Agenda

- 8:30am Welcome
- 8:35am Shaheen III Hardware Overview
- 8:55am How to apply on Shaheen III
- 9:05am Getting Started on Shaheen III
- 9:15am Software Environment
- 9:35am Job Scheduling
- 10:00am Coffee Break
- 10:15am Storage overview & Best practices
- 10:30am Applications software example: VASP workflow
- 10:50 am Applications software example: CFD applications
- **11:10 am Applications software example: Bioinformatics workflow**
- 11:20-11.30am Q&A and Open Discussion

Shaheen III HPC Training
Applications software example: Bioinformatics workflow

February 4, 2025
Location: Auditorium between bldgs. 2 & 3

Outline

- BioApps: Software Ecosystem
- Reference data directory
  - E.g.: Application reference
- Computing Resources
  - E.g.: Shared and large memory partitions
- Bioinformatics Workflows
  - Power of pipe operator
  - Example job script

BioApps: Software Ecosystem

BioApps ecosystem

Two options to install the BioApps:

Option #1: Install your own software at $PROJECT directory.

(Like Shaheen II, we will be continue providing the support for installing the software)

Example use cases:

(a) Early development versions
(b) Frequent changes in the software development
(c) Uncommon software for multiple PI's

Option #2: Install the software as a "module"

(We will install most commonly used software as a module)

Example use cases:

(a) Stable version
(b) Strong developer support on the application
(c) Commonly used software by multiple PI's

Contact us for support at: help@hpc.kaust.edu.sa and we will suggest the best optimal solution!

BioApps Software modules

Important: **DON'T purge the modules at Shaheen III (like Ibex)**

#1. Ensure the default modules (as always) are available at Shaheen III

```
$ module list
1) craype-x86-genoa      5) xpmem/2.8.4-1.0_7.2__ga37cbd9.shasta
2) libfabric/1.20.1      6) cce/18.0.0         9) cray-mpich/8.1.30
3) craype-network-ofi    7) craype/2.7.32          10) cray-libsci/24.07.0
4) perftools-base/24.07.0     8) cray-dsmml/0.3.0       11) PrgEnv-cray/8.5.0
```

Required modules for Shaheen III

#2. Check the list of BioApps (commonly used) available at Shaheen III

```
$ export MODULEPATH=/scratch/project/software/ex111genoa/modulefiles:$MODULEPATH
$ module avail -S bio
---------------- /sw/ex111genoa/modulefiles -------------
bio/blast+/2.15.0
    bio/bwa/0.7.17
    bio/gatk/4.1.6
    bio/java/8u401
    bio/java/21.0.2
    bio/samtools/1.18
```

Bioinformatics database and Application reference

Application reference

- Most commonly used application reference (E.g.: At-least 2 PI's common reference data) may include at:
  `/scratch/reference/bio`
- The structure of reference directory will be:
  `<Application name>/<reference version>`

Example:

`/scratch/reference/bio/ncbi/v5/nt`

Computing Resources (E.g.: Shared and large memory partitions)

Compute nodes

- **Exclusive** nodes
- **Shared** nodes (e.g. like Ibex)
- **ppn** (Pre and Post-processing) nodes
  Large memory nodes (3TB RAM, part of PPN)
- **dtn** (Data transfer Nodes)

Wall-clock time

- **24 Hours**, as a default partition
- **72 Hours**, extended wall-clock time and this partition will be granted based on justifications

Bioinformatics: Example Workflows

Power of pipe (|) operator

BWA mem: Genome alignment
samtools view : SAM -> BAM file conversion
samtools sort: Sort the BAM file alignment

```
bwa mem -M -k 30 -t $CORES
$REFERENCE/human_g1k_v37_decoy.fasta
$INPUT/$PREFIX.trimmed.P1.fastq.gz
$INPUT/$PREFIX.trimmed.P2.fastq.gz | samtools view -@
$CORES -b -S -h -q 30 - | samtools sort - >
$OUTPUT/$PREFIX.sorted.bam
```

Example job script

```
#!/bin/bash
#SBATCH -N 1
#SBATCH -p shared
#SBATCH -c 4
#SBATCH -J bwa_samtools_job
#SBATCH --time=4:00:00
#SBATCH -o bwa_samtools.out
#SBATCH -e bwa_samtools.err

export PREFIX=mom;
export WORK=/scratch/kathirn/BWA ;
export INPUT=$WORK/data ;
export REFERENCE=$WORK/ref ;
export OUTPUT=$PWD/output;
export MODULEPATH=/scratch/project/software/ex111genoa/modulefiles:$MODULEPATH
module load bio/bwa/0.7.17  bio/samtools/1.8
export CORES=4;

bwa mem -M -k 30 -t $CORES $REFERENCE/human_g1k_v37_decoy.fasta
$INPUT/$PREFIX.trimmed.P1.fastq.gz $INPUT/$PREFIX.trimmed.P2.fastq.gz | samtools view -@ $CORES -b
-S -h -q 30 - | samtools sort - > $OUTPUT/$PREFIX.$CORES.sorted.bam
```

Contact us

https://www.hpc.kaust.edu.sa

https://docs.hpc.kaust.edu.sa

help@hpc.kaust.edu.sa