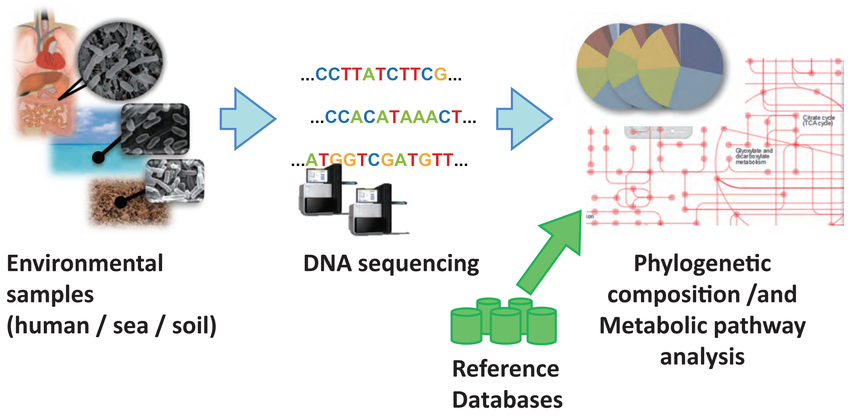
Big Data
Application & Cloud Infrastructure

TOKYO TECH Pursuing Excellence

Ultra-fast Metagenomic Analysis

Emerging Needs

Metagenome analysis

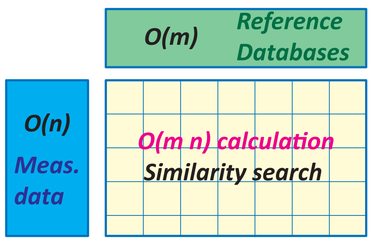


Environmental samples (human / sea / soil)

DNA sequencing

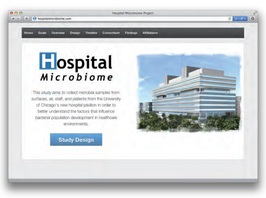
Reference Databases

Phylogenetic composition /and Metabolic pathway analysis

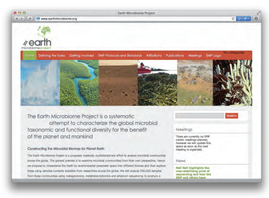


A typical EBD x EBD calculation

Large-scale projects



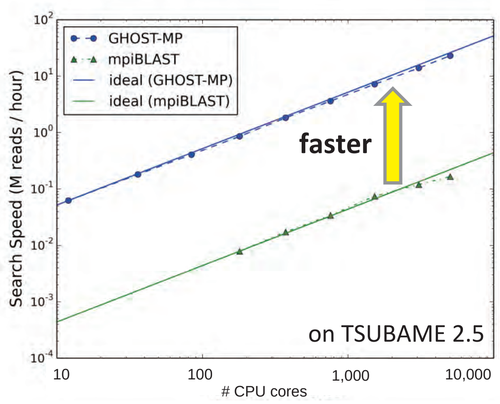
Hospital Microbiome Project



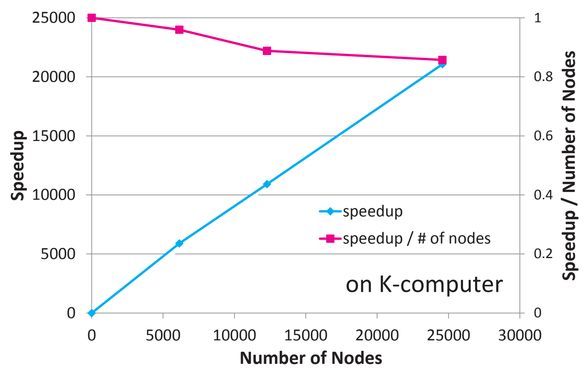
Earth Microbiome Project

Home Microbiome Study

Extreme Performance



GHOST-MP is x96 - x138 faster than mpi-BLAST

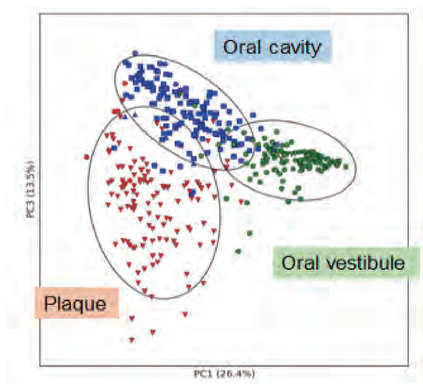


GHOST-MP shows extreme scalability

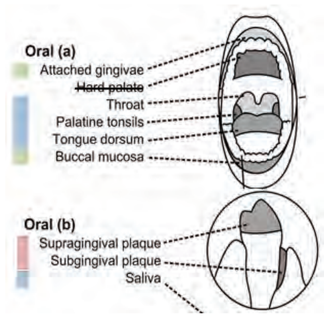
Weak scaling: 0.85 on 24,576 nodes (= 196,608 cores)
85% efficiency compared to 1 node, on K-computer.

Medical Applications

Homology Search for 18 billion sequences: Whole oral microbiome data in HMP DB



Discovery of functional clusters by PCA for relative frequency of KEGG Ontology



Site	Subsite	# of samples	# of reads (x 10^6)	Total file size (GB)
Oral cavity	Saliva	5	278	56
	Keratinized gingiva	6	361	73
	Buccal mucosa	123	7658	1521
	Hard palate	1	54	11
	Palatine tonsils	7	373	74
	Subgingival plaque	8	517	104
	Supragingival plaque	128	7965	1595
	Throat	7	393	79
	Tongue dorsum	137	8815	1765
Total		422	26290	5253

Using GHOST-MP system, we have performed **high-sensitive homology search** for whole human oral microbiome data **(18 billion reads, 5.2TB)** in HMP database.

NIH HUMAN MICROBIOME PROJECT

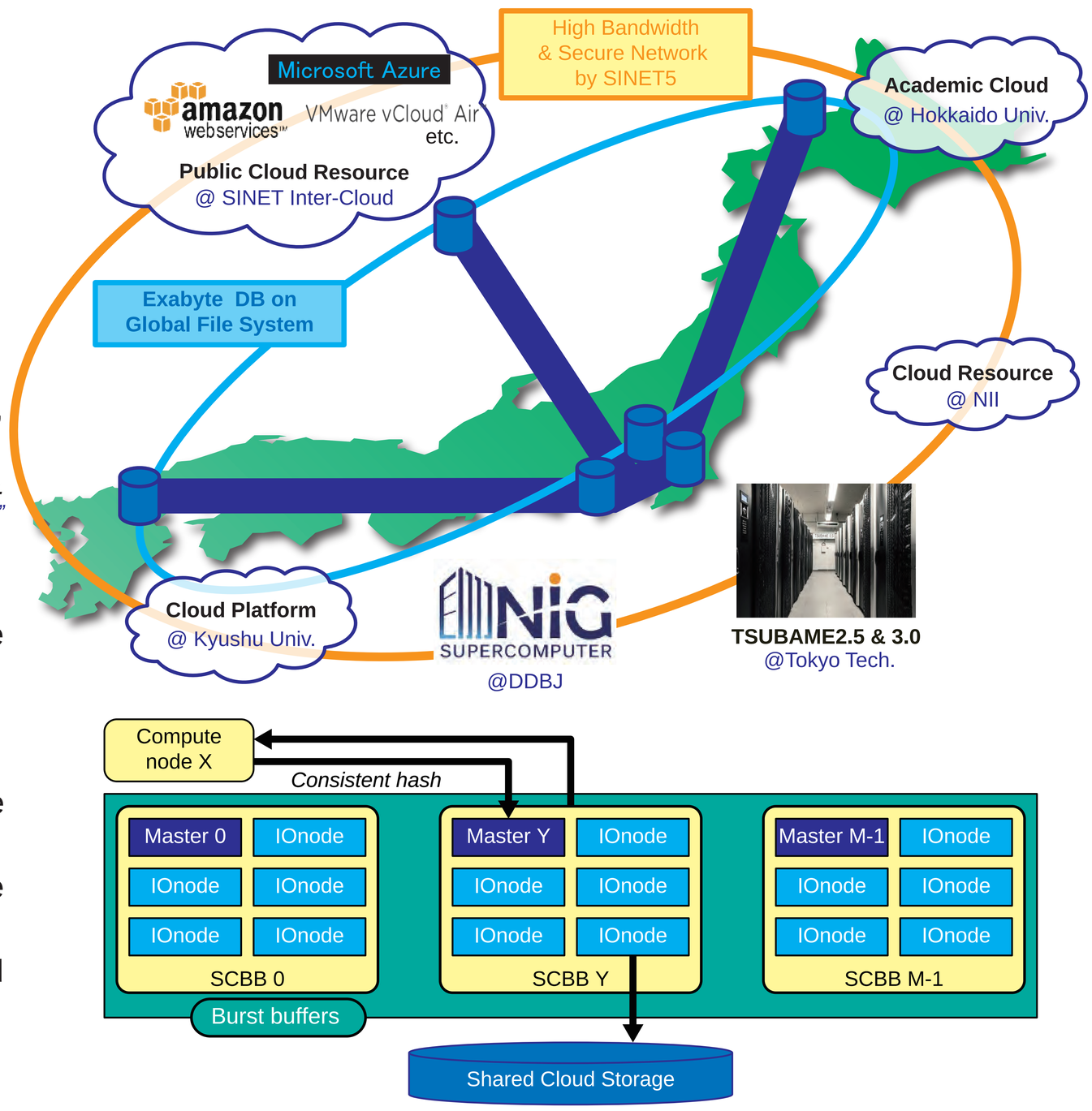
[1] Shuji Suzuki, Masanori Kakuta, Takashi Ishida, Yutaka Akiyama, “GHOSTX: An improved sequence homology search algorithm using a query suffix array and a database suffix array", PLOS ONE, 9(8): e103833, 2014.
[2] Shuji Suzuki, Masanori Kakuta, Takashi Ishida, Yutaka Akiyama, "Faster sequence homology searches by clustering subsequences”, Bioinformatics, 31(8):1183-1190, 2015.
[3] Masanori Kakuta, Shuji Suzuki, Takashi Ishida, Yutaka Akiyama, “A massively parallel sequence similarity search for metagenomic sequencing data", (submitted).

Cloud Infrastructure for Big Data Analysis

Building a Testbed Infrastructure on Overlay Cloud

- Using SINET5 network infrastructure (100Gbps Network)
- Cooperation with various computing resource
  - Private cloud
  - Public cloud
  - Super computer
- Science BigData Repository
  - Testbed: PetaBytes, Real System: ExaBytes
- Stored public scientific database:
  - DNA Sequence DB, Astrophotography, Metrological Data, etc.

Collaboration work with NII, NIG, Hokkaido Univ. and Kyushu Univ.
For more details, Please go to booth #2908, “National Institute of Informatics (NII)”



Cloud-Based Burst Buffer

- Burst buffers are several dedicate nodes to provide remote data cache with high throughput and low latency.
- Our system consists of several SCBBs (Sub CloudBB). In each SCBB, there is a Master and several IOnodes
  - Masters control the IOnodes in the same SCBB, manage file metadata and handle I/O requests from Compute Nodes.
  - IOnodes store actual data and transfer data with Compute Nodes.
- Consistent hash of file path is used to distribute workload among each SCBBs

Poster at SC15:
“Design and Modelling of Cloud-Based Burst Buffers”, Tianqi Xu, Kento Sato, Satoshi Matsuoka.

This research is supported by CREST, JST.

http://www.gsic.titech.ac.jp/sc15