Extreme Big Data: 次世代ビッグデータとスパコンの必然的統合

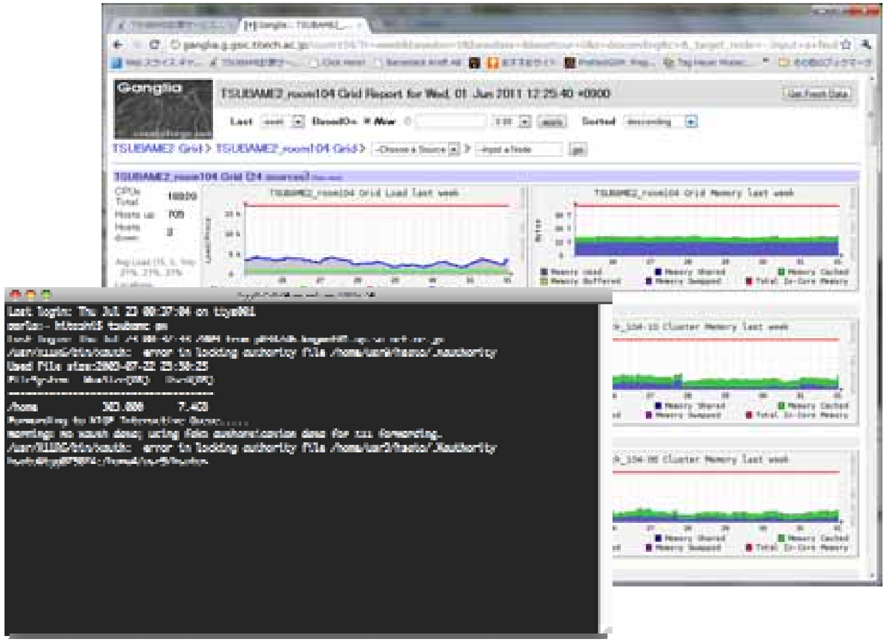
松岡聡・東工大
Satoshi Matsuoka
Tokyo Institute of Technology
ACM / ISC Fellow

統数研講演

20141104

スーパーコンピューター（スパコン）

- 内部の演算処理速度がその時代の一般的なコンピュータより極めて高速な計算機
- 例: 東工大TSUBAME2.0



使うときの見た目は普通のPCとあまり変わらず．．

1957

China Wrests Supercomputer Title From U.S.

By ASHLEE VANCE
Published: October 28, 2010

A Chinese scientific research center has built the fastest supercomputer ever made, replacing the United States as maker of the swiftest machine, and giving China bragging rights as a technology superpower.

Enlarge This Image

The computer, known as Tianhe-1A, has 1.4 times the horsepower of the current top computer, which is at a national laboratory in Tennessee, as measured by the standard test used to gauge how well the systems handle

2010

スパコン世界一奪

議で20日、計算速度を競う世界ランキング「TOP500」が発表され、理化学研究所と富士通が神戸市内で共同開発中の「京」＝写真、同研究所提供＝が、1秒当たり8162兆回の計算能力を示して第1位となった。日本のスパコンが首位に立ったのは、2004年6月の「地球シミュレー

TOP500の上位5機

	スパコンの名称	計算速度
1	京（理化学研究所）	8162
2	天河1A（天津スパコンセンター）	[illegible]
3	ジャガー（オークリッジ国立研究所）	1759
4	星雲（深圳スパコンセンター）	1271
5	TSUBAME2.0（東京工業大）	1192

日本の「京」中国抜

"Reclaimed No.1 Supercompu[illegible] Rank in the World"

2011

U.S. reclaims top spot in supercomputer race

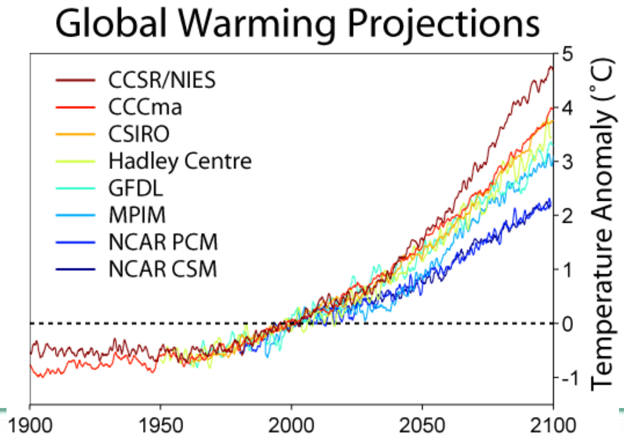
By Erin Kim @CNNMoneyTech June 18, 2012: 3:34 PM ET

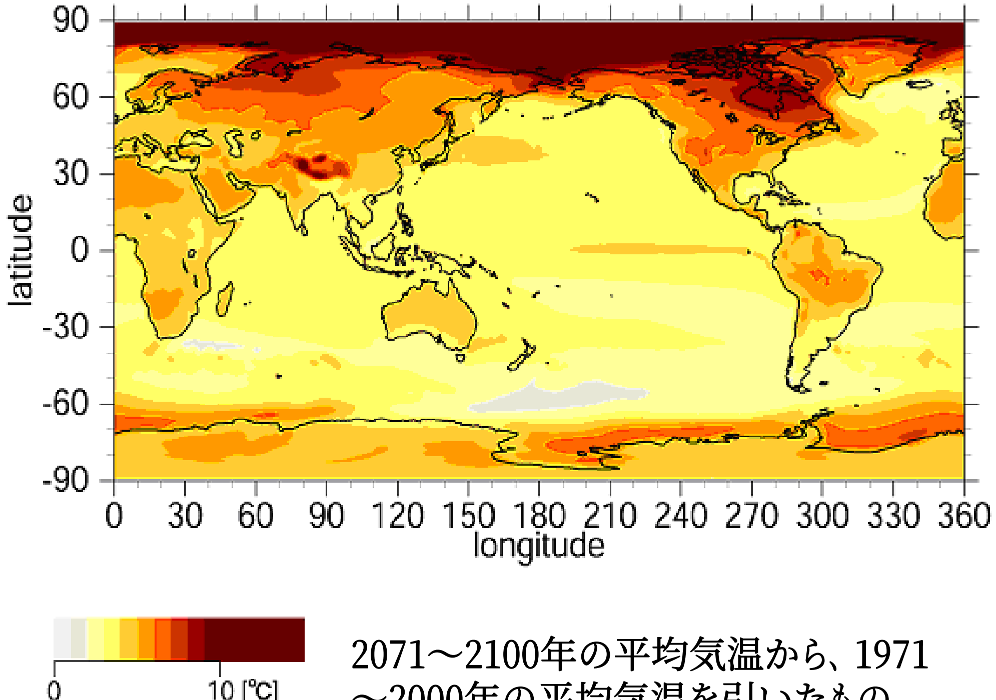
2012

スパコンのシミュレーションは未来を予測する技術

地球温暖化問題、CO_2の増加が地球環境に重大な影響を与えることを人々に認識させたのは、スパコンによるシミュレーションと可視化である

Al Gore's Keynote Presentation at SC09



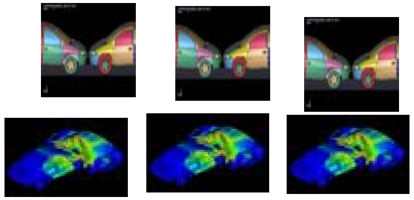


2071～2100年の平均気温から、1971～2000年の平均気温を引いたもの

ペタスケールのスパコンで何が実現できるのか?

大量の計算

→ 入力データを変えた計算を大量に実行する事により、新しい知見を得る



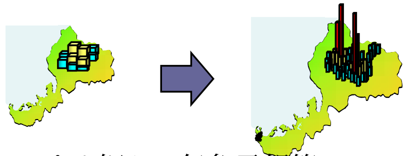
最適な設計パラメータ探索

→ Drug Design 候補を選ぶ

精密な計算

正確に方程式を解く

→ より精密なシミュレーション

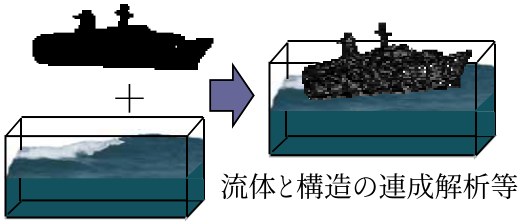


きめ細かい気象予報等

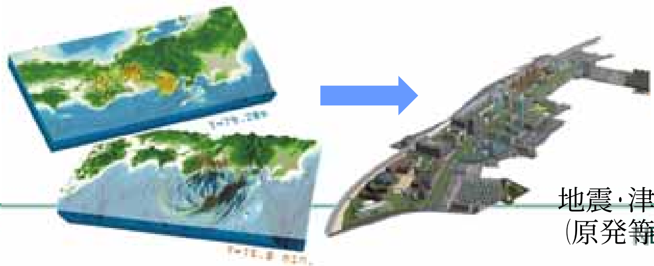
計算機能力の拡大

計算の組合せ

→ 異なる種類の計算を組み合わせる事により、より現実の世界に近い解析を



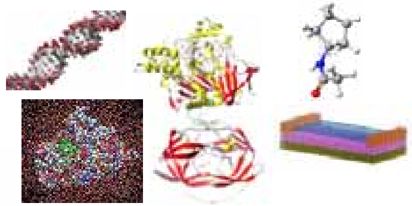
流体と構造の連成解析等



地震・津波による構造物（原発等）の被害予測

複雑な計算

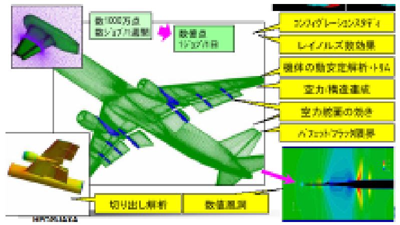
→ より精密・正確な計算手法により、より現実の世界に近い解析を



蛋白質の相互作用解析等

大規模な計算

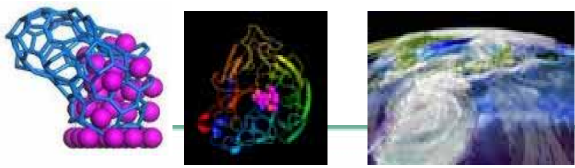
→ より大規模な計算モデルにより、より現実の世界に近い解析を



航空機まるごと解析等

©JAXA

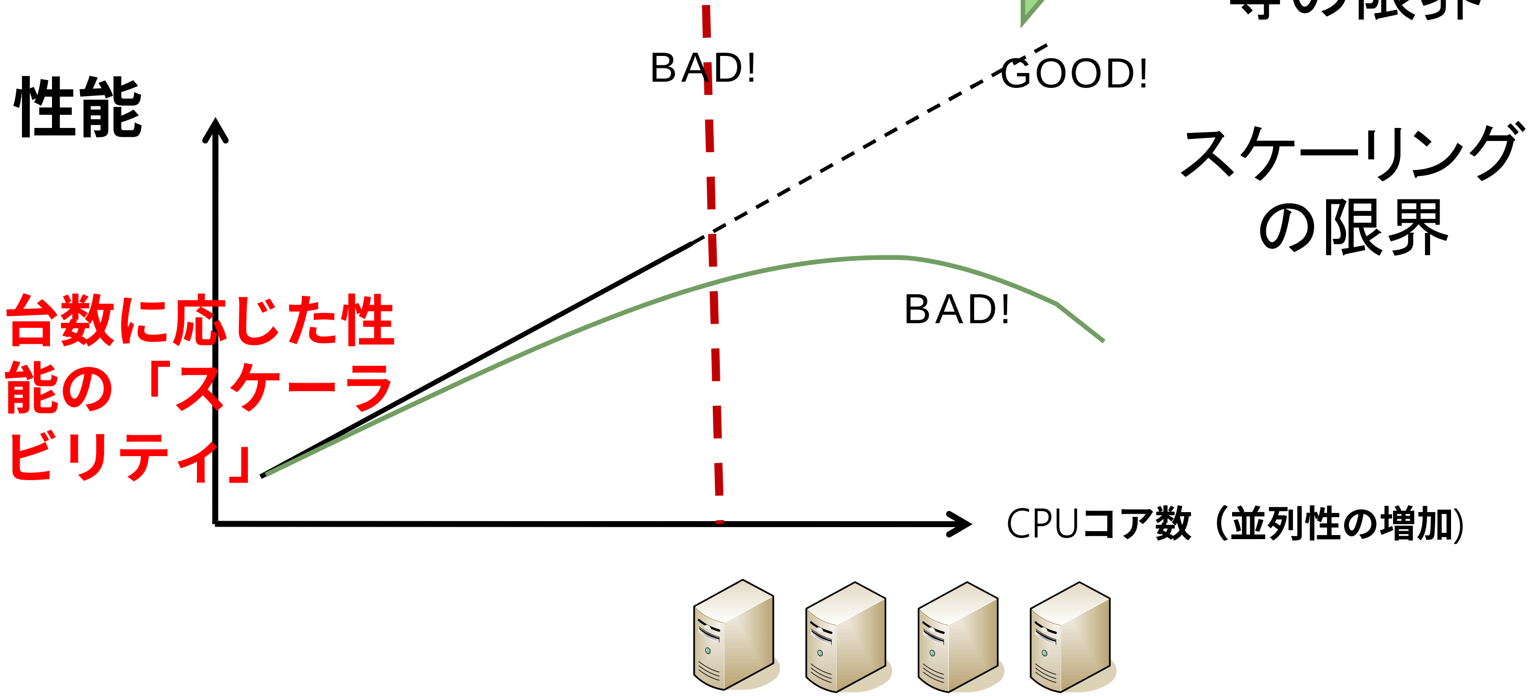
→ より長時間の時間発展



スケーラビリティと(超)並列性

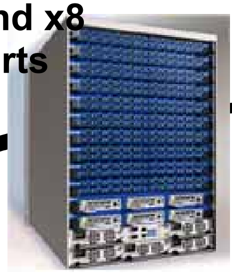
- マシンの台数を増やす⇒並列性の増加
    - ◆より早く結果を得るシミュレーション
    - ◆より大きなシミュレーション
    - ◆質的に違うシミュレーション

電力・コスト等の限界

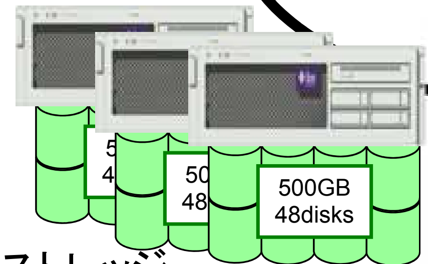


2006年4月東工大スパコン ”TSUBAME1.0” クラスタ・グリッド研究の集大成

Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)



10Gbps+外部
ネットワーク



ストレッジ
1 Petabyte (Sun “Thumper”)
0.1Petabyte (NEC iStore)
Lustre ファイルシステム
>400Gbps

2006年6月
アジア No.1, 世界No.7
38.18Teraflops
(Top500計測値)

Sun/AMD高性能計算クラスタ
(Opteron Dual core 8-Way)
10480core/655ノード
50.4TeraFlops
OS(現状) Linux
(検討中) Solaris, Windows
NAREGIグリッドミドル

ClearSpeed CSX600
SIMD accelerator
360 boards,
30TeraFlops

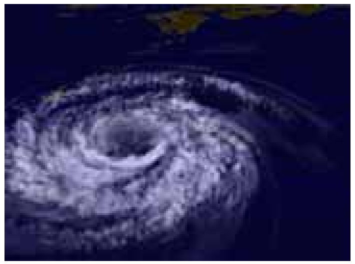
TSUBAMEの4年の運用成果:全目標達成

1. 東工大のシンボル:世界トップレベルの情報インフラ

多数の内外報道・訪問者

＊Top500＊
4期連続日本一
4期連続性能向上
(世界初)

2. 研究推進:莫大な計算パワー・ストレッジ(1ペタバイト以上)・みんなのスパコン



3. 産学連携等の推進、大型プロジェクトへの呼び水、アライアンスを組む他大学計算ニーズホスティング

MITSUBISHI CHEMICAL / NUMERICAL TECHNOLOGIES / SUMITOMO CHEMICAL

文部科学省先端研究施設共用イノベーション創出事業【産業戦略利用】

CompView Global COE「計算世界観」

AMD / NEC / Sun microsystems / Microsoft

企業との包括collaboration

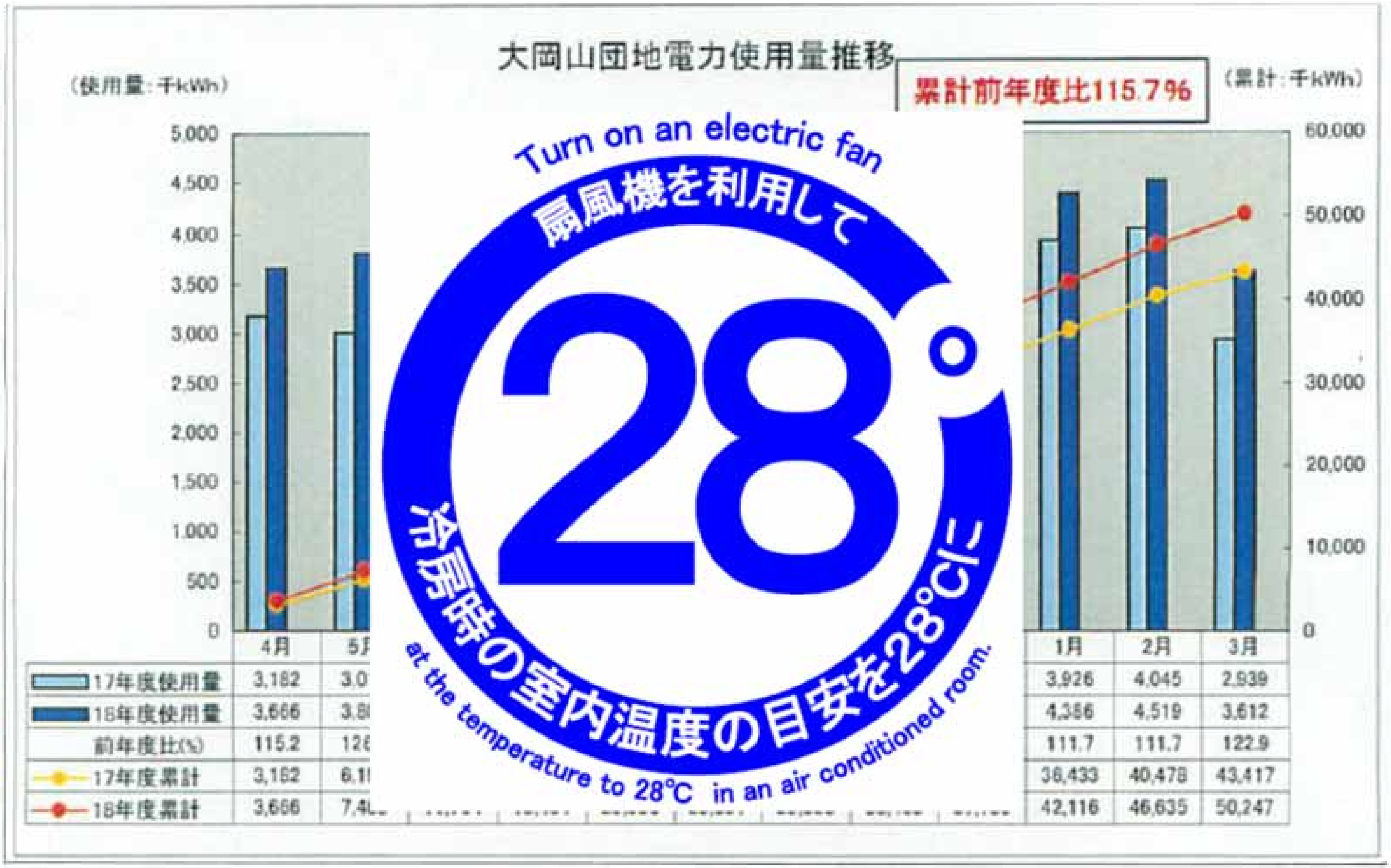
Research Organization of Information and Systems 国立情報学研究所 National Institute of Informatics / NAREGI

NAREGI/NII-CSI
全国サイバーサイエンスインフラ
＊NAREGI 開発への貢献
＊阪大−東工大 NAREGI β2連携

4. 学内の分散した情報基盤の集約化・ホスティング

·TSUBAME「みんなのスパコン」
- ·新概念の課金利用法によるユーザ数増加 => 1300人へ倍増
- · SE運用業務の追加(アプリ・性能評価・グリッド試験運用など)
- ·各種ITサービスのホスティング

現代のスパコンは電力が問題

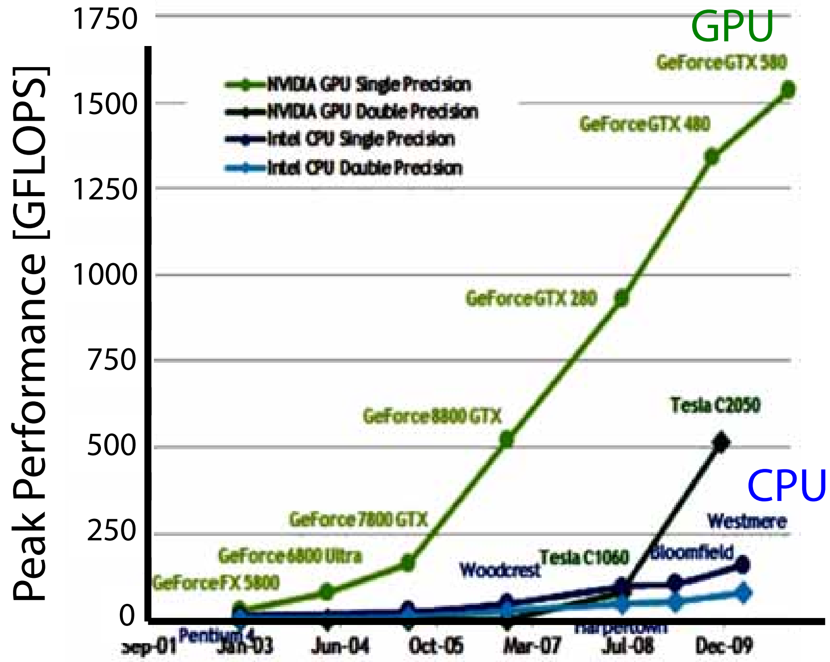


TSUBAME1.0 2006年6月
アジア No.1, 世界7位
38.18Teraflops(Top500)
運用電力1MW

2010年TSUBAME2.0
ペタフロップスの性能目標
<u>電力一定→25倍の電力性能向上</u>

CPU対GPUの性能比較

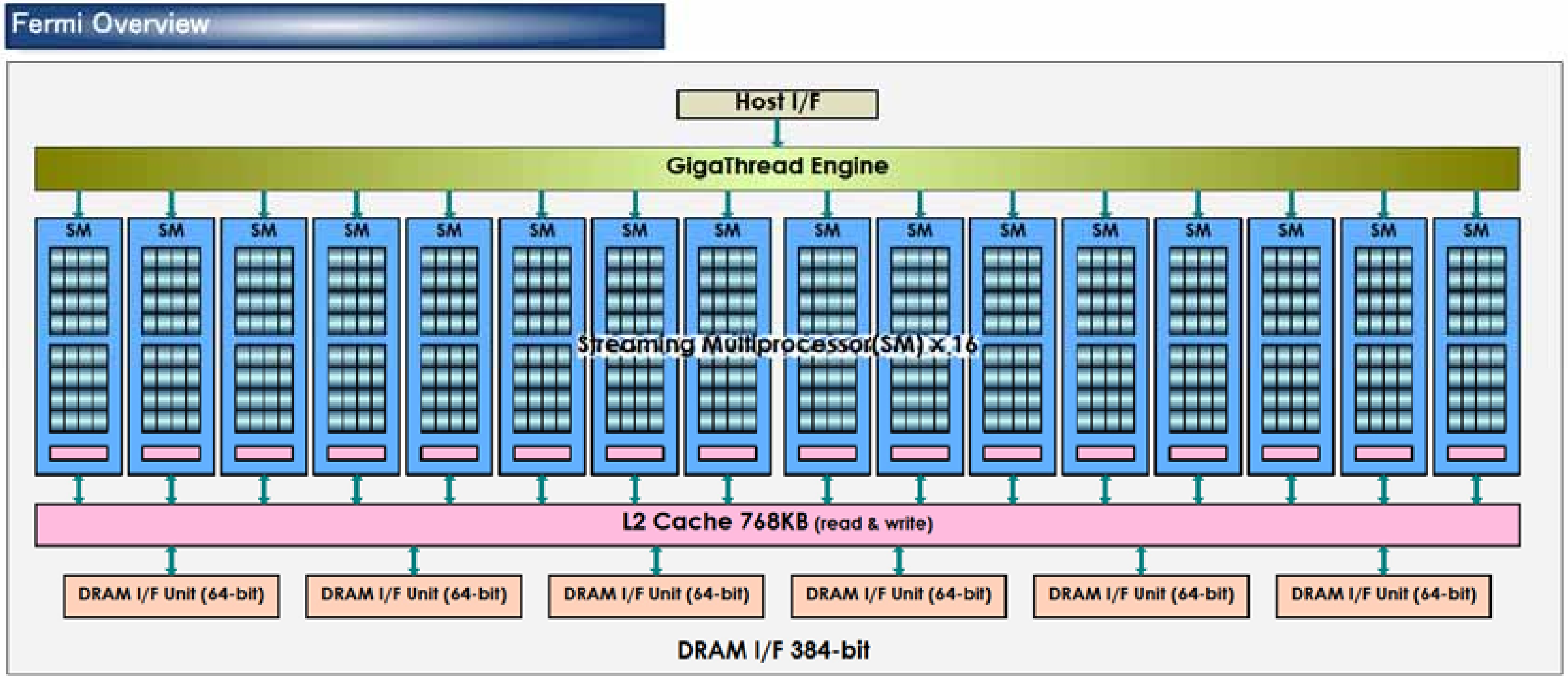
JST-CREST ULP-HPCプロジェクト等での電力性能評価



計算性能・メモリバンド幅とも 5-6倍
システム実質電力はほぼ同一
TSUBAME1.2にて運用テスト

GPUは「メニーコアプロセッサ」: 数百の処理コア

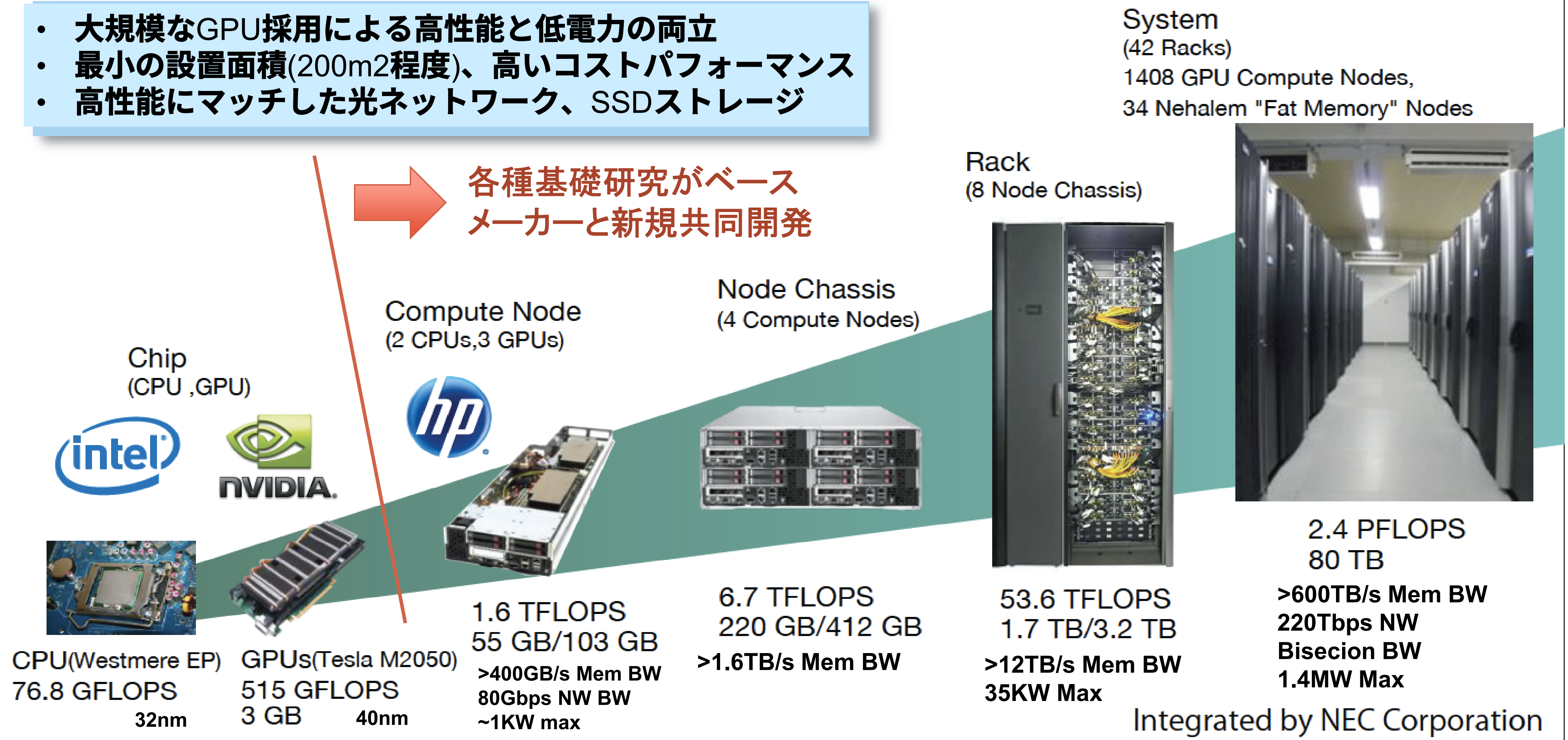
Many Core, Multhreaded, SIMD-Vector, MIMD Parallel Architecture



(Figure by Kazushige Goto)

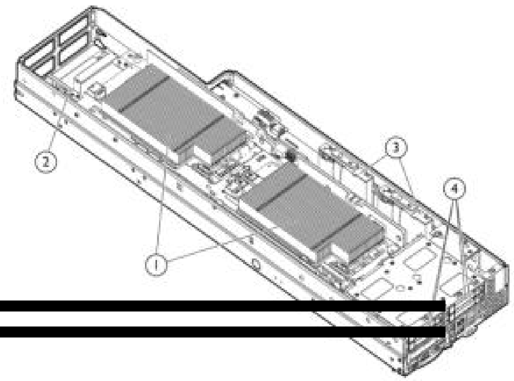
TSUBAME2.0 2010年11月1日稼働開始
世界最小のペタフロップス･省電力スパコン

- 大規模なGPU採用による高性能と低電力の両立
- 最小の設置面積(200m2程度)、高いコストパフォーマンス
- 高性能にマッチした光ネットワーク、SSDストレージ



TSUBAME2.0 計算ノードの詳細

Thin 計算ノード



Infiniband QDR x2 (80Gbps)

1.6 Tflops
400GB/s
Mem BW
80GBps
NW
~1KW max

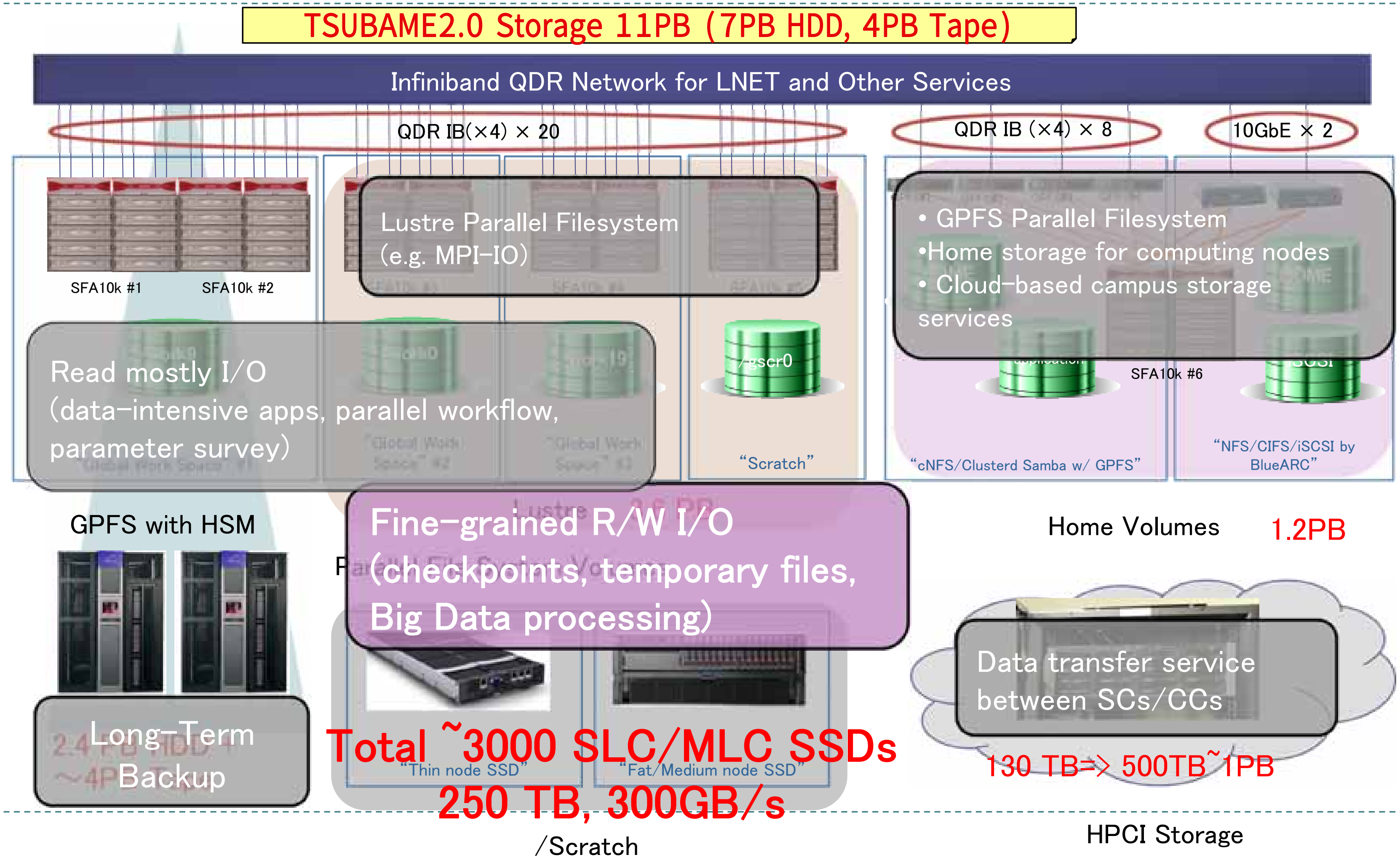
システム合算性能
2.4PFlops
メモリ: ~100TB
SSD: ~200TB
HDD: ~7PB
光ネットワーク: ~200Tbps
運用電力: 1MW以下

TSUBAME 2.0用にHP社と共同開発

GPU: NVIDIA Fermi M2050 x 3
515GFlops, 3GByte memory /GPU
CPU: Intel Westmere-EP 2.93GHz x2 (12cores/node)
Multi I/O chips, 72 PCI-e lanes --- 3GPUs + 2 IB QDR
Memory: 54, 96 GB DDR3-1333
SSD:60GBx2, 120GBx2


TSUBAMEの光ネットワーク
Oversubscribed Full Fat-Tree
3500 Fiber Cables > 100Km
w/DFB Silicon Photonics
End-to-End 7.5GB/s, ~2us
Non-Blocking 220Tbps Bisection

TSUBAME2.0ストレージ：スパコンとしては世界初の大規模シリコンストレージ(SSD)採用→高速性と低コスト・低電力(サーバ数大幅減少)の両立




11ペタ(10^{15})バイト複
合型のストレージ
HDD 4000台
SSD 3000台

2010年11月世界トップランクスパコンへ

- Green500省エネ性能 958MFlops/W ⇒世界3位!!
  - Greenest Production Supercomputer in the World賞
- 演算性能1.192PFlops ⇒ 世界4位!!
  - 「京」登場後も世界5位を一年間維持

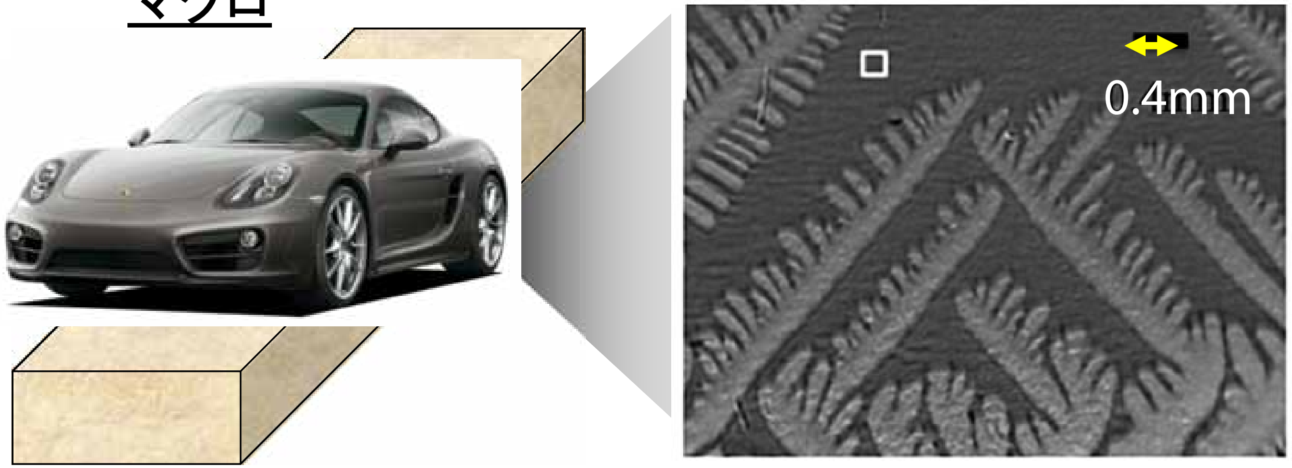
TOP500 SUPERCOMPUTER SITES

Rank	Site	Computer/Year Vendor				
1	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C / 2010 NUDT	186368	2566.00	4701.00	4040.00
2	DOE/SC/Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
3	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU / 2010 Dawning	120640	1271.00	2984.30	2580.00
4	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.61
5	DOE/SC/LBNL/NERSC United States	Hopper - Cray XE6 12-core 2.1 GHz / 2010 Cray Inc.	153408	1054.00	1288.63	2910.00
6	Commissariat a l'Energie Atomique (CEA) France	Tera-100 - Bull bullx super-node S6010/S6030 / 2010 Bull SA	138368	1050.00	1254.55	4590.00
7	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.50

金属材料の機械的強度のシミュレーション [下川辺、青木ら]

マクロ

ミクロ(~μm)



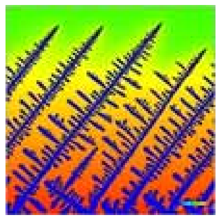
マクロな特性は
ミクロ構造に依存

Spring-8 による実際の観察写真

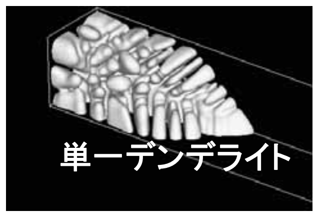
金属材料の機械的強度や
特性の予測

従来の計算では
意味のある規模の
構造のシミュレーションが
不可能

2次元計算



ミクロな組織構造に基づく
大規模シミュレーションが必要



Spring-8 (http://user.spring8.or.jp/sp8info/?p=17393)

TSUBAME2.0のアプリケーションの受賞


SC11

ACM Gordon Bell Prize

Special Achievements in Scalability and Time-to-Solution

Takashi Shimokawabe, Takayuki Aoki,
Tomohiro Takaki, Akinori Yamanaka,
Akira Nukada, Toshio Endo,
Naoya Maruyama, Satoshi Matsuoka

Peta-Scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer

Scott Lathrop
SC11 Conference Chair

Thom H. Dunning, Jr.
Gordon Bell Chair

ACM Gordon Bell Prize 2011
ACM ゴードンベル賞
2.0ペタフロップス達成
(京コンピュータと同時受賞)

Special Achievements in Scalability and Time-to-Solution

"Peta-Scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer"

TSUBAMEによるスパコン応用分野とIT技術の共生的イノベーション

2000人以上の「スパコンユーザ」とアプリ

以下の「国民の高い関心事」であるアプリも多々

1. 環境・防災 Disaster & Environment
2. 医療・創薬 Medical & Pharmaceutical
3. ものづくり・素材 Manufacturing & Materials

に加えて、TSUBAMEの研究開発によりスパコンに限らず、IDCやビッグデータなど、現代のIT産業へ対するインパクトを伴うコデザイン

先進的なLattice-Boltzmann法による東京全域の気流シミュレーション[小野寺・青木]

東京の10km四方領域を<u>1m解像度</u>でモデル化

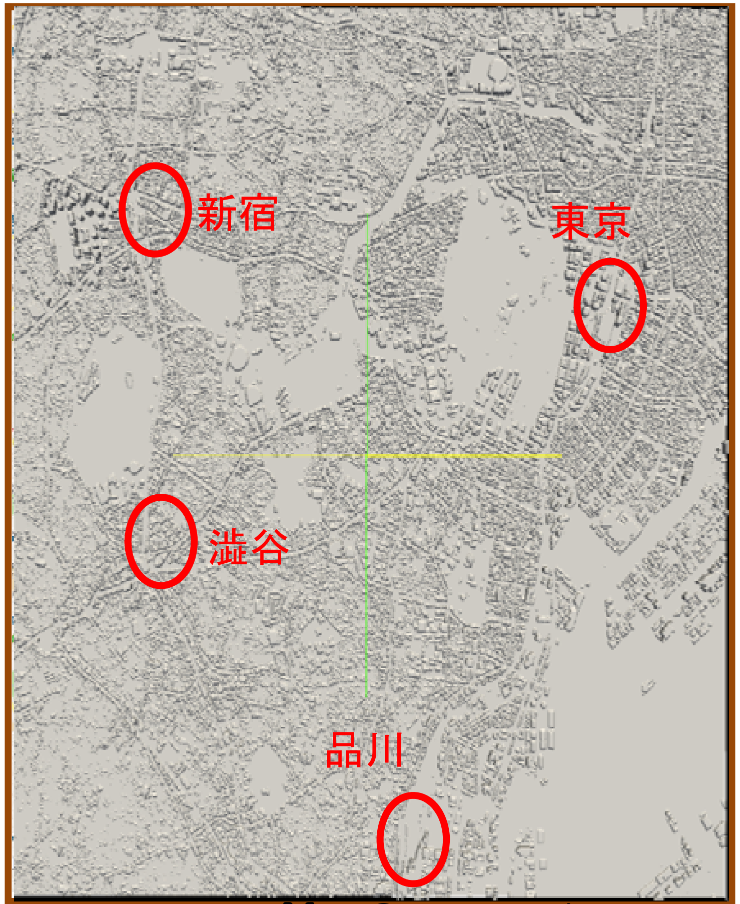
地図データ: ゼンリン、**Google**

建物データ: Pasco Co. Ltd.TDM 3D

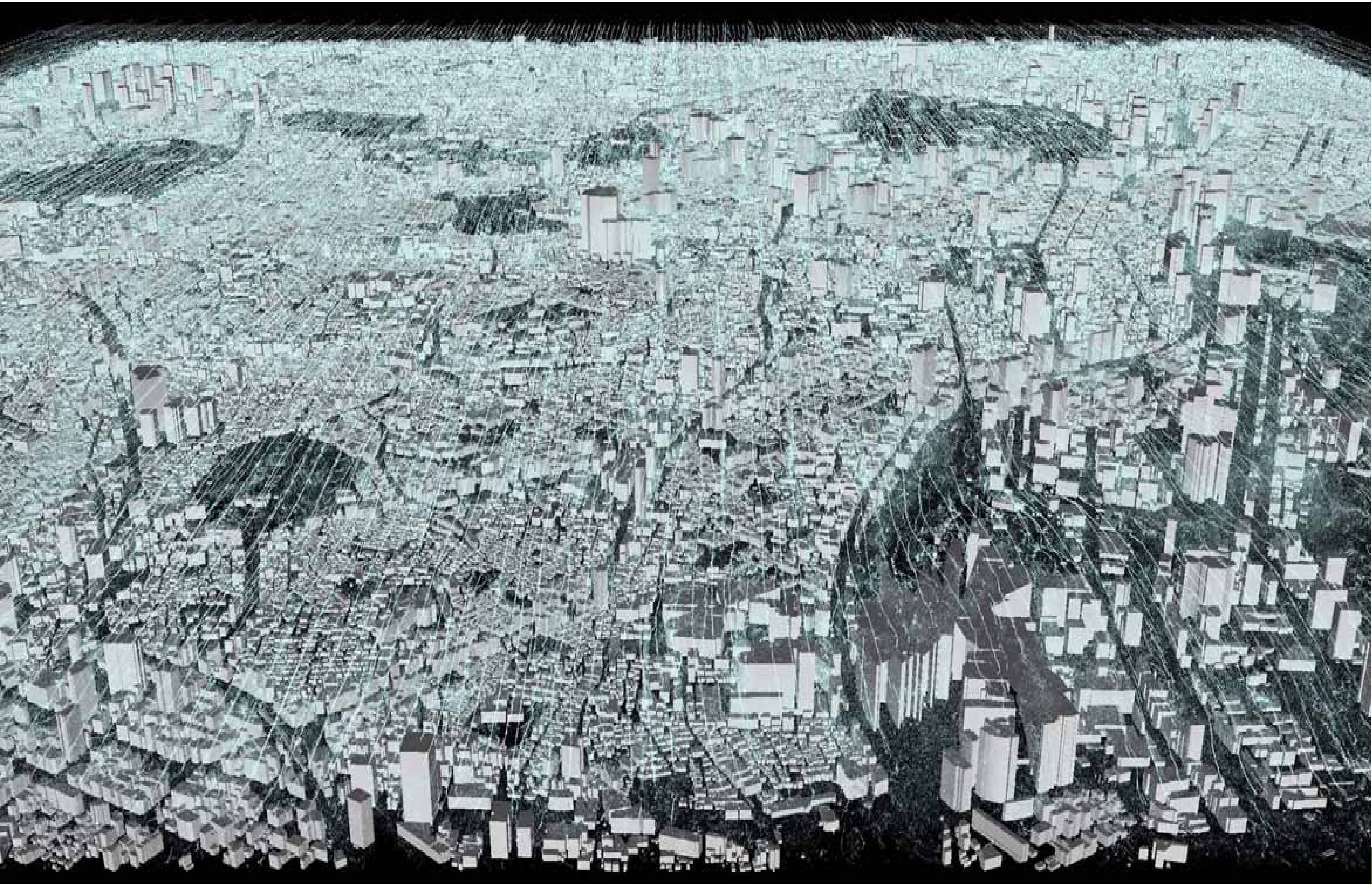
乱流を有効に扱えるCoherent Structured SGS モデルを用いた先進的Lattice-Boltzmann法で超並列計算

TSUBAME2.0全系の4000GPUを用い、0.592 Petaflopsを達成 (効率15%)

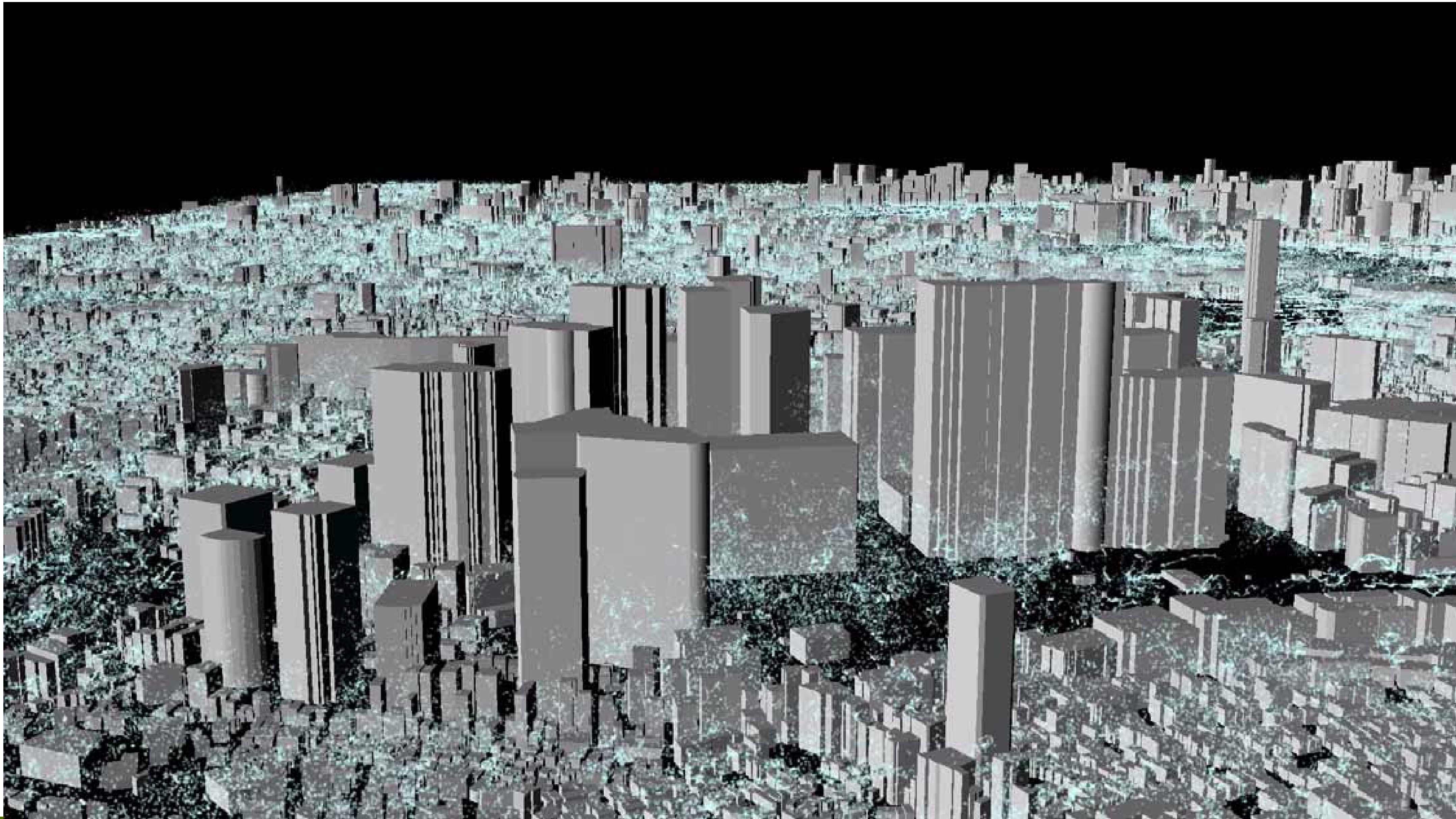
ヒートアイランド現象の解明、汚染物質の拡散、高層ビル建築時の都市計画シミュレーションなど、多くの用途

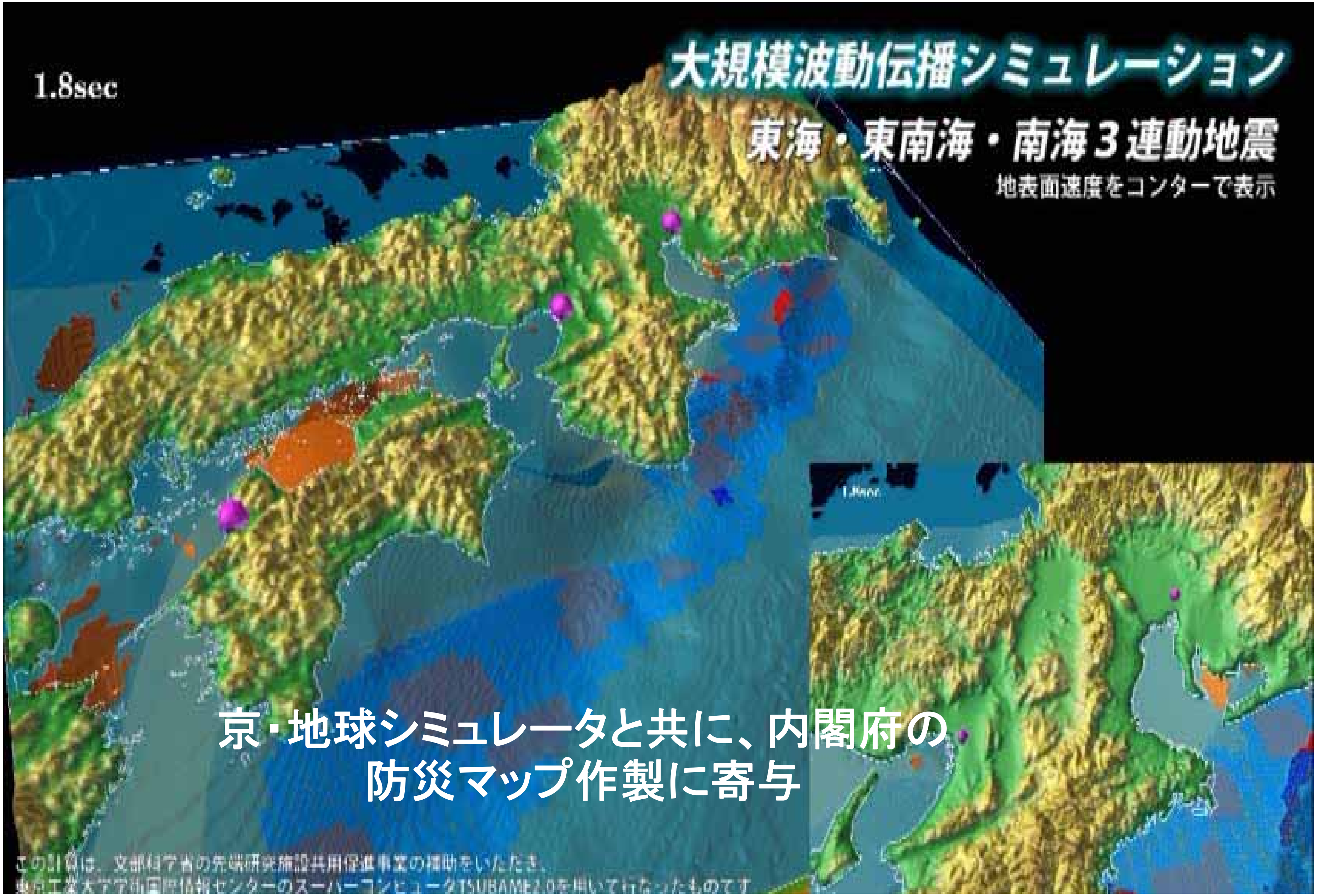


Map ©2012 Google, ZENRIN

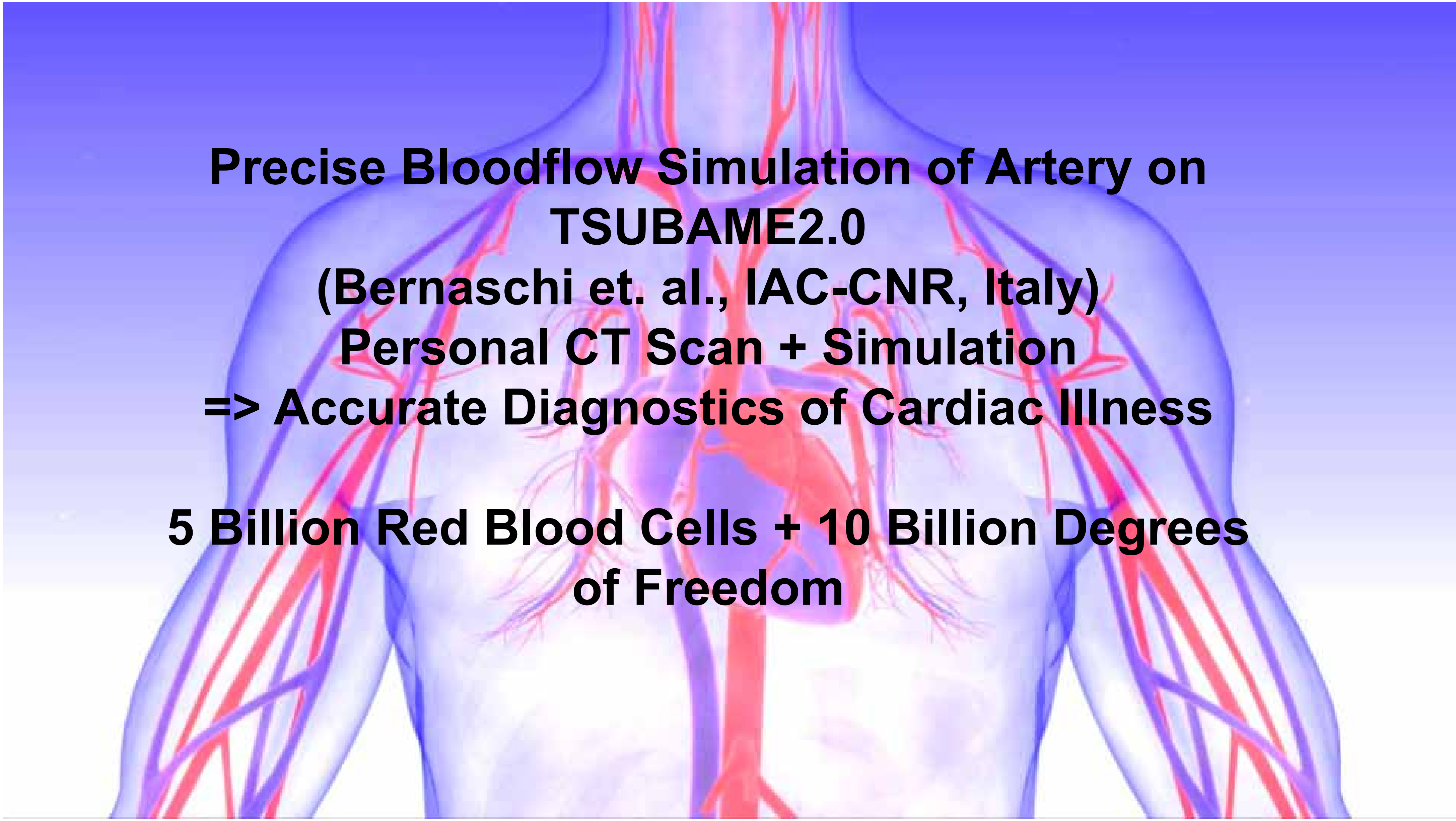


AOKI Lab.




1.8sec
大規模波動伝播シミュレーション
東海・東南海・南海3連動地震
地表面速度をコンターで表示
京・地球シミュレータと共に、内閣府の
防災マップ作製に寄与
この計算は、文部科学省の先端研究施設共用促進事業の補助をいただき、
東京工業大学学術国際情報センターのスーパーコンピュータTSUBAME2.0を用いておこなったものです

産業利用TOTO株式会社[池端]



Precise Bloodflow Simulation of Artery on TSUBAME2.0

(Bernaschi et. al., IAC-CNR, Italy)

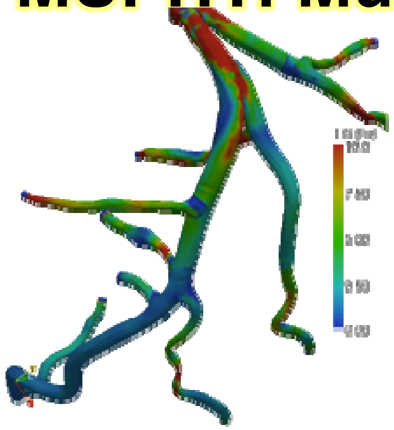
Personal CT Scan + Simulation

=> Accurate Diagnostics of Cardiac Illness

5 Billion Red Blood Cells + 10 Billion Degrees of Freedom

MUPHY: Multiphysics simulation of blood flow

(Melchionna, Bernaschi et al.)



Multiphyics simulation with *MUPHY* software

Combined Lattice-Boltzmann (LB) simulation for plasma and Molecular Dynamics (MD) for Red Blood Cells

Realistic geometry (from CAT scan)

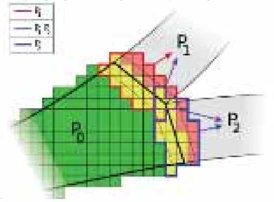
Two-levels of parallelism: CUDA (on GPU) + MPI

• 1 Billion mesh node for LB component

•100 Million RBCs

4000 GPUs, 0.6Petaflops

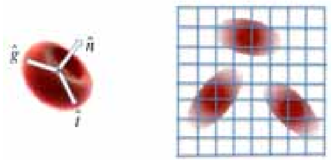
Fluid: Blood plasma → Lattice Boltzmann



Irregular mesh is divided by using PT-SCOTCH tool, considering cutoff distance

coupled

Body: Red blood cell → Extended MD



Red blood cells (RBCs) are represented as ellipsoidal particles

ACM Gordon Bell Prize 2011 Honorable Mention

顧みられない熱帯病の治療薬探索

astellas
Leading Light for Life

TOKYO TECH
Pursuing Excellence

関嶋准教授ら

リーシュマニア症、シャーガス病、
アフリカ睡眠病を引き起こす
寄生原虫の治療薬探索を
アステラス製薬と共同で実施

TSUBAME2.0を用いることで

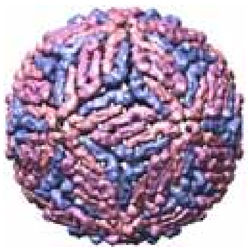
- データマイニング（ターゲットタンパク質探索）
- インシリコスクリーニング（薬候補化合物探索）

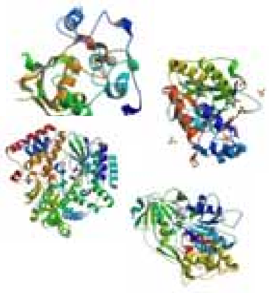
のプロセスで創薬を効率化

リーシュマニア症
（蚊が媒介する
寄生原虫が原因）

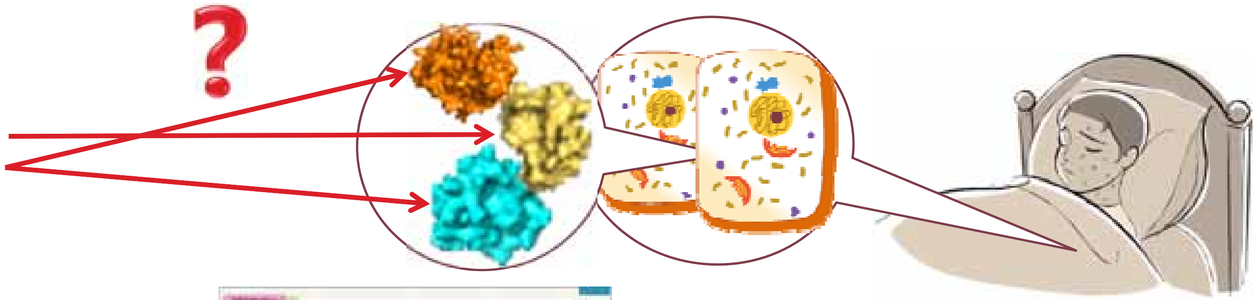
GPUを用いた超高速タンパク質ドッキングによるデング熱特効薬の開発

デング熱の酵素





人間全てのタンパク質



タンパク質名	タンパク質構造 (PDB ID)
Protease	3U1I
Methyltransferase	1R6A
Polymerase	3VWS
Helicase	2JLR

Human protein structures were collected from the public database PDB using the following criteria:

- ✓ >25 residues
- ✓ X-ray resolution better than 3.25 Å
- ✓ No mutation

#Structures (PDB-chains)	30,544
#Proteins (UniProt IDs)	3,353

June 15, 2013

$4 \times 30{,}544 = 122{,}176$ ドッキング候補

アステラス製薬とのデング熱等の熱帯病の特効薬の創薬

Business Wire
A Berkshire Hathaway Company

HOME ▾ SERVICES ▾ NEWS ▾ EDUCATION ▾ ABOUT US ▾ Search Log In Sign Up

astellas
Leading Light for Life

TOKYO TECH
Pursuing Excellence

いいね！ Tweet 3 Share 0

Release Versions

- English
- Chinese
- EON: Enhanced Online News

Company Information Center

ASTELLAS PHARMA INC.

TOKYO:4503

March 21, 2013 03:09 AM Eastern Daylight Time

Tokyo Institute of Technology and Astellas Launch Collaborative Research for New Anti-Dengue Virus Drugs for Neglected Tropical Diseases

- IT drug-discovery research through use of Tokyo Tech's Supercomp

TOKYO--(BUSINESS WIRE)--Tokyo Institute of Technology ("Tokyo Tech"; Tokyo, Japan; I
Astellas Pharma Inc. ("Astellas")(TOKYO:4503)(President and CEO: Yoshihiko Hatanaka) t
a joint research agreement for drug discovery research utilizing Tokyo Tech's TSUBAME2.0
candidates for the treatment of neglected tropical diseases ("NTDs") caused by dengue viru

NTDs, prevalent mainly among the poor in tropical areas of developing countries, are infectic
bacteria. As it is estimated that approximately one billion people are affected with NTDs wor
healthcare issue that is being addressed on a global scale. Among them, diseases caused b
fever/dengue hemorrhagic fever are with high unmet medical needs for treatment and develc
There is no existing drug to treat dengue fever/dengue hemorrhagic fever in the market as w
effectiveness of some vaccines to prevent dengue virus currently under development is uncle

Under the collaborative agreement, Tokyo Tech which has cutting-edge computation techniq

---

熱帯病薬

スパコンで探索研究

情報収集、3次元構造解析

アステラス 東工大

アステラス製薬は30日、東京工業大学のスーパーコンピューター「TSUBAME2・0」を活用して熱帯病治療薬の探索研究を実施すると発表した。アステラスは、東工大と2010年から スパコン活用の創薬技術確立を進めてきた。これまでの研究成果も使いながら、治療薬が限られている熱帯病分野で新薬を見つけ出す。

アステラスと東工大学術国際情報センターの関嶋政和准教授らの研究グループが熱帯病の一種であるリーシュマニア症、シャーガス病、アフリカ睡眠病の治療薬創出を進める。具体的にはスパコンを活用して特許や文献などの公開情報を網羅的に調べ、有用な情報を抽出。そこから得た情報を活用しながら、標的分子の3次元構造をスパコンで解析し、活性を示す化合物を試薬など市販化合物から選び出す。取得した化合物の商業化のスキームは未定としている。

東工大が運営するTSUBAME2・0は、世界で最も高速なスパコンの1つ。アステラスは効率的な創薬技術の確立を目指し、10年から東工大と共同研究を開始し、IT創薬における新規計算技術の開発、プログラムの高速化に取り組んできた。これまでの共同研究で得た成果なども判断し、熱帯病薬でもスパコン活用の創薬研究を推し進めることにした。

アステラス薬 と 東工大

スパコンで創薬研究

寄生原虫治療薬に着手

アステラス製薬と東京工業大学は30日、東工大のスーパーコンピューター「TSUBAME（ツバメ）2・0」（写真）を使い、寄生原虫治療薬の創薬研究を共同で始めると発表した。計算速度が世界トップレベルのスパコンを活用、候補となる化合物の探索など高速化できると期待する。製薬業界は研究開発の期間が10年以上と長く、効率化が課題。第一三共や中外製薬などがスパコンを使った創薬の効率化に着手している。

創薬研究の対象はリーシュマニア症、シャーガス病といった寄生虫や細菌による感染症。世界保健機関（WHO）が発展途上国などで流行する病気として重視する17の熱帯病の中でも効果的な治療薬がなく、創薬ニーズが高いとされる。

まずスパコンを使って特許や文献など関係するデータを見つけ出すデータマイニングにより有用な知識を探索。その後、活性のある化合物を高速計算で探す。スパコンの稼働率が下がる夏休み期間を利用するため、なるべく早く着手する。

アステラス薬と東工大学術国際情報センターの関嶋政和准教授らは、2010年4月から12年3月まで共同でツバメ2・0を使った創薬の新規計算技術とプログラム高速化の研究を進めてきた。このノウハウも活用して効率的な創薬ができるのではないかと期待する。

日本で実際の創薬に大規模スパコンが用いられるのはTSUBAMEが初

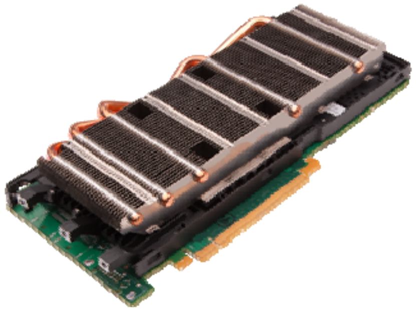
2013年9月:HPCI補正予算により TSUBAME2.0 => 2.5に進化

- 性能が 2~3倍に
  - 理論性能2.4(倍精度)/4.8(単精度) Petaflops => 5.76(x 2.4)/17.1(x3.6)
- 高速GPUメモリの速度向上・容量倍化
  - GPUあたり3GB=>6GB, バンド幅 150GB/s => 250GB/s
- 高信頼化
  - GPUとPCI-eが絡んだマイナーなハードウェアバグを解消、ノードダウンの解消
- 低電力化
  - 約10~20%の運用電力削減
- より高機能なGPUのプログラミング機能
  - Dynamic tasks, HyperQ, CPU/GPU shared memory
- TSUBAME2 の寿命を1-2年延長
  - TSUBAME3.0 2014年11月 => 2016年4月以降に

TSUBAME2.0⇒2.5 計算ノードの進化

- 全4224GPUを最新のKepler GPUに交換
- 幾つかの技術上・運用上の問題をメーカーと共同で克服
- 低コスト・短期間でマシンの能力を2-3倍に向上に成功

NVIDIA Fermi
M2050
1039/515GFlops
3GBメモリ



NVIDIA Kepler
K20X
3950/1310GFlops
6GBメモリ

アプリケーション名 性能値	TSUBAME2.0 性能	TSUBAME2.5 性能	速度向上比
Top500/Linpack 4131 GPUs (PFlops)	1.192	2.843	2.39
Green500/Linpack 4131 GPUs (GFlops/W)	0.958	3.068	3.20
Semi-Definite Nonlinear Programming 4080 GPUs (PFlops)	1.019	1.713	1.68
Gordon Bell Dendrite Stencil 3968 GPUs (PFlops)	2.000	3.444	1.72
LBM LES Whole City Airflow 3968 GPUs (PFlops)	0.592	1.142	1.93
Amber 12 pmemd 4 nodes 8 GPUs (nsec/day)	3.44	11.39	3.31
GHOSTM Genome Homology Search 1 GPU (Sec)	19361	10785	1.80
MEGADOC Protein Docking 1 node 3GPUs (vs. 1CPU core)	37.11	83.49	2.25

TSUBAMEと京との比較(1)

東京工業大学
Tokyo Institute of Technology
GSIC

TSUBAME2.0(2010)
→ TSUBAME2.5(2013)
単精度17.1 Petaflops(最速)
倍精度5.76 Petaflops
約50億円/6年 (電気代等含)

性能≒
コスト<<

独立行政法人理化学研究所
計算科学研究機構
RIKEN RIKEN Advanced Institute for Computational Science

京コンピュータ (2011)

単精度11.4Petaflops
倍精度11.4Petaflops(最速)

約1500億円?/6年
(電気代等含)

TSUBAMEと京との比較(2)
(TSUBAME2は時代の最先端技術採用)

	TSUBAME2.5	BG/Q Sequoia	K Computer
単精度ピーク性能	17.1 Petaflops	20.1 Petaflops	11.3 Petaflops
Green500 (MFLOPS/W)	**3,068.71 (6th)**	**2,176.58 (26th)**	**830.18 (123rd)**
運用時電力(冷却含む)	~1MW	5~6MW?	> 10MW
ハードウェアアーキテク	Many-Core (GPU) + Multi-Core Heterogeneous	Multi-Core Homo	Multi-Core Homo
最大スレッド数	1億以上	6百万	70万
メモリ技術	GDDR5+DDR3	DDR3	DDR3
ネットワーク技術	Luxtera シリコンフォト光	通常光	銅線
不揮発性メモリ/ SSD	全ノードフラッシュSSD, ~250TBytes	なし	なし
電力制御・アクティブ電力キャップ	ノード・CPU/GPUおよびの電力キャップ	ラックレベル計測のみ	ラックレベル計測のみ
仮想化	KVM(G&Vキュー資源分離)	なし	なし

TSUBAME3.0:世界をリードする最先端技術

- 現在設計中:2016年度初頭〜中頃ごろ稼働
- 高演算性能 **〜20 ペタフロップス, メモリ性能 ~5 ペタバイト/秒** (京の2倍)
- 超高密度: ラック毎0.6ペタフロップス以上(TSUBAME2比10倍、京の60倍)
- 超省電力: 10ギガフロップス/W以上(TSUBAME2比10倍以上)
  - 最先端の電力制御・温水自然冷却・エネルギー回生
- **超高速ネットワーク: 1ペタビット/秒以上の容量**
  - **全世界インターネットの通信容量以上**
- **次世代の科学ビッグデータ: 数ペタバイト不揮発メモリ、5-10テラバイト秒(京の3~6倍)、1億 IOPS 以上、数十ペタバイトの総合容量**
- 先進的仮想化と資源マネジメント: 世界初のペタフロップス高性能仮想化、電力最適化スケジューリング、超高信頼耐故障性など

TSUBAME-KFC: ウルトラグリーン・スパコン研究設備

（文部科学省概算要求・2011-2015・約2億円)

液浸冷却＋高温大気冷却＋高密度実装＋電力制御のスパコン技術を統合
TSUBAME3.0のプロトタイプ

高密度実装・油浸冷却
210TFlops (倍精度)
630TFlops (単精度)

高温冷却系
冷媒油 35~45℃
⇒ 水 25~35℃
(TSUBAME2は7~17℃)

冷却塔:
水 25~35℃
⇒ 自然大気へ

コンテナ型研究設備
20フィートコンテナ(16m²)
無人自動制御運転

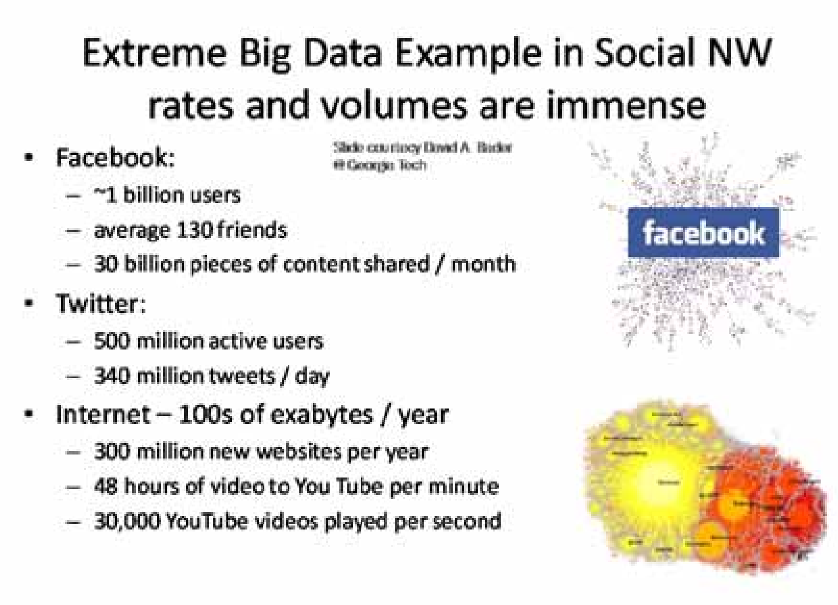
2013年11月 Green500ランキング TSUBAME2.5 世界一(日本初)

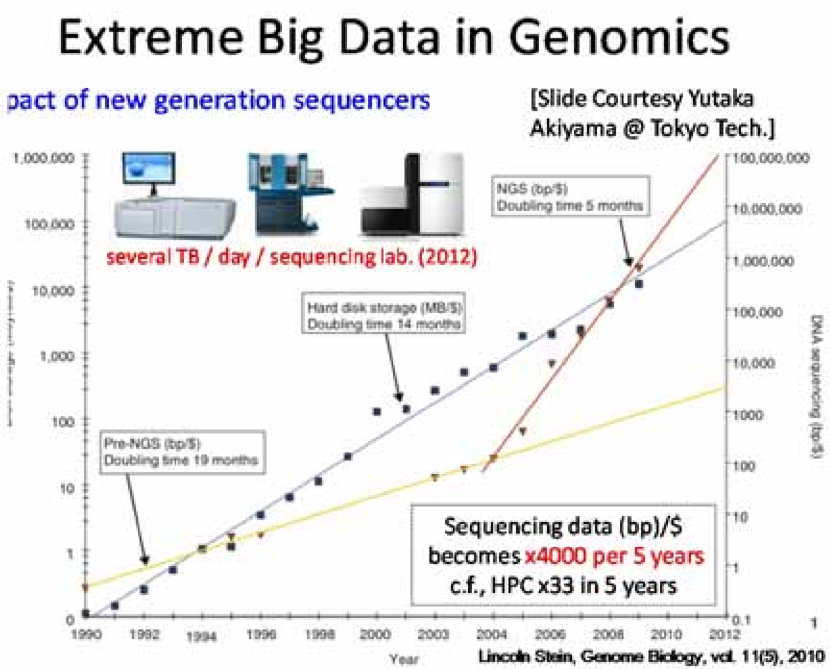
Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 Infiniband QDR, NVIDIA K20x	
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, I FDR14, NVIDIA K20x	
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, NVIDIA K20x	
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, NVIDIA K20x	
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C Infiniband FDR, Nvidia K20m	

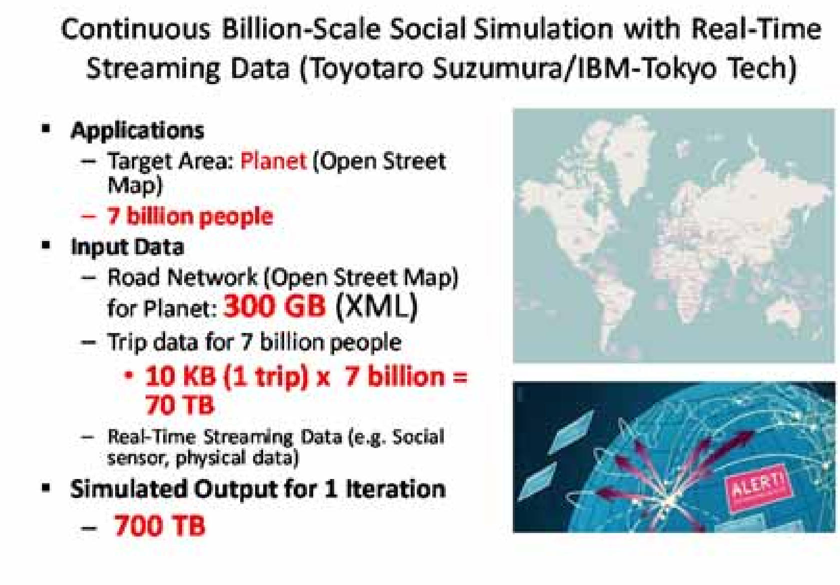
2014/6月にも世界一防衛成功

The current “Big Data” are not really that Big… 今の「ビッグデータ」はビッグではない

- Typical “real” definition: “Mining people’s privacy data to make money”「企業がプライバシーをマイニングして金儲け」
- Corporate data Gigabytes~Terabytes, seldom Petabytes. せいぜいギガ～テラバイト級
  - Processing involve simple O(n) algorithms, or those that can be accelerated with DB-inherited indexing algorithms 処理や処理量も少ない
- Executed on re-purposed commodity “web” servers linked with 1Gbps networks running Hadoop/HDFS ウェブ用のサーバのHadoop程度
- Vicious cycle of stagnation in innovations…このままでは進歩がない
- **⇒ Convergence with Supercomputing with Extreme Big Data スパコンとの「コンバージェンス」による次世代ビッグデータ**







Future "Extreme Big Data"

- **<u>NOT mining Tbytes Silo Data</u>**
- **Peta~Zetabytes of Data**
- **Ultra High-BW Data Stream**
- **Highly Unstructured, Irregular**
- **Complex correlations between data from multiple sources**
- **Extreme Capacity, Bandwidth, Compute All Required**

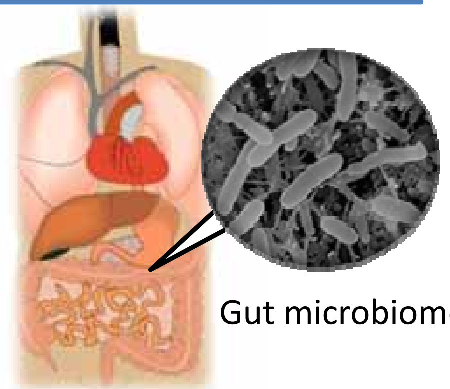
We will have tons of unknown genes

Metagenome analysis

[Slide Courtesy Yutaka Akiyama @ Tokyo Tech.]

- Directly sequencing uncultured microbiomes obtained from target environment and analyzing the sequence data
  - Finding novel genes from unculturable microorganism
  - Elucidating composition of species/genes of environments

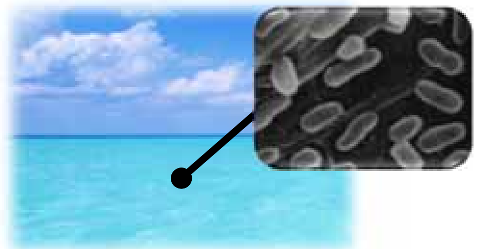
Examples of microbiome



Gut microbiome

Human body

Soil



Sea

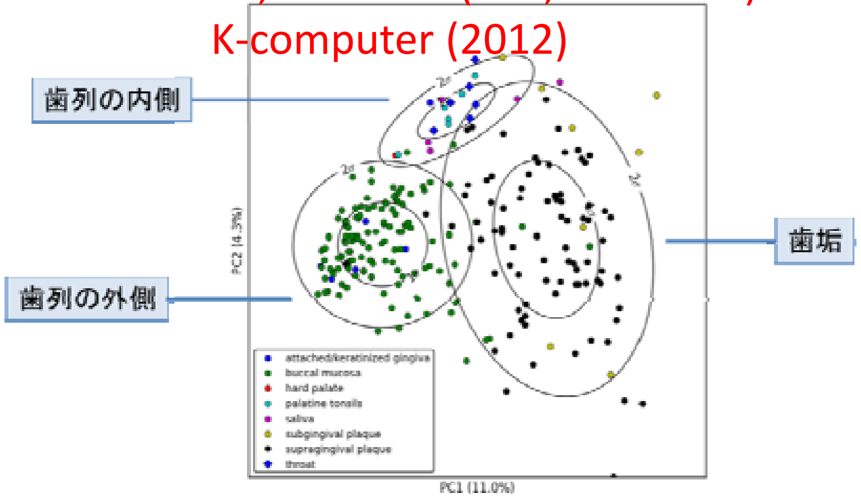
Results from Akiyama group@Tokyo Tech

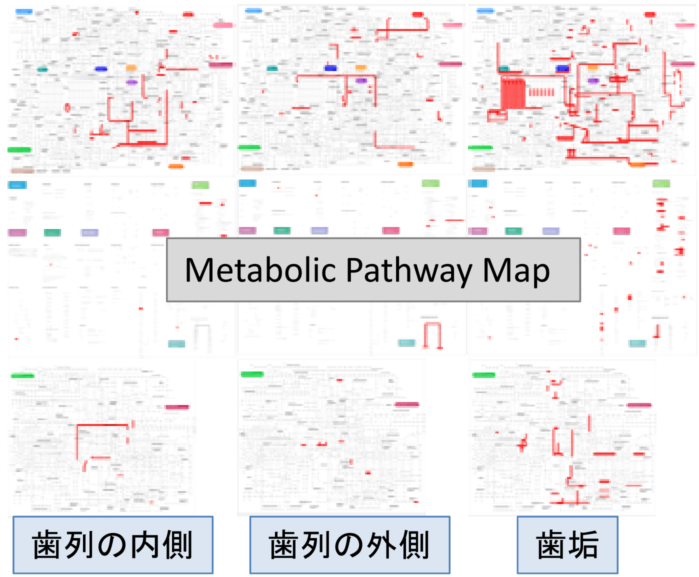
Ultra high-sensitive “big data” metagenome sequence analysis of human oral microbiome

- Required > **1 million node*hour product** on K-computer
- World’s most sensitive sequence analysis (based on amino acid similar

- Discovered at least three microbiome clusters with <u>functional</u> differences.
  (Integrated 422 experiment samples taken from 9 different oral parts)

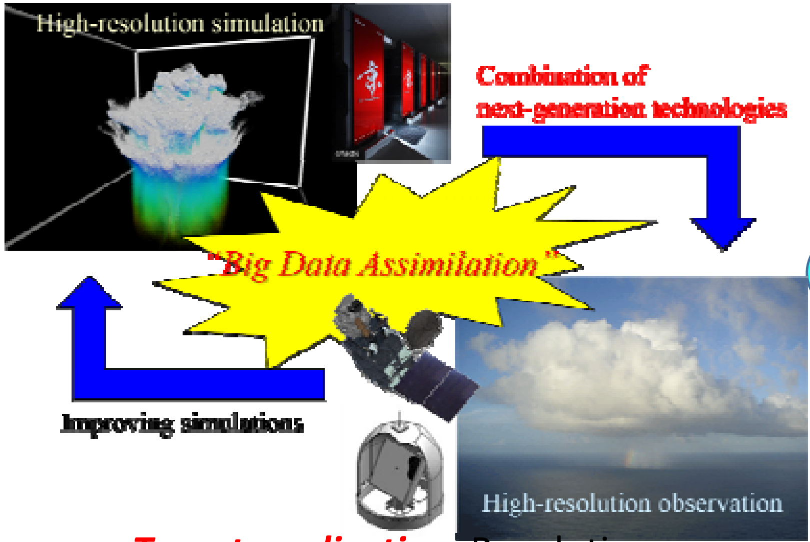
572.8 M Reads / hour
82,944 node (663,552 Cores)
K-computer (2012)





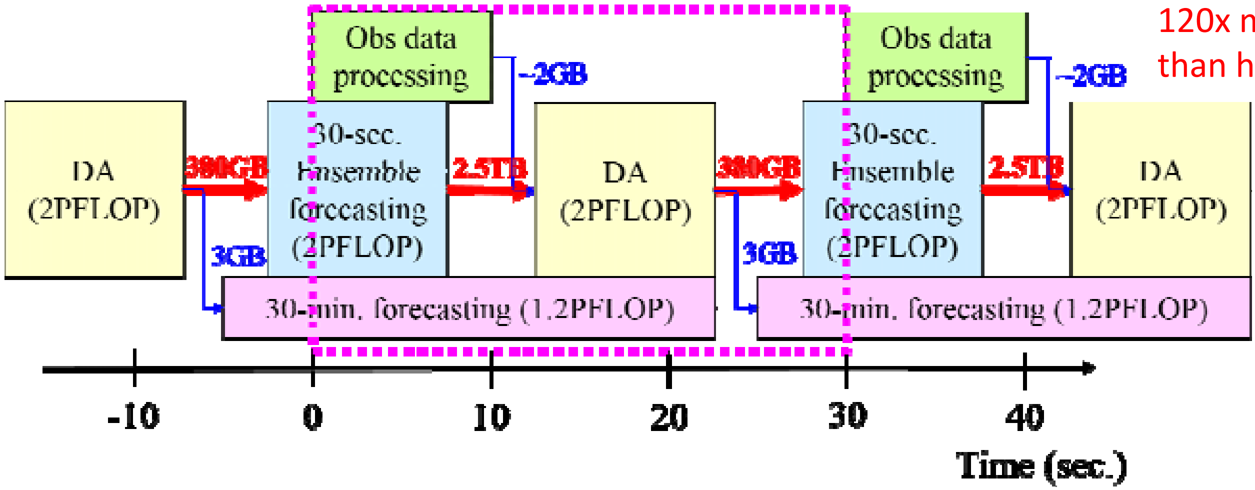
"Big Data Assimilation"-
気象におけるシミュレーションと観測の融合

EBD Miyoshi Group



莫大なデータの転送量
超高速なマッチング
高信頼性の確保

Target application: Revolutionary super-rapid 30-second update forecasting

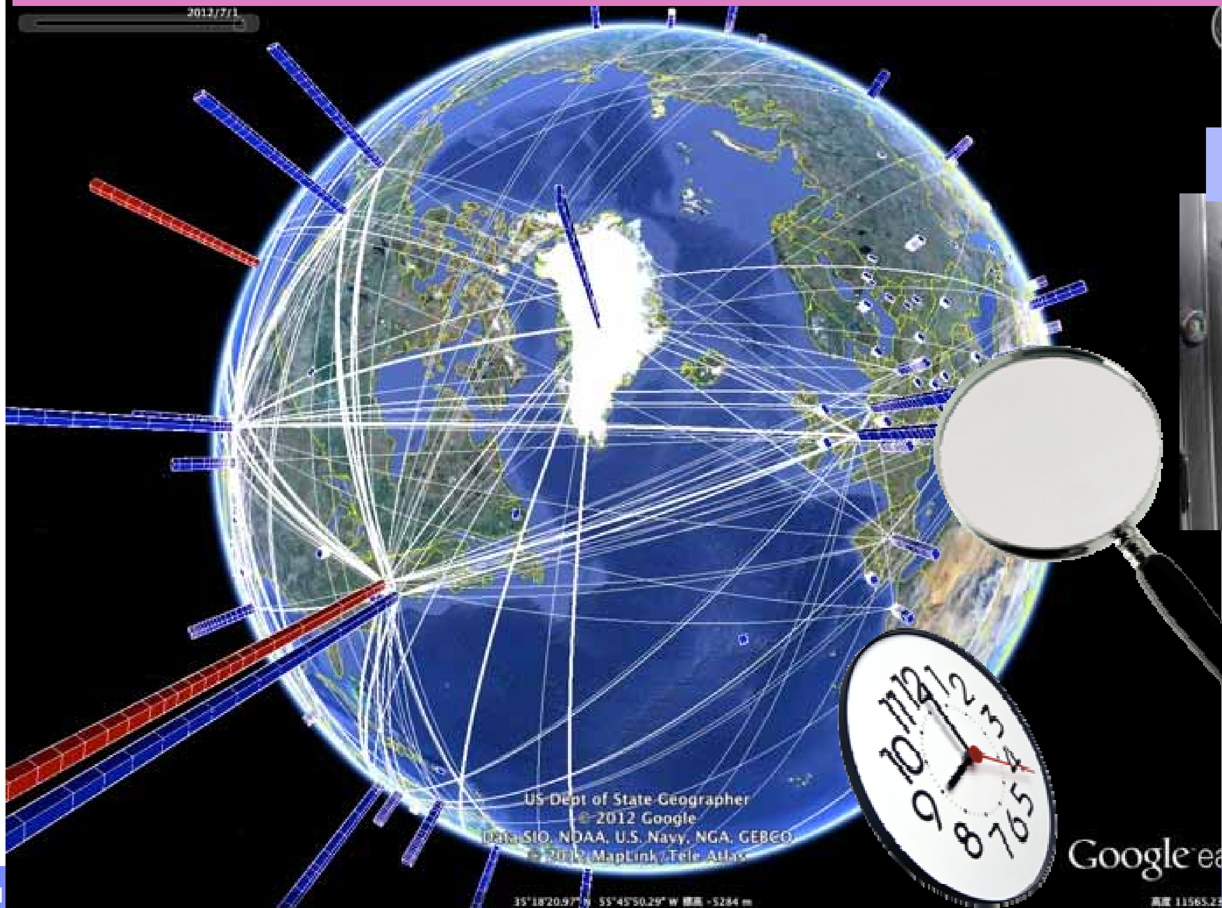


120x more rapid than hourly update

巨大なソーシャルネットワークの理解 → グラフ構造の解析

(e.g. separation of degree, diameter, clustering, ..) EBD Suzumura Group

Crawled the entire Twitter follower/followee network of **826.10 million vertices and 29.23 billion edges. How could we analyze this gigantic graph ?**

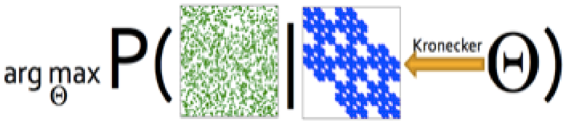


Supercomputers

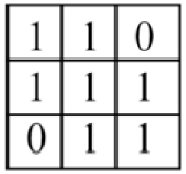
8.26億個の頂点(人)
292億個の片(フォロー・フォロワー関係)

Graph500 “Big Data” Benchmark

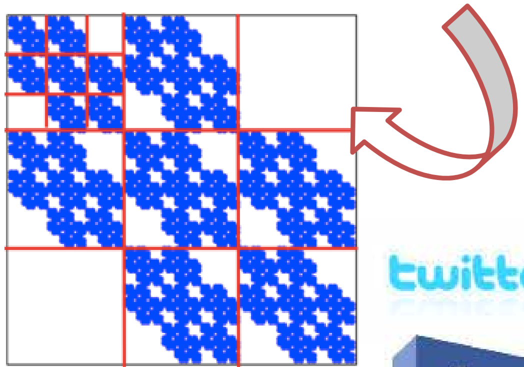
Kronecker graph BSP Problem



A: 0.57, B: 0.19
C: 0.19, D: 0.05



G_1



G_4 adjacency matrix

November 15, 2010

Graph 500 Takes Aim at a New Kind of HPC
Richard Murphy (Sandia NL => Micron)
“ **I expect that this ranking may at times look very different from the TOP500 list. Cloud architectures will almost certainly dominate a major chunk of part of the list.**” 予想：クラウドが検討？

The 4th Graph500 List (Jun2012) TSUBAME #4 w/GPUs

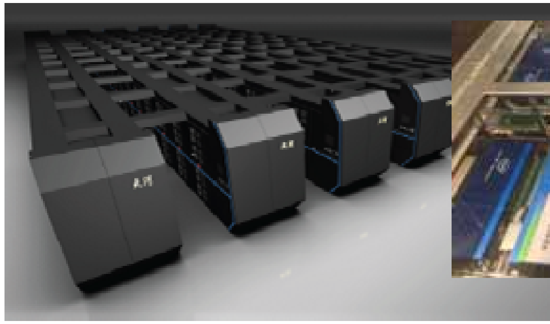
Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology

Rank	Installation Site	Machine	Number of nodes	Number of cores	Problem scale	GTEPS
1	DOE/SC/Argonne National Laboratory	Mira/BlueGene/Q	32768	524288	38	3541.00
1	LLNL	Sequoia/Blue Gene/Q	32768	524288	38	3541.00
2	DARPA Trial Subset, IBM Development Engineering	Power 775, POWER7 8C 3.836 GHz	1024	32768	35	508.05
3	Information Technology Center, The University of Tokyo	Oakleaf-FX (Fujitsu PRIMEHPC FX 10)	4800	76800	38	358.10
4	GSIC Center, Tokyo Institute of Technology	TSUBAME	1366	16392	35	317.09
5	Brookhaven National Laboratory	BLUE GENE/Q	1024	16384	34	294.29
6	DOE/SC/Argonne National Laboratory	Vesta/BlueGene/Q	1024	16384	34	292.36

Reality: Top500 Supercomputers Dominate
No Cloud IDCs at all 現実：クラウドはランク外
TSUBAME2.0 #3(Nov.2011) #4(Jun.2012)

Top Supercomputers vs. Global IDC

K Computer (#1 2011-12) Riken-AICS
Fujitsu Sparc VIII-fx Venus CPU
88,000 nodes, 800,000CPU cores
~11 Petaflops (10^{16})
1.4 Petabyte memory, 13 MW Power
864 racks、3000m^2



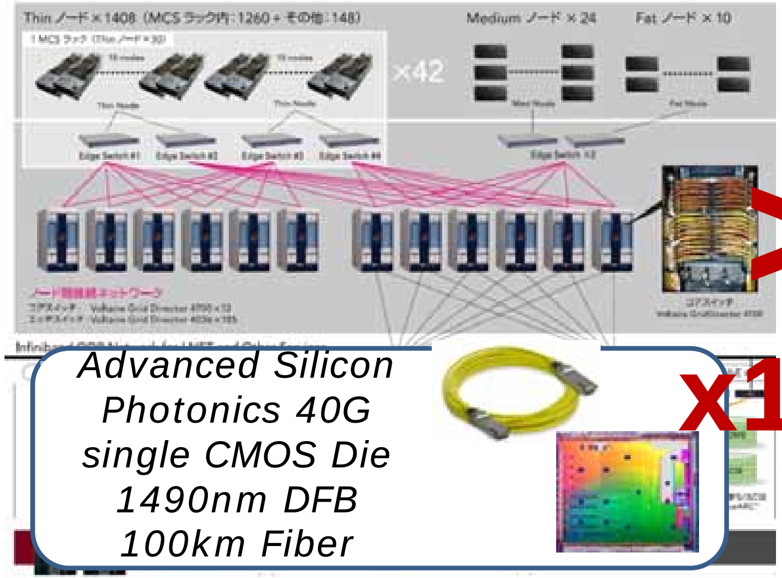
Tianhe2 (#1 2013) China Gwanjou
48,000 KNC Xeon Phi + 36,000 Ivy Bridge Xeon
18,000 nodes, >3 Million CPU cores
54 Petaflops (10^{16})
0.8 Petabyte memory, 20 MW Power
??? racks、???m^2

C.f. Amazon ~= 500,000 Nodes, ~5 million Cores

#1 2012 IBM BlueGene/Q “Sequoia”
Lawrence Livermore National Lab
IBM PowerPC System-On-Chip
98,000 nodes, 1.57million Cores
~20 Petaflops
1.6 Petabytes, 8MW, 96 racks

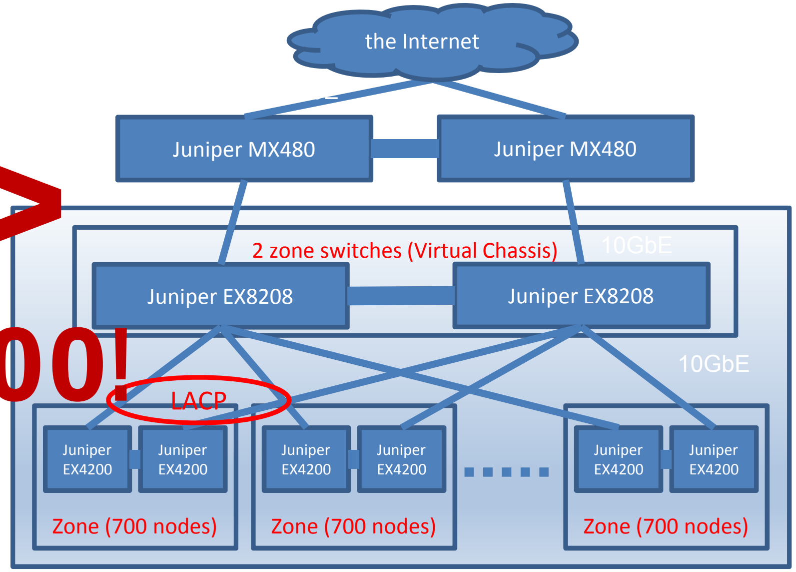
DARPA study
2020 Exaflop (10^{18})
100 million~
1 Billion Cores

Supercomputer Tokyo Tech.
Tsubame 2.0
#4 Top500 (2010)



>> x1000!

A Major Northern Japanese Cloud Datacenter (2013)



~1500 nodes compute & storage
Full Bisection Multi-Rail
Optical Network
Injection 80GBps/Node
Bisection 220Terabps

8 zones, Total 5600 nodes,
Injection 1GBps/Node
Bisection 160Gigabps

But what does “220Tbps” mean?
220テラビット/秒とは?

Global IP Traffic, 2011-2016 (Source Cicso)							
	2011	**2012**	**2013**	**2014**	**2015**	**2016**	**CAGR 2011-2016**
By Type (PB per Month / Average Bitrate in Tbps)							
Fixed Internet	23,288	32,990	40,587	50,888	64,349	81,347	28%
	71.9	101.8	125.3	157.1	198.6	251.1	
Managed IP	6,849	9,199	11,846	13,925	16,085	18,131	21%
	21.1	28.4	36.6	43.0	49.6	56.0	
Mobile data	597	1,252	2,379	4,215	6,896	10,804	78%
	1.8	3.9	7.3	13.0	21.3	33.3	
Total IP traffic	30,734	43,441	54,812	69,028	87,331	110,282	29%
	94.9	134.1	169.2	213.0	269.5	340.4	

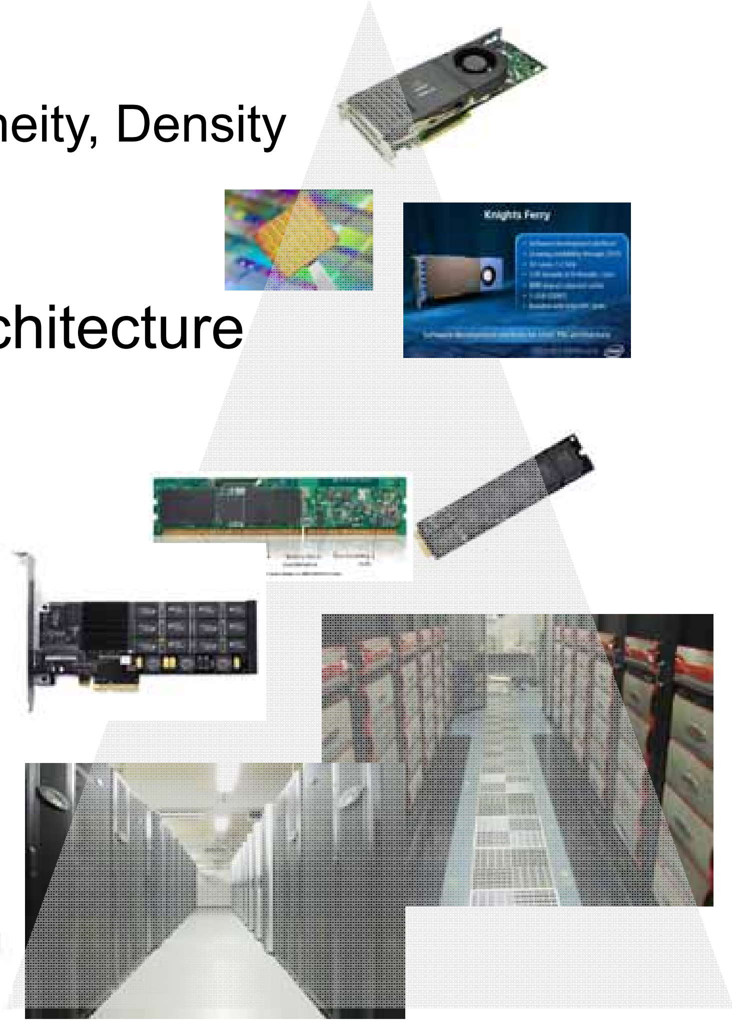
TSUBAME2のネットワーク容量は全世界の
インターネット全体の平均トラフィックに匹敵

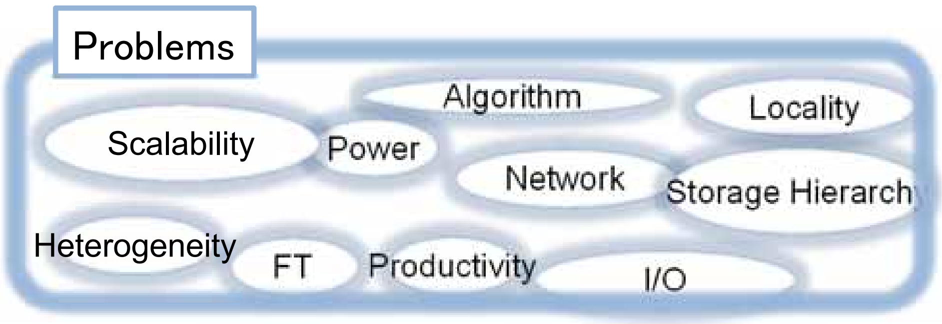
今後:スパコンとビッグデータ基盤の統合

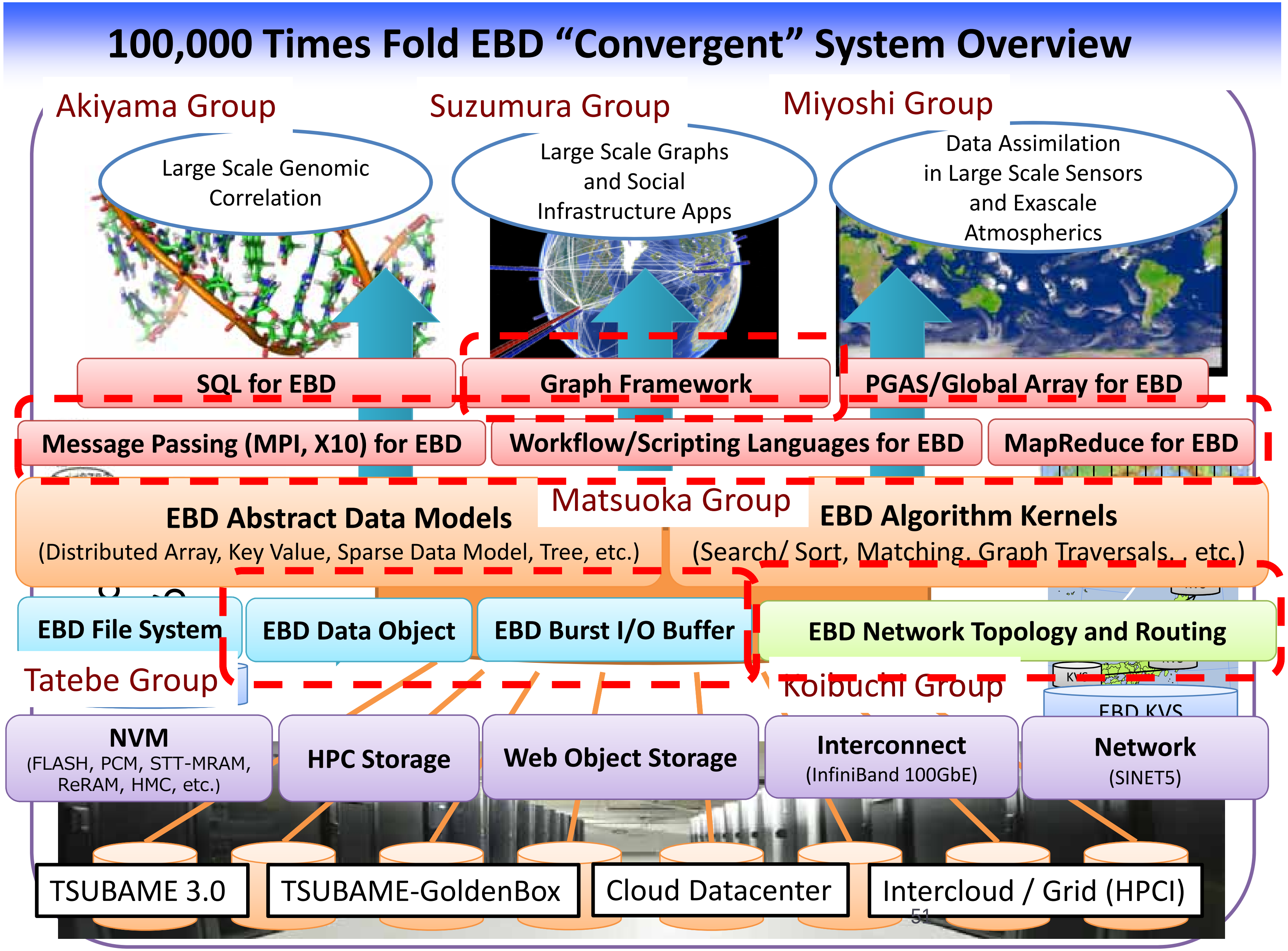
JST-CREST Extreme Big Data (2013-)

Towards Extreme-scale BigData Machines
将来のEBDスパコン・IDCビッグデータ統合マシン

- Computation
  - Increase in Parallelism, Heterogeneity, Density
    - Multi-core, Many-core processors
    - Heterogeneous processors
- Hierarchial Memory/Storage Architecture
  - NVM (Non-Volatile Memory), SCM (Storage Class Memory)
    - FLASH, PCM, STT-MRAM, ReRAM, HMC, etc.
  - Next-gen HDDs (SMR), Tapes (LTFS)



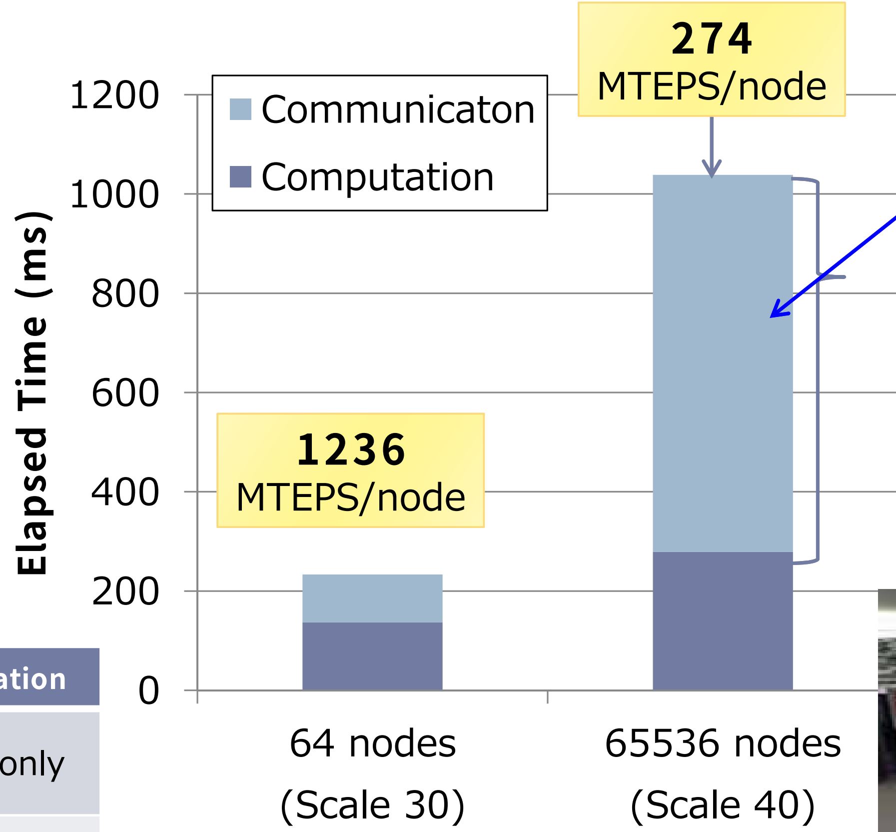




100,000 Times Fold EBD "Convergent" System Overview
Akiyama Group
Suzumura Group
Miyoshi Group
Large Scale Genomic Correlation
Large Scale Graphs and Social Infrastructure Apps
Data Assimilation in Large Scale Sensors and Exascale Atmospherics
SQL for EBD
Graph Framework
PGAS/Global Array for EBD
Message Passing (MPI, X10) for EBD
Workflow/Scripting Languages for EBD
MapReduce for EBD
Matsuoka Group
EBD Abstract Data Models
(Distributed Array, Key Value, Sparse Data Model, Tree, etc.)
EBD Algorithm Kernels
(Search/ Sort, Matching, Graph Traversals, , etc.)
EBD File System
EBD Data Object
EBD Burst I/O Buffer
EBD Network Topology and Routing
Tatebe Group
Koibuchi Group
EBD KVS
NVM
(FLASH, PCM, STT-MRAM, ReRAM, HMC, etc.)
HPC Storage
Web Object Storage
Interconnect
(InfiniBand 100GbE)
Network
(SINET5)
TSUBAME 3.0
TSUBAME-GoldenBox
Cloud Datacenter
Intercloud / Grid (HPCI)

K Computer #1 東工大[EBD CREST]、九大[Fujisawa Graph CREST]、理研などの共同成果

List	Rank	GTEPS	Implementation
November 2013	4	5524.12	Top-down only
June 2014	1	17977.05	**Efficient hybrid**



73% total exec time wait in communication

*Problem size is weak scaling

2013/11 Green Graph500 ランキング

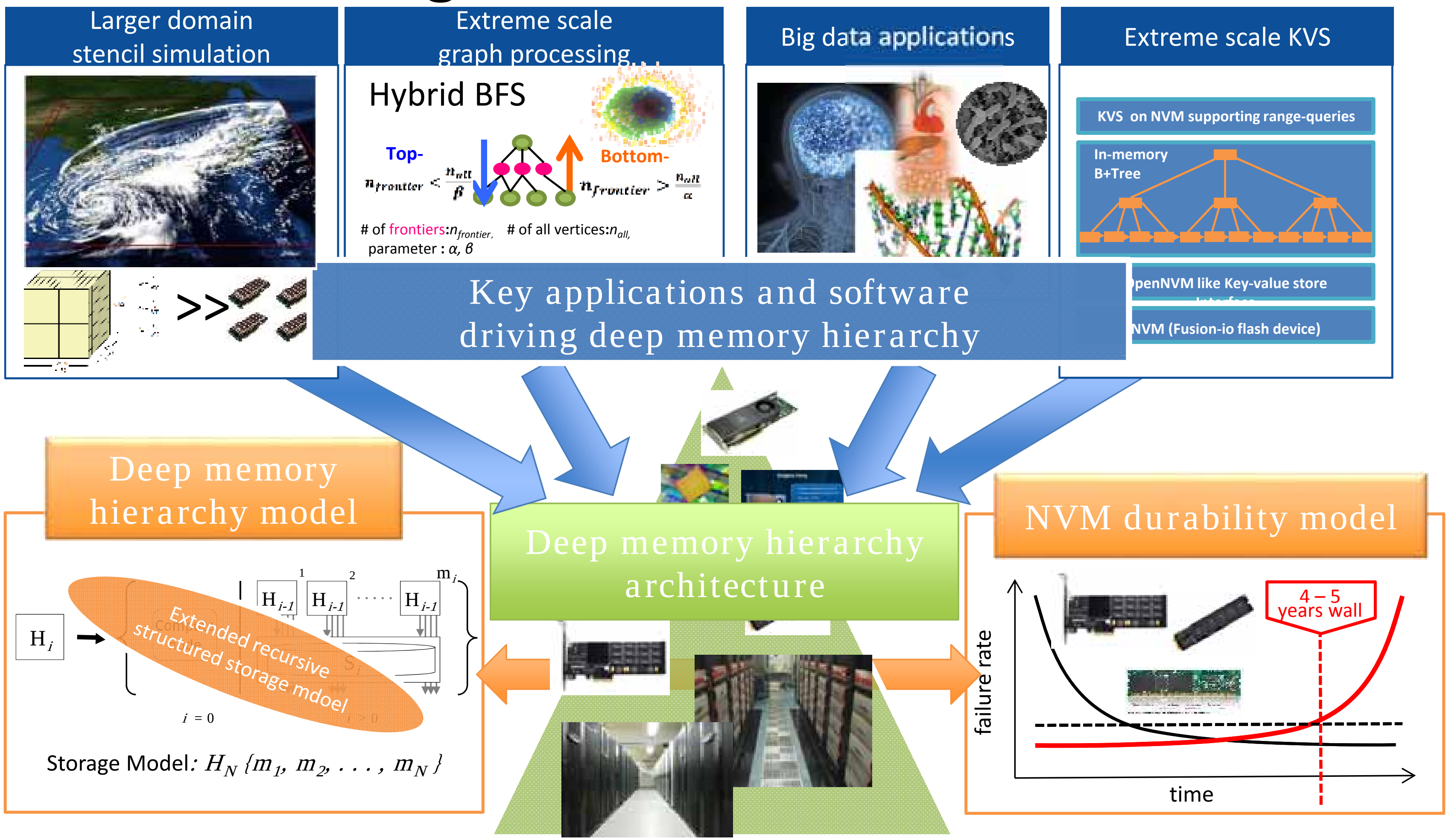
- ビッグデータグラフの単位電力あたりの処理能力 **TEPS/W** 値で評価
- http://green.graph500.org
- Green500と共にTSUBAME-KFCは世界一 ⇒ 二冠達成！

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							6384
3			DO				2768
4			E				1
5			DO				5536
6							1
7							64

TSUBAME-KFCは二冠獲得！

Future: Big Data & Deep memory hierarchy and modeling



TSUBAME4 2021-22年「K in a Box」

「京」を一箱にする技術の研究

大きさ1/500

電力 1/150

コスト 1/500

より高い汎用性

10ペタフロップス

メモリ10ペタバイト（京は1.6)

ひと箱1万ノード

巨大スパコンだけでなく、世界の巨大インターネット・データセンターを
ひと箱に集約、1/1000にコストダウン

2兆円スパコン市場だけでなく、20兆円以上のIDCビッグデータインフラの革命

GoldenBox Proto1 (NVIDIA K1-based) IEEE/ACM Supercomputing 2014 東工大ブースにて展示

- 36 Node Tegra K1, ~11TFlops SFP
- ~700GB/s BW
- 100-700Watts
- Integrated mSata SSD, ~7GB/s I/O
- Ultra dense, Oil immersive cooling
- Same SW stack as TSUBAME2

2022: x10 Flops, x10 Mem Bandwidth, silicon photonics, x10 NVM, x10 node density, with new device and packaging technologies

専門家向けバックアップ資料

EBD Programming Framework

Hamar (Highly Accelerated Map Reduce)

- 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

EBD Programming Framework

Out-of-core GPU-MapReduce for Large-scale Graph Processing [Cluster 2014]

Emergence of large-scale graphs

- SNS, road network, smart grid, etc.
- Millions to trillions of vertices/edges

→ Need for fast graph processing on supercomputers

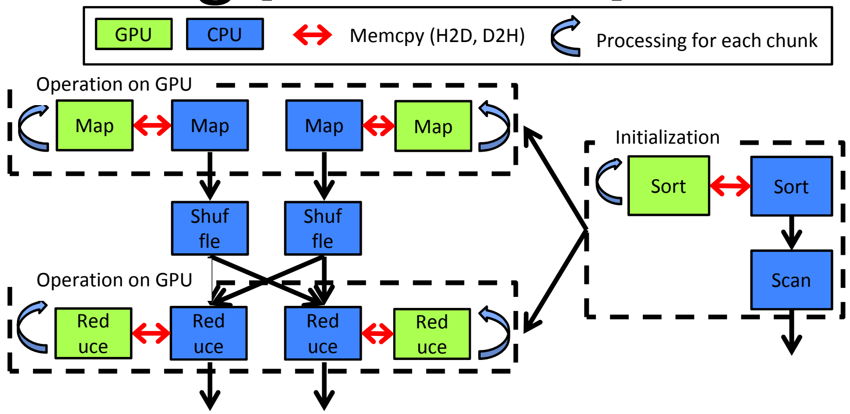
Problem: GPU memory capacity limits scalable large-scale graph processing

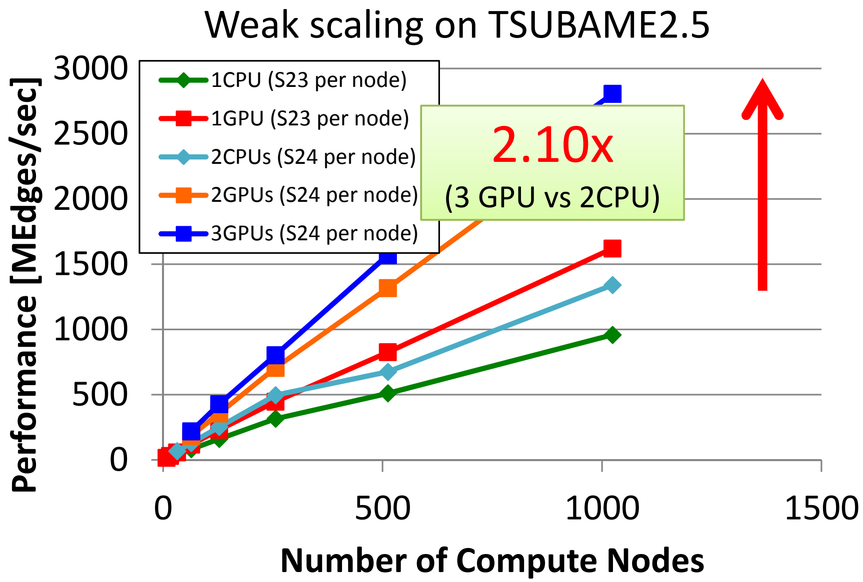
Proposal: Out-of-core GPU memory management on MapReduce

- Stream-based GPU MapReduce
- Out-of-core GPU sorting

Experimental Results:
performance improvement over CPUs

- Map: 1.41x, Reduce: 1.49x, Sort: 4.95x speedup
- Overlapping communication effectively





EBD Algorithm Kernels

GPU-HykSort [IEEE BigData2014]

Motivation

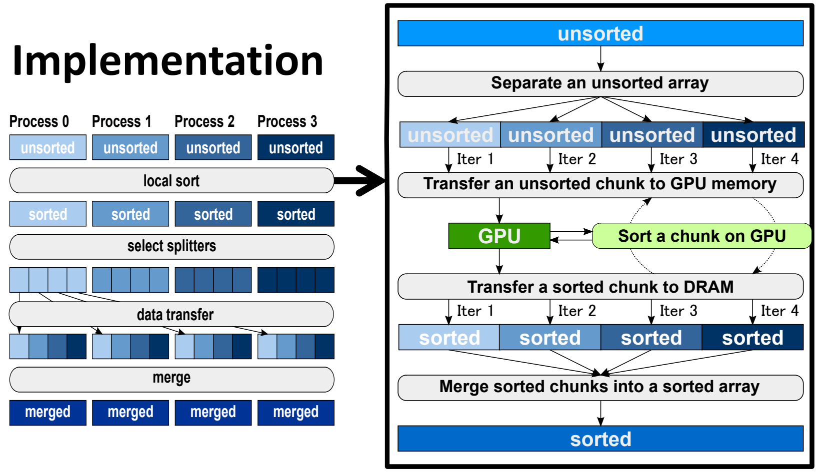
Effectiveness of sorting for large-scale GPU-based heterogeneous systems remains unclear

- Appropriate selection of phases to be offloaded to GPU is required
- Handling GPU memory overflow is required

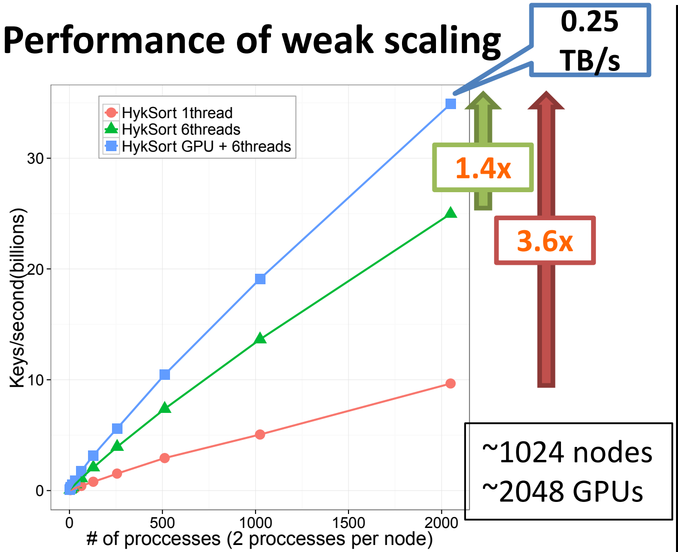
Approach

Offload local sort, the most time-consuming phase, to GPU accelerators

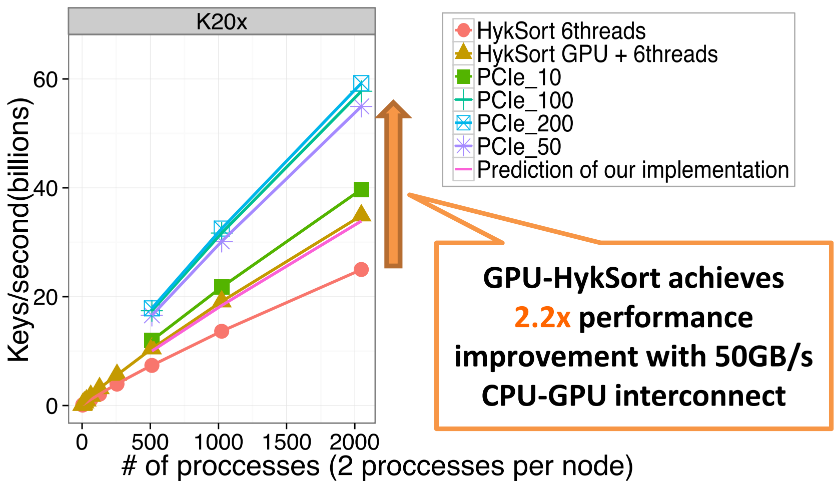
Implementation



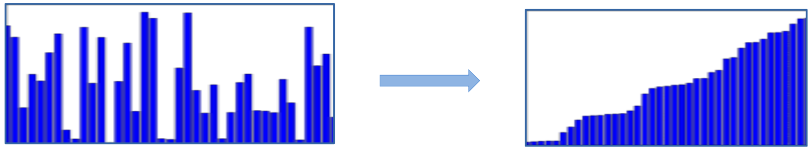
Performance of weak scaling



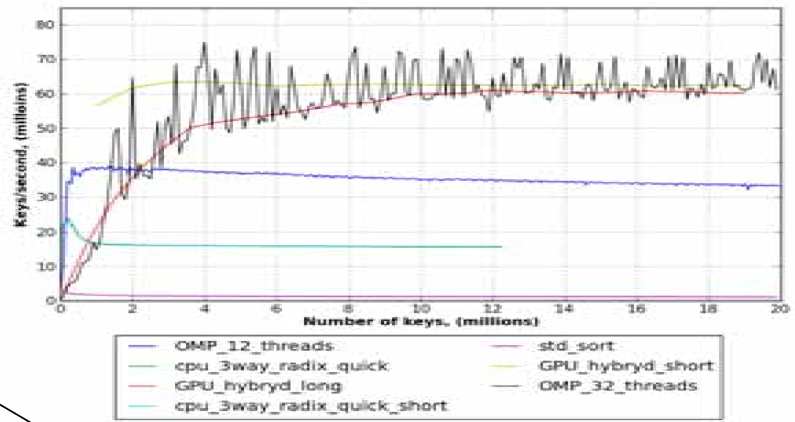
Performance prediction



Efficient Parallel Sorting Algorithm for Variable-Length Keys



- Comparison-based sorts inefficient for long/variable-length keys (like strings)
- Better way: examining individual characters (based on MSD Radix sort algorithm)
- Hybrid parallelization scheme: combining data-parallel and task-parallel stages



apple
apricot
banana
kiwi

70 M keys/second sorting throughput on 100bytes strings

Aleksandr Drozd, Miquel Pericàs, Satoshi Matsuoka. Efficient String Sorting on Multi- and Many-Core Architectures. *in Proceedings of IEEE 3rd International Congress on Big Data. Anchorage, USA, August 2014*

Aleksandr Drozd, Miquel Pericàs, Satoshi Matsuoka. MSD Radix String Sort on GPU: Longer Keys, Shorter Alphabets *in proceedings of 第142回ハイパフォーマンスコンピューティング合同研究発表会 (HOKKE-21)*

EBD Algorithm Kernels

Scalable Distributed Memory BFS

What's the best algorithm for the distributed memory Breadth First Search?

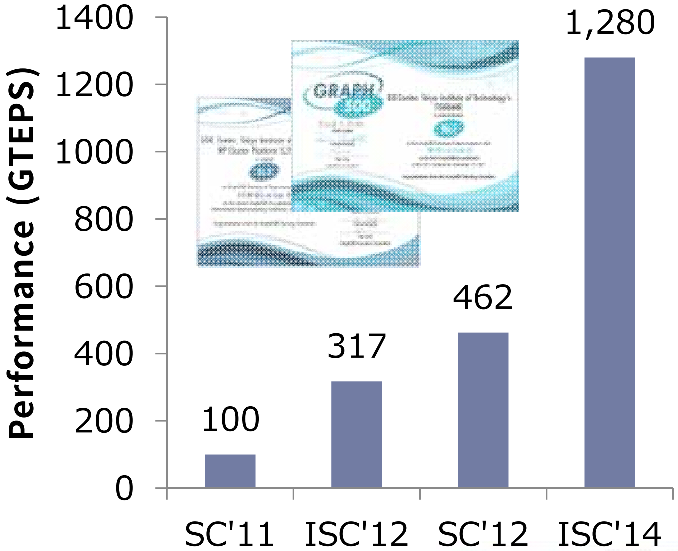
Proposal

- Efficient CSR with Bitmap
- Adaptive Data Representation
- And Many Other Optimizations

Optimizations	SC11	ISC12	SC12	ISC14
2D decomposition	✓	✓	✓	✓
vertex sorting	✓			
direction optimization				✓
data compression	✓	✓	✓	
sparse vector with pop counting				✓
adaptive data representation				✓
overlapped communication	✓	✓	✓	✓
shared memory				✓
GPGPU		✓	✓	

✓ Utilization for each version

Graph500 score history of TSUBAME2



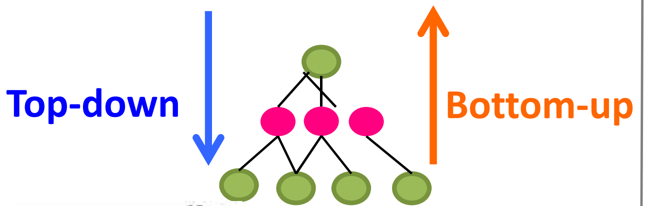
We achieved **17,977GTEPS** on K computer and ranked **1st** in the June 2014 **Graph500** List

Large Scale Graph Processing Using NVM

EBD Algorithm Kernels

1. Hybrid-BFS (Beamer'11)

Switching two approaches

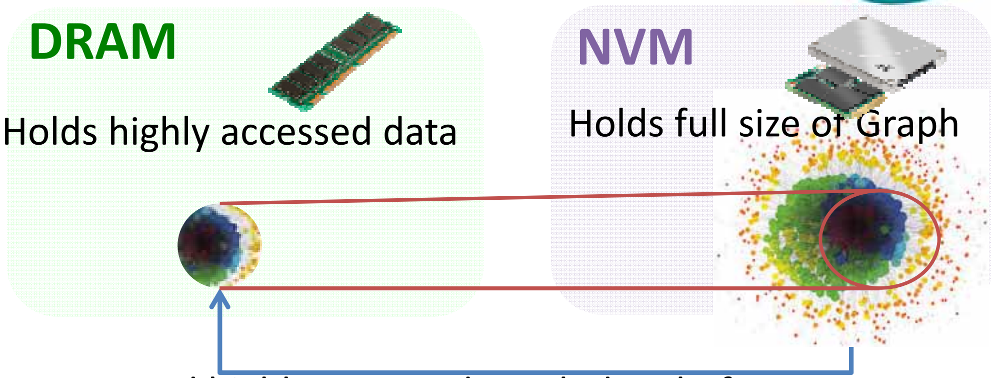


$n_{frontier} < \frac{n_{all}}{\beta}$ $n_{frontier} > \frac{n_{all}}{\alpha}$

# of frontiers: $n_{frontier}$, # of all vertices: n_{all}, Parameter : α, β

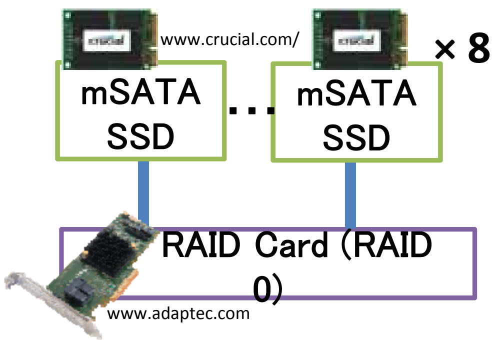
2. Proposal

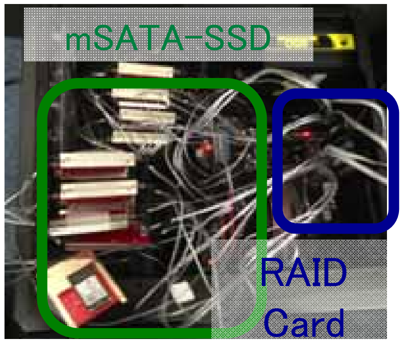
[BigData2014]

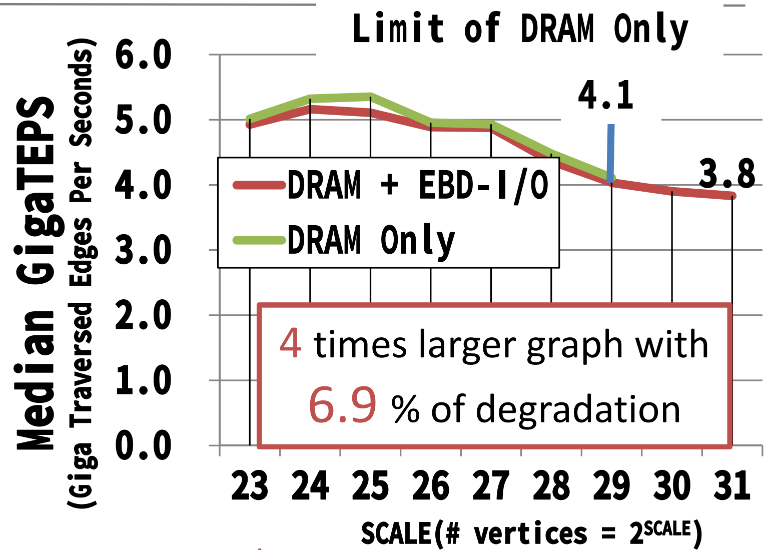


3. Experiment

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





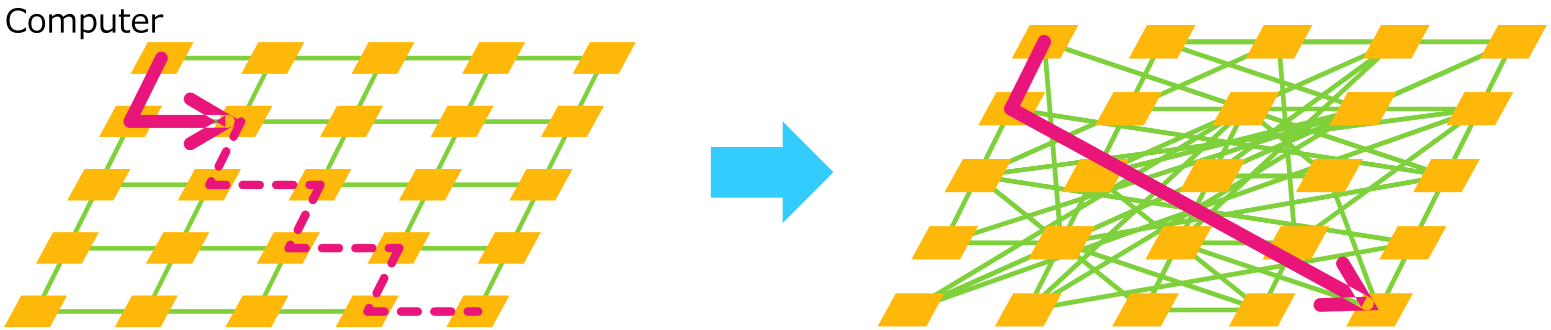


Ranked 3rd
in **Green Graph500 (June 2014)**

Random network topology

EBD Interconnect

Enabling short path hops [Koibuchi G]



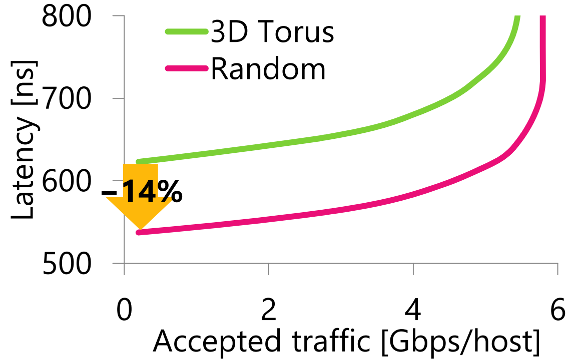
(a) Existing supercomputers (torus topology)
long communication delay t

(b) Random network topology
a lot of shortcut paths

Low Communication Latency [<1us] is one of crucial topic on Exascale supercomputer networks

[NSF WS2008 Reports]

▼ Latency (Y-axis) and Throughput (X-axis)



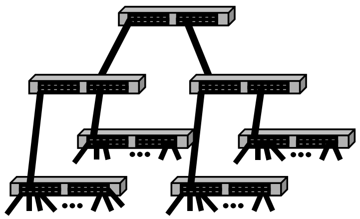
Michihiro Koibuchi, Hiroki Matsutani, Hideharu Amano, D. Frank Hsu, Henri Casanova, ``A Case for Random Shortcut Topologies for HPC Interconnects'', The 39th International Symposium on Computer Architecture (**ISCA**), pp.177-188, June 2012

Our **Random** Network Topology fits to **EBD Irregular Workload**

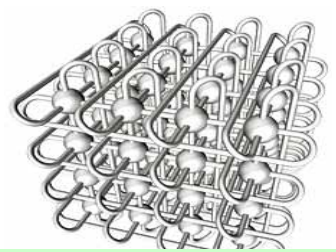
Graph Analysis (e.g.: Graph500 Benchmark)

Nonneighboring-communication scientific workload

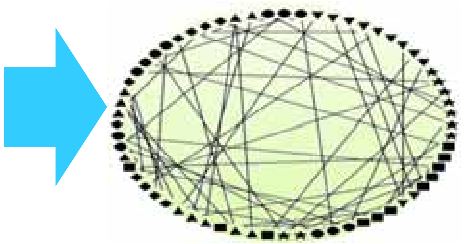
Observation data (Miyoshi G), ….

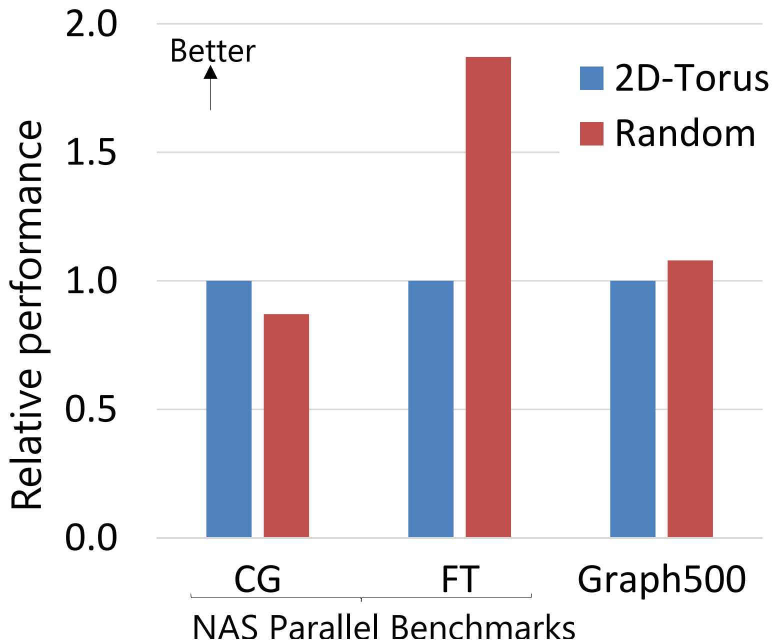


Data Centers



Supercomputers

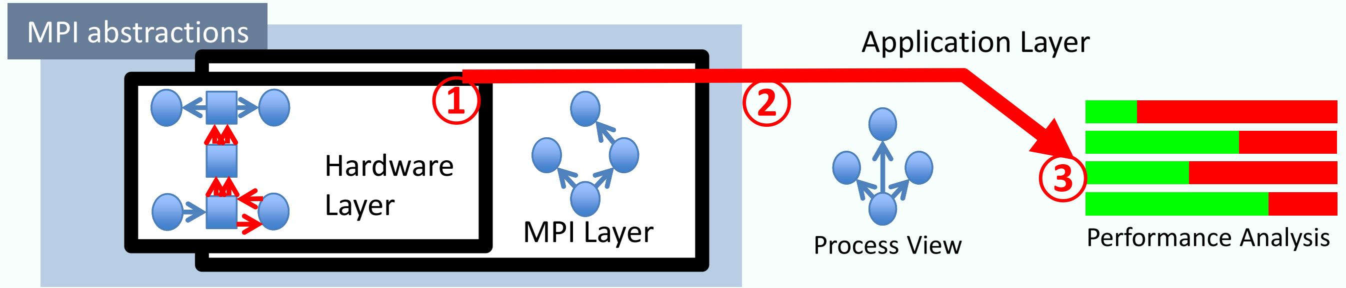




EBD Interconnect

Network Performance Visualization [EuroMPI/Asia 2014 Poster]

OUR GOALS

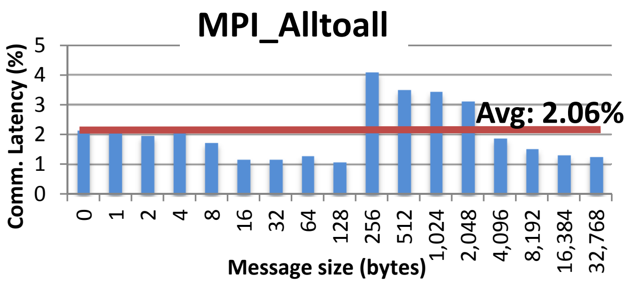


① Portably expose MPI's internal performance

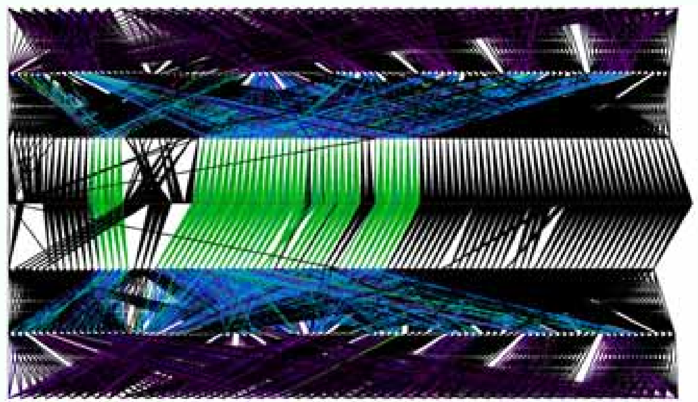
② Non-intrusively profile low-level metrics

③ Flexible hardware-centric performance analysis

Overhead of our profiler (named ibprof):



NAS Parallel FT Benchmark
Runtime overhead = less than 0.02%
(12.1919s -> 12.1935s)

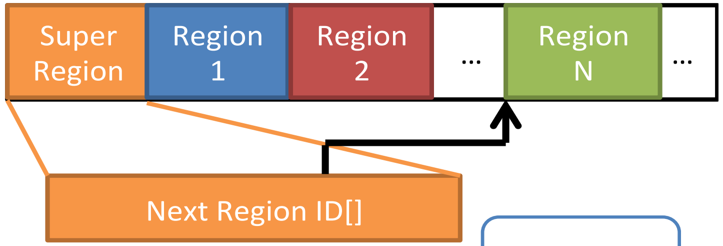


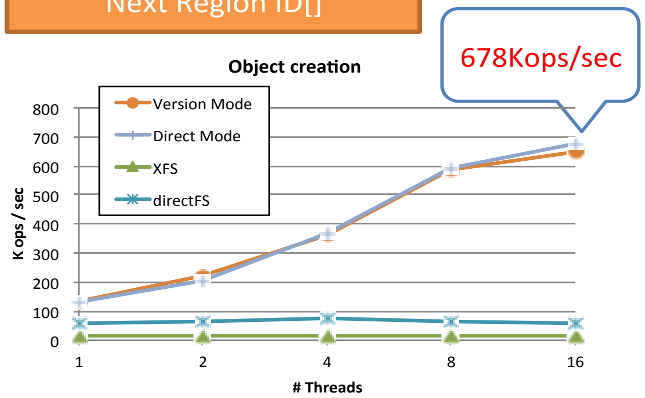
Network visualization of TSUBAME 2.5 running the Graph500 benchmark on 512 nodes

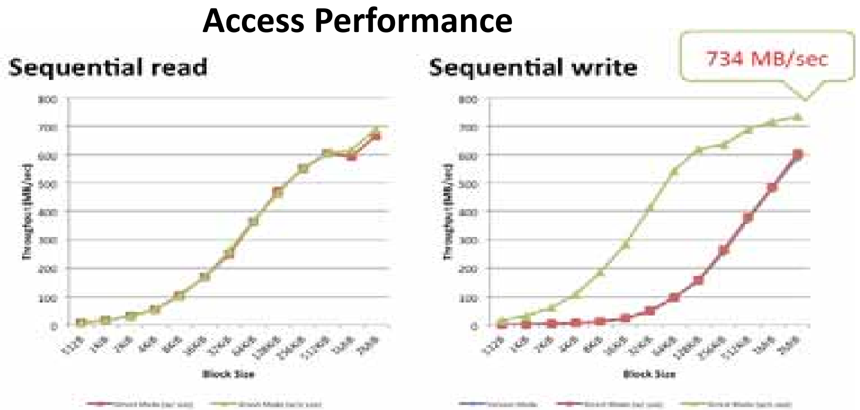
Object Storage for OpenNVM flash primitives [Tatebe Group, Takatsu]

EBD NVM System Software

- Object storage design for high bandwidth and high IOPS in OpenNVM
  - OpenNVM flash primitives: sparse address space and atomic batch write
- Simple design based on fixed-size Region
  - One object for one object
  - Enough region size to store one object
  - Object ID = region number
- Simple region number management in super region
  - Sequential assignment
  - Free'ed by persistent trim and no reuse
  - Blocked assignment to mitigate access conflict to the super block







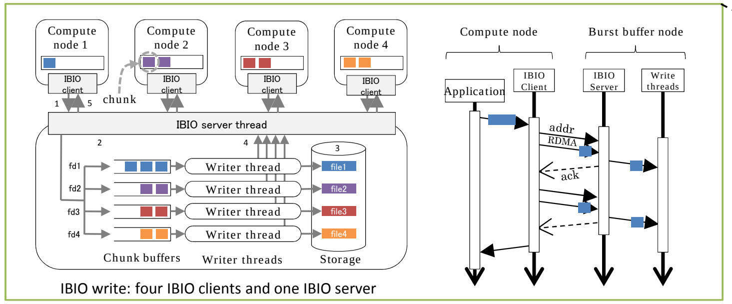
EBD I/O and C/R modeling for extreme scale[CCGrid2014 Best Paper]

EBD NVM System Software

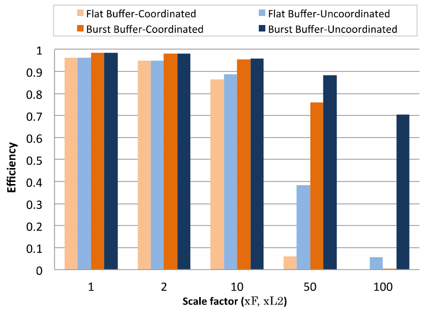
Extreme scale I/O for Burst Buffer

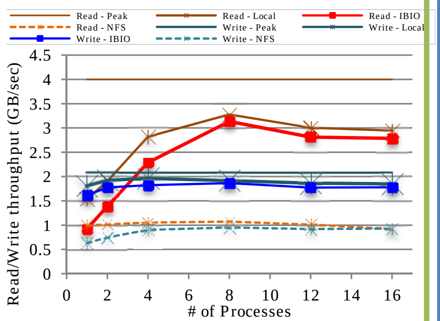
CPU	Intel Core i7-3770K CPU (3.50GHz x 4 cores)
Memory	Cetus DDR3-1600 (16GB)
M/B	GIGABYTE GA-Z77X-UD5H
SSD	Crucial m4 msata 256GB CT256M4SSD3 (Peak read: 500MB/s, Peak write: 260MB/s)
SATA converter	KOUTECH IO-ASS110 mSATA to 2.5' SATA Device Converter with Metal Fram
RAID Card	Adaptec RAID 7805Q ASR-7805Q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

EBD I/O



IBIO write: four IBIO clients and one IBIO server

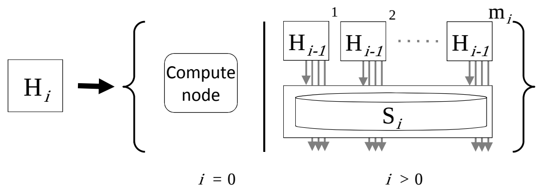




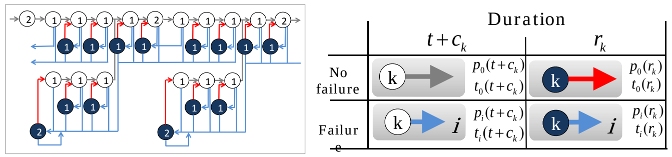
Extreme scale C/R modeling

$$O_i = \begin{cases} C_i + E_i & \text{(Sync.)} \\ I_i & \text{(Async.)} \end{cases} \qquad L_i = C_i + E_i$$

$$C_i \text{ or } R_i = \frac{\text{< C/R date size / node >} \times \text{<\# of C/R nodes per } S_i^* \text{ >}}{\text{< write perf. (} w_i \text{) > or <read perf. (} r_i \text{) >}}$$



Storage Model: $H_N \{m_1, m_2, \ldots, m_N\}$



t : Interval

c_c : c-level checkpoint time

r_c : c-level recovery time

λ_i : i-level checkpoint time

$$p_0(T) = e^{-\lambda T}$$
$$t_0(T) = T$$
$$p_i(T) = \frac{\lambda_i}{\lambda}(1 - e^{-\lambda T})$$
$$t_i(T) = \frac{1 - (\lambda T + 1)\cdot e^{-\lambda T}}{\lambda \cdot (1 - e^{-\lambda T})}$$

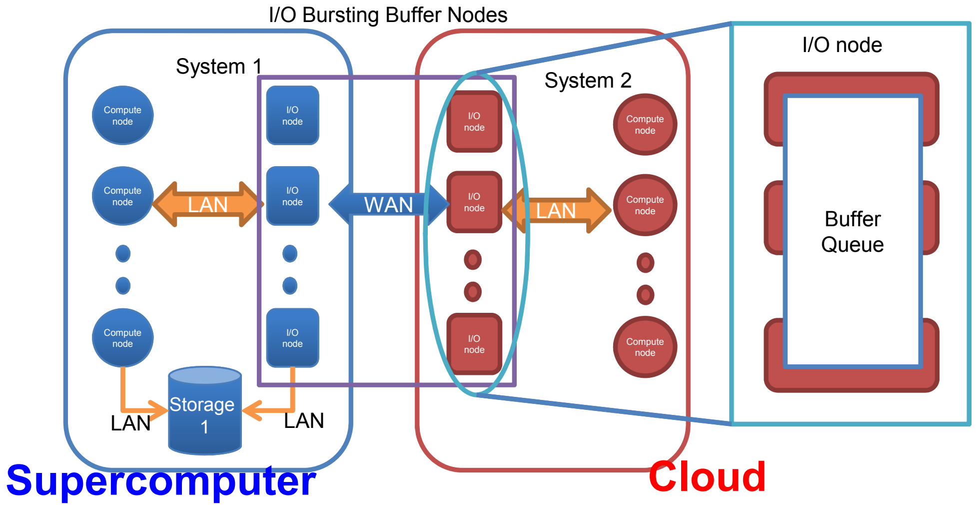
$p_0(T)$: No failure for T

$t_0(T)$: Expected time

$p_i(T)$: i - level failure for T

$t_i(T)$: Expected time

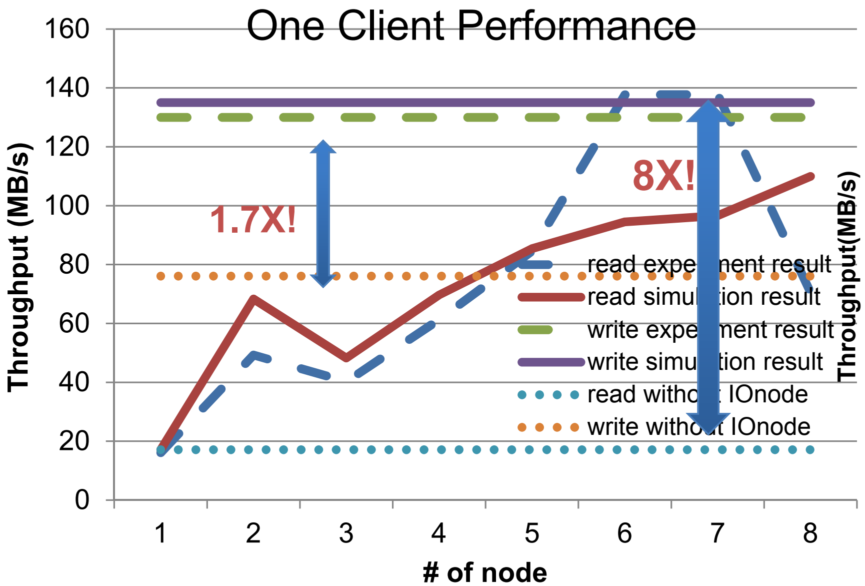
Cloud-based I/O Burst Buffer Architecture (I/O Burst Buffer)
In collaboration talks with Amazon EC2

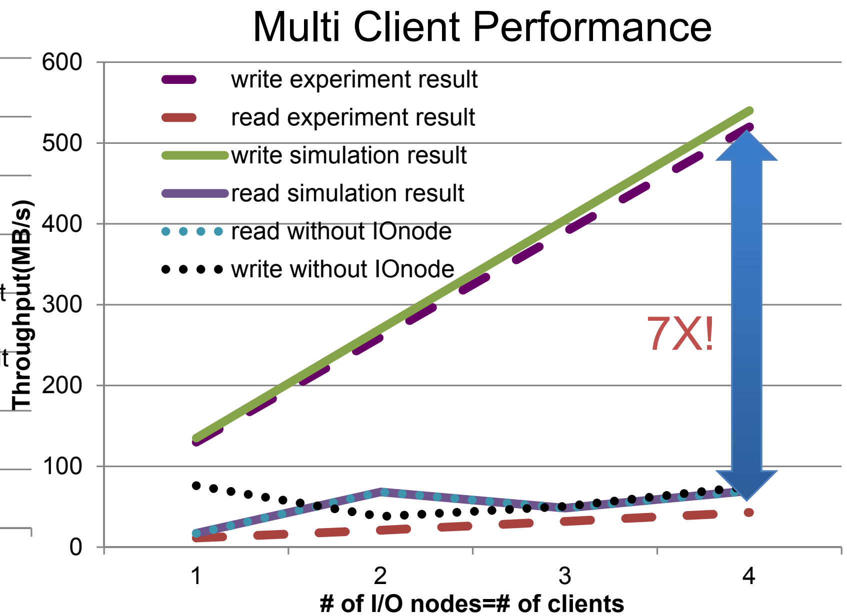


Main idea: using several compute nodes in public cloud as I/O nodes

- Buffer I/O data in the main memory of I/O nodes
- All I/O nodes maintain a on-memory buffer queue
- dynamic burst buffer, # of I/O nodes can be dynamically decided

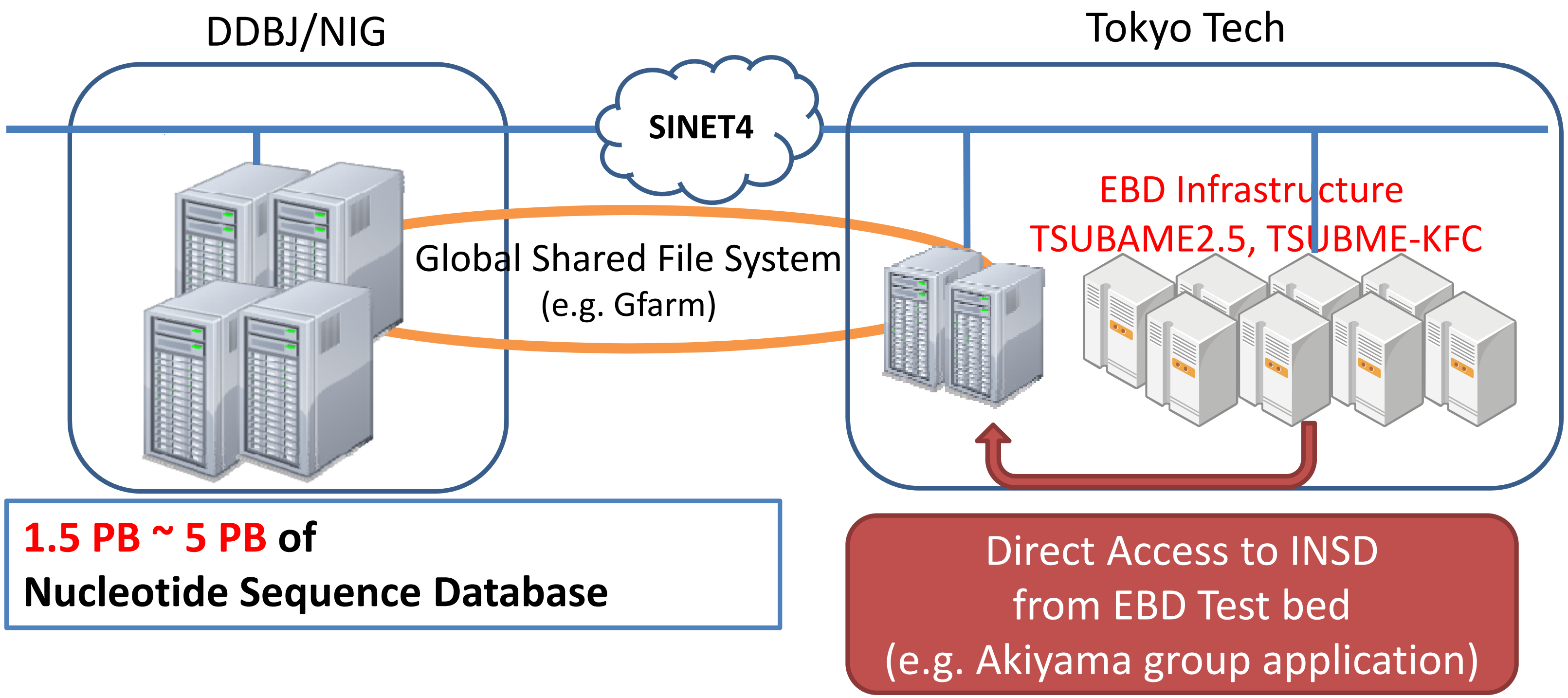
Taking advantage of high throughput of LAN inside public cloud





Extreme Big Data Federation
For Real Analysis

- Target Data Set:
  International Nucleotide Sequence Database
  at DNA Data Bank of Japan



Co-design Example: Genome Science and EBD

Akiyama Group

Metagenome analysis

300% increase per year

Metagenome sequences

O(n)

Meas. data

O(m) ***Reference Database***

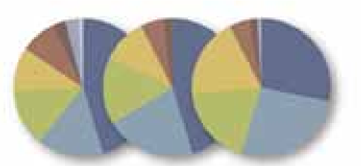
O(m n) calculation

Correlation, Similarity search

Important issues:

1) Allocation of big Measured data

2) Allocation of big References (DB)

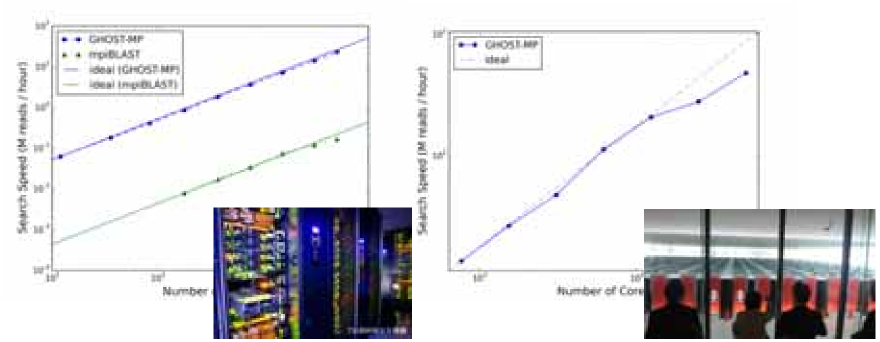


Profiles of environment

GHOST-MP

Ultra-fast pipeline for metagenomic analysis

- OpenMP / MPI
  - load-balancing
  - data dispatcher
- GHOSTX (much faster than BLAST) algorithm



>100 times faster than BLAST and good scaling (49,152 nodes on K computer)

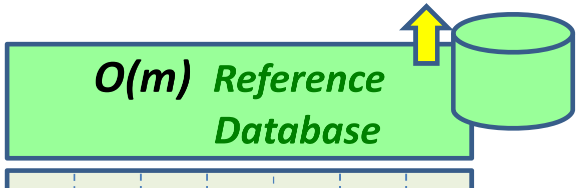
Co-design Example: Genome Science and EBD

Akiyama Group

Metagenome analysis

300% increase per year

Metagenome sequences



Important issues:

1) Allocation of
big Measured data

2) Allocation of

Problem:

Applicability to wider range of infrastructures and software stacks from Clouds to Supercomputers

- Acceleration of BigData kernel operations using many-core accelerators
  - Sort, PrefixSums, Set Operations, etc.
- I/O acceleration using NVM-devices
- Next-gen EBD-oriented interconnects

GHOST-MP

Ultra-fast pipeline for metagenomic analysis

- OpenMP / MPI
  - load-balancing
  - data dispatcher
- GHOSTX (much faster than BLAST) algorithm

Performance analysis on popular Big Data software sub-states/tool-chains

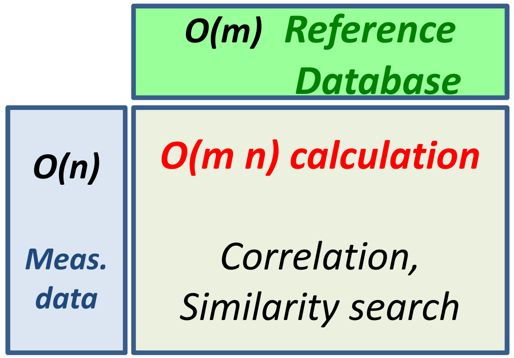
- Convert GHOST-MP to MapReduce
- Benchmarking on variety of high-performance MapReduce implementations on EBD-oriented infrastructures

Size of metagenomic sequencing data

Sequencing data of human oral metagenome

(Subset of Human Microbiome Project data)

Site	# of samples	# of reads (x 10^6)	Total file size (GB)
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



We have performed <u>sensitive homology search against KEGG Genes DB</u> for whole reads (26 billion reads, 5.2TB)

GHOST-MP

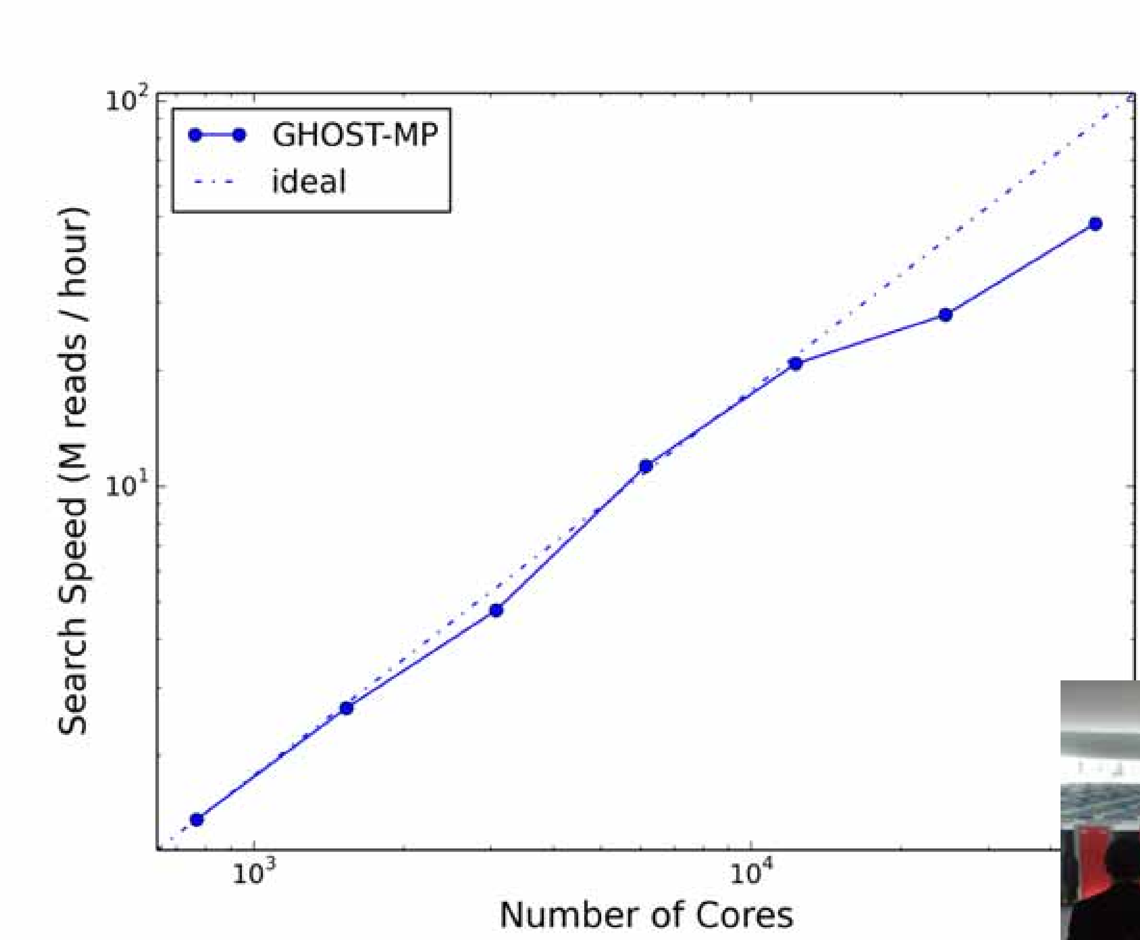
- Massively parallel metagenomic analysis software
  - Thread-level parallelism: OpenMP (same DB, different reads)
  - Node-level parallelism: MPI (different DBs, different read sets)

- **Automatic Load balancer** “mpidp” module included
- **Sophisticated I/O scheme**

TSUBAME2.5 at Tokyo Tech
17304 CPU cores, 4224 GPUs

K-computer at RIKEN
705024 CPU cores, No GPUs

GHOST-MP scalability to extreme scale



Query:
Human tongue dorsum metagenomic shotgun sequencing (SRS078182)

Database:
KEGG Genes

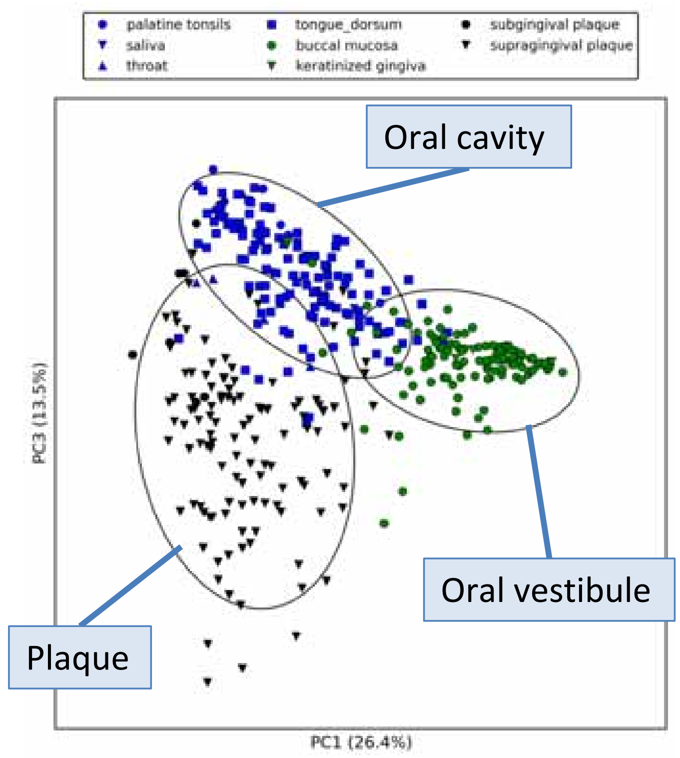
Computer:
K computer
(SPARC64 VIIIfx, 16GB mem)

Weak scaling up to 49152 nodes (= 393216 cores)

Preliminary results: oral microbiome

Principal component analysis of relative frequency profiles

- First three PCs account for 58.7% of total variance
- The samples from same oral sites tend to cluster and the clusters can be clearly distinguished with PC1 and PC3
- Some relative abundance of orthologous groups related to the specific biological function reveals (negative) correlations to the PCs

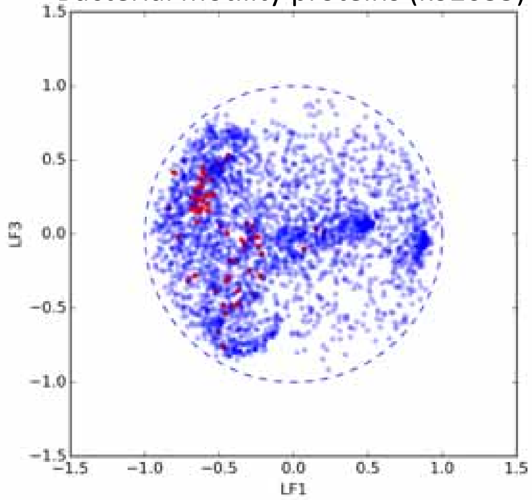


Preliminary results: oral microbiome

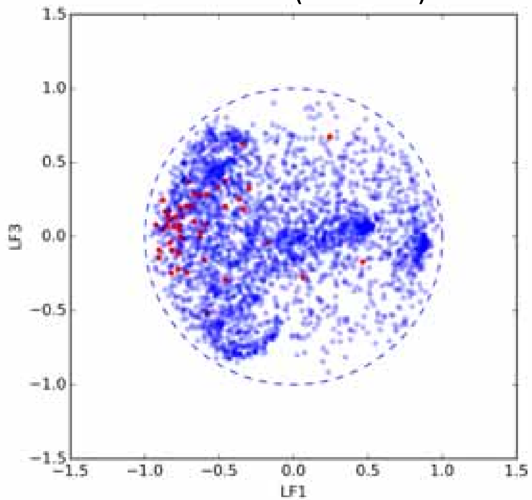
Principal component analysis of relative frequency profiles

- First three PCs account for 58.7% of total variance
- The samples from same oral sites tend to cluster and the clusters can be clearly distinguished with PC1 and PC3
- Some relative abundance of orthologous groups related to the specific biological function reveals (negative) correlations to the PCs

Bacterial motility proteins (ko2035)



Ribosome (ko03011)



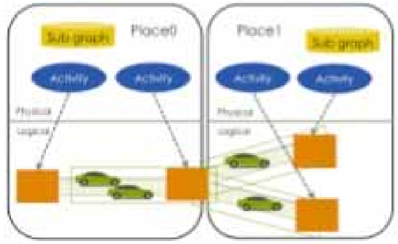
Towards EBD-Driven Social Simulation

A Study on Scalable Architecture and Optimization Methods for Billion-scale Social Simulation

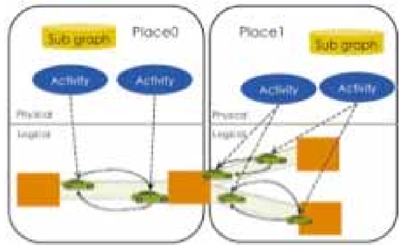
Suzumura Group

- **Motivation & Goal**: Our previous design (ABCA) cannot cope with billion-scale social simulation due workload imbalance and global synchronization issues
- This study is to propose the best architecture that can deal with real-time billion-scale social simulation on the future hardware designed for extremely big data processing
- We optimized the ABCA architecture using active subspace technique in which only active subspace are monitored and processed
- According to the evaluation result, Autonomous Agent (AA) outperforms ABCA at the beginning of simulation but as the simulation progresses, the elapsed time of AA grows exponentially
- In both both strong-scaling and weak-scaling analysis, ABCA shows obviously better performance and scalability over AA.
- We achieved running the simple traffic flow simulation with one billion of agents in almost real time (1.92 second/step) on 1,536 cores of total 128 machines of TSUBAME Supercomputer 2.5

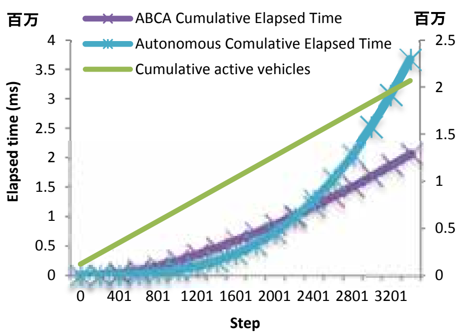
Agent-Based Cellular Automata Architecture



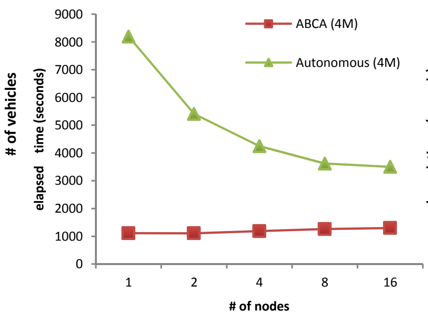
Autonomous Agent Architecture



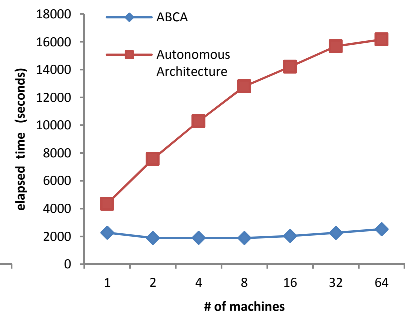
Elapsed at each step
(3600 steps, 4M of agents on Tokyo road network)



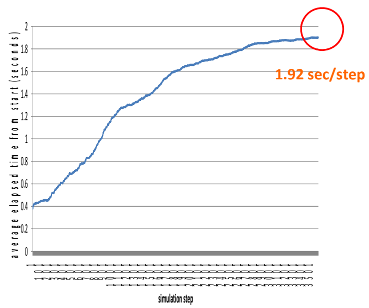
Strong-scaling performance analysis
(3600 steps, 4M of agents on India road network)



Weak-scaling performance analysis
(3600 steps on Tokyo road network)
problem size: 4M, 8M, 16M, 32M, 64M, 128M, and 256M of agents



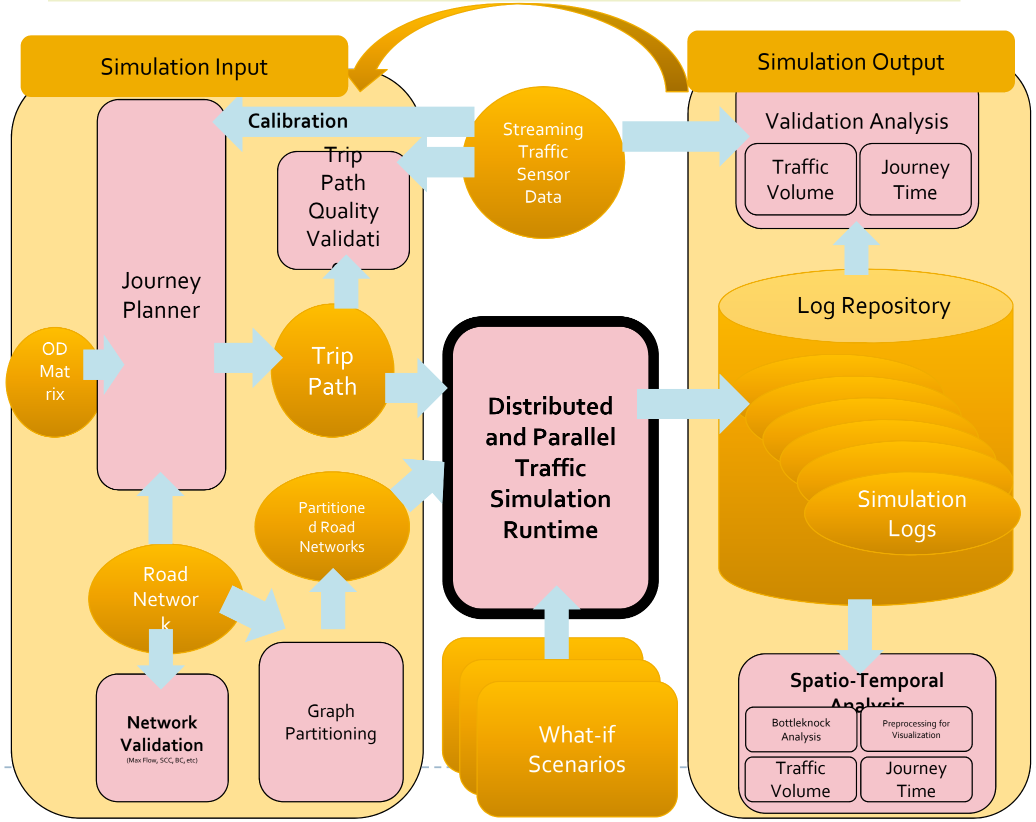
The average elapsed time at each step of optimized ABCA architecture on bill-scale simulation



Designing Large-Scale Traffic Simulation Platforms on EBD Software Stack

Design the next-generation large-scale traffic simulation on top of EBD Software Stack

Iterative Simulation and Data flow



- Based on our experiences on building large-scale traffic simulation platforms, we are currently in the middle of designing the next generation architecture on top of EBD Software Stack
- The left diagram shows a data flow for iterative traffic simulation.
- Currently designing how the simulation platform could be built on top of spatio-temporal enabled EBD Object Data stores.