Queueing System安裝與設定

國立臺灣師範大學物理學系 陳俊明

chunming@ntnu.edu.tw

排程系統(Queueing System) 軟體

- Condor
- Sun Grid Engine (SGE)
- LSF(Load Sharing Facility), IBM
- PBS(Portable Batch System)為目前功能最為齊全的排程系統之一
  - PBS Pro - PBS的商業版本，功能最為豐富
  - openPBS (Open Portable Batch System) -最早的PBS系統
  - Torque –目前已商業化，目前尚有開源版本以專案形式放置於github上
- Slurm

PBS – 組成元件

- commands
  以命令列方式，讓透過socket通訊協定來讓使用者進行提交(submit)、監督 (monitor)、修改 (modify)和刪除(delete) 工作
- pbs_server
  接收、產生、管理及保護使用者的批次工作
- pbs_mom
  接收pbs_server 給予的批次工作，並呼叫對應的程式來執行，完成後將結果回報給pbs_server
- pbs_sched
  負責排程工作、資源分配及節點管理

安裝 OpenPBS

- 填寫資料下載 OpenPBS RPM https://www.openpbs.org/
- 下載原始碼自行編譯 https://github.com/openpbs/openpbs
- PBS 文件：https://www.altair.com/pbs-works-documentation/
- 伺服器：openpbs-server-23.06.06-0.x86_64.rpm
- 運算節點：openpbs-execution-23.06.06-0.x86_64.rpm
- 用戶端：openpbs-client-23.06.06-0.x86_64.rpm

OpenPBS 下載原始碼編譯

- 編譯環境準備套件安裝，讀 INSTALL 檔案

```
[root@master ~]# yum install -y gcc make rpm-build libtool hwloc-devel libX11-devel libXt-devel libedit-devel libical-devel ncurses-devel perl postgresql-devel postgresql-contrib python3-devel tcl-devel tk-devel swig expat-devel openssl-devel libXext libXft autoconf automake libtool-ltdl-devel git
```

- 使用 git 指令下載原始碼或直接下載壓縮檔

```
[root@master ~]# git clone https://github.com/openpbs/openpbs.git
```

```
[root@master ~]# wget https://github.com/openpbs/openpbs/archive/refs/tags/v23.06.06.tar.gz
```

- 如果要編譯某特定版本，就用 git 指令指定，建議使用 v19

```
[root@master ~]# git clone –b release_19_1_branch https://github.com/openpbs/openpbs.git
```

OpenPBS 下載原始碼編譯

- 修改安裝路徑openpbs.spec openpbs.spec.in

```
[root@master ~]# tar zxvf v23.06.06.tar.gz
[root@master ~]# vi ~/openpbs-23.06.06/openpbs.spec
%define pbs_prefix /usr/local/pbs
```

- 編譯環境準備，建目錄及壓縮複製檔案

```
[root@master ~]# mkdir -p ~/rpmbuild/{SPECS,SOURCES}
[root@master ~]# cp ~/openpbs-23.06.06/openpbs.spec ~/rpmbuild/SPECS/
[root@master ~]# mv v23.06.06.tar.gz openpbs-23.06.06.tar.gz
[root@master ~]# mv openpbs-23.06.06.tar.gz ~/rpmbuild/SOURCES/
[root@master ~]# cd ~/rpmbuild/SPECS/
```

- 開始編譯

```
[root@master ~]# rpmbuild -bb openpbs.spec
```

PBS 使用通訊埠 (Ports)

- pbs_server : 15001
- pbs_mom : 15002
- pbs_resmom : 15003
- pbs_sched : 15004
- pbs_datastore : 15007
- altair license server : 6200

PBS 伺服器安裝

- 環境準備套件安裝，讀 INSTALL 檔案

```
[root@master ~]# yum install expat libedit postgresql-server postgresql-contrib python3 sendmail sudo tcl tk libical perl-Env perl-Switch
```

- 建立一個管理帳戶（非必要）

```
[root@master ~]# useradd -m pbsadmin
[root@master ~]# passwd pbsadmin
[root@master ~]# make -C /var/yp
```

- 安裝伺服器套件

```
[root@master ~]# rpm -ivh openpbs-server-23.06.06-0.x86_64.rpm openpbs-devel-23.06.06-0.x86_64.rpm
```

或

```
[root@master ~]# yum install openpbs-server-23.06.06-0.x86_64.rpm openpbs-devel-23.06.06-0.x86_64.rpm
```

PBS 伺服器安裝

- 設定 pbs.conf

```
[root@master ~]# vi /etc/pbs.conf
PBS_EXEC=/usr/local/pbs
PBS_SERVER=master
PBS_START_SERVER=1
PBS_START_SCHED=1
PBS_START_COMM=1
PBS_START_MOM=0
PBS_HOME=/var/spool/pbs
PBS_CORE_LIMIT=unlimited
PBS_SCP=/bin/scp
PBS_RCP=/bin/false
PBS_RSHCOMMAND=/usr/bin/ssh
```

PBS 伺服器安裝

- 啟動 PBS 服務

```
[root@master ~]# systemctl start pbs
```

- 設定每次開機都啟動 PBS

```
[root@master ~]# systemctl enable pbs
```

- 重新開機

```
[root@master ~]# init 6
```

PBS 伺服器設定

- 檢查目前伺服器設定

```
[root@master ~]# qmgr -c "list server"
```

- 設定可以查歷史工作及效期

```
[root@master ~]# qmgr -c "set server job_history_enable = True"
[root@master ~]# qmgr -c "set server job_history_duration = 1440:00:00"
```

- 改變郵件寄送來源，還要額外設定 sendmail

```
[root@master ~]# qmgr -c "set server mail_from = root@master.cluster"
```

- 設定 pbsadmin 為管理者

```
[root@master ~]# qmgr -c  "set server managers+=pbsadmin@*"
```

PBS 伺服器設定

- 允許使用者從運算節點送 jobs

```
[root@master ~]# qmgr -c "set server flatuid=true"
```

- 設定紀錄的 logs levels

```
[root@master ~]# qmgr -c "set server log_events=2047"
## 0 : 不紀錄
## 511 : 預設
## 2047 : 最多資訊
```

- 新增運算節點

```
[root@master ~]# qmgr -c "create node cn1"
```

PBS 運算節點安裝

- 環境準備套件安裝

```
[root@cn1 ~]# yum config-manager --set-enable powertools
[root@cn1 ~]# yum install epel-release
[root@cn1 ~]# yum update
[root@cn1 ~]# yum install hwloc-libs libX11-common python3 python3-setuptools tcl tk libSM libICE perl-Env perl-Switch
```

- 安裝運算節點套件

```
[root@cn1 ~]# rpm -ivh openpbs-execution-23.06.06-0.x86_64.rpm
or
[root@cn1 ~]# yum install openpbs-execution-23.06.06-0.x86_64.rpm
```

PBS 運算節點設定

- 設定 pbs.conf

```
[root@cn1 ~]# vi /etc/pbs.conf
PBS_EXEC=/usr/local/pbs
PBS_SERVER=master
PBS_START_SERVER=0
PBS_START_SCHED=0
PBS_START_COMM=0
PBS_START_MOM=1
PBS_HOME=/var/spool/pbs
PBS_CORE_LIMIT=unlimited
PBS_SCP=/bin/scp
PBS_RCP=/bin/false
PBS_RSHCOMMAND=/usr/bin/ssh
```

- 重新啟動並設定每次開機都啟動 PBS

```
[root@cn1 ~]# systemctl start pbs && systemctl enable pbs
```

PBS 運算節點設定

- openmpi 必須重新編譯並加入 **--with-tm=/usr/local/pbs** 選項

```
[root@master ~]# ./configure --prefix=/opt/openmpi/4.1.6_gcc_8.5.0 --enable-mpi-cxx CC=gcc CXX=g++
FC=gfortran F77=gfortran --with-tm=/usr/local/pbs
or
[root@master ~]# source /opt/intel/oneapi/setvars.sh intel64
[root@master ~]# ./configure --prefix=/opt/openmpi/4.1.6_intel_2024.2.0 ---enable-mpi-cxx CC=icx
CXX=icpx FC=ifx F77=ifx  --with-tm=/usr/local/pbs
```

設定 Queues

指令	說明
qmgr -c "del queue batch"	刪除Queues
qmgr -c "create queue workq queue_type=execution“ qmgr -c "set queue workq enabled = True“ qmgr -c "set queue workq started = True"	建立Queues
qmgr -c "set queue workq resources_default.walltime=1:00:00"	運算最長時間
qmgr -c "set queue workq max_running=10"	設定Queues最多同時作業數量
qmgr -c "set queue workq max_user_queuable=20"	設定使用者最多提交作業數量
qmgr -c "set queue workq max_user_run=5"	設定使用者最多同時作業數量
qmgr -c "set queue workq resources_default.nodes=1"	設定Queues預設的nodes數量
qmgr -c "set queue workq resources_max.nodes=4"	設定Queues最大使用的nodes數量

設定 Queues

指令	說明
qmgr -c "set queue workq acl_hosts=cn1+cn2+…“ qmgr -c "set queue workq acl_host_enable=true"	設定Queues使用的nodes群組
qmgr -c "set queue workq acl_users = user1" qmgr -c "set queue workq acl_users += user2" qmgr -c "set queue workq enabled=true"	指定Queues可以使用的帳號
qmgr -c "s q queue priority = 100"	設定Queues的優先等級

設定 Queues 範例

- 建立Queue

```
qmgr -c "create queue workq"
qmgr -c "set queue workq queue_type = Execution"
qmgr -c "set queue workq started = True"
qmgr -c "set queue workq resources_default.nodes = 1"
qmgr -c "set queue workq resources_default.walltime = 01:00:00"
qmgr -c "set queue workq enabled = True"
qmgr -c "set queue workq started = True"
```

Submit Job 範例

- 一般的提交

```
[user1@master ~]$ qsub submit.sh
```

- 指定 queue 名稱

```
user1@master ~]$ qsub -q vip submit.sh
```

- 指定執行時間

```
$ qsub -l walltime=24:00:00 submit.sh
```

- 指定 Job 名稱

```
$ qsub -N hello submit.sh
```

Submit Job 範例

```
#!/bin/bash
#PBS -N job
#PBS -o folder path
#PBS -e folder path
#PBS -q workq
#PBS -l nodes=1:ppn=2
cd $PBS_O_WORKDIR

echo “PBS_O_WORKDIR : ” $PBS_O_WORKDIR > job.out
echo "pwd : " `pwd` >> job.out
echo $PBS_JOBID
hostnamectl >> hostnamectl.out
echo "job started“ >> job.out
sleep 60
echo "wake up“>> job.out
```

跑跑看HPL

```
#!/bin/bash
#PBS -N job
#PBS -o folder path
#PBS -e folder path
#PBS -q workq
#PBS -l nodes=1:ppn=2
cd $PBS_O_WORKDIR

source /opt/intel/oneapi/setvars.sh intel64
export PATH=/opt/openmpi/4.1.6_intel_2024.2.0/bin:$PATH
export LD_LIBRARY_PATH=/opt/openmpi/4.1.5_intel_2024.2.0/lib:$LD_LIBRARY_PATH

if [ -n "$PBS_NODEFILE" ]; then
                if [ -f $PBS_NODEFILE ]; then
                echo "Nodes used for this job:"
                cat ${PBS_NODEFILE}
                NPROCS=`wc -l < $PBS_NODEFILE`
                fi
fi

echo "Starting on `hostname` at `date`"
mpirun -hostfile $PBS_NODEFILE -n $NPROCS $HOME/hpl/bin/xhpl > HPL.out
```

Jobs 執行監控及操作

- 查詢所有 Jobs 狀態

```
$ qstat
```

- 查詢所有 Jobs 更多狀態

```
$ qstat -a
```

- 顯示某特定 Job 詳細狀態

```
$ qstat -f JOBID
```

- 查詢某使用者 Jobs 狀態

```
$ qstat -u USERID
```

- 查詢 Jobs 歷史紀錄

```
$ qstat -x
```

Jobs 執行監控及操作

- 查詢某使用者歷史紀錄

```
$ qstat -xu USERID
```

- 顯示某 Job 使用哪些機器

```
$ qstat -xf JOBID | grep exec_host
```

- 顯示系統目前的 queues 狀態

```
$ qstat -Q
```

- 顯示系統目前的 queues 詳細狀態

```
$ qstat -fQ
```

- 顯示系統目前的某 queue 詳細狀態

```
$ qstat -fQ workq
```

Jobs 執行監控及操作

- 暫停某 Job 執行

```
$ qhold JOBID
```

- 釋放某 Job 狀態

```
$ qrls JOBID
```

- 停止執行某 Job

```
$ qdel JOBID
```

- 強制停止某 Job

```
# qdel -W force JOBID
```

- 執行某 Job

```
# qrun JOBID
```

Jobs 執行監控及操作

- 變更某 Job 優先權

```
# qalter JOBID -p 150
```

- 變更某 Job 使用時間

```
# qalter JOBID -l walltime=02:00:00
```

- 重新跑某 Job

```
# qrerun JOBID
```

Nodes 維護及操作

- 列出某 node 的狀態

```
# pbsnodes cn1
```

- 列出有狀況的 nodes

```
# pbsnodes -l
```

- 使某 node 離線進行維護

```
# pbsnodes -o cn1
```

- 恢復某 node 運算服務

```
# pbsnodes -c cn1
```

設定檔及 logs 路徑

- PBS 伺服器使用紀錄路徑

```
/var/spool/pbs/server_priv/accounting/
```

- PBS 伺服器 logs 路徑

```
/var/spool/pbs/server_logs
```

- PBS 排程設定檔

```
/var/spool/pbs/sched_priv/sched_config
```

- PBS 排程 logs 路徑

```
/var/spool/pbs/sched_logs
```

設定檔及 logs 路徑

- PBS 運算節點設定檔

```
/var/spool/pbs/mom_priv/config
```

- PBS 運算節點 logs 路徑

```
/var/spool/pbs/mom_logs
```