RC Account Registration

Processing Data at Scale

Research Computing
UNIVERSITY OF COLORADO BOULDER

Be Boulder.

View the Slides

https://github.com/ResearchComputing/Processing_Data_At_Scale

Meet the User Support Team

Layla Freeborn

Brandon Reyes

Andy Monaghan

Michael Schneider

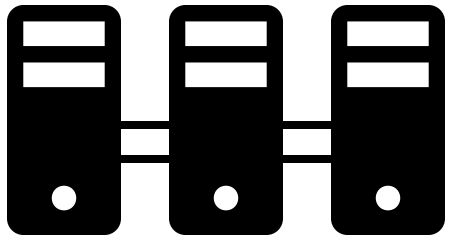
John Reiland

Dylan Gottlieb

Mohal Khandelwal

Ragan Lee

Workshop Overview



HPC Basics

Slurm Jobs

Dask Library

Ask Questions

Discuss Ideas

Practice Project – Counting Words

- Project Gutenberg –
  - Free e-book repository
  - Started by Michael Hart (creator of first e-book)
  - Structured, but poorly formatted

Accessing Gutenberg Files

[Gutenberg Project]

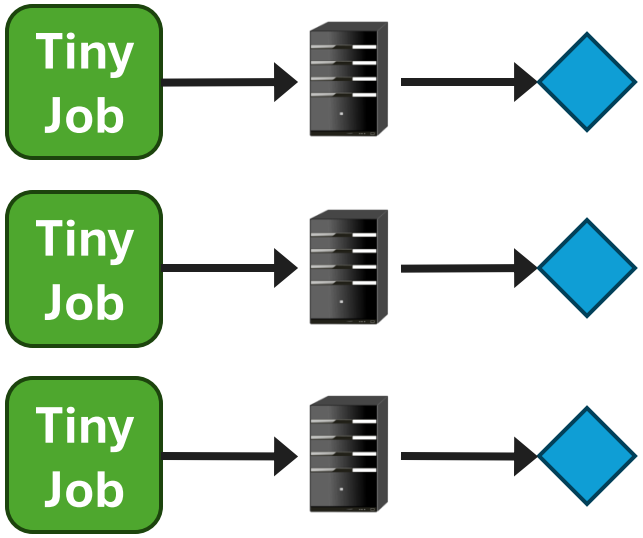
[Web Scraping Tips]

[Downloading Files]

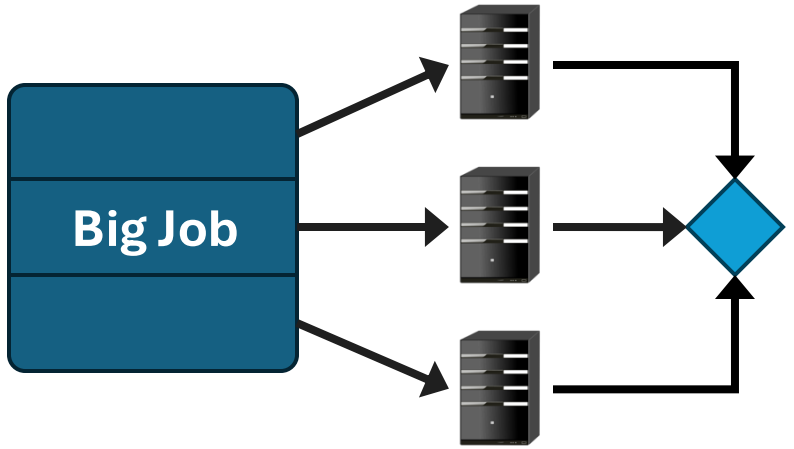
What is HPC?

Scale **vs** **Speed**

What can I use HPC for?

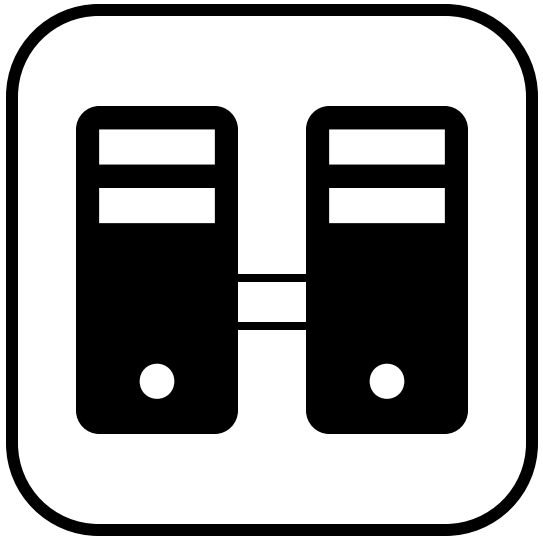


Serialized Jobs

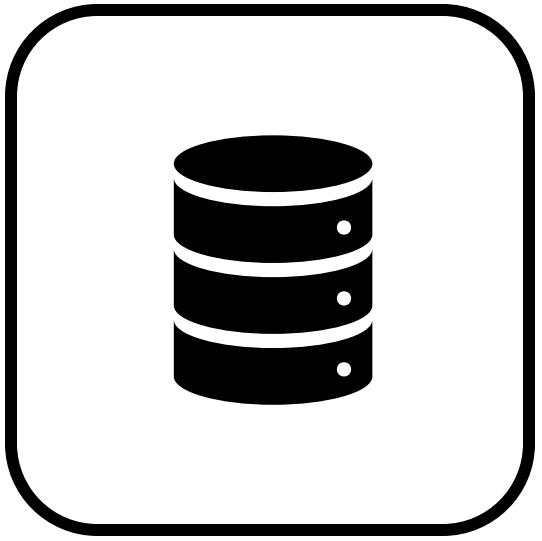


Parallelized Job

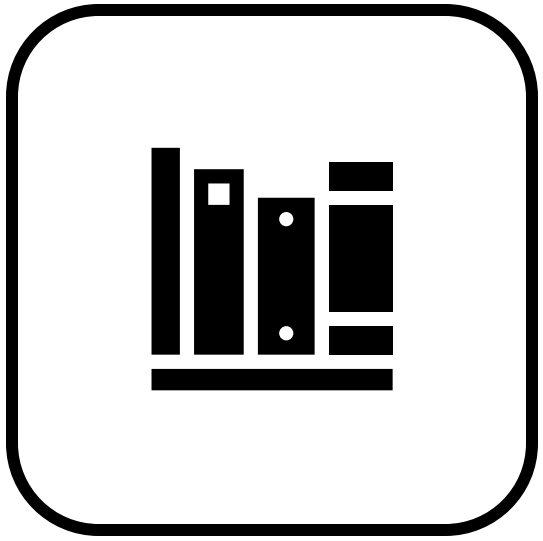
CURC HPC Resources



Alpine Cluster



Data Storage



Software Modules

CURC HPC Resources

Alpine Cluster

Data Storage

Software Modules

Alpine Partitions

amilan

General Usage

Alpine Partitions

amilan

General Usage

amem

High Memory

Alpine Partitions

amilan

General Usage

amem

High Memory

aa100

Nvidia GPU's

Alpine Partitions

amilan
General Usage

amem
High Memory

aa100
Nvidia GPU's

ami100
AMD GPU's

CURC Web Portal

Data Storage

Core

- Personal Storage
- Includes 3 Directories
  - /home (2 GB)
  - /projects (250 GB)
  - /scratch (10 TB)

PL

- PetaLibrary
- Tiered Storage
  - Active, Archive
- Requires Funding
- Starts at 1 TB

Copy Files

- CP – Copy command

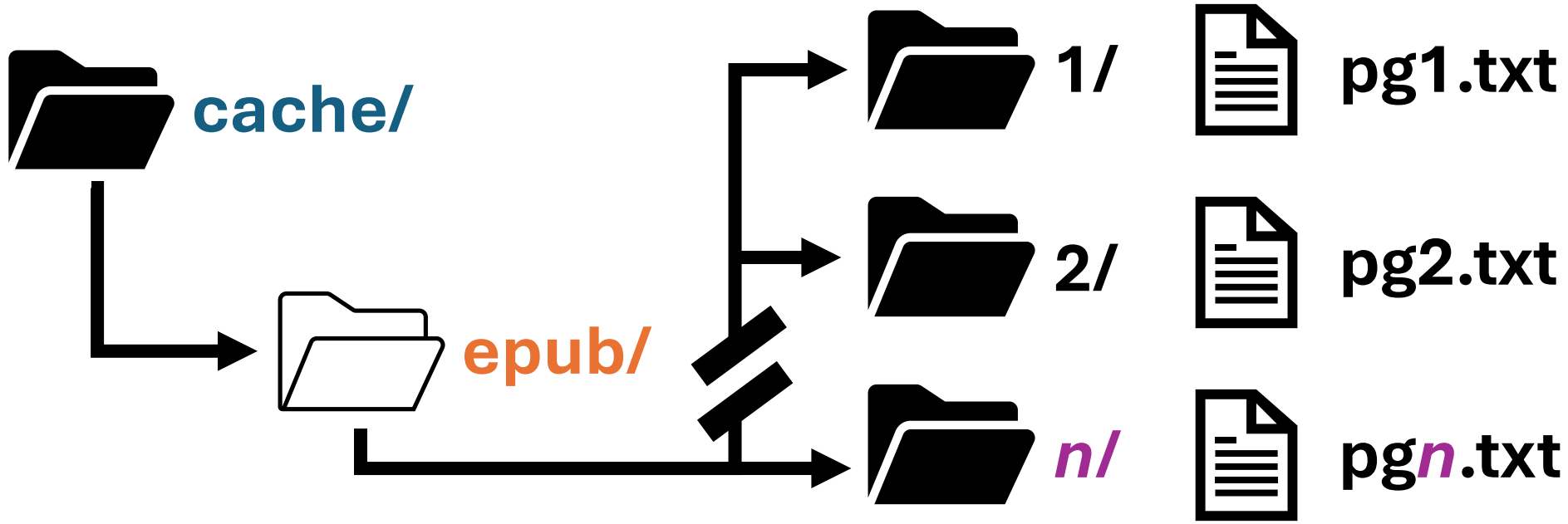
```
$ cp /pl/active/courses/2025_spring/CMCI_LL/txt-files.tar
     /scratch/alpine/$USER/txt-files.tar

$ cp /pl/active/courses/2025_spring/CMCI_LL/code.tar
     /projects/$USER/code.tar
```

Extract Files

- tar – extract or “unzip” files command

```
$ cd /scratch/alpine/$USER
$ tar -xvf txt-files.tar

$ cd /projects/$USER
$ tar -xvf code.tar
```

Dataset Structure



/scratch/alpine/$USER/cache/epub/*n*/pg*n*.txt

Manually counting words

wc -w <file name>

Anatomy of a job script

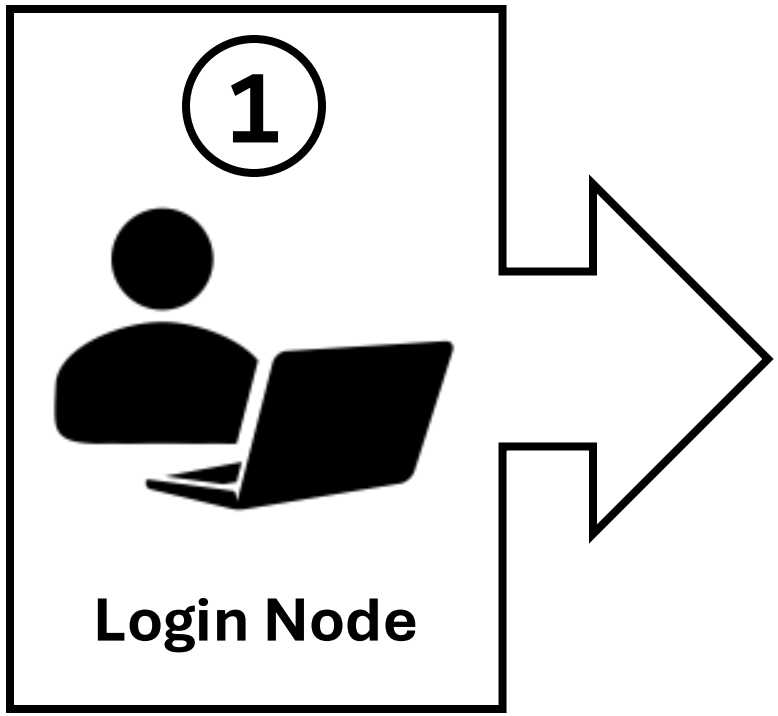
```
#!/bin/bash

## Directives
#SBATCH --<option>=<value>

## Software
module load <software>

## User scripting
<command>
```

Batch Jobs

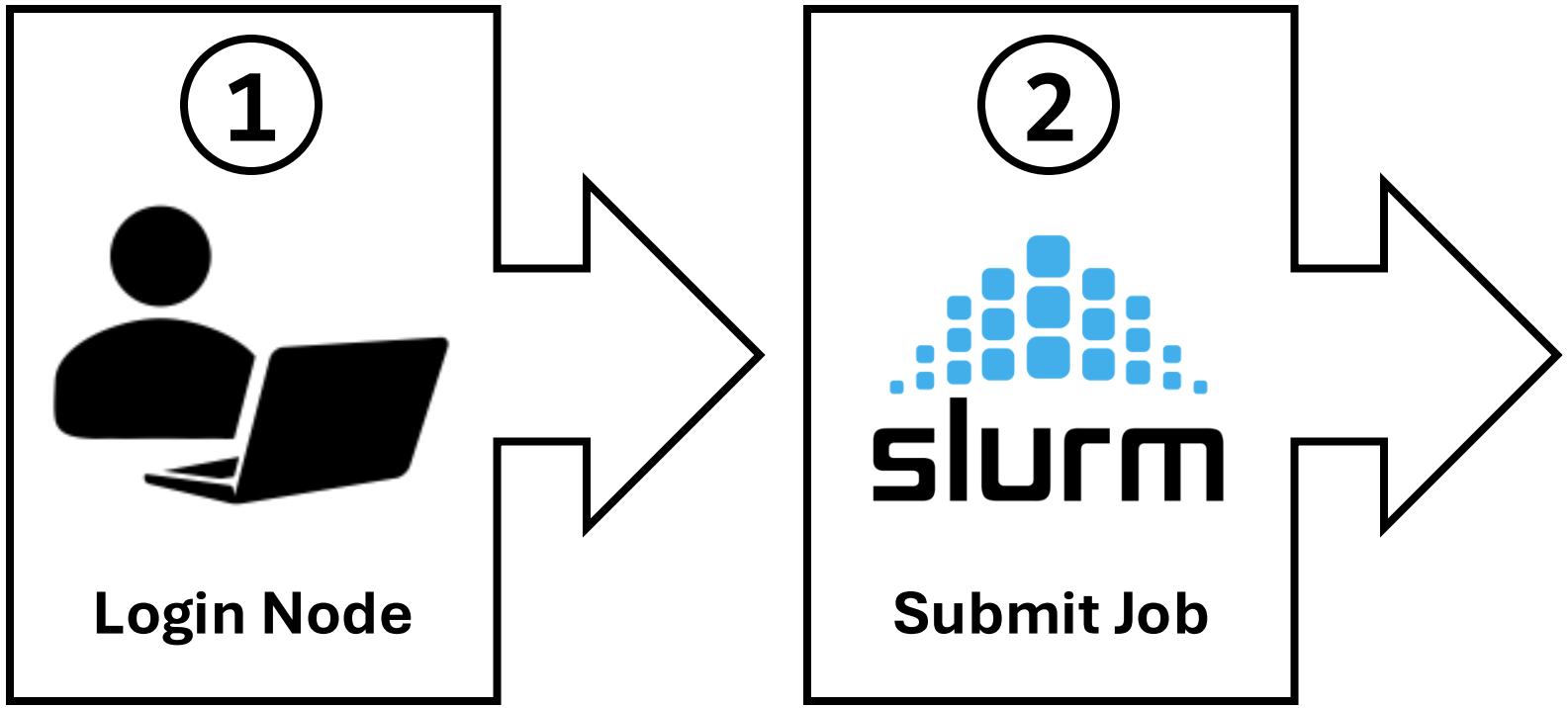
job.sh

```
#!/bin/bash

#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --time=00:20:00
#SBATCH --partition=amilan
#SBATCH --output=slurm_logs/serial-%j.out

JOB=$SLURM_JOB_ID
TASK=0
START=0
END=10

./count_words.sh "$JOB" "$TASK" "$START" "$END"
```

Submitting a Batch Job

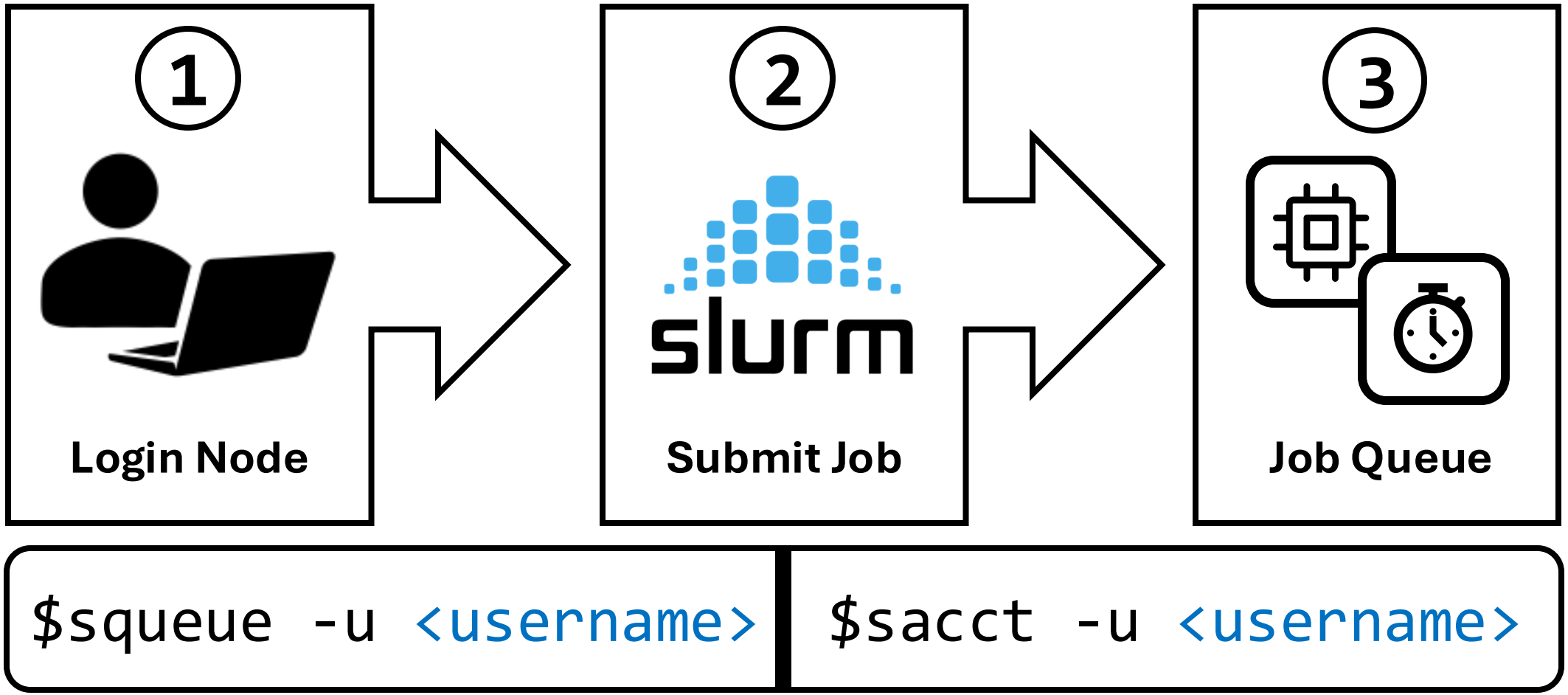


Submitting a Batch Job

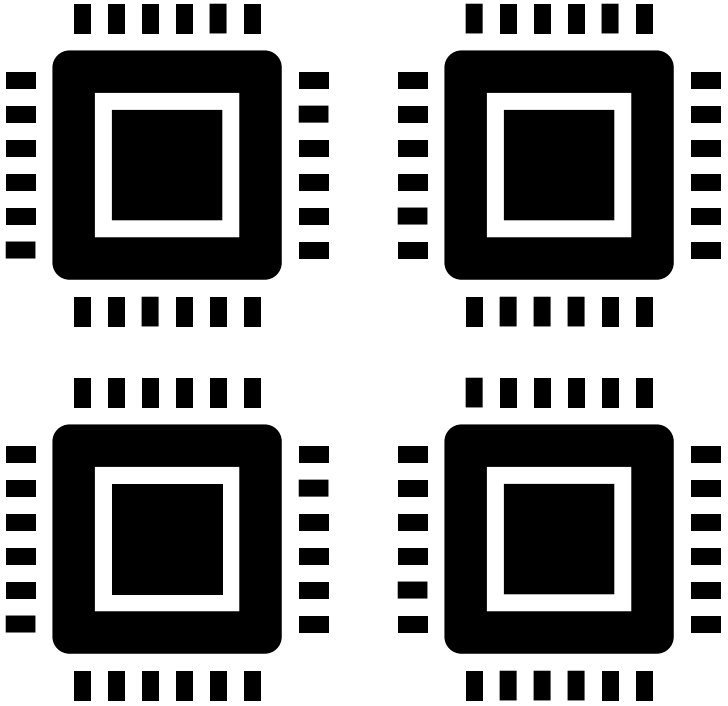



```
$sbatch <job_file> <other-directives>
```

Submitting a Batch Job



Cores != Performance



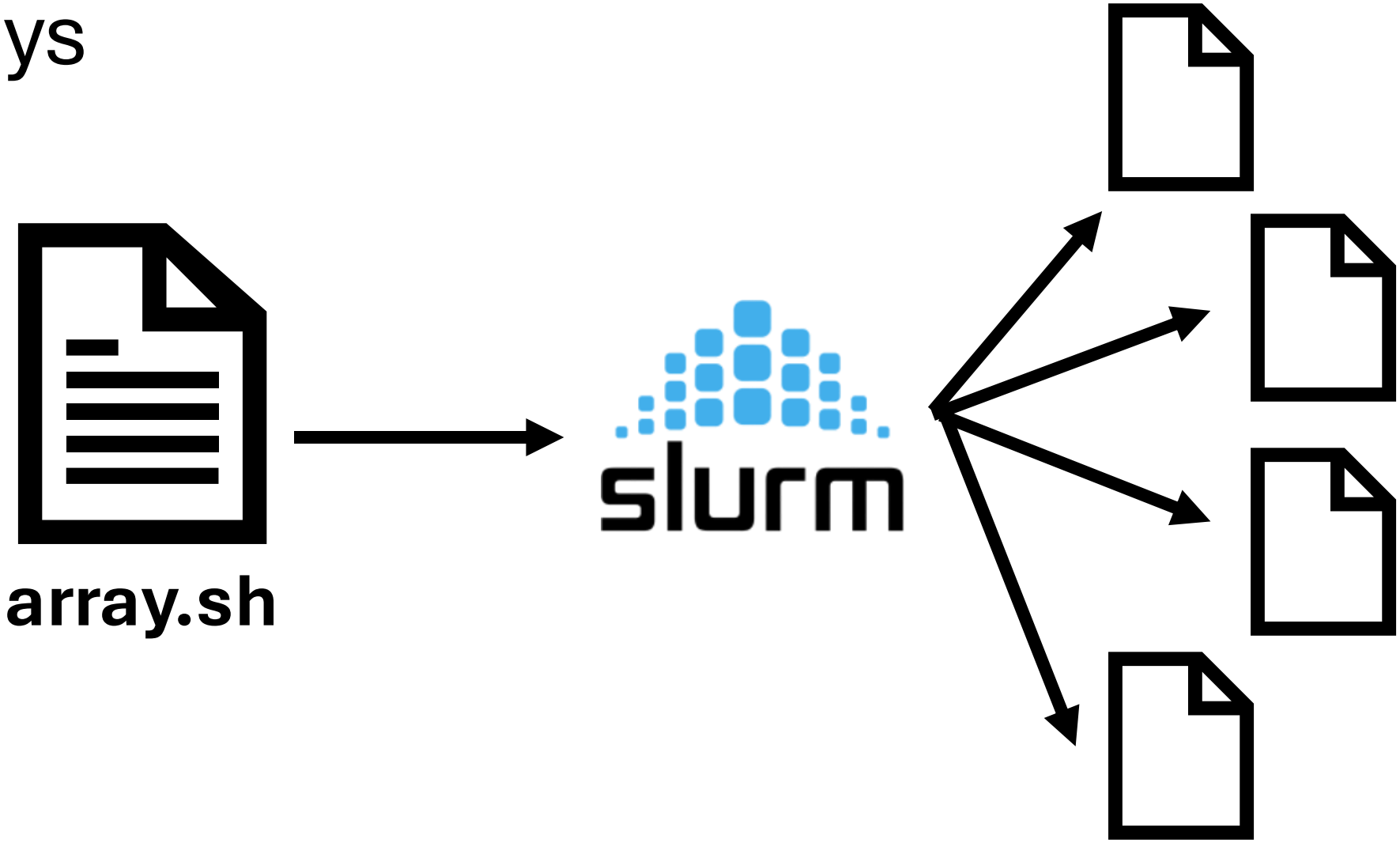
Checking Job Performance

```
$ module load slurmtools
$ seff <job number>
```

```
Job ID: 8636572
Cluster: alpine
User/Group: ralphie/ralphiegrp
State: COMPLETED (exit code 0)
Nodes: 1
Cores per node: 24
CPU Utilized: 04:04:05
CPU Efficiency: 92.18% of 04:24:48 core-walltime
Job Wall-clock time: 00:11:02
Memory Utilized: 163.49 MB
Memory Efficiency: 0.14% of 113.62 GB
```

Monitoring Resources

Job Arrays



Scaling with Dask

Creating an Anaconda Environment

```
$ module load anaconda
$ conda create -n dask
$ conda activate dask
$ conda install dask -c conda-forge
$ conda install -c conda-forge jupyterlab
```

Jupyter Session

Kernel:

```
$ conda install -y ipykernel
$ python -m ipykernel install --user --name dask --display-name dask
```

Execute Jupyter Notebook

```
jupyter execute <notebook.ipynb>
```

Where to go next?

- Discuss python libraries:
  - Multiprocessing
  - Cuda and optimized ml libraries for mpi
- MPI enabled libraries and compiling c++ code
- R libraries – futures
- CRDDs office hours and other workshops.

CRDDS Events & Office Hours

Documentation

https://curc.readthedocs.io/en/latest/

Survey and feedback

Survey: http://tinyurl.com/curc-survey18