

次世代スパコン技術を用いた 超大規模グラフ解析と実社会への応用

藤澤克樹

中央大学理工学部 & JST CREST

2012年9月4日 : FIT 2012(法政大学)

e-サイエンス: 超大規模実問題に
挑戦するアルゴリズムと計算技術

スパコンに関する様々な疑問

- 大量に電力を消費するので環境に優しくない？（最近多い疑問）
- 何の役立つのかわからない？（一般の方から）
- 研究開発では使われてもビジネスの世界では必要ないのでは？（ビジネス界から）
- 新規応用分野があまり無い？（理工系一般から）
- スパコンが無くてクラウドで十分？（IT系関係者から）
- スパコンの速度向上はそろそろ限界？（材料、物理、電気系等の方から）



日本以外の国でスパコンの話題が一般向けのテレビ番組、新聞等で扱われることはほとんど無い。実は日本人はスパコン好きな国民性なのでは？

意外と？エコなスパコン



6.6万倍高速

3倍省エネ

<<

4.4万倍データ



Laptop: SONY Vaio type Z (VPCZ1)
CPU: Intel Core i7 620M (2.66GHz)
MEMORY: DDR3-1066 4GBx2
OS: Microsoft Windows 7 Ultimate 64bit
HPL: Intel(R) Optimized LINPACK
Benchmark for Windows (10.2.6.015)
256GB HDD

18.1 ギガ(10^9)フロップス
369 メガ(10^6)フロップス / Watt

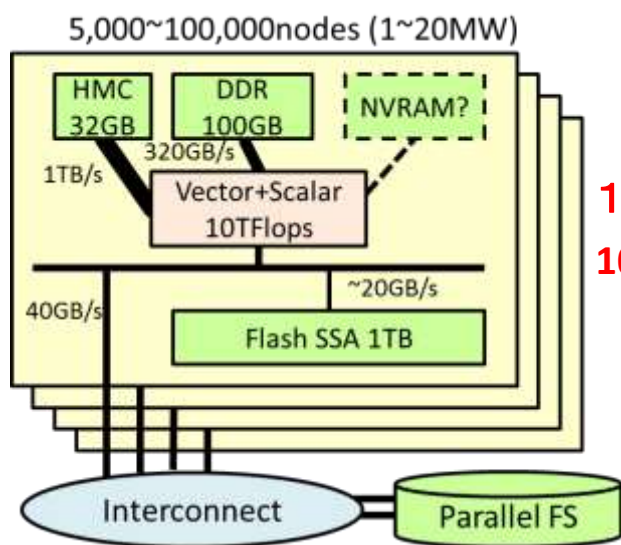
Supercomputer: TSUBAME 2.0 CPU: 2714
Intel Westmere 2.93 Ghz GPU: 4071 nVidia
Fermi M2050 MEMORY: DDR3-1333 80TB +
GDDR5 12TB OS: SuSE Linux 11 + Windows
HPC Server R2 HPL: Tokyo Tech
Heterogeneous HPL 11PB Hierarchical
Storage

1.192 ペタ(10^{15})フロップス
1037 メガ(10^6)フロップス / Watt

大量のデータの高速 & 省電力処理ではやはりスパコンが有利

次世代(ポストペタ)の予想と課題

- 並列数の爆発的増大、不均質化、高密度化
 - プロセッサのマルチコア・メニーコア化
 - ヘテロジニアスプロセッサ(ベクトル・スカラ)の台頭
- 記憶装置の多階層化・大容量化
 - 次世代メモリ、Flash/NVRAM, 並列FS
 - ノード内外のデータ転送性能バランスの逆転



1ノードあたりで
10TFlops, 200W

次世代(100ペタフロップス級)スパコンの
想定アーキテクチャ(2015年頃?)

技術的課題

数千万並列規模の
スケーラビリティ

並列アルゴリズム

局所性の活用

生産性

通信最適化

ストレージの
階層性の深化

不均質性

耐故障

省電力

大規模データI/O

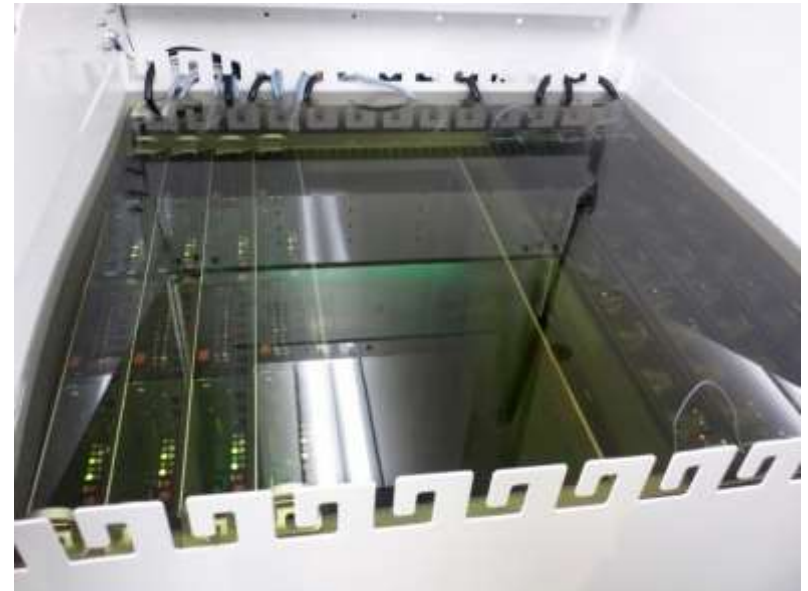
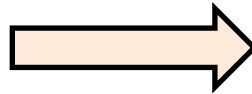


次世代スパコンでの課題をまとめると

- 電力が十分に無いので省電力性が大事
 - 次々世代スパコン(エクサ)までには現在の1000倍の性能を10倍の消費電力で達成する必要がある
 - 発熱の問題も重要なので、効率の良い冷却方法が必要となる



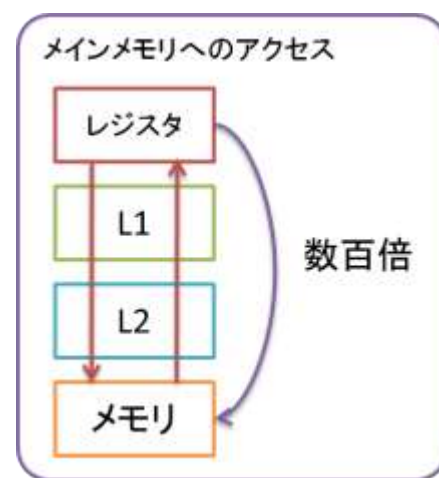
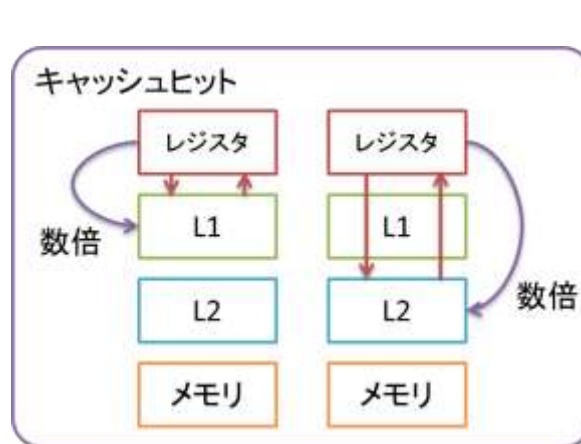
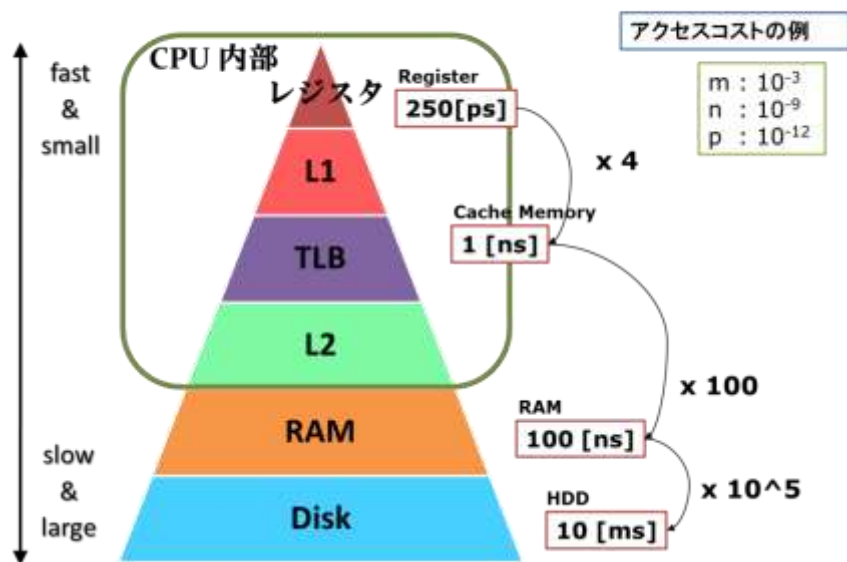
水冷、空冷から
油冷なども採用



メモリアクセスに必要な電力 >> 演算に必要な電力

- 演算はレジスタで行われる
 - メインメモリへのアクセスはコスト大 → キャッシュメモリで改善

メモリの階層構造

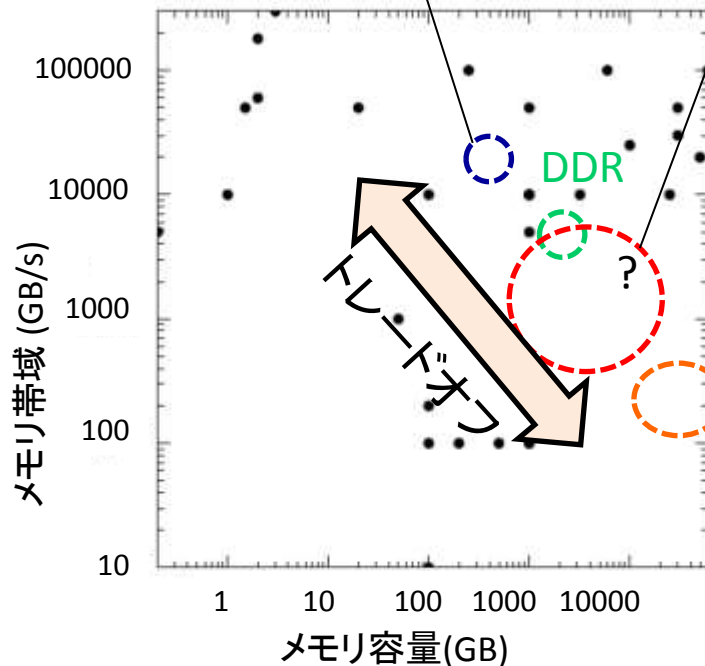
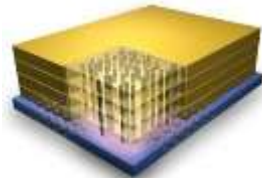


- 1個のデータ移動に必要な電力は、1個のデータの演算に必要な電力の100倍になることもある。
 - 必要以上にデータ移動させたら負け
 - 今後は既存のアルゴリズムの一大転換が必要(高速性と省電力性の両立)**
 - 今後のスパコンの主要なアプリケーションの一つはビッグデータ処理(グラフ解析等も含む)

多様化するメモリアーキテクチャ技術の活用

Hybrid Memory Cube (HMC)

- DRAMチップの3D積層化による高帯域化
- DDRより容量/電力は不利か
- Micron/Intelなど

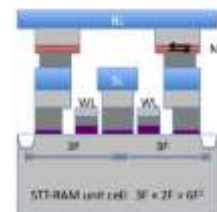


2018年頃の見積もり

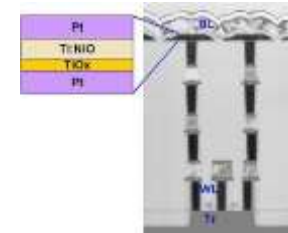
次世代 不揮発メモリ(NVRAM)

- DRAMと異なる記憶方式
- アクセス速度・密度・write耐性まちまち

STT-MRAM



ReRAM



他、PCM,
FeRAM...

高速Flashメモリ

- PCI-Express直接接続・デバイス並列化によりO(GB/s)の帯域
- Solid State Accelerator(SSA)とも



FusionIO社ioDrive

既存HW/SW技術の延長で
多様な要求に応えられるか？



- 科学技術振興機構の戦略創造事業のうち、全体の規模としては最大
- 我が国の社会的・経済的ニーズの実現に向けた戦略目標に対して設定
- 公募により採用された強力な研究チームが国の政策実現に向け研究を推進

■ 研究領域：ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出 ……**スパコンに関する領域**

- 研究総括：米澤 明憲 理化学研究所 計算科学研究機構 副機構長
- 採択件数： H22年度 5 件、H23年度 5 件、H24年度 4 件

■ 研究課題：ポストペタスケールシステムにおける超大規模グラフ最適化基盤

- H23年10月～平成28年3月
- 研究代表：藤澤 克樹 中央大学 理工学部 教授
- 拠点リーダー：鈴木 豊大郎 東京工業大学 大学院理工学研究科 客員准教授
佐藤 仁 東京工業大学 学術国際情報センター 特任助教
脇田 建 東京工業大学 大学院理工学研究科 准教授

大規模グラフ最適化(解析)の重要性

1: 実社会における**本質的課題の解決**に必須な手法

- **大規模災害等で突発的に発生**してリアルタイムに状況が変化し早急な解決が望まれる問題
- 最先端理論(Algorithm Theory) + 大規模実データ(Practice) + 最新計算技術(Computation)による超大規模問題の解決

2: 適用が期待される多くの分野

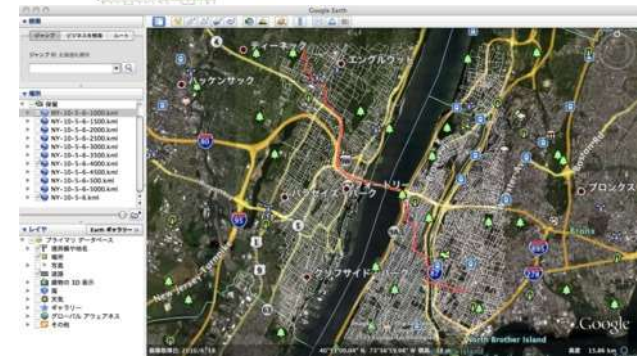
- **交通データに対する経路探索**: 動的に変化する交通量等から高速な重要度判定を行うことによって、交通管制等に活用する
- **ソーシャルネットワークデータに対する解析**: 動的な重要度、影響度の判定。各点の周辺、及び広域内における影響(情報の伝播力)を推定する
- その他: 疫病の拡散、人口の増減、経済動向等の分析。ライフライン等の基盤計画(電力、水、食料)。生命科学系(創薬、遺伝子)。ビジネス系(金融、データマイニング)。安全保障分野(組織構成の解明、事件事故の事前予測)。

3: HPC分野における注目のアプリケーション

- **Graph500 などの新しいベンチマークテスト**が提案。各国ともグラフ探索等のプログラム開発とスパコンでの実行を重視

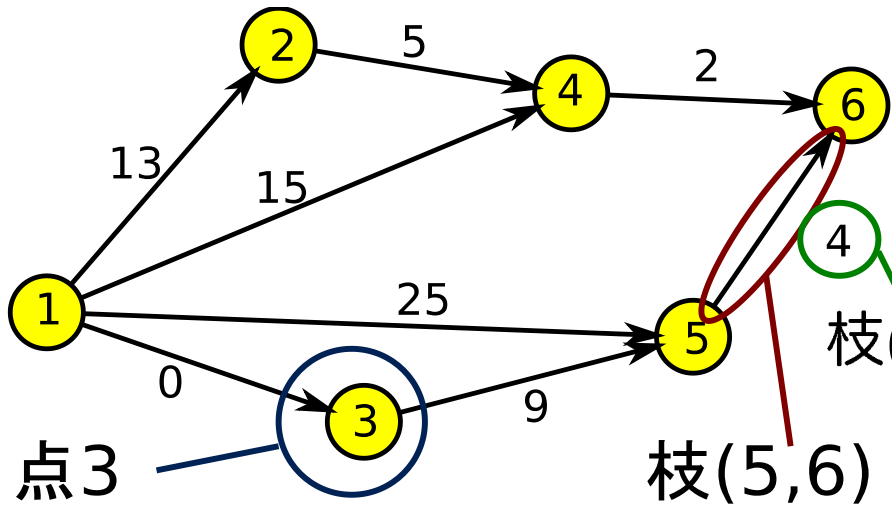


ポストペタスパコン上での
超大規模グラフ最適化システムの提案と開発



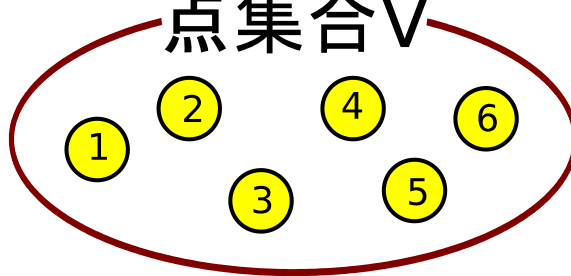
グラフとは？

- グラフは **点** と **枝** で構成される

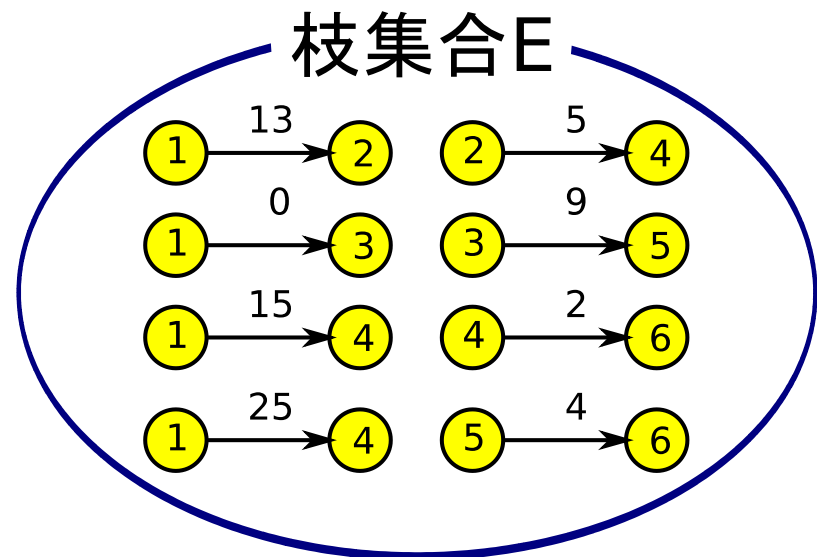


グラフの例（点数6、枝数8）

点集合V

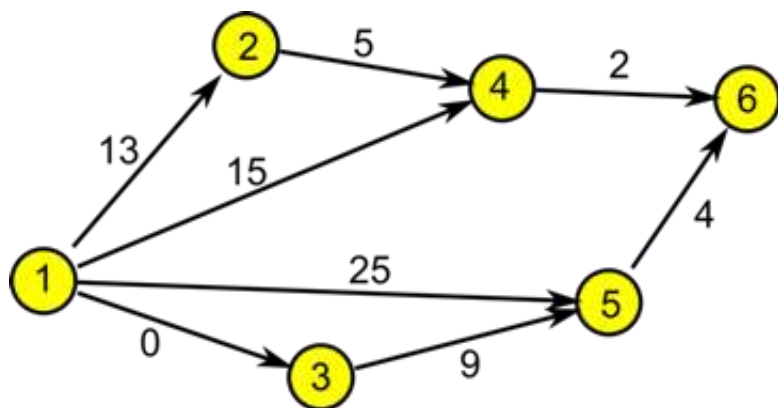


枝集合E

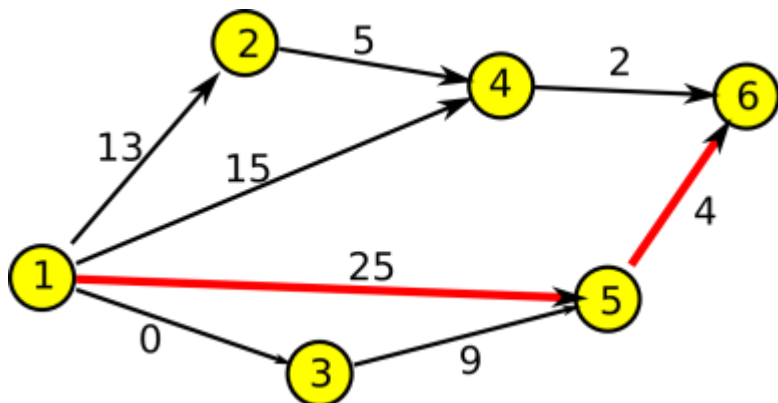


最短路問題とは？

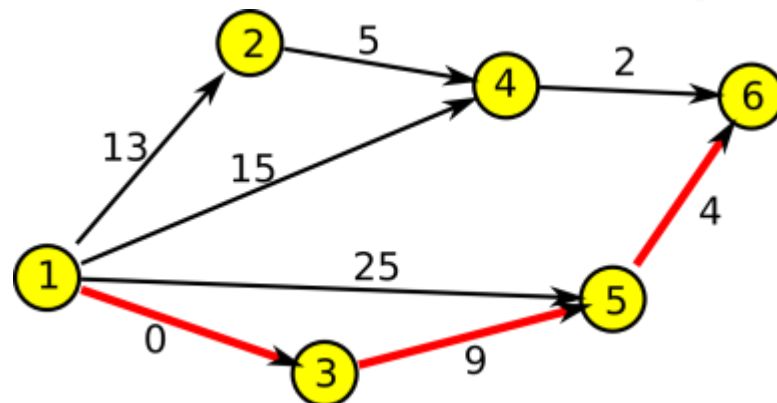
- グラフが与えられたとき、ある2点間の最も短い経路を求める問題
- グラフとは点と枝の集合（道路：点は交差点、枝は交差点間の道路）
- 枝に長さ(重み)を持たせることができる（道路：所要時間、移動距離）



グラフの例（点数：6、枝数：8）



始点1、終点6の最短路ではない経路



始点1、終点6の最短路

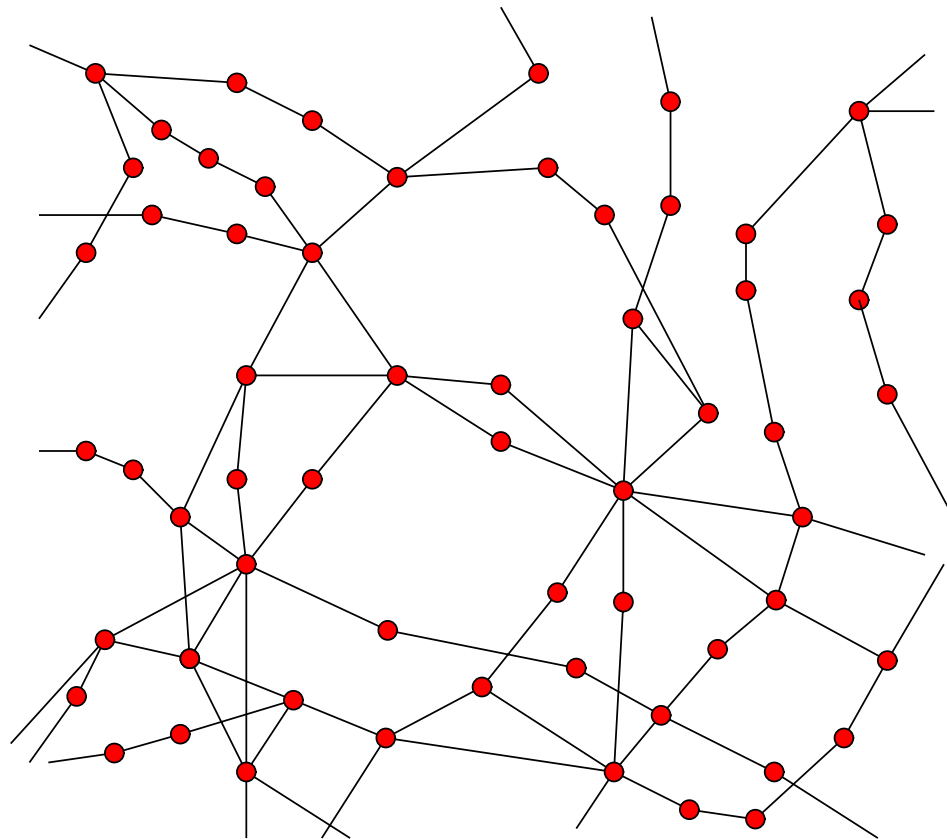
実社会ネットワークのグラフ化

- 道路ネットワーク
 - NY のジョージワシントン橋周辺



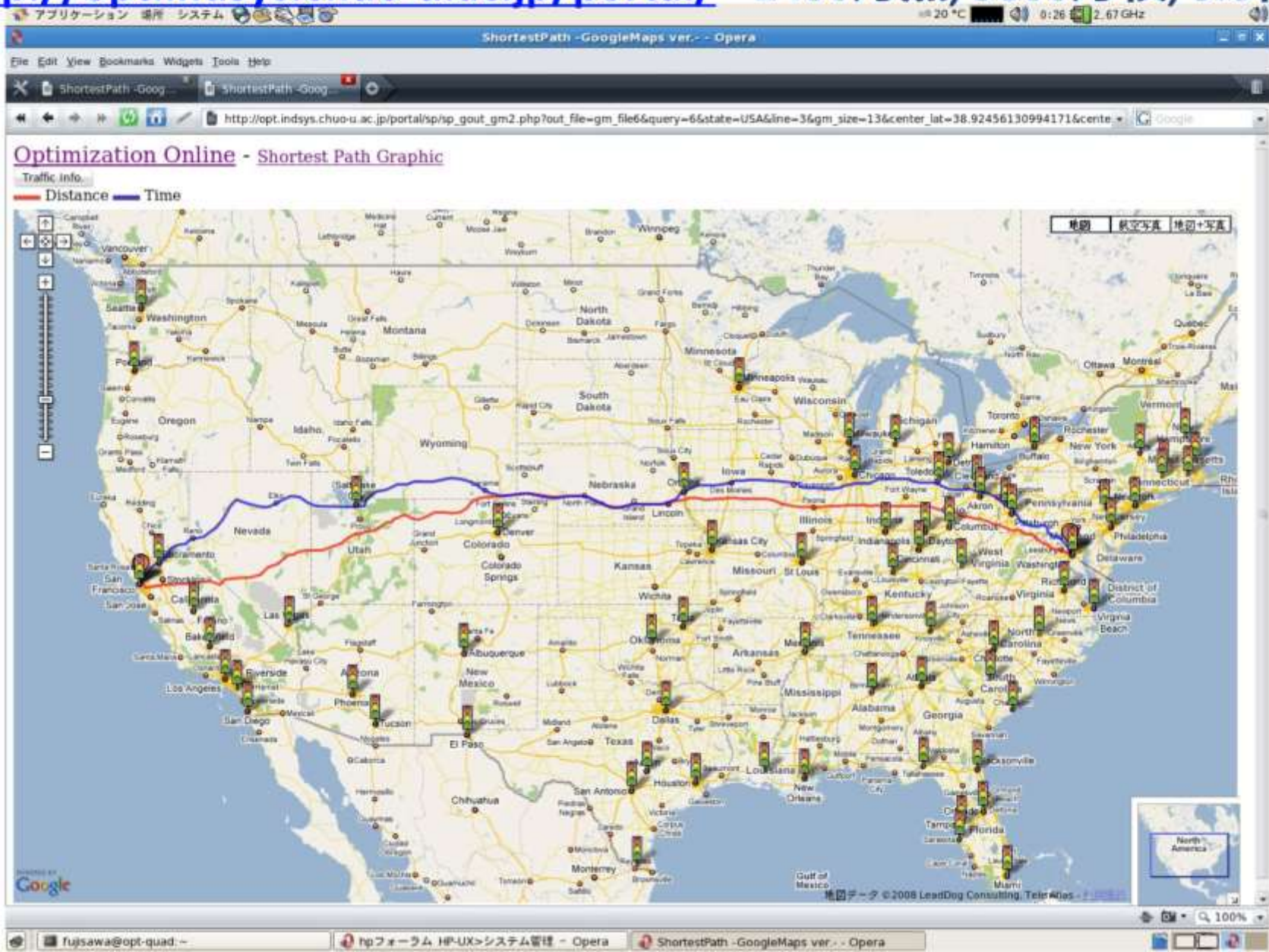
実社会ネットワークのグラフ化

- 鉄道ネットワーク
 - 東京メトロの路線図



全米横断の最短路(赤: 距離; 青: 時間)

<http://opt.indsys.chuo-u.ac.jp/portal/> 2400万点, 5800万枝, 5.8秒

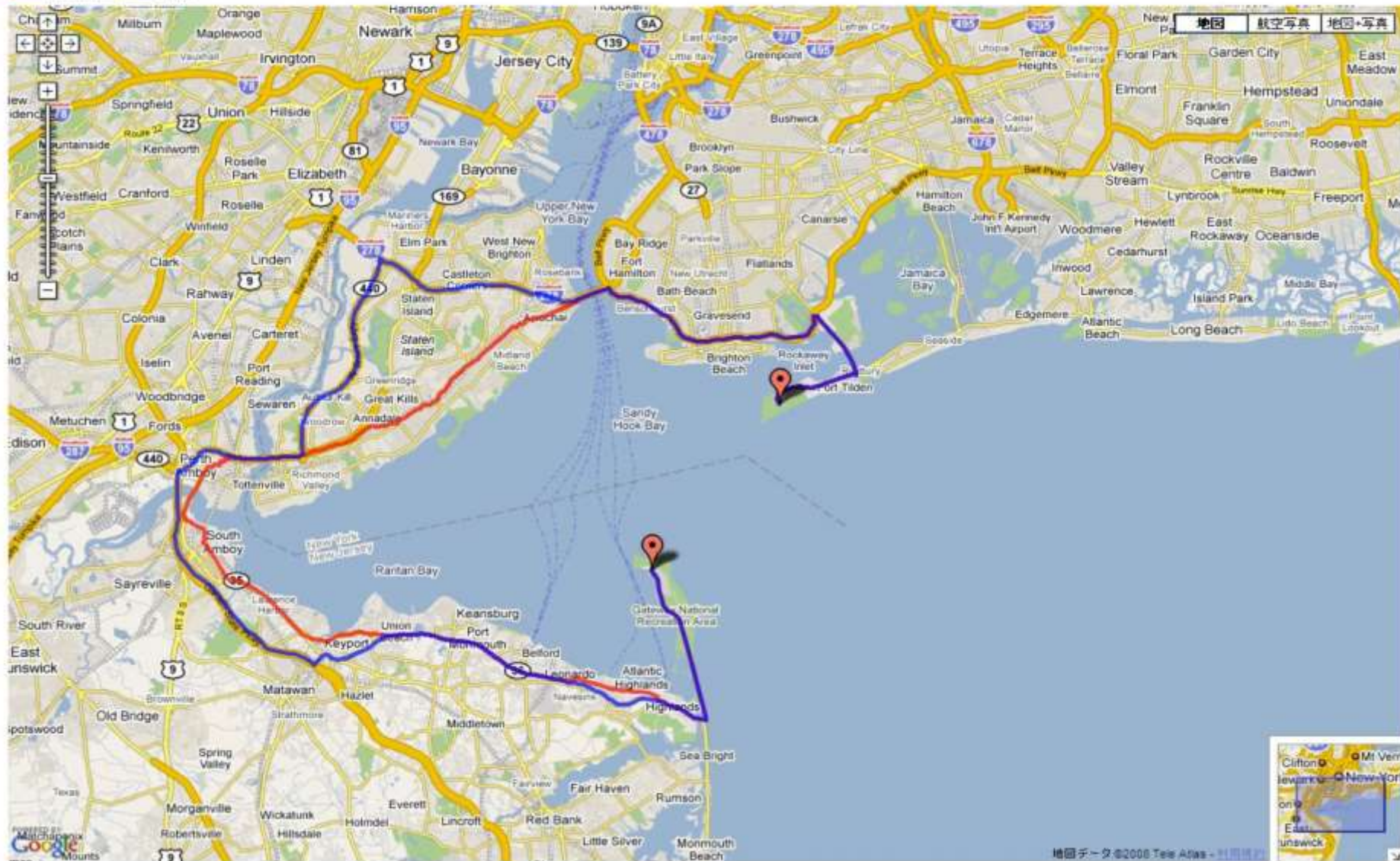


オンライン・ソルバーの出力ページ (距離優先・時間優先の両方を同時に出力)

[Optimization Online](#) - [Shortest Path Graphic](#)

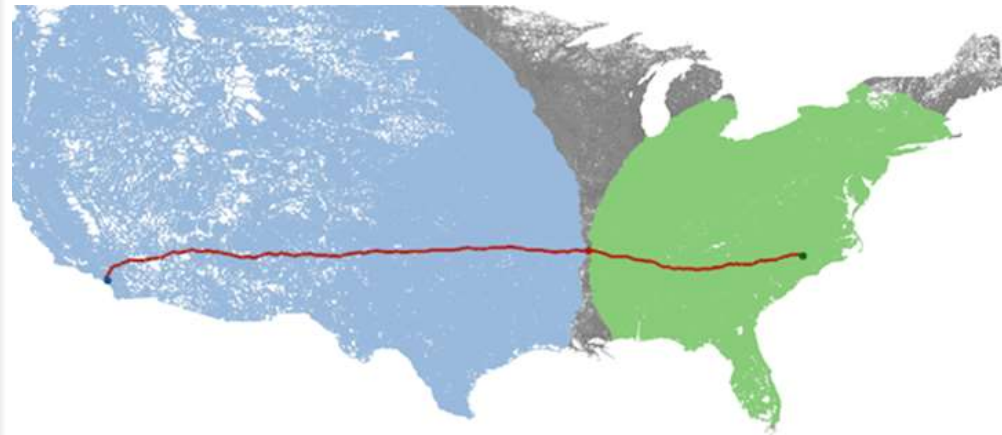
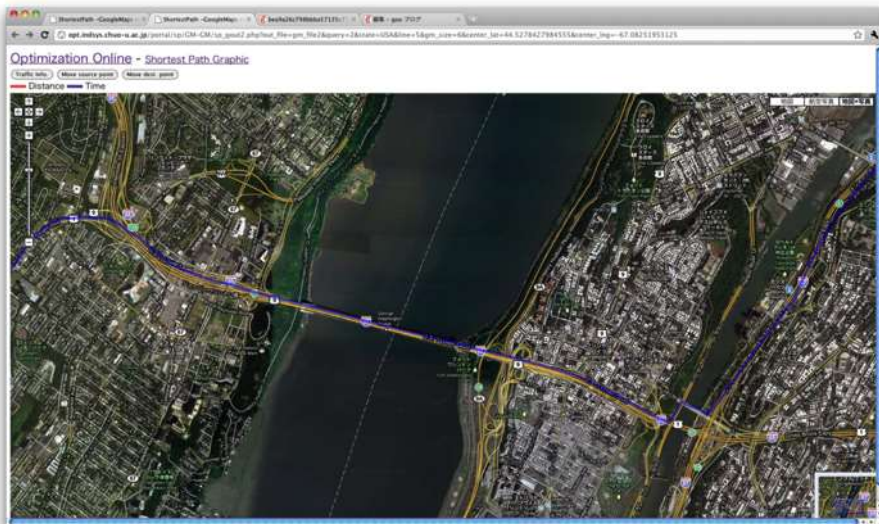
Traffic Info

Distance Time



超大規模ネットワークにおける高速探索技術の開発

- 超大規模ネットワークデータ (点数10億以上)が対象
- 最短路(1対全), グラフ探索(深さ優先、幅優先等)
- グラフ探索の局所的な評価では優先キュー(ヒープ木)を用いる⇒実行時間、メモリ消費量が安定的
- 実行時間(1対全最短路計算: CPU : Intel Xeon 5670 2.93GHz)
 - 全米グラフ 3秒 (点数 $n = 23,947,347$ 、枝数 $m = 58,333,344$)
 - 超大規模グラフ 870秒 (点数 $n = 10$ 億、枝数 $m = 20$ 億)
- 多くの応用(交通ネットワーク, e-コミュニティのネットワーク、ロジスティクスネットワーク)



ソーシャルネットワーク

- 問題

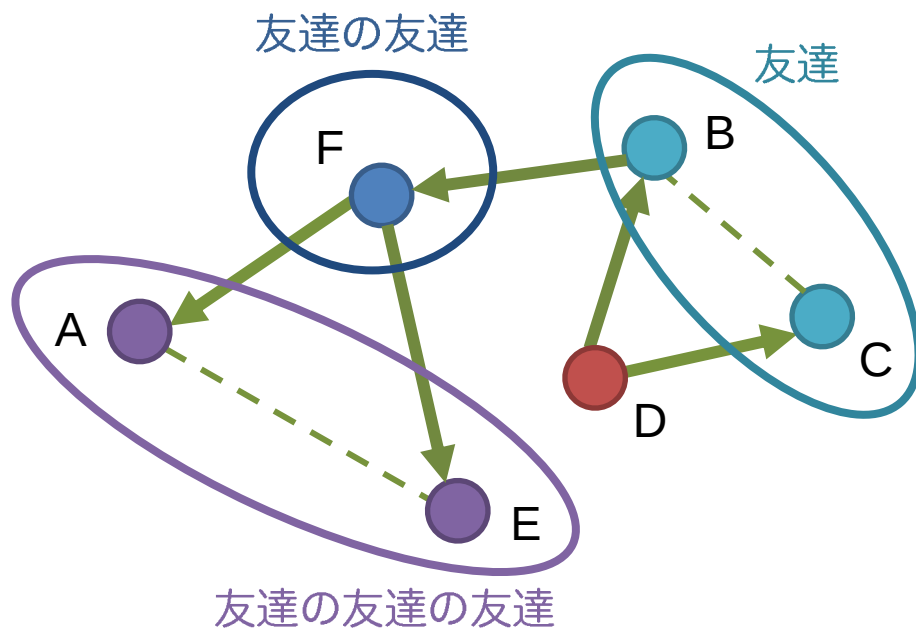
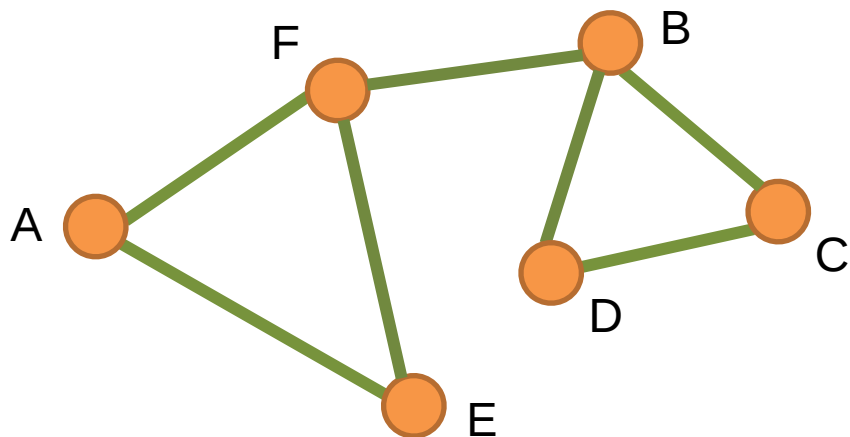
- A 君の友達は E さんと F 君
- B 君の友達は C さんと D さんと F 君
- C 君の友達は D さん
- E 君の友達は F 君
- さて D さんの友達の友達は誰？

ソーシャルネットワーク

- 問題

- A 君の友達は E さんと F 君
- B 君の友達は C さんと D さんと F 君
- C 君の友達は D さん
- E 君の友達は F 君
- さて D さんの友達の友達は誰？

答) F 君

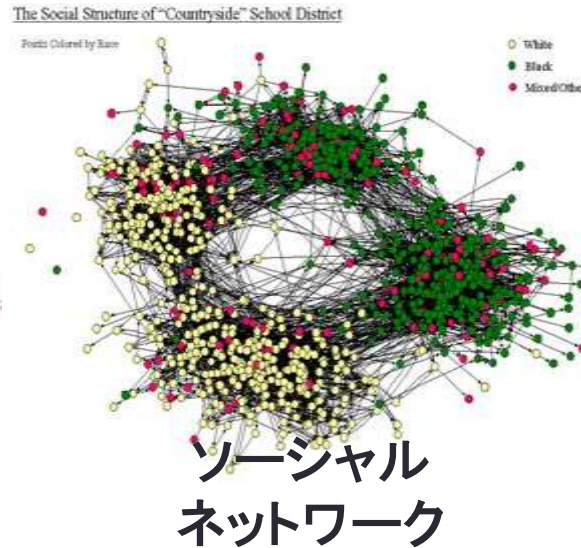


大規模グラフ処理の必要性



インターネット

lumeta.com



ソーシャル
ネットワーク



lumberjaph.net

- バッチ的な大規模グラフ処理の研究
- 一方で、リアルタイムでの大規模グラフ処理の要求
 - 市場調査、株式市場、など ➡ データストリーム処理による解決



商品-ユーザー、株式-売買人、有名人-ファン

データストリーム処理とは

ネットワーク、その他の技術の発達による
膨大なリアルタイムデータ...

例：ウェブ、スマートフォン、株式市場

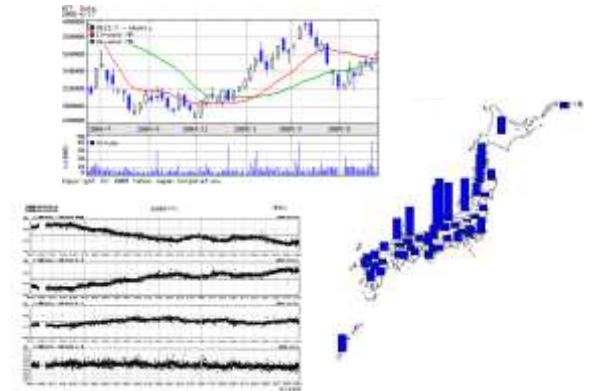


リアルタイム
で活用したい



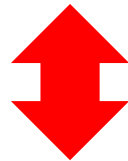
ストレージに蓄積せずに
メモリ上のみで逐次処理

即時性の高い
情報が手に入る！



バッチ処理

蓄積したデータに
詳細な解析処理



データストリーム処理

バッチ処理

即時の応答が可能

蓄積後の応答

データの一部を参照
=> 解析精度の限界

データ全体を参照
=> 詳細な解析

ストレージは不要

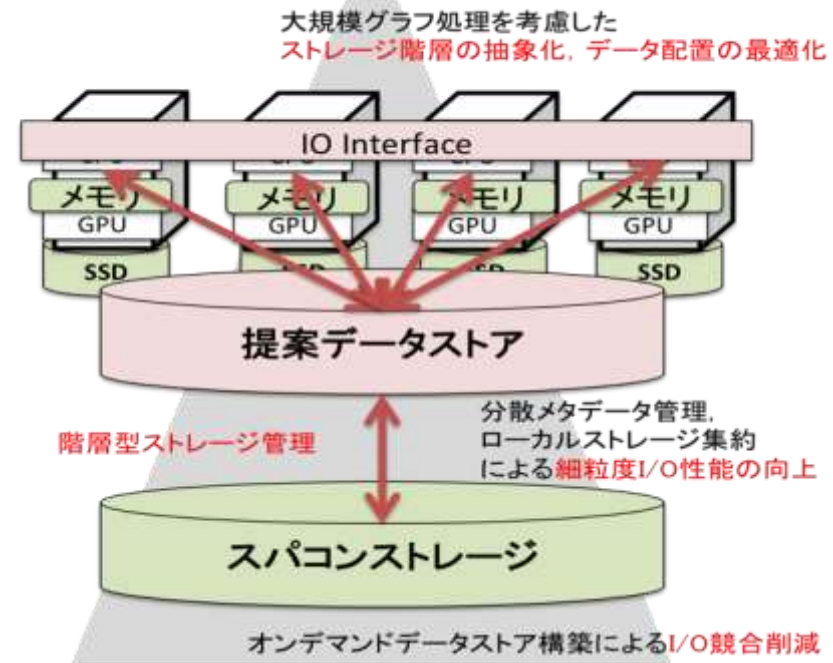
大規模ストレージ

大規模グラフ処理向け階層型データストア

- 大規模グラフに特化したI/Oインターフェース・DB
 - 頂点・辺情報への高速アクセス
 - 隣接頂点・辺情報の**ダイレクトポインタ・属性情報管理**
 - アルゴリズムに応じた**データ配置の最適化**
 - ストレージ階層の**抽象化**
 - 次世代メモリ, Flash/NVRAM, 並列FS (Lustre, GPFS, etc.)
- 階層型ストレージ管理
 - スパコンスケジューラと協調し**オンデマンド**にデータストアを構築
 - ローカルストレージ (Flash/NVRAM) の集約
 - 参照データのみローカルストレージに保持
 - I/Oの分離によるスパコンFS全体に対する**I/O競合を解消**
- 細粒度I/O性能の向上
 - 分散メタデータ管理による**IOPS向上**, **メタデータアクセス遅延の削減**
 - ノード内Flash/NVRAMの集約による**スループットの向上**

- メモリ帯域を最大限活用 (ノードあたり**0.2TB/s~1TB/s**)
- ファイルのストアはローカルFlashへ (ノードあたり**1M~1G IOPS**)

オンデマンドにノードを集約
合算により性能向上
(10Kノード規模)

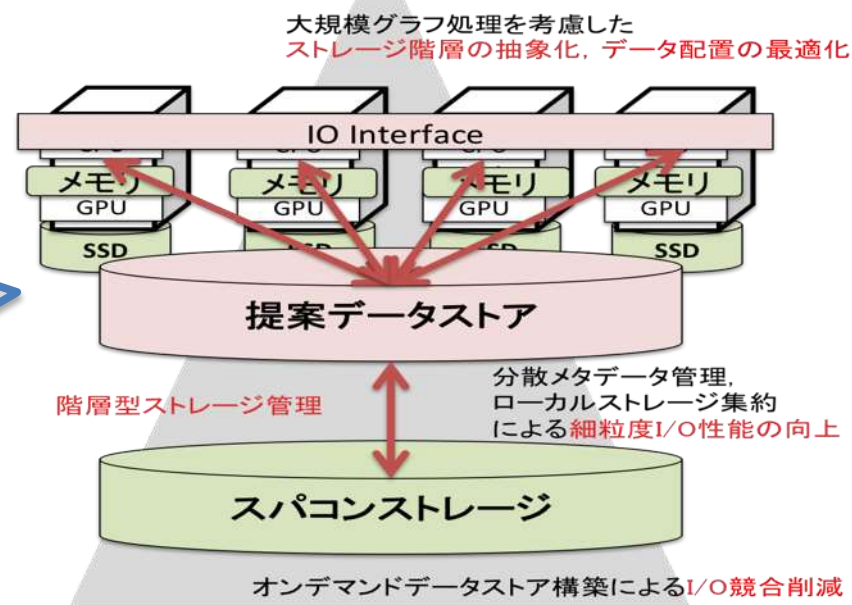
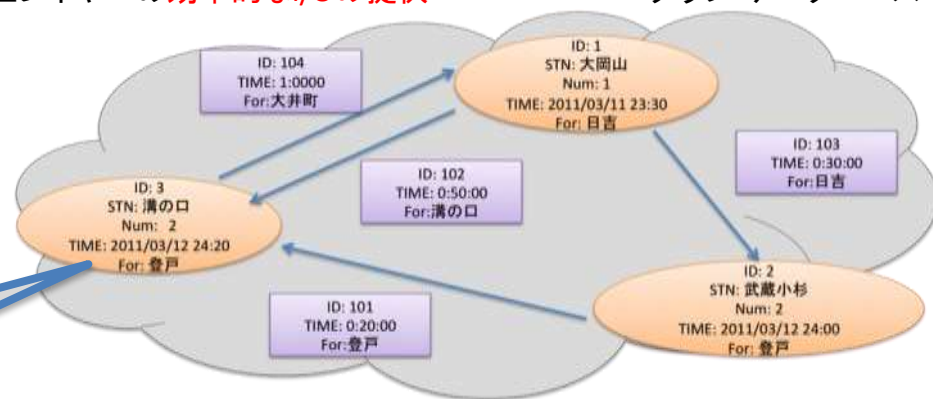


- 2^{42} 頂点、数PB以上のワーキングデータセット
- 現実的には**属性情報**や**複数グラフデータ**を格納するので**数十~数百PB**規模のストアが必要

GraphStore と全体の中での位置付け

大規模グラフに対する属性情報の取得
上位レイヤへの効率的なI/Oの提供

グラフ・データベース



大規模グラフ解析：緊急に取り組むべき課題と実社会へのインパクト

防災計画策定

交通・災害復興・
避難・ロジスティクス

エネルギー・省電力

ソーシャル
ネットワーク解析

インターネットマップ

Image: Illustration by Mirko Ilic
ニューラルネットワーク

ソーシャルネットワーク

サイバーセキュリティ

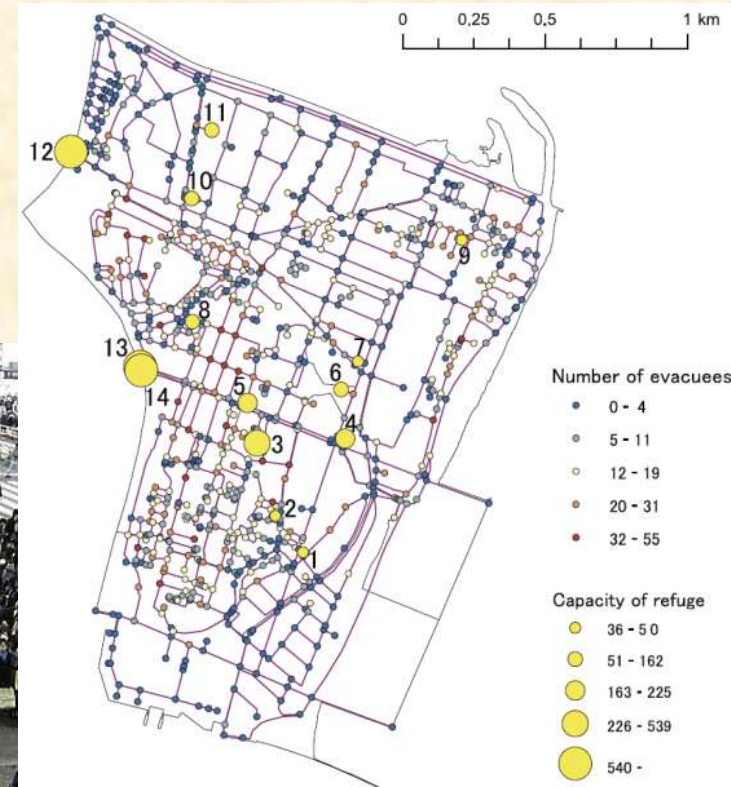
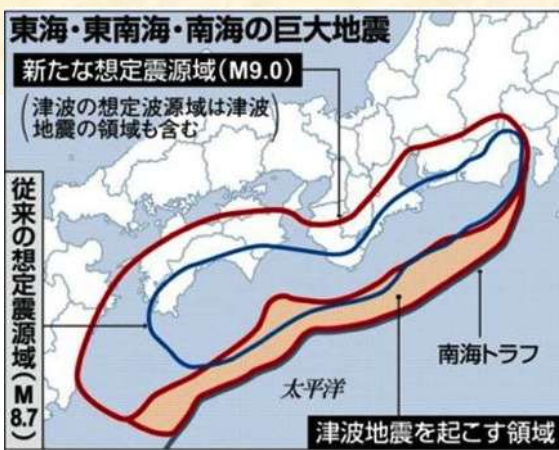
交通(時空間)ネットワーク

スモールワールドやスケールフリー性などの特性を持つ超巨大グラフ(1兆点以上)上での
高速探索技術(アルゴリズム、実装技術、高速計算): 国際会議 ISC12 での
Graph500 において世界第三位(8.8兆枝のグラフに対する幅優先探索で約12秒)

緊急避難計画

- 想定外の規模の超巨大地震に対する津波を考慮した安全性向上が急務
- 大地震(東海, 東南海, 南海)に対して, 大都市を中心とした避難シミュレーション
- 津波避難ビル等への最速避難時間や避難経路の計算
- 超大規模データ
 - 空間: 大都市の道路ネットワークを細街路まで考慮(東京都: 50万頂点)
 - 時間: 細かな時間単位(1秒単位)
 - 対象: 数百万の人の動きを考慮
- マルチシナリオ
 - 被災予測を確率的に利用(モンテカルロ法)
 - 時間を考慮したシミュレーション

大規模ネットワークに対する高速な最大流
& 最短路アルゴリズムの開発が必要



緊急避難計画の数理モデル

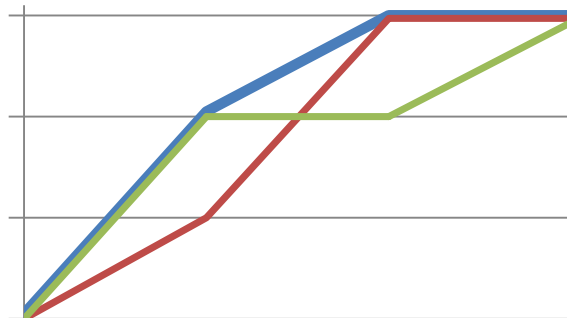
Universally Quickest Flow

どの時点でもそれまでの避難完了人数が最大となる実行可能な動的フロー

— Universally Quickest Flow
— Quickest Flow
— Flow

避難完了人数

時間



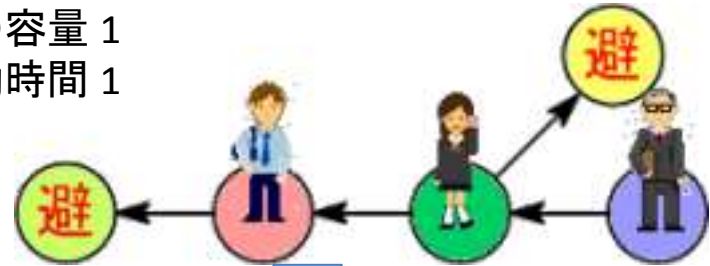
避難所の容量制約がない

⇒ Universally Quickest Flow は常に存在

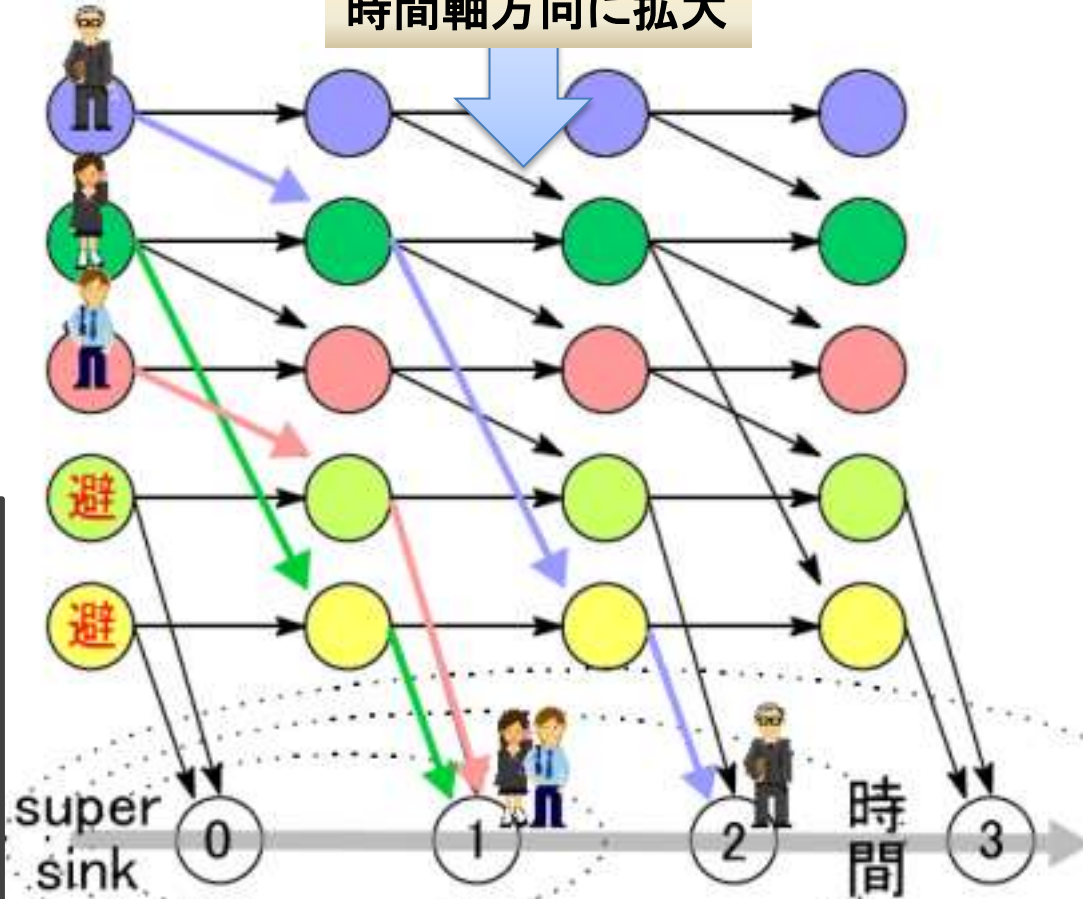
⇒ 時間拡大ネットワークにおける
Lexicographic Maximum Flow に帰着

大規模ネットワークに対する高速な最大流
& グラフ探索アルゴリズム開発が重要

枝の容量 1
移動時間 1



時間軸方向に拡大

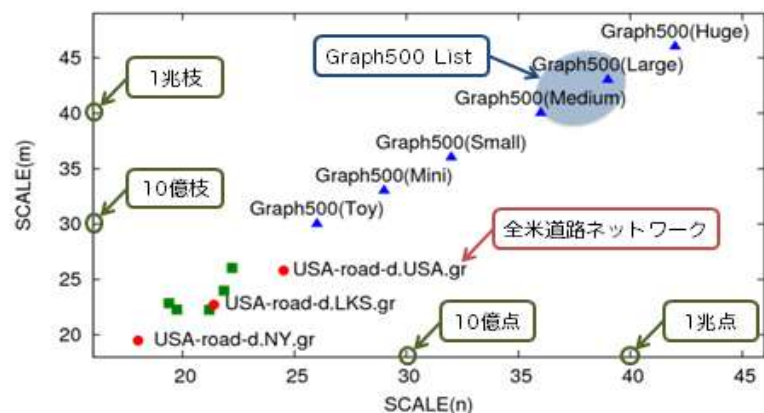




Graph500 (<http://www.graph500.org/>)

超大規模グラフの探索能力で計算機を評価する新しいベンチマーク

Graph500 List

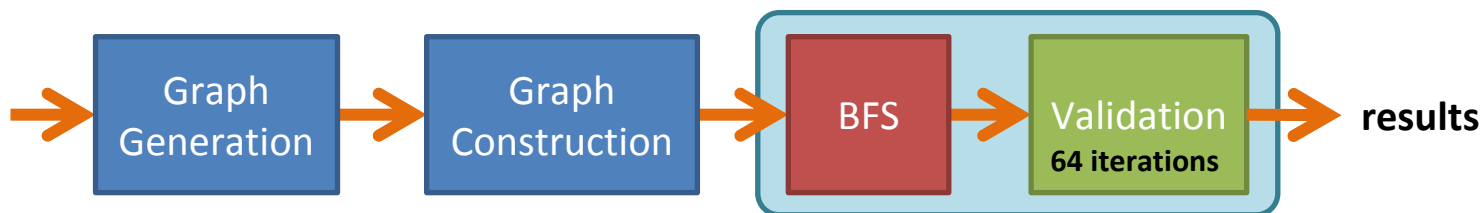


- 現在の指標 TEPS(Traversed Edges Per Second)
- 多様な応用分野 (Cybersecurity, Medical Informatics, Social Networks, Data Enrichment, Symbolic Networks)
- 三つのカーネル
 - concurrent search(Breadth First Search : BFS)
 - optimization (Single Source Shortest Path)
 - edge-oriented (Maximal Independent Set)
- 超大規模グラフへの適用
- 省電力性を競う Green Graph500 ベンチマーク
 - <http://green.graph500.org/>

- 人工的に生成した Kronecker Graph に対する幅優先探索 (BFS) 性能
 - 平均次数が 16 ($=m/n$) の重みなし無向グラフ(平均次数 32 の有向グラフとして扱う)
 - パラメータ **SCALE** を用いてグラフ規模を **点数** 2^{SCALE} **枝数** $2^{\text{SCALE} + 4}$ と決定する
 - 例) SCALE30 のとき、10 億点 172 億枝の無向グラフ(有向: 344 億枝)

Input parameters

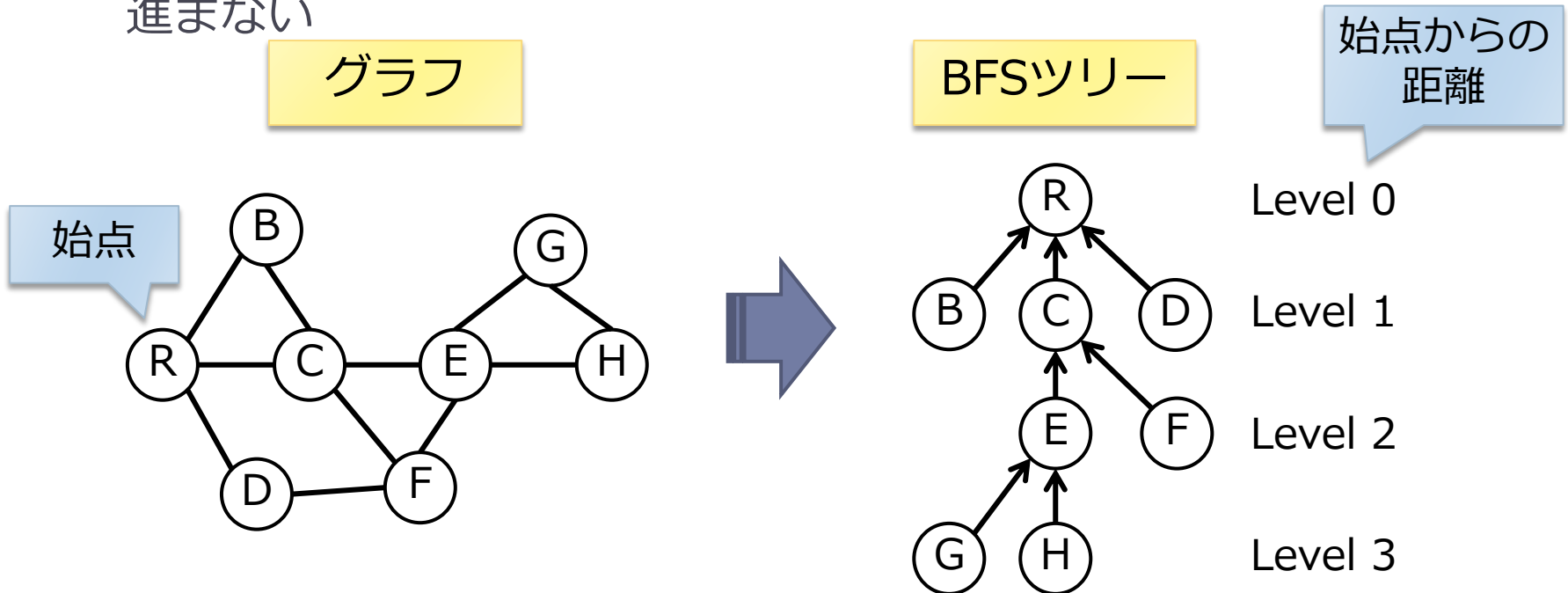
- SCALE
- edgefactor (=16)



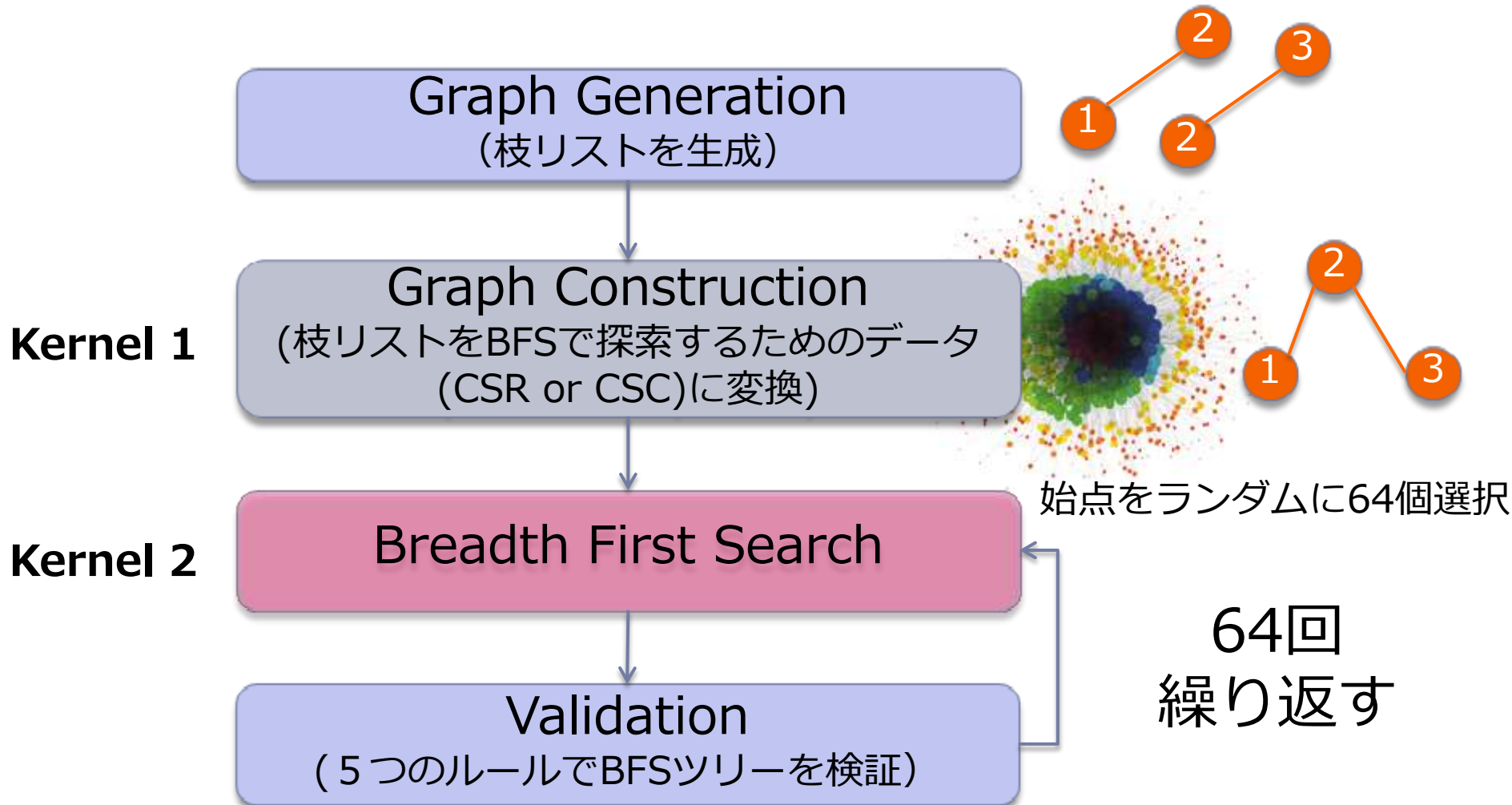
- メインメモリに対する質(速度)・量の両面への厳しい要求
 - 高速化だけでなく、省メモリ化も十分に考慮する必要がある

Level Synchronized BFS

- ▶ BFSの出力: BFSツリー
 - ▶ Level: 始点からの距離
- ▶ Level Synchronized BFS
 - ▶ Graph500のすべてのリファレンス実装がベースとしているBFS(Breadth First Search)アルゴリズム
 - ▶ Level Synchronized: 現在のレベルが終了するまで次のレベルに進まない



Graph500ベンチマークの実行順



※CSR: Compressed Sparse Row, CSC: Compressed Sparse Column

■ Cybersecurity

- 150億ログエントリー / 日
- Full data scan required

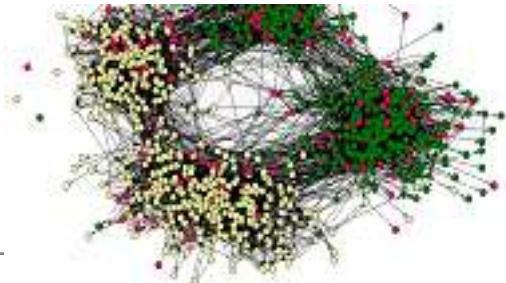


■ Medical Informatics

- 患者数 5千万 × 20 ~ 200 項目 / 患者 = 数十億データ
- Entry resolution important

■ Social Networks

- 例) facebook twitter
- データ量に上限無しの状態



■ Data Enrichment

- ペタ(1000兆)バイト級データ
- 例) Maritime Domain Awareness
 - Hundreds of Millions of Transponders
 - Tens of Thousands of Cargo Ships
 - Tens of Millions of Pieces of Bulk Cargo
 - May involve additional data

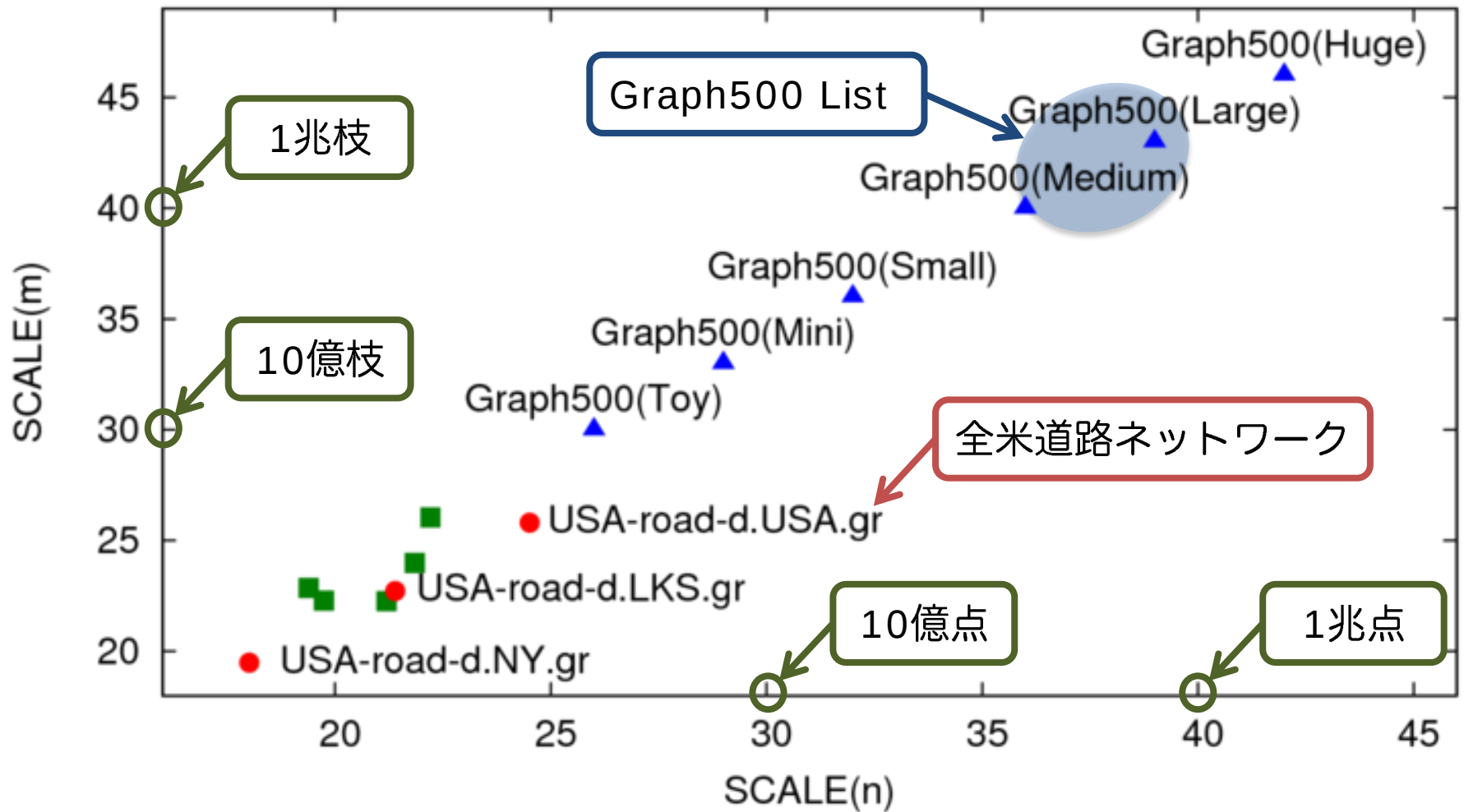
■ Symbolic Network

- 例) Human Brain Project : 人間の脳
- 890億個のニューロンとその接続 100兆枝
- 2023年エクサスパコン上に再現



Image: Illustration by Mirko Ilic

Graph500 List



The 4th Graph500 List (June)

JST CREST チームの成果

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
7	NASA-Ames / Parallel Computing Lab, Intel Labs	Pleiades - SGI ICE-X, dual plane hypercube FDR infiniband, E5-2670 "sandybridge"	1024	16384	34	270.33
8	NERSC/LBNL	XE6	4817	115600	35	254.07
9	NNSA and IBM Research, T.J. Watson	NNSA/SC Blue Gene/Q Prototype II	4096	65536	32	236.00
33	Chuo University	Intel(R) Xeon(R) CPU E7-4870 @ 2.40GHz (4 sockets, 40 cores) / 1 node	1	40	29	5.68



- 東工大 TSUBAME 2.0**
- 世界3位(3rd Graph500 List)
 - 世界4位(4th Graph500 List)
 - 世界初の GPU によるグラフ探索の大規模実装



Graph500 と電力性能



Graph500 ベンチマーク

「1秒あたり幅優先探索性能 TEPS (traversed edges per second)」

Green Graph500 ベンチマーク

「消費電力あたりの幅優先探索性能 TEPS/kW」

- 最新の HPC 用プロセッサは消費電力あたりの幅優先探索性能 TEPS/kW が非常に高い
- 低消費電力プロセッサは必ずしも消費電力あたりの性能が良いとは限らない

SCALE 26 (点数 $67,108,864 \div 2^{26}$ 、枝数 $2,147,475,066 \div 2^{31}$)

計算機(並列数)	探索枝数/秒 (TEPS)	実行時間* (秒)	有効電力 (W)	皮相電力量 (kVAh)	CO2排出量 (kg)	TEPS/kW
WestmreEX(80)	10.066 G	207.21	1000.6	0.06	0.02	10.060 G
SandyBridgeEP(32)	5.607 G	370.34	362.0	0.04	0.01	15.489 G

SCALE 21 (点数 $2,097,152 \div 2^{21}$ 、枝数 $67,105,930 \div 2^{26}$) 全米道路ネットワーク規模に相当

計算機(並列数)	探索枝数/秒 (TEPS)	実行時間* (秒)	有効電力 (W)	皮相電力量 (kVAh)	CO2排出量 (kg)	TEPS/kW
WestmreEX(80)	3.967 G	6.06	854.1	0.00	0.00	4.645 G
SandyBridgeEP(32)	5.293 G	6.23	310.3	0.00	0.00	17.058 G
Atom2(2)	0.163 G	925.16	94.6	0.03	0.01	1.723 G
MacbookAir(4)	0.782 G	375.93	20.3	0.00	0.00	38.133 G

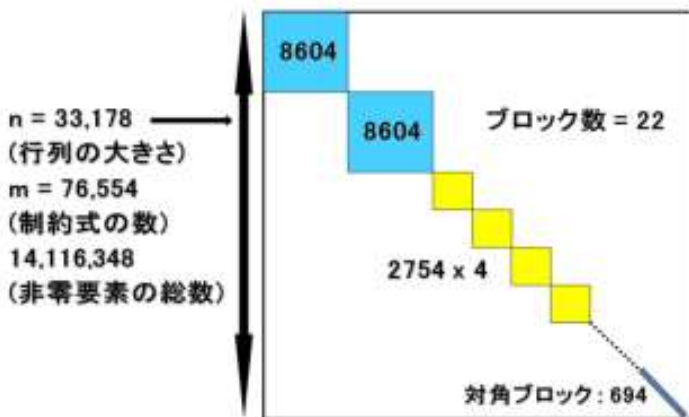
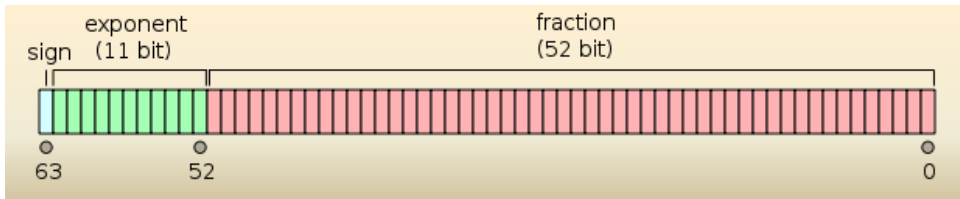
* 実行時間には 64 回の BFS とその検証時間が含まれる

超大規模最適化問題に対する高速計算システムの構築と評価

- **世界最高レベルの性能を持つ最適化ソルバー**(グラフ探索(最短路、幅優先等), 半正定値計画問題(SDP), 混合整数計画問題(MIP)) の開発
- グラフ問題等の良質な近似解を高速に探索する高性能最適化基盤の開発
- 高精度線形代数演算ライブラリ: 線形代数演算は問題規模が大きくなれば数値的に不安となるので, 高精度ライブラリをアクセラレータにより高速化する

高精度線形演算ライブラリの開発と高速化

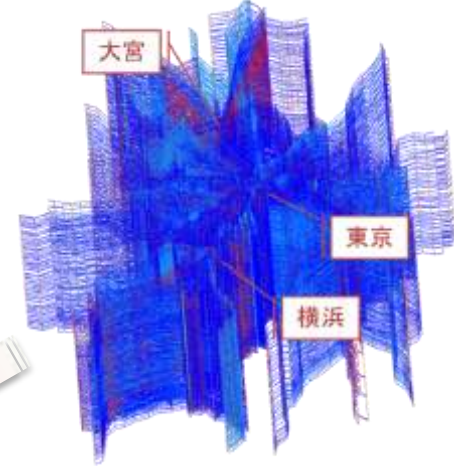
超大規模ネットワークにおける高速探索技術の開発



巨大 SDP のデータ構造

疎性の追求と前処理さらに並列計算の適用
(大規模スレッド並列 & CPU + GPU による高速化) 変数と制約条件 100 万以上

数千万頂点のグラフ
(1スレッドあたりのメモリ要求量1Gbytes)
に対して、同時に数百万スレッド単位での最短路計算



スーパーコンピュータ
上の大規模グラフ処理
基盤での実行

大規模グラフデータ
数理計画問題
並列化支援
高精度量子化学計算

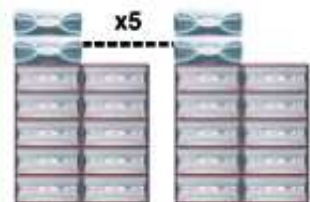
前処理による変数の削除と並列計算の適用
(大規模スレッド並列) 整数変数10万以上

IP に対する分枝操作

TSUBAME 2.0 System Configuration

Petascale storage: Total 7.13PB (Lustre + home)

Parallel file system area : 5.93PB



MDS,OSS
HP DL360 G6 30nodes
Storage
DDN SFA 10000x5
(10 enclosures x5)
Lustre (5 Filesystems)
OSS:20 OST:5.9PB
MDS:10 MDT:30TB

OSS MDS

Users' home space : 1.2PB



Storage Server
HP DL380 G6 4nodes
BlueArc Mercury 100 x2
Storage
DDN SFA10000 x1
(10 enclosures x1)

NFS,CIFS x4

NFS,CIFS,iSCSI
Acceleration x2

Existing system

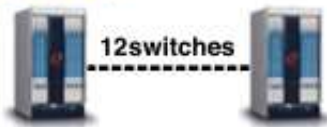
**Sun SL8500
Tape system**

Super TITANET

SINET 3

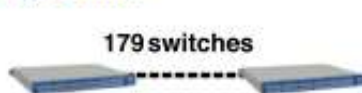
Inter-node connection network: full bisection/non-blocking

Core Switch



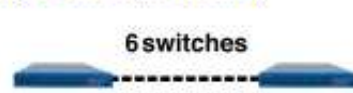
Voltaire Grid Director 4700 12switches
IB QDR: 324ports

Edge Switch



Voltaire Grid Director 4036 179switches
IB QDR: 36ports

Edge Switch(w/10GbE)



Voltaire Grid Director 4036E 6switches
IB QDR: 34ports
10GbE: 2ports

Administrative servers

**1360 nodes, 2720 CPUs,
4080 GPUs**

Compute nodes : 2.4PF (CPU+GPU) / 224.69TF(CPU)

Thin compute node



1408nodes (32node x44 racks)

HP servers featuring GPUs 1408nodes
CPU Intel Westmere-EP 2.93GHz
(Turbo boost 3.196GHz) 12Cores/node
Mem:55.8GB (=52GiB) or 103GB (=96GiB)
GPU NVIDIA M2050 515GFlops.3GPUs/node
SSD 60GB x2 120GB *55.8GB memory
120GB x2 240GB *103GB memory
OS: SUSE Linux Enterprise + Windows HPC

CPU(Total):215.99TFLOPS (Turbo boost 3.196GHz)
CPU+GPU:2391.35TFlops
Memory (Total):80.55TB
SSD (Total):173.88TB

Medium compute node



HP 4Socket Server 24nodes
CPU Intel Nehalem-EX 2.0GHz
32Cores/node
Mem:137GB (=128GiB)
SSD 120GB x4 480GB
OS:SUSE Linux Enterprise Server

CPU(Total): 6.14TFLOPS

Fat compute node



HP 4Socket Server 10nodes
CPU Intel Nehalem-EX 2.0GHz
32Core/ node
Mem:274GB (=256GiB) 8nodes
549GB (=512GiB) 2nodes
SSD 120GB x4 480GB
OS: SUSE Linux Enterprise Server

CPU(Total): 2.56TFLOPS

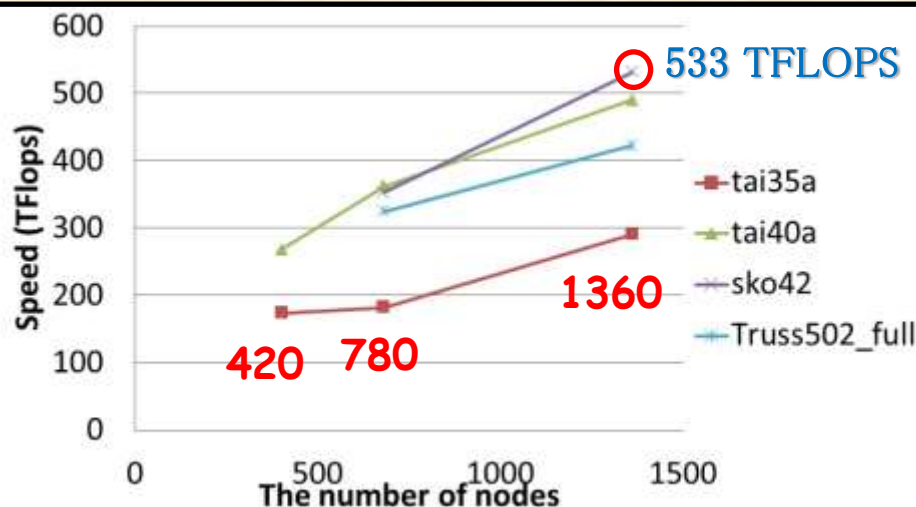
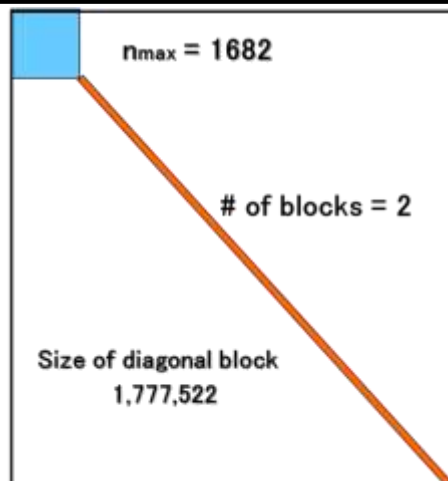
PCI-E gen2 x16 x2slot/node

GSIC:NVIDIA Tesla S1070GPU (34 units)

SDPARA can solve the largest SDP problem which has over **1.48 million constraints** (DNN relaxation problem : sko42a QAPLIB, and create **a new world record in 2012**.

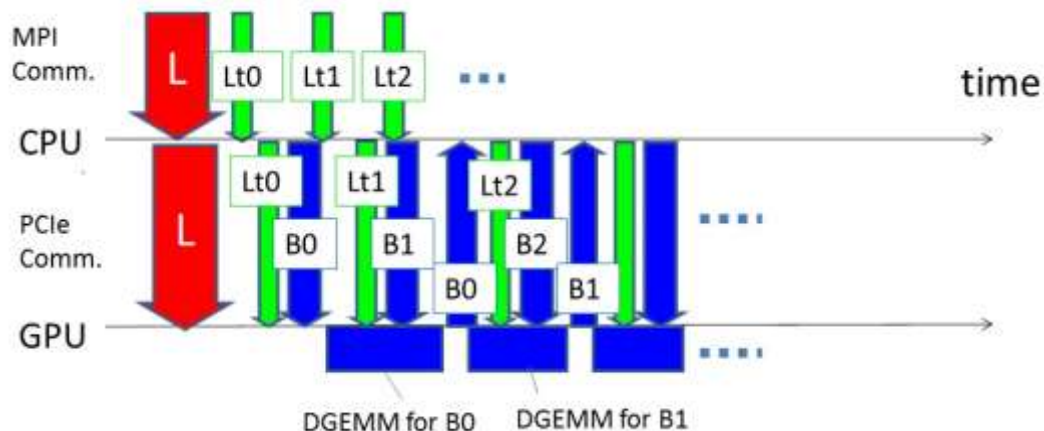
- Using **1360 nodes, 2720 CPUs, 4080 GPUs**
- Computation Time : **2700 sec./iteration** => About 30 iterations are needed
- Performance of **CHOLESKY (= 1 Exa Flops : 1.48m x 1.48m)** :
=> **2045 sec./iteration --> 533 TFLOPS**
- One of the **fastest and largest result** as mathematical optimization problems

$n = 1,779,204$
(matrix size)
 $m = 1,484,406$
(# of constraints)
 $nnz = 23,476,376$
(# of nonzero elements)



Accelerate **CHOLESKY** by using massively parallel GPUs

- In order to achieve scalable performance with thousands of GPUs
- Utilize high performance BLAS kernel coupled with optimization techniques to **overlap computation, PCI-Express communication and MPI communication.**



1. 解析対象

twitter	Wikipedia	facebook
2億ユーザ 2億呟き/日	4.7億更新 19更新/頁	7.5億ユーザ 130 friends 300億記事

2. 高次元グラフィック

大規模グラフデータ	10億頂点, 200辺/頂点
クラスタリング > 次元圧縮	1階層, 5億次元 > 3階層, 40,000次元
高次元レイアウトデータ	1.2PB

3. 対話的グラフフィルタリング

検索キーワード

原発 推進 反対

ノード

- ☒ 記事
- ☐ 議論
- ☐ 紹介

クラスタ

関係性

- ☒ 同期的編集関係
- ☒ 参照関係
- ☐ 共著関係

検索対象期間

2011.3.8 ~

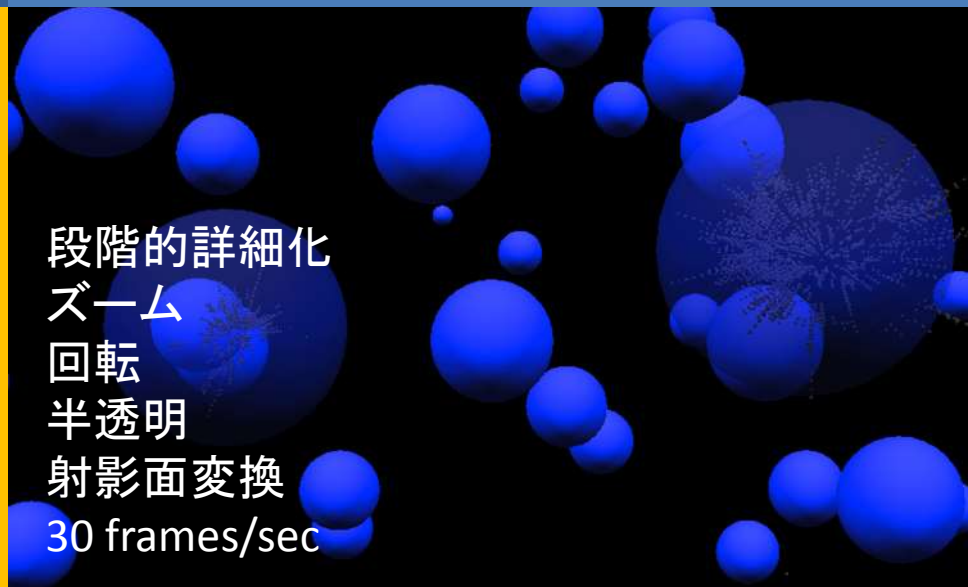
2011.3.25

10億頂点
200辺/頂点

数100万頂点
数10辺/頂点

4. 対話処理: 射影演算

段階的詳細化
ズーム
回転
半透明
射影面変換
30 frames/sec



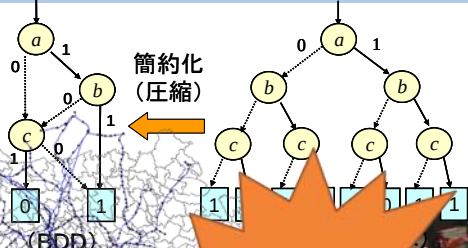
まとめ

- ポストペタスケールスパコン上での高性能な**超大規模グラフ処理技術**の達成
 - リアルタイム大規模グラフストリーム処理基盤
 - 高速グラフ最適化ライブラリ・ソルバー
 - 大規模グラフストア
 - 大規模グラフストリームデータの対話型閲覧システム
- **最重要カーネル**のひとつである超大規模グラフ処理の研究開発を通じて**ポストペタ・エクサスケールスーパーコンピューティング**に向けた要素技術の実現に大きく貢献
 - 数千万規模の並列性・不均質性の克服
 - 計算量・データ移動の最適バランスの実現



研究開発を通じて得られた知見は
他のカーネルへ適用・応用

最先端理論(Algorithm Theory)



有機的な融合



超大規模実データ(Practice)

最新計算技術(Computation)

HPCによる安心安全な社会基盤の実現

- 応用例
 - ✓ 防災計画の策定
 - ✓ 災害時避難と誘導及び情報収集と解析
 - ✓ スマートグリッドによる高度かつ安定な電力供給
 - ✓ ソーシャルネットワーク解析
- 新しいスパコン利用法
 - ✓ 災害時に稼働可能(免震、独立電源)な状態で整備
 - ✓ 緊急時には災害対策プログラムが稼働
→ 国民の生命財産の保護