A Scalable Runtime for the ECOSCALE Heterogeneous Exascale Hardware Platform

Paul Harvey

Konstantin Bakanov, Ivor Spence, Dimitrios S. Nikolopoulos


Thurso
OUTER HEBRIDES
Ullapol
ISLE OF SKYE
Inverness
Aberdeen
Perth
Dundee
Stirling
Glasgow
EDINBURGH
SCOTLAND
Londonderry
NORTHERN IRELAND
Stranraer
Newcastle-upon-Tyne
Durham
Carlisle
BELFAST
ISLE OF MAN
York
Blackpool
Preston
Leeds
Kingston-upon-Hull
REPUBLIC OF IRELAND
DUBLIN
ANGLESEY
LIVERPOOL
Holyhead
MANCHESTER
Caernarfon
Llangollen
ENGLAND
Leicester
Norwich
Aberystwyth
BIRMINGHAM
WALES
Brecon
Stratford-upon-Avon
Cambridge
Ipswich
Pembroke
Oxford
LONDON
Cardiff
Bristol
Bath
Stonehenge
Dover
EUROTUNNEL
Salisbury
Folkestone
Exeter
Portsmouth
Calais
Plymouth
Brighton
Penzance
ISLE OF WIGHT


Thurso
OUTER HEBRIDES
Ullapol
ISLE OF SKYE
Inverness
Aberdeen
Perth
Dundee
Stirling
Glasgow
EDINBURGH
SCOTLAND
Londonder
NORTH IRELAND
Stranraer
Newcastle-upon-Tyne
Durham
Carlisle
BELFAST
ISLE OF MAN
York
Blackpool
Preston
Leeds
Kingston-upon-Hull
REPUBLIC OF IRELAND
DUBLIN
ANGLESEY
LIVERPOOL
Holyhead
MANCHESTER
Caernarfon
Llangollen
ENGLAND
Leicester
Norwich
Aberystwyth
BIRMINGHAM
WALES
Brecon
Stratford-upon-Avon
Cambridge
Ipswich
Pembroke
Oxford
LONDON
Cardiff
Bristol
Bath
Stonehenge
Dover
EUROTUNNEL
Salisbury
Folkestone
Calais
Exeter
Portsmouth
Plymouth
Brighton
Penzance
ISLE OF WIGHT

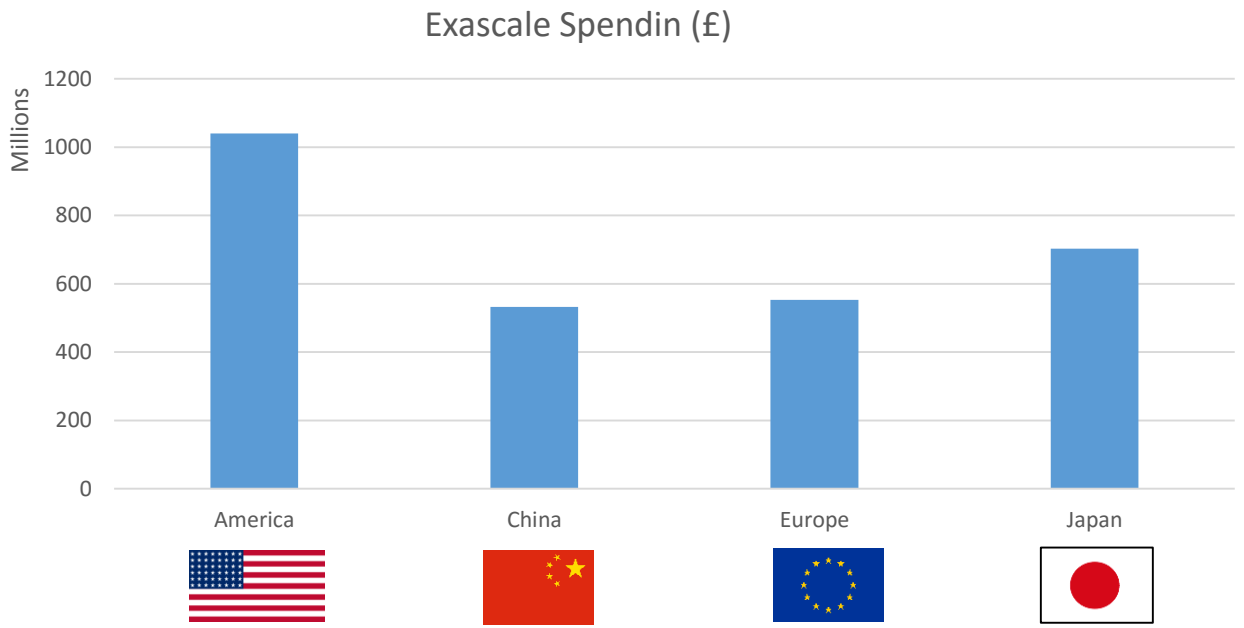
Looking To Discuss and Share Ideas

- No implementation
- No results

- Just design!
  - Intro & Context
  - Hardware
  - Language
  - Runtime Architecture

Exascale: Money

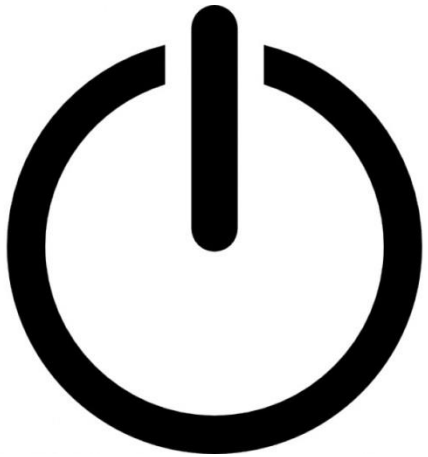
- America : ~$1500 Million
- Europe : €700 million
- China : 5000 million CNY
- Japan : 110 Billion JPY



http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/3-BDEC2015-ishikawa.pdf
http://www.hpcwire.com/2016/02/12/obama-budget-reveals-new-elements-exascale-program/
http://www.scientific-computing.com/news/news_story.php?news_id=2732
http://www.exascale.org/mediawiki/images/b/b8/Talk25-zjin.pdf

Exascale: Brains

Exascale: Problems



http://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf

Exascale: Problems

http://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf

Ecoscale - ecoscale.eu

- Funded till October 2018
- ~£4,000,000
- Building new Hardware
  - Exascale prototype with FPGA focus
- Queen's University working on Software

TSI

Queen's University Belfast

life.augmented

acciona

MANCHESTER 1824
The University of Manchester

POLITECNICO DI TORINO

CHALMERS
UNIVERSITY OF TECHNOLOGY

SYNELIXIS

FPGA

FFT

BitCoin

Matrix Mul

FPGA: Floating point Intensive Calculation

Platform	Time (ns)	W	Energy/Step (nJ)	Obtained By
HD 4400 (GPU)	3.13	15	46.9	Measurement
GTX 960 (GPU)	0.163	120	19.56	Measurement
Quadro K4200 (GPU)	0.204	105	21.42	Measurement
GTX Titan (GPU)	0.0389	375	14.61	Extrapolation
Virtex 7 (FPGA)	0.315	24.4	7.69	Measurement

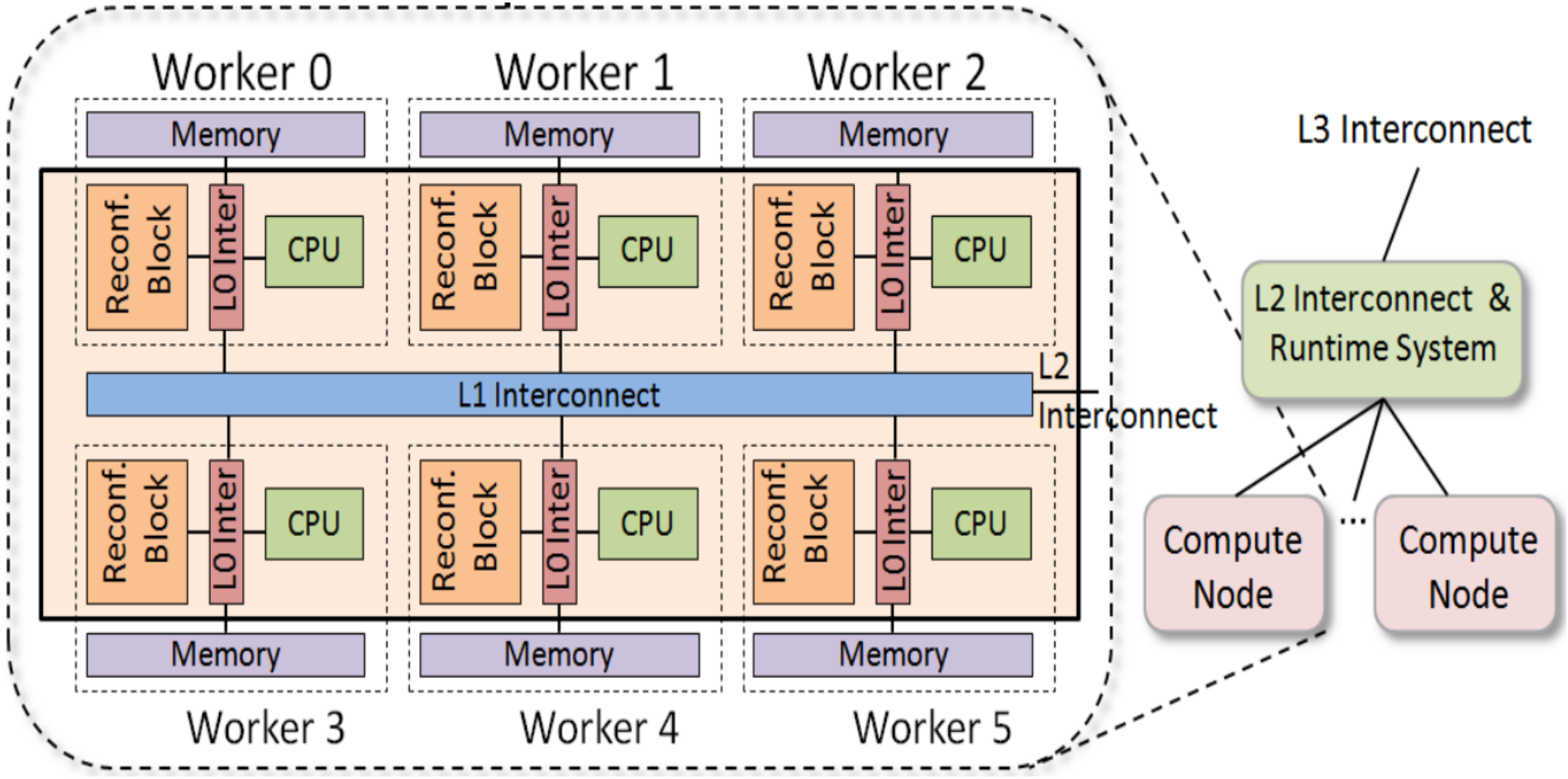
- Compute-intensive, not using global memory
- GPU memory bandwidth is >> FPGA memory bandwidth
  - GPU DDR4 ~8x more than FPGA DDR3

FPGA: Floating point Intensive Calculation

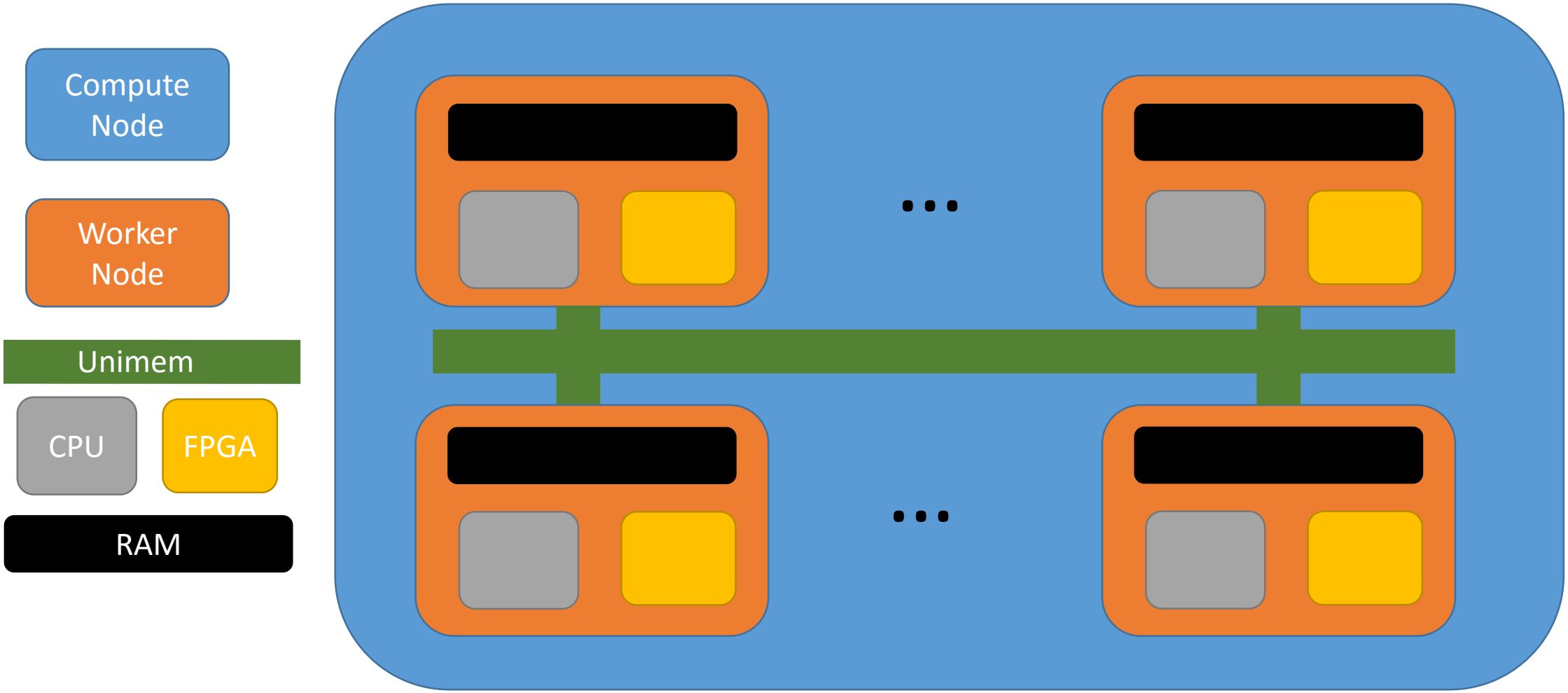
Platform	Time (ns)	W	Energy/Step (nJ)	Obtained By
HD 4400 (GPU)	3.13	15	46.9	Measurement
GTX 960 (GPU)	0.163	120	19.56	Measurement
Quadro K4200 (GPU)	0.204	105	21.42	Measurement
GTX Titan (GPU)	0.0389	375	14.61	Extrapolation
Virtex 7 (FPGA)	0.315	24.4	7.69	Measurement

- Compute-intensive, not using global memory
- GPU memory bandwidth is >> FPGA memory bandwidth
  - GPU DDR4 ~8x more than FPGA DDR3

Architecture



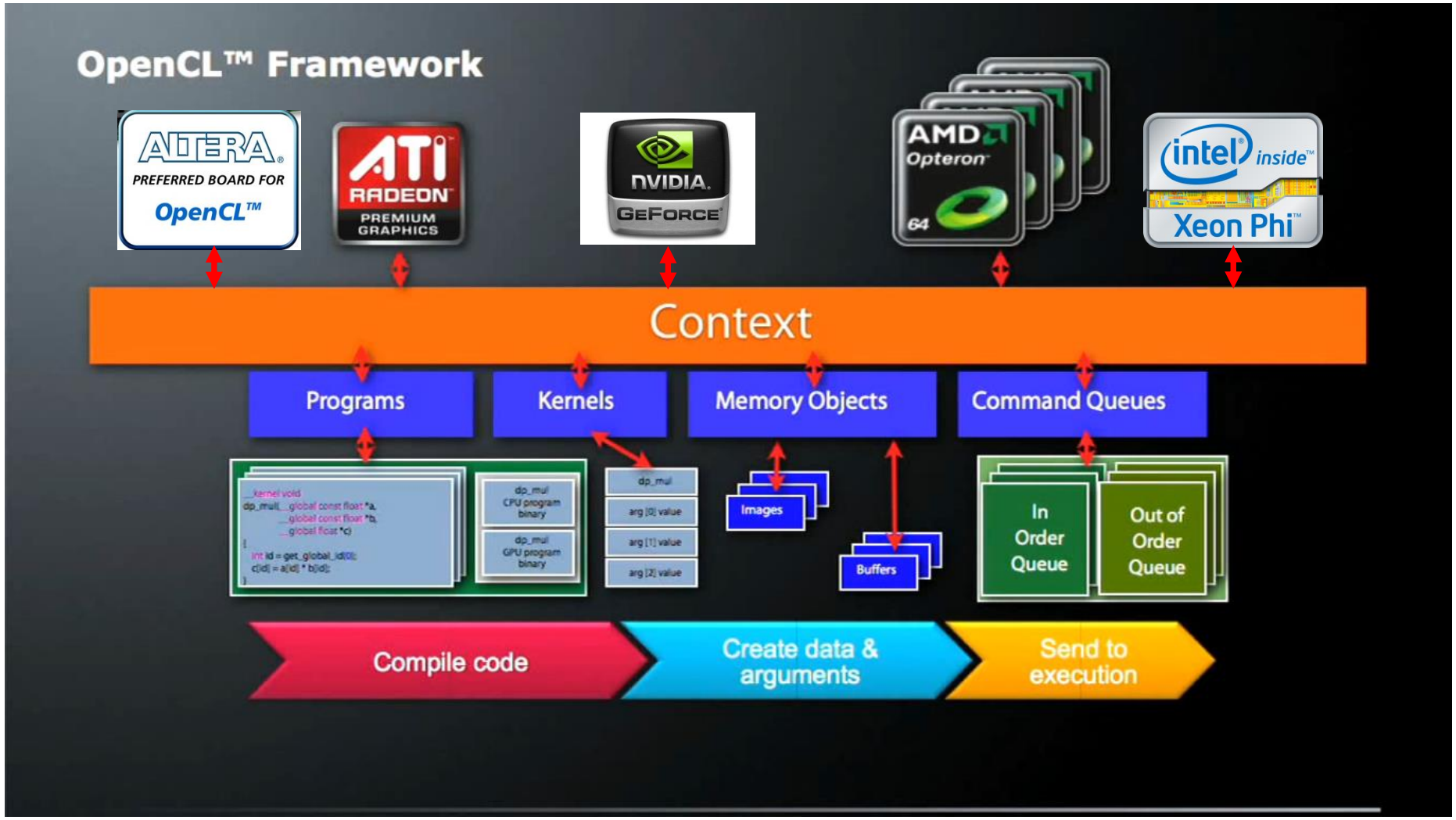
Simplified Architecture



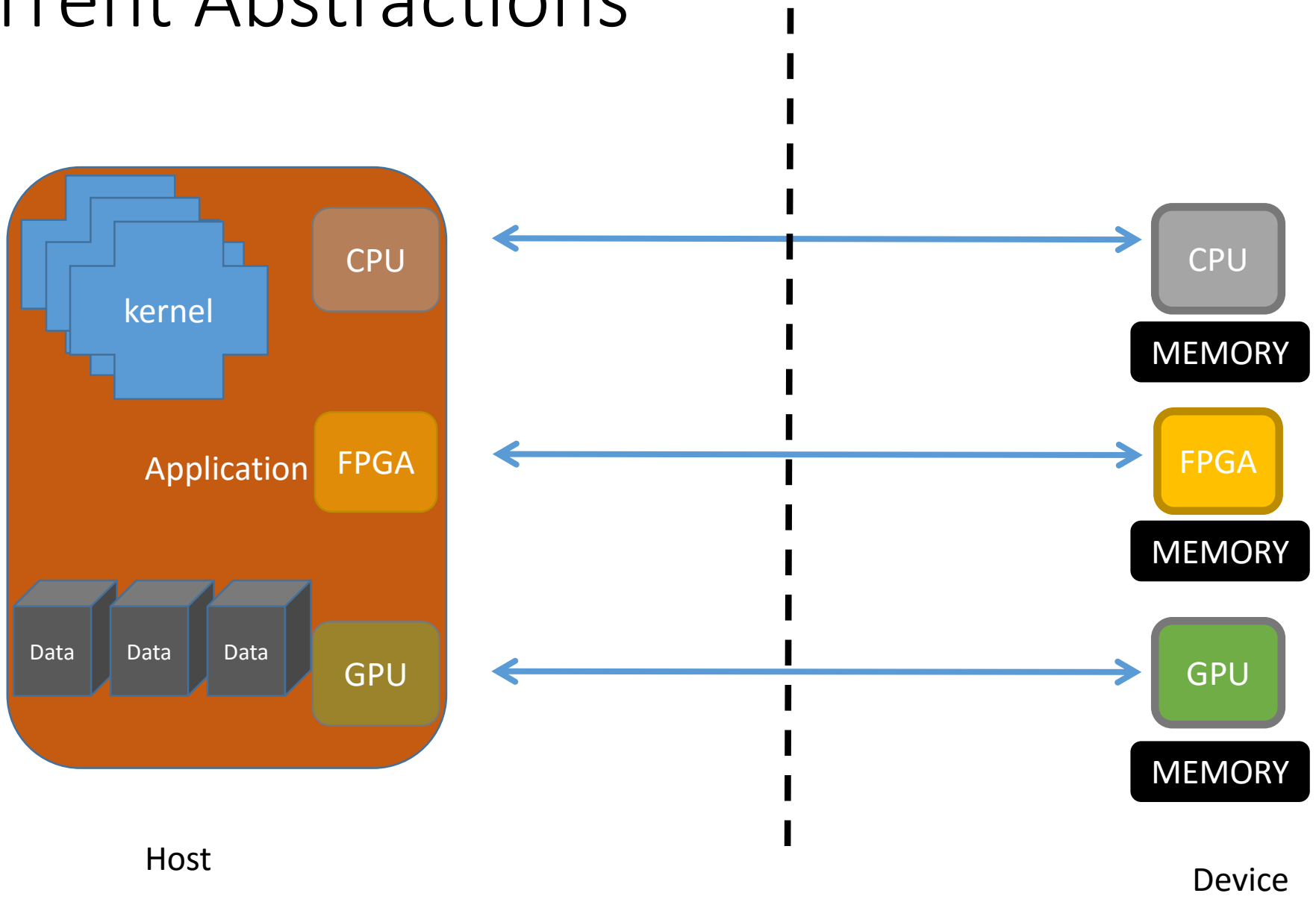
Unimem

- RDMA
- PGAS Address Space
  - One or more single address spaces

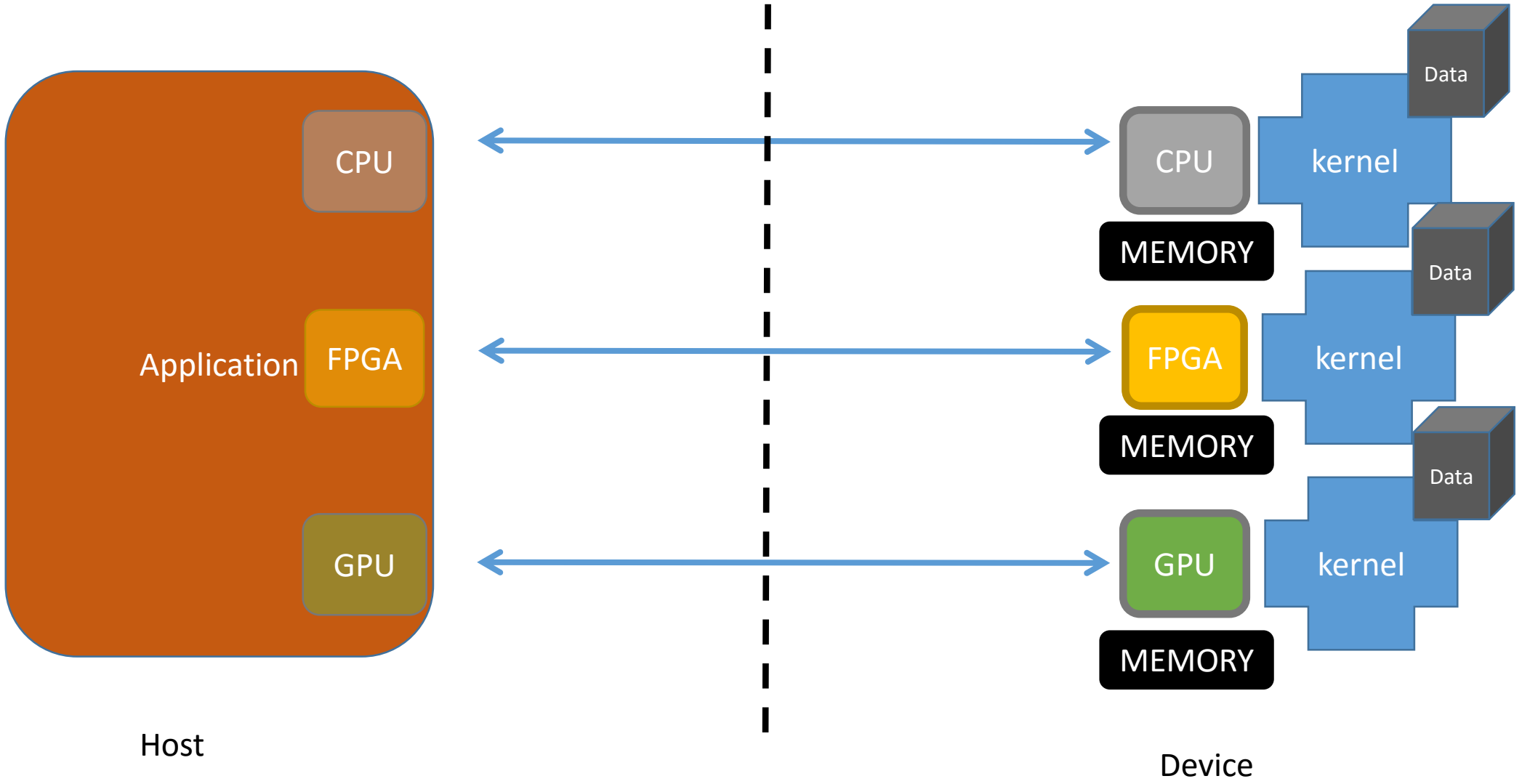
OpenCL


OpenCL™ Framework
ALTERA
PREFERRED BOARD FOR
OpenCL™
ATI
RADEON
PREMIUM
GRAPHICS
NVIDIA
GeForce
AMD
Opteron
64
intel
inside
Xeon Phi
Context
Programs
Kernels
Memory Objects
Command Queues
__kernel void
dp_mul(__global const float *a,
__global const float *b,
__global float *c)
{
int id = get_global_id(0);
c[id] = a[id] * b[id];
}
dp_mul
CPU program
binary
dp_mul
GPU program
binary
dp_mul
arg [0] value
arg [1] value
arg [2] value
Images
Buffers
In
Order
Queue
Out of
Order
Queue
Compile code
Create data &
arguments
Send to
execution

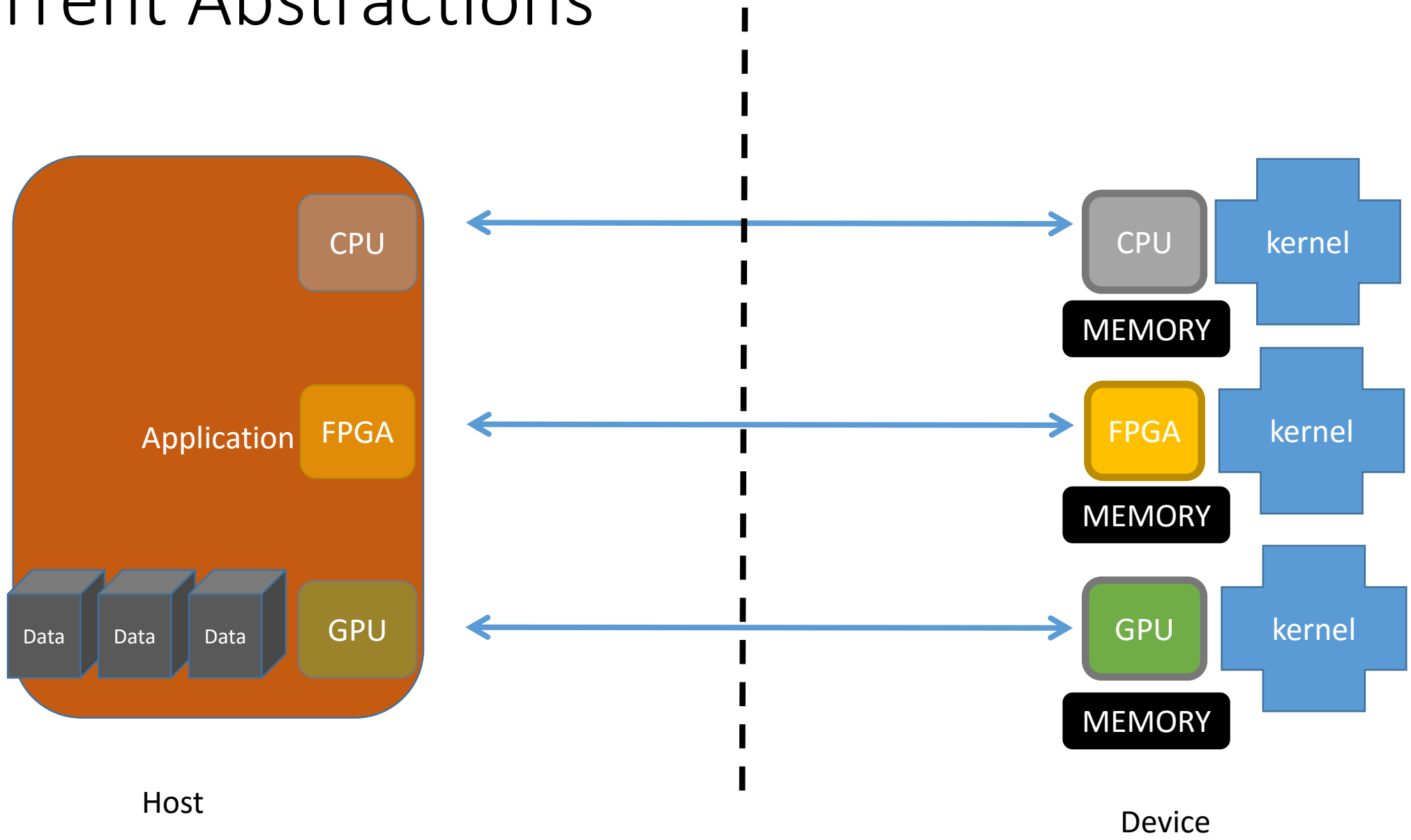
Current Abstractions



Current Abstractions



Current Abstractions



OpenCL

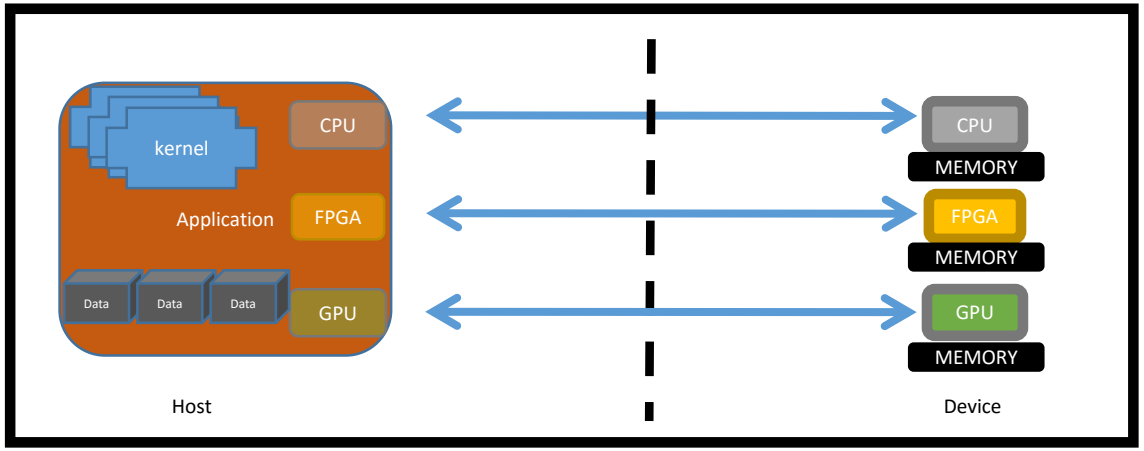
- Simple model
- Widely used in non-hpc
- Standardised
- Lots of activity
  - Industry
  - Academia
- Non-proprietary

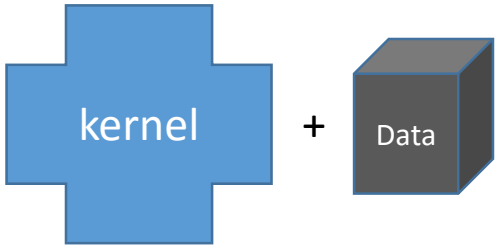
Extensions

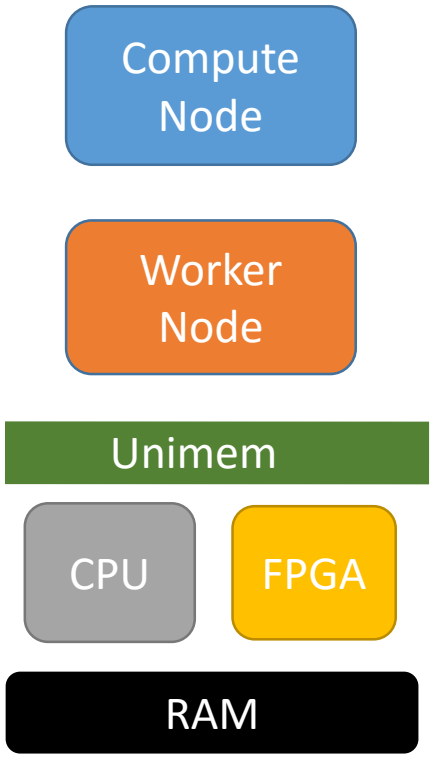
1. New abstractions of multiple hardware devices
    1. Enables scheduler to dynamically go after performance or power

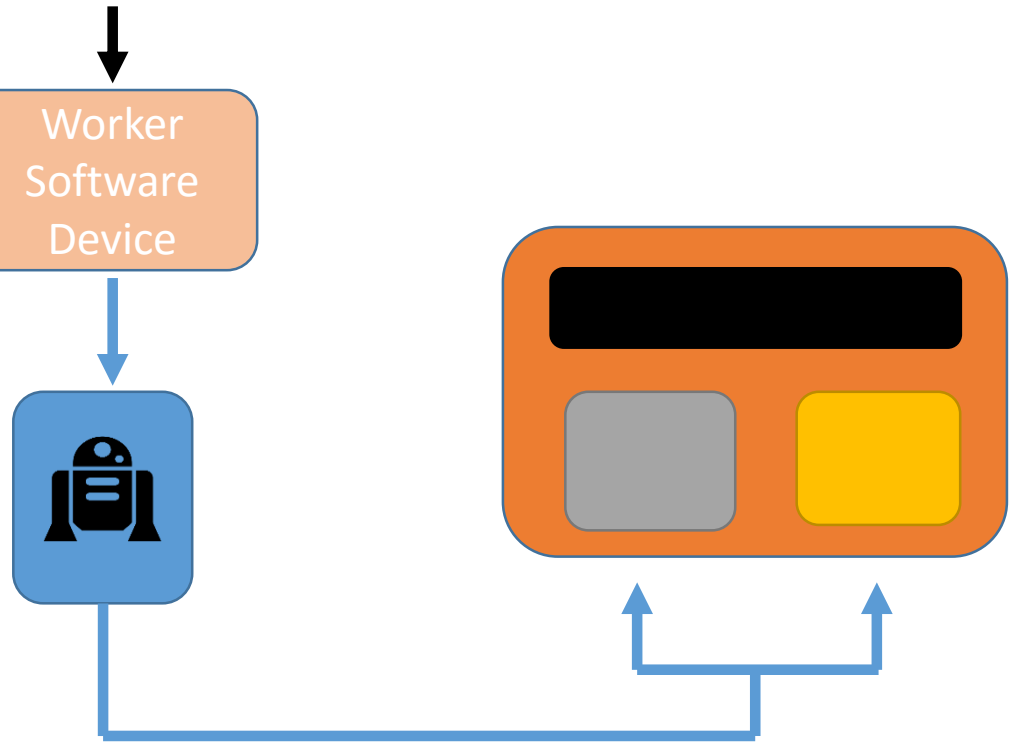
2. New fundamental unit of scheduling
    1. Better scaling across multiple compute devices
    2. Enables kernels to run where a single device has insufficient resources

Worker Abstraction



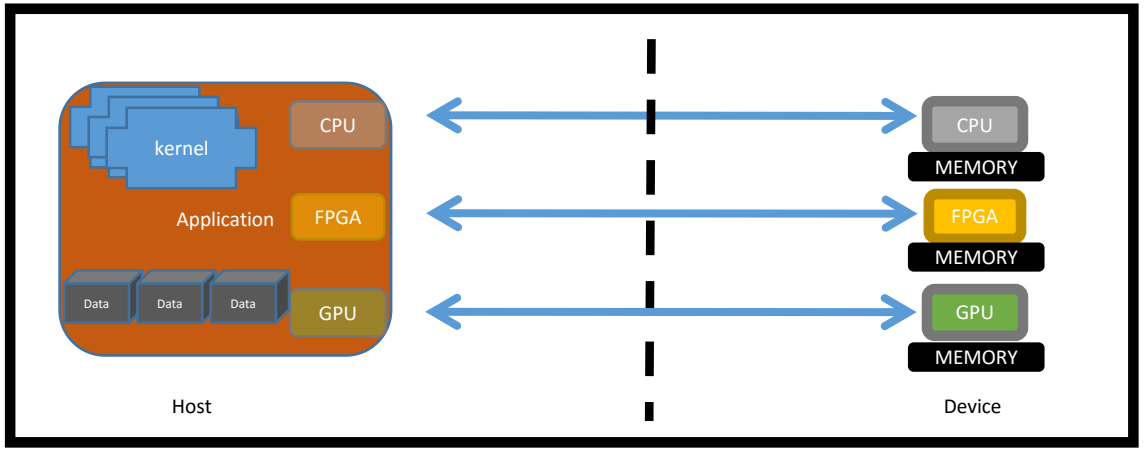


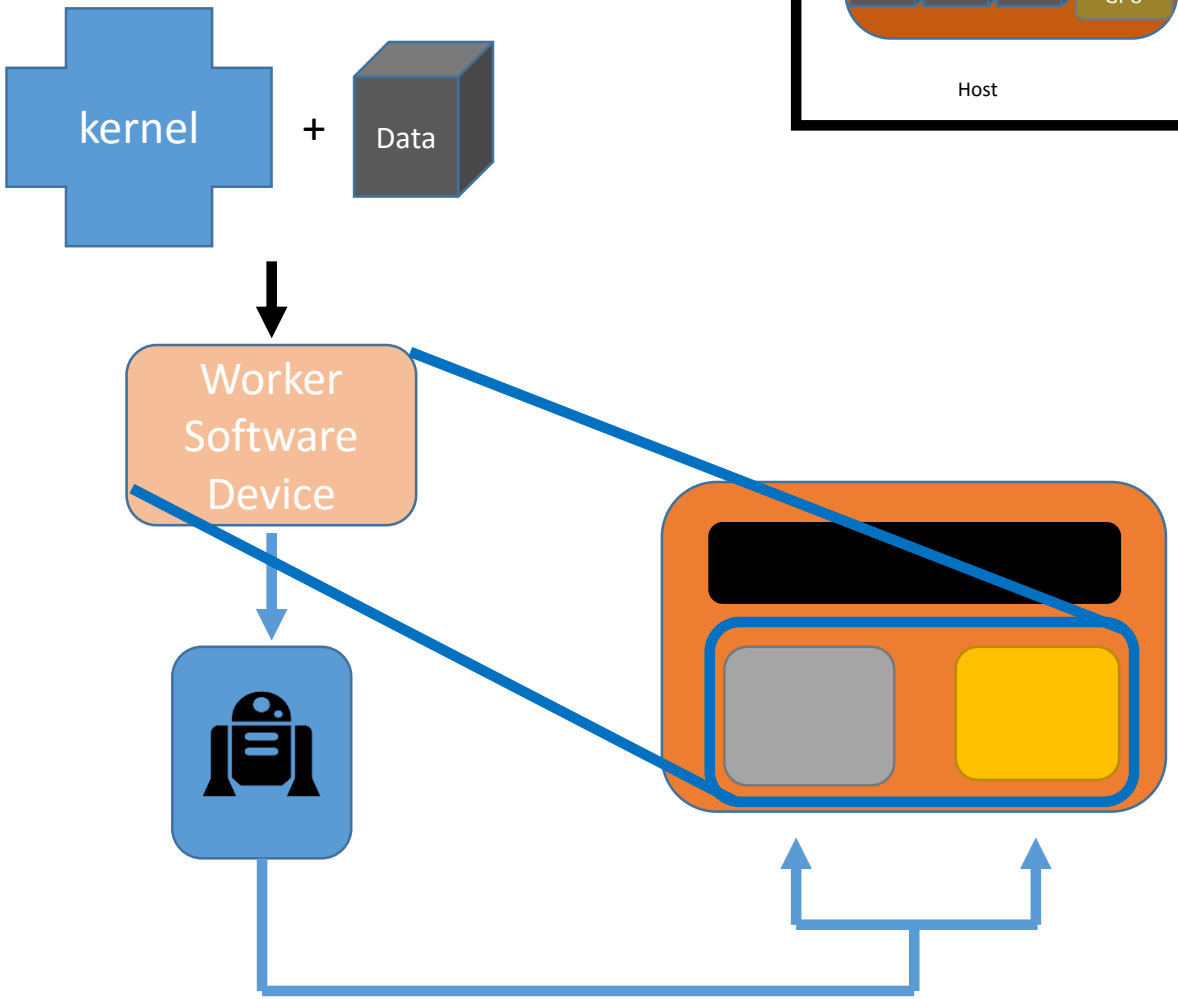




- No change for Programmer
- Scheduler control for power vs. Performance

Worker Abstraction





Compute Node

Worker Node

Unimem

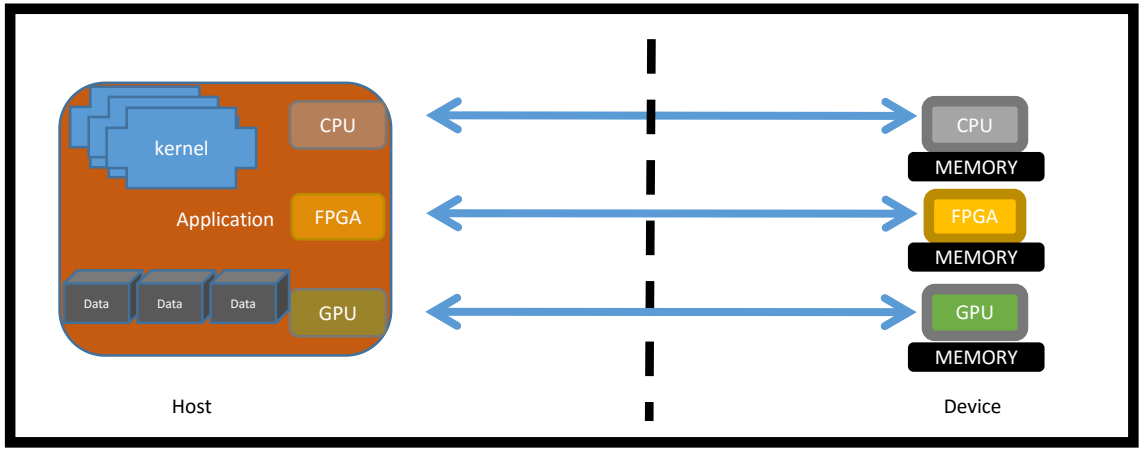
CPU

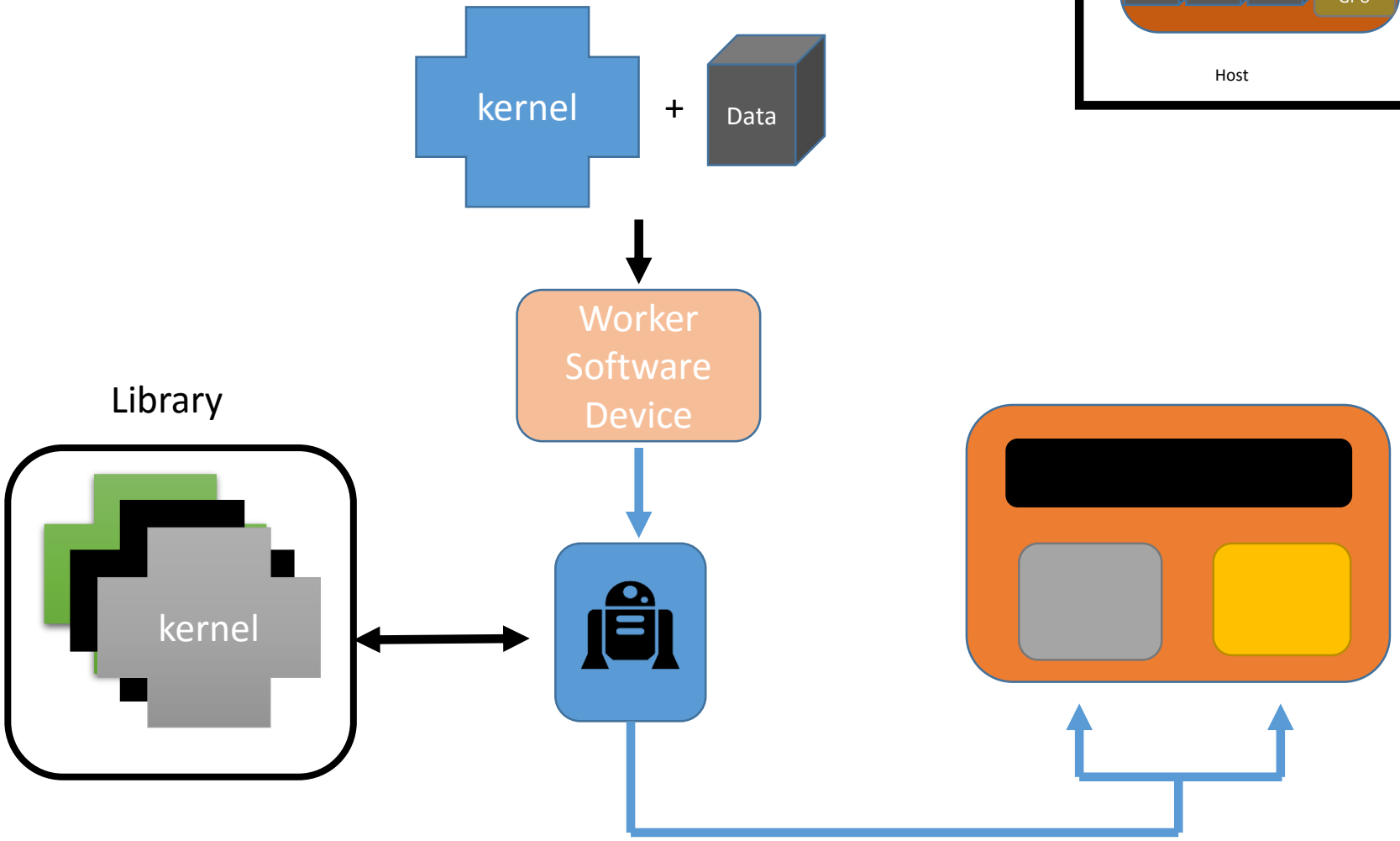
FPGA

RAM

- No change for Programmer
- Scheduler control for power vs. Performance

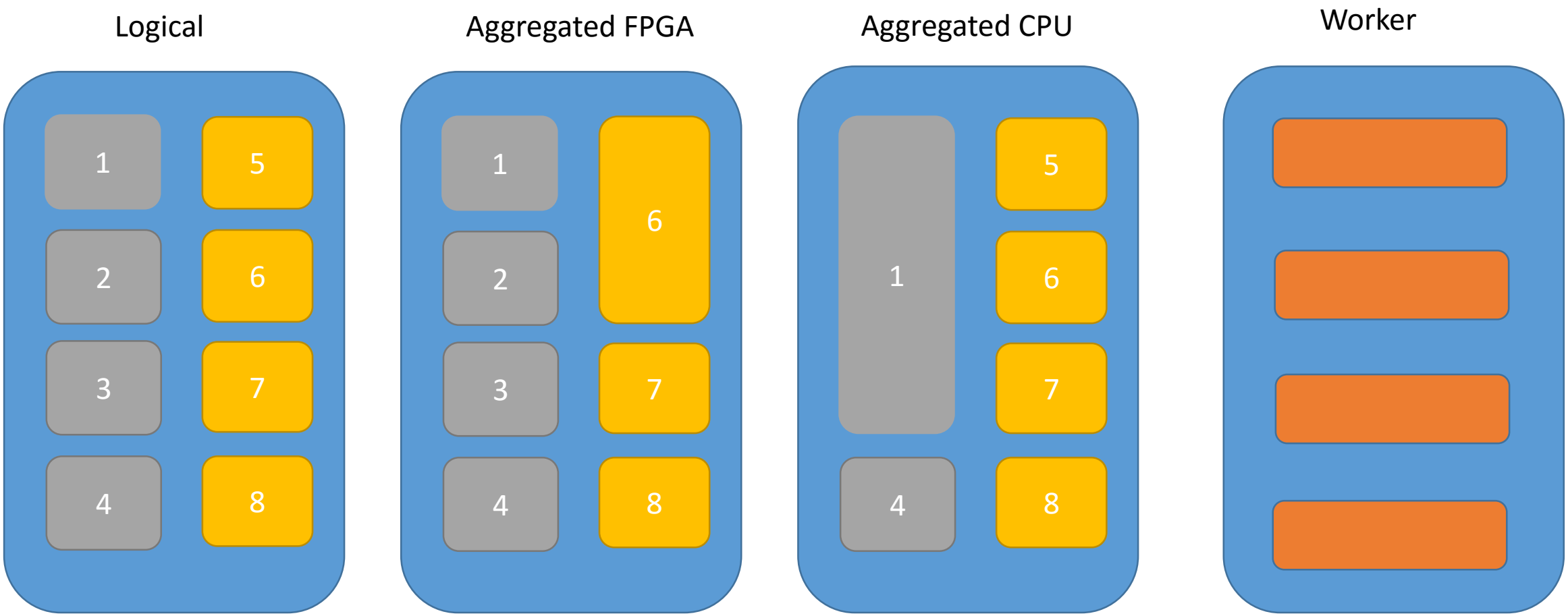
Worker Abstraction





- No change for Programmer
- Scheduler control for power vs. Performance

Abstraction Configurations

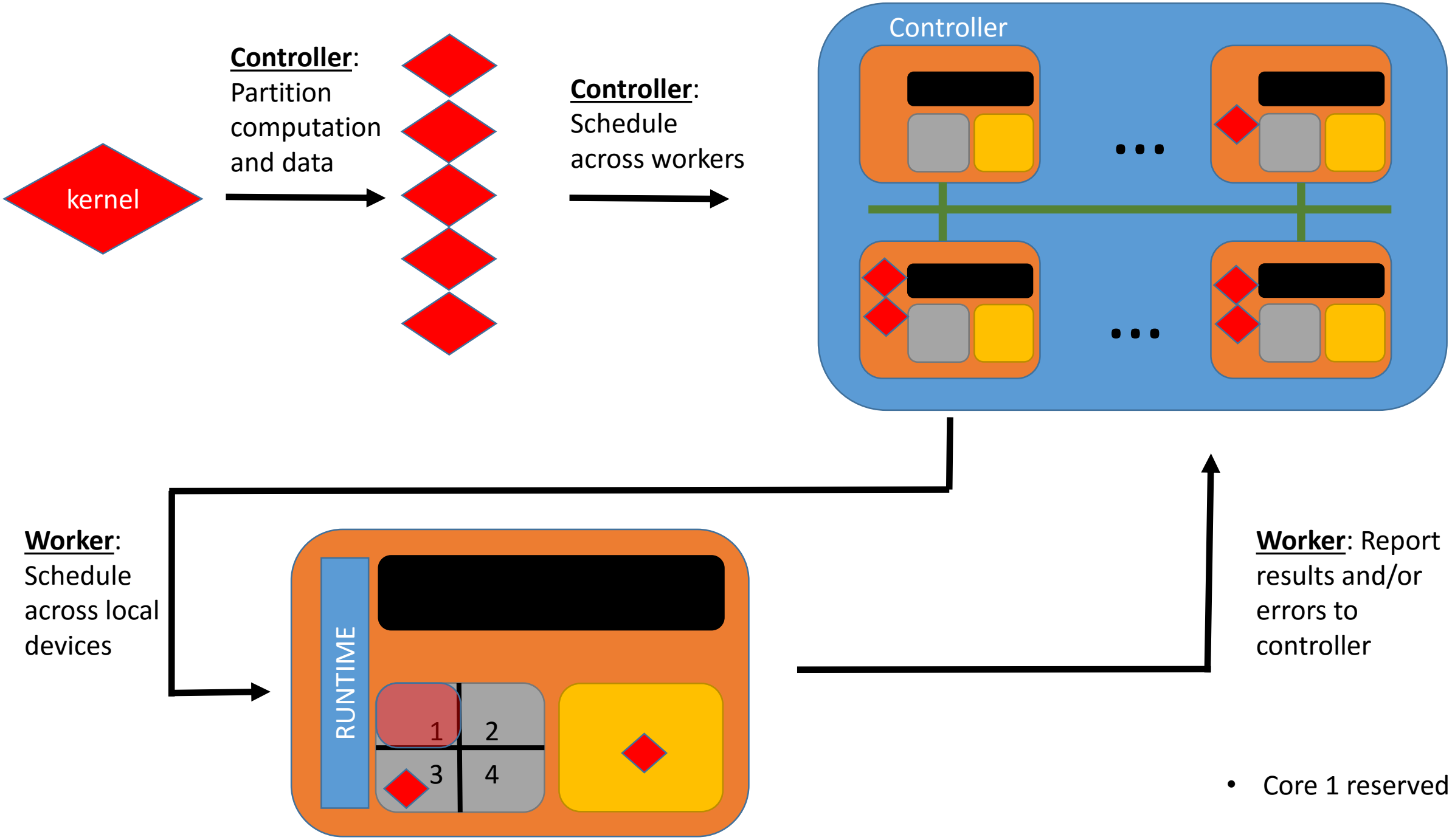


Scheduling: CPU vs. FPGA

- Machine Learning based on:
  - Runtime performance
  - Kernel input data size
  - CPUF/FPGA power consumption
  - Data locality
  - #global memory accesses
  - #branches and loops

- Is a cost model enough?

- How do we determine:
  - a power budget?
    - 100^{th} of current GPU?
  - A performance budget?
    - Current best GPU?

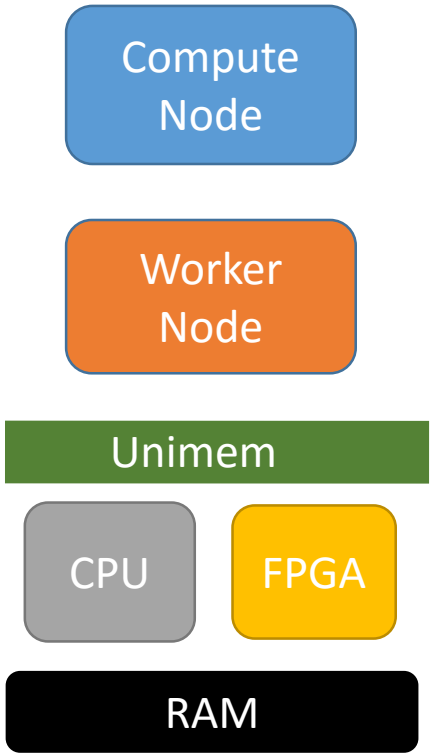


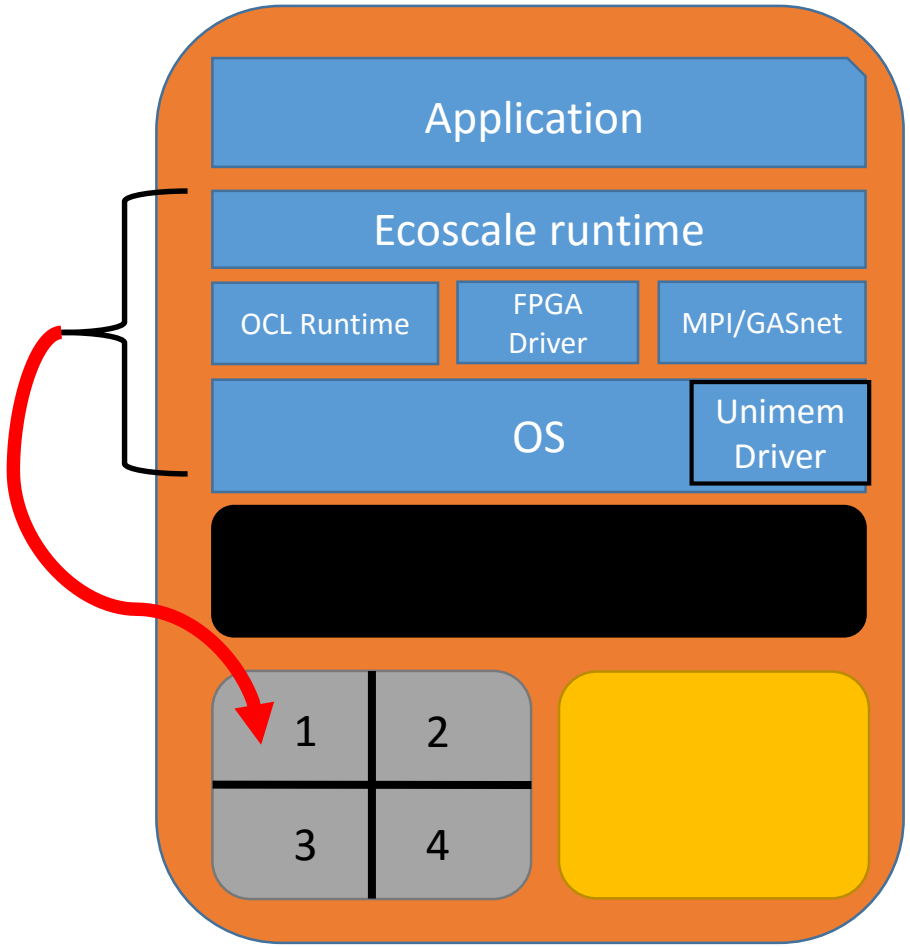
- Core 1 reserved for OS

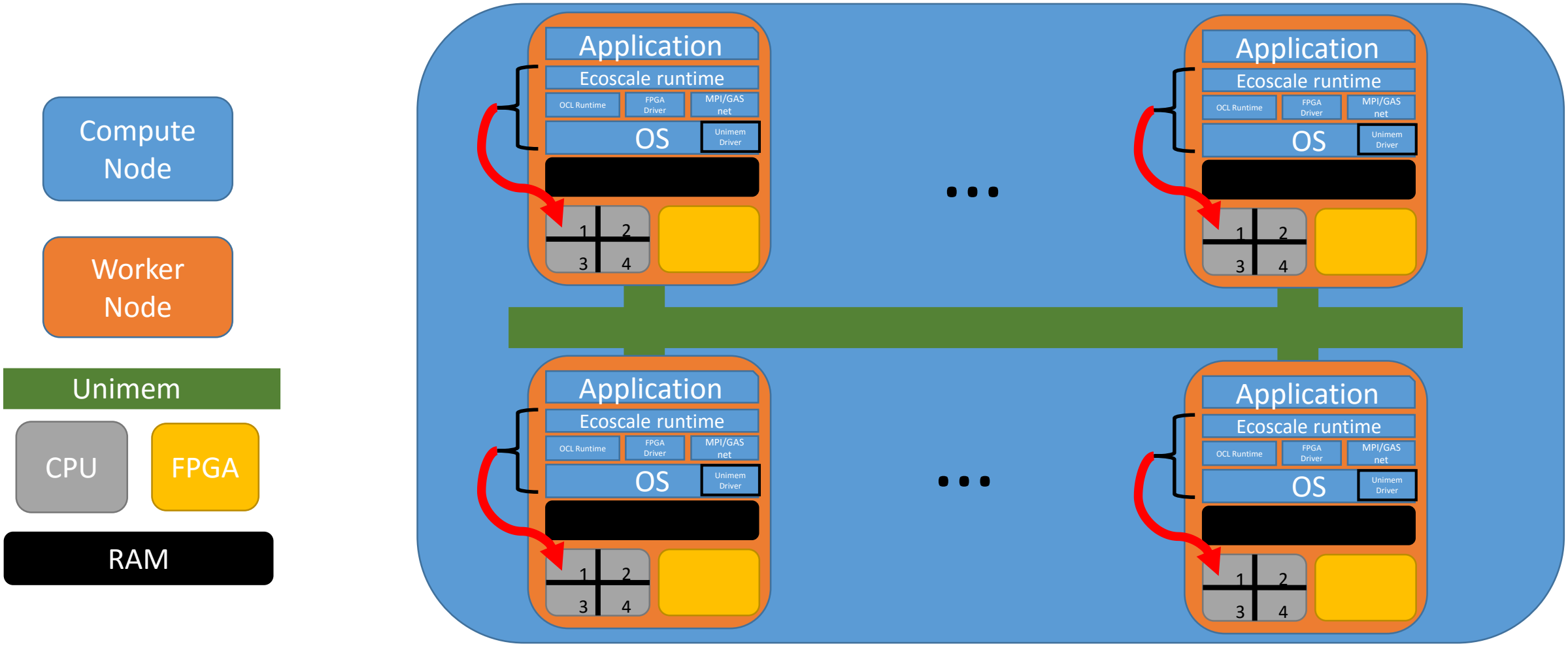
Language – Data Partitioning

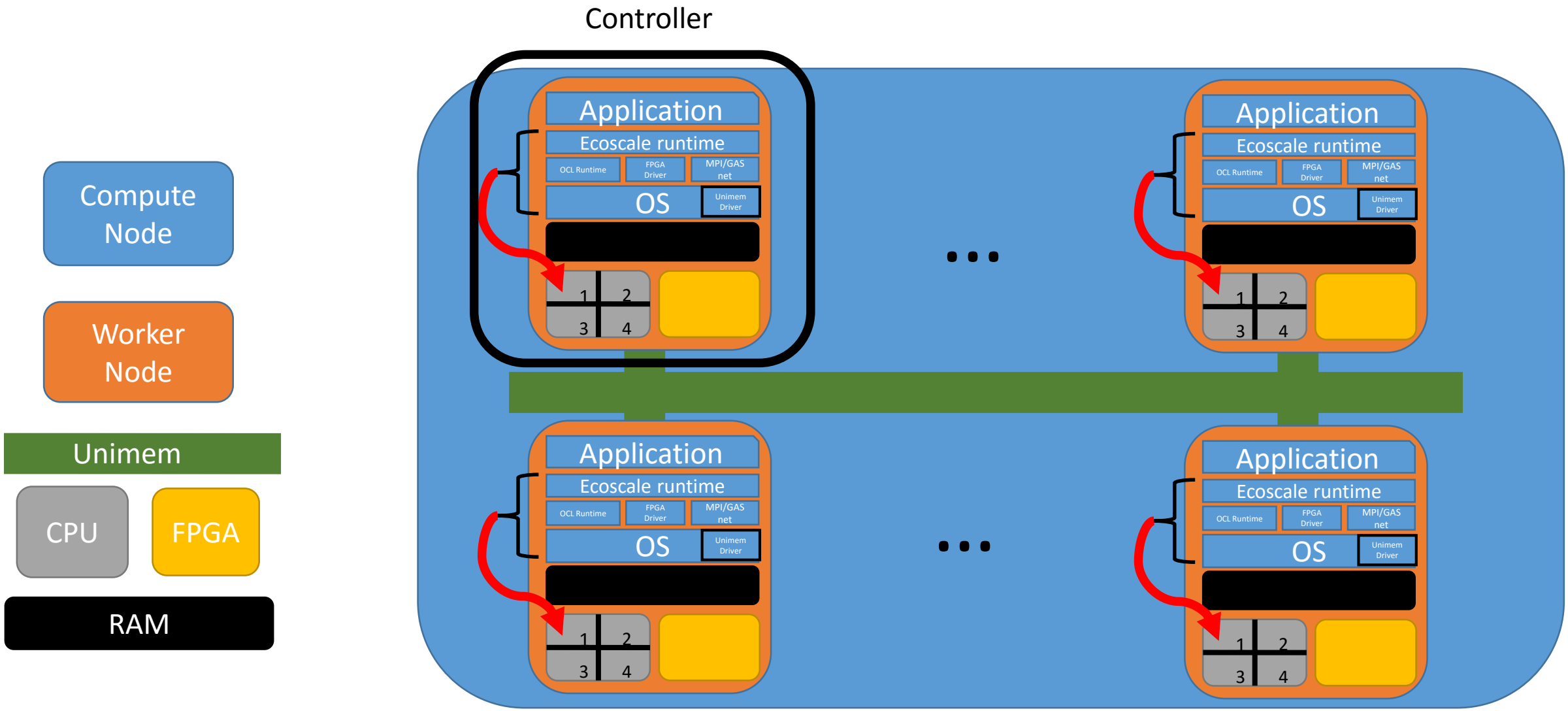
```
d_m1 = clCreateBuffer(context,
        CL_MEM_READ_WRITE,
        matrix_dim*matrix_dim*sizeof(double),
        NULL,
        ecoscale_partition(d_m1, REPLICATE, 0),
        &errcode);
```

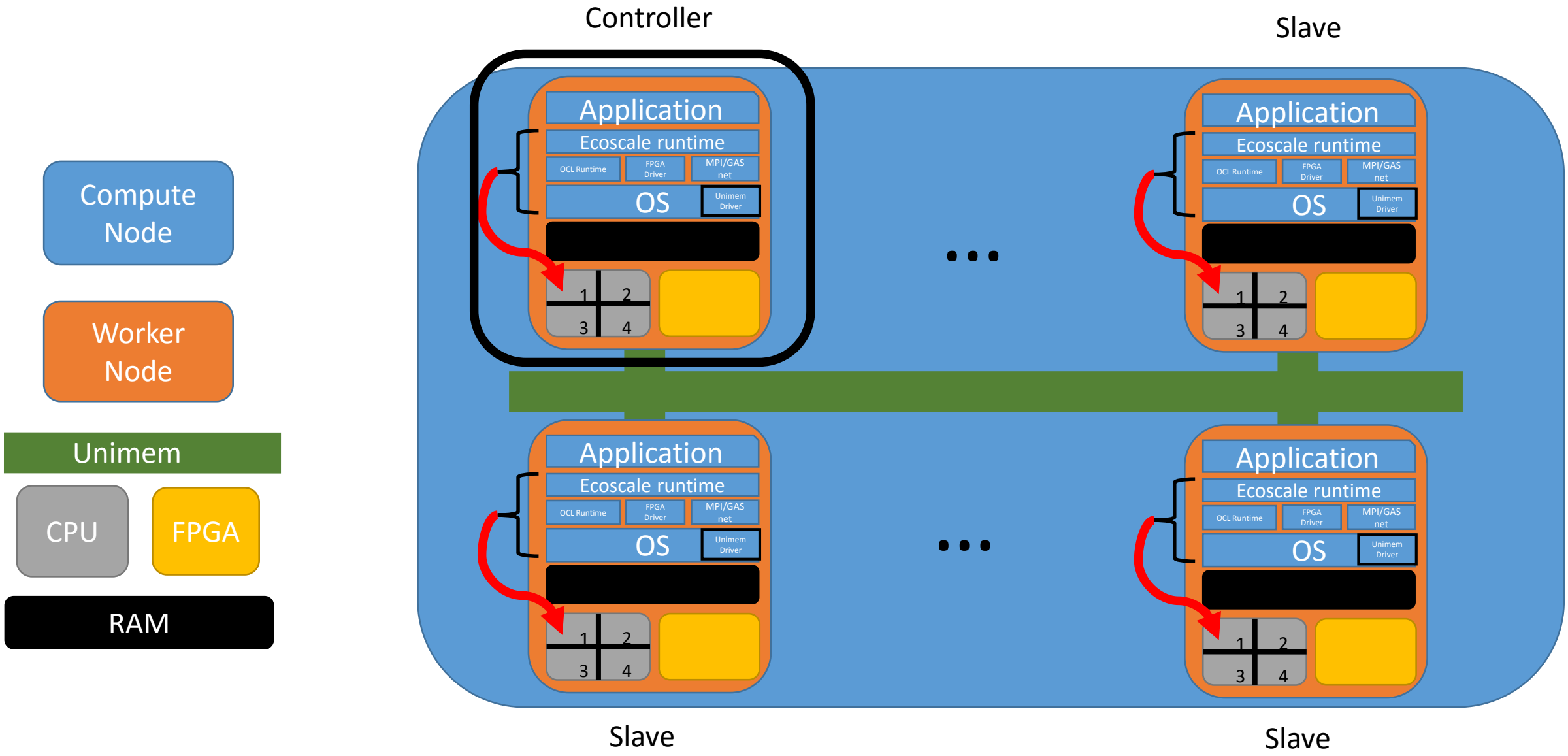
Architecture





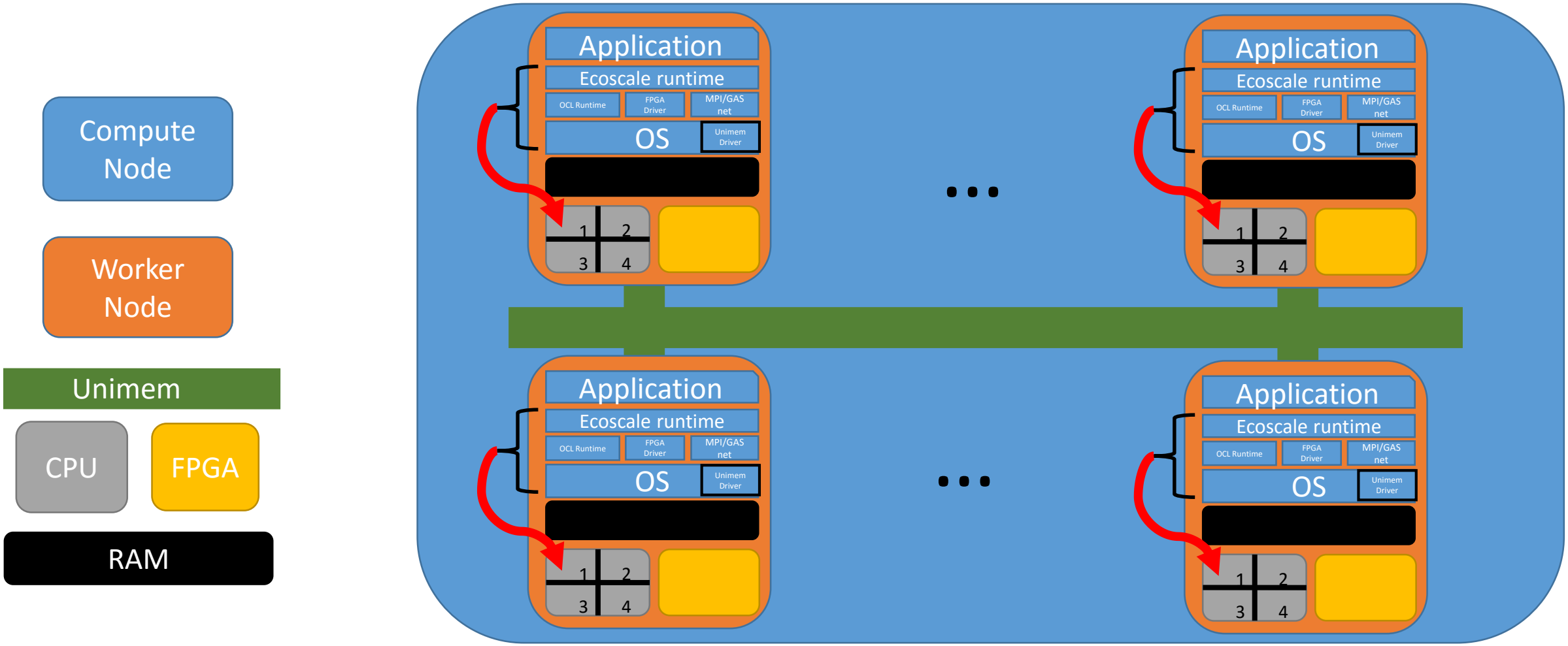

Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4


Controller
Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4

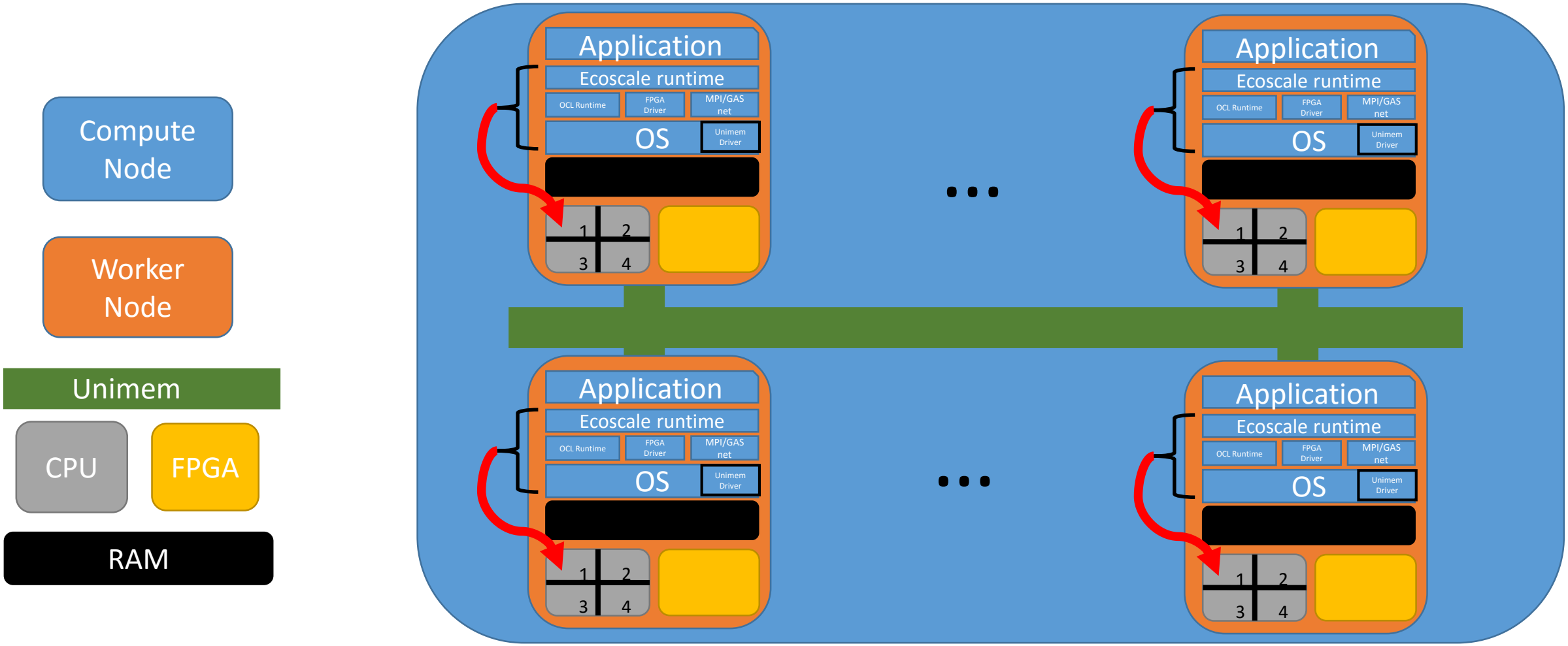

Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Controller
Slave
Slave
Slave
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4

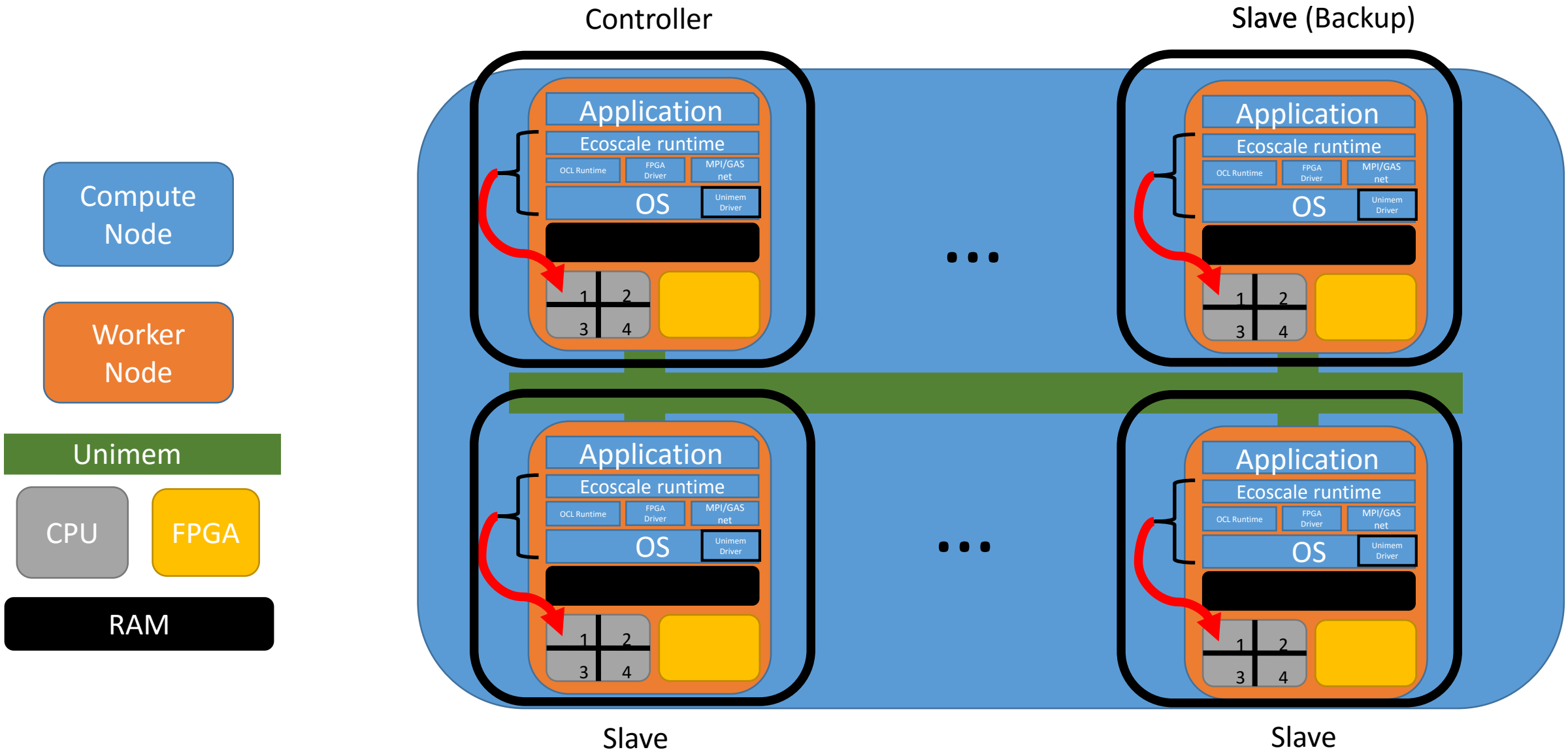
Resilience

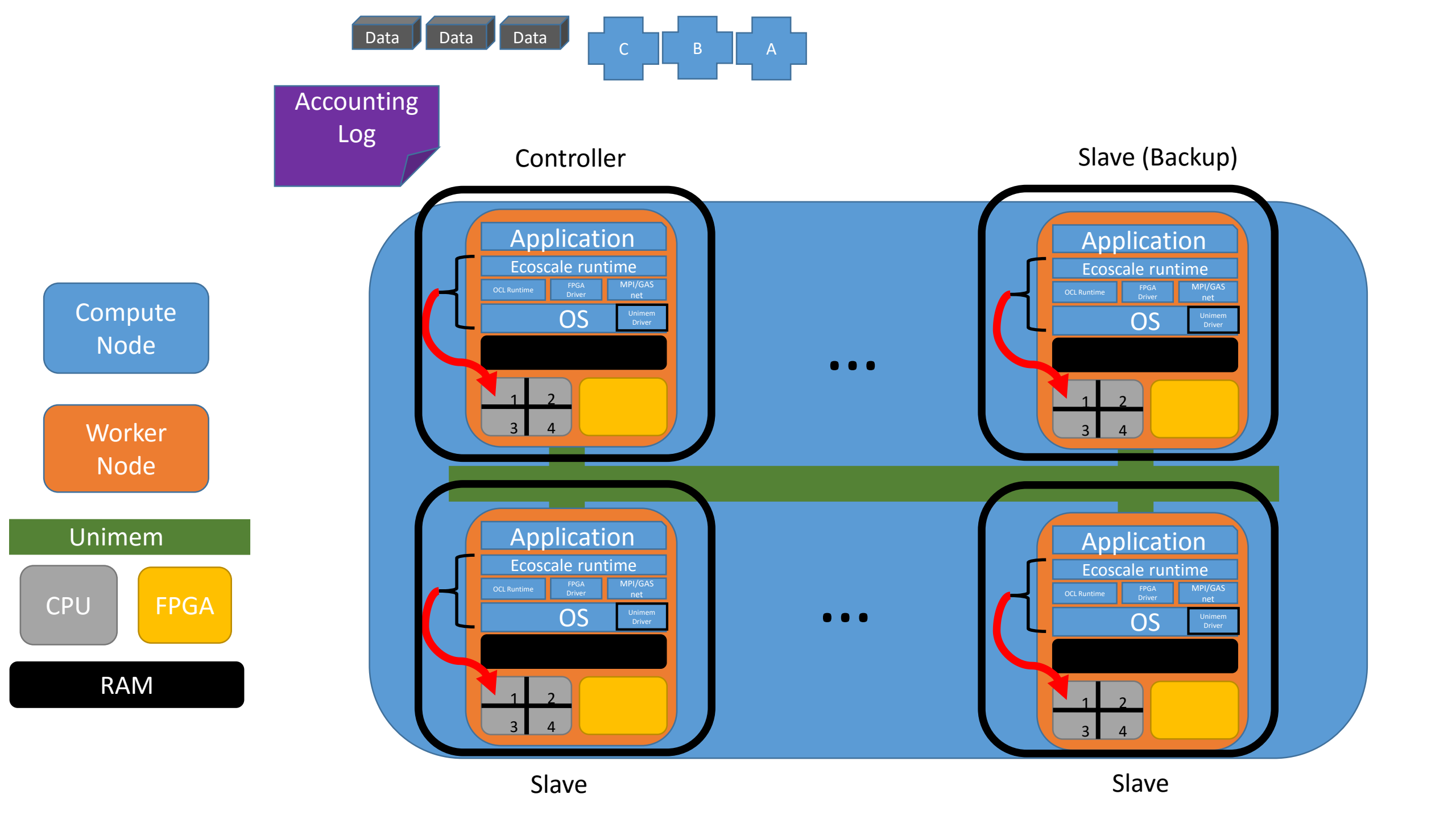
- Leaders & slaves
- Heatbeats messages
- Checkpointing

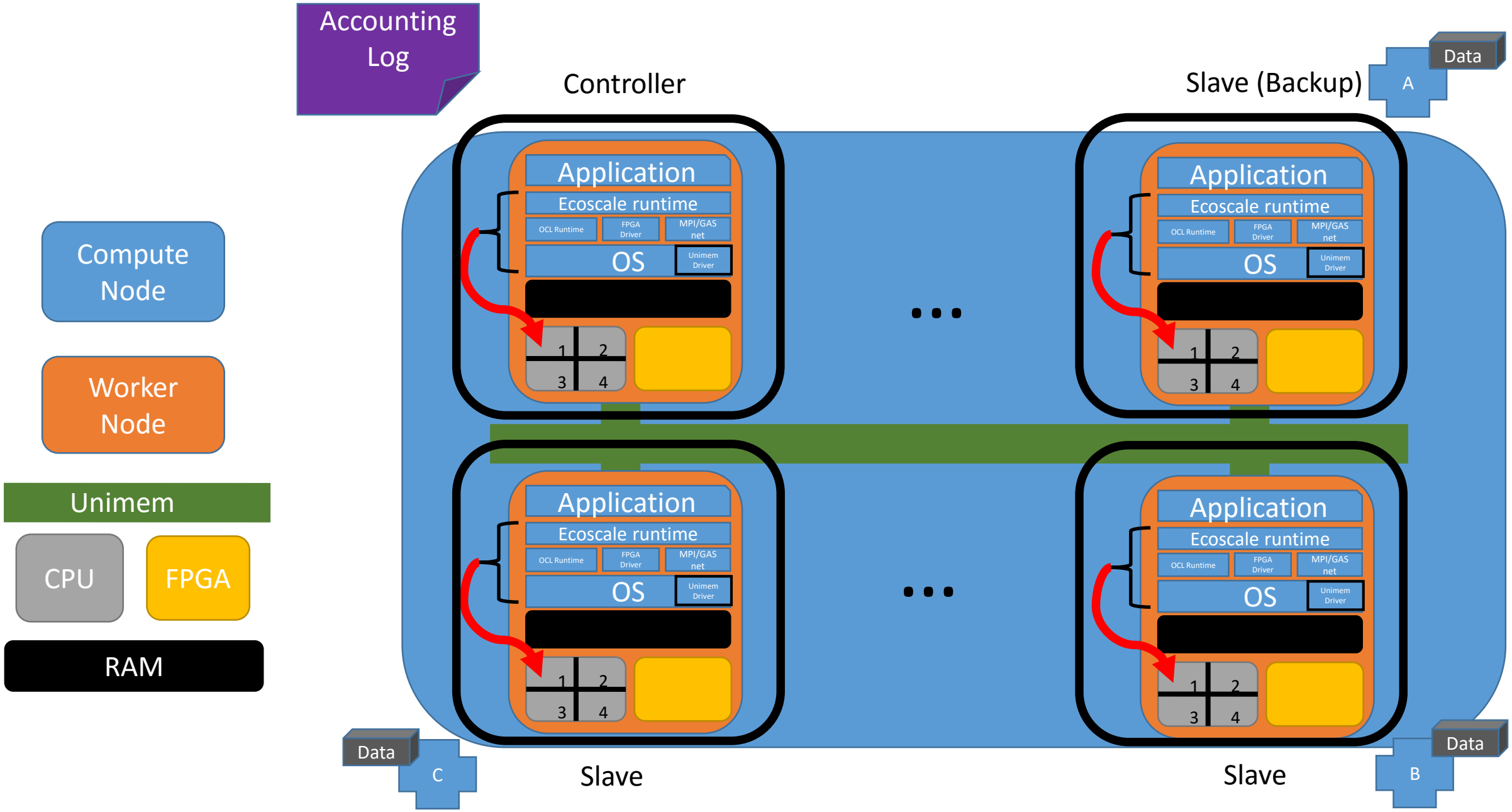

Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4

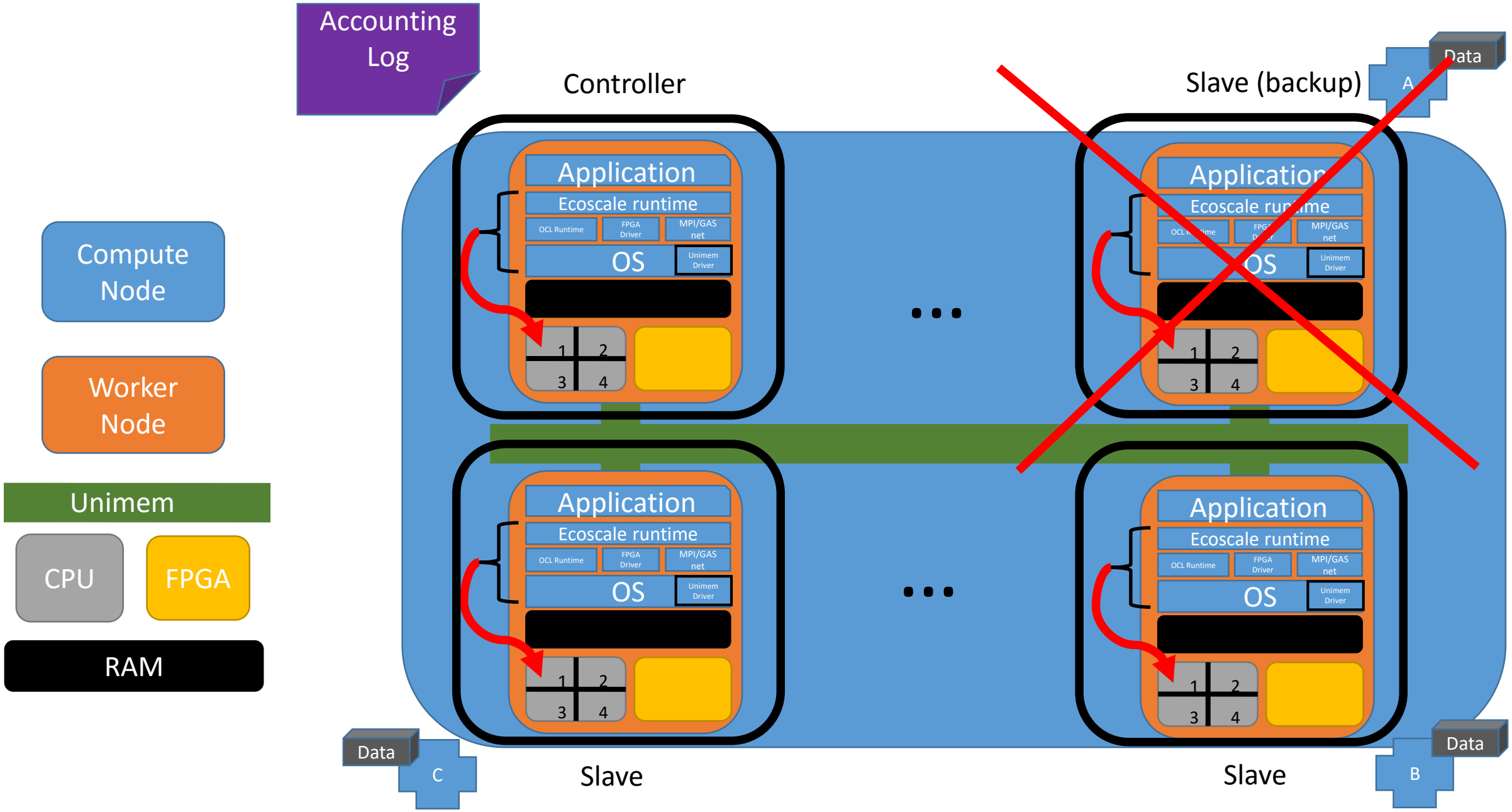
Leadership Election

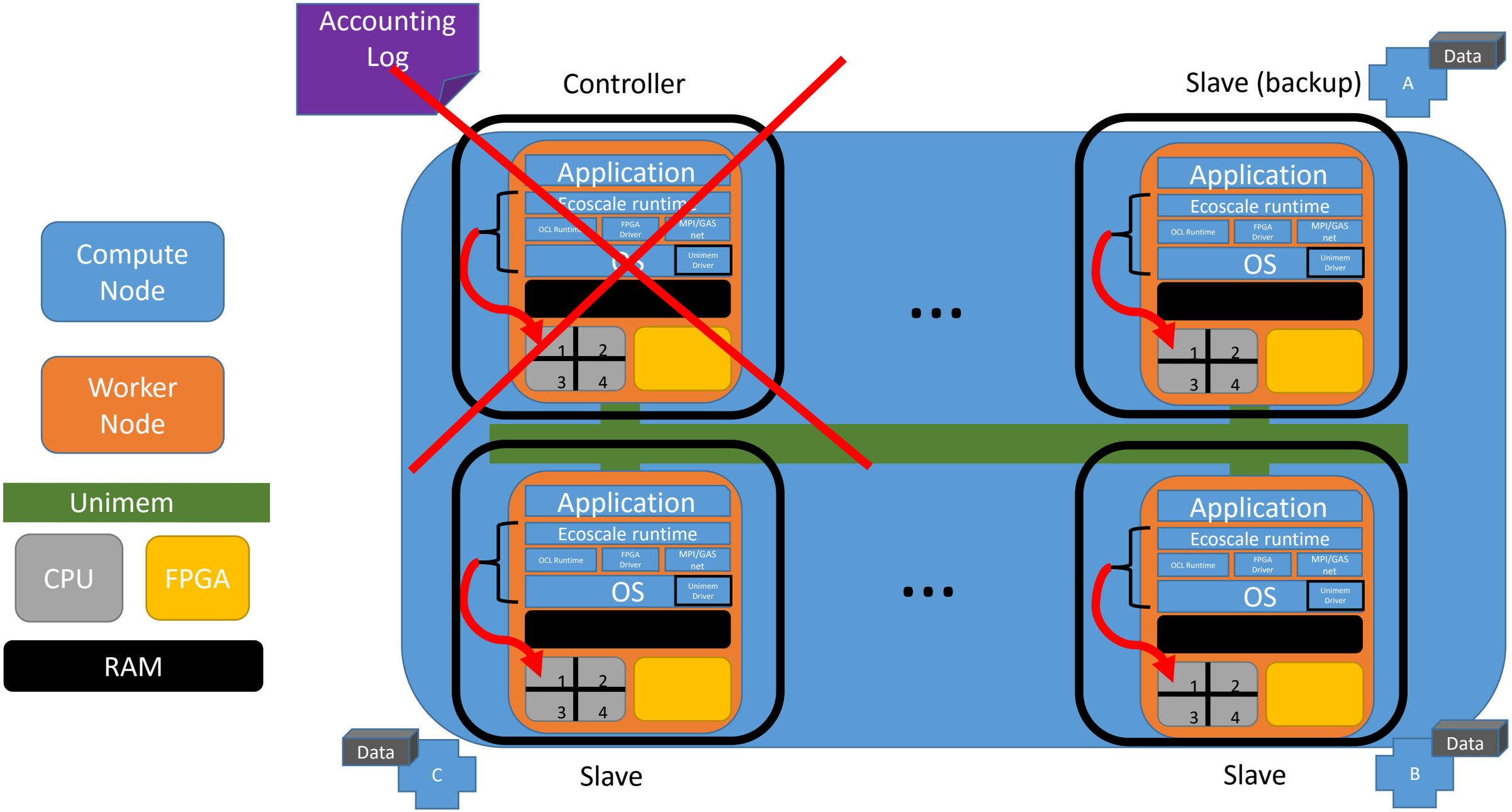


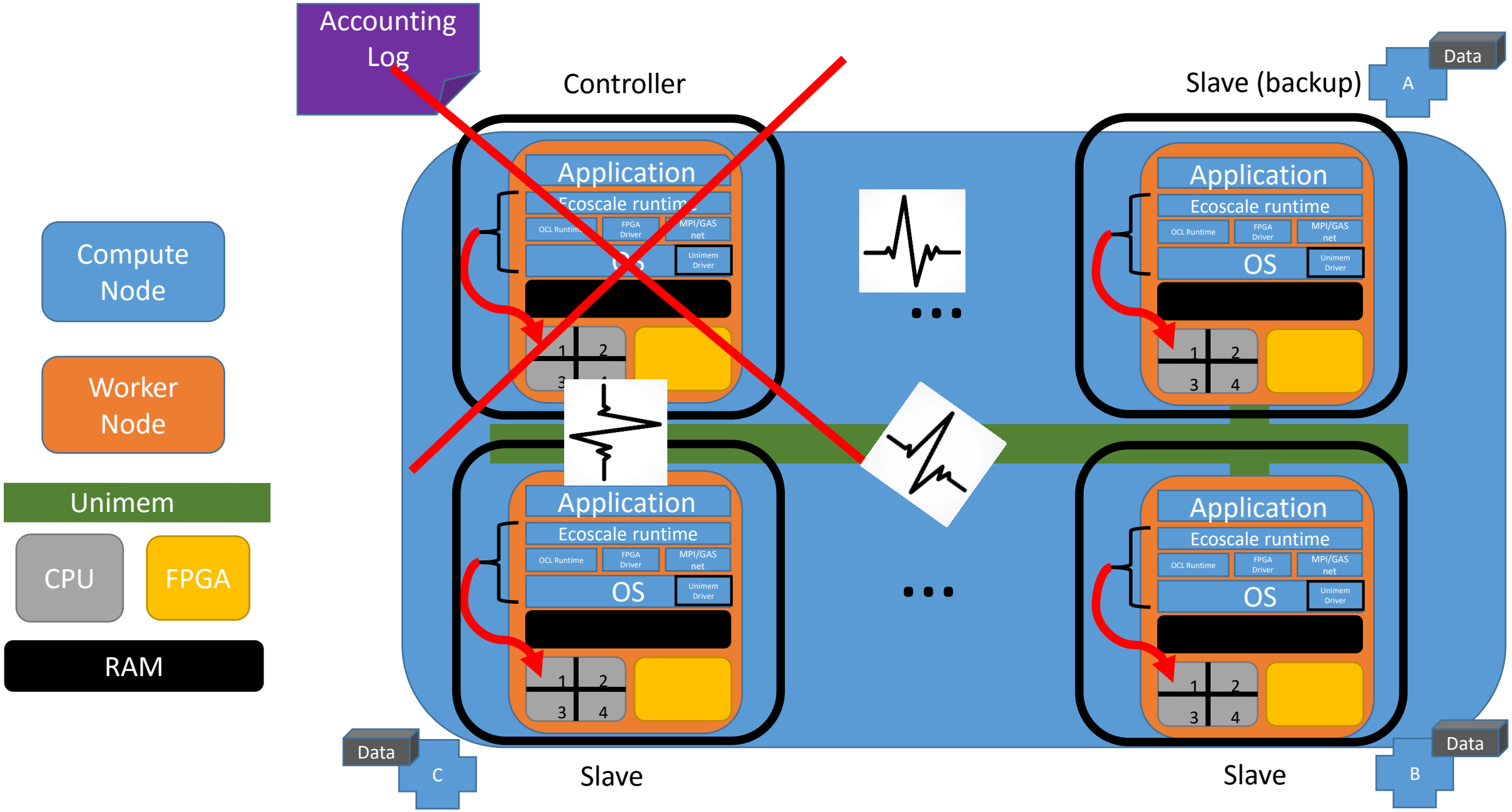

Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Controller
Slave (Backup)
Slave
Slave
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4


Data
Data
Data
C
B
A
Accounting Log
Controller
Slave (Backup)
Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4
Slave
Slave

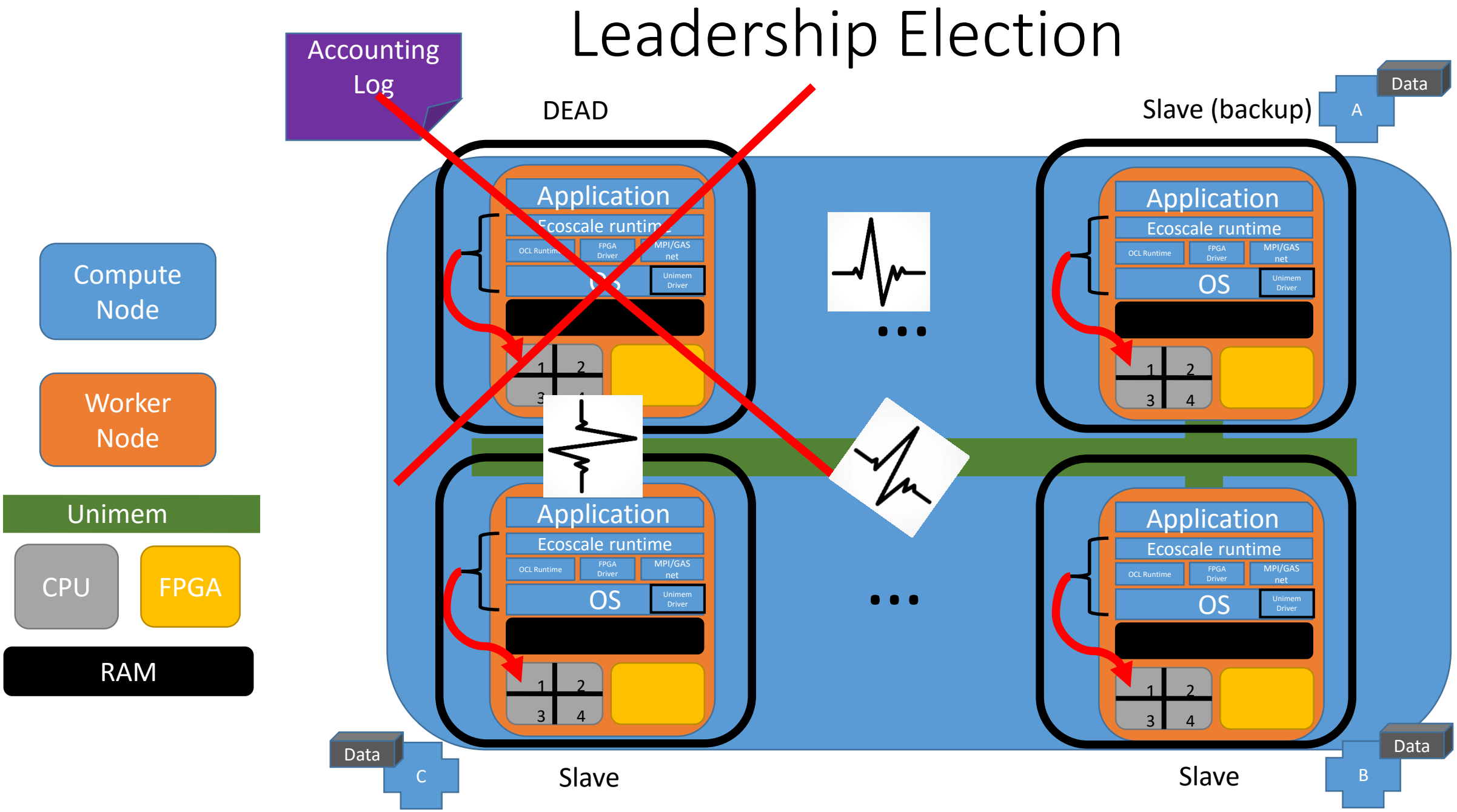

Accounting Log
Controller
Slave (Backup)
Data
A
Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4
Data
C
Slave
Slave
Data
B

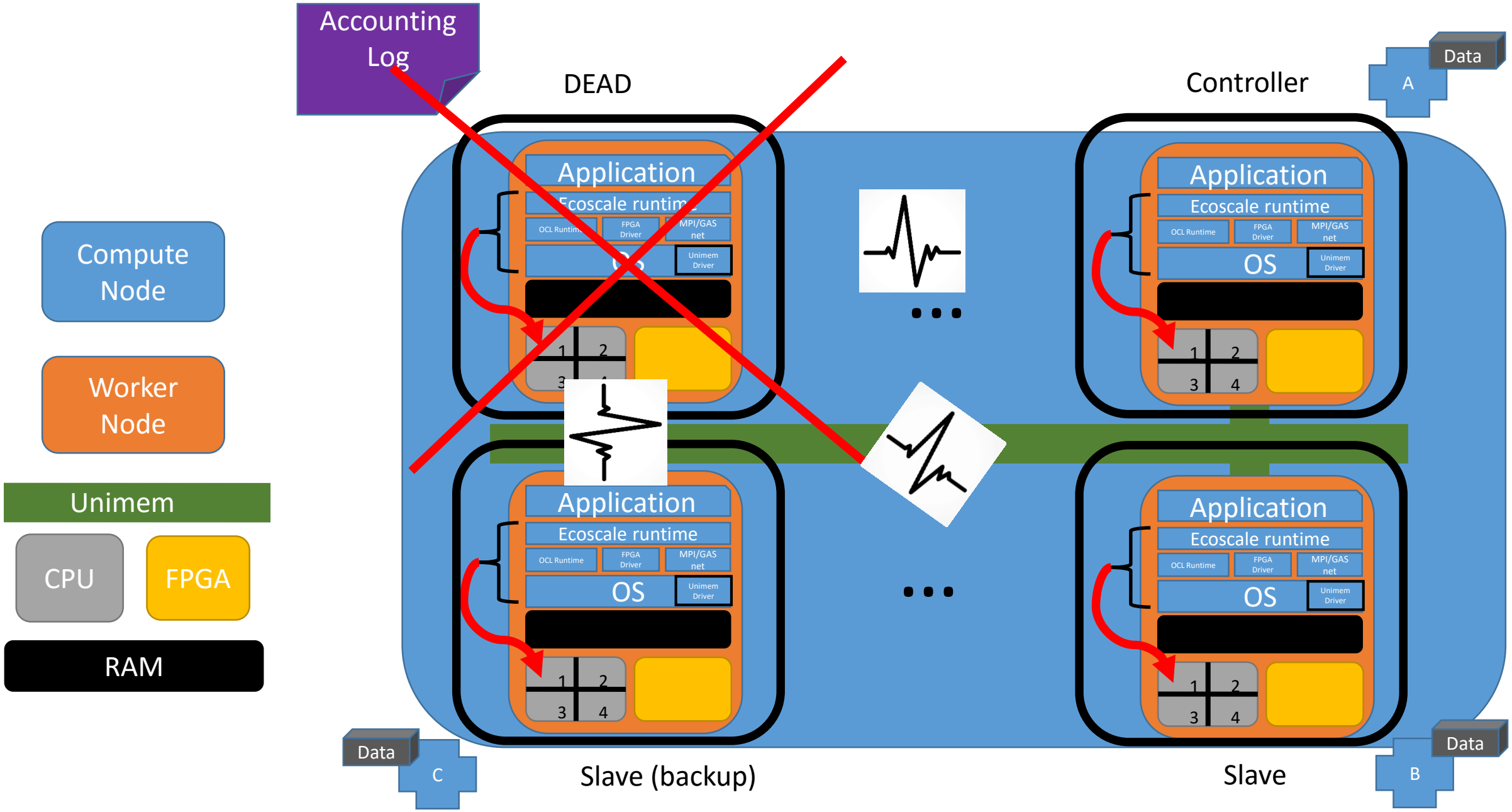

Accounting Log
Controller
Slave (backup)
Data
A
Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4
Data
C
Slave
Slave
Data
B


Accounting Log
Controller
Slave (backup)
Data
A
Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4
Data
C
Slave
Slave
Data
B


Accounting Log
Controller
Slave (backup)
Data
A
Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4
Data
C
Slave
Slave
Data
B

Leadership Election

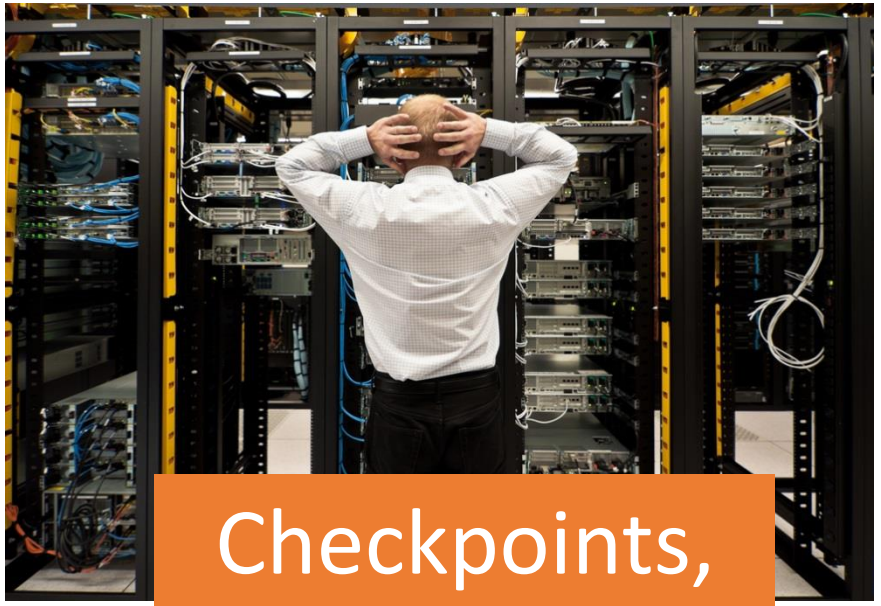



Accounting Log
DEAD
Controller
Data
A
Compute Node
Worker Node
Unimem
CPU
FPGA
RAM
Application
Ecoscale runtime
OCL Runtime
FPGA Driver
MPI/GAS net
OS
Unimem Driver
1
2
3
4
Data
C
Slave (backup)
Slave
Data
B

Exascale: Problems Solved?

Extended OpenCL

FPGA

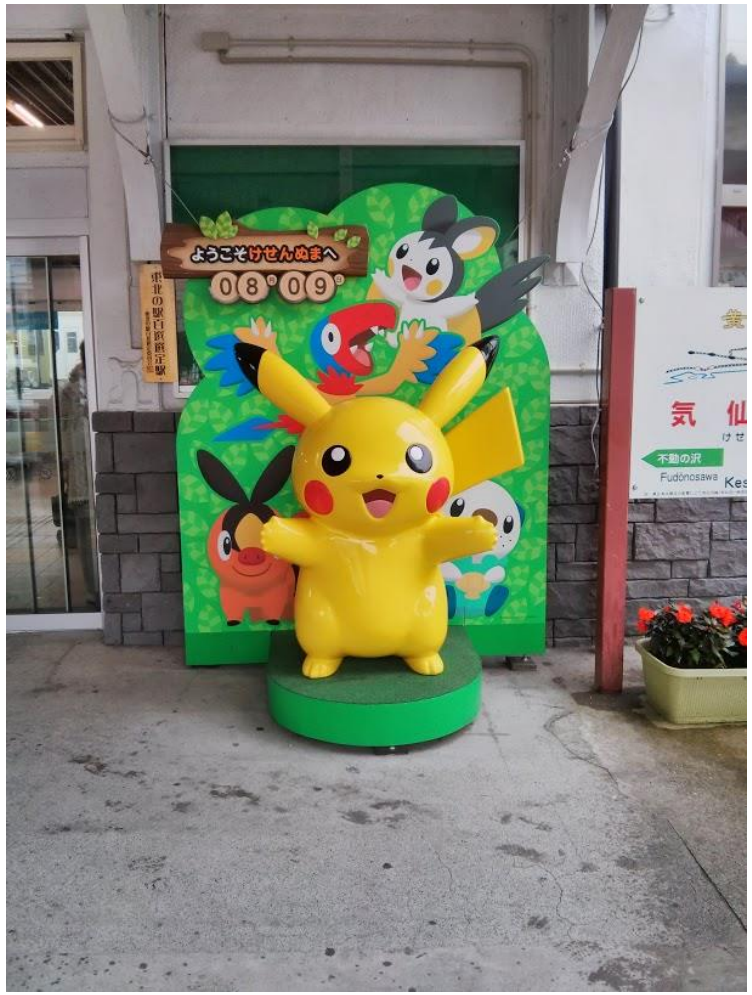


Checkpoints, Heartbeats, and internal monitors

http://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf

Ideas?

ありがとうございました!



質問はありますか

@jhebus
Paul-Harvey.org