Sentieonの実行(1)
sentieon_quickstart.shの実行

2023/06/21

横浜市大
三澤計治

この資料の目的

目的

- 遺伝研スパコンでSentieonを実行する。

内容

- 遺伝研スパコンの特性を理解する。
- sentieon_quickstart.shを改変する。

学習目標

- sentieon_quickstart.sh実行に成功する。

sentieon_quickstart.sh実行に必要なこと

以下のことを指定する必要がある。

a. 使用するメモリの指定
b. CPUの数の指定
c. Sentieonをインストールした場所
d. ライセンスの場所
e. 入力・出力データの場所

これらのことを指定するためには、普通のコンピュータと遺伝研スパコンの違いを理解する必要がある。

普通のコンピュータと遺伝研スパコンの違い

普通のコンピュータ

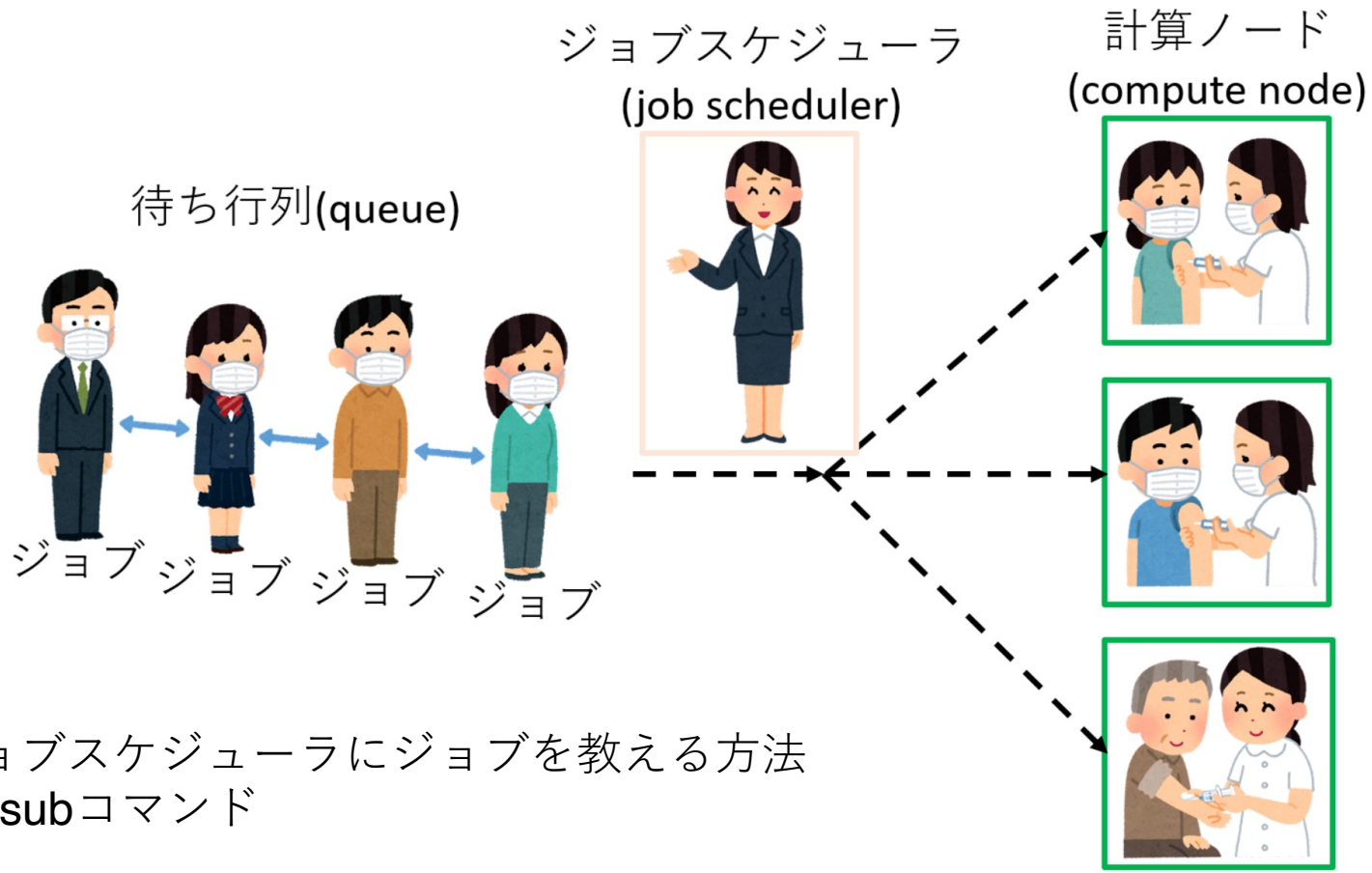
- 利用しているコンピュータで実行

遺伝研スパコン

- 計算ノードとログインノードは別
- ログインノード
  - ログインした時に見ているコンピュータ

ジョブスケジューラと並列処理

- 並列処理を行うために、ジョブスケジューラが、待ち行列にあるジョブを計算ノードに割り振って実行する



ジョブスケジューラにジョブを教える方法がqsubコマンド

ジョブスケジューラUGE

- SGE: Sun Girid Engine (2000-)
  - Sun が開発し2001 年にオープンソース化
- OGE: Oracle Girid Engine (2010-)
  - Oracle がSun を買収し有償化
- UGE: Univa Grid Engine (2013-)
  - 分離
- AGE: Altair Grid Engine (2021-)
  - 2020 年にAltair 社がUniva 社を買収

- UGEではqsubでジョブスケジューラにジョブを伝える

qsub実行時のオプションの指定方法(1)

- コマンドqsubを実行する

  ```
  qsub -cwd work.sh
  ```

- 解説
  - qsub 今使っているコンピュータに対し、「これから伝えるコマンドをジョブ管理システムの待ち行列に加えてくれ」と命じるコマンド
  - -cwd ジョブ管理システムに対し、「今使っているディレクトリ(current working directory) で実行するよう実行ノードに伝えてくれ」と命じるオプション
  - work.sh 実行ノードが実行するスクリプト

qsub実行時のオプションの指定方法(1)

- 実行するスクリプト内にオプションを書くことも可能。
- #$は、実行ノードではなく、ジョブスケジューラへのオプション

注意点

- 最初の文字が$で始まる行をコメントアウトしようとして#を挿入すると、#$となってしまい、ジョブスケジューラが読もうとして失敗する。
- 以下のようなエラーとなる。
- `Unable to read script file because of error: invalid option argument`
- このエラーメッセージからは上述の原因がわかりにくい
- 対策
  - コメントアウトの時は#と$の間に1文字スペースを入れる

変更点の解説

- sentieon_quickstart.tar.gzを展開
  - gunzip sentieon_quickstart.tar.gz
  - tar xvf sentieon_quickstart.tar
- sentieon_quickstartフォルダに移動
- sentieon_quickstart.shをバックアップ
- sentieon_quickstart.shを編集
  - a. 使用するメモリの指定、b. CPUの数の指定、c. Sentieonをインストールした場所、d. ライセンスの場所、e. 入力・出力データの場所の5つ

変更点(1) a.使用するメモリのサイズ

sentieon_quickstart.shの最初に以下の青文字部分を入れる

```
#$ -S /bin/bash
#$ -cwd
#$ -V
#$ -l medium
#$ -l s_vmem=35G
#$ -l mem_req=35G
```

- シェルスクリプトの指定
- ディレクトリの指定
- 環境を引きつく
- mediumというqueで実行
- メモリサイズの指定(1)
- メモリサイズの指定(2)

解説

- コマンドラインインターフェース
  - ユーザーがテキストベースのコマンドを入力し、それに対してコンピュータが応答するタイプのインターフェース
- シェルスクリプト
  - コンピュータのコマンドラインインターフェースで実行されるスクリプト言語
  - 遺伝研のログインノードのシェルスクリプトはbashだが、計算ノードはcshになっている
  - sentieon_quickstart.shを実行するため、計算ノードのシェルスクリプトをbashに変更する必要がある。

解説（続き）

- 大きなメモリが必要なため、mediumという名前のqueを利用
  - CPU
    - Intel Xeon Gold 6148
    - 80 CPU cores/node
  - メモリ
    - 38.4GB memory/CPU core

変更点(2) b. CPUの指定

- 並列処理を128 coreで実行

Other settings

nt=128 #number of threads to use in computation

変更点(3) c. ライセンスの場所

ライセンスファイルの場所を書き込む

```
export SENTIEON_LICENSE=/home/kazumisawa/data/license/Yokohama_City_University_eval.lic
```

変更点(4) d. Sentieonをインストールした場所

sentieon-genomics-202112.07.tar.gzを展開してできるsentieon-genomics-202112.07フォルダの場所を指定する。

```
# Update with the location of the Sentieon software package
SENTIEON_INSTALL_DIR=/home/kazumisawa/bin/sentieon-genomics-202112.07
```

変更点(5) e.データのあるディレクトリの指定data_dir

data_dirを変更する
sentieon_quickstart.tar.gzを展開してできたフォルダのpath

```
# Update with the fullpath location of your sample fastq
set -x
#data_dir="$( cd -P "$( dirname "$0" )" && pwd )"
data_dir=/home/kazumisawa/bin/sentieon_quickstart
```

変更点(6) オプションにならないように変更

- #$は、実行ノードではなく、ジョブスケジューラへのオプション
- 対策
  - コメントアウトの時は#と$の間に1文字スペースを入れる

実行

- 編集したsentieon_quickstart.shをqsubで実行

  qsub sentieon_quickstart.sh
- 実行されているかどうか確認

  qstat
- 終了後確認

  cd result

  ls -lh
  - 240Kのoutput-hg.vcf.gzができていたら成功