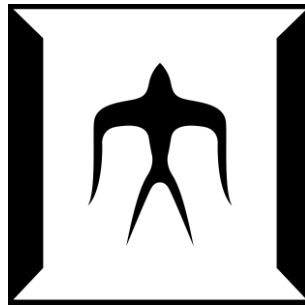
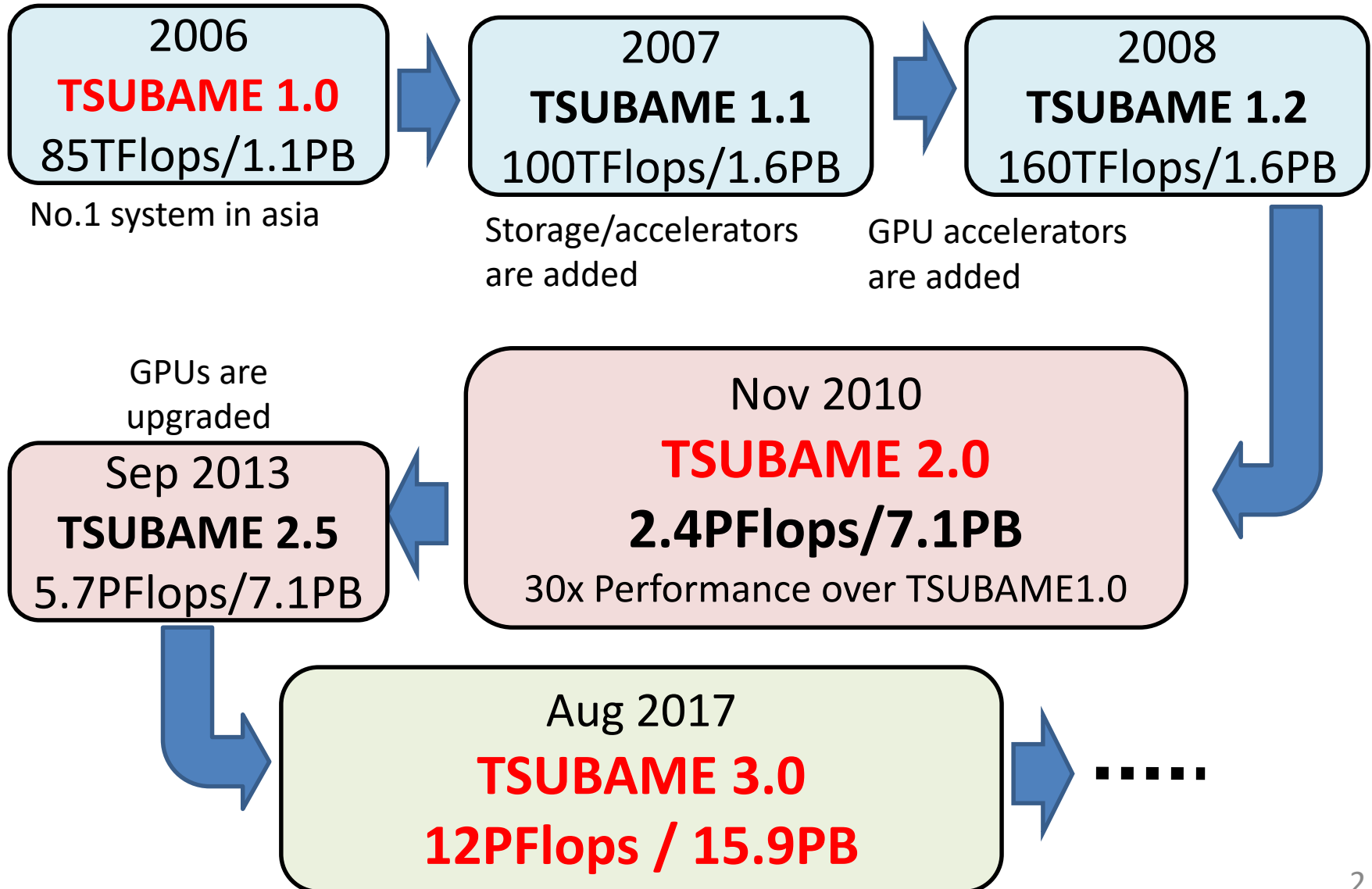


TSUBAME2.5 Supercomputer Short Guidance

This is a simplified version of
material by GSIC



History of TSUBAME



Information on TSUBAME

- <http://tsubame.gsic.titech.ac.jp>
 - Manual
 - FAQ
 - Statistics (node usage etc.)
 - Maintenance schedule
- TSUBAME query ML:
 - soudan@cc.titech.ac.jp



The screenshot shows the homepage of the TSUBAME calculation service. The browser address bar displays tsubame.gsic.titech.ac.jp. The page header features the GSIC logo (Global Scientific Information and Computing Center) and the title "TSUBAME計算サービス" (TSUBAME Calculation Service) with the subtitle "High-Performance Computing at Tokyo Tech". Below the header is a navigation bar with links: Top, 初めての方へ (For first-time users), 利用について (About usage), TSUBAME2システム構成 (TSUBAME2 system configuration), and お問い合わせ (Contact). A left sidebar menu lists various links, including Top, お知らせ (Notice), FAQ, 利用について, TSUBAME2計算サービス種別 (TSUBAME2 calculation service types), アカウント取得方法 (Account acquisition method), 有料サービス利用方法 (Paid service usage method), ログイン方法 (Login method), TSUBAMEポータル (TSUBAME portal), 各種利用の手引き (Various usage guides), TSUBAME1利用者の方へ (For TSUBAME1 users), TSUBAME2システム構成 (TSUBAME2 system configuration), TSUBAME2ハードウェア構成 (TSUBAME2 hardware configuration), and TSUBAME2ソフト (TSUBAME2 software). The main content area is divided into two sections: "重要なお知らせ" (Important notices) and "最新のお知らせ" (Latest notices). The "重要なお知らせ" section contains three items: a notice from 2011.10.02 about the current TSUBAME operation status (Grand Challenge implementation), a notice from 2011.08.15 about the suspension of services during the 2011 autumn Grand Challenge implementation, and a notice from 2011.07.22 about group disk quota usage. The "最新のお知らせ" section contains several items, including notices from 2011.9.27, 2011.9.22, 2011.9.13, 2011.9.12, 2011.9.11, 2011.9.06, and 2011.9.02, all reporting system outages or maintenance.

TSUBAME計算サービス
High-Performance Computing at Tokyo Tech

Top 初めての方へ 利用について TSUBAME2システム構成 お問い合わせ

メニュー

- ▶ Top
- ▶ お知らせ
- ▶ FAQ
- ▼ 利用について
 - ▶ TSUBAME2計算サービス種別
 - ▶ アカウント取得方法
 - ▶ 有料サービス利用方法
 - ▶ ログイン方法
 - ▶ TSUBAMEポータル
 - ▶ 各種利用の手引き
 - ▶ TSUBAME1利用者の方へ
- ▼ TSUBAME2システム構成
 - ▶ TSUBAME2ハードウェア構成
 - ▶ TSUBAME2ソフト

重要なお知らせ

- ▶ 2011.10.02
現在のTSUBAMEの運用状況はこちら(グランドチャレンジ実施中)
- ▶ 2011.08.15
平成23年度秋期グランドチャレンジ制度実施に伴うサービス休止のお知らせ(更新)
- ▶ 2011.07.22
グループディスク課金本運用開始について
- ▶ 2011.06.27
7月以降の夏季のTSUBAME運用・停止について(9/29更新)

最新のお知らせ

- ▶ 2011.9.27 : 【障害報告】2011.9.16発生: /work0ストレージ障害
- ▶ 2011.9.22 : Hキュー予約システムメンテナンスのお知らせ(9/26実施)
- ▶ 2011.9.22 : Yキューの終了について
- ▶ 2011.9.13 : 【障害報告】2011.9.12発生: ジョブスケジューラ障害
- ▶ 2011.9.13 : 【障害報告】2011.9.9発生: ジョブスケジューラ障害
- ▶ 2011.9.06 : Materials Studio 5.5.3, Discovery Studio 3.1公開、学内配布について
- ▶ 2011.9.02 : 【障害報告】2011.9.2発生: ジョブスケジューラ障害

Features of TSUBAME2.5 (1)

- High computation performance of 5.7PFlops
 - Total CPUs: 0.2PFlops
 - **Total GPUs: 5.5PFlops**
- Large storage capacity of 7.1PByte
- High speed network
 - 80Gbps per node
 - 200Tbps bisection bandwidth

Features of TSUBAME2.5 (2)

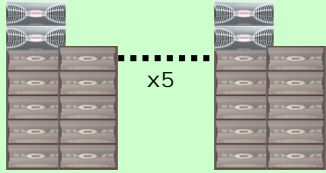
- Easy to use with commodity technologies
 - Intel CPUs + Linux OS
 - Commodity software work on TSUBAME
 - Lots of parallel programming environments
 - MPI, OpenMP, CUDA, OpenACC...



System Overview of The TSUBAME2.5 Supercomputer

Petascale HDD Storage: Total **7.2PB** (Lustre+ home)

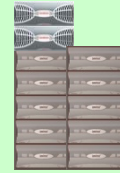
Parallel FS Partition 6.0PB



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

Home Partition 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
Tape System
~8PB

HPCI storage 0.6PB

Storage
DDN SFA12000
(5 enclosure)
File Sytem
GFarm



Node Interconnect: **Optical, Full Bisection, Non-Blocking, Dual-Rail QDR InfiniBand**

Core Switch



Voltaire Grid Director 4700 12switches
IB QDR: 324port

Edge Switch



Voltaire
Grid Director 4036 179switches
IB QDR : 36 port

Edge Switch (with 10GbE port)



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

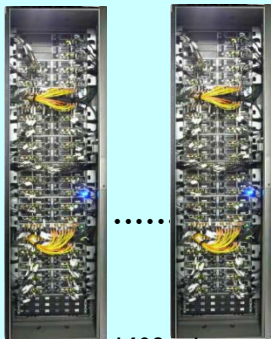
Mgmt Servers

Titenet3

SINET5

Compute Nodes: **5.76PFLOPS** (CPU+GPU), **225TFLOPS** (CPU), **~120TB** memory, **>200TB** SSD

Thin Nodes



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 node
120GB x 2 240GB *96GiB 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 Nodes



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 Nodes



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

30node x 42MCS racks, others 148nodes

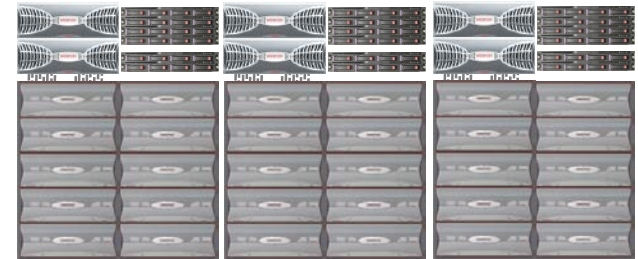
Compute Nodes

- There are (mainly) 1408 compute nodes
 - HP Proliant SL390s G7
 - CPU: Intel Xeon 2.93GHz 6 cores $\times 2 = 12$ cores
 - With Hyperthreading, 24 hyperthreads
 - GPU: NVIDIA Tesla K20X 3GPU
 - Memory: 54GB (some has 96GB)
 - SSD: 120GB (some has 240GB)
 - Network: QDR InfiniBand x 2 = 80Gbps



TSUBAME2.5 Shared Storage

- Home directory (/home)
 - Every user can use up to 25GB
- Parallel filesystem (/work0, /work1)
 - “TSUBAME group” need to buy use rights



First Access to TSUBAME2.5

- Application for new account
 - Login to Tokyo Tech Portal → “TSUBAME Portal” → Application
Tokyo Tech Portal : <http://portal.titech.ac.jp>
 - Temporary password will be informed by e-mail, and then change it to real password in TSUBAME portal
- Login to TSUBAME
 - Connect to “login-t2.g.gsic.titech.ac.jp” by SSH protocol
 - From inside of campus, public key and password are OK
 - From outside of campus, public key is OK, password is NG

System Software

OS	SUSE Linux Enterprise Server 11 SP3
Job Scheduler	PBS Professional

Compiler	Intel Compiler, PGI CDK, gcc 4.3.4
MPI	OpenMPI, MVAPICH2
GPGPU	CUDA

- You can switch compilers by environment variables
 - See manuals in TSUBAME web
- Software versions may be upgraded in maintenance

Illustration of TSUBAME Usage

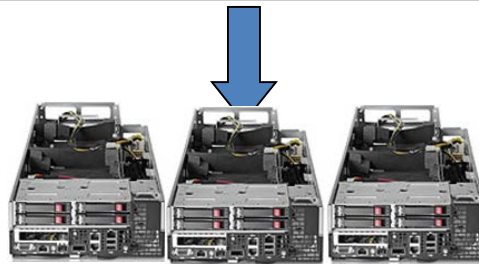


(1) SSH

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

(2) You are connected to one of “interactive” nodes. You can edit/compile your programs

(3) Throw jobs via the job scheduler



>1000 Compute nodes

TSUBAME 2.5 Guidance

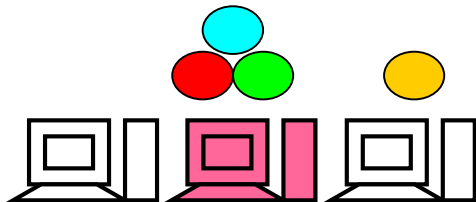
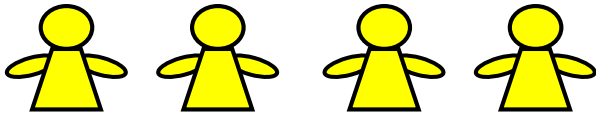
Login to TSUBAME2

- From Linux, Mac
 - `ssh login-t2.g.gsic.titech.ac.jp`
 - From terminals in Windows (putty, ttssh, etc.)
 - Host: `login-t2.g.gsic.titech.ac.jp`
 - Protocol: SSH
 - Port: 22
 - Input your account name and password
- If you see a prompt like below, your login is successful
- `10B12345@t2a006163:>`

What is Job Scheduler?

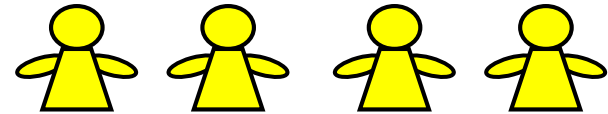
- You have to use the job scheduler (PBSPro on TSUBAME2), when you execute programs for a long time (such as >10minutes)
- The job scheduler does “traffic control” of many programs by many users

Without scheduler

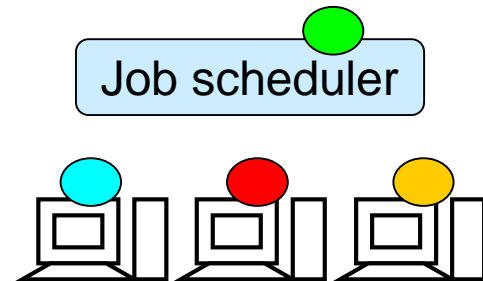


If users execute programs without control, there will be congestions

With scheduler



Job scheduler



Scheduler determines nodes for each job.
Some program executions may be “queued”

Using Job Scheduler

Basics of t2sub command

- You are going to execute your program named “myprog”
- (1) Please make a script file on the same directory (In this example, the script name is “job.sh”)

Contents of
job.sh file

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

- (2) Make the script “executable” by `chmod u+x job.sh` once
- (3) Throw a new job with **t2sub command**

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

- q xxx: Specify queue name (refer to manuals)
- W group_list=xxx: TSUBAME group name (consult to your instructor)

Using Job Scheduler

Basics of t2sub command (cont'd)

(4) You will see the following messages (If not, something is wrong)

Checking accounting informations...

:

Submitting a job to PBS...

12345.t2zpbs01

This is ID
of your jobs

(5) Wait until job execution is finished

(6) You will see new files, whose names are look like

– OTHERS.o12345

– OTHERS.e12345

} Depend on the job ID

The output by myprog has been stored in the files

Using Job Scheduler for Thread (OpenMP) Programs

(1) You are going to execute a program, myprog

job.sh file

```
#!/bin/sh  
cd $PBS_O_WORKDIR  
export OMP_NUM_THREADS=8  
./myprog
```

(3) Throw a job with t2sub

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

- myprog will be executed with 8 threads on a single node

Using Job Scheduler for MPI Programs

(1) Here you are going to execute a MPI program, myprog
job.sh file

```
#!/bin/sh  
cd $PBS_O_WORKDIR  
mpirun -n [number-of-processes] -hostfile $PBS_NODEFILE ./myprog
```

(3) Throw a job with t2sub

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

- In this case, 10 nodes are allocated, and 12 processes per node are invoked. Number-of-process is $10 \times 12 = 120$

Other Options of t2sub

- -l walltime=10:00:00

Maximum execution time of the job. Default is 1hour

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

Maximum memory size used by the job (per node). Default is 1GB

- -o /xxx/yyy.txt

Output file name for “stdout”

- -e /xxx/yyy.txt

Output file name for “stderr”

For more detail, “t2sub -h” and manuals on web will be helpful

Other Commands related to Job Scheduler

- `t2stat`

See status of jobs. In default, your own jobs are listed

cf) `t2stat -all`: See jobs by all users

- `t2del`

Delete your job before it finishes

cf) `t2del 12345.t2zpbs01`

Rules in TSUBAME Usage

- On interactive nodes, do not execute long running (and CPU centric) programs
 - You can execute editors (vi, emacs, etc.) or visualization tools longer than 10 minutes, since they do not consume CPUs much
 - With long CPU centric programs, you should use job scheduler
- Use the account of yourself. Do NOT lend/borrow accounts.

