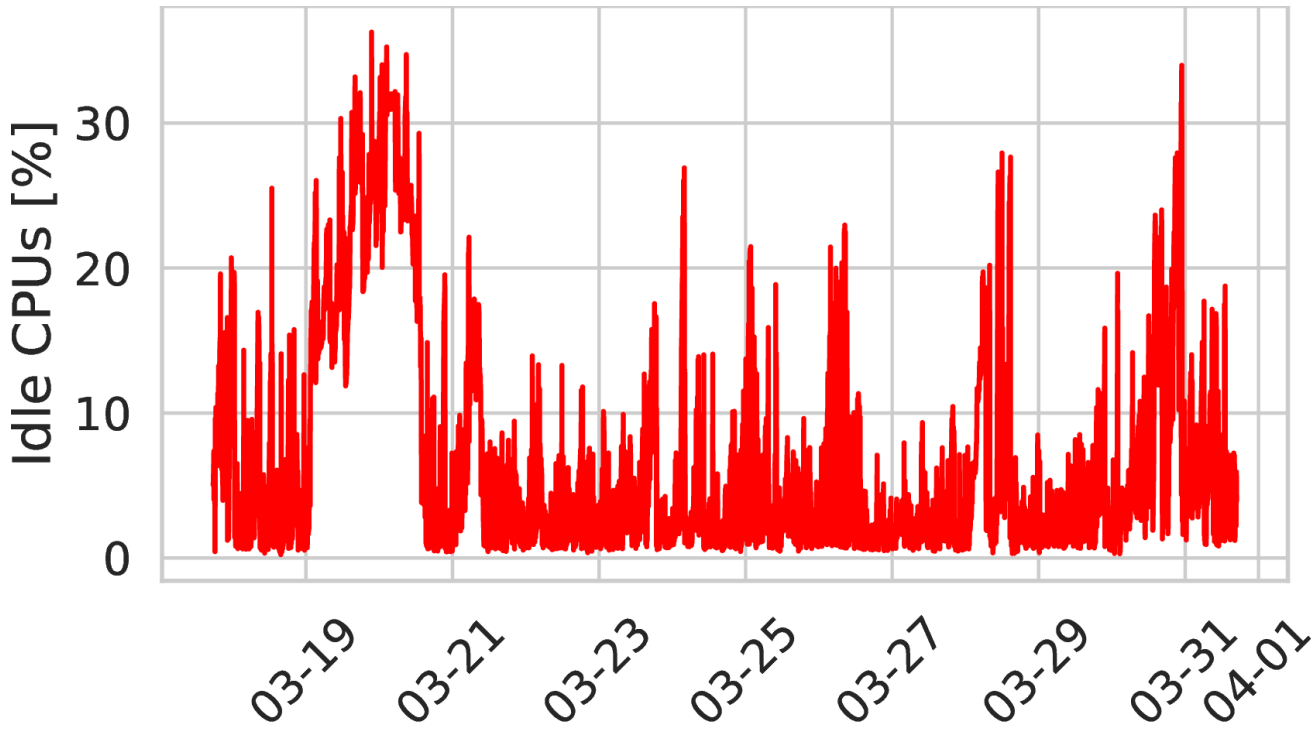
Software Resource Disaggregation for HPC with Serverless Computing

Marcin Copik, Marcin Chrapek, Alexandru Calotoiu, Torsten Hoefler

HPC System Utilization

Piz Daint, April 2022.
- XC50 nodes – CPU + GPU, 64 GB memory.
- XC40 nodes – CPU, 64/128 GB memory.

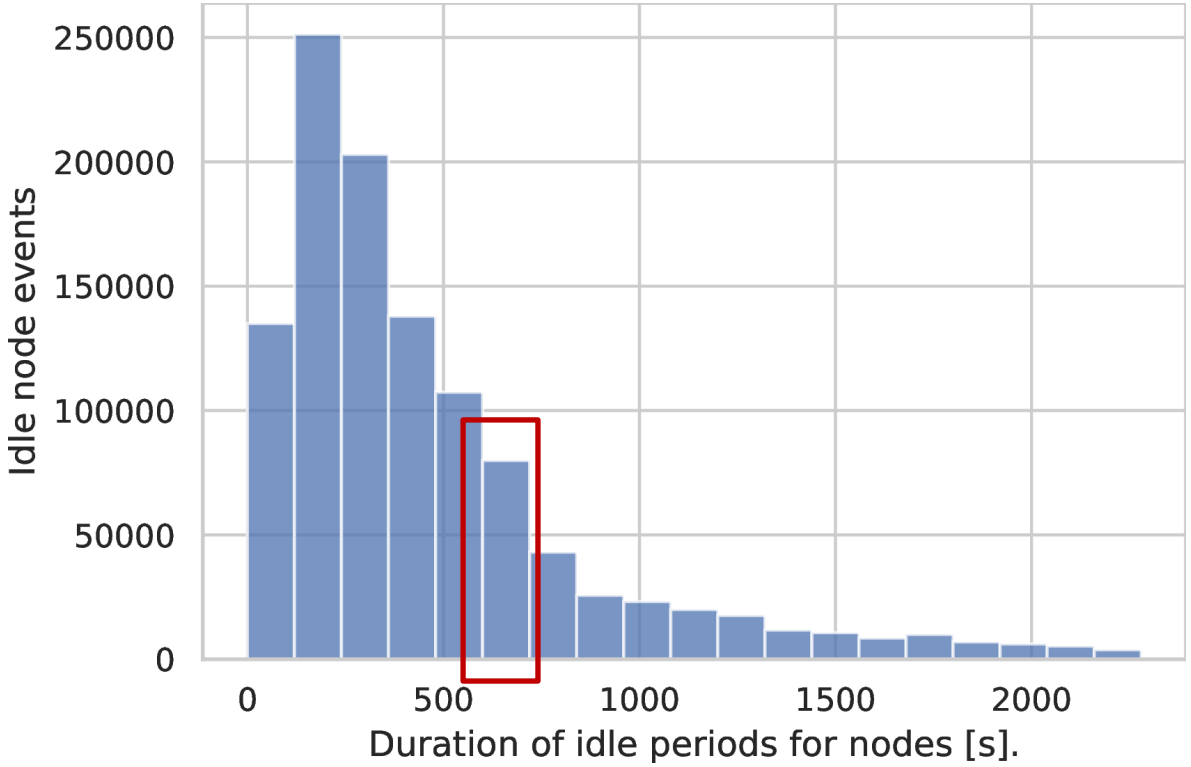
Query SLURM info every two minutes.



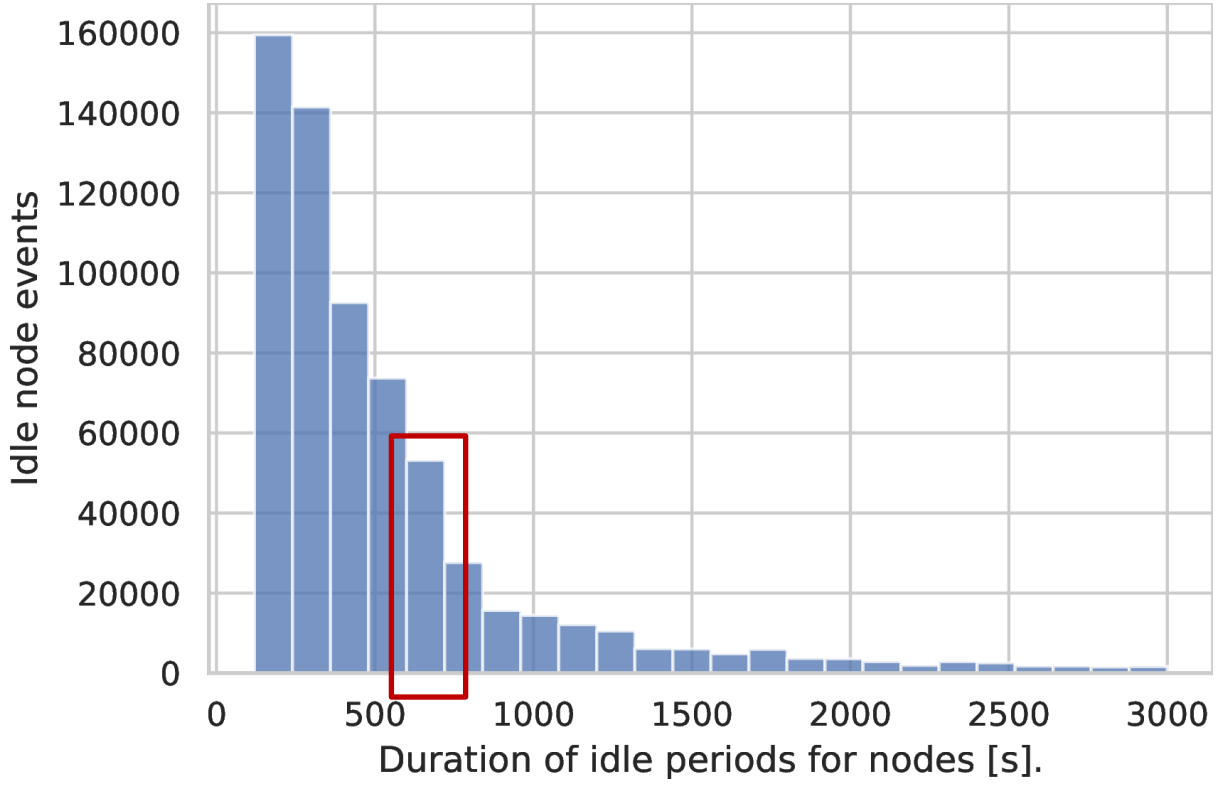
How long do nodes stay idle?

HPC System Utilization - CPU

Minimum

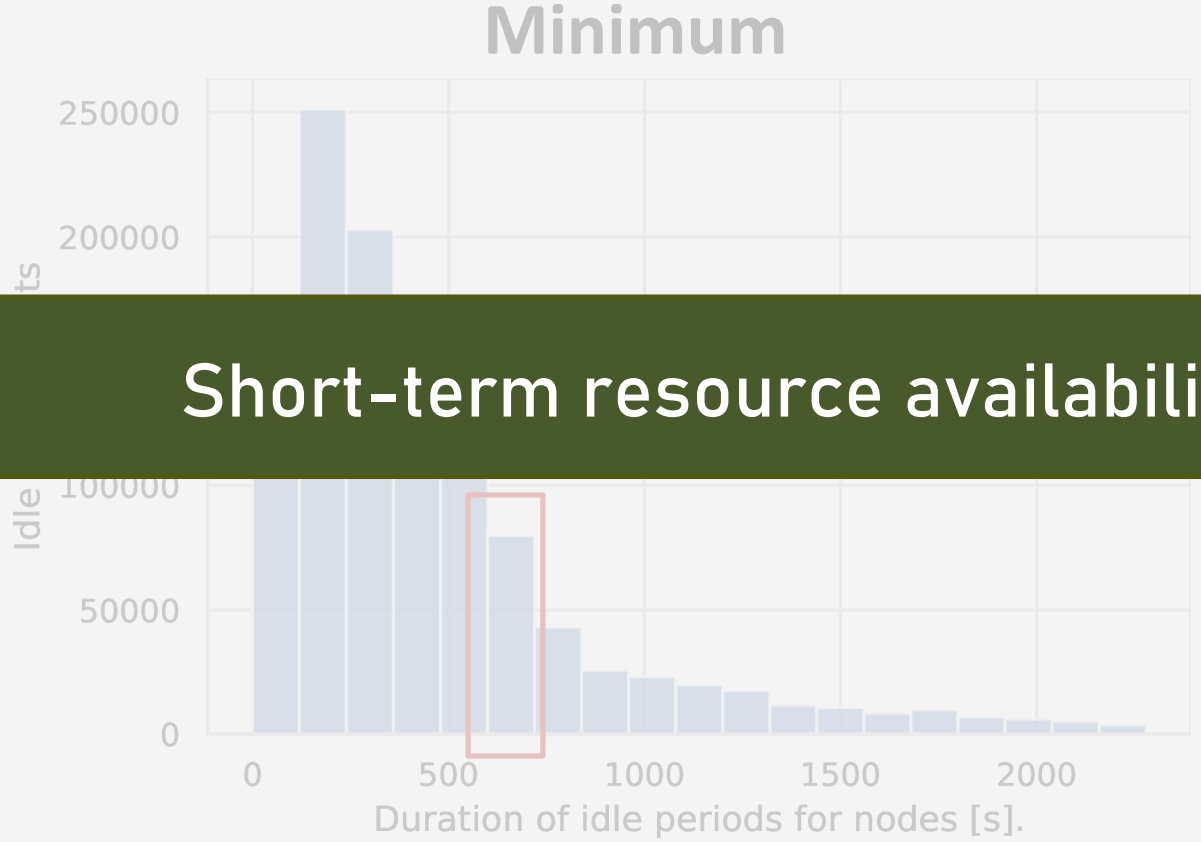


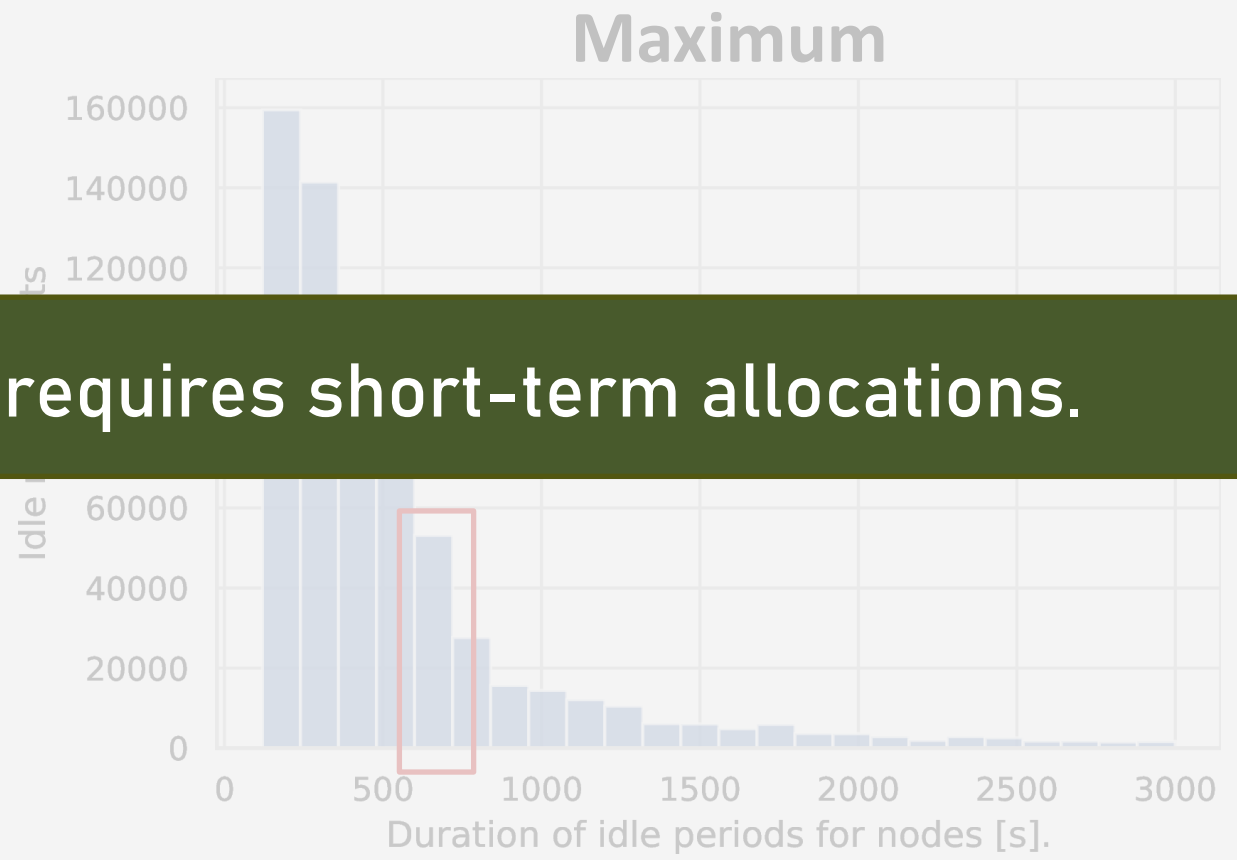
Maximum



80% and 70% of idle node events last less than 10 minutes.

HPC System Utilization - CPU

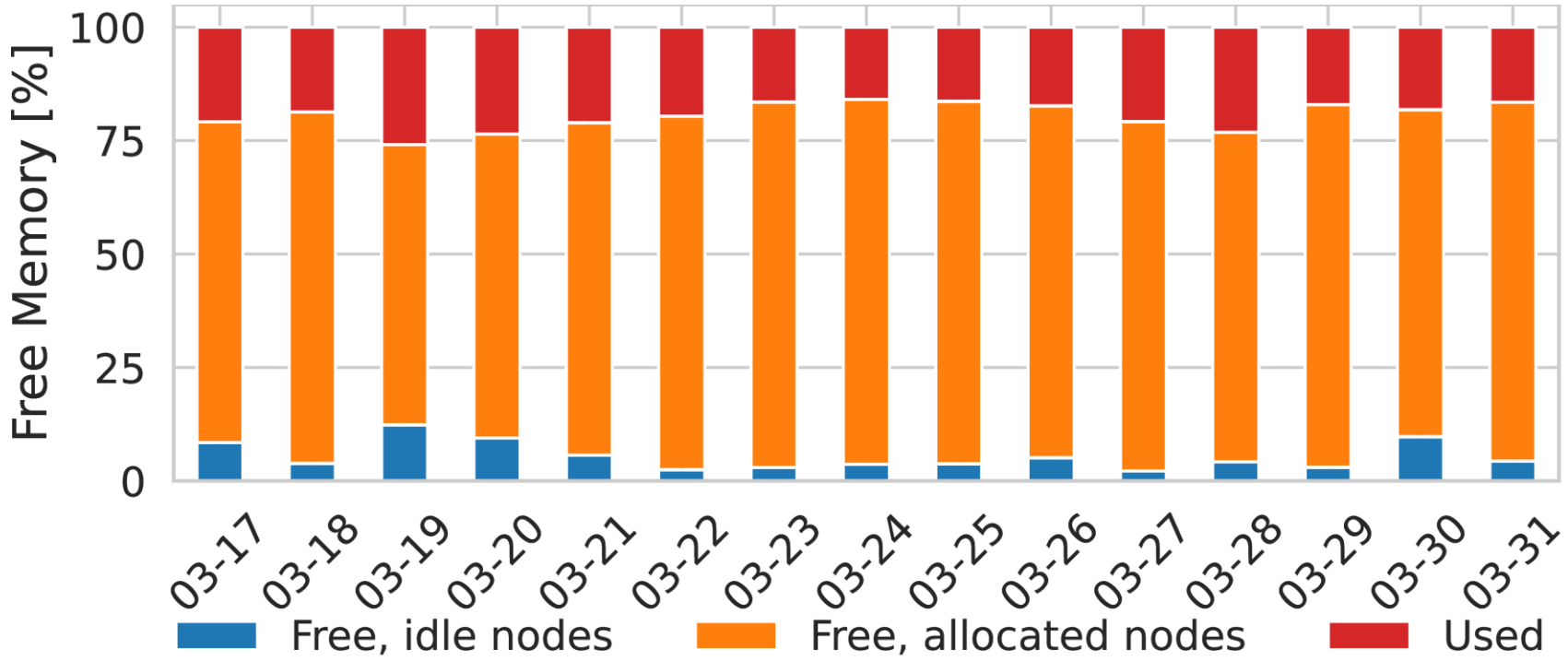




Short-term resource availability requires short-term allocations.

80% and 70% of idle node events last less than 10 minutes.

HPC System Utilization - Memory

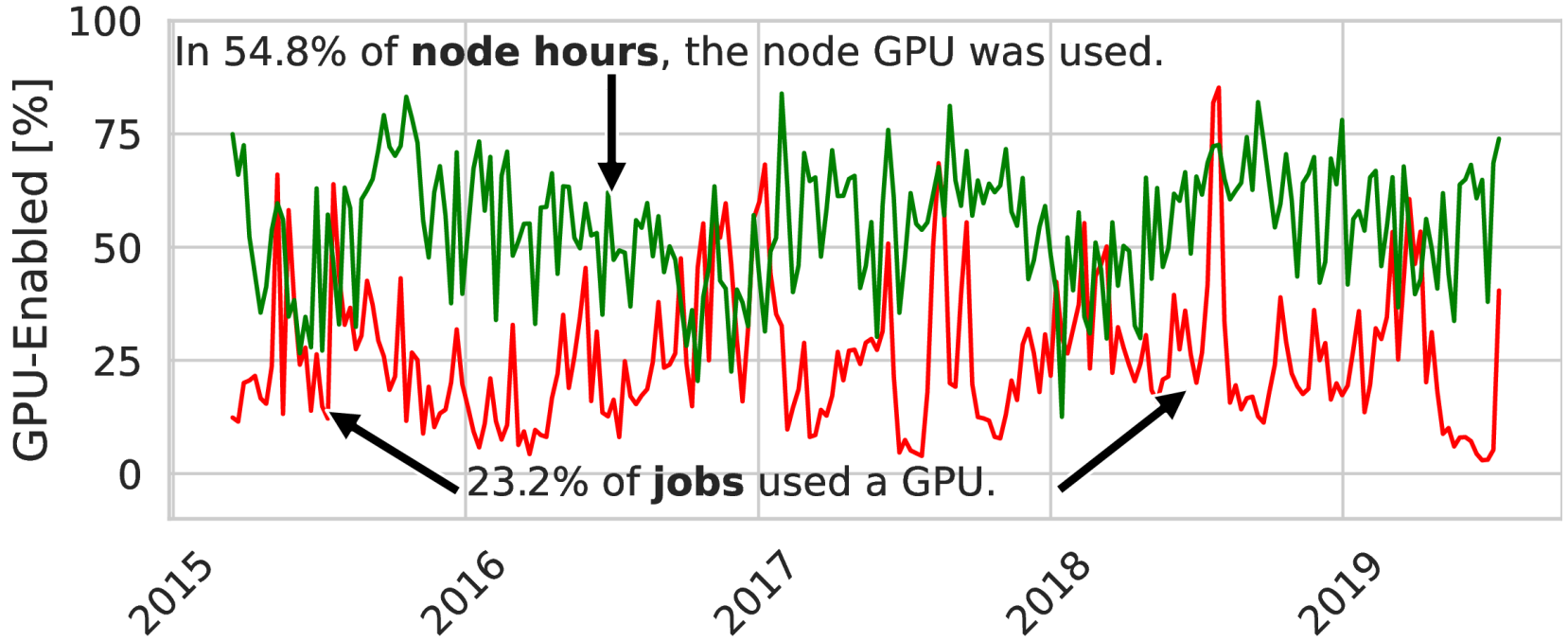


HPC System Utilization - GPU

Learning from Five-year Resource-Utilization Data of Titan System

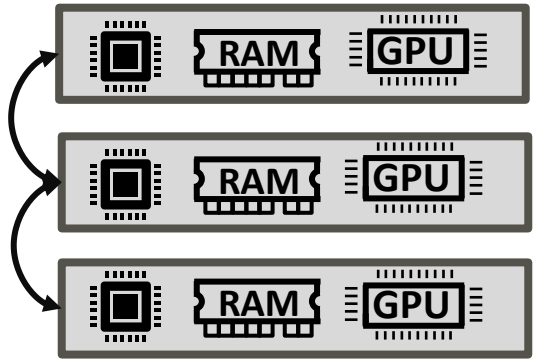
Feiyi Wang*, Sarp Oral†, Satyabrata Sen ‡ and Neena Imam§
Oak Ridge National Laboratory

CLUSTER, 2019



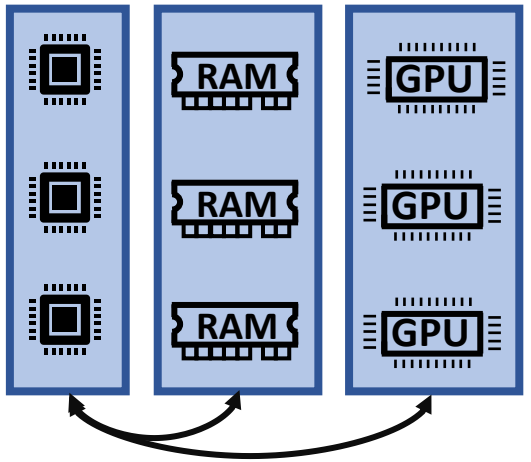
Software Solution

Standard HPC Node



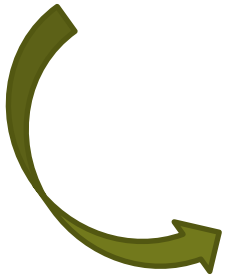
- High performance
- Inflexible architecture

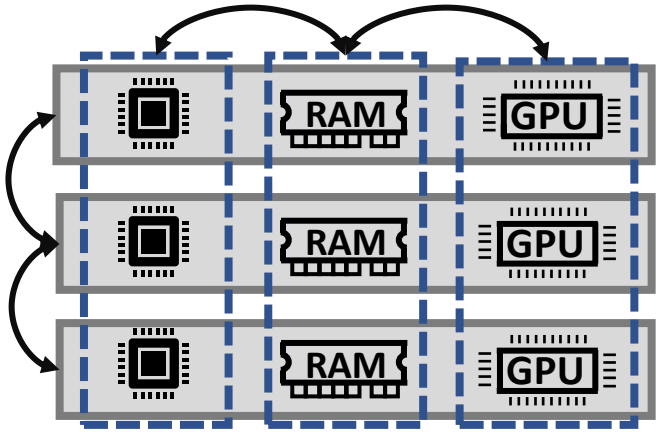
Hardware Disaggregation

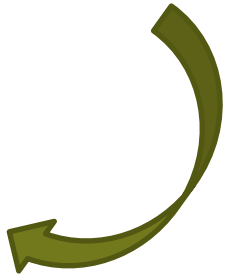


- High efficiency
- Cost, performance penalty

Existing Coupled Hardware Systems







Software Abstraction for Disaggregation

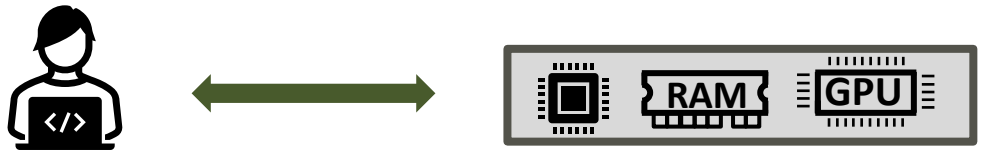
We propose a software disaggregation approach to share node resources between
coarse-grained, long-running, and static batch jobs
and
fine-grained, short-term, and dynamically allocated serverless functions.

Serverless as an Answer

Serverless as an Answer

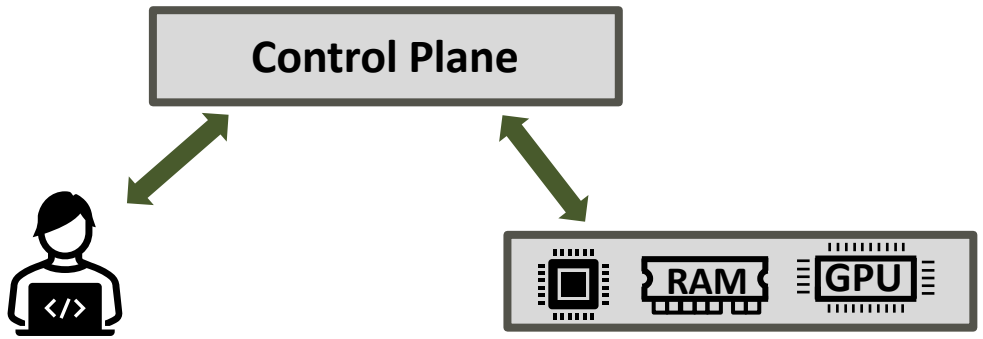
Hardware Abstraction

Software Abstraction



Serverless as an Answer

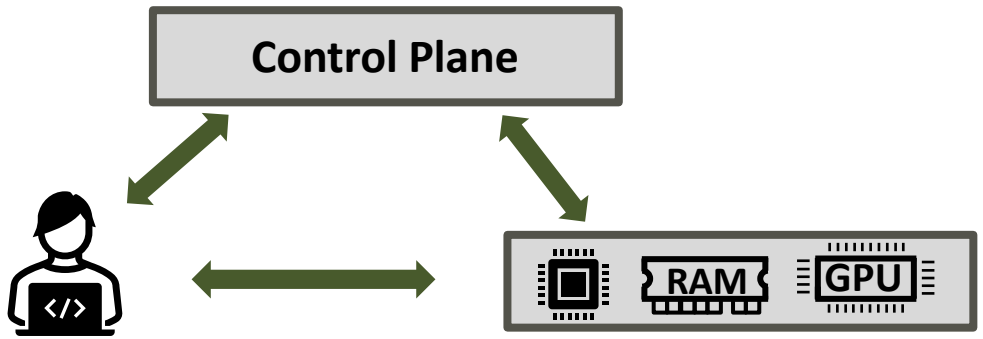
Hardware Abstraction



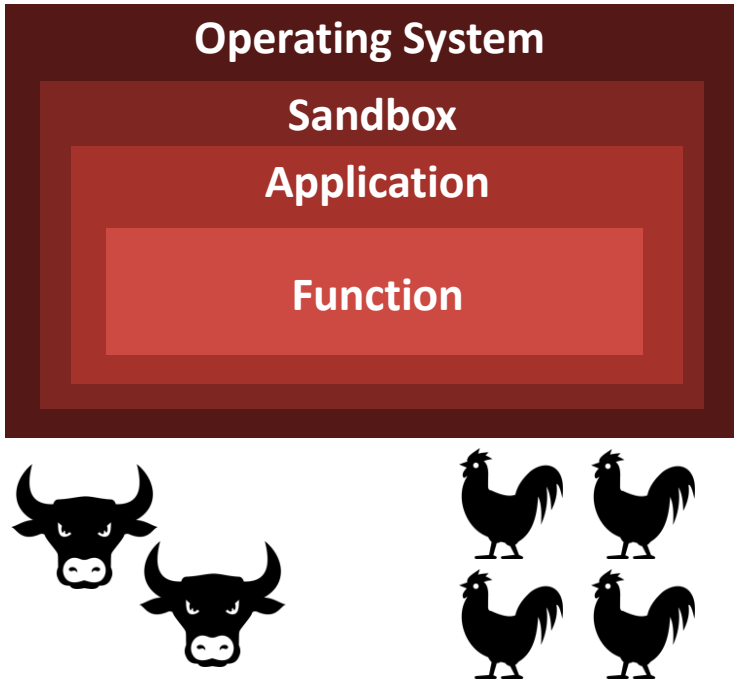
Software Abstraction

Serverless as an Answer

Hardware Abstraction

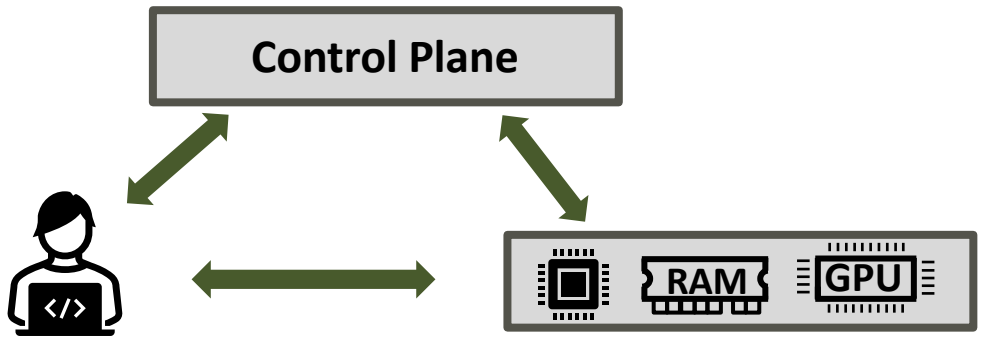


Software Abstraction



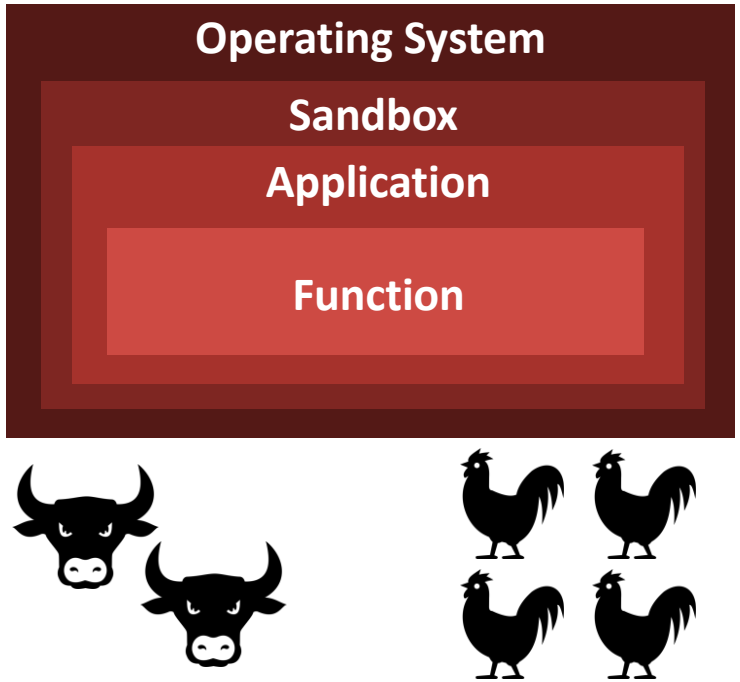
Serverless as an Answer

Hardware Abstraction



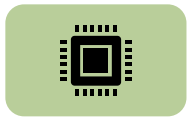
Pay-as-you-go billing

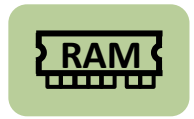
Software Abstraction

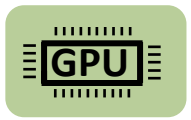


Granular computing

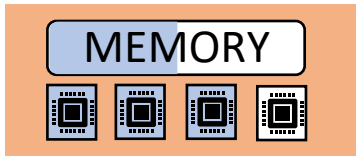
Serverless Disaggregation



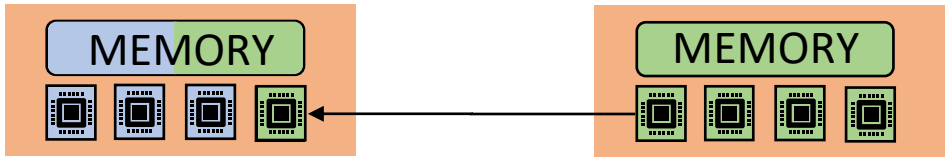




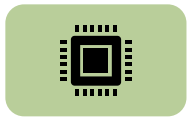
Batch jobs

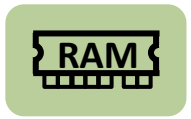


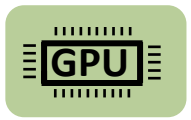
Batch jobs + serverless functions



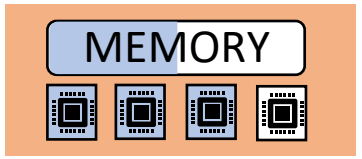
Serverless Disaggregation

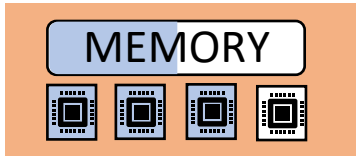




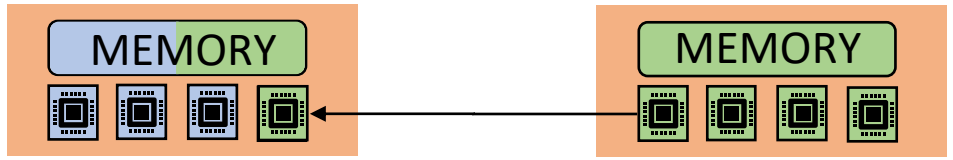


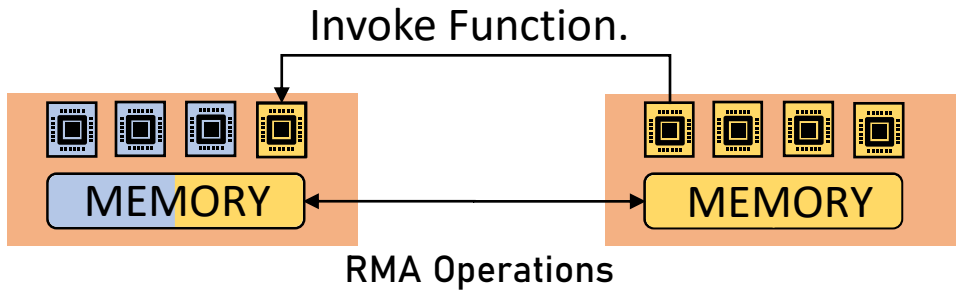
Batch jobs



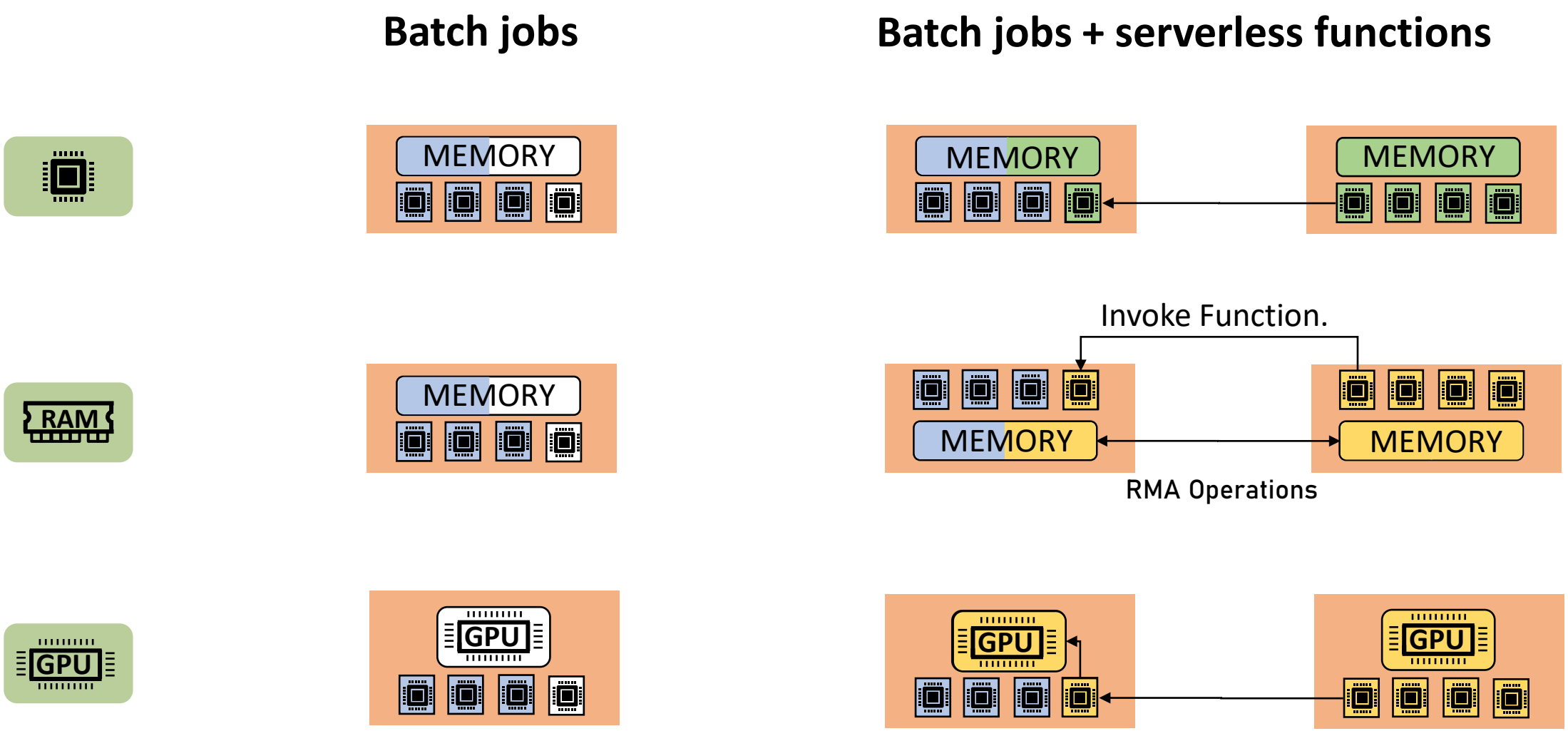


Batch jobs + serverless functions

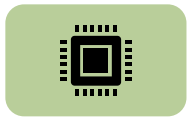


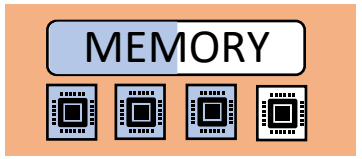


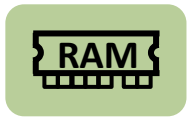
Serverless Disaggregation

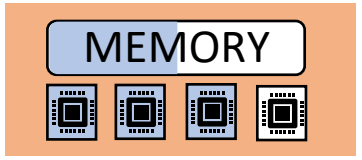


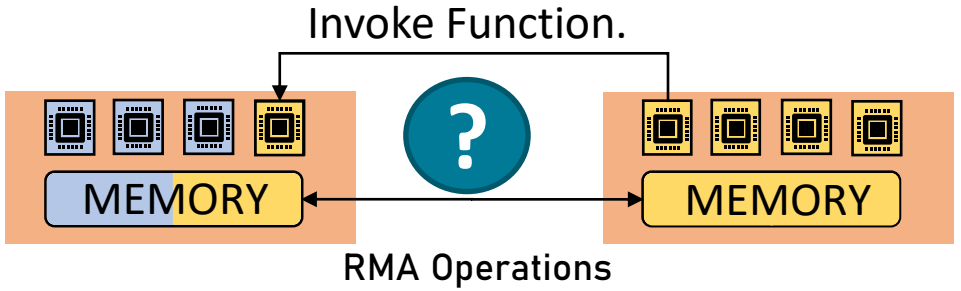
Serverless Disaggregation

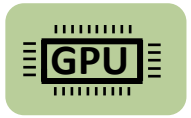


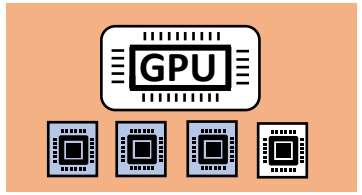


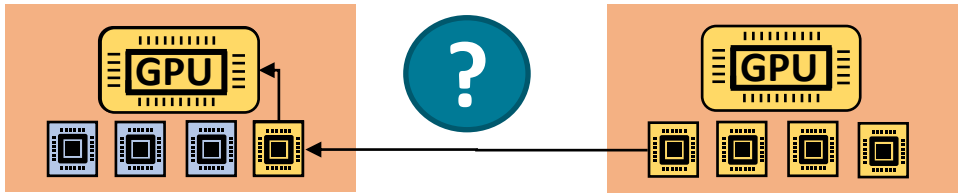








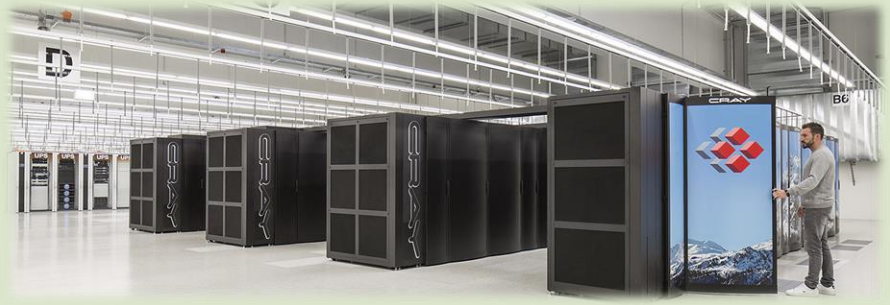




Batch jobs

Batch jobs + serverless functions

Evaluation



XC50 nodes - 12 CPU cores, GPU, 64 GB memory.

XC40 nodes - 36 CPU cores, 64/128 GB memory.

Cray Aries interconnect.

36 CPU cores, 377 GB memory.
Ethernet with RoCEv2 support.

#1 CPU Sharing

collocated benchmark type	Mean utilisation: Disaggregation	Mean utilisation: Ideal Non-sharing	Mean utilisation: Realistic	Total time: Disaggregation	Total time: Realistic	Core hours: Disaggregation	Core hours: Ideal Non-sharing	Core hours: Realistic
BT, A -	0.937	0.893	0.693	0.877	1.0	0.968	1.0	1.29
BT, W -	0.903	0.89	0.64	0.981	1.0	0.994	1.0	1.39
CG, B -	0.992	0.901	0.65	0.94	1.0	0.908	1.0	1.39
EP, B -	0.915	0.891	0.661	0.901	1.0	0.98	1.0	1.35
LU, A -	0.941	0.893	0.674	0.929	1.0	0.964	1.0	1.33
MG, A -	0.903	0.89	0.625	1.01	1.0	1.01	1.0	1.42
MG, W -	0.903	0.89	0.638	1.01	1.0	1.0	1.0	1.39

LULESH

64 ranks, 2 nodes

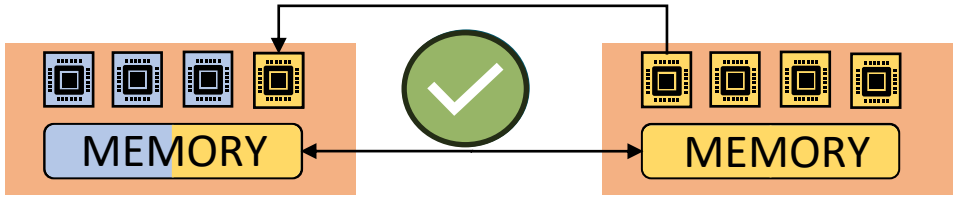
32 out of 36 cores allocated.

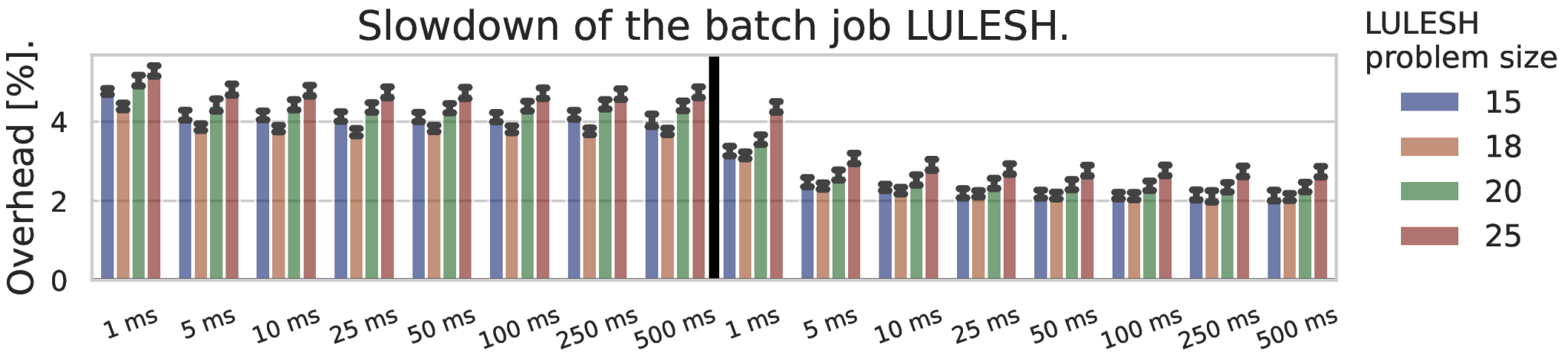
NAS

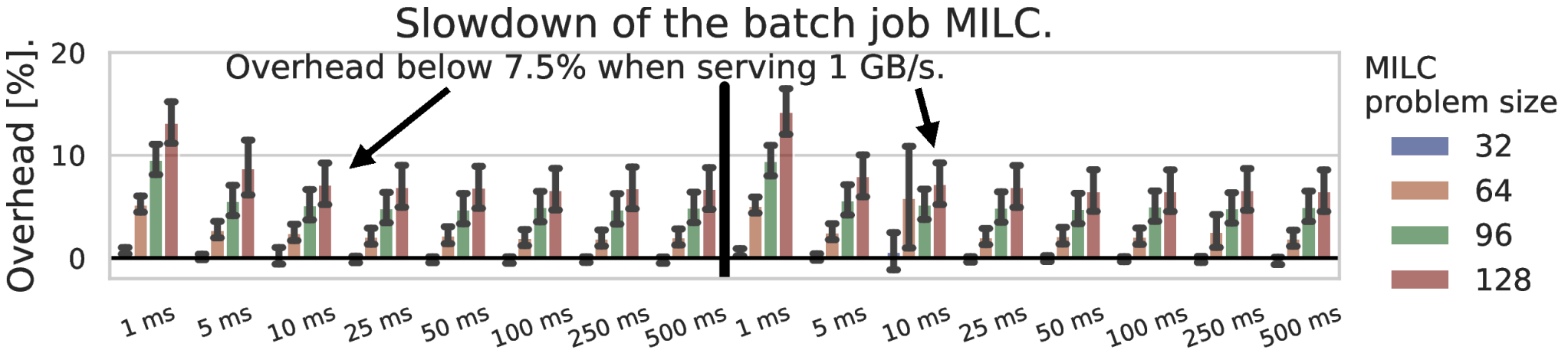
1 – 4 ranks

Distributed across nodes.

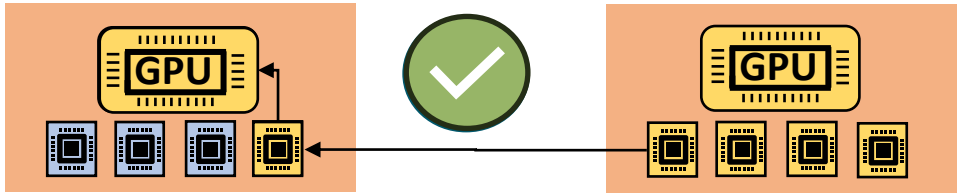
#2 Serving Remote Memory

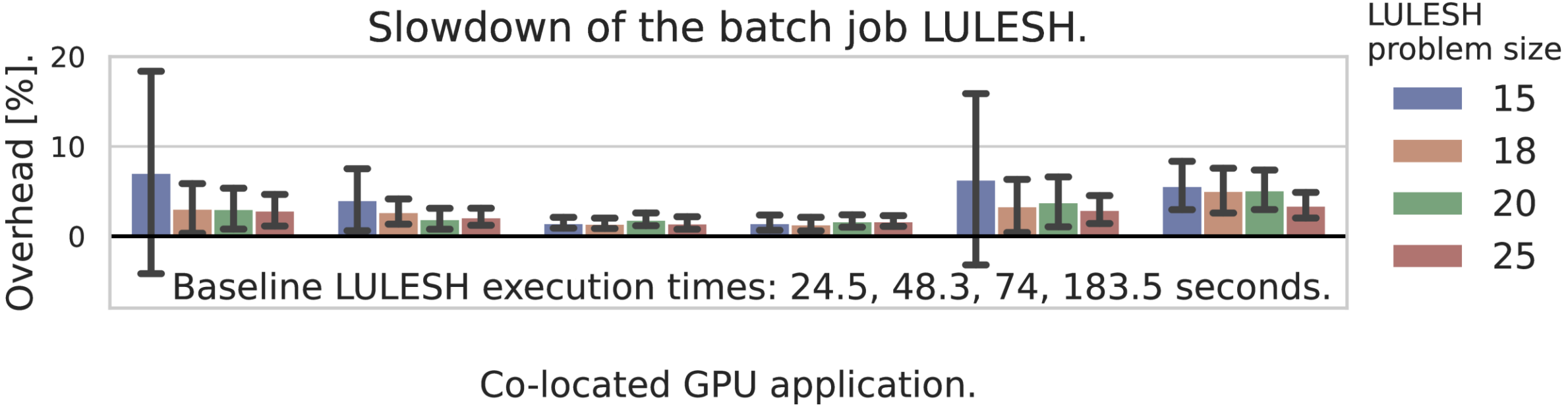






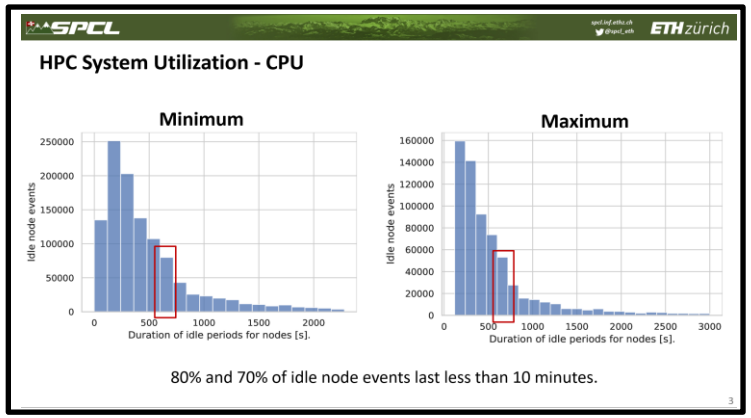
#3 Co-locating GPU and CPU workloads

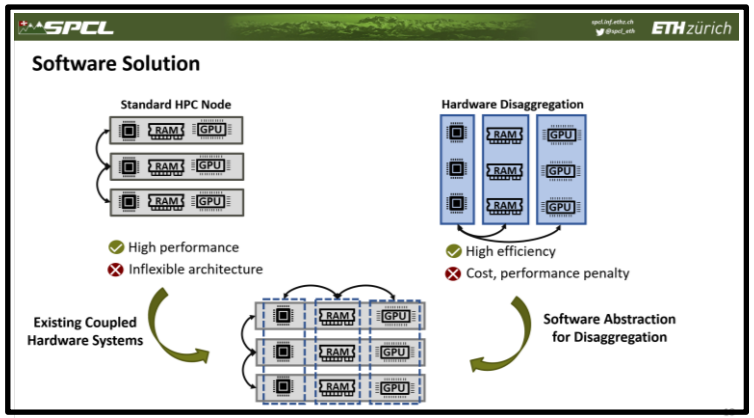


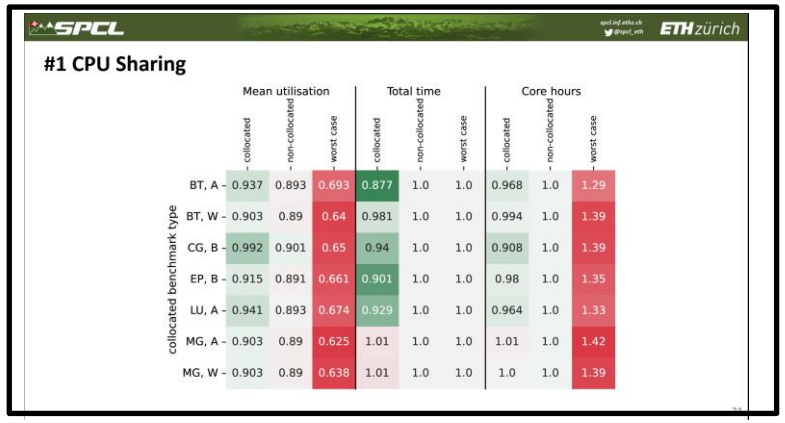


LULESH – 27 ranks, 3 nodes, 9 out of 12 cores allocated.
Rodinia – 1 MPI rank, 1 GPU.

Summary







spcl/rFaaS

"the goal of achieving near 100% utilization while supporting a real parallel supercomputing workload is unrealistic"

Scheduling for Parallel Supercomputing: A Historical Perspective of Achievable Utilization

James Patton Jones[1] and Bill Nitzberg[1]

MRJ Technology Solutions
NASA Ames Research Center, M/S 258-6
Moffett Field, CA 94035-1000

jjones@nas.nasa.gov

1999