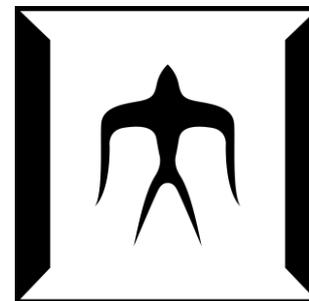
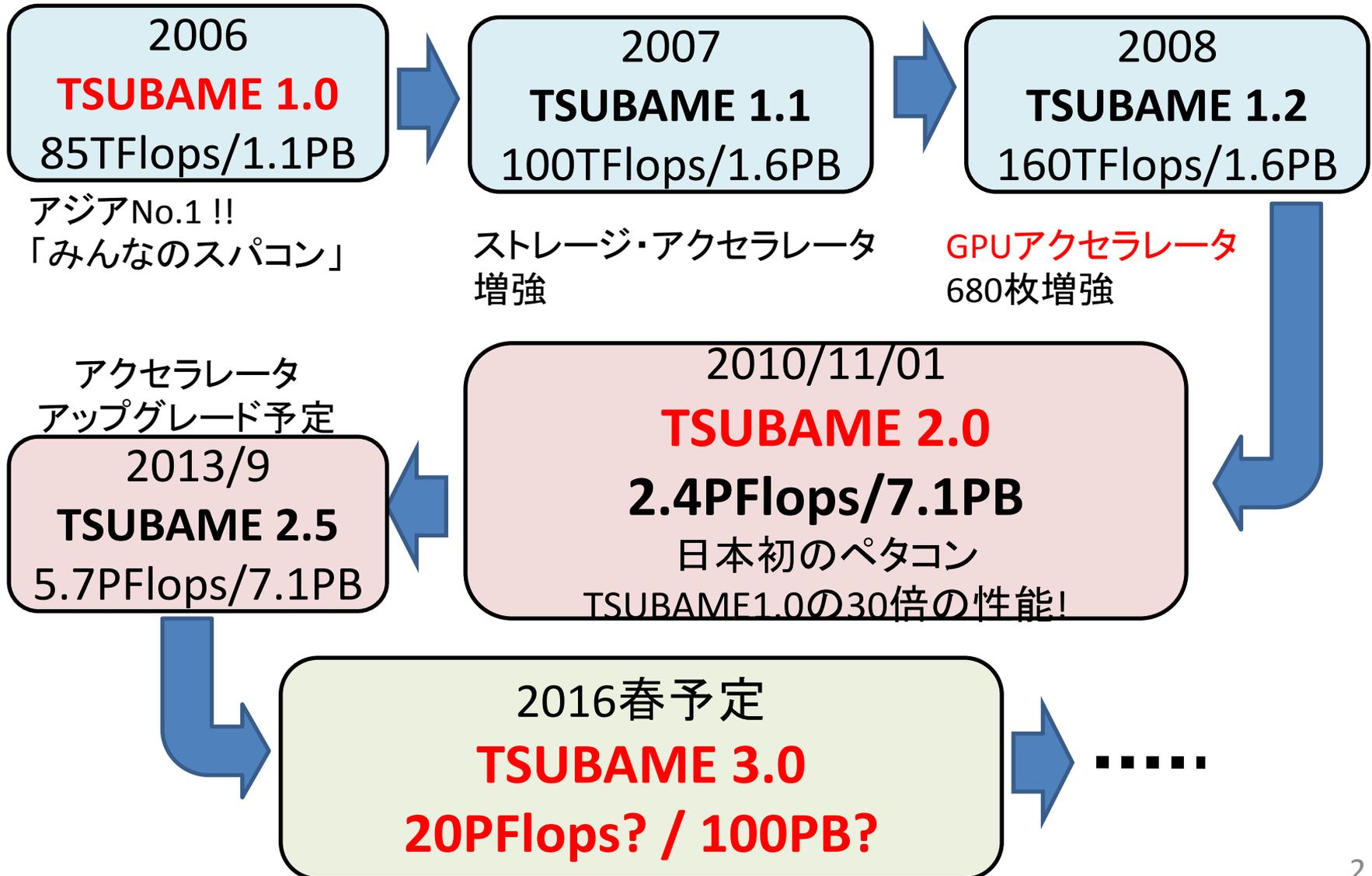


# スーパーコンピュータ TSUBAME 2.5利用ガイドンス

GSICのガイドンス資料を抜粋・改変



# TSUBAMEの歴史と今後



# TSUBAME2.5の特徴(1)

- 理論値5.7PFlopsのばく大な演算性能
  - CPU合計性能: 0.2PFlops
  - GPU合計性能: 5.5PFlops
- 合計容量7.1PByteの巨大ストレージ
- バイセクションバンド幅200Tb/sの高速光ネットワーク

# TSUBAME2.5の特徴(2)

- ソフトウェア資産の継続性と新規運用
  - 既存のMPI, OpenMP, CUDAなどで記述されたプログラムの利用
    - GPU向けにOpenACCも利用可能
  - 既存ISVアプリの大部分の利用
  - SUSE Linux Enterprise 11
  - 新たにWindows HPC Serverの運用 (今回は説明対象外)
- GPU対応アプリも採用、ぜひ使ってください
  - CPUよりも計算が短時間で済む⇒課金も少なくてすむ

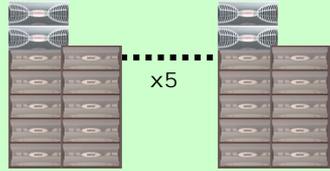


# ハードウェア構成

# スーパーコンピュータTSUBAME2.5 システム構成

ペタバイト級HDD ストレージ: Total **7.2PB** (Lustre+ home)

## 並列ファイルシステム領域 6.0PB



Storage  
DDN SFA10000 x5  
(10 enclosure x5)  
Storage Server  
HP DL360 G6/G8 x26nodes  
HP DL380 G6 x4nodes  
File Sytem  
Lustre 3.6PB, GPFS 2.4PB

Storage Servers x30

## ホーム領域 1.2PB



Storage  
DDN SFA10000 x1  
(10 enclosure x1)  
Storage Server  
HP DL380 G6 x4nodes  
BlueArc Mercury 100 x2  
File System  
NFS, CIFS, iSCSI

NFS,CIFS servers x4 NFS,CIFS,iSCSI x2

StorageTek  
SL8500  
テープシステム  
~8PB

## HPCIストレージ 0.6PB

Storage  
DDN SFA12000  
(5 enclosure)  
File Sytem  
GFarm



## ノード間相互結合網: フルバイセクション ノンブロッキング 光 QDR InfiniBand ネットワーク

### Core Switch



12switches

Voltaire Grid Director 4700 12switches  
IB QDR: 324port

### Edge Switch



179switches

Voltaire  
Grid Director 4036 179switches  
IB QDR: 36 port

### Edge Switch (10GbE port付き)



6switches

Voltaire  
Grid Director 4036E 6 switches  
IB QDR:34port  
10GbE: 2port

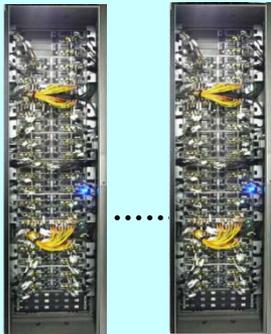
管理サーバ群

Titenet3

Sinet4

計算ノード: **5.76PFLOPS (CPU+GPU)**, **225TFLOPS (CPU)**, **~120TBメモリ**, **>200TB SSD**

## Thin計算ノード



1408nodes

30node x 42MCS racks, 他148nodes

HP Proliant SL390s G7 1408nodes  
CPU Intel Westmere-EP X5670 2.93GHz  
(Turbo boost 3.2GHz) 12Core/node  
Mem: 58GB (54GiB) x1367nodes  
103GB (96GiB) x41nodes  
GPU NVIDIA Tesla K20X 1.31TFlops,3GPU/node  
SSD 60GB x 2 120GB ※54GiBメモリ搭載node  
120GB x 2 240GB ※96GiBメモリ搭載node  
OS: SUSE Linux Enterprise / Windows HPC Server

CPU Total Speed: 216TFLOPS (w/Turbo boost)

Total Speed: 5750TFLOPS

Memory Total:83.5TB (CPU) + 27.2TB (GPU)

SSD Total: 173.9TB

## Medium計算ノード



HP DL580 G7 24nodes  
CPU Intel Nehalem-EX 2.0GHz  
32Core/node  
Mem:137GB (128GiB)  
SSD 120GB x 4 480GB  
OS: Suse Linux Enterprise Server

CPU Total Speed: 6.14TFLOPS

## Fat計算ノード



HP DL580 G7 10nodes  
CPU Intel Nehalem-EX 2.0GHz  
32Core/node  
Mem:274GB (256GiB) x8nodes  
548GB (512GiB) x2nodes  
SSD 120GB x 4 480GB  
OS: Suse Linux Enterprise Server

CPU Total Speed: 2.56TFLOPS

PCI-E gen2 x16 x2slot/node

GSIC:NVIDIA Tesla S1070GPU

# 計算ノード (1)

- Thinノード, Mediumノード, Fatノードの三種類
- Thinノード: **1408台** [一番良く使われる計算ノード]
  - HP Proliant SL390s G7
  - CPU: Intel Xeon 2.93GHz 6コア × 2 = **12コア**
    - Hyperthreadingのために24コアに見える
  - GPU: NVIDIA Tesla K20X **3GPU**
  - Memory: **54GB** (一部は96GB)
  - SSD: 120GB (一部は240GB)
  - ネットワーク: QDR InfiniBand x 2 = 80Gbps



# 計算ノード(2)

- Medium/Fatノード:M24台 + F10台  
[大容量メモリが必要なジョブ向け]
  - HP Proliant DL580 G7
  - CPU: Intel Xeon 2.0GHz 8コア × 4 = 32コア
    - Hyperthreadingのために64コアに見える
  - Memory: 128GB (Medium), 256/512GB (Fat)
  - SSD: 480GB
  - ネットワーク: QDR InfiniBand x 1 = 40Gbps



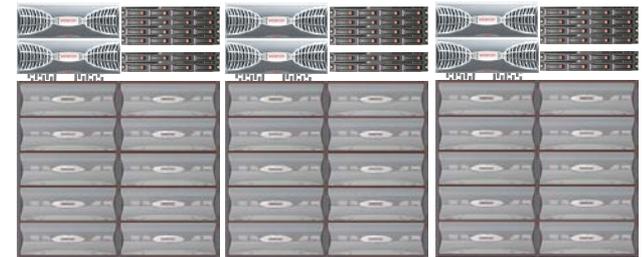
# TSUBAME2のストレージ

- ホームディレクトリ用 (/home)
  - 全ユーザが最初から利用可能
    - NFS, CIFS, iSCSI
      - BlueArc Mercury 100 (一部GridScaler)
      - DDN SFA 10K × 1, SATA × 600 disks



- 並列ファイルシステム
  - グループ購入必要

- Lustre (/work0, /work1)
  - MDS : HP DL360 G6 × 6
  - OSS : HP DL360 G6 × 20
  - DDN SFA 10K × 3, 2TB SATA × 3550 disks, 600GB SAS × 50 disks
- 他. アーカイブ向けの/data0



# 実際の利用について

- 利用開始までの流れ
- 課金とTSUBAMEグループについて

# TSUBAME2の利用開始

- 利用申請(必須)
  - 東工大ポータルにログインして、メニューからTSUBAME利用ポータルにシングルサインオン(SSO)で申請  
東工大ポータル: <http://portal.titech.ac.jp>
  - メールで仮パスワード発行、TSUBAME利用ポータルで本パスワードを設定して利用開始
  - ペーパーレスで即日利用が可能
- TSUBAME2へのログイン
  - SSHによるログイン
  - 学外からは鍵認証○、パスワード×
  - 学内からは鍵認証○、パスワード○ (場所による)

# TSUBAME利用ポータル

- 以下のサービスが利用可能なwebページ
  - アカウント新規利用申請、利用者情報の変更、利用停止  
(利用者自身)
  - TSUBAMEグループの作成、管理
  - 予算の追加、登録(予算管理者のみ)
  - Hキューの予約(グループ参加者)
  - 有償サービス利用履歴閲覧(利用者ごと、管理者)
  - 課金請求データの閲覧(予算管理者のみ)
- 入り方(1): 東工大ポータルから
- 入り方(2): <http://tsubame.gsic.titech.ac.jp/> からTSUBAME portalリンク、TSUBAMEアカウントでログイン

# TSUBAME2上で利用できるサービス

- 無償サービス
  - インタラクティブ、デバッグ専用ノードの利用
  - 小規模の計算試験(2ノード10分間まで)
  - 個人用ストレージサービス(home領域、全学ストレージ、学内ホスティング)
- 有償サービス
  - 研究目的の大規模計算(従量制、定額制)
  - Work領域, Data領域(グループ利用、月額制)
  - 追加ISVアプリケーション利用(予定)

# 有償サービス

- 研究室、研究プロジェクト単位でグループ作成 (TSUBAMEグループ)
- TSUBAMEポイントによるプリペイド従量制
  - 1ポイントで従来の1ノード・時間を利用できるポイント制
  - 1口=6000円/600ポイント
- 定額制の仮想ノード計算サービス
- グループ共有の大規模work領域サービス

# ソフトウェア構成と使い方

- システムソフトウェア・ストレージ
- バッチキューの構成と使い方
- アプリケーション

# System Software

	<b>TSUBAME 2</b>
Linux OS	SUSE Linux Enterprise Server 11 SP1
Windows OS	Windows HPC Server 2008 R2
Job Scheduler for Linux	PBS Professional
Job Scheduler for Windows	Windows HPC Server

# Compilers & Libraries

	TSUBAME 2
Compiler	Intel Compiler 2013 PGI CDK 14 gcc 4.3.4
MPI	OpenMPI 1.6.3 MVAPICH2 1.5.1
CUDA	5.5

- コンパイラの切り替えは環境変数の設定で可能
  - 利用の手引をご参照ください
- CUDA C/FortranによるGPUプログラミング可能
  - CUDA+MPIの場合はコンパイラの組み合わせについてご相談を
- バージョンアップの可能性あり

# TSUBAME2へのログイン(1)

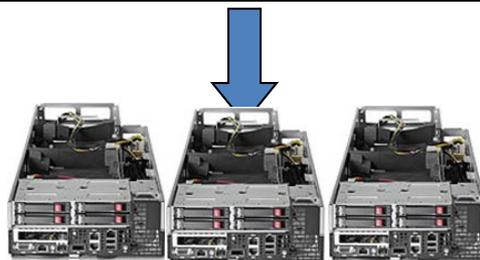


SSH

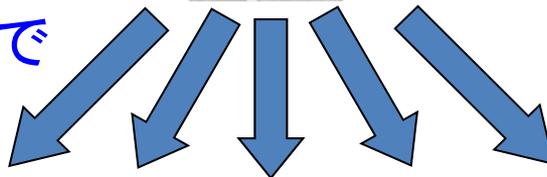


login-t2.g.gsic.titech.ac.jp

インタラクティブノードに  
自動で移動、作業可能



バッチキューシステムで  
ジョブ投入



計算ノードたち(1000ノード以上)

TSUBAME 2.5 ガイダンス

# TSUBAME2へのログイン(2)

- Linuxなどからの場合
    - ssh [アカウント名]@login-t2.g.gsic.titech.ac.jp
  - Windowsの端末ソフトからの場合(putty, ttsshなど)
    - ホスト名 : login-t2.g.gsic.titech.ac.jp
    - プロトコル : SSH
      - ポート : 22
  - ユーザ名(アカウント名)・パスワードを正しく入力
- 様々なメッセージの後に以下のように表示されればログイン成功

```
10B12345@t2a006163:>
```

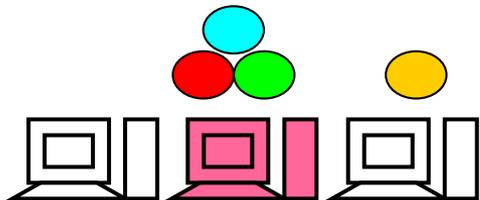
# バッチキューシステムとは

- TSUBAME2ではPBSProというバッチキューシステムでジョブ(プログラム)を投入
- 多数のプログラムの「交通整理」
  - OSはノード内、バッチキューシステムはノード間の管理

システムなし



システムあり



ユーザが自分でノード決定  
混雑すると実行が遅くなる

Batch system



システムが自動決定  
ジョブ開始が待たされることあり

# 主要キュー一覧

- インタラクティブノード
- バッチキュー
  - [S] ノード占有系: 12CPUコア、3GPUのノード利用
  - [U] 仮想マシン占有系: 8CPUコア(16hyperthread) の仮想ノード利用
  - [V] 仮想マシン共有系: 8CPUコア(16hyperthread) の仮想ノード利用
  - [G] GPU系: 4CPUコア、3GPUのノード利用
  - [H] 予約系: Thinノードをノード数、期間を予約して利用
- グランドチャレンジ(超大規模並列)制度
  - 数千～万の超大規模並列計算のための利用(要審査、年に2回+ $\alpha$ )

# ノード占有系：Sキュー・Lキュー

- Sキュー：12CPUコア, 3GPU, 54GBメモリを持つノードを利用
  - 多数CPUまたはGPUによる並列性や、I/O(ディスク・通信)性能が必要なジョブ向け
  - ノード内のジョブ混在は起こらない
  - 従量制課金
- 大容量メモリが必要なジョブには、S96, L128, L256, L512キュー
  - 数字はメモリ容量(GB)
  - Sに比べ1.5倍、2倍...の課金
  - L系はMedium/Fatノード。CPUコア数が多く、GPUが古い

# 仮想マシン系：U・Vキュー

- ノードあたり8CPUコアを利用
  - 比較的小規模なジョブ向け
  - KVM仮想マシン技術により、以下のようなノードに見える
    - 8CPUコア (hyperthreadingで16コアに見える)
    - 23GBメモリ
    - GPUは無し
  - Vキューではノード内にジョブは混在しうる(共有)
  - I/O速度は他キューより下がるので注意

# GPU系：Gキュー

- ノードあたり3GPU+4CPUコアを利用
  - GPUジョブに適している
  - 以下のようなノードに見える
    - 4CPUコア
    - 3GPU
    - 25GBメモリ
  - Vキュージョブと仮想マシン技術によりノードを共有
  - 従量制課金、Sに比べ0.5倍 (お買い得)
  - 定期的にGSICがGPU講習会開催 (ほぼ毎回満員御礼)

# 予約系：Hキュー

- 予約した期間ノードを占有して利用
  - 従来のHPCキューに相当
  - 1000CPUコアレベルの並列性が必要なジョブ向け
  - Webから日程・ノード数を予約
  - バッチキューを介さない利用も可
  - 従来よりも、柔軟な予約が可能
    - ノード数は16以上自由、期間は一日単位で最大7日

# 主要サービス比較

<b>S</b> ノード占有系 S96, L128なども	従量	約330台	並列度・I/O速度重視 MPI, GPUジョブもOK
<b>U</b> 仮想マシン占有系	定額	約220台	ノード内並列ジョブ向け CPUのみ、I/Oはやや弱め
<b>V</b> 仮想マシン内共有系	定額	約220台	比較的小規模ジョブ向け CPUのみ、I/Oはやや弱め
<b>G</b> GPU系	従量	約480台 (U,Vと共有)	GPUジョブ向け MPI, GPUジョブもOK CPUコア数とメモリが少な目
<b>H</b> 予約系	従量	約420台	大規模並列向け 1日単位1ノード単位で予約が可能に
グラウンド チャレンジ		400~1400台	超大規模ジョブ向け 審査制、年数回予定

# バッチキューの使い方

## t2subコマンドの基本

- PBS Proというバッチキューシステムを用いて計算ノードにジョブ投入します
  - myprogというプログラムを、Sキューで実行する場合
- (1) 同じディレクトリに**スクリプトファイル**を作っておく (たとえばjob.shというファイル) ⇒ `chmod 755 job.sh` などにより「実行可能」の必要

```
#!/bin/sh
cd $PBS_O_WORKDIR
./myprog
```

job.shファイル

- (2) **t2subコマンド**で投入

```
t2sub -W group_list=xxx -q S ./job.sh
```

-q xxx: キュー名を指定

-W group\_list=xxx: TSUBAMEグループ番号を指定

# バッチキューの使い方

## MPI並列ジョブの場合

(1) myprogがMPIプログラムとする。スクリプトは以下のよう:

job.shファイル

```
#!/bin/sh
cd $PBS_O_WORKDIR
mpirun -n 並列数 -hostfile $PBS_NODEFILE ./myprog
```

(2) t2subコマンドで投入

```
t2sub -q S -W group_list=xxx -l select=10:mpiprocs=12 ¥
-l place=scatter ./job.sh
```

- この場合、ノードあたり12並列 × 10ノード = **120並列**で実行

# バッチキューの使い方

## SMP並列(スレッド,OpenMP)ジョブ

(1) myprogがプログラムとする。スクリプトは以下:

```
#!/bin/sh  
cd $PBS_O_WORKDIR  
./myprog
```

job.shファイル

(2) t2subコマンドで投入

```
t2sub -W group_list=xxx -l select=1:ncpus=8 -q S ./job.sh
```

- この場合、1ノード内で、**8並列**で実行

# T2subのその他のオプション

- `-l walltime=10:00:00`

ジョブの最大実行時間。省略すると1時間

- `-l select=~~:mem=40gb`

ジョブが利用するメモリサイズ(ノードあたり)。省略すると1GB

- `-o /xxx/yyy.txt`

標準出力の出力先ファイル名

- `-e /xxx/yyy.txt`

標準エラー出力の出力先ファイル名

詳細は「`t2sub -h`」のヘルプ表示や、web上の「利用の手引」をご参照ください

# バッチキュー関係コマンド

- t2stat

ジョブの状態を確認。通常は自ジョブのみ

例) t2stat -all: 他ユーザのジョブも表示

例) t2stat V: 指定したキュー(V)の情報のみ表示

- t2del

ジョブの終了を待たずに削除

例) t2del 147.t2zpbs03

# ユーザが利用可能なストレージ構成

## Home領域

- 用途
  - 計算ノードのホームディレクトリ(NFS)
  - (学内ストレージサービス(CIFS))
  - (学内ホスティングサービス(iSCSI))
- 利用方法
  - 1ユーザあたり25GBまで無料
  - ~ユーザ名/ でアクセス可能

## Work領域

- 用途
  - 大規模データ格納
  - Linux計算ノードからアクセス可能 (Lustre)
  - グループ単位で利用可能
- 利用方法
  - TSUBAMEグループ単位で要申請。TB×月で課金
  - /work1, /work0

テープライブラリと連携した階層型ファイルシステム(GPFSによる/data0)もあり

# Work領域の利用方法

- [グループ管理者] TSUBAMEグループを登録、ディスクオプションを有効に ⇒ /work1または/work0以下にグループ名のディレクトリが生成
- [各ユーザ] 生成されたディレクトリ内に自分の作業ディレクトリを作成する

例: /work1/t2g-group-name/USER01, USER02, ....

# 利用上の注意

- インタラクティブノードでは長時間CPUを独占するプロセスを走らせない  
ください (数分が目安)
  - エディタ、コンパイラ、可視化ツール等はok
- 大量にディスクI/Oを行う場合は/homeではなく/work1, /work0を利用し  
てください
- **アカウントの貸し借り禁止**



**積極的にTSUBAME2を  
使ってみましょう**

TSUBAME 2.5 ガイダンス

# TSUBAME2の情報入手

## TSUBAME2 WWWサイト

<http://tsubame.gsic.titech.ac.jp/>

特に大事なものは、メニュー⇒利用について⇒各種利用の手引き⇒TSUBAME2.5利用の手引き

Top⇒「Current Status」で、今の混雑具合やシステム利用電力を閲覧

## TSUBAME2についての問い合わせ先

[soudan@o.cc.titech.ac.jp](mailto:soudan@o.cc.titech.ac.jp)

数人のGSICメンバー+常駐員が数千人のユーザからの質問を受け付けています。FAQも一度読んでから！