東京大学宇宙線研究所
電子計算機システム
利用講習会

2008年4月24日

日本SGI株式会社

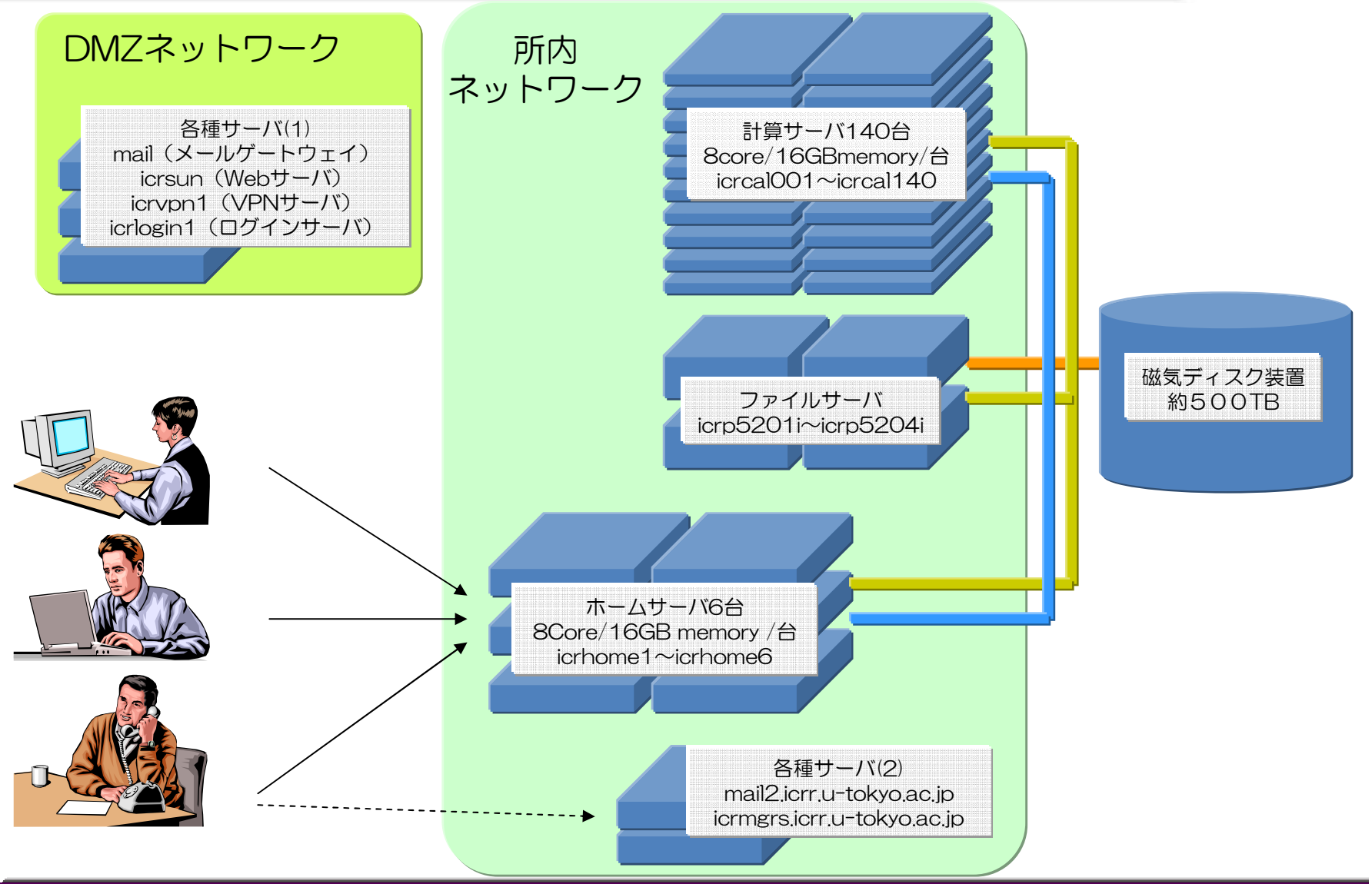
Agenda

- この講習会の目的
- システムの概要
- バッチシステムとその利用
  - ファイルシステムの共有
  - キューの構成
  - ジョブの投入
  - I/Oの多いジョブ
- その他ホームサーバで出来ること

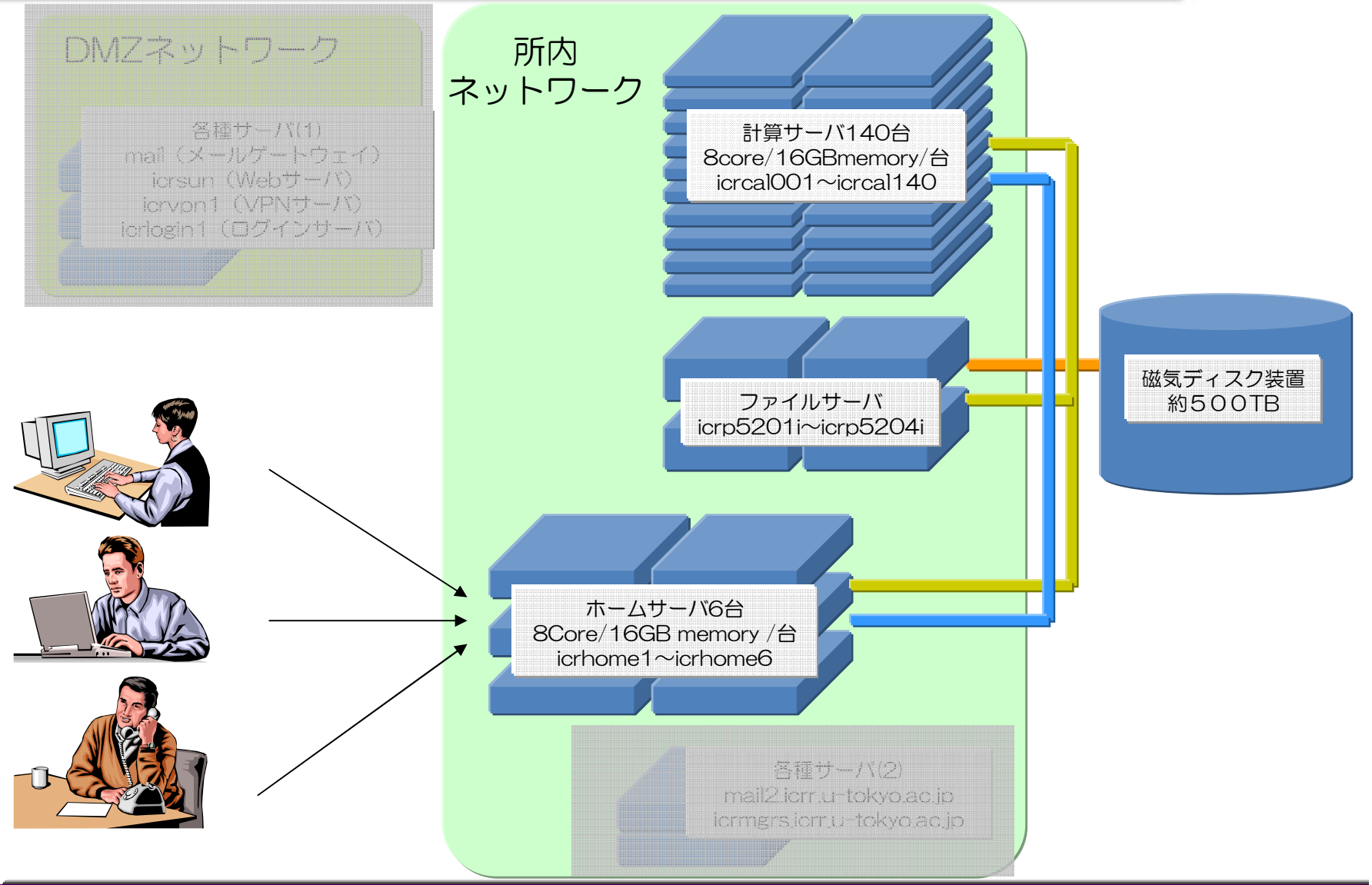
この講習会の目的

- 主に使用する機会が多いホームサーバ icrhome1～icrhome6 にて行なえることの習得
  - バッチジョブの投入
  - その他機能の紹介
    - メールの転送
    - Web コンテンツ更新
    - パスワード変更

システム構成概要


DMZネットワーク
各種サーバ(1)
mail（メールゲートウェイ）
icrsun（Webサーバ）
icrvpn1（VPNサーバ）
icrlogin1（ログインサーバ）
所内
ネットワーク
計算サーバ140台
8core/16GBmemory/台
icrcal001～icrcal140
ファイルサーバ
icrp5201i～icrp5204i
磁気ディスク装置
約５００TB
ホームサーバ6台
8Core/16GB memory /台
icrhome1～icrhome6
各種サーバ(2)
mail2.icrr.u-tokyo.ac.jp
icrmgrs.icrr.u-tokyo.ac.jp

本日の講習会で主に紹介する要素（計算サーバ・バッチジョブの投入）



バッチシステムについて

合計140台（1120コア）のシステムを全て
手動で利用する場合、各サーバの負荷状況を
各自把握し、どのサーバにて計算を実行するか
を調べる必要があります。（面倒・煩雑）

そのため、バッチソフトウェア『Moab』を利用し
合計140台（1120コア）のシステムを効率よく
使用します。

バッチシステムの利用の流れ

- バッチシステム構成
  - バッチジョブを投入するサーバ（ホームサーバ：icrhome1~icrhome6）
  - バッチジョブを実行するサーバ（計算サーバ　：icrcal001~icrcal140）
- バッチジョブスクリプトをホームサーバにて作成
  - 一般的なバッチジョブスクリプトの作成手順
- ホームサーバからバッチジョブを投入
  - 投入されたジョブの確認
  - ジョブをキャンセルするには
- ディスクアクセス（ランダムI/O）が多いジョブ

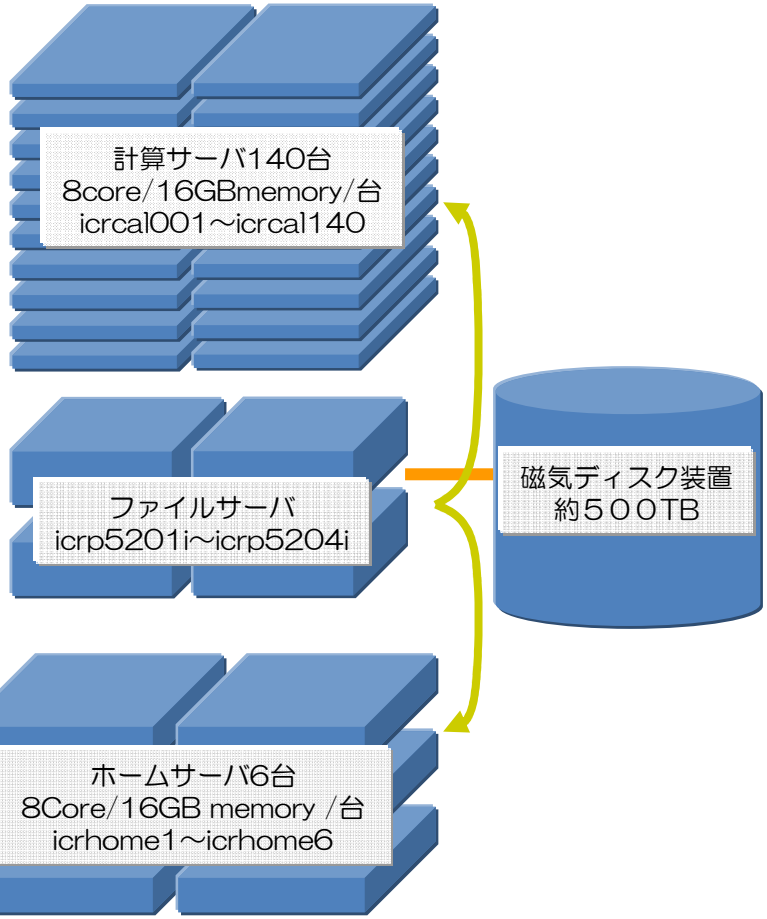
バッチシステム構成（１）

- ホームサーバはユーザの所属するグループごとに異なるサーバを使用します。
  - icrhome1：Ashraグループ
  - Icrhome2：CANGAROO グループ
  - Icrhome3：TA グループ
  - Icrhome4：Tibet グループ
  - Icrhome5：GR（重力波）、SDSS、TH（理論） グループ
  - Icrhome6：KAM（神岡）、 ICRRグループ

- ジョブスクリプトの作成・ジョブの実行の方法は共通です。
  本日はicrr2002というユーザ(ICRRグループ：icrhome6)を使用します。

バッチシステム構成（2）

- ホームサーバ・計算サーバはGPFS（General Parallel File System）を使用しディスクのデータを共有しています。
- そのためホームサーバ・計算サーバから見たデータは同一の場所に存在しているように見ることができます。



ホームサーバ(icrhome6)から

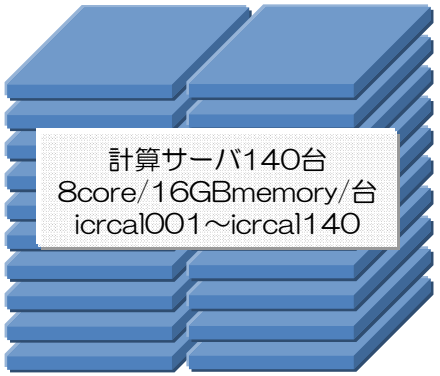
```
[sgi2002@icrhome6 sgi2002]$ pwd
/icrr/work/sgi2002
[sgi2002@icrhome6 sgi2002]$ ls -l
total 8
-rw-r--r--  1 sgi2002 icrr 157 Apr 21 13:50 sample.sh
[sgi2002@icrhome6 sgi2002]$
```

計算サーバから(icrcal001)から

```
[sgi2002@icrcal001 sgi2002]$ pwd
/icrr/work/sgi2002
[sgi2002@icrcal001 sgi2002]$ ls -l
total 8
-rw-r--r--  1 sgi2002 icrr 157 Apr 21 13:50 sample.sh
[sgi2002@icrcal001 sgi2002]$
```

バッチシステム構成（3）

- 計算を実行する計算サーバにはログインすることはできませんが Moabがバッチジョブを実行する際に自動的に最適なホストを選択し 実行します。
- バッチジョブは以下のクラスに分かれており ユーザはこのクラスへジョブを投入することで 計算サーバを利用することができます。



キュー名	ジョブ本数制限	制限時間	グループ制限	備考
A	200	1時間	無し	
B	880	24時間	無し	
C	100	7日間	無し	
ashra	10	無制限	ashraグループのみ利用可能	
can	10	無制限	canグループのみ利用可能	
gr	10	無制限	grグループのみ利用可能	
kam	10	無制限	kamグループのみ利用可能	
sdss	10	無制限	sdssグループのみ利用可能	
ta	10	無制限	taグループのみ利用可能	
th	10	無制限	thグループのみ利用可能	
tibet	10	無制限	tibetグループのみ利用可能	

一般的なバッチジョブスクリプトの作成とジョブの投入（1）

- 一般的なバッチジョブスクリプトの作成手順
  →実演による作成

- ホームサーバからバッチジョブを投入
  - ジョブのサブミット

```
[sgi2002@icrhome6 ~]$ msub ./sample.sh

1318950　※
[sgi2002@icrhome6 ~]$
```

※この値がジョブ投入時にMoabより付与されたJOBIDです。
ジョブのプロセス内部では環境変数：$PBS_JOBIDとして
この値を取得することができます。
この場合、 1318950ではなく、 1318950.icrhome1.icrr.u-tokyo.ac.jp
として表示されます（Moabのコントロールサーバがicrhome1であるため）

一般的なバッチジョブスクリプトの作成とジョブの投入（2）

- ホームサーバからバッチジョブを投入（続き）
  - 投入されたジョブの確認方法　（ジョブの投入本数の確認方法）

```
[sgi2002@icrhome6 ~]$ showq
active jobs------------------------
JOBID              USERNAME      STATE PROCS   REMAINING            STARTTIME
****active jobs     ****of **** processors in use by local jobs (18.60%)
                          ***of *** nodes active      (100.00%)
eligible jobs----------------------
JOBID              USERNAME      STATE PROCS     WCLIMIT            QUEUETIME
0 eligible jobs
blocked jobs-----------------------
JOBID              USERNAME      STATE PROCS     WCLIMIT            QUEUETIME
0 blocked jobs
Total jobs:  231
```

  - 投入されたジョブのキャンセル方法

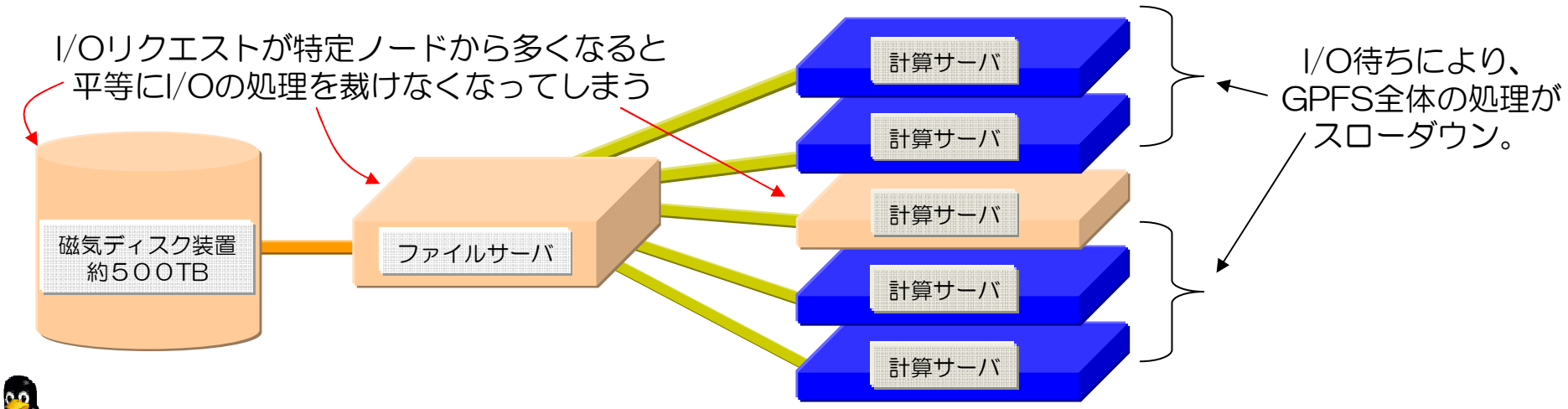
```
[sgi2002@icrhome6 ~]$ canceljob 1318950

job '1318950' cancelled

[sgi2002@icrhome6 ~]$
```

ディスクアクセスが多いジョブ（１）
～ その問題点 ～

- ディスクアクセス（ランダムI/O）が多いジョブ
  - ランダムI/Oが多いプログラムを計算サーバで実行する場合注意が必要です。

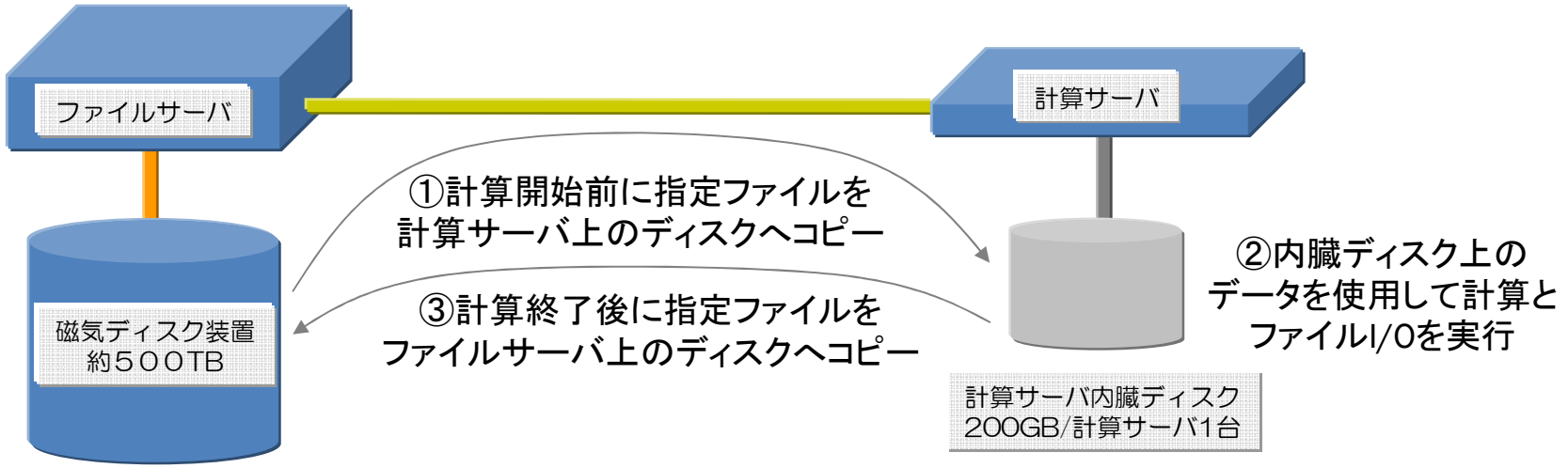
    注意が必要なプログラムの例）
    - ✓ 単一のファイルのOpen/Closeのリクエストが多い
    - ✓ 単一ファイルの部分書き換えなどが多い

上記の条件に合致するジョブは該当のデータがGPFSの共有ディスク上に存在する場合、ネットワーク経由で頻繁にデータの更新リクエストを発生します。
このようなとき、ファイルサーバの応答が遅くなる場合があることが報告されています。



ディスクアクセスが多いジョブ（２）
～ 対策：ステージイン・ステージアウト ～

- ディスクアクセス（ランダムI/O）が多い場合バッチジョブは、以下のような仕組みを利用することでGPFS 全体への負荷を軽減することができます。



※一般的には、ステージイン・ステージアウト処理といわれるものです。

ディスクアクセスが多いジョブ（３）
～ SI/SO ジョブスクリプト#1～

- 通常のスクリプトに以下の処理を加えます。
  - 通常スクリプト

```
#!/bin/sh
#PBS -j oe
#PBS -q B


cd /icrr/work/sgi2002




./a.out test_DATA > test_DATA.out
```

  - ステージイン・ステージアウト付加

```
#!/bin/sh
#PBS -j oe
#PBS -q B

mkdir /tmp/$PBS_JOBID
cd /tmp/$PBS_JOBID
cp /icrr/work/sgi2002/test_DATA .
cp /icrr/work/sgi2002/a.out .

./a.out test_DATA > test_DATA.out
cp ./test_DATA.out /icrr/work/sgi2002/.
cd /tmp/
rm -rf ./$PBS_JOBID
```

下線部がGPFS（ファイルサーバ経由）の
ディスク領域です。

ディスクアクセスが多いジョブ（４）
～ SI/SO ジョブスクリプト#2 ～

- スクリプトの解説
  - ステージイン・ステージアウト付加後のスクリプト

```
#!/bin/sh
#PBS -j oe
#PBS -q B

mkdir /tmp/$PBS_JOBID
cd /tmp/$PBS_JOBID
cp /icrr/work/sgi2002/test_DATA .
cp /icrr/work/sgi2002/a.out .

./a.out test_DATA > test_DATA.out
cp ./test_DATA.out /icrr/work/sgi2002/.
cd /tmp/
rm -rf ./$PBS_JOBID
```

作業用のディレクトリを作成します。
$PBS_JOBIDは、実行しているジョブに
付与されるユニークな名称です。
この番号のディレクトリを作成することで
他のユーザとの混同を避けます

作業用のディレクトリへ
元データをコピーします。

処理の結果ファイルをオリジナルの
ディレクトリへコピーします。

処理終了に伴い、不要になった
ディレクトリを削除します。

バッチジョブ（計算）以外の機能

バッチジョブ以外の
機能についていくつか紹介します。

ホームサーバの作業ディレクトリ

ホームサーバ icrhomeX には、用途別に３つのユーザ用領域が存在します。

- 普段の作業を行うには ・・・

  /<GROUP>/home/<USER>/

- メール関連の設定を行うには ・・・

  /<GROUP>/mailhome/<USER>/

- Web コンテンツの更新を行うには ・・・

  /<GROUP>/wwwhome/<USER>/public_html/

メールの転送 (1)
～ まずは設定ファイル置き場所 ～

ここでは、<user>@icrr.u-tokyo.ac.jp 宛のメールを別のメールアドレスに転送する方法について説明します。まずは前提から。

- 各グループのホームサーバには、メールの設定ファイル置き場として、以下のディレクトリが存在します

icrhomeX:/<GROUP>/mailhome/<USER>　(ex. icrhome6:/icrr/mailhome/sgi2002)

NFS マウント

mail2:/<GROUP>/home/<USER> （ディレクトリの実体はこちら）

- mail2 にはログインできません
- この領域は非常に小さい（各グループ 1GB 程度）ので、大きなデータを置かないで下さい

メールの転送 (2)
～ 設定ファイル .forward” ～

- **メール転送の設定方法**

  icrhomeX:/<group>/mailhome/<user> ディレクトリで .forward ファイルを作成・編集する。

- **.forward ファイルの記述フォーマット**

  .forward ファイルには、転送先メールアドレスを記述します。
  複数指定する場合は、各行１アドレスで複数行に渡って指定します。

- **記述例（以下のメールアドレスは実在のアドレスとは関係ありません）**

hogehoge@gmail.com	← *転送先メールアドレスを指定*
#nanchara@yahoo.co.jp	← *無効にするには # でコメントアウト*
~/Maildir/	← *転送しつつ、mail2 にもメールを残す場合*
sucharaka-mailing-list	← *メーリングリストも指定可能*

Web コンテンツの更新方法

- Web コンテンツは以下の場所に置くことにより、外部に公開することができます

icrhomeX:/<GROUP>/wwwhome/<USER>/public_html/

NFS マウント

icrsun:/<GROUP>/home/<USER>/public_html/ （実体はこちら）

- icrsun にはログインできません
- public_html/ に置かれたファイルは、以下の URL で参照できます

  http://www.icrr.u-tokyo.ac.jp/~<USER>/<FILE>

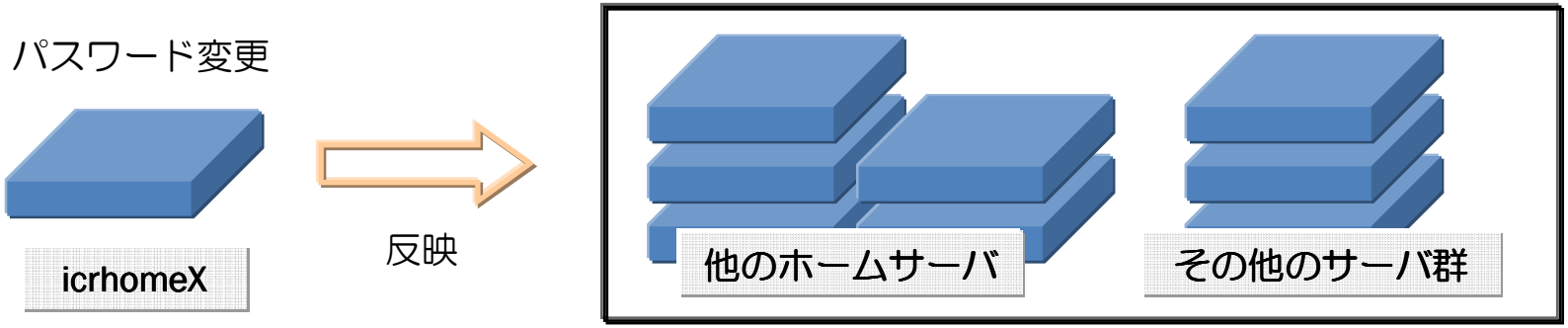
- セキュリティの観点から、上記ディレクトリでは CGI を使用できない設定になっています

パスワード変更 (1)
～ ホームサーバ編 ～

- ホームサーバ (icrhome1~6) でパスワードの変更を行うには、
  以下のようにします

```
$ chpasswd
Changing password for <USER>.
(current) UNIX password:        ← 現在のパスワードを入力してください
New password :                  ← 新しいパスワードを入力してください
Retype new password :           ← 再度新しいパスワードを入力してください
```

- ホームサーバ (icrhome1~6) を含む各種サーバ (icrvpn1, mail2, etc.)
  におけるパスワードは LDAP により一元管理されており、
  ホームサーバでのパスワード変更は、他のサーバにも反映されます



パスワード変更 (2)
～ ログインサーバ編 ～

- ログインサーバ (icrlogin1, icrlogin2) は LDAP の管理下にありませんので、<u>ホームサーバのパスワード変更とは連動しません</u>
- ログインサーバでパスワード変更を行うには、以下のようにします

```
> passwd
Changing password for <USER>.
Old Password:                 ← 現在のパスワードを入力してください
New Password:                 ← 新しいパスワードを入力してください
Reenter New Password:         ← 再度新しいパスワードを入力してください
Password changed.
```

- なお、ログインサーバ同士でもパスワード変更の連動はしませんので、<u>icrlogin1, icrlogin2 個別に設定する必要があります</u>
- ログインサーバは研究所外部から直接アクセス可能なため、<u>単純なパスワードを設定すべきではありません</u>

ユーザ向け利用の手引き

- オンラインマニュアル

  http://www.icrr.u-tokyo.ac.jp/sgi/manual/