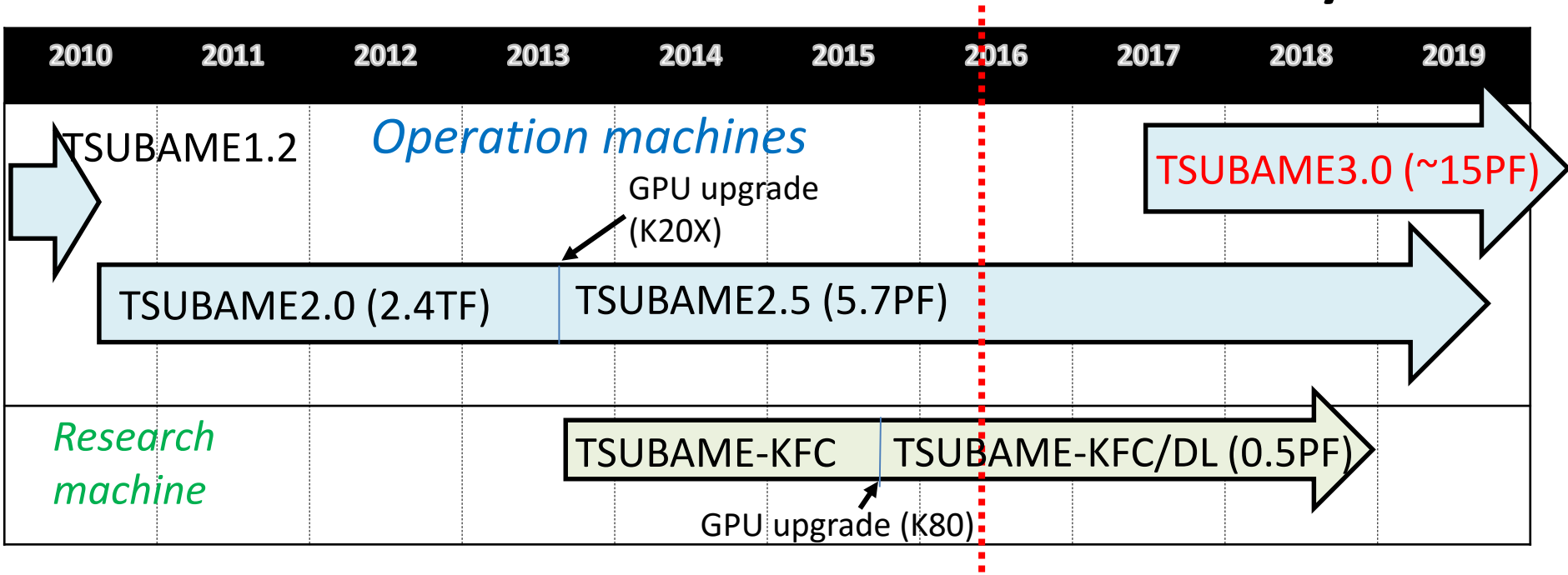
Operation Experience with SSDs and GPUs

Toshio Endo

GSIC, Tokyo Institute of Technology

Resource Management breakout, ADAC workshop

Overview of TSUBAME Family



- TSUBAME2.5
  - Current main system
- TSUBAME-KFC (/DL)
  - For low-power/big-data research, No.1 in Green 500 (Nov 13, Jul 14)
- TSUBAME3.0
  - Operation will be started on Q3 2017

TSUBAME2.5 Supercomputer (NEC/HP)

Compute nodes:

- 1,408 nodes/ 4224 K20X GPUs

→5.7PFlops (DP), 1.1PB/s mem BW

Storage:

- 7.2PB (raw) HDD, 4PB Tape
- Lustre, GPFS, NFS

Network:

- 4x QDR InifiniBand (40Gbps per link)
- Fat tree topology with ~200TB/s Bisection BW

HP ProLiant SL390s G7

CPU:	Intel Xeon X5670 (Westmere-EP, 2.93GHz) ×2 sockets six-core per socket, total 12 cores per node.
GPU:	NVIDIA Tesla K20X (GK110) ×3. 1.31TFLOPS and VRAM 6GB per GPU
Memory:	58GB DDR3 1333MHz, partly 103GB
SSD:	120GB(60GB×2), partly 240GB(120GB×2)
Network:	4X QDR InfiniBand ×2

Power consumption including cooling

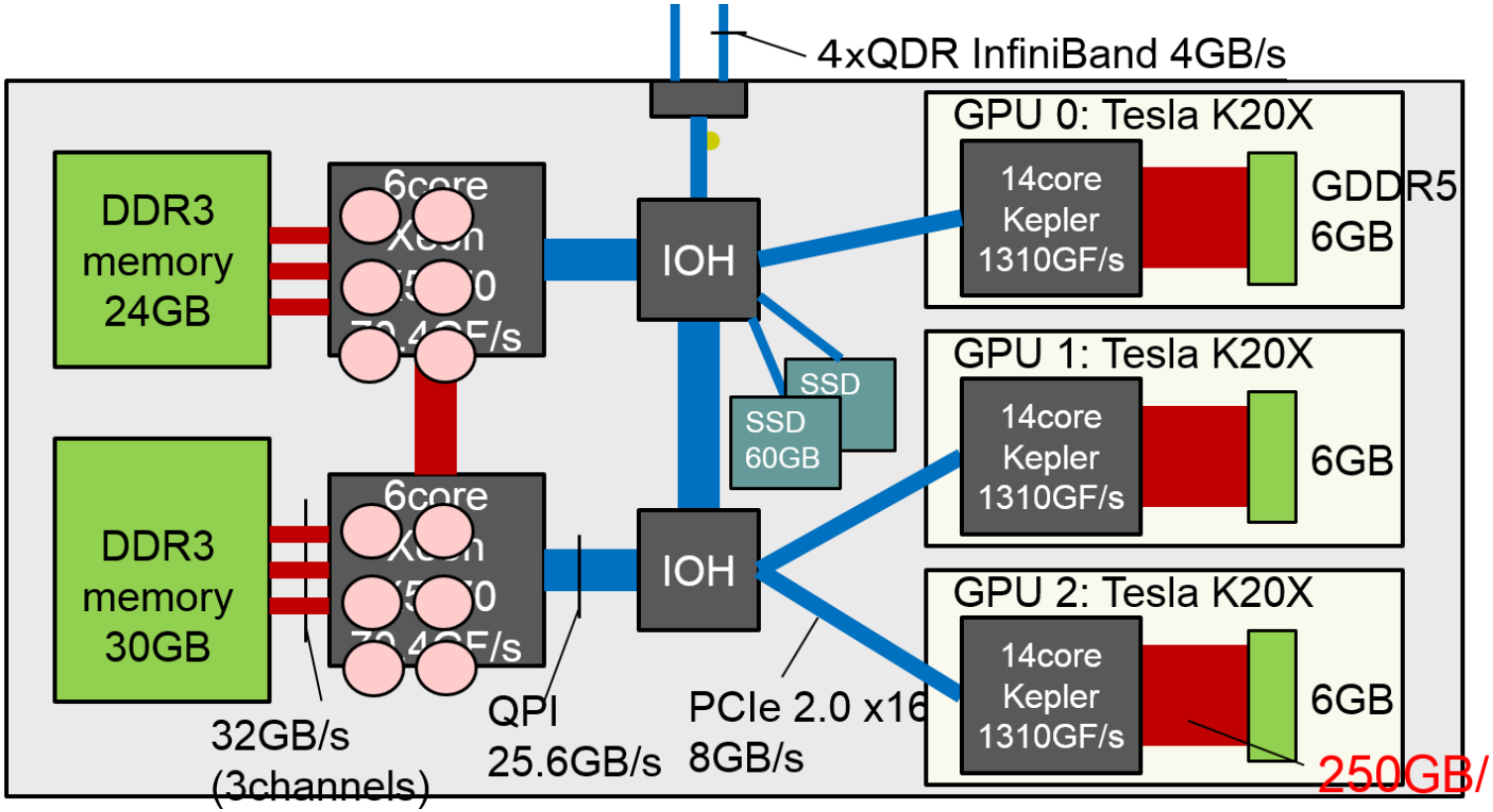
→ Average: ~1.0MW, Peak: ~1.8MW

Cooling: Chiller + HP MCS Racks

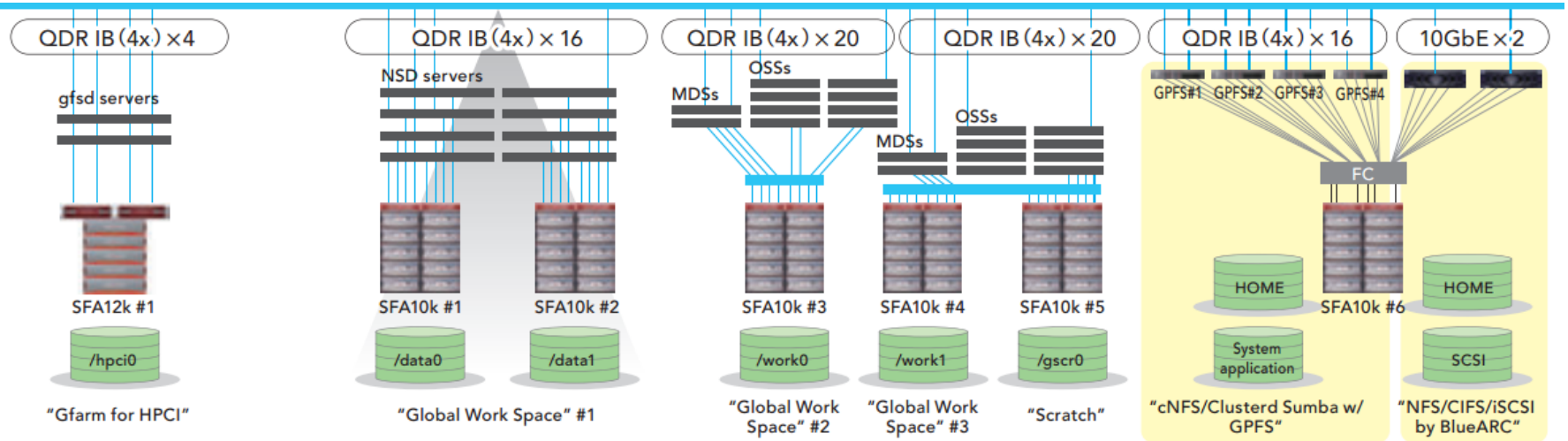
Footprint: 180m^2

Node Architecture of TSUBAME2.5

- 2x Intel 6C Xeon CPUs (Westmere) (2010)
- 3x Tesla K20X GPUs (with 6GB mem) (2013)
- 2x HP SATA 60GB SSDs (2010)
- 2x 4xQDR InfiniBand (2010)



Shared Storage of TSUBAME2.5



HPCI

Gfarm: ~ 600TB

gsfd server:
HP ProLiant DL360p Gen8 × 2
Intel XeonE5 2640 × 2,
64 GB Mem,
QDR IB (4x) × 2

Storage:
DDN SFA12k × 1
4TB SAS HDD × 155 disks

GPFS with HSM

GPFS 2.4 PB

NSD server: HP ProLiant DL380 G6 × 4
Intel Westmere EP × 2,
48GB Mem, QDR IB (4x) × 2

HP ProLiant DL360 G6 × 4
Intel Westmere EP × 2,
24GB Mem,
QDR IB (4x) × 2

Storage: DDN SFA10k × 2
2TB SATA HDD × 1180 disks
600GB SAS HDD × 20 disks

Lustre

Lustre 3.6 PB

MDS : HP ProLiant DL360 G6 × 4
Intel Westmere-EP × 2,
48GB Mem, QDR IB (4x) × 2

OSS : HP ProLiant DL360 G6 × 16
Intel Westmere-EP × 2,
24GB Mem,
QDR IB (4x) × 2

Storage: DDN SFA 10k × 3,
2TB SATA HDD × 1770 disks,
600GB SAS HDD × 30 disks

Home

Home 1.2 PB

cNFS (GridScaler)/Clustered Sumba w/ GPFS:
HP ProLiant DL380 G6 × 4
Intel Westmere EP × 2,
48GB Mem, QDR IB (4x) × 2

NFS/CIFS/iSCSI:
BlueArc Mercury 100 × 2
10Gbps × 2

Storage: DDN SFA 10k × 1
2TB SATA HDD × 600 disks

Considerations in Operation (2010~)

- How to model users' jobs?
  - How do they use CPUs and GPUs?
  - How do (should) they use SSDs and Shared FS?
- Are GPUs and SSDs robust for SC operation?
  - SSD's lifetime
  - GPU's lifetime and/or double bit ECC error
- Are there new usage methods of GPUs and SSDs?

TOPIC 1:
HOW TO MODEL USERS' JOBS?

Modelling Usage of Shared FS/Local SSDs

- Large scale data are stored in O(PB) shared FS
- Compared with shared FS, SSDs are "fast, but small"
  - Also files on SSDs created by a job should not occupy capacity in long time

→ We have configured so that SSDs are used for:

(1) Linux OS installation

(2) Temporary files

- If they want, users can either
  - Explicitly create files or copy from/to shared FS
  - Staging in/out by batch queue system
- Lifetime of files is job running time

Modelling Usage of GPUs/CPUs

- (3GPUs + 2CPUs)/node: → "Ideal" jobs would use
  - CUDA for intra-GPU parallelism
  - OpenMP for intra-node multicore CPU and multi-GPU parallelism
  - MPI for inter-node parallelism

  → ***Just too hard!***
- 3 Types of actual workloads are assumed
  - CPU only – Intra-node and internode parallelism (OpenMP and/or MPI)
  - Intra-node multiple GPUs – strong scaling & ensembles (OpenACC/CUDA + OpenMP)
  - Inter-node many GPUs – MPI weak scaling (OpenACC/CUDA + MPI, Single GPU per MPI Rank)

How to support them all in a single system?

=> Introduced node partition (partly)

Node Partition Policies

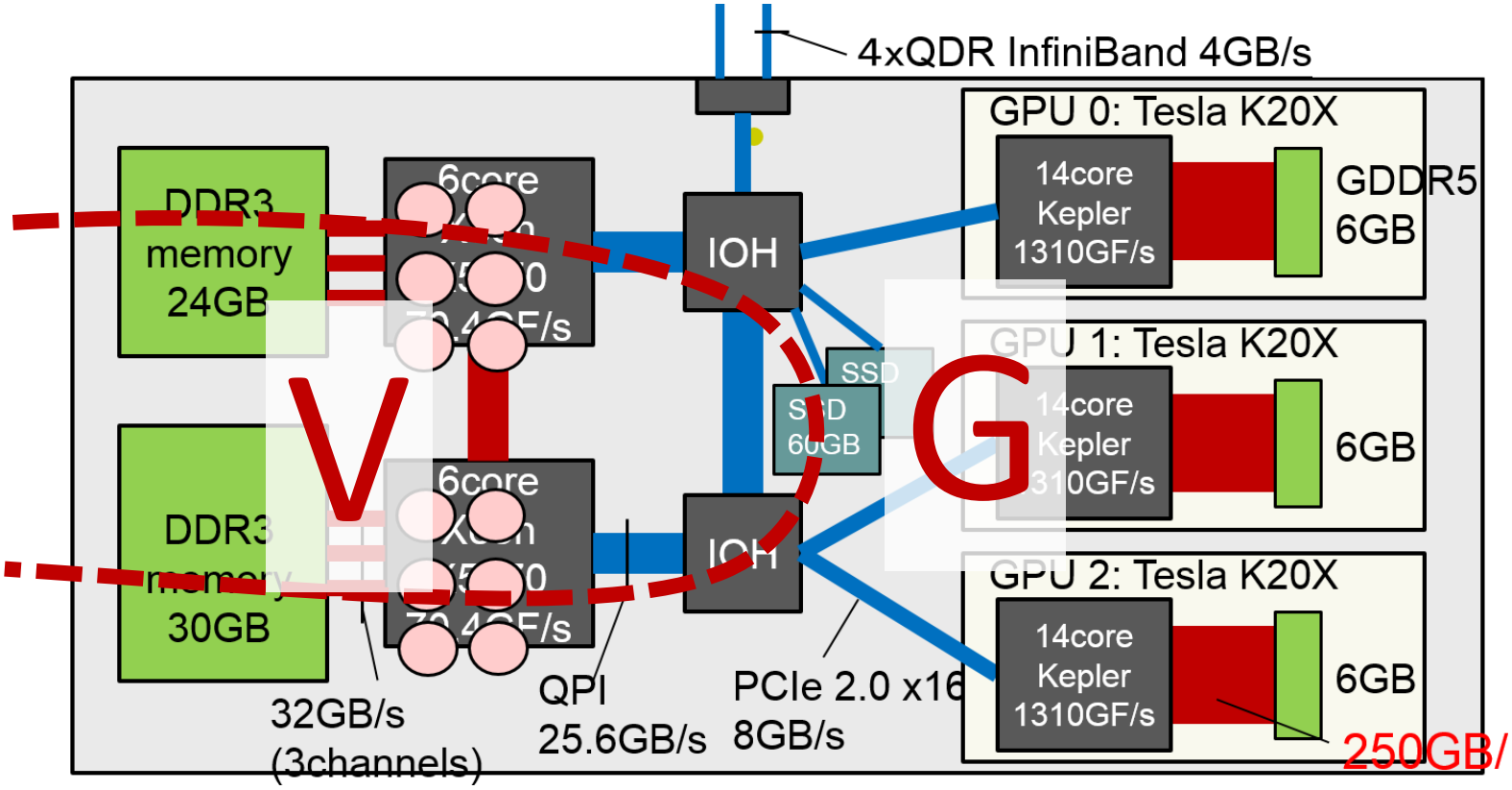
- Node Group 1: Per node allocation (~800nodes)
  →Mainly for CPU+GPU hybrid jobs
  – 400 nodes are for smaller jobs (S queue)
  – Other 400 nodes are for large jobs with reservation (H queue)
- Node Group 2: VM (KVM) based node partition (~480nodes)
  – A node is divided into “G partition” and “V partition”
  – “G partition” for GPU-centric jobs
  – “V partition” for CPU-centric jobs

VM Based Node Partition in TSUBAME2.5

Node = 12CPU cores + 54GiB mem + 3GPUs + 120GB SSD

V partition= 8 cores + 23GiB mem + 30GB SSD (OS)

G partition= 4 cores + 25GiB mem + 3GPUs + 90GB SSD



Current Issue and Future Plan

- Current: VM-based G/V node partition
  - Partitions are fixed → wide variety of job types will decrease resource utilization
  - V partition on top of KVM suffers from performance overhead, and limitation in access to GPUs, Infiniband and shared FS
- TSUBAME3 plan:
  - For more flexibility and performance, dynamic deployment configuration is required
  - Docker/Shifter containers are good candidates

TOPIC 2:
ARE GPUS AND SSDS ROBUST FOR SC OPERATION?

Consideration before TSUBAME2.0 Installation

- Robustness of ~2800 SSDs
  - Lifetime
  - Is wear leveling (as of 2010) ok?
- Robustness of ~4200 GPUs
  - Lifetime
  - Effects of high temperature
    - TSUBAME2 is dense: up to 30kW per rack
  - Memory error
    - We had used S1070 GPUs without ECC on TSUBAME1.2 (2008-2010) and observed some data corruption
    - M2050 on TSUBAME2.0 (2010-2013) have ECC, but are they ok?

Recording Faults

- We have recorded faults/events continuously
  - Not only for GPU/SSD, but for main board, NW, FS…
  - List is available on Web

TSUBAME2.5 障害履歴[Failure History of TSUBAME2.5]

csv形式で出力

1.障害発生

Issues:

- No statistical information
- Events are sometimes categorized badly
- Only in Japanese ☹

Mainly due to that lists are maintained by human (engineers)

ホスト/機器								ゴリ
t2a002100								
t2a002039 t2a003062 t2a003103								
t2a005016								
t2a005126								
t2a006175								
t2a000143								
t2a005107								
t2a000063								temBoard
t2a005025 t2a006134 t2a006155								erHW
t2b010120								erHW
t2zmaregi02		17:00						erHW
t2zpbs06	管理	2016/06/05 09:00-17:00	2016/06/06 15:00	Battery障害	-	CacheBattery交換	該当ノード	OtherHW
t2s007033 t2s007043 t2s007045 t2s007049		2016/06/05 09:00-17:00	2016/06/06 15:00	Battery障害	-	CacheBattery交換	該当ノード	OtherHW
t2s007029		2016/06/05 09:00-17:00	2016/06/06 15:00	HDD障害	-	HDD Bay1交換	該当ノード	Disk
t2a003041	GV	2016/06/02 10:00-11:00	2016/06/07 15:00	GPU0 ECC Error SingleBit:85493 DoubleBit:0	ECC Error	ECC Errorのクリア後、12時間のストレステストにて問題無し。	該当ノード	GPU
t2a003180	HH	2016/05/31 13:00-14:00	2016/06/02 09:00	起動不可	システムボード障害	システムボード交換	該当ノード	SystemBoard

Statistics during Aug 2012 – Jul 2013

Obtained before upgrade to TSUBAME2.5; GPUs were still M2050

	SUM	Boot Failures	Fail-stop Failures	Unaffected	Multi Failures	Repairs	GPU Repairs	DIMM Repairs
Unknown Boot Failure	35	35				0		
CPU	16	4	12			14		
Disk/Storage (wo SSD)	17	5		12		5		
Unknown Node Failure	15		15			0		
Fan	100			100		13		
GPUs								
GPU-PCI	362		362			96	39	
GPU-Link	16			16		10	3	
GPU-ECC	10		10			10	10	
GPU-Unknown	10	6	4			9	5	
Memory	12	1	7	4		11		14
Network								
Infiniband	22		4	16	2	4		
Other Networks	78		55	23		0		
Other HW	27	22		5		24		
Batch System (PBS Pro)	13		3	10		0		
PSU	33		26	7		33		
Rack	6			5	1	4		
SSD	34	12	22			32		
System Board	22	9	13			22		
Total	828	94	533	198	3	287	57	14
Corrected Total	498	94	210	191	3	209	57	14

Counted and categorized by Prof. Matsuoka

Conditions Peculiar to TSUBAME2.0

	SUM	Boot Failures	Fail-stop Failures	Unaffected	Multi Failures	Repairs	GPU Repairs	DIMM Repairs
GPUs								
GPU-PCI	362		362			96	39	
GPU-Link	16			16		10	5	
GPU-ECC	10		10			10	10	
GPU-Unknown	10	6	4			9	5	

- We saw 362 fail-stop failures owing to GPU-PCI link
  - But >60% of them have been recovered only with reboot
  - This was due to initial lot of M2050, whose power consumption sometimes exceeds assumed one

→ # of HW Repairs was 135 (including preventive ones)

This is peculiar to M2050, and now the situation on TSUBAME2.5 is improved (needs aggregation of event logs)

Current Issue and Future Plan

Robustness itself is good, but

- Current: Event logs created by engineers
  - No statistical information; aggregation requires considerable efforts
  - Events are sometimes categorized badly
- TSUBAME3 plan:
  - Logging in more automatic, and clear ways
  - Using NVIDIA DCGM

TOPIC 3:
ARE THERE NEW USAGE METHODS OF GPUS AND SSDS?

Usage of SSDs

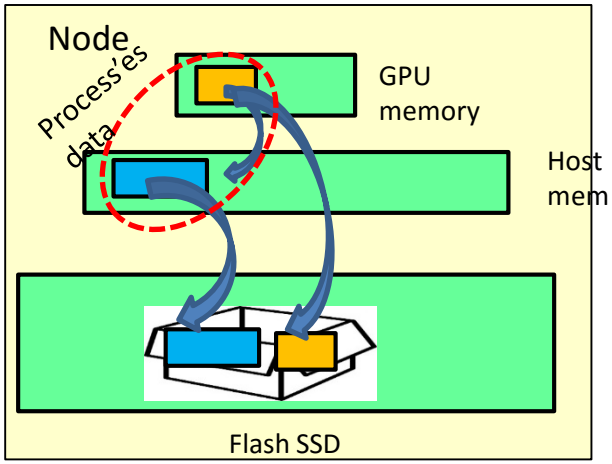
- Typical usage of SSDs on TSUBAME are:
  (1) Linux OS installation
  (2) Temporary files
  - They are natural when we regard SSDs as better replacements of local HDDs

- Is it feasible to use SSDs as extension to GPU/CPU memory?
  - Currently we are exploring with HPC (stencil) computation
  - We need careful quantitative discussion on SSDs' access performance and capacity

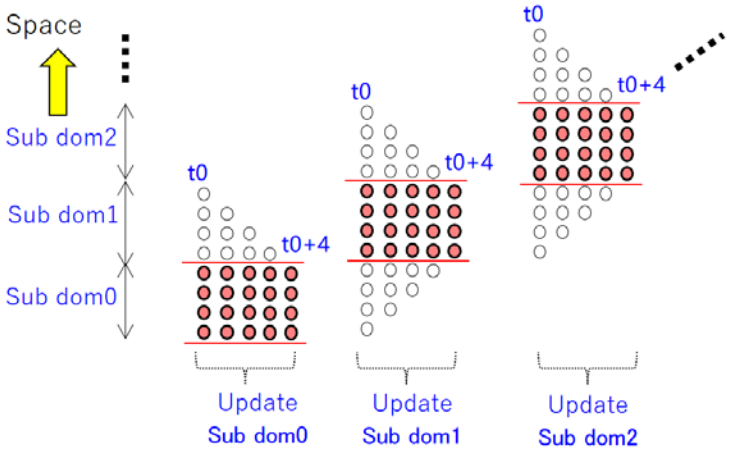
Our Method to Achieve "Out-of-Core" Stencil Computation

- OS swapping is insufficient due to
  - Larger overhead
  - GPU memory is not supported
- Instead, we combine the following two:

(1) HHRT: Runtime library for data swapping between memory hierarchy



(2) Temporal blocking technique to improve access locality of stencil computations



For basic concepts, please refer to CLUSTER14, ICPADS15 papers

Parameters of SSDs are Heavily Divergent

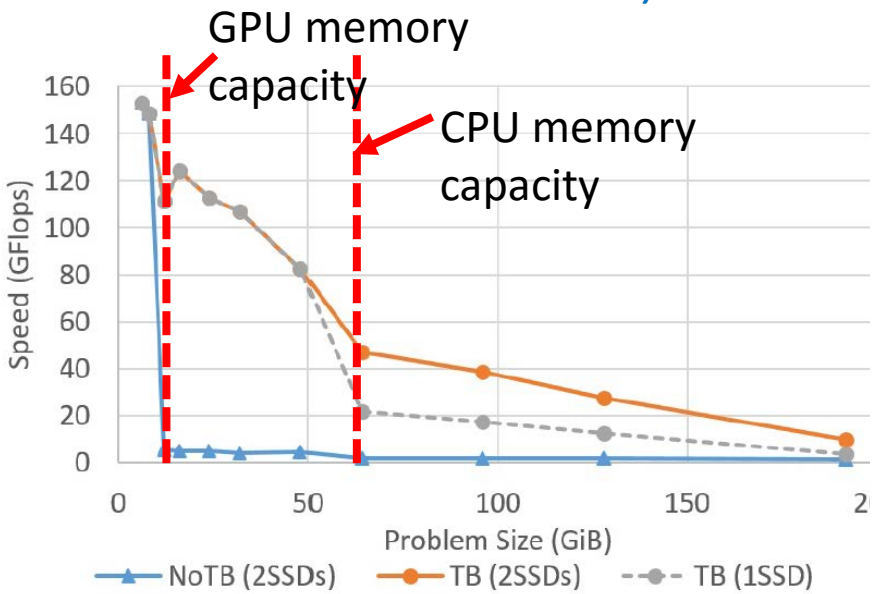
	TSUBAME2.5	TSUBAME-KFC	Test PC Server
SSD	HP 572071-B21 x2 (2010)	Intel S3500 x2 (2013)	Samsung m.2 950PRO x1 (2015)
Capacity	60 GB x 2	480 GB x 2	512GB
IOR Read BW	325 MB/s (RAID0)	480 MB/s x 2	2480 MB/s
IOR Write BW	240 MB/s (RAID0)	460 MB/s x 2	1500 MB/s

Unfortunately, we found SSDs on TSUBAME2.5 are insufficient for the current objectives, both in capacity and BW

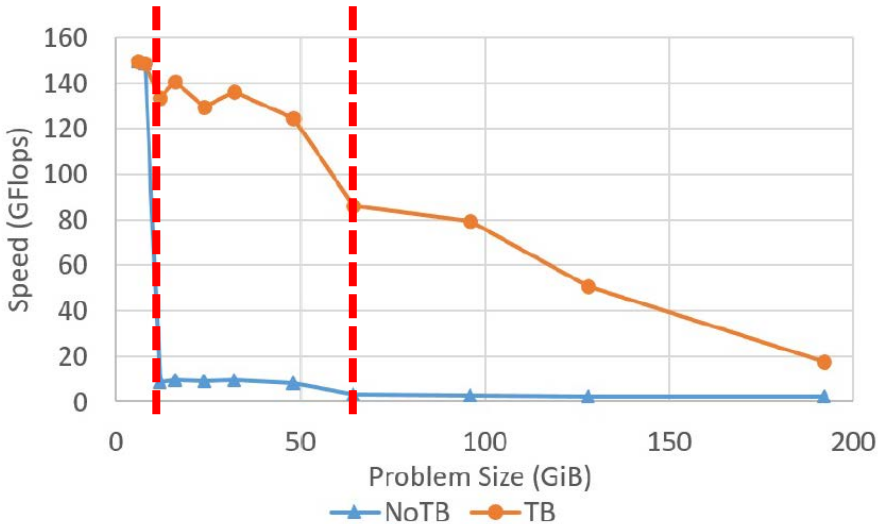
Performance of "Out-of-Core" Stencil

- 7-point stencil is measured on a TSUBAME-KFC node and test PC server w/ m.2

A TSUBAME-KFC Node w/ a K80 GPU



Test PC server w/ a K40 GPU



- With our method, we achieve "out-of-core" stencil
- Performance with "Out-of CPU memory" is directly affected by SSD performance

Current Issue and Future Plan

- Current:
  - Due to performance/capacity limitation of TSUBAME2.5 SSDs, it is hard to use them as extension to memory
- TSUBAME3 plan:
  - Adaption of SSDs with O(TB) capacity and O(GB/s) bandwidth
  - Adaption of “Burst buffers” (shared SSD) possibly
  - Support software stack to support “Extremely Big Data” applications

TSUBAME3.0 in 2017

Leading Machine Towards Exa & Big Data

1. "Everybody's Supercomputer" - High Performance (13~25 Petaflops, ~4PB/s Mem, ~1Pbit/s NW), innovative high cost/performance packaging & design, in mere 130m²…
2. "Extreme Green" – 9~1 TOPIC 3 efficient architecture, system-wide power control, advanced cooling, futu load leveling & energy recovery
3. "Big Data Convergence" – Extreme high BW &capacity, deep memory hierarchy, extreme I/O acceleration, Big Data SW Stack for machine learning, graph processing, …
4. "Cloud SC" – dynamic deployment, container-based node co-location & dynamic configuration, resource elasticity, assimilation of public clouds…
5. "Transparency" - full monitoring & user visibility of machine & job state, accountability via reproducibility

TOPIC 1

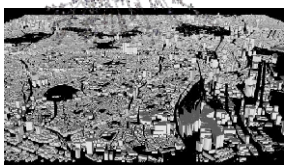
TOPIC 2

2013 TSUBAME2.5 upgrade 5.7PF DFP

2017 TSUBAME3.0 .3~25PF(DFP) ~4PB/s Mem BW 9~10GFlops/W power efficiency Big Data & Cloud Convergence

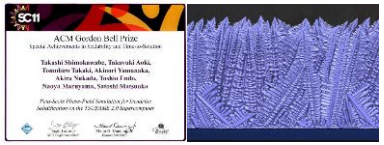
2010 TSUBAME2.0 2.4 Petaflops #4 World "Greenest Production SC"

2013 TSUBAME-KFC #1 Green 500



Large Scale Simulation Big Data Analytics Industrial Apps

5 TSUBAME1.0 ps, #1 Asia #7 World Everybody's Supercomputer"



2011 ACM Gordon Bell Prize