TSUBAME2.0 NAREGI環境 利用の手引き

東京工業大学学術国際情報センター
2011.01
version 1.0

目次

NAREGIでのTSUBAME2.0利用の手引き

1. はじめに

本書は、NIIポータルからTSUBAME2.0にジョブを投入する方法を記しています。はじめにTSUBAME2.0の概要、次いでジョブ投入、そして並列ジョブの投入、メモリの指定方法、特定アプリケーションの利用方法について述べています。本書はNAREGI(9大学学術グリッド)連携試験参加者のためのガイドです。NIIポータルのログイン、Sign on等NAREGI共通分は、「NAREGI上でのジョブ実行」を参照してください。

また、TSUBAME2.0の構成、詳細利用方法は以下のページを参照してください。
TSUBAME 2.0 利用の手引き
FAQ

※TSUBAMEのNAREGI連携環境に関しての注意点
NAREGI用の計算クライアントの構成により、使用できるキュー構成が変化する場合があります。

2010年11月～ の環境では、以下の様になっております。

性能保証 ・・・・・・・・ 全体で375ノード:

```
S キュー    300ノード
S96キュー   41ノード
L128キュー  24ノード
L256キュー  8ノード
L512キュー  2ノード
```

2. TSUBAME2.0 概要

TSUBAME2.0は，HP社製「SL390s G7」を中心としたクラウド型グリーンスーパーコンピュータシステムです。演算サーバ「SL390s G7」は12CPU(cores)/node, Naregi連携実験用のノードは54GBメモリを実装しています。そのため、TSUBAME2.0に投入するジョブでは、並列数は24以下(Hyper-Threadingによる最大値)、メモリサイズは52GB以下を指定してください。この制限を越えると、投入されたジョブは計算資源不足のエラー、または、無限時間に計算資源を 待つ状態になる可能性があります。十分にご注意ください。

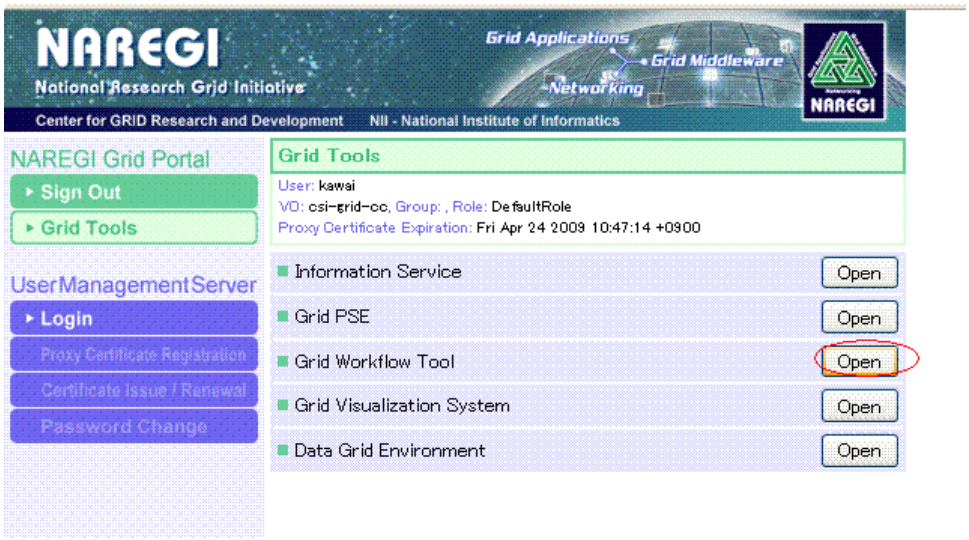
並列ジョブについて、TSUBAME2.0はMPI,OpenMPI,SMP,Linda,DDIなどをサポートしますが、GridMPIは実装されておりません。そのため、シングル、並列ジョブに係わらず、WFTでは必ず「Add program Icon」を利用してジョブを投入してください。
並列ジョブの指定などは環境変数で指定します。詳細は 4 並列ジョブ を参照して下さい。また、ジョブ投入時に環境変数で使用するメモリ量を指定できます。無指定の場合はプロセスあたり1GBになりますので、これ以上のメモリを使用する場合はメモリサイズを指定してください。詳細は 4.4 メモリ指定 を参照してください。

特定のアプリケーションを利用したい場合には、アプリの指定および必要な環境設定などは、WFTにて環境変数を指定することにより利用可能となります。詳細は 5 特定アプリの利用 を参照してください。

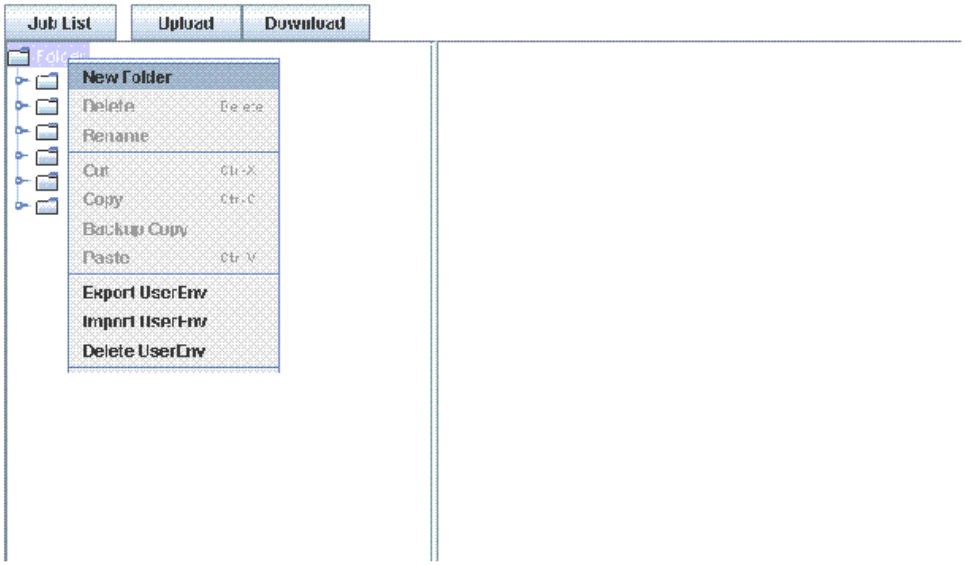
3. NAREGIポータルからのTSUBAME2.0利用方法

3.1 WFT起動

NII Portalにログインしてsign onまでの操作はここでは述べませんので、「NAREGI上でのジョブ実行」のP11までをご参照ください。「Sign on」ができたら、ポータルメニュー画面を表示します。(下図参照)



上記画面から、「Grid Work Tool」の「Open」をクリックすると、WFTが起動されます。(下図参照) 「Folder」を右クリックして、メニューをの中から 「New Folder」を選択(クリック)して、テスト用フォルダー(ディレクトリ)を作成します。



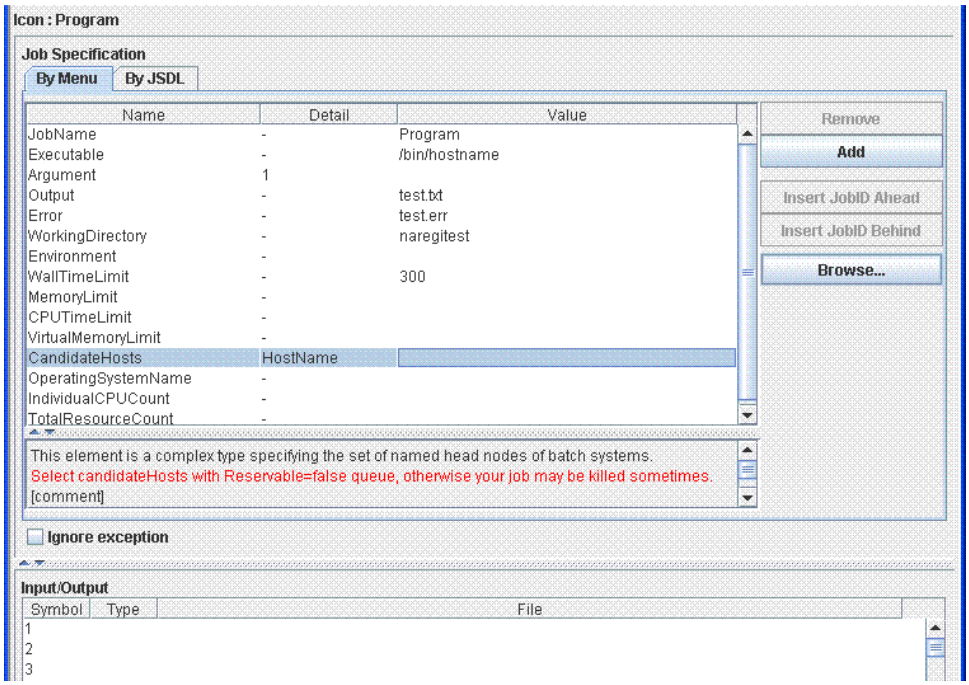
作成された新規フォルダは好きな名前に変更できます。テストディレクトリをクリックすると、ファイルが存在する場合は 以下のような画面が表示されますが、何もない場合は、何も表示されませんので、

画面の右側を右クリックしてメニューから 「New Workflow Icon」選択して、WFTアイコン作成画面(下図参照)を表示します。
※ 計算資源TSUBAME2.0を選択の場合、シングル、並列ジョブに係わらずに 「New Workflow Icon」を選択してください。

ここで、「Add program Icon」をクリックして、ジョブ投入用「Program Icon」作成画面を表示します。

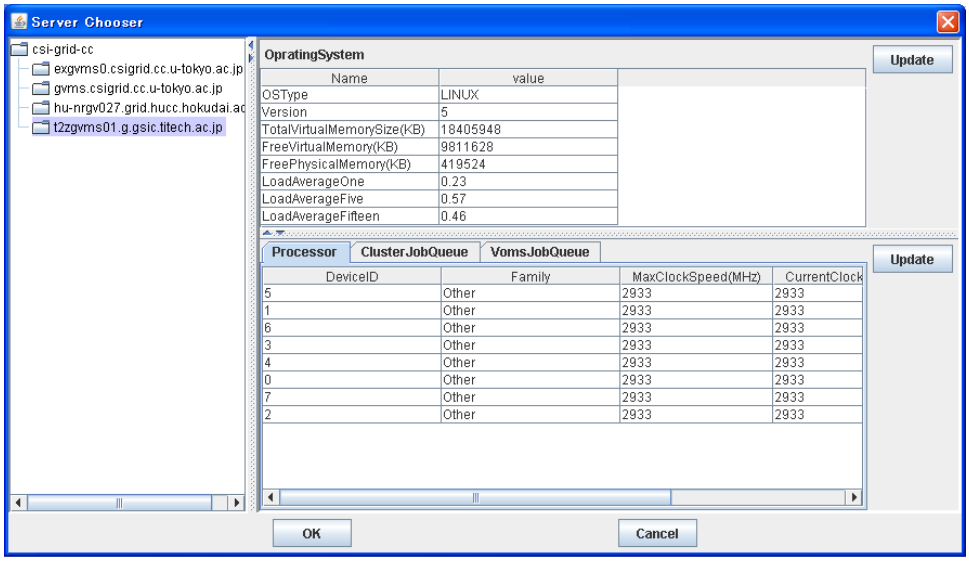
3.2 TSUBAME2.0指定

「Program Icon」作成画面で、ジョブ投入に必要な項目を入力します。起動時点では、JobNameとWallTimeLimitのみが入っています。以下は、"hostname"コマンドの表示ジョブを投入しようとしている例です。

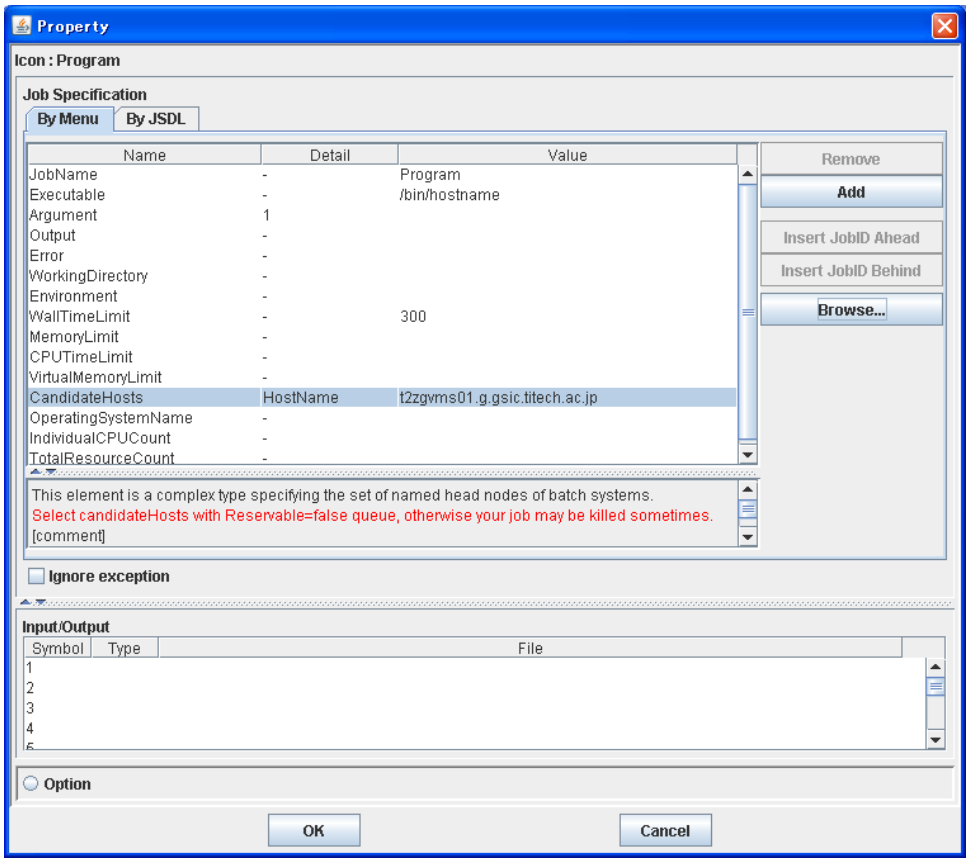


上の画面において、「CandidateHosts」を選択して、右の「Browse」クリックすると、計算資源の選択画面を表示します。TSUBAME2.0を利用したい場合は、t2zgvms01.g.gsic.titech.ac.jpを選択してください。(下図参照)

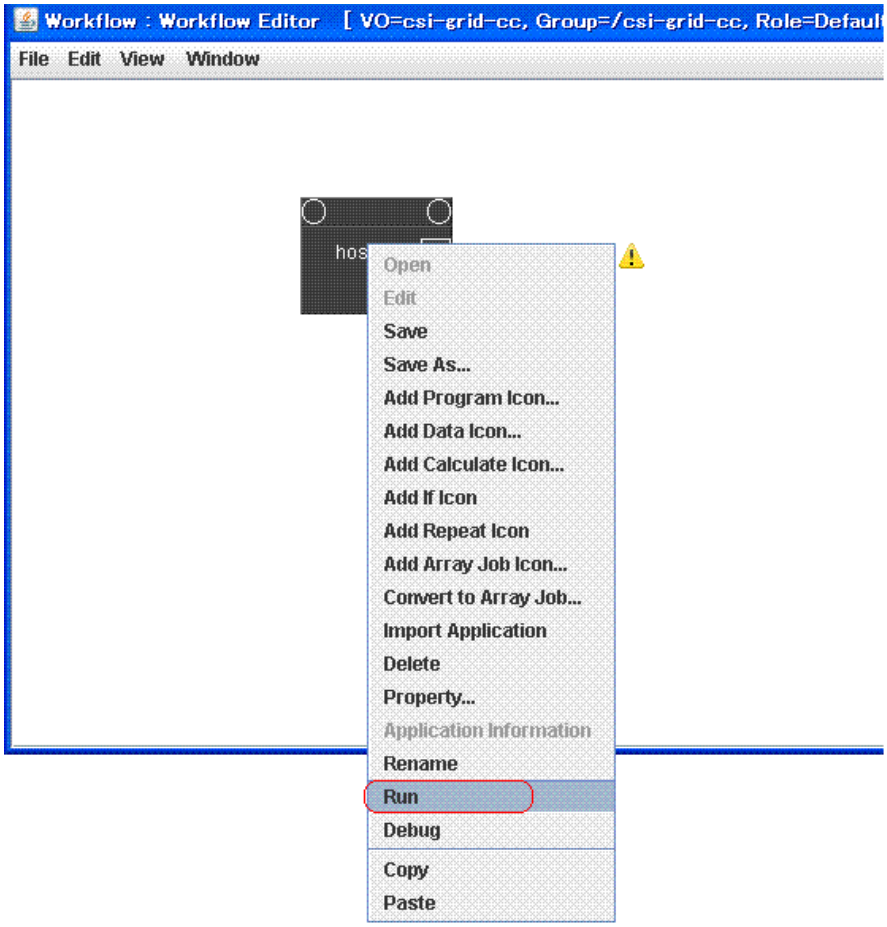
3.2 TSUBAME2.0指定



この状態で「OK」をクリックすると、「Program Icon」は以下のように TSUBAME2.0のスケジューラを表示した状態になります。

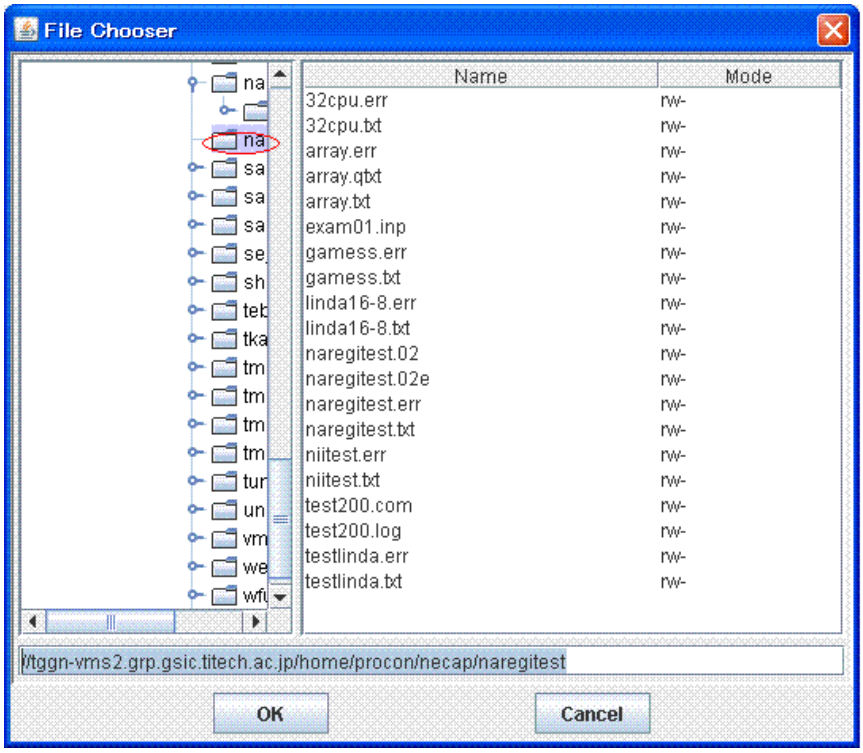


「OK」ボタンをクリックすると、「Program Icon」を作成して選択画面が終了します。作成されたアイコンを右クリックして、「Run」サブメニューをクリックすると、作成されたジョブがTSUBAME2.0に投入されます。



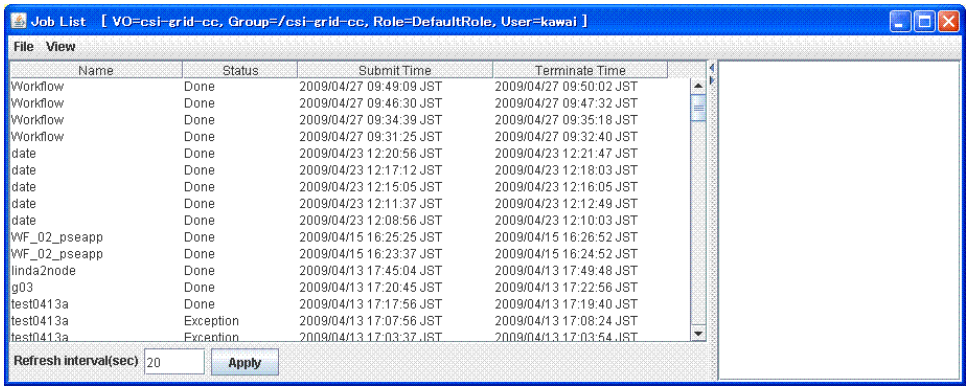
3.3 入力ファイルがある場合

上の例では入力ファイルが必要ない場合でした。ジョブ実行時に入力ファイルが必要な場合には、ジョブ投入前に入力ファイルをローカルで用意してから、WFTのアップロード 機能を利用してt2zgvms01にアップロードしてください。手順は、NAREGI上でのジョブ実行のPage14-15を参照してください。但し、「File Chooser」のアップロード先については、「t2zgvms01.g.gsic.titech.ac.jp」を 選択してください。またアップロード先のディレクトリは「Program icon」作成時に指定された Directoryと同じにしてください。

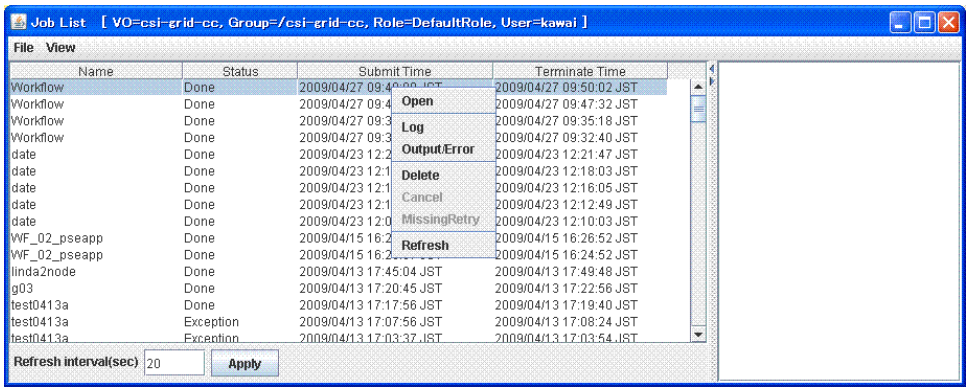


3.4 ジョブ確認

投入されたジョブを確認したい場合に、他のnaregiジョブと共通で、「WFT」画面の「Job List」ボタン をクリックすると、ジョブのリスト画面(一覧)が出ます。確認したいジョブが出ない場合は、暫くお待ちいただいて、改めて「Apply」ボタンをクリックしてください。最新の投入ジョブが一番上に表示されます。



確認したいジョブを選択して右クリックすると、ジョブの「Log」「Out/Err」を選択できます。ここでジョブログ、標準出力/エラー情報を表示できます。



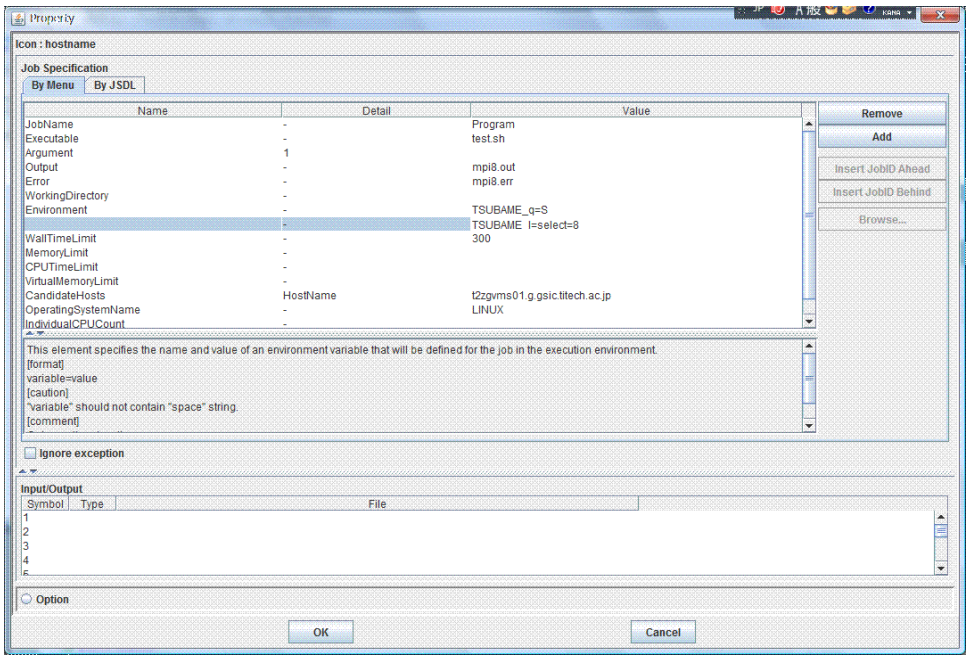
プログラムの実行結果などをファイルに出力する場合には、「WFT」の「DownLoad」機能を 利用してください。「DownLoad」の操作方法は Naregi Middleware WFT利用ガイドの12.2(Page 116-118) を参照してください。

4 並列ジョブ、メモリ指定等

並列ジョブもシングルジョブと同様に「Add program」でジョブiconを作成しますが、環境変数のところで並列数などを指定します。

4.1 MPIジョブ

TSUBAME2.0 はPBSProのバッチキューシステムを採用しています。TSUBAME2.0 にMPIジョブを投入する場合は、の-l select [引数] オプションでMPI ジョブを識別しています。NII portalからMPIジョブを投入する場合には、専用環境変数TSUBAME_l=selectでCPU数を指定の上、実行シェルにmpirunを記述すれば、MPI並列ジョブになります。指定フォーマットはTSUNAME_l=select=:mpiprocs=[] です。例えば、メモリ無指定(1GB/proc利用)、MPI 8並列ジョブの場合、「Program Icon」の入力は以下になります。



ジョブスクリプトは以下となります。

```
cat test.sh
#!/bin/sh
cd $PBS_O_WORKDIR
mpirun -np 8 -hostfile $PBS_NODEFILE ./a.out
```

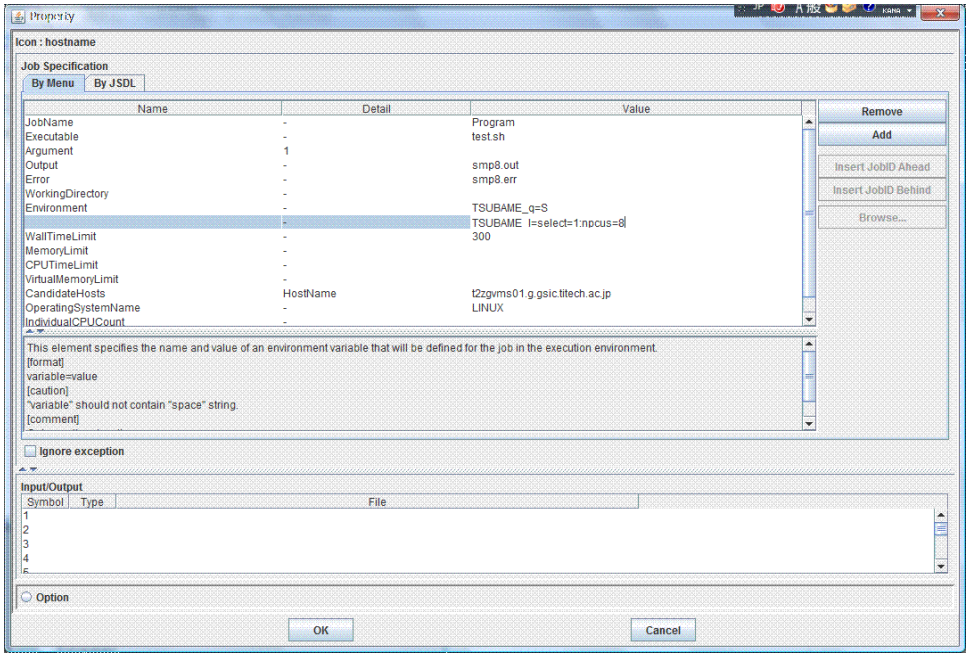
4.2 他のMPI環境利用

投入画面の設定は同じですが、MPIの選択はジョブスクリプトにより指定できます。無指定の場合には、OpenMPI(intel)の利用となります。たとえば、プログラムがopenmpi + PGIで作成された時に、ジョブスクリプトは次のようになります。

```
#!/bin/sh
cd $PBS_O_WORKDIR
source /usr/apps/free/env/setompi-pgi.sh
mpirun -np 4 -hostfile $PBS_NODEFILE ./a.out
```

4.3 SMP/proc並列

TSUBAME2.0ではSMP並列ジョブ、同時実行ジョブ(proc)もサポートしています。SMP/procのジョブ投入時のオプション指定例は以下の画面の様になります。



4.3.1 SMP並列

SMP並列ジョブはシングルジョブと同様に実行するモジュール名を指定してください。また、使用するスレッド数につきましては、明示的に指定するようにお願いします。

```
#!/bin/sh
#
# sample jobscript
export OMP_NUM_THREADS=2
export NCPUS=2
#
cd $HOME/test
./myprog < input_data
```

4.3.2 同時実行ジョブ

この方法は、シングルジョブを同一ノード内で同時に実行させる方法で、TSUBAME1では「proc」オプションを使用していたユーザーに同様にご利用いただける方法です。

指定するオプションとしては，ノード内並列と同様の指定になります．

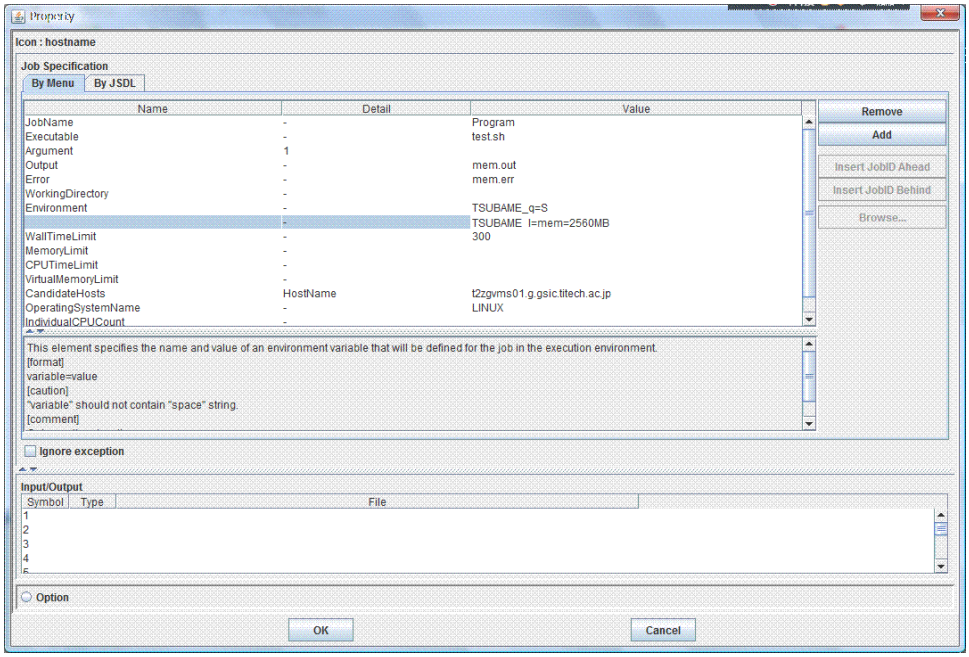
```
++++ job.sh ここから
#!/bin/bash

for n in 1 2 3 4 5 6 7 8; do
  ./a.out < input$n > output$n &
done
wait
++++ job.sh ここまで
```

4.4 メモリ指定

TSUBAME2.0にジョブを投入する場合に、プロセスあたり1GB以上のメモリが必要になれば、任意のメモリ数の指定ができます。(ただし、52GB以下) portalからジョブを投入する場合、TSUBAME_l=mem=「メモリ値」でというオプションでメモリを予約します。このメモリ値の指定はbyte,kb,mb,gb単位で指定できます。

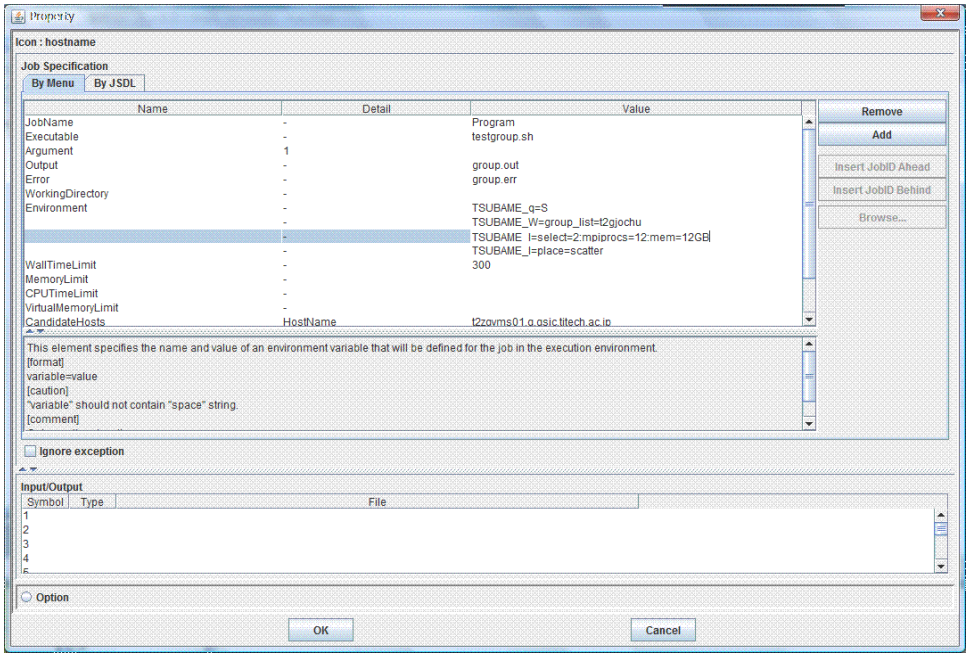
例えば、シングルジョブで、2.5GBのメモリを要求する場合の指定は以下のような画面となります。環境変数が複数の場合には、マウスにて「Enviroment」の行を選択して、画面右の「Add」ボタンを押して、環境変数を追加できます。



4.5 課金グループ指定

TSUBAME2.0バッチシステムでは、課金グループ制度が導入されています。グループを無指定の場合は、無料で短時間ジョブを利用できますが、初心者、プログラムテストなどに限定しています。詳細は キュー構成 、`TSUBAME FAQ <http://tsubame.gsic.titech.ac.jp/ja/faq>`_ を参照してください。

naregiテストジョブも同じで、大規模ジョブ利用の場合、課金グループ を指定して利用してください。利用方法はTSUBAME_W=group_list=「課金グループ」で課金グループ を指定しての利用になります。課金グループを指定して、2ノード24CPUを利用、ノードあたり12GBメモリの指定の場合は以下となります。



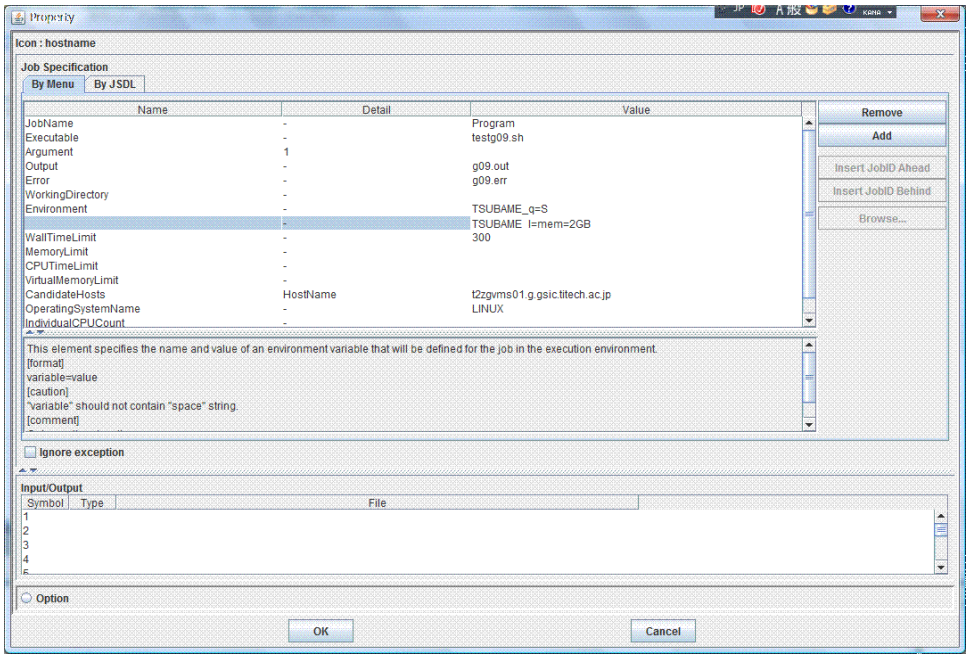
5 TSUBAME2.0アプリ利用

TSUBAME2.0では，様々なISV アプリケーション、フリーソフトが利用可能です。詳細は、TSUBAME2.0利用の手引きの 3.3 アプリケーション をご参照ください。但し、ライセンスの関係で外部利用者は、多くのISVアプリケーションを利用できません。ここでは利用できる数少ないISVアプリケーションのGaussian、freeアプリケーションのGAMESSの ジョブの投入方法を説明します。

5.1 Gaussianジョブ

Gausssianジョブ投入する場合は、基本的にはインプットファイルと実行シェルを用意してから、「Program Icon」の作成画面で、実行コマンド行に実行シェル、実行シェル中にGaussianを指定して、Gaussianジョブを投入できます。またはSMP並列、あるいはLindaで並列ジョブを投入できます。

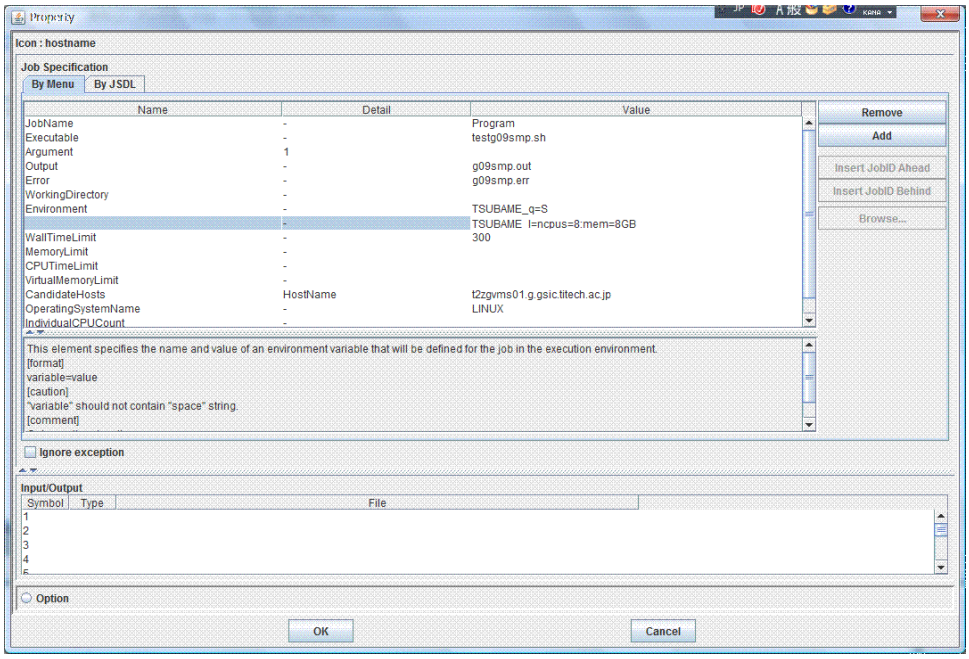
1). Gaussianシングルジョブ



ジョブスクリプトは以下となります。

```
cat testg09.sh
#!/bin/sh
cd $PBS_O_WORKDIR
export GAUSS_SCRDIR=$TMPDIR
g09 test.com
```

2). Gaussian SMPジョブ ※ この例では、シングルジョブで1GBの実行に対して、SMP8並列として、使用メモリが8倍になると想定してのオプションを指定しています。※ SMP並列では、使用する総メモリ数を指定します。



ジョブスクリプトは以下となります。

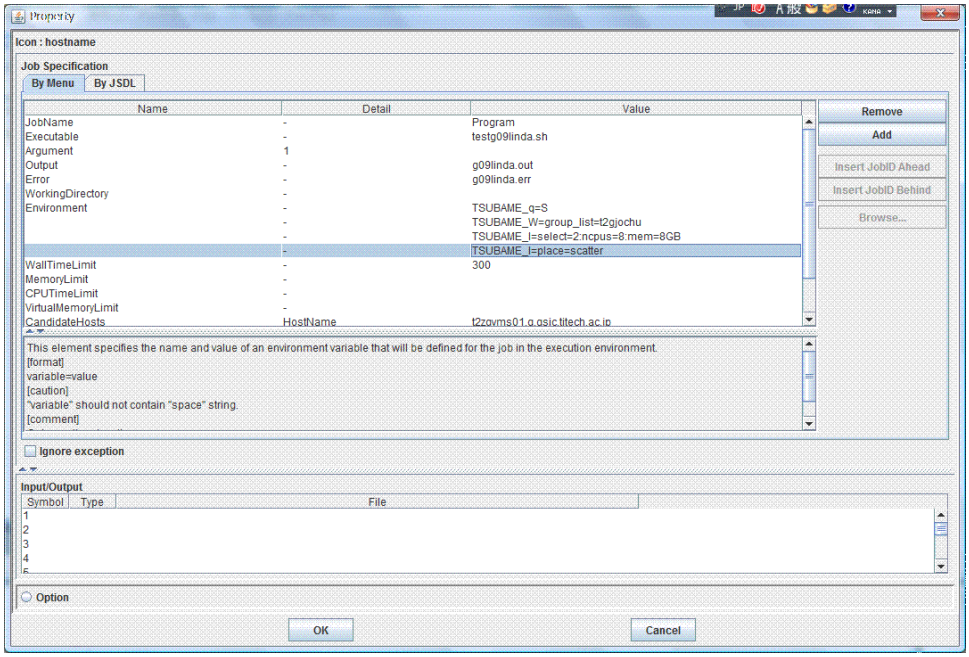
```
cat testg09smp.sh
#!/bin/sh
cd $PBS_O_WORKDIR
export GAUSS_SCRDIR=$TMPDIR
unset NCPUS
unset OMP_NUM_THREADS

g09 test.com
```

またSMP指定するために、インプットファイルの先頭に以下の行を追加してください。:

```
%NprocShared=8
```

3). Gaussian Linda並列 TSUBAME2.0でのGaussian Linda並列について Gaussianの利用手引き を参照してください。以下の例は、16並列で、ノード内8並列になります。(mpi 16:8と等価)



ジョブスクリプトは以下となります。:

```
cat testg09linda.sh
#!/bin/sh
cd $PBS_O_WORKDIR
export GAUSS_SCRDIR=$TMPDIR
unset NCPUS
unset OMP_NUM_THREADS

export g09root=/usr/apps/isv/gaussian_linda/gaussian09.B01
source $g09root/g09/bsd/g09.profile
export GAUSS_LFLAGS='-opt "Tsnet.Node.lindarsharg:ssh"'
export GAUSS_LFLAGS="$GAUSS_LFLAGS -nodelist ''`cat $PBS_NODEFILE`'' -mp 2"

g09 test.com
```

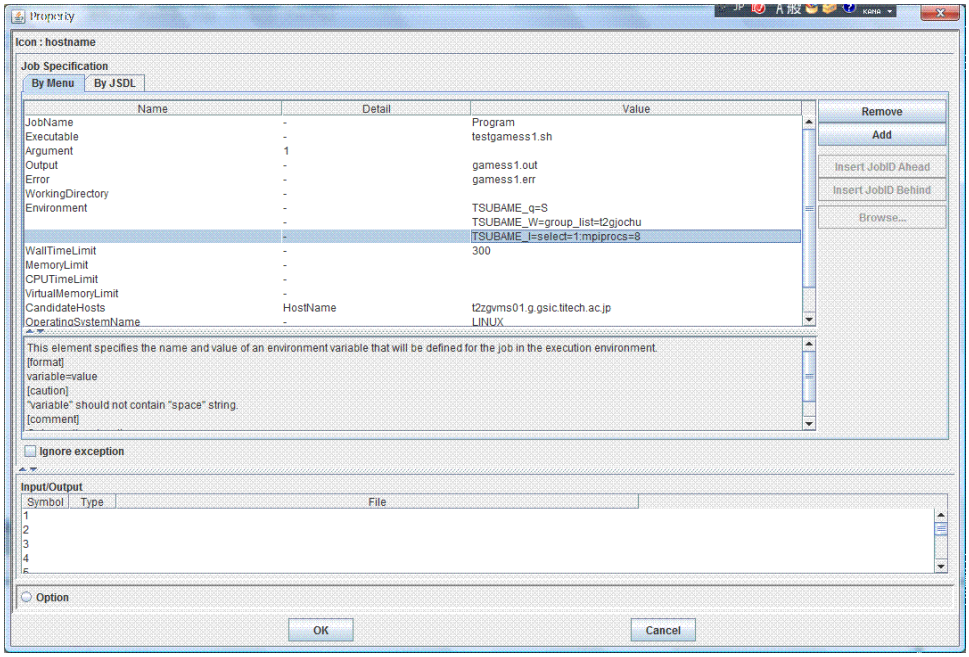
またLinda/SMP指定するために、インプットファイルの先頭に以下の行を追加してください。:

```
%NprocLinda=2
%NprocShared=8
```

5.2 Gamessジョブ投入

TSUBAME2.0のfreeアプリケーションは、GUIによる動作以外は基本的にNAREGIで利用できます。ここではGamessを例として、投入方法を説明します。

1). Gamessジョブ(ノード内並列)
ノード内8並列ジョブを投入する場合の「Program Icon」作成画面は以下となります。* シングルジョブの場合は並列の指定は必要ありません。

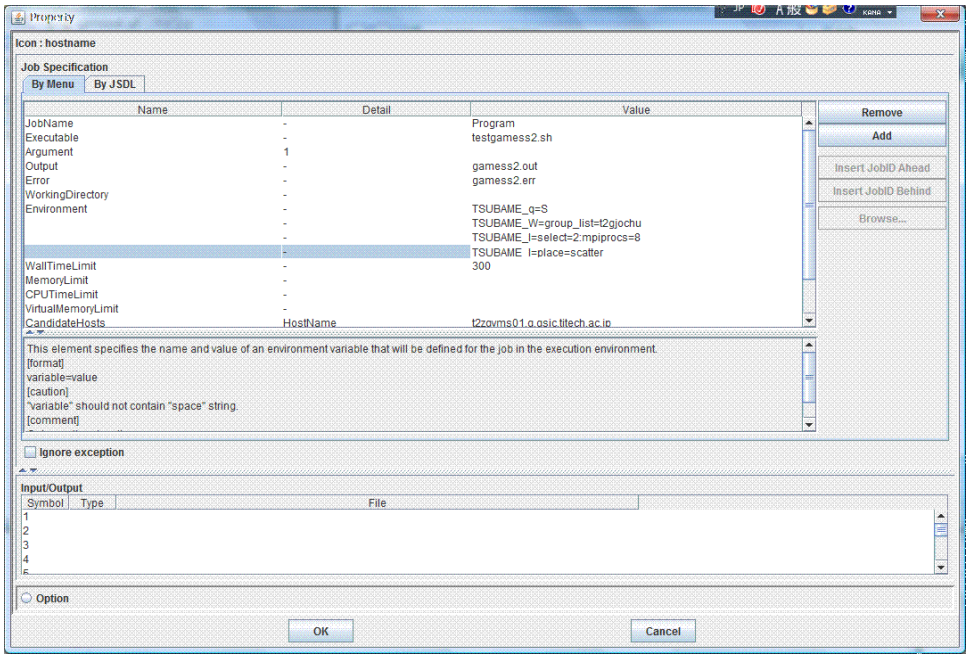


ジョブスクリプトは以下となります。:

```
cat testgamess1.sh
#!/bin/sh
cd $PBS_O_WORKDIR
export SCR=$TMPDIR

rungms test01 00 8 > test01.log
```

2). Gamessジョブ(ノード間並列)
2ノードを使用した16並列(ノードあたり8並列)ジョブを投入する場合は、
「Program Icon」作成画面は以下となります。



ジョブスクリプトは以下となります。:

```
cat testgamess2.sh
#!/bin/sh
```

```
cd $PBS_O_WORKDIR
export SCR=$TMPDIR

rungms test02 00 16 > test02.log
```