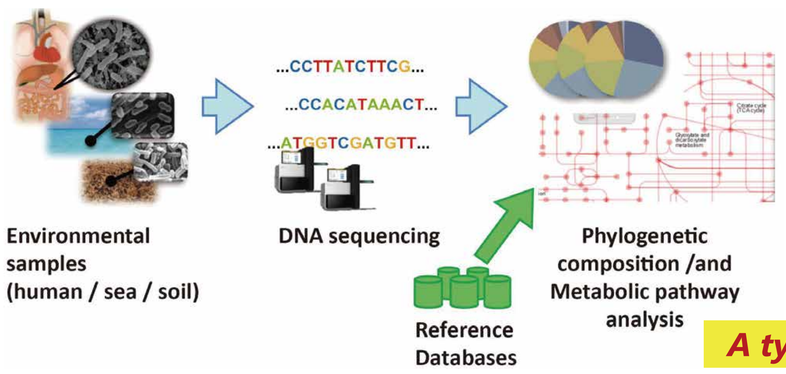
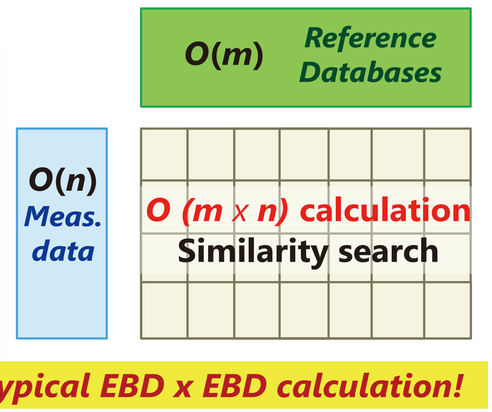
Extreme Big Data - Applications

Next Generation Big Data Infrastructure Technologies
Towards Yottabyte / Year

Ultra-fast Metagenome Analysis

Metagenome Analysis

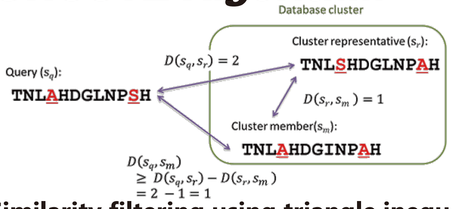




A typical EBD x EBD calculation!

GHOSTZ-MP: Ultra-fast Sequence Homology Search

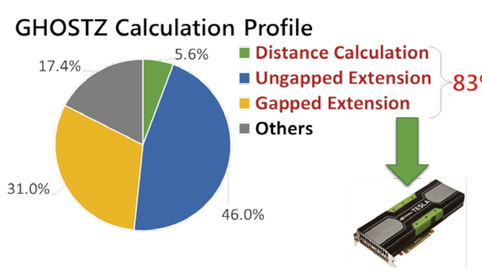
GHOSTZ Algorithm [1][2]



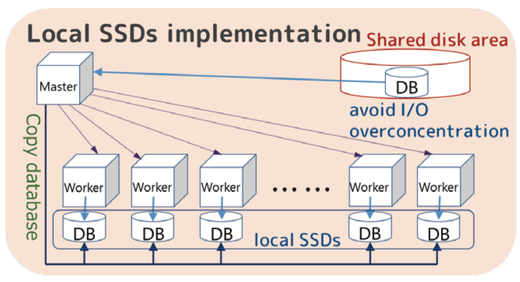
Similarity filtering using triangle inequality among subsequences

GHOSTZ is ~185-261x faster than BLASTX.

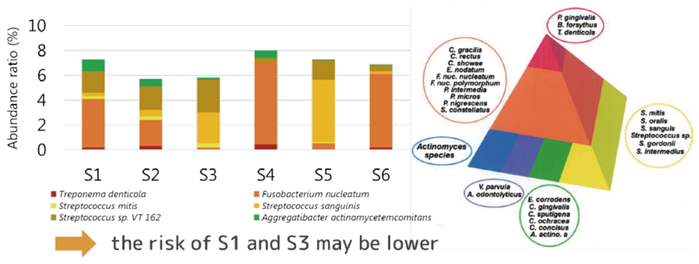
GHOSTZ-GPU [3]*: GPU Acceleration*



GPU/MPI Parallelization



Oral Metagenome Application



[1] Suzuki S, Kakuta M, Ishida T, Akiyama Y. *PLoS ONE*, **9**(8): e103833, 2014.
[2] Suzuki S, Kakuta M, Ishida T, Akiyama Y. *Bioinformatics*, **31**(8): 1183-1190, 2015.
[3] Suzuki S, Kakuta M, Ishida T, Akiyama Y. *PLoS ONE*, **11**(8): e0157338, 2016.

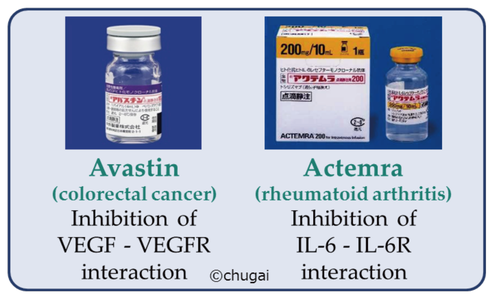
Acknowledgments. This work was partly supported by the Strategic Programs for Innovative Research (SPIRE) Field 1 Supercomputational Life Science of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan and Core Research for Evolutional Science and Technology (CREST) "Extreme Big Data" from the Japan Science and Technology Agency (JST).

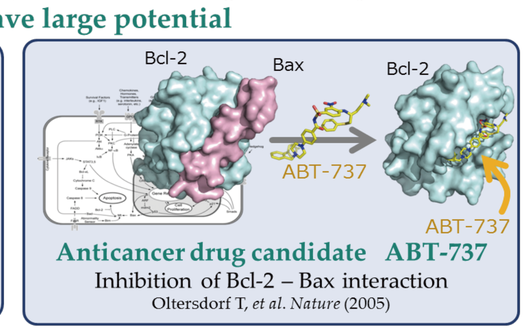
Exhaustive PPI Predictions

Protein-Protein Interactions (PPIs)

Elucidation of protein-protein interactions(PPIs) is important for understanding disease mechanisms and for drug discovery

PPI Inhibitors have large potential

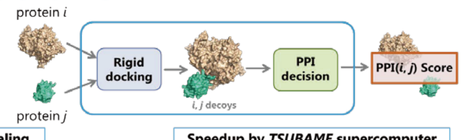


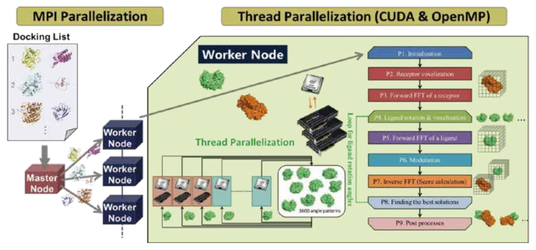


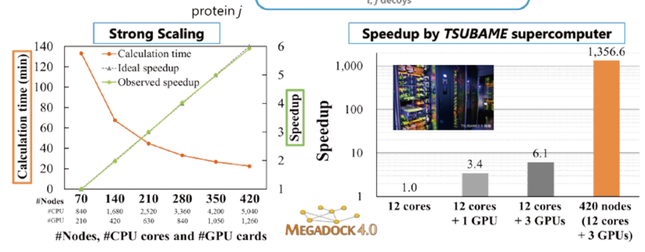
MEGADOCK: Ultra-fast PPI Predictions

MEGADOCK [4]

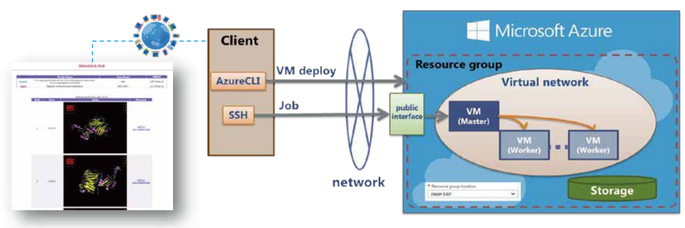
- 1-million PPIs can be predicted in a day with 3,600 CPU cores + 1,200 K20x GPUs
- Available on CPU cluster, GPU cluster[5], MIC (Xeon Phi) cluster[6]

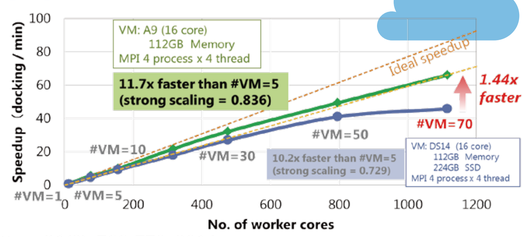






Azure-based PPI Prediction Platform





[4] Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y. *Protein and Peptide Letters*, **21**(8): 766-778, 2014.
[5] Ohue M, Shimoda T, Suzuki S, Matsuzaki Y, Ishida T, Akiyama Y. *Bioinformatics*, **30**(22): 3281-3283, 2014.
[6] Shimoda T, Suzuki S, Ohue M, Ishida T, Akiyama Y. *BMC Systems Biology*, **9**(Suppl 1): S6, 2015.

Acknowledgments. This work was partly supported by the Next-generation Integrated Living Matter Simulation (ISLiM) project of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) Japan, Core Research for Evolutional Science and Technology (CREST) "Extreme Big Data" from the Japan Science and Technology Agency (JST). HPCI system research project (hp120131), Microsoft Japan Co., Ltd. and Leave a Nest research grant.

Predicting Statistics of a Distributed DL System

Collaborative work with DENSO CORPORATION and DENSO IT LABORATORY, INC

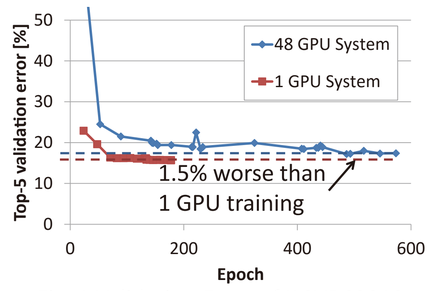


Fig. 1 Validation Error of ILSVRC2012 Classification Task on Two Platforms

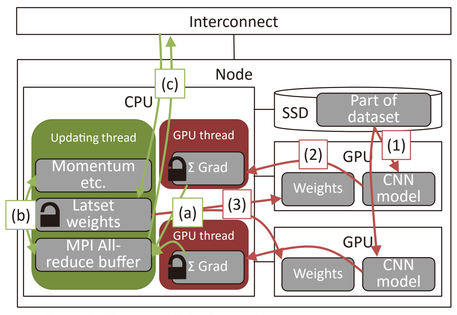


Fig. 2 Target DL App. Architecture

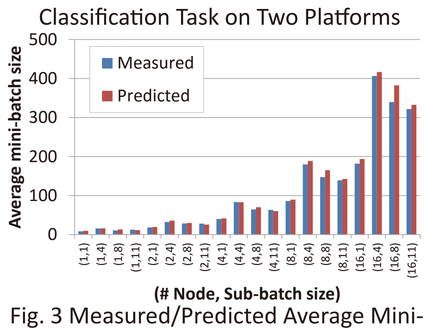


Fig. 3 Measured/Predicted Average Mini-batch Size On TSUBAME-KFC/DL

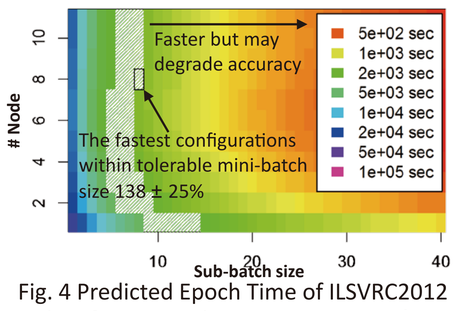


Fig. 4 Predicted Epoch Time of ILSVRC2012 Classification Task on TSUBAME-KFC/DL

Many studies have shown Deep Convolutional Neural Networks (DCNNs) exhibit great accuracies given large training datasets in image recognition tasks. Optimization technique known as mini-batch Stochastic Gradient Descent (SGD) is widely used for deep learning because it gives fast training speed and good recognition accuracies when mini-batch size is set in an appropriate range. We propose a performance model of a distributed DCNN training system called "SPRINT", which uses asynchronous GPU processing based on mini-batch SGD, with considering average mini-batch size that is averaged number of training samples used in a single weight update. Our performance model takes DCNN architecture and machine specifications as input parameters, and predicts time to sweep entire dataset and average mini-batch size with 8% error in average on certain supercomputer. Experimental results on two different supercomputers show that our model can steadily choose the fastest machine configuration that nearly meets a target mini-batch size.

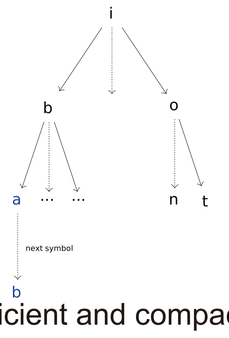
Bibliography
[1] Yosuke Oyama, Akihiro Nomura, Ikuro Sato, Hiroki Nishimura, Yukimasa Tamatsu, and Satoshi Matsuoka, "Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers", in proceedings of 2016 IEEE International Conference on Big Data (IEEE BigData 2016), Washington D.C., Dec. 5-8, 2016 (to appear)

Acknowledgments. This research was partly supported by JST, CREST (Research Area: Advanced Core Technologies for Big Data Integration).

Algorithms and Tools for Computational Linguistics

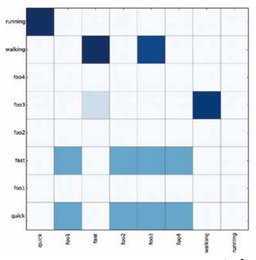
Process billion-words scale corpora in-memory, distributed, fast

Corpus: large collection of texts

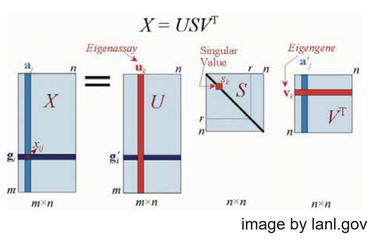


Efficient and compact intra-node data structure based on ternary trees

Distributed implementation with MPI library



Co-occurrence matrix in Compressed Sparse Row format stored to parallel storage in hdf5



Integration with SLEPC high performance distributed sparse linear algebra library for dimensionality reduction

Various machine learning algorithms

Bibliography
[1] A. Drozd, A. Gladkova, and S. Matsuoka, "Word embeddings, analogies, and machine learning: beyond king - man + woman = queen," accepted for COLING 2016.
[2] A. Gladkova, A. Drozd, and S. Matsuoka, "Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.," in Proceedings of the NAACL-HLT SRW, San Diego, CA, June 12-17, 2016, pp. 47–54.
[3] A. Gladkova and A. Drozd, "Intrinsic evaluations of word embeddings: what can we do better?," in Proceedings of The 1st Workshop on Evaluating Vector Space Representations for NLP, Berlin, Germany, 2016, pp. 36–42.
[4] A. Drozd, A. Gladkova, and S. Matsuoka, "Discovering Aspectual Classes of Russian Verbs in Untagged Large Corpora," in Proceedings of 2015 IEEE International Conference on Data Science and Data Intensive Systems (DSDIS), 2015, pp. 61–68.
[5] A. Drozd, A. Gladkova, and S. Matsuoka, "Python, Performance, and Natural Language Processing," in Proceedings of the 5th Workshop on Python for High-Performance and Scientific Computing, New York, NY, USA, 2015, p. 1:1–1:10.

The whole pipeline orchestrated from Python

For more details, please visit http://vsm.blackbird.pw

Acknowledgments. This research was supported by JST, CREST (Research Area: Advanced Core Technologies for Big Data Integration).

http://www.gsic.titech.ac.jp/sc16