学際大規模情報基盤共同利用・共同研究拠点 萌芽型共同研究 採択課題
EX17312 (東京大学情報基盤センター推薦課題)
9th Symposium
Joint Usage / Research Center for Interdisciplinary Large-scale Information Infrastructures

佐藤　仁 (産業技術総合研究所　人工知能研究センター)

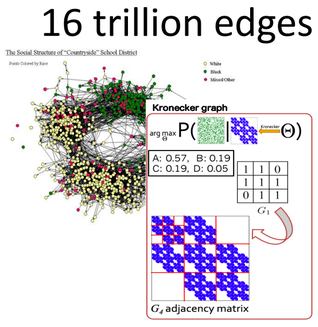
将来の大規模メニーコアプロセッサ環境に向けたビッグデータ基盤処理の性能評価

Extreme Big Data Examples

Large-scale Graphs

Graph500:

- ✓ New benchmark for Big Data processing on supercomputers
- ✓ 0.45 sec for a Kronecker graph w/ 1 trillion vertices and 16 trillion edges using 82,944 nodes, 663,552 processes



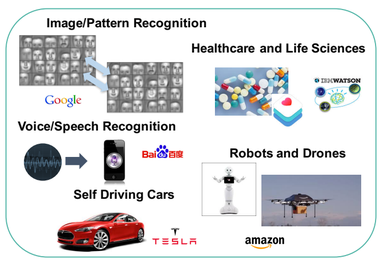
July 2015

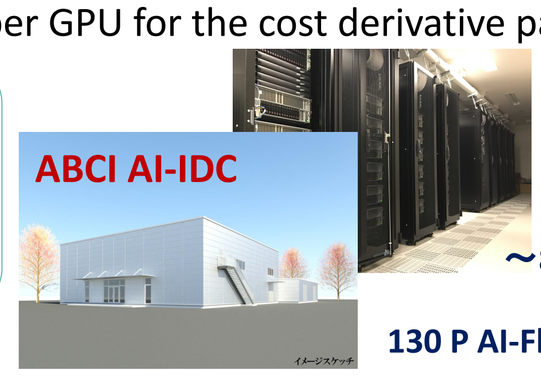
No.	Rank	Machine	Installation Site	Number of nodes	Number of cores	Problem scale	GTEPS
1	1	K computer (Fujitsu Custom)	RIKEN Advanced Institute for Computational Science (AICS)	82944	663552	40	38621.4
2	2	DOE/NNSA/LLNL Sequoia (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Lawrence Livermore National Laboratory	98304	1572864	41	23751
			tional	49152	786432	40	14982
			entrum	16384	262144	38	5848
				8192	131072	37	2567

AI / Machine Learning

Deep Learning Training:

- ✓ 96 GPUs are used for a single DNN model in ImageNet dataset using a node-distributed asynchronous machine learning implementation
- ✓ 1146 GPUs are used in Total for the training
- ✓ Reaching 1 Tflops (SFP) per GPU for the cost derivative parts





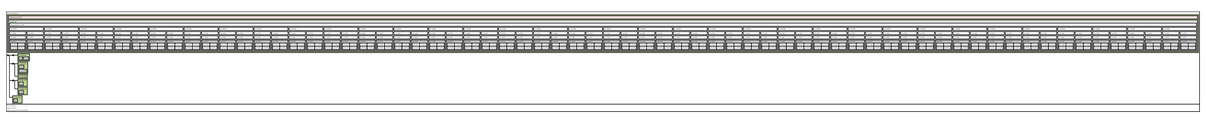
AIST AAIC
AI-Cloud
～8.4P AI-Flops
130 P AI-Flops～

Towards HPC and Big Data Convergence

- ✓ Dropping available memory capacity per core for achieving efficient bandwidth by increasing in parallelism, heterogeneity, density of processors, e.g.) Multi-core CPUs, Many-core accelerator, Post-Moore Era
- ✓ Deeping Memory/Storage Architectures, such as NVM (Non-Volatile Memory), SCM (Storage Class Memory), e.g. Flash, MCDRAMM, PCM, STT-MRAM, ReRAM, HMC, etc.

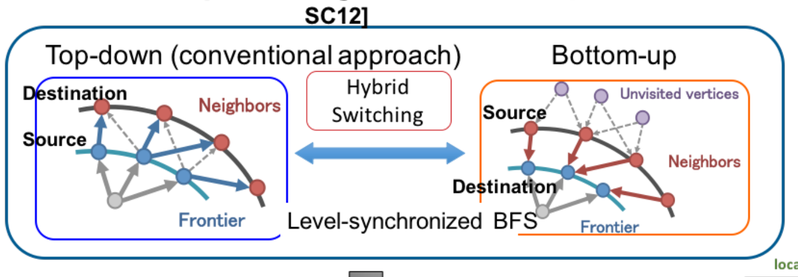
Motivation:

- ✓ How to utilize many-core environments for accelerating Extreme Big Data applications, such as Large-scale Graphs, AI, etc.
  - How to overcome memory capacity limitation?
  - How to offload bandwidth oblivious operations onto low throughput memory volumes?



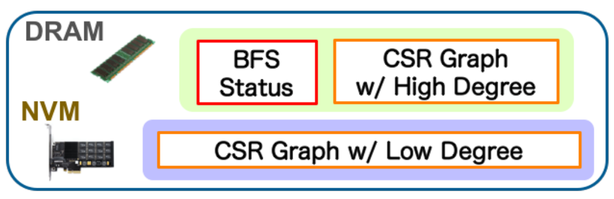
Out-of-core BFS for Graph500

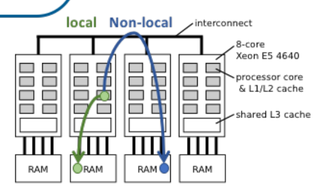
Direction Optimizing BFS Search [Beamer SC12]



NUMA-optimized Implementation [Yasui BigData2013]

Proposed Technique

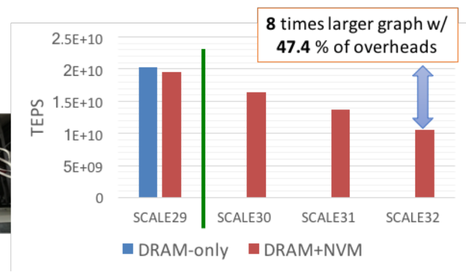




Optimized data between placement using vertex sorting

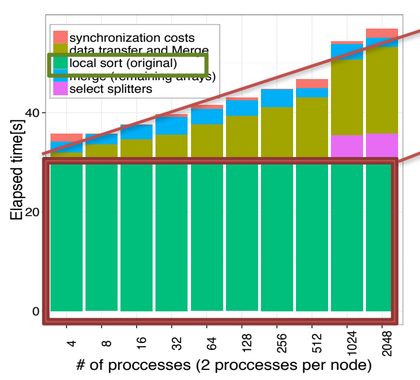
Experiment

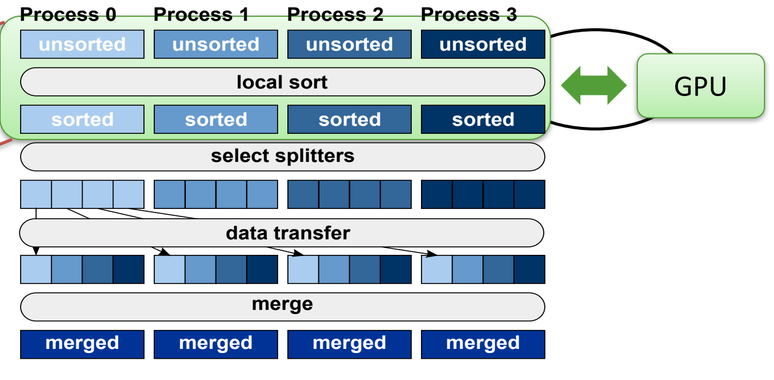
CPU	Intel Xeon E5-2690v2 × 2
DRAM	256 GB
NVM	EBD-I/O 4TB (2TB×2)
Data	RMAT Graph w/ Edge factor 16 (Graph500)



Scalable Large-scale Sorting

- Effectiveness of large-scale sorting on supercomputers w/ massive amounts of throughput processors remains unclear
  - ✓ What operations are accelerated or bottlenecked by GPUs in large-scale sorting
  - ✓ Handling data overflow on hierarchical memory volumes may also be required
- Based on sample sorting, offloading the most time-consuming local sort onto GPUs and network communications



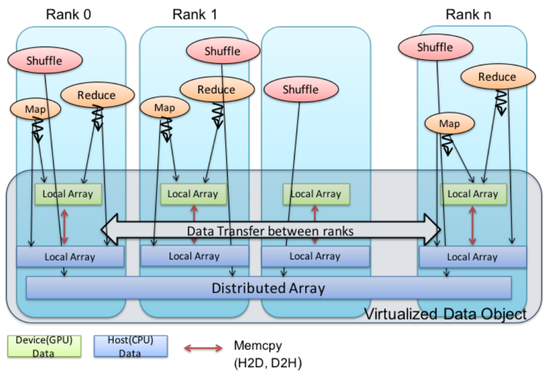


MapReduce for Scalable Supercomputers

A Software Framework for Large-scale Supercomputers w/ Many-core Accelerators and Burst Buffer NVMs

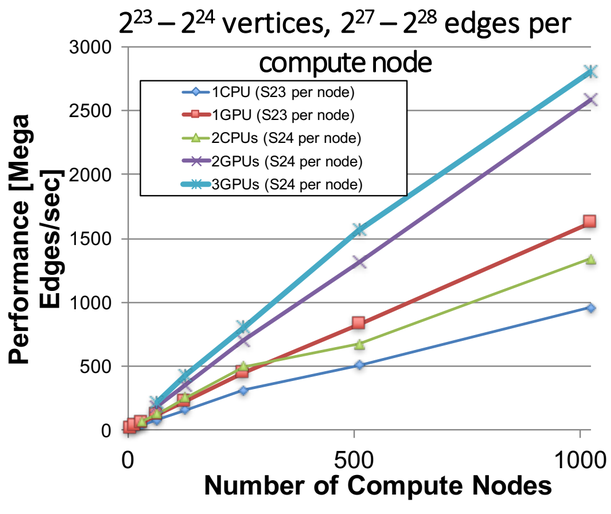
- ✓ Abstraction for deep memory hierarchy, e.g., GPU, CPU, NVM
- Object-oriented (C++)
- Weak-scaling over 1000 GPUs
- Out-of-core GPU data management
- Optimized data formats for many-core accelerators

How to apply similar techniques for many-core environments



Weak Scaling on K20X-based TSUBAME2.5
MapReduce-based Page Rank Application for RMAT Graph (Graph500)

2.81 Giga Edges/sec on 3072 GPUs
(17.2 bn vertices, 274 bn edges)



The size of graph exceeds the total GPU memory capacity

2.10x Speedup
(3 GPU vs 2CPU)

Performance improvement by sorting, prefix-sum, data-parallel operations