Score＋SGE 利用の手引き

第１版

改版履歴

日付	版数	変更内容	変更者
2002/09/27	第１版	新規作成	武澤　慎

目次

１．Score とは （ http://www.pccluster.org ）

SCore Cluster System Software はワークステーションおよび PC クラスタ用の高性能並列プログラミング環境です。特徴としては、1)高性能通信、2)SSI による効率的なシステム管理、3)高利用性および高可用性、4)シームレスなクラスタ環境、5)並列プログラミング環境ということがあります。

２．Sun Grid Engine（ SGE ）とは

Sun Grid Engine とは Sun Microsystems 社が公開しているグリッドコンピューティングシステム構築ソフトウェアです。クラスタシステムにおけるジョブ受付、実行ホスト・キューでのリソース使用状況監視、ジョブが要求するリソースに見合った最適な実行キューを割り当て、実行ホストでジョブの実行をします。

３．Score＋SGE とは

SCore の特徴である高速クラスタ機能を損なわず、SGE のジョブスケジューリングを適用した運用を可能とします。

４．Score＋SGE の運用設定

4.1【 概要 】

SCore と SGE の連携を可能とします。具体的には、SGE の PE(Parallel Environment) においてスレーブキューを持つホスト群と SCore のホストグループを対応させ、SCore の実行ジョブを SGE でコントロールすることを可能とします。

4.2【 システム構成 】

○ SCore、SGE　実装バージョン

SCore：4.2
SGE　：5.3b2

○ ホスト構成

マスターホストを設定し、それに対して複数の実行ホストを持つ構成とします。マスターホストは、SCore+SGE においてマスターサーバとなります。

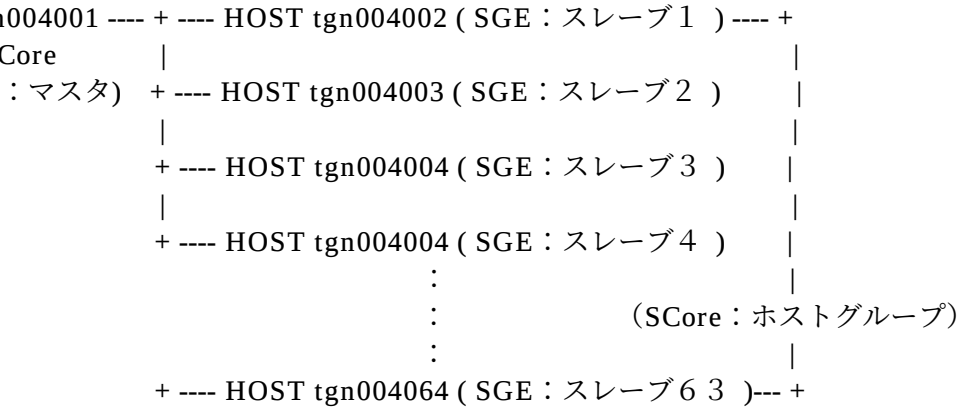
SGE：
各ホストにおいて、PE を構成するためにキューを 1 つ設定します。

SCore：

SGE の PE におけるスレーブキューとなる各ホストを一つのグループとして設定します。

センタ B2 システムでは、以下の構成となります。

センタ B2 システム環境（　tgn004001 ～ tgn004064 (64node) ）



HOST tgn004001:

SCore ： SCored server (scoreboard)

SGE ： SGE master server (sge_qmaster)

HOST tgn0040xx:

SCore ： 実行ホスト

SGE ： 実行ホスト (sge_execd)

4.3【 環境設定 】

ユーザ管理、PE 作成などの実行環境構築のための各設定を示します。

なお、各 SGE のコマンドの実行には、最初に環境変数を設定するスクリプトを実行しておく必要とがあります。

設定スクリプトファイル： SGE_ROOT_DIR /default/common/settings.csh
settings.sh
（センタ B2 システム　SGE_ROOT_DIR：/home/cc/gusr20/necst/SGE5.3）

Ex.)
 # source　SGE_ROOT_DIR /default/common/settings.csh

○ SGE - ユーザ管理

SCore+SGE を使用するユーザの設定（登録）を行います。SCore+SGE では、SGE のユーザ管理に従っています。SGE のインストール時にもユーザの設定はできますが、ここで

は後から設定する方法を説明します。なお、ユーザ管理は、Manager の権限が必要です。

設定可能なユーザタイプは以下のとおりです。

・Manager

Manager は、SGE を操作する十分な能力を持っています。デフォルトで、マシンをホスティングするキューのスーパーユーザは Manager 権限を持っています。また、すべてのユーザを追加、削除することができます。

・Operator

Operator は、キューを追加、削除、修正する、設定変更を加える例外を除いて、Manager と同じく、多くのコマンドを実行することができます。ユーザの追加は、User を追加、削除することができます。

・User

User は、アクセス権は持っていますが、クラスタやキューの管理能力はありません。ユーザの設定もできません。ジョブの投入、状態確認、削除(自分で投入したジョブ)等はできます。

通常、SGE をインストールしたユーザは Manager となりますので、そのユーザ権限にて設定を開始します。 設定法には、コマンドからと GUI からの 2 種類があります。
なお、センタ B2 システムにおきましては、root が、Manager となっております

コマンドによる設定）

qconf コマンドによりをユーザ設定を行います。

形式： qconf [option] user[,user….] [user_list]

ユーザの追加時の option には Manager は -am 、Operator は -ao、User は -au を指定します。なお、User タイプの登録時には、user_list の指定が必要です。user_list はユーザリスト名として任意に指定可能です。

Ex.)

```
#qconf  -au  nec1  group1
added “nec1” to access list “group1”
#qconf  -au  nec2  group1
added “nec2” to access list “group1”
```

設定ユーザを確認する場合は、qconf の option に Manager は -sm 、Operator は -so、User は -su および user_list を指定します。

Ex.)

```
#qconf  -su  group1
name      group1
entries     nec1,nec2
```

GUI による設定)

qmon コマンドで ブラウザを起動し、UserConfiguration ボタンを押してください。各ユーザタイプでの設定が可能となります。

Ex.)

```
# qmon
```

○ SGE - PE(Parallel Environment)設定

SGE-PE の設定を行います。PE の設定をするには、Manager 権限が必要です。
なお、センタ B2 システムにおきましては、以下の PE を設定済みです。

PE 名： test1 （マスタキュー ： tgn004001.q
スレーブキュー： tgn004002.q～tgn004064.q (63node)）

コマンドによる設定）

qconf コマンドで PE(Parallel Environment)を追加します。ここで追加する PE 名を test とすると、エディタが起動され、次のように表示されます。

```
# qconf –ap test

pe_name           test
queue_list        all
slots             0
user_lists        NONE
xuser_lists       NONE
start_proc_args   /bin/true
stop_proc_args    /bin/true
allocation_rule   $pe_slots
control_slaves    FALSE
job_is_first_task TRUE
```

queue_list に追加したいキューを入力してください。すべてのキューを追加するには all と指定します。

```
queue_list      all
```

slots には総プロセス数を入力してください。

user_lists には、使用を許可するユーザ、ユーザリストを入力してください。
設定しない場合は NONE のままにしてください。

SCore+SGE 連携用の PE スタートアッププロシジャ(開発物件)を PE のスタートアッププロシジャとして設定します。その際に SGE の選択ホストを取得できるように設定します。
($pe_hostfile：SGE での並列ジョブ選択ホスト名の一覧ファイル)

start_proc_args root@/xx/xx/pe_startup $pe_hostfile

（センタ B2 システム　PE：test1　の設定
root@/home/cc/gusr20necst/SGE5.3/default/common/pe_startup　$pe_hostfile）

ジョブにおいて、1CPU あたり 1 プロセスと設定します。なお、この設定は、SGE でのホスト割り当て制御のためであり、実際のホスト単位の実行プロセス数には影響しません。

allocation_rule 1

マスターキューを持つマスターホストを、実行ホストに割り当てないように設定します。

Control Slaves TRUE
Job is first task TRUE

これらを設定すると、以下のように表示となります。

qconf -ap test

pe_name test
queue_list all
slots 100
user_lists NONE
xuser_lists NONE
start_proc_args root@/< pe_startup へのパス >/pe_startup $pe_hostfile
stop_proc_args /bin/true
allocation_rule 1
control_slaves TRUE
job_is_first_task TRUE

設定を保存（書き込み）して終了すると PE が追加されます。

GUI による設定）

qmon コマンドで qmon ブラウザを開いてください。

PE Configuration ボタンを押してください。そして、add ボタンを押して新規に PE を作成します。

queue_lists には、必要なキューを選択して追加してください。すべてのキューを選択するには、all にチェックを入れください。

slots には、必要なプロセス数を入力してください。

user_lists には、使用を許可するユーザ、ユーザリストを入力してください。

start_proc_args には、root@/< pe_startup へのパス >/pe_startup $pe_hostfile を入力してください。

ジョブにおいて、1CPU あたり 1 プロセスと設定しますので、allocation_rule を 1 に設定してください。
マスターキューを持つマスターホストを、実行ホストに割り当てないように Control Slaves、Job is first task のチェックボックスをオンにしてください。

そして OK ボタンを押すと設定が追加されます。修正したい場合には、modify ボタンを押して、設定内容を修正してください。

○ SCore

・SGE の PE と連携する実行ホストのグループを作成します。
(以降、この作成したグループ名を' pcc 'とします。)

なお、センタ B2 システムにおきましては、以下のグループを設定済みです。

グループ名： pccall2 （tgn004002～tgn004064 (63node)）

4.4【 システム起動 】

○ SCore

ホストグループ(pcc)をマルチユーザモードで起動します。その際に、SCore のコンソールメッセージ出力を保存するように設定します。

センタ B2 システムにおきましては、SCore 起動、およびコンソールメッセージ出力を保存設定するスクリプトを以下に格納していますのでご利用ください。

コマンドスクリプトファイル ：/opt/score/bin/scored_syslog
コンソールメッセージ格納場所：tgn004001:/tmp/scored.message

ex.)

```
# scbcast   syslog
scbcast started
```

コンソールメッセージを格納するホスト名、格納場所（ファイル）を設定します。

```
#sc_syslog   host_name   /tmp/scored.message
sc_syslog started

# scout   -g   pcc

SCOUT: Spawning done.
SCOUT: session started.
```

コンソールメッセージ出力先を host_name に指定して起動します。

scored -syslog host_name

コンソールメッセージを確認し、以下ように表示されていれば scored サーバが正常起動しています。

```
SYSLOG: /opt/score5.0.0/deploy/scored
SYSLOG: SCore-D 5.0.0 $Id: init.cc,v 1.66 2002/02/13 04:18:40 hori Exp $
SYSLOG: Compile option(s):
SYSLOG: SCore-D network: ethernet/ethernet

                    :
SYSLOG: Operated by: root
SYSLOG: ========= SCore-D (5.0.0) bootup in SECURE MODE ========
```

○ SGE

・起動方法は、インストール時にシステムの起動スクリプトとして登録されるために通常は、システム立ち上げに起動されます。

SGE の起動確認は、以下の方法で確認することができます。確認結果、各ホストで必要なプロセスが起動していれば正常です。

% ps -A | grep sge

マスターホスト)

```
sge_commd
sge_qmaster
sge_schedd
sge_execd
```

実行ホスト)

```
sge_commd
sge_execd
```

もし、起動していない場合は、起動スクリプトを実行してください。

/etc/rc.d/init.d/rcsge start

4.5【 SCore+SGE のシャットダウン 】

○ SCore

別のコンソールでマスターサーバにログインして、root になります。

下のコマンドを実行して scored サーバをシャットダウンします。
sc_console scored_server_name -c shutdown

そうすると scored サーバを立ち上げたコンソール画面に

SYSLOG: CONSOLE connected from popeye.hpc.necst.nec.co.jp
CONSOLE: >> shutdown
SYSLOG: SCore-D shutting down in 0 seconds.
SYSLOG: Login disabled.
SYSLOG: Waiting for all job terminates.
SYSLOG: CONSOLE shutdown
SYSLOG: SCore-D shutdown.

と表示されます。つぎに、scored サーバだったコンソール画面で
ロックしていたホストを解除します。

exit

SCOUT: Session done.

と表示されると SCore のシャットダウン完了です。

○ SGE

各ホストで、以下の起動スクリプトを実行してください。

/etc/rc.d/init.d/rcsge stop

５．Score＋SGE の利用

5.1【 利用の流れ 】

SCore+SGE を利用してジョブを実行する流れは下記の通りです。

ユーザはジョブ投入ホストにログインし、SGE の環境変数を設定します。
（詳細は、4.3【環境設定】を参照願います）
↓
ジョブ実行のスクリプトファイルの作成します。
↓
qsub2 コマンドを用いて、PE にジョブを投入します。
↓
ジョブ実行結果を取得します。

5.2【 ジョブ実行について 】

○ ジョブ投入方法

qsub2 コマンドにより、ジョブを投入します。 qsub2 コマンドは、本システムでマスタとなるホストで実行してください。

qsub2 コマンド形式：

形式：qsub2 -pe PE_name n[-proc m]-masterq masterq_name[options]jobscript

機能：-pe　PE_name　PE 名
　　n　実行総プロセス数
　-proc　m　ホスト単位の実行プロセス数
　-masterq　masterq_name　PE のマスタキュー名
　jobscript　ジョブスクリプト

scrun コマンドの-node オプションとの対応は以下のとおりです。

scrun		qsub2				
-nodes	○x□	：	-pe PE_name	◎	–proc	□　（◎ ＝ ○ ＊ □）
-nodes	△	：	-pe PE_name	△		（–proc 指定は省略）

また、実際に SGE が割り当てるホスト数は、-proc の指定の有無により以下のようになります。

-proc 有り：n/m (n%m != 0 であれば、n/m+1)
-proc 無し：n/ホスト単位の CPU 数 (n%CPU 数 != 0 であれば、n/CPU 数+1)

なお、センタ B2 システムでは、ホスト tgn004001 にて以下の PE 名、マスタキューの指定でご利用頂けます。

PE 名　：　test1
マスタキュー　：　tgn004001.q

Ex.)

```
# qsub2  -pe  test1  10  -masterq  tgn004001.q  test.sh
```

○ ジョブスクリプトファイル

スクリプトファイルのイメージは以下の通りです。なお、SCore+SGE 連携ジョブの判断/環境設定は、scrun スクリプトファイル内で実行します。
scrun の形式に変更はありませんが、-nodes オプションの指定は不要です。

形式：scrun [-SCoreOptions] file [program_options]

機能：file　実行プログラム

ex.) /xx/xx/a.out を実行する。

```
-----------------------------------------------------------------
#!/bin/sh
#$ -S /bin/sh
scrun   /xx/xx/a.out
-----------------------------------------------------------------
```

○ ジョブ実行結果の確認

ジョブの標準出力/エラー出力は、SGE の各ジョブ結果ファイルに格納されます。
ファイルの出力先は、既定値では各ユーザのホームディレクトリとなります。

5.3【 ジョブ制御コマンド 】

○ ジョブの実行状態の確認

qstat コマンドを使用します。

ex.)

・qstat コマンドで状態を表示する。

```
$ qstat -f

queuename                qtype used/tot. load_avg arch          states
---------------------------------------------------------------------------------------------------------------
tgn004001.q              BIP   3/20       0.52     glinux
      6      0 score.sh      user         t      09/24/2002 17:54:32 MASTER
      7      0 score.sh      user         t      09/24/2002 17:54:32 MASTER
      8      0 score.sh      user         t      09/24/2002 17:54:32 MASTER
---------------------------------------------------------------------------------------------------------------
tgn004002.q              BIP   1/10       0.02     glinux
      7      0 score.sh      user         t      09/24/2002 17:54:32 SLAVE
---------------------------------------------------------------------------------------------------------------
tgn004003.q              BIP   1/10       0.02     glinux
      7      0 score.sh      user         t      09/24/2002 17:54:32 SLAVE
---------------------------------------------------------------------------------------------------------------
tgn004004.q              BIP   2/10       0.00     glinux
      6      0 score.sh      user         t      09/24/2002 17:54:32 SLAVE
      8      0 score.sh      user         t      09/24/2002 17:54:32 SLAVE
---------------------------------------------------------------------------------------------------------------
tgn004005.q              BIP   2/10       0.00     glinux
      6      0 score.sh      user         t      09/24/2002 17:54:32 SLAVE
      8      0 score.sh      user         t      09/24/2002 17:54:32 SLAVE
                                  :
                                  :
#############################################################################
```

```
- PENDING JOBS - PENDING JOBS - PENDING JOBS - PENDING JOBS - PENDING
############################################################################
          9      0 score.sh      user         qw      09/24/2002 17:54:26
         10      0 score.sh      user         qw      09/24/2002 17:54:27
```

○ ジョブの削除

qdel コマンドを使用します。

・qdel コマンドで qstat で表示された削除したいジョブ ID を指定します。

$ qdel 10

user has deleted job 10

これで削除されます。

６．コマンド一覧

SCore+SGE で利用可能なコマンドを下記に示します。

・qsub2：submit a job to SCore

SCore にジョブを投入します。

qsub2 -pe PE_name n [-proc m] -masterq masterq_name
[options] jobscript

・qdel：delete Grid Engine jobs from queues

グリッド・エンジンのジョブをキューから削除します。

qdel [-f] [-help] [-verify] [job/task_id_list]

・qconf：Grid Engine Queue Configuration

グリッド・エンジンのキューの設定をします。

qconf options

・qhost：show the status of Grid Engine hosts, queues, jobs

グリッド・エンジン・ホスト、キュー、仕事の状態を示します。

qhost [-F [resource_name,...] [help] [h host_list]
[j] [l resource=val,...] [u user,...]

・qstat：show the status of Grid Engine jobs and queues

グリッド・エンジンのジョブおよびキューのステータスを示します。

qstat [ext] [f] [-F [resource_name,...]] [g d]
[help] [j [job_list]] [l resource=val,...] [ne]
[pe pe_name,...] [q queue,...] [r]
[s {r|p|s|z|hu|ho|hs|hj|ha|h}[+]]] [t] [U user,...]
[u user,...]

・qmod：modify a Grid Engine queue

グリッド・エンジンのキューを修正します。

qmod [options] [job/task_id_list | queue_list]

・qmon：XWindows OSF/Motifgraphical user's interface for Grid Engine

グリッド・エンジンの XWindows OSF/Motif のグラフィカルユーザーインターフェースです。

qmon [options]

また、各ユーザタイプのコマンド能力は以下のようになっています。

Command	Manager	Operator	Owner	User
qdel	FULL	FULL	Own jobs only	Own jobs only
qhost	FULL	FULL	FULL	FULL
qconf	FULL	No system setup modifications	Show only configurations and access permissions	Show only configurations and access permissions
qmod	FULL	FULL	Own jobs and owned queues only	Own jobs only
qmon	FULL	No system setup modifications	No configuration changes	No configuration changes

qstat	FULL	FULL	FULL	FULL
qsub2	FULL	FULL	FULL	FULL

７．注意／制限事項

7.1【 注意事項 】

○ SGE

・PE の設定で user_lists でユーザを指定すると、それ以外のユーザはジョブを投入できても、そのジョブは PENDING されたままになり実行されません。そのジョブは qdel コマンドで削除してください。

・（qmon コマンドによる）GUI からのジョブ投入はできません。

・ジョブ実行時にメモリなどのリソース不足で異常終了した場合、該当実行ホストの SGE デーモンプロセスも影響を受けて終了する場合があります。その場合は、該当ホストの SGE デーモンプロセスを再起動してください。SGE デーモンプロセスの起動確認／再起動につきましては、「4.4【システム起動】」を参照願います。

○ SCORE

・SCore の起動時に、scored コマンドで-server 指定はしないでください。

7.2【 制限事項 】

○ SGE

・チェックポイント／リスタート機能は使用できません。