Effective Use of HPC

Common misconception

Reality

- It doesn’t magically make your program run super fast
- We get many of small computers and run parallel by
  - Code change
  - Distributing loads

Good “utilization” of resource

- Make all the resource as busy as possible
- Unfortunately, not always easy : often a problem is inherently not parallel-izeable.
  - Eg. You can’t have 9 women pregnant to have a baby in 1 month
- If can’t do quicker, can do more?
  - Eg. Can we have 9 babies in 9 month?

Better utilization of computing resource

- Understand your problem
- Understand your HPC systems

What can happen

Mahuika vs Maui (482nd (11/2018) @top500)

Mahuika		Maui
8,136 (16,272 with Hyperthreading)	Total cores	18,560 (37,120 with Hyperthreading)
36 cores per node (2.1Gh Broadwell)	CPU per node	40 cores per node (2.4 GHz Skylake)
Capacity: Designed to solve many small problems from many users	Purpose	Capability : Designed to solve a few large problems with very fast interconnect between nodes
CentOS 7.4 Generally better open source support	OS	Cray Linux Environment Can be quirky
128Gb +	Memory per node	96Gb (232 nodes) 182 Gb (232 nodes)
IOPS optimized (good for many small data?)	IO	High Bandwidth to disk (good for large data?)
1 core	Minimum size	1 node (40 cores)

Policy

- Login node is NOT HPC : Never run anything heavy on the login node !
- Submit a job via SLURM
- Mahuika
  - 20,000 CPU-hours maximum, No user put 1000 jobs in the queue
- Maui
  - 66 nodes maximum, 168-node-hour maximum, 20 jobs in queue

Hyperthreading: fake core doubling

- 1 physical core becomes 2 logical cores
- 3 GB per physical core becomes 1.5 GB per logical core

Storage

Filesystem	/home	/nesi/project	/nesi/nobackup	/nesi/nearline (Not yet available)
Default disk space quota	20 GB	100 GB (per project)	No limit	No limit
Default file count quota (inodes)	100,000 files	100,000 files	1,000,000 files	500,000 files, each no smaller than 5 MB
Intended use	User-specific files such as configuration files, environment setup, source code, etc.	Persistent project-related data	Data created by compute jobs that is intended to be temporary	Long-term archive storage
Total capacity	175 TB	1,590 TB	4,400 TB	>100 PB (media funded by projects)
Expiry	When the user is no longer a member of any active project	90 days after the end of the project	When the Librarian service is available files will deleted after being untouched for 60 days (or earlier if space is required)[3]	365 days after the end of the project (unless agreed otherwise)
Data Backup	Daily, last 10 versions of any given file retained for up to 90 days.	Daily, last 10 versions of any given file retained for up to 90 days.	None	Replicated to offsite tape library
Snapshots	Daily (retention period, 7 days)	None	None	None
Access Speed	Moderate	Moderate	Fast	Slow. Only accessible via the Librarian Service.

Mahuika Partitions

Name	Max Walltime	Nodes	CPUs/Node	Available Mem/CPU	Available Mem/Node	Fairshare Weight	Description
large	3 days	226	72	1500 MB	108 GB	1	Standard partition.
long	3 weeks	69	72	1500 MB	108 GB	1	For jobs that need to run for longer than 3 days.
prepost	3 hours	5	72	6800 MB	480 GB	1	Use for pre and post processing tasks in a workflow.
bigmem	7 days	4	72	6800 MB	480 GB	2	Partition for jobs requiring large amounts of memory.
hugemem	7 days	0.5	128	30 GB	4,000 GB	4	Can be used to run jobs that need up to 2 TB of memory.
gpu	3 days	4	8	13500 MB	108 GB	56 / GPU	See below for more info.

Maui Partitions

Māui (XC50) Slurm Partitions

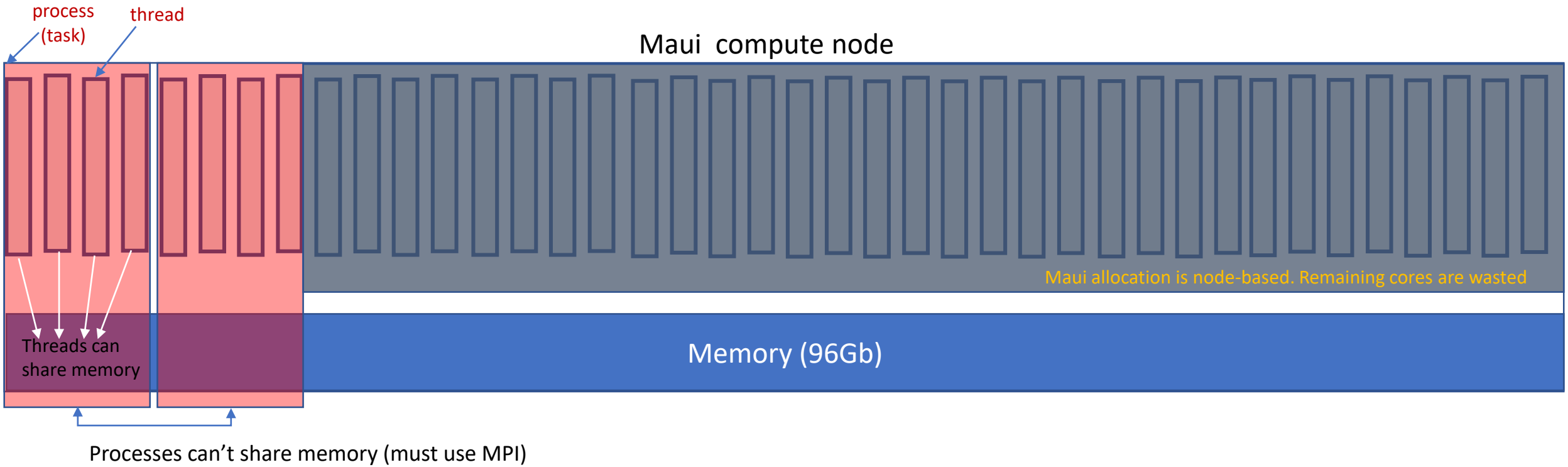
Name	Nodes	Max Walltime	Avail / Node	Max / Job	Max / User	Max in Queue	Description
nesi_research	264	24 hours	80 CPUs 80 or 160 GB RAM	66 nodes 168 node-hours	8 jobs	20 jobs	Standard partition for all jobs.

maui_ancil Slurm Partitions

Partition	Nodes	Max Walltime	Avail / Node	Max / Job	Max / User	Description
nesi_prepost	4	24 hours	80 CPUs 720 GB RAM	20 CPUs 700 GB RAM	80 CPUs 700 GB RAM	Pre and post processing tasks.
nesi_gpu	4 to 5	72 hours	4 CPUs 12 GB RAM 1 GPU	4 CPUs 12 GB RAM 1 GPU	4 CPUs 12 GB RAM 1 GPU	GPU jobs and visualisation.

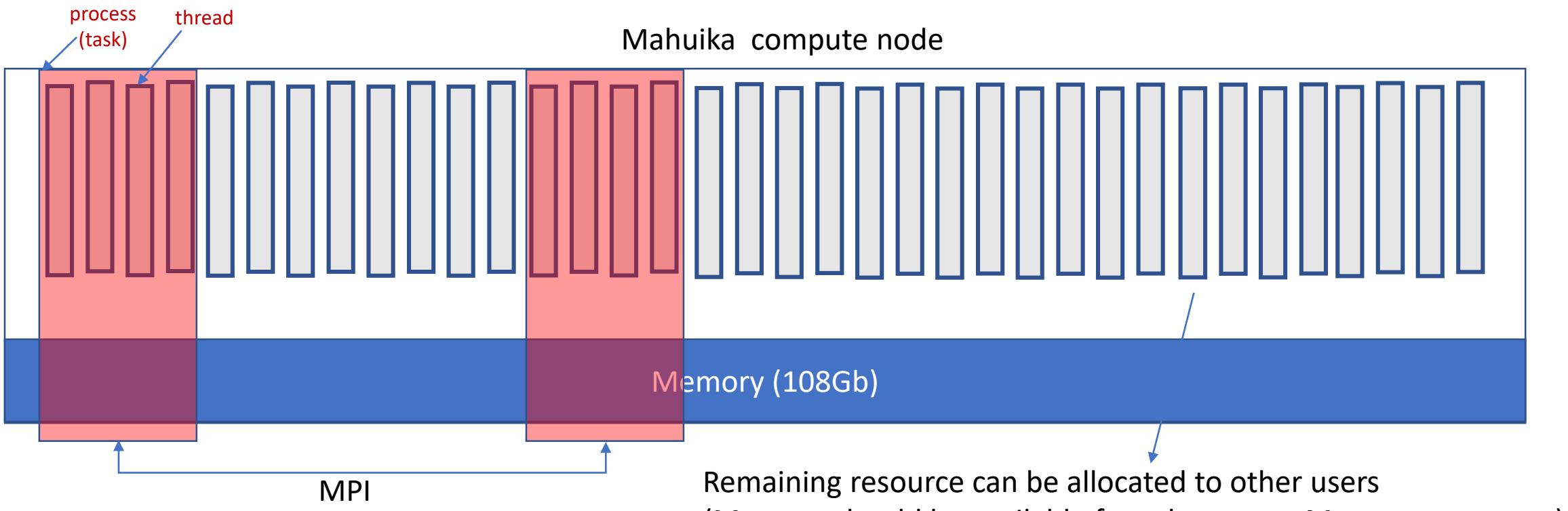
SLURM

```
#!/bin/bash
#SBATCH --job-name=JobName # job name (shows up in the queue)
#SBATCH --account=nesi99999 # Project Account
#SBATCH --time=03:00:00 # Walltime (HH:MM:SS)
#SBATCH --mem-per-cpu=1500 # memory/cpu (in MB)
#SBATCH --ntasks=2 # number of tasks (e.g. MPI)
#SBATCH --cpus-per-task=4 # number of cores per task (e.g. OpenMP)
#SBATCH --partition=nesi_research # specify a partition
#SBATCH --hint=nomultithread # don't use hyperthreading srun [options]
<executable> [options]
```



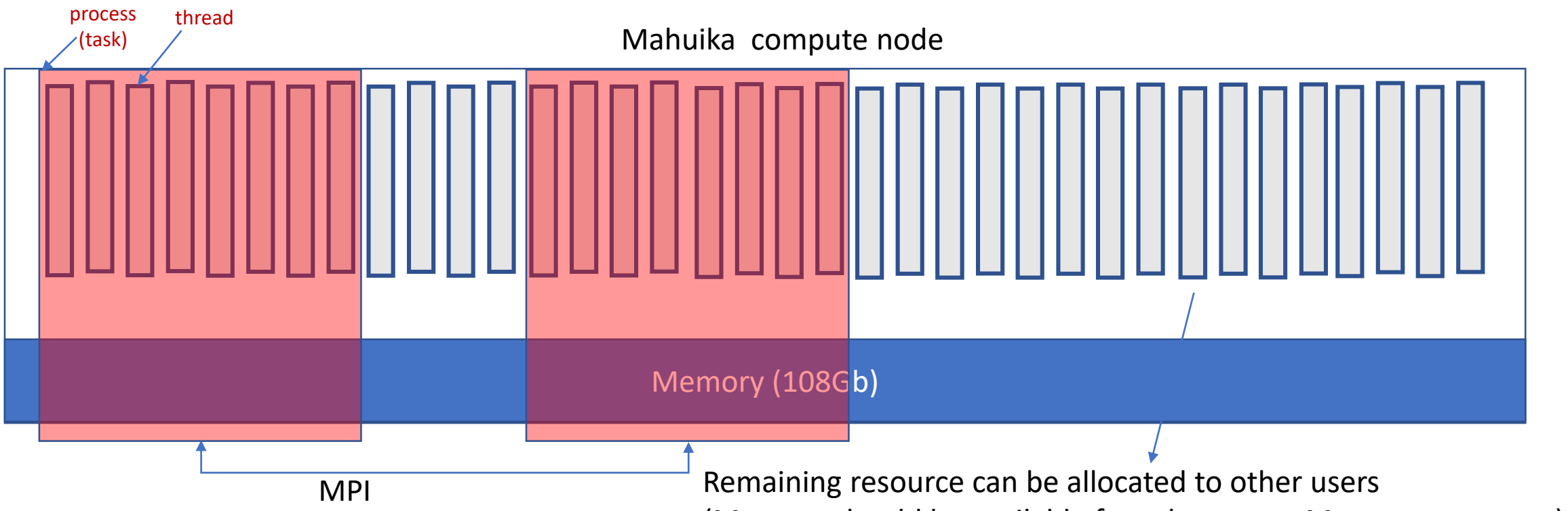
SLURM

```
#!/bin/bash
#SBATCH --job-name=JobName # job name (shows up in the queue)
#SBATCH --account=nesi99999 # Project Account
#SBATCH --time=03:00:00 # Walltime (HH:MM:SS)
#SBATCH --mem-per-cpu=1500 # memory/cpu (in MB)
#SBATCH --ntasks=2 # number of tasks (e.g. MPI)
#SBATCH --cpus-per-task=4 # number of cores per task (e.g. OpenMP)
#SBATCH --partition=large # specify a partition
#SBATCH --hint=nomultithread # don't use hyperthreading srun [options]
<executable> [options]
```



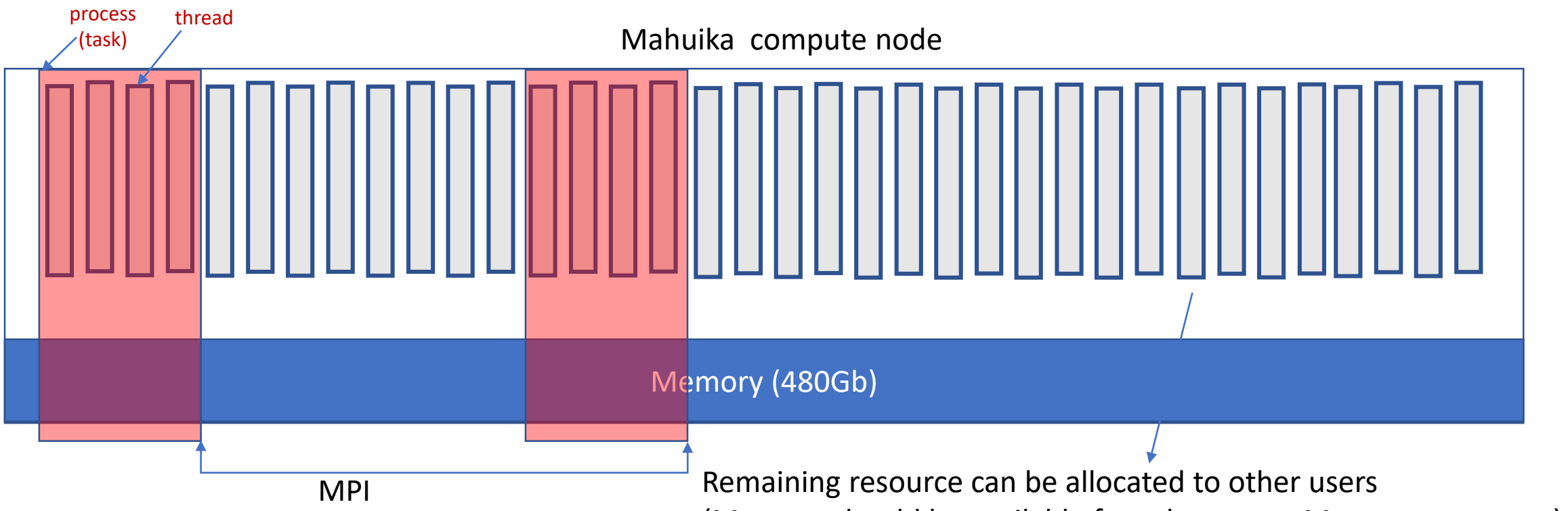
SLURM

```
#!/bin/bash
#SBATCH --job-name=JobName # job name (shows up in the queue)
#SBATCH --account=nesi99999 # Project Account
#SBATCH --time=03:00:00 # Walltime (HH:MM:SS)
#SBATCH --mem-per-cpu=3000 # memory/cpu (in MB)
#SBATCH --ntasks=2 # number of tasks (e.g. MPI)
#SBATCH --cpus-per-task=4 # number of cores per task (e.g. OpenMP)
#SBATCH --partition=large # specify a partition
#SBATCH --hint=nomultithread # don't use hyperthreading srun [options]
<executable> [options]
```



SLURM

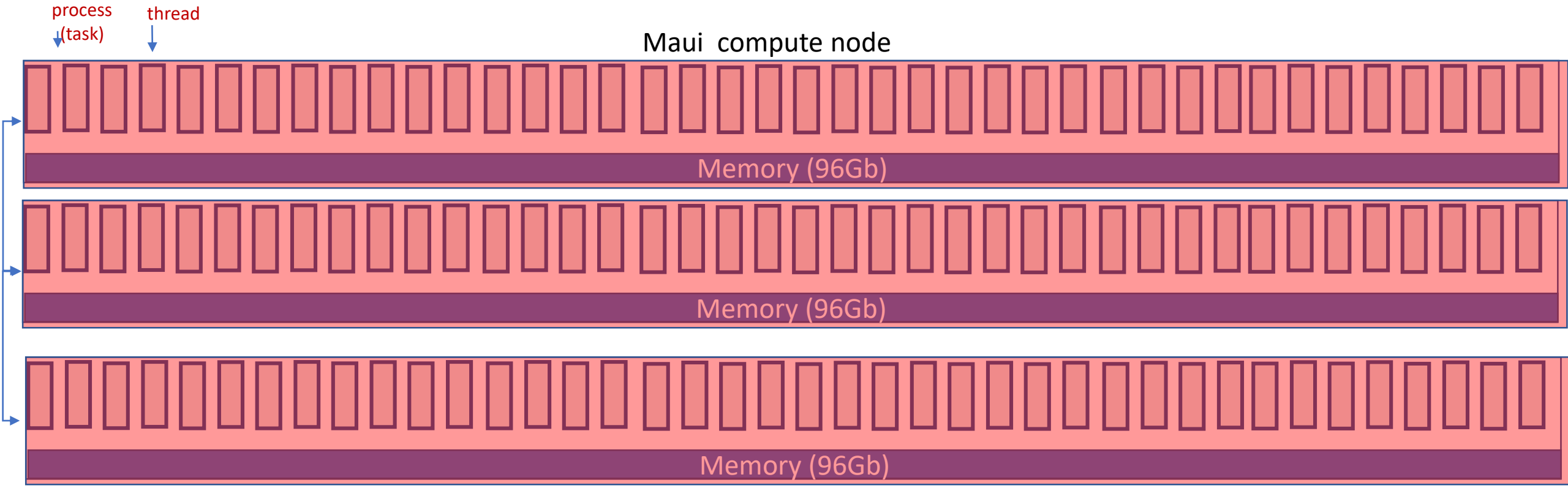
```
#!/bin/bash
#SBATCH --job-name=JobName # job name (shows up in the queue)
#SBATCH --account=nesi99999 # Project Account
#SBATCH --time=03:00:00 # Walltime (HH:MM:SS)
#SBATCH --mem-per-cpu=3000 # memory/cpu (in MB)
#SBATCH --ntasks=2 # number of tasks (e.g. MPI)
#SBATCH --cpus-per-task=4 # number of cores per task (e.g. OpenMP)
#SBATCH -partition=prepost # specify a partition
#SBATCH --hint=nomultithread # don't use hyperthreading srun [options]
<executable> [options]
```



SLURM

```
#!/bin/bash
#SBATCH --job-name=JobName # job name (shows up in the queue)
#SBATCH --account=nesi99999 # Project Account
#SBATCH --time=03:00:00 # Walltime (HH:MM:SS)
#SBATCH --mem-per-cpu=1500 # memory/cpu (in MB)
#SBATCH --ntasks=3 # number of tasks (e.g. MPI)
#SBATCH --cpus-per-task=40 # number of cores per task (e.g. OpenMP)
#SBATCH --partition=nesi_research # specify a partition
#SBATCH --hint=nomultithread # don't use hyperthreading srun [options]
<executable> [options]
```

Processes can't share memory (must use MPI)



Available SW modules