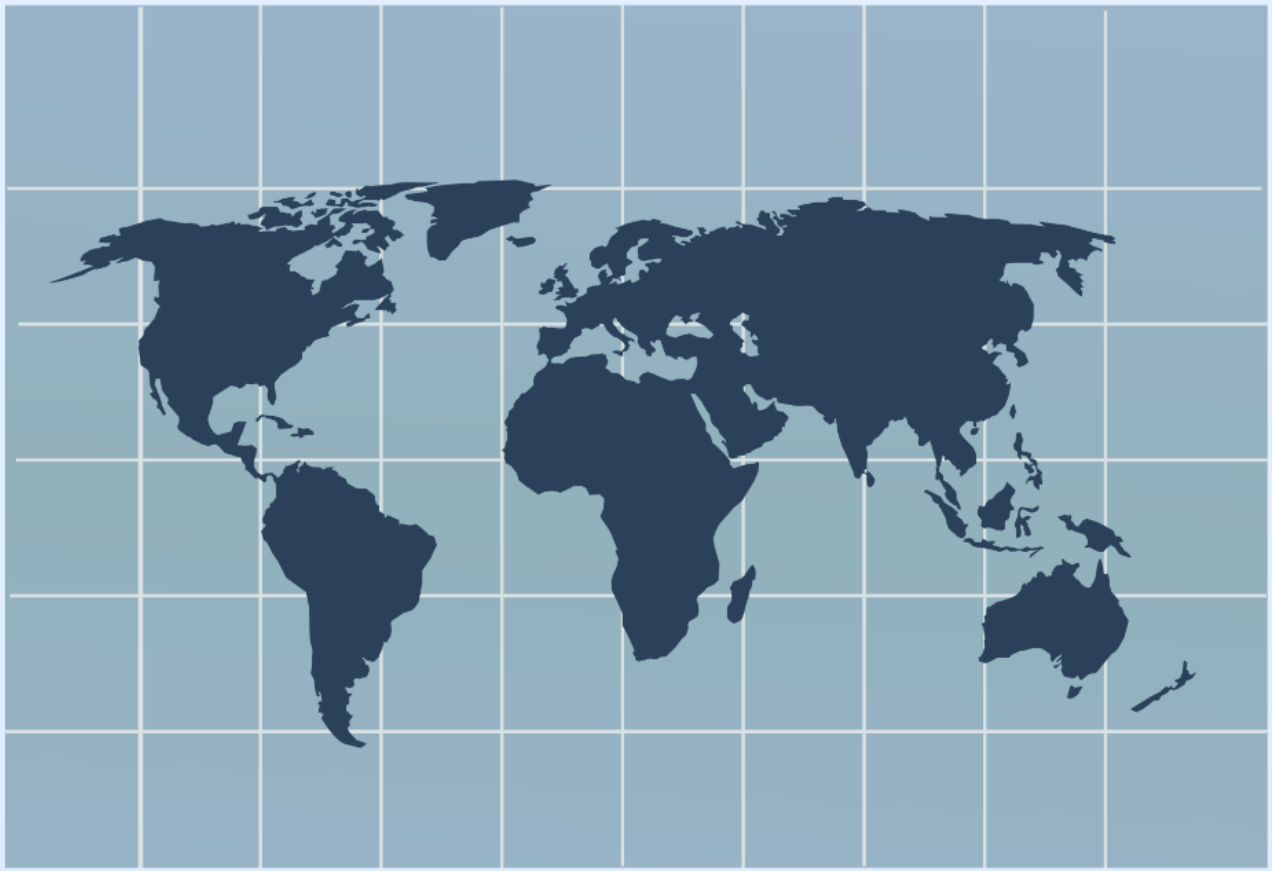
Typically tens of PB



Scheduler

Receives input from
- Users
- Computers

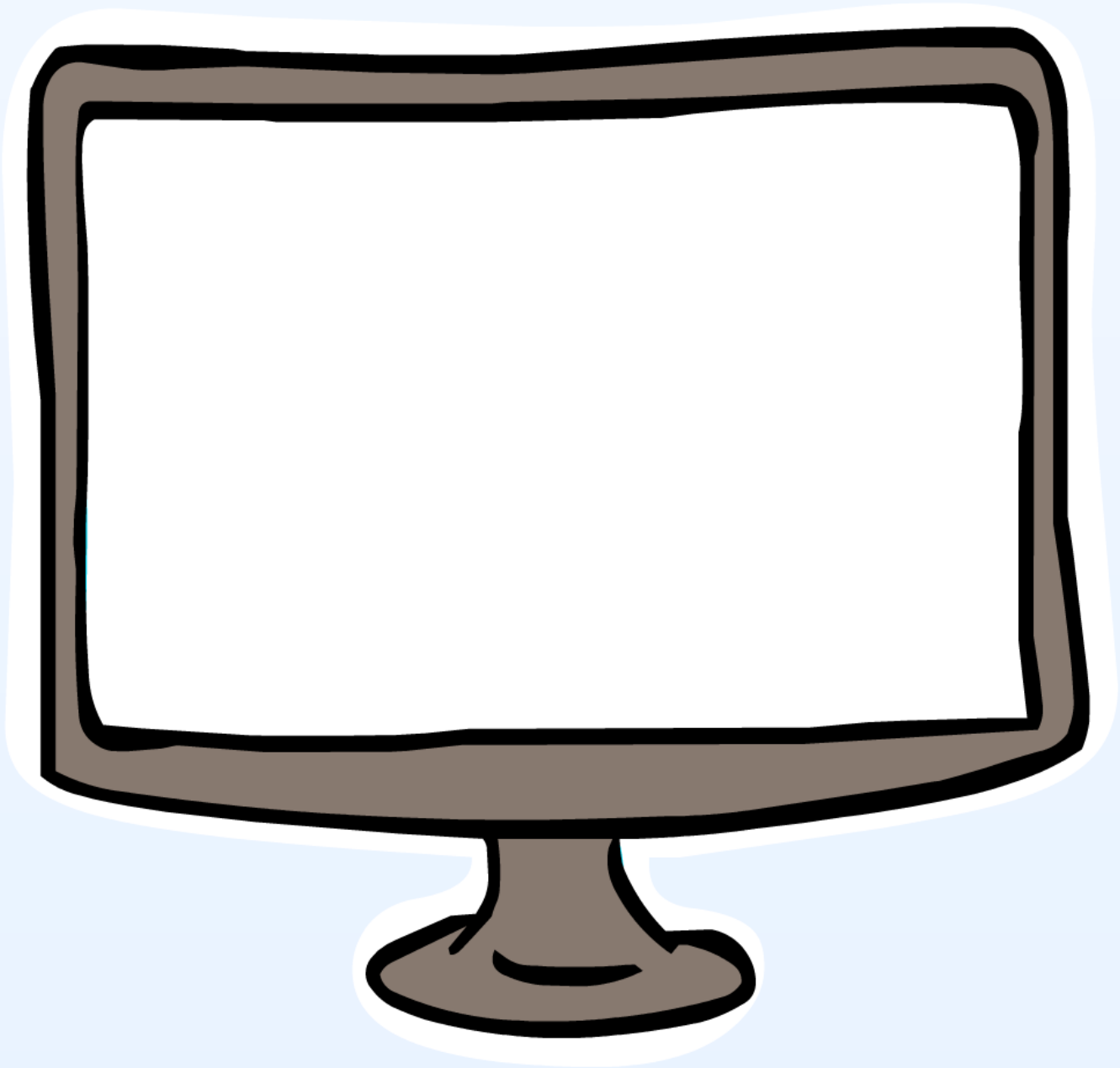
Depends on
- Computers
- Staging from tape
- Real-time reservations
- Grid



Grid

Receives input from
- External Schedulers
- Computers
- Tape

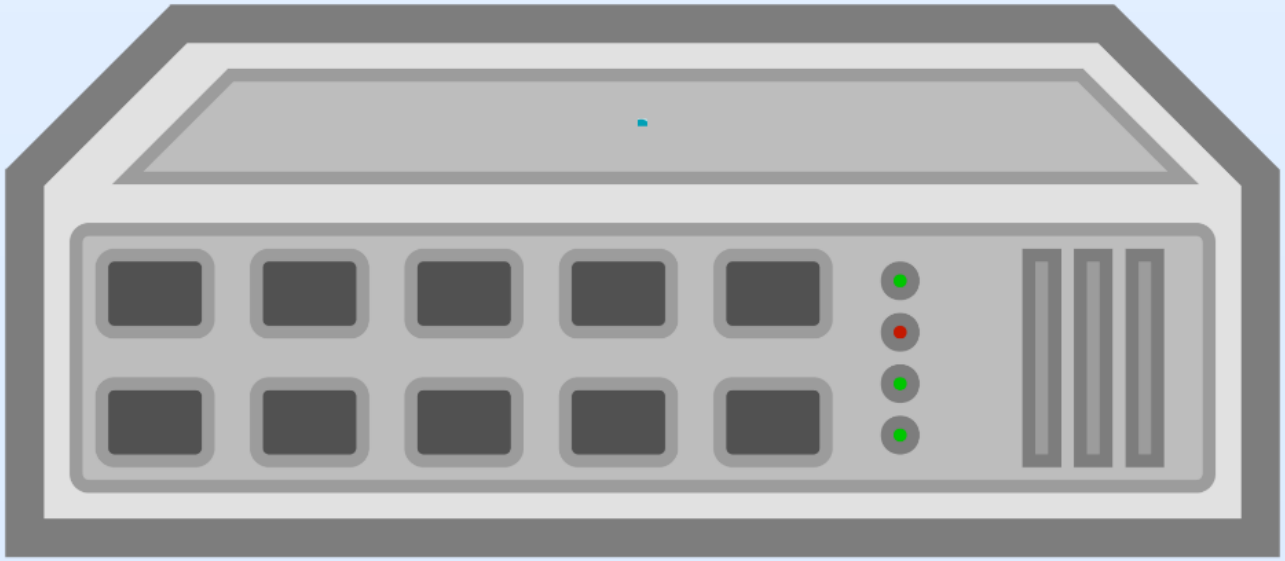
Depends on
- Computers
- Staging from tape
- External schedulers



Visualization

Depends on

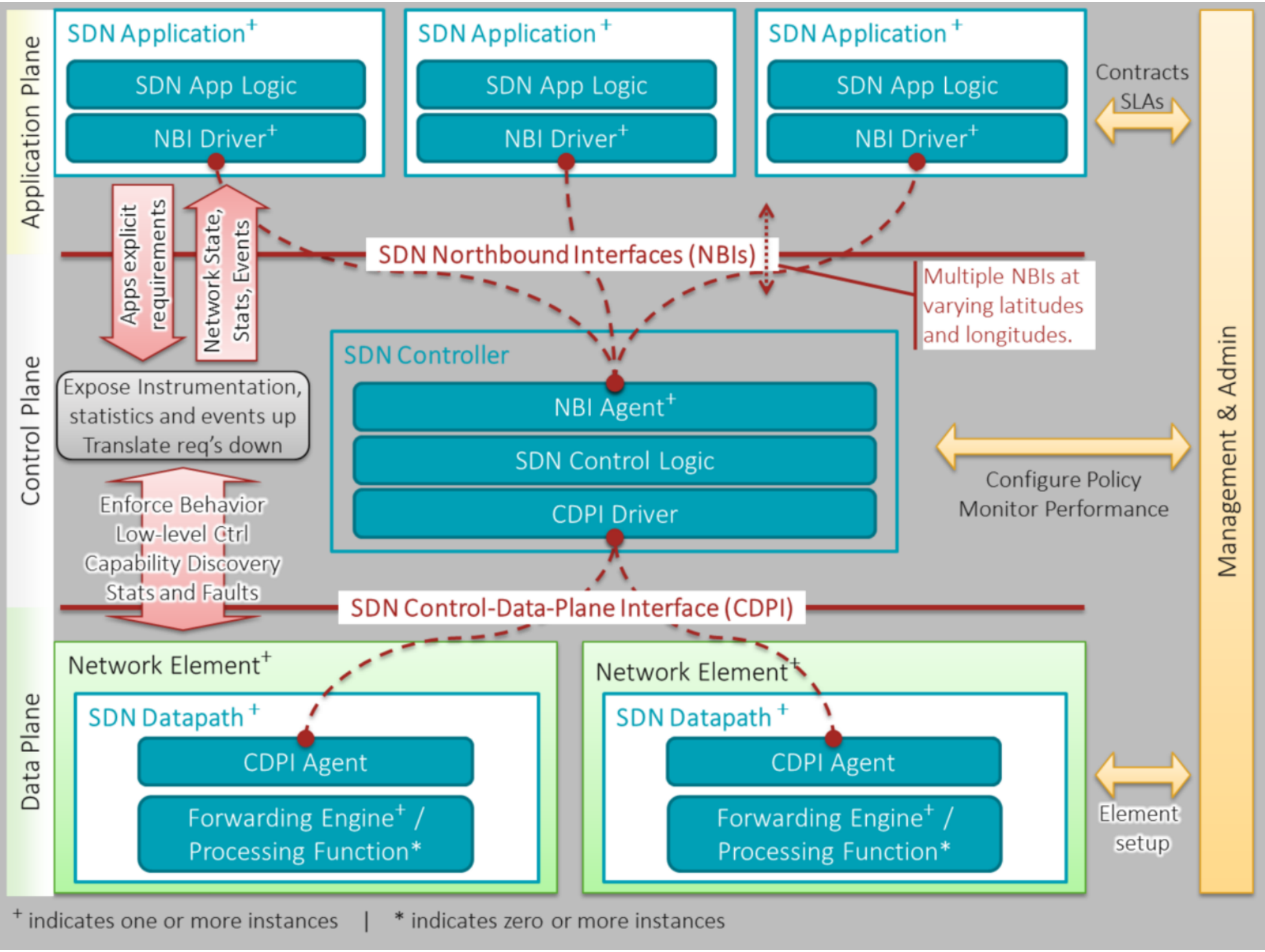
- Scheduler
- Computers
- Network



Gateway

100 Gbps primary connection

Hundreds of concurrent connections


Application Plane
Control Plane
Data Plane
SDN Application+
SDN App Logic
NBI Driver+
SDN Application+
SDN App Logic
NBI Driver+
SDN Application+
SDN App Logic
NBI Driver+
Contracts
SLAs
Apps explicit requirements
Network State, Stats, Events
SDN Northbound Interfaces (NBIs)
Multiple NBIs at varying latitudes and longitudes.
SDN Controller
NBI Agent+
SDN Control Logic
CDPI Driver
Expose Instrumentation, statistics and events up
Translate req's down
Configure Policy
Monitor Performance
Management & Admin
Enforce Behavior
Low-level Ctrl
Capability Discovery
Stats and Faults
SDN Control-Data-Plane Interface (CDPI)
Network Element+
SDN Datapath+
CDPI Agent
Forwarding Engine+ / Processing Function*
Network Element+
SDN Datapath+
CDPI Agent
Forwarding Engine+ / Processing Function*
Element setup
+ indicates one or more instances | * indicates zero or more instances

User Communities

All-hills simulations

Post-processing

Data analysis

Storage

Disk Storage

Ready

Input files for running jobs

Results files for viewing jobs

Files targeted for tape

Files staged from tape

Typically on one of the slow

Visualization

Depends on

- Scheduler
- Computers
- Filesystem

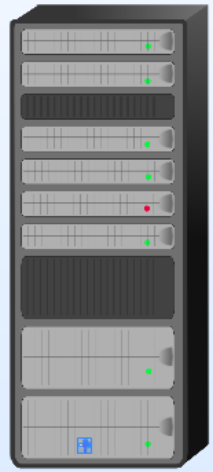
Scheduler

Receives input from

- Users
- Computers

Depends on

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



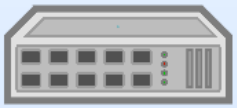
Computers

Typically 1-5 exaflop/s systems

10K - 1M Cores

High Bandwidth, Low Latency Interconnect (2-5 TB/sec and 0.5 us)

Increasingly often uses accelerators (GPGPUs, FPGAs, etc)



Gateway

100 Gbps primary connection

Hundreds of concurrent connections

Grid

Receives input from

- External Schedulers
- Computers
- Tape

Depends on

- Computers
- Staging from tape
- External schedulers

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]