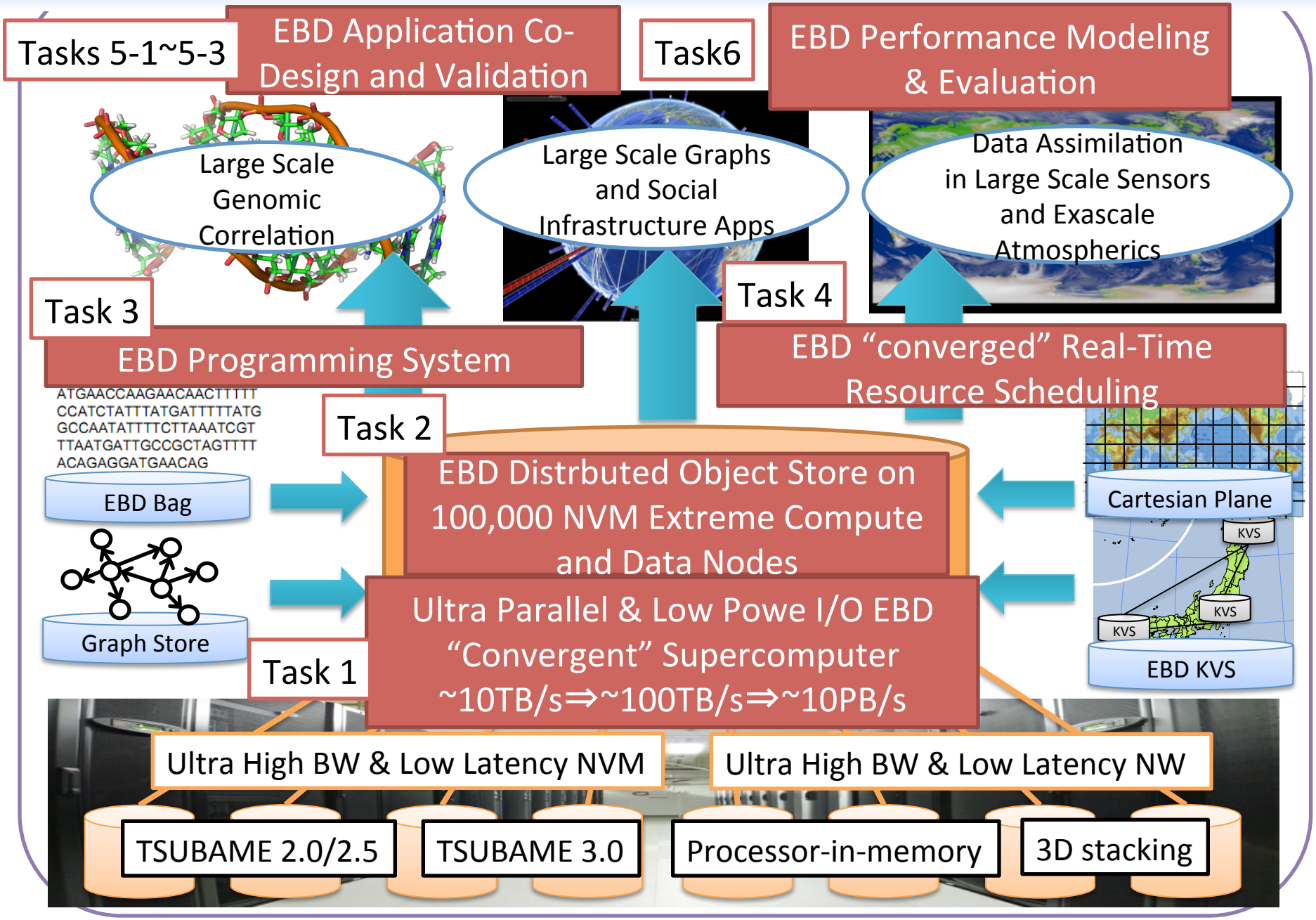
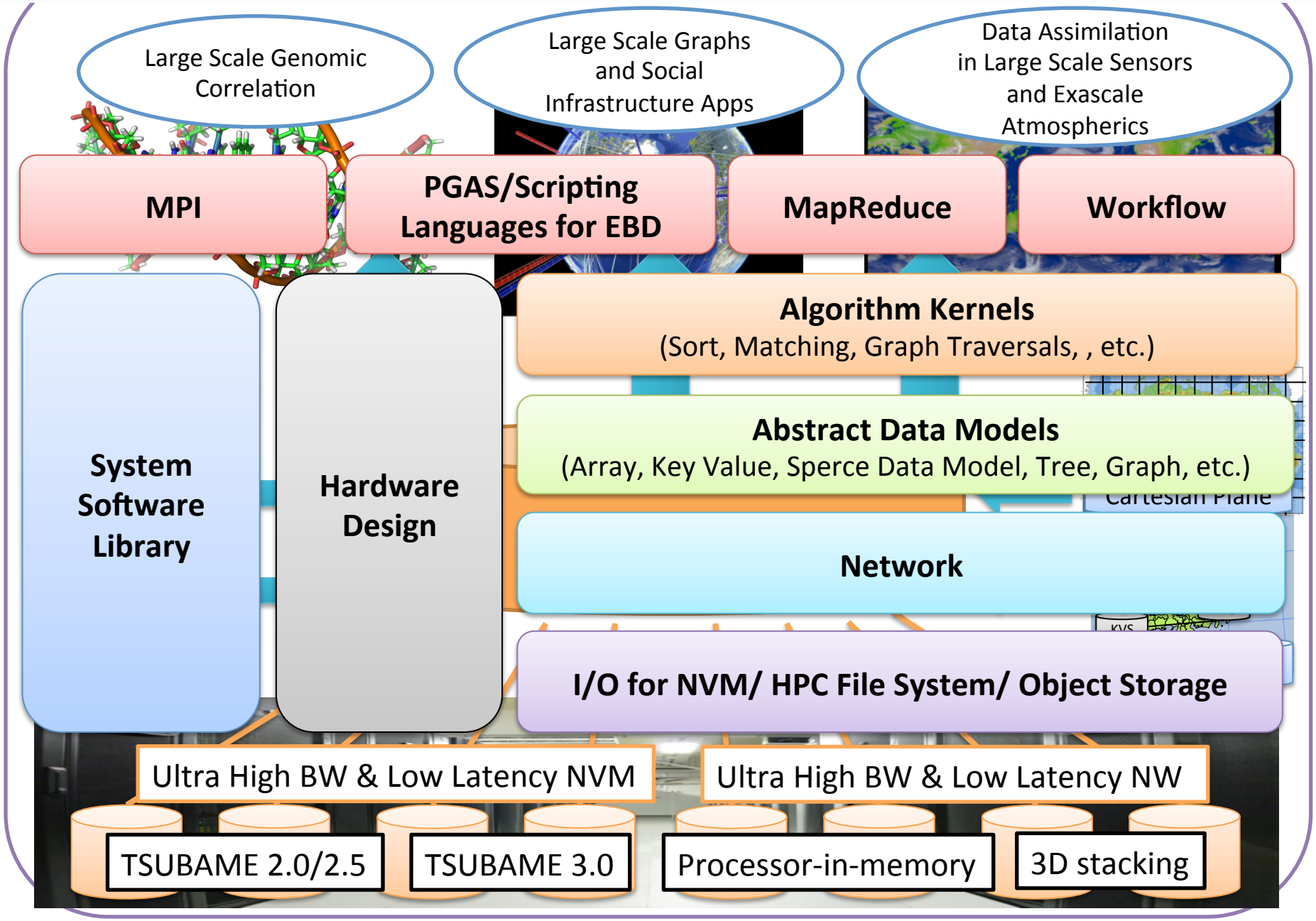
100,000 Times Fold EBD "Convergent" System Overview

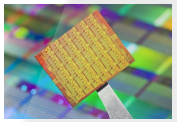


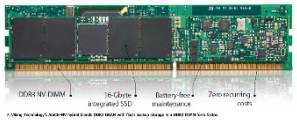
100,000 Times Fold EBD "Convergent" System Overview

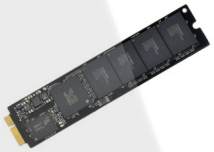


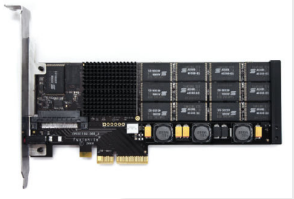
Hamar (Highly Accelerated Map Reduce) [Shirahata, Sato et al. Cluster2014]

- A software framework for large-scale supercomputers w/ many-core accelerators and local NVM devices
  - Abstraction for deepening memory hierarchy
    - Device memory on GPUs, DRAM, Flash devices, etc.
- Features
  - Object-oriented
    - C++-based implementation
    - Easy adaptation to modern commodity many-core accelerator/Flash devices w/ SDKs
      - CUDA, OpenNVM, etc.
  - Weak-scaling over 1000 GPUs
    - TSUBAME2
  - Out-of-core GPU data management
    - Optimized data streaming between device/host memory
    - GPU-based external sorting
  - Optimized data formats for many-core accelerators
    - Similar to JDS format

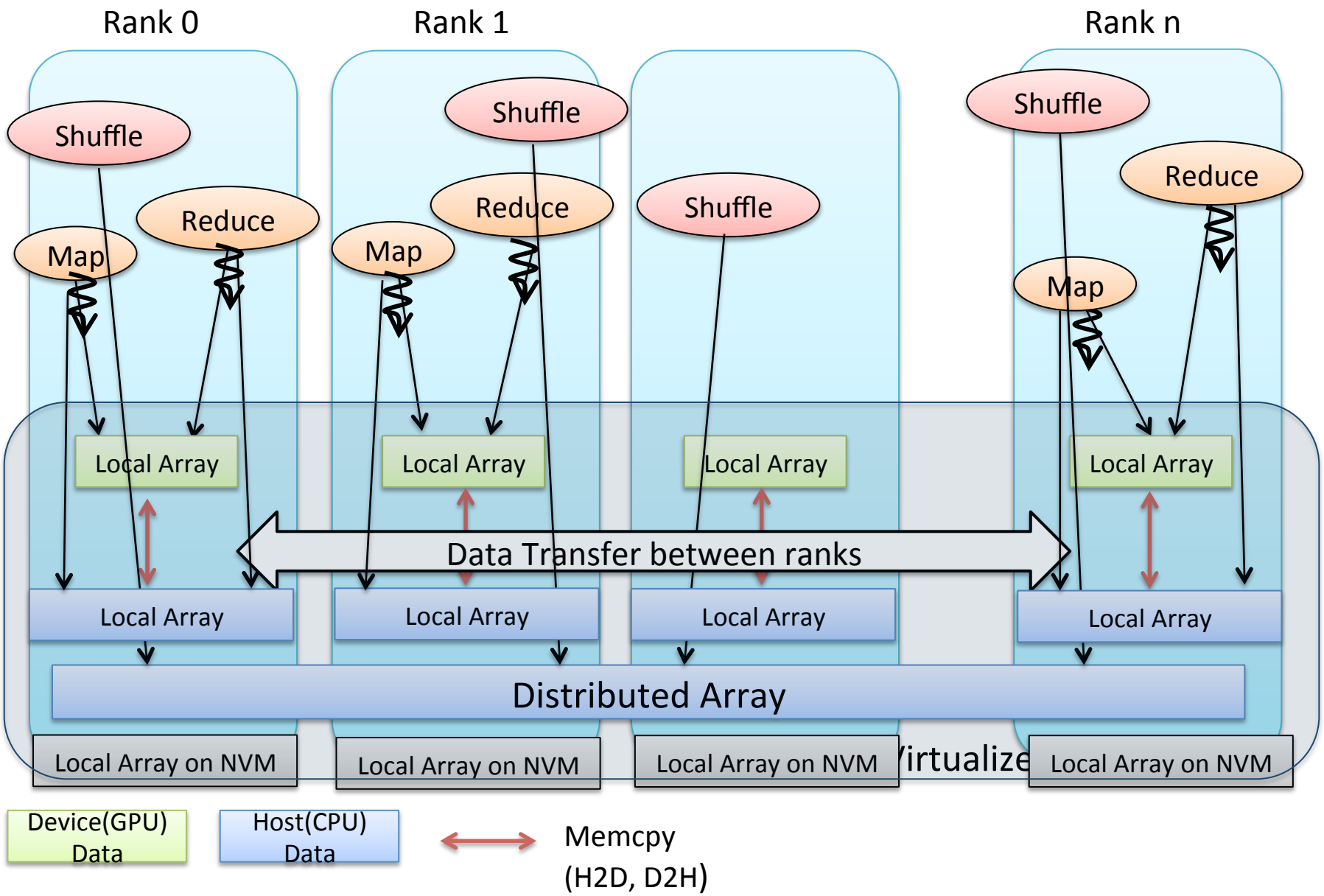








Hamar Overview

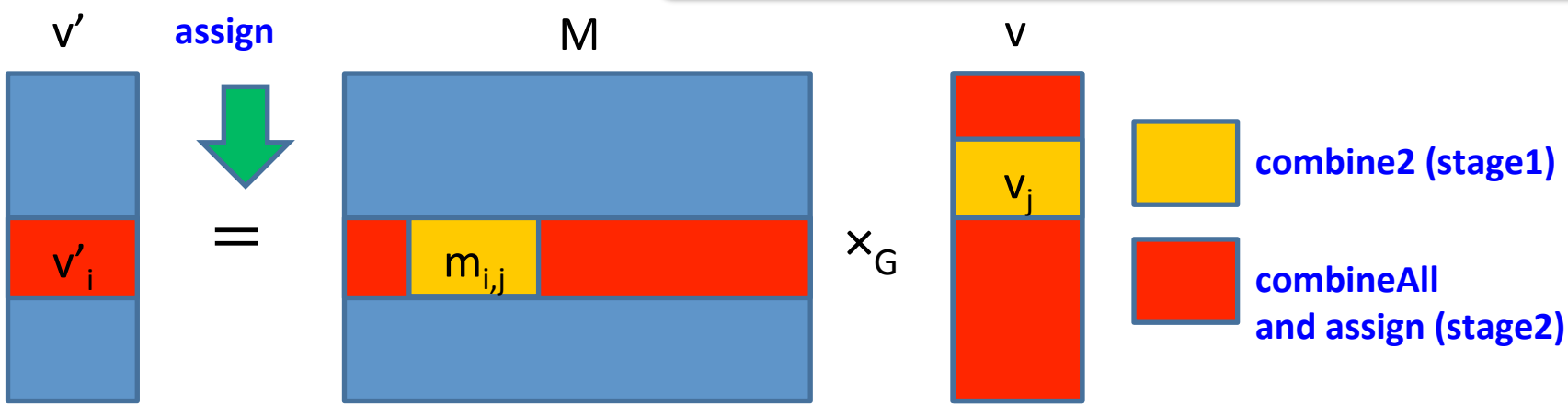


Application Example : GIM-V

Generalized Iterative Matrix-Vector multiplication*1

- Easy description of various graph algorithms by implementing combine2, combineAll, assign functions
- PageRank, Random Walk Restart, Connected Component
  - $v' = M \times_G v$ where
    $v'_i = \text{assign}(v_j, \text{combineAll}_j(\{x_j \mid j = 1..n, x_j = \text{combine2}(m_{i,j}, v_j)\}))$ (i = 1..n)
  - Iterative 2 phases MapReduce operations

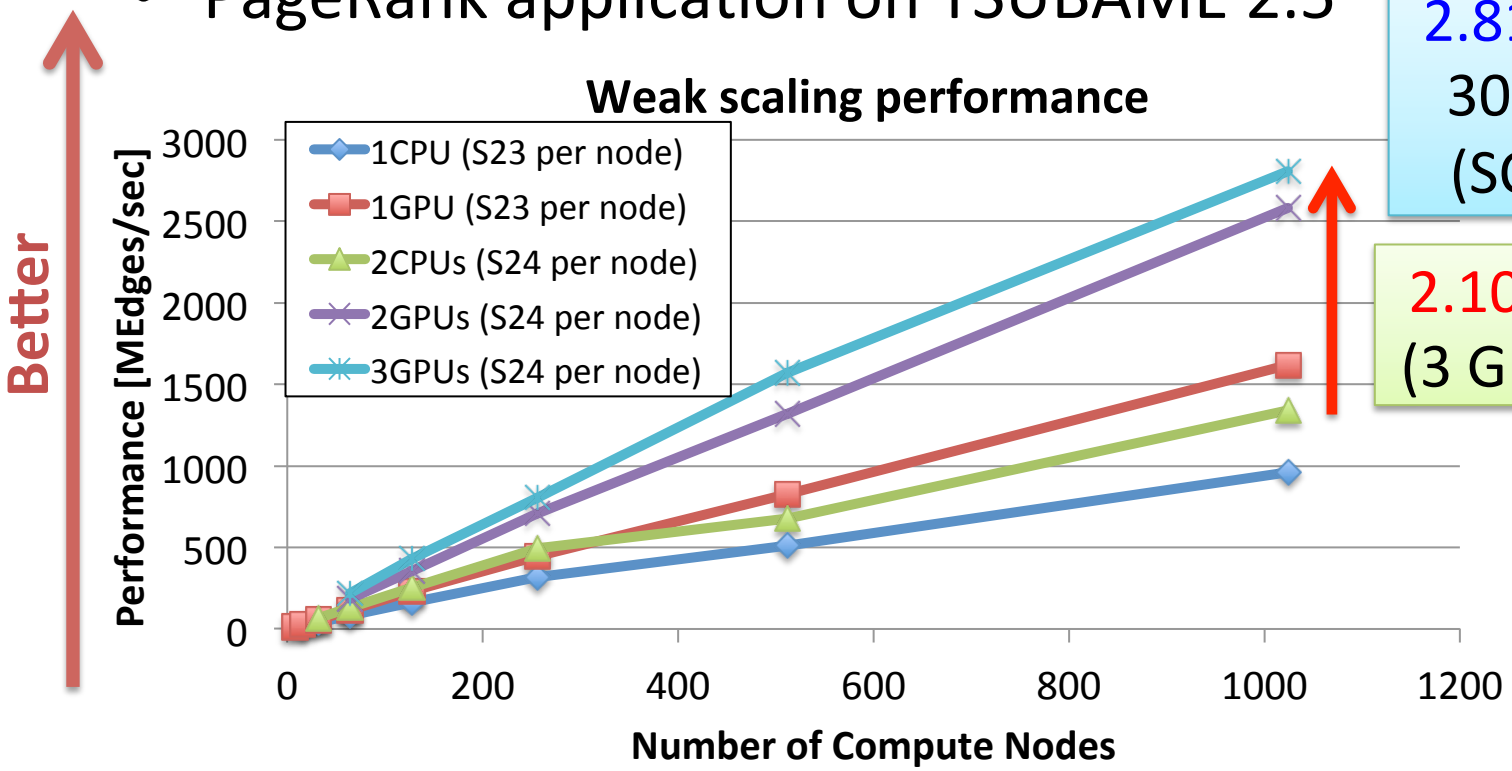
Straightforward implementation using Hamar



*1 : Kang, U. et al, "PEGASUS: A Peta-Scale Graph Mining System- Implementation and Observations", IEEE INTERNATIONAL CONFERENCE ON DATA MINING 2009

MapReduce-based Graph Processing with Out-of-core Support on GPUs

- Hierarchical memory management for large-scale graph processing using multi-GPUs
  - Support out-of-core processing on GPU
  - Overlapping computation and CPU-GPU communication
- PageRank application on TSUBAME 2.5



2.81 GE/s on 3072 GPUs (SCALE 34)

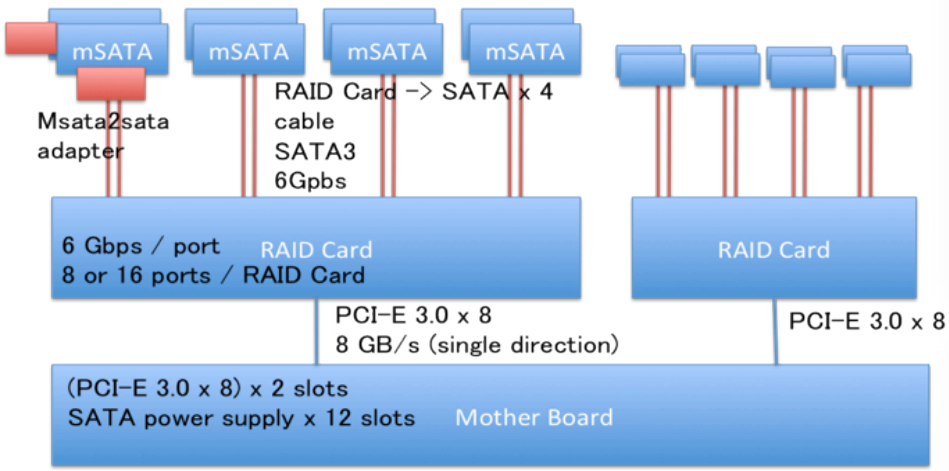
2.10x Speedup (3 GPU v 2CPU)

GPUアクセラレータと不揮発性メモリを考慮したI/O構成法 [Shirahata, Sato et al. HPC141]

16 枚の mSATA SSD を用いたプロトタイプマシンの設計

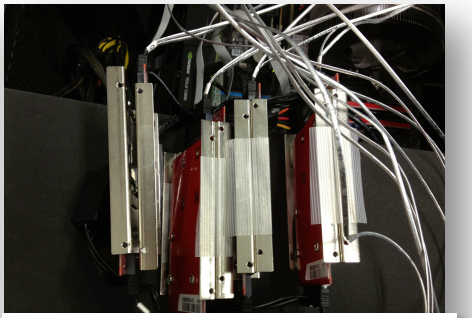
容量: 256GB x 16枚 → 4TB　Readバンド幅: 0.5GB/s x 16枚 → 8 GB/s

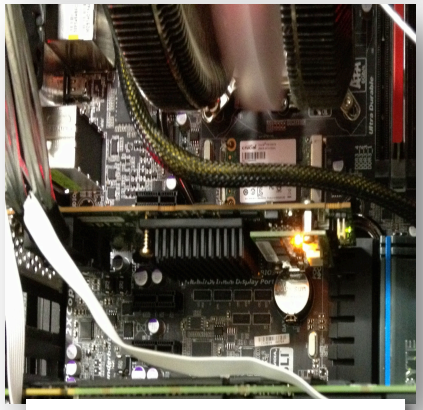
mSATA x 16 (= 8 (per RAID card) x 2 (RAID cards))
256GB x 16
500MB/s per mSATA (sequential)
-> 8 GB/s (theoretical peak) (= 0.5GB/s x 16)



CPU	Intel Core i7-3770K CPU (3.50GHz x 4 cores)
Memory	Cetus DDR3-1600 (16GB)
M/B	GIGABYTE GA-Z77X-UD5H
SSD	Crucial m4 msata 256GB CT256M4SSD3 (Peak read: 500MB/s, Peak write: 260MB/s)
SATA converter	KOUTECH IO-ASS110 mSATA to 2.5' SATA Device Converter with Metal Fram
RAID Card	Adaptec RAID 7805Q ASR-7805Q Single


A single mSATA SSD


8 integrated mSATA SSDs

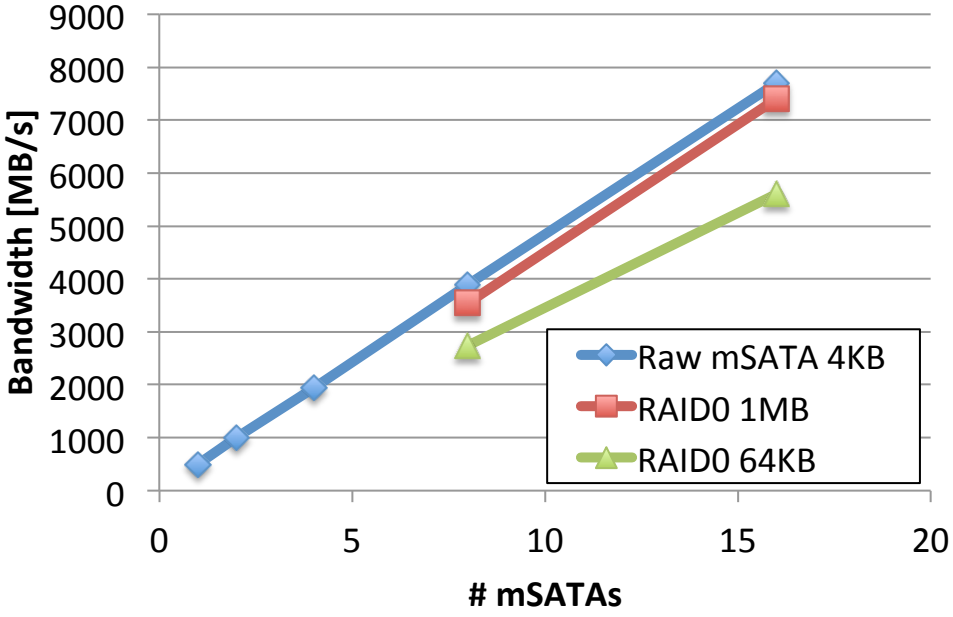

RAID cards


Prototype/Test machine

GPUから複数 mSATA SSD への I/O性能の予備評価

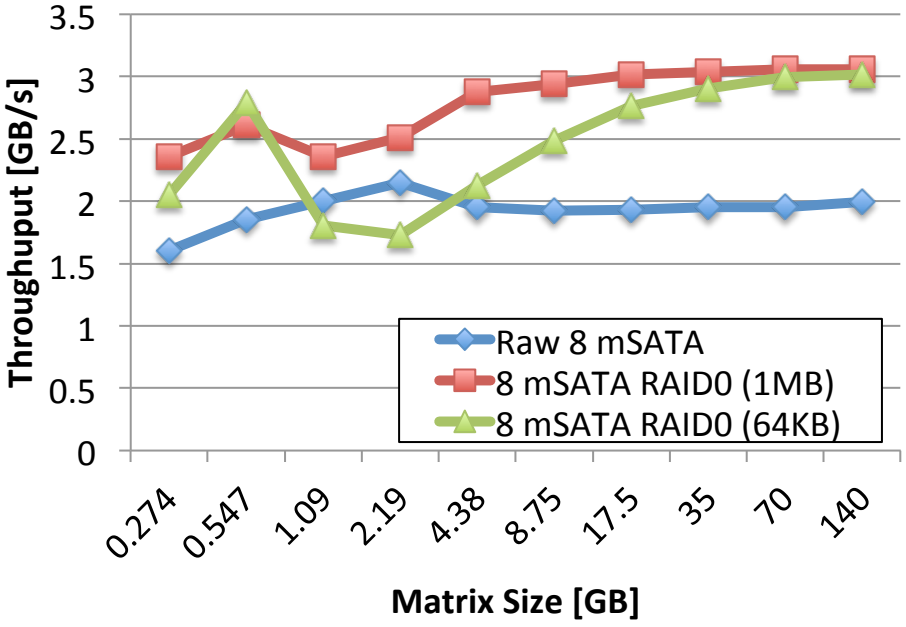
複数 mSATA SSD の I/O 性能
～ 7.39 GB/s (RAID0 を使用)

FIO による複数 mSATA SSD の I/O 性能



GPU への I/O 性能
～ 3.06 GB/s (PCI-E 上限)

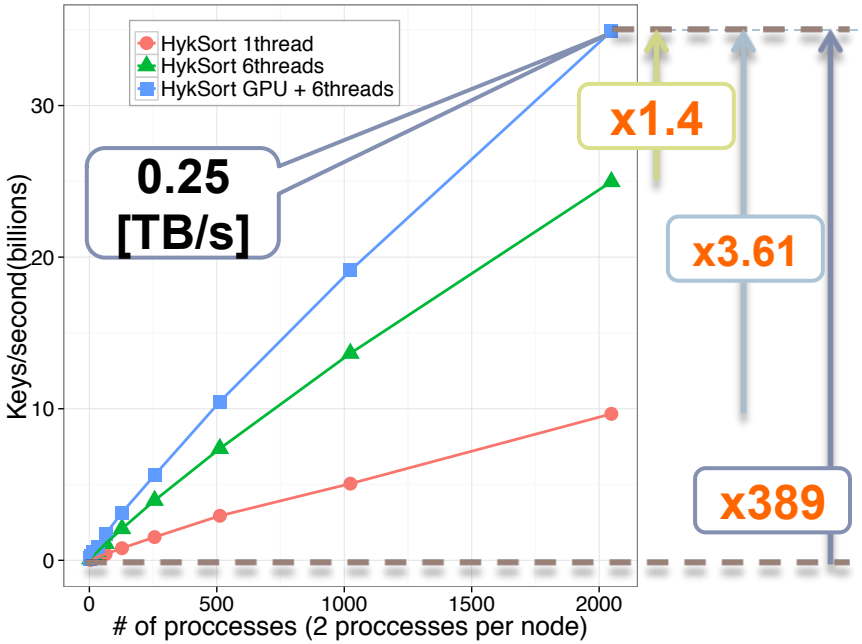
行列ベクトル積による GPU への I/O 性能



Sorting for EBD

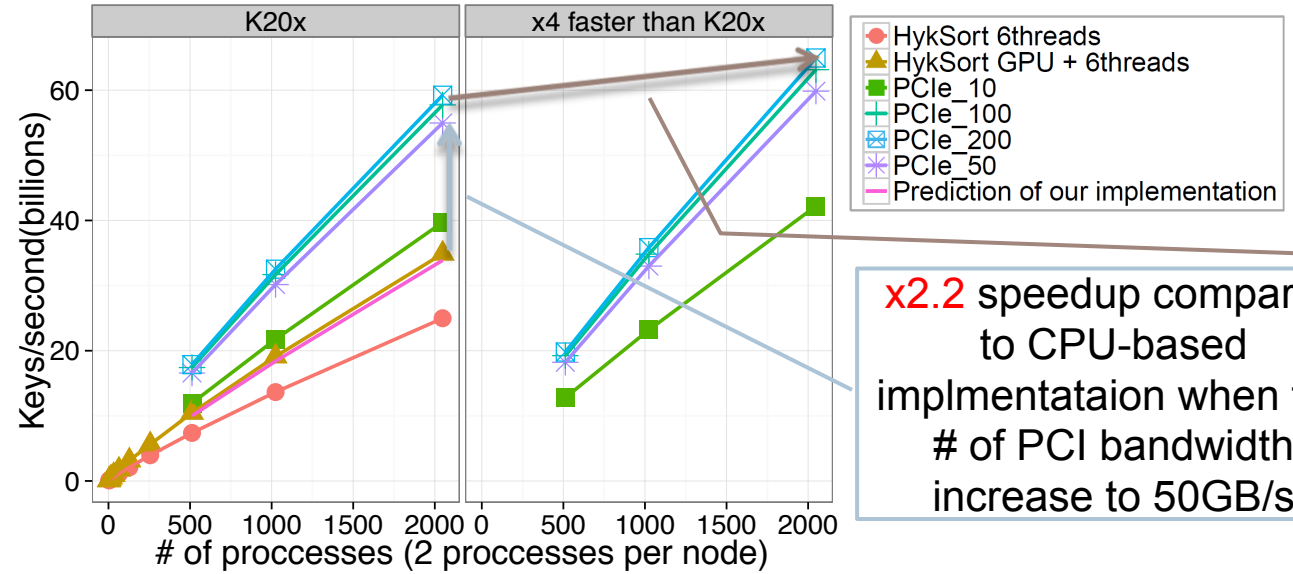
Plugging in GPUs for large-scale sorting

[Shamoto, Sato et al. BigData 2014]



- **GPU implementation of splitter-based sorting** (HykSort)
- Weak scaling performance (Grand Challenge on TSUBAME2.5)
  - 1 ~ 1024 nodes (2 ~ 2048 GPUs)
  - 2 processes per node and each node has 2GB 64bit integer
- Yahoo/Hadoop Terasort: 0.02[TB/s]
  - Including I/O

- **Performance prediction**



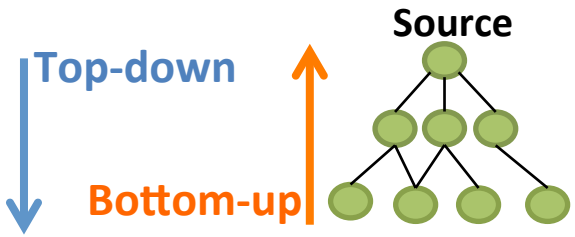
- PCIe_#: #GB/s bandwidth of interconnect between CPU and GPU

x2.2 speedup compared to CPU-based implmentataion when the # of PCI bandwidth increase to 50GB/s

8.8% reduction of overall runtime when the accelerators work 4 times faster than K20x

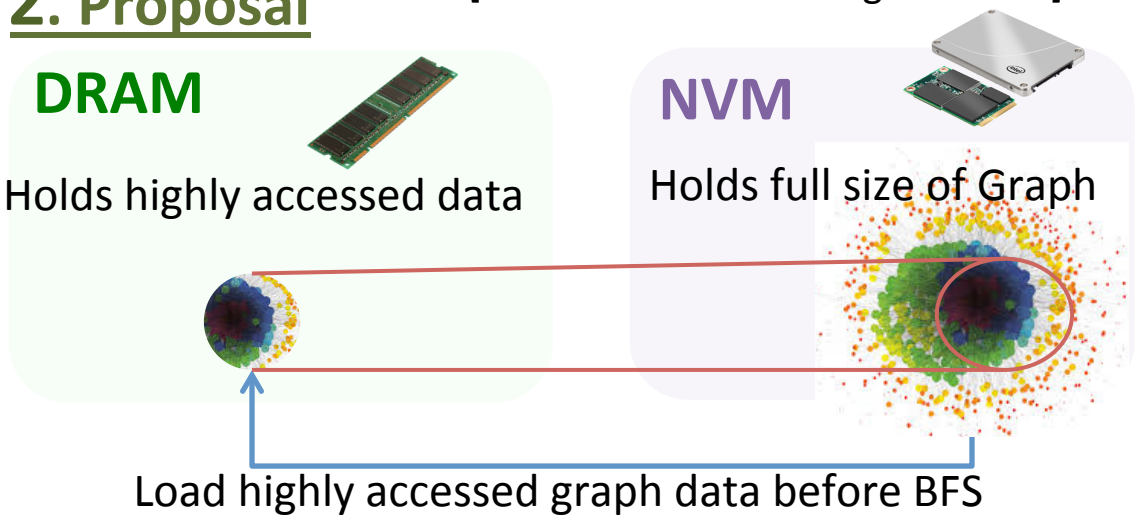
Large Scale Graph Processing Using NVM

1. Hybrid-BFS (Beamer'11)



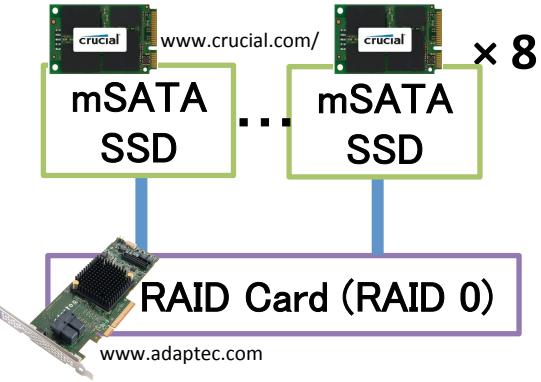
2. Proposal

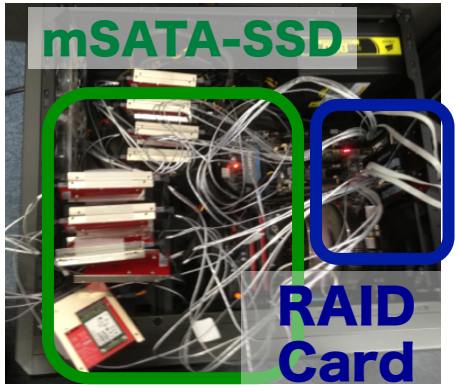
[Iwabuchi Sato et al. BigData2014]

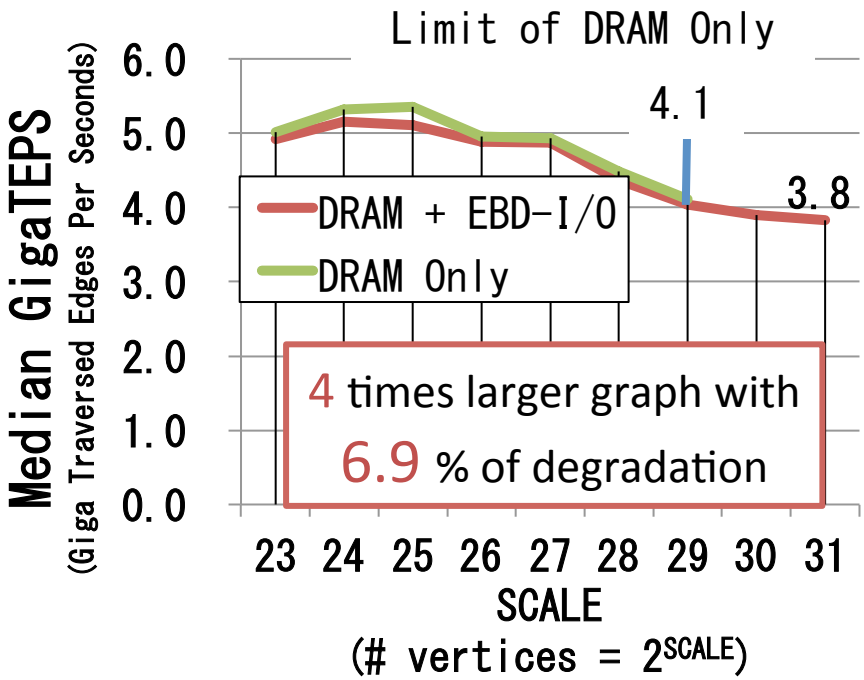


3. Experiment

CPU	Intel Xeon E5-2690 × 2
DRAM	256 GB
NVM	EBD-I/O 2TB × 2







Results : BFS Performance

The Graph 500 2014 June
DRAM + NVM model

	MEM-CREST Node #2 (Supermicro 2027GR-TRF)	GraphCrest Node #1	EBD-RH5885v2 (Huawei Tecal RH5885 V2)
DRAM	128 GB	256 GB	1024 GB
NVM	ioDrive2 1.2 TB × 2	EBD-I/O 2TB × 2	• Tecal ES3000 800GBx2,1.2TBx2 • EBD-I/O 4TB × 2
SCALE (Total Data Size)	30 (500GB)	31 (1TB)	33 (4TB)
GTEPS	7.98	13.80	3.11
MTEPS / W	28.88	35.21	3.42

The 2nd Green Graph500 list on Nov. 2013

- Measures power-efficient using **TEPS/W** ratio
- Results on various system such as **Huawei's RH5885v2 w/ Tecal ES3000 PCIe SSD 800GB * 2 and 1.2TB * 2**
- **http://green.graph500.org**

In the **Big Data** category:

Rank	MTEPS/W	Site	Machine	G500 rank	Scale	GTEPS	Nodes
1	**6.72**	Tokyo Institute of Technology	TSUBAME KFC	47	32	44.01	32
2	**5.41**	Forschungszentrum Julich (FZJ)	JUQUEEN	3	38	5848	16384
3	**4.42**	Argonne National Laboratory	DOE/SC/ANL Mira	2	40	14328	32768
4	**4.35**	Tokyo Institute of Technology	EBD-RH5885v2	96	30	3.67	1
5	**3.55**	Lawrence Livermore National Laboratory	DOE/NNSA/LLNL Sequoia	1	40	15363	65536
6	**1.89**	Research Center for Advanced Computing Infrastructure	altix	50	30	37.66	1
7	**0.73**	Mayo Clinic	grace	68	31	10.32	64

Third Green Graph 500 List (released June 2014)

In the **Big Data** category:

Rank	MTEPS/W	Site	Machine	G500 rank	Scale	GTEPS	Nodes	Cores
1	**59.12**	Kyushu University	GraphCREST-SandybridgeEP-2.4GHz	57	30	28.48	1	32
2	**48.29**	Kyushu University	GraphCREST-Sandybridge-EP-2.7GHz	59	30	31.95	1	32
3	**35.21**	Tokyo Institute of Technology	GraphCREST-Custom #1	71	31	13.8	1	32
4	**23.88**	Tokyo Institute of Technology	MEM-CREST Node #2	85	30	7.98	1	16
5	**17.24**	Kyushu University	GraphCREST-Bulldozer	72	31	13.63	1	64
6	**14.06**	Tokyo Institute of Technology	TSUBAME-KFC	46	32	104.31	32	384
7	**12.48**	The Institute of Statistical Mathematics	ismuv2k2	42	32	131.43	1	640
8	**5.41**	Forschungszentrum Julich (FZJ)	JUQUEEN	4	38	5848	16384	262144
9	**4.42**	Argonne National Laboratory	DOE/SC/ANL Mira	3	40	14328	32768	524288
10	**4.35**	Tokyo Institute of Technology	EBD-RH5885v2	105	30	3.67	1	48
11	**3.55**	Lawrence Livermore National Laboratory	DOE/NNSA/LLNL Sequoia	2	40	15363	65536	1048576

Expectations for Next-Gen Storage System (Towards TSUBAME3.0)

- Achieving high IOPS
  - Many apps with massive small I/O ops
    - graph, etc.
- Utilizing NVM devices
  - Discrete local SSDs on TSUBAME2
  - How to aggregate them?
- Stability/Reliability as Archival Storage
- I/O resource reduction/consolidation
  - Can we allow a large number of OSSs for achieving ~TB/s throughput
  - Many constraints
    - Space, Power, Budget, etc.

Current Status

- New Approaches
  - Tsubame 2.0 has pioneered the use of local flash storage as a high-IOPS alternative to an external PFS
  - Tired and hybrid storage environments, combining (node) local flash with an external PFS
- Industry Status
  - High-performance, high-capacity flash (and other new semiconductor devices) are becoming available at reasonable cost
  - New approaches/interface to use high-performance devices (e.g. NVMexpress)