グリッド･クラウドコンピューティング

佐藤　仁　　東京工業大学　学術国際情報センター

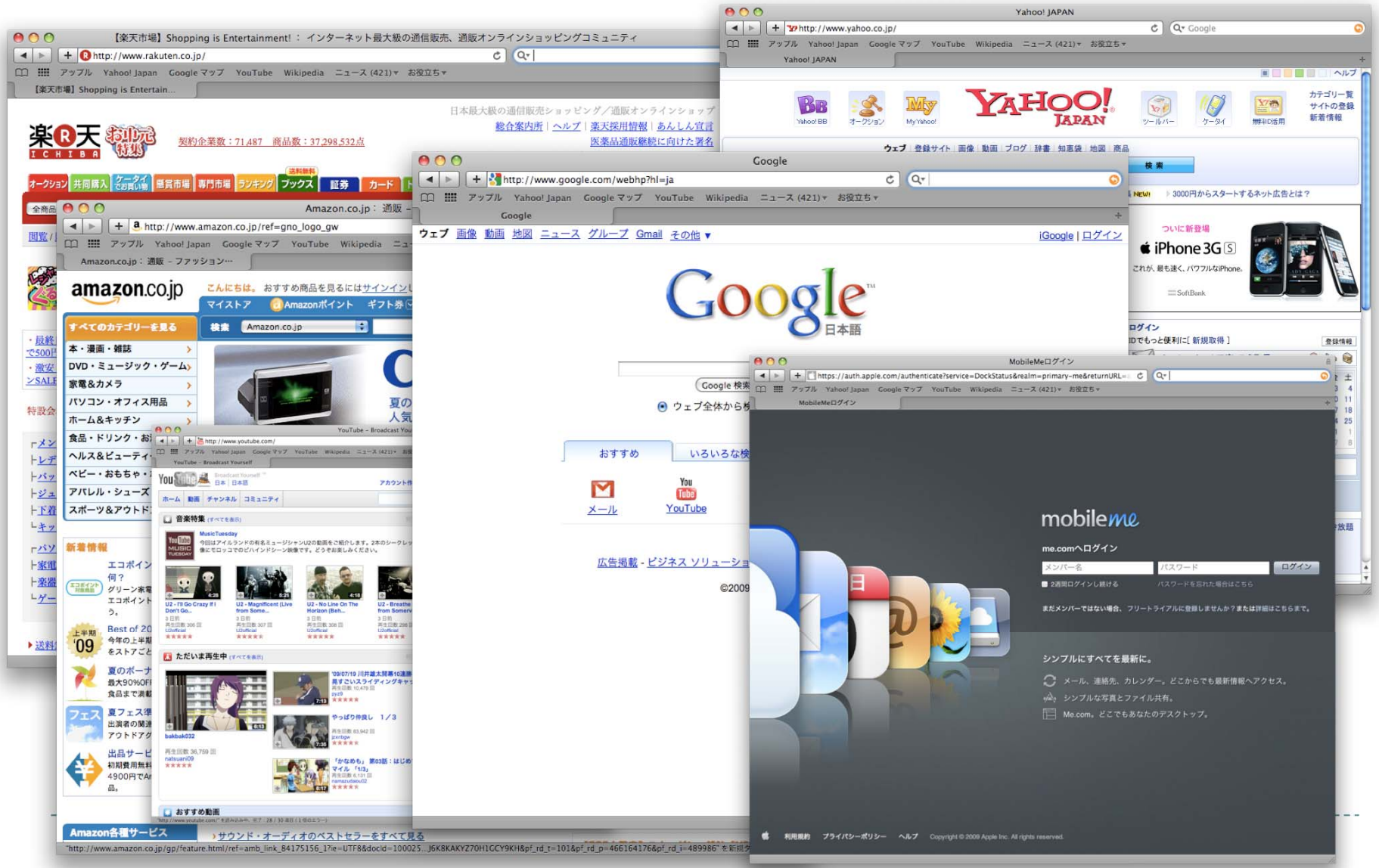
簡単な自己紹介

- 佐藤　仁
  - 学術情報国際センター　産学官連携研究員
  - スーパーコンピュータの運用･管理
  - 大規模データ処理に関する研究，クラウド，グリッド

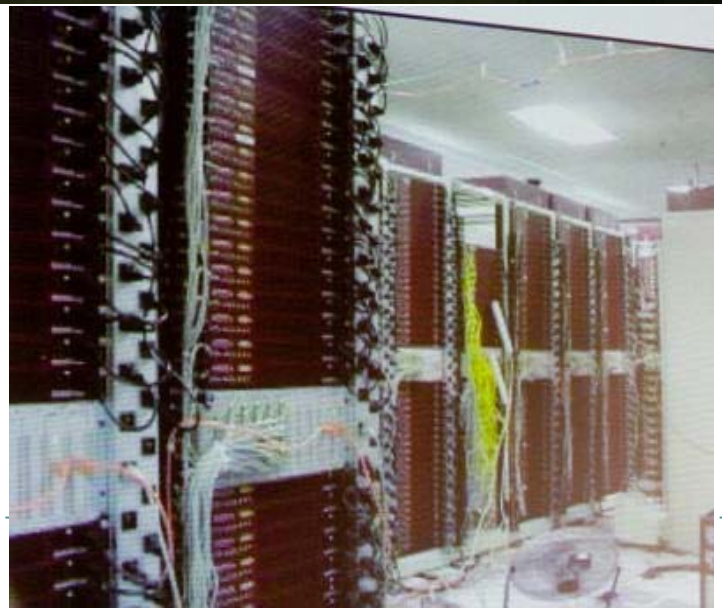
アウトライン

- 大規模な計算環境に関する一般的な話
- グリッドコンピューティング
- クラウドコンピューティング
- 将来の課題

いろいろなWeb Site



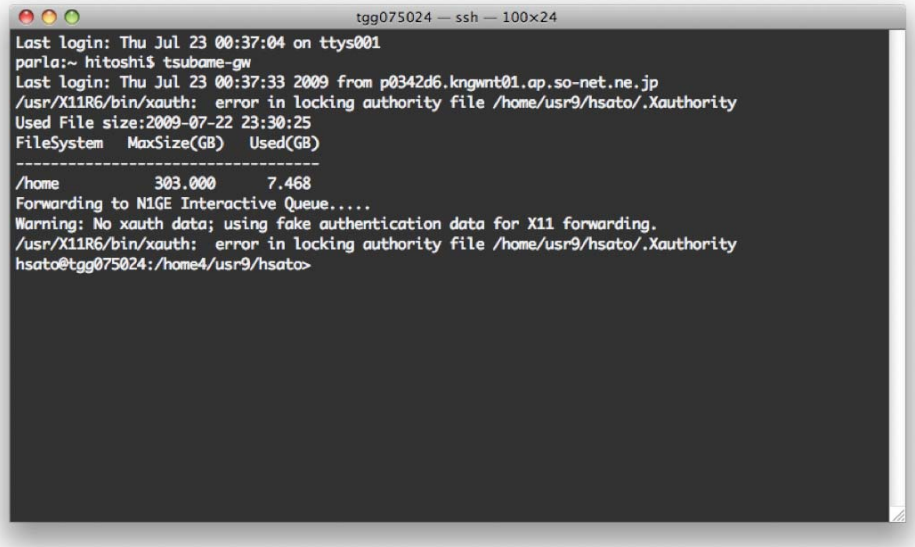
裏側の世界



スーパーコンピューター

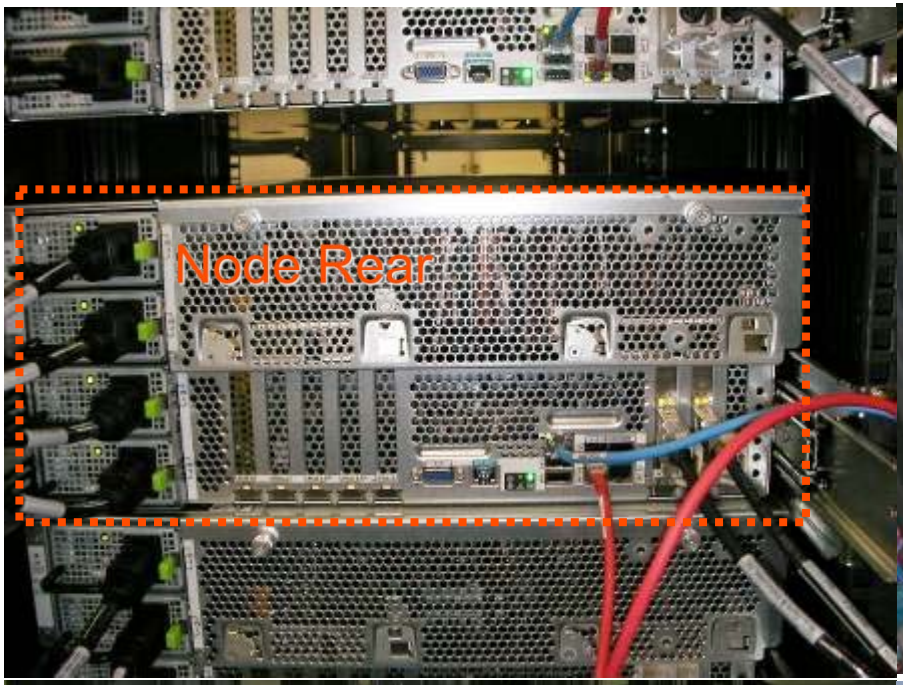
- 内部の演算処理速度がその時代の一般的なコンピュータより極めて高速な計算機
- 例: 東工大TSUBAME

大岡山キャンパス　図書館の隣



使うときの見た目は普通の端末とあまり変わらず．．

裏側の世界


Node Rear


Local Infiniband Switch
(288 ports)
Currently
2GB/s / node
Easily scalable to
8GB/s / node

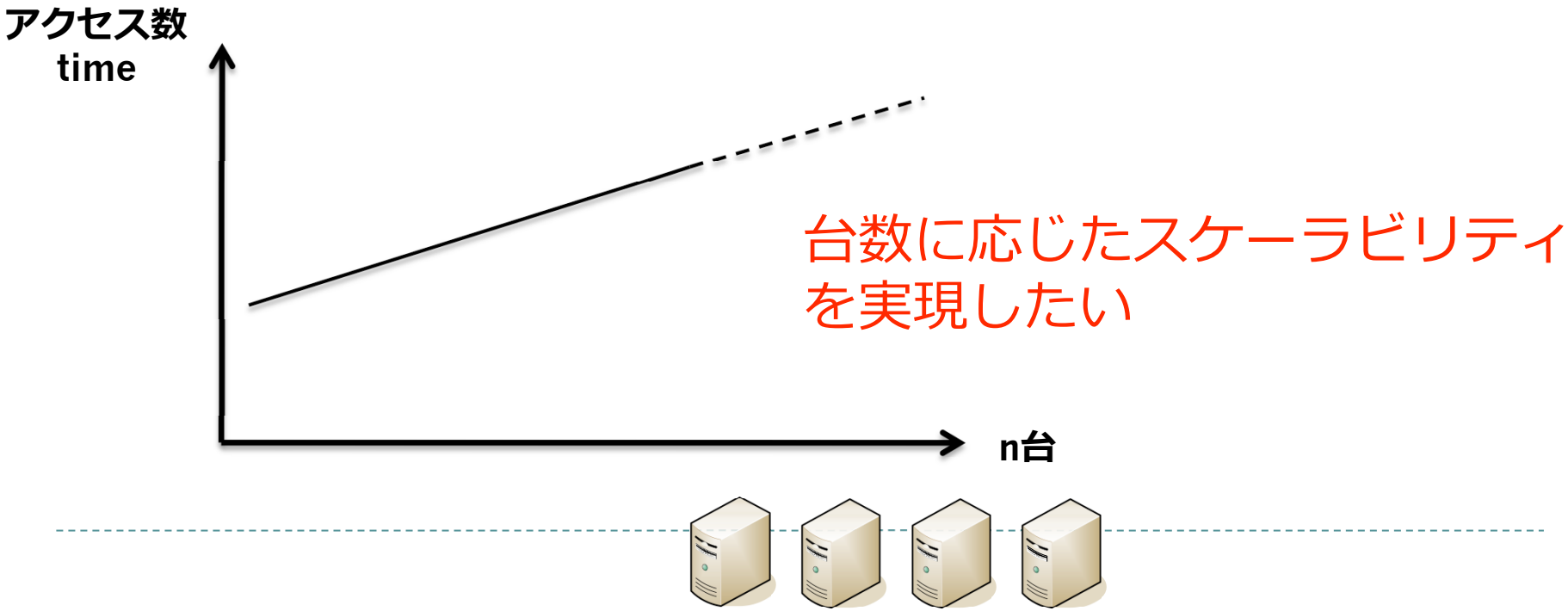

~500 TB out of 1.1PB


Cooling Towers (~32 units)

なぜ，こんなにマシンを並べているのか？

スケーラビリティ

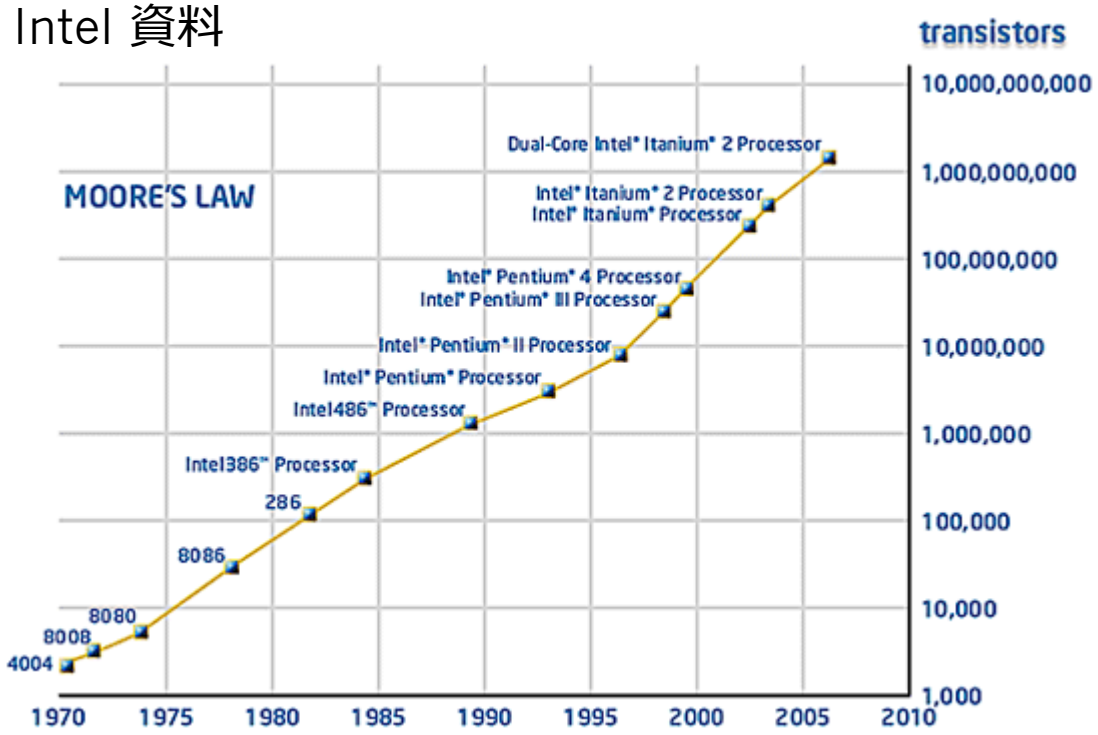
- 利用者，仕事の増大に適応する能力･度合い
- 例
  - マシンの台数を増やしたら，，，
    - 多くのウェブアクセス要求を裁けるようになった
    - アプリケーションの実行時間が短くなった



Mooreの法則 (1965年)

- 「ICの集積度(+性能)は3年で4倍になる」

Intel 資料



限界はあるか？
→ 物理的限界

マイクロプロセッサの性能向上の凄さ
→ これらをネットワークで接続し，並列処理させれば
さらなる性能向上が見込める(はず)！！

計算機の性能を表す指標

- FLOPS
  (Floating point number operations per second)
  - 一秒間に浮動小数点数(実数の近似値)演算が何回できるかという能力を表したもの

PSP
9.6GFlops

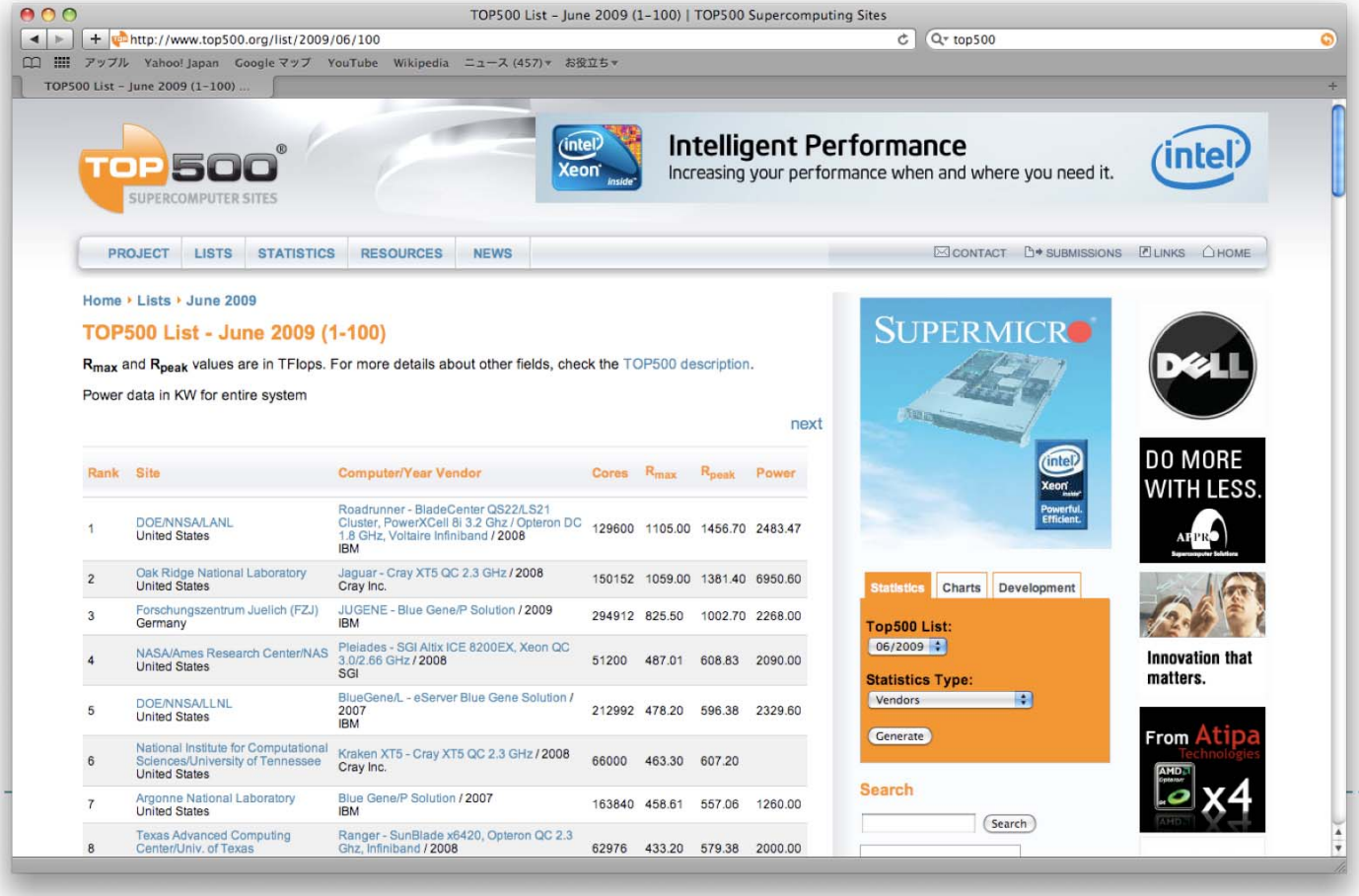
Xbox 360
1TFlops

PS3
2TFlops

PC (Core 2 Quad)
102.4GFlops

Top 500

- スーパーコンピュータのランキング
  - LINPACKベンチマークにより線形方程式系をガウス消去法により解く速度(FLOPS)を測定



世界最高峰のスーパーコンピューター

2008 Q2 LANL/IBM "Roadrunner"
12,240 IBM PowerXCell
6120 AMD Dual Core Opterons
> 100,000 Cell SPE Cores
> 1.3 Petaflops Peak, ~1 Petaflop Linpack
98TB Memory, 2.4MB Power
278 racks, ~500m2 floorspace, 250 tons
~100KM Infiniband

2008Q4 ORNL/Cray XT5 "Jaguar"
~180,000 AMD"Barcelona" Opteron
CPU Cores, 1.64 Petaflops Peak,
1.06 Petaflops Linpack
~200 racks, ~580m2 floorspace
362TB Memory, ~7MW Power,
10 Petabytes HDD

2008年にはじめて1ペタフロップスを超えるマシンが登場

TSUBAME

159 TFlops (Peak)
87.01 TFlops (Linpack)
Top 500 41st machine

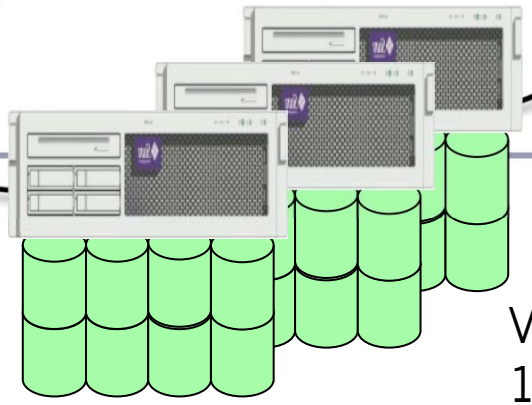
Sun Fire X4600
Opteron 16 cores/node
× 655 nodes

ClearSpeed CSX600
SIMD accelerator × 360

TSUBAME

TSUBAME 1.1

Tesla S1070
× 170



0.1PB
Storage

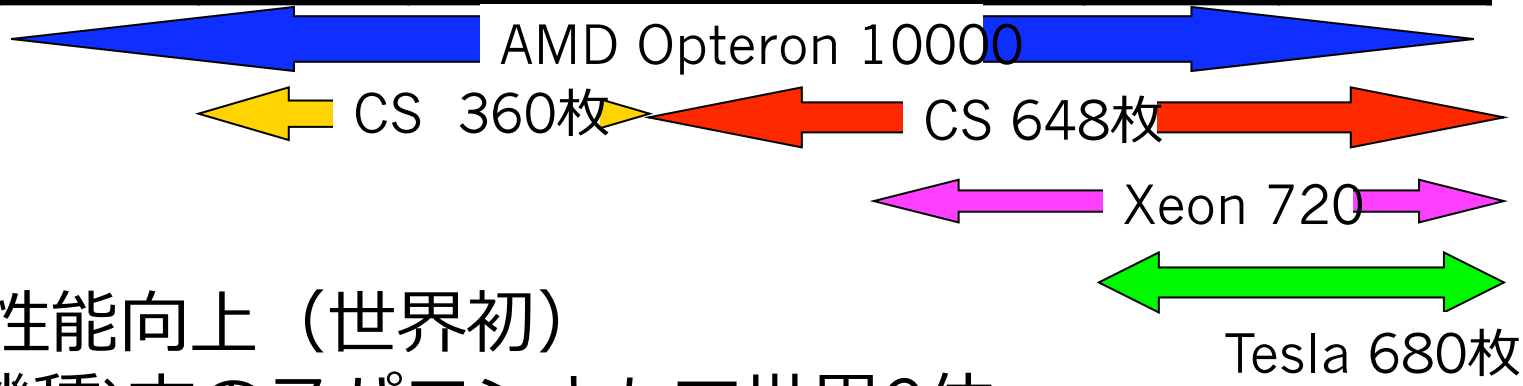
1.6PB
Lustre Storage

Voltaire ISR9288 Infiniband
10Gbps x～1400 ports

TSUBAME 1.2

TSUBAME とTop500

	Jun06	Nov06	Jun07	Nov07	Jun08	Nov08	Jun 09
性能（テラフロップス）	38.18	47.38	48.88	56.43	67.70	77.48	87.01
世界ランク	7	9	14	16	24	29	41
国内ランク	1	1	1	1	3	2	4
(参考：地球シミュレータ世界ランク)	10	14	20	30	49	73	-

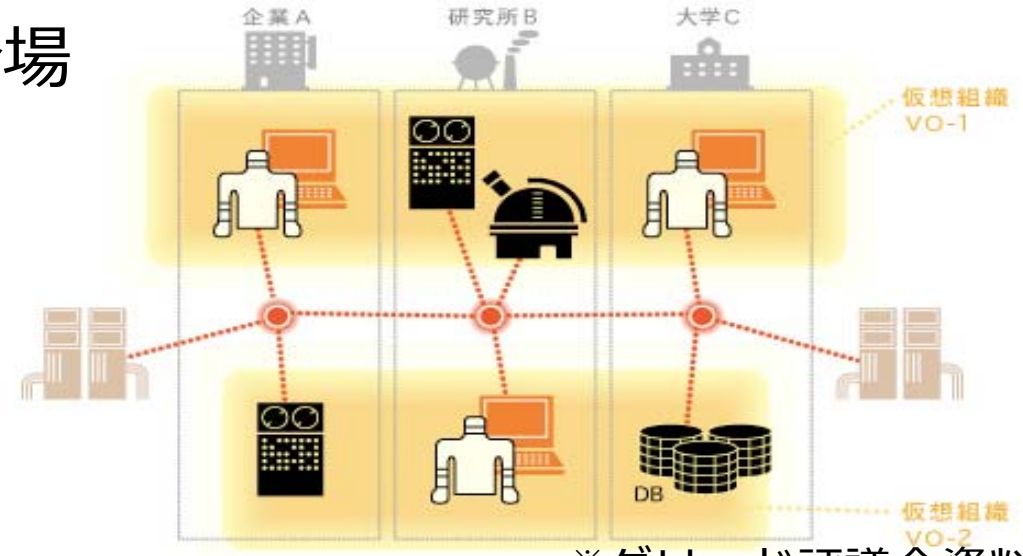


- 6回連続の性能向上（世界初）
- ヘテロ(異機種)方のスパコンとして世界2位
- GPGPUスパコンとして世界初のTop500ランクイン

もっとマシンを集めて使いたい!?

グリッドコンピューティング

- (アカデミカルな)定義
  - 地理的に分散した計算，ストレージ，ソフトウェア，実験装置，観測器などの資源をネットワークで接続して仮想化･統合し，必要に応じて仮想的な計算機や仮想組織を動的に形成するための技術 – **I. Foster, The Anatomy of the Grid**
- 単に広域分散環境を指す人も....
- 1990年代後半から登場

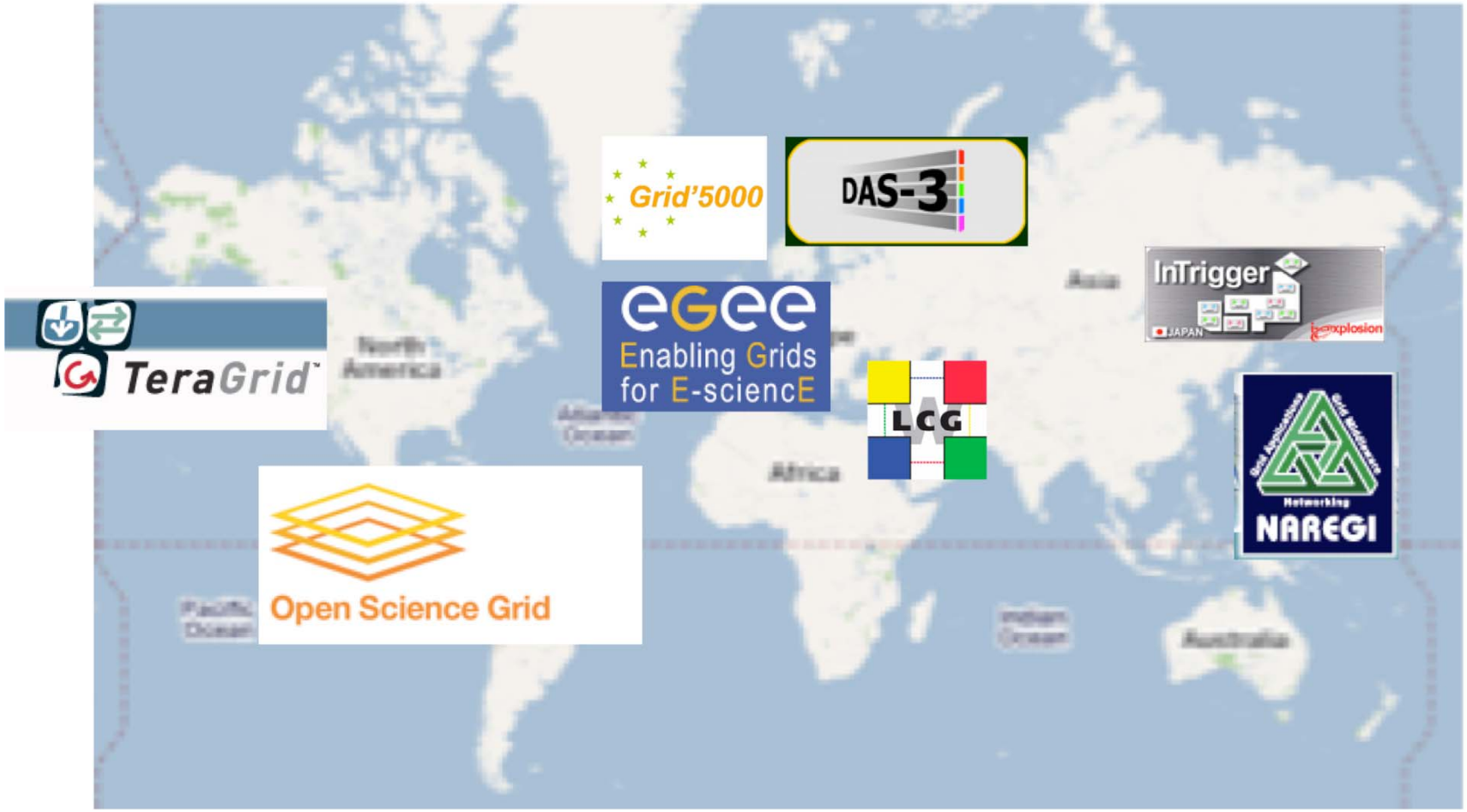


※グリッド評議会資料

利用シナリオ

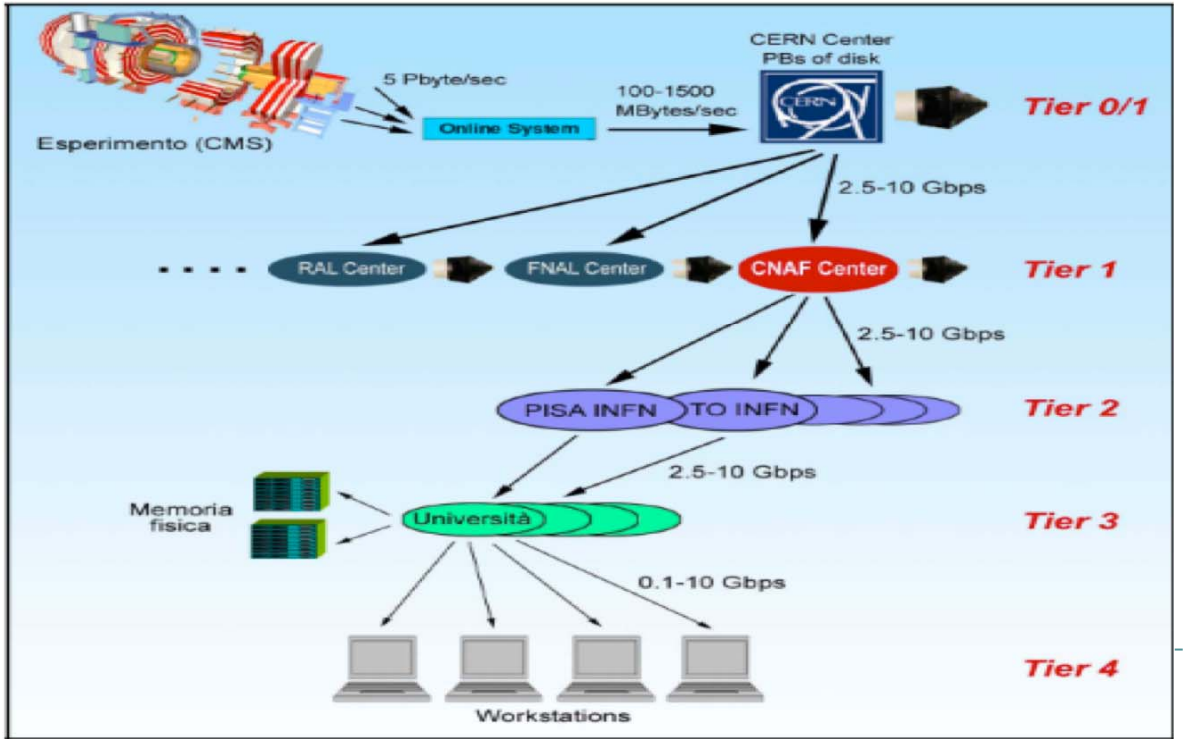
- 自動車製造会社が新しい工場計画のシナリオを評価するためにコンサルタント，アプリケーションサービスプロバイダ，ストレージサービスプロバイダ，計算資源プロバイダが一時的な仮想組織を構成する
- 危機管理チームが非常事態が発生したときの計画するためにデータベースとシミュレーションシステムからなる仮想組織を構成する
- 高エネルギー物理学研究のコラボレーションのために国際的な研究機関が複数年にわたり仮想組織を構成する

いろいろなグリッド


Grid'5000
DAS-3
InTrigger
JAPAN
explosion
TeraGrid
eGee
Enabling Grids
for E-sciencE
LCG
NAREGI
Networking
Open Science Grid
North America
Africa
Asia
Australia

事例　LCG Grid

- 欧州CERNによるLHC加速器実験
  - 20カ国3000人規模の研究者と階層的な計算センターからなる仮想組織
  - 年間数ペタバイトの実験データ解析およびシミュレーションによる検証



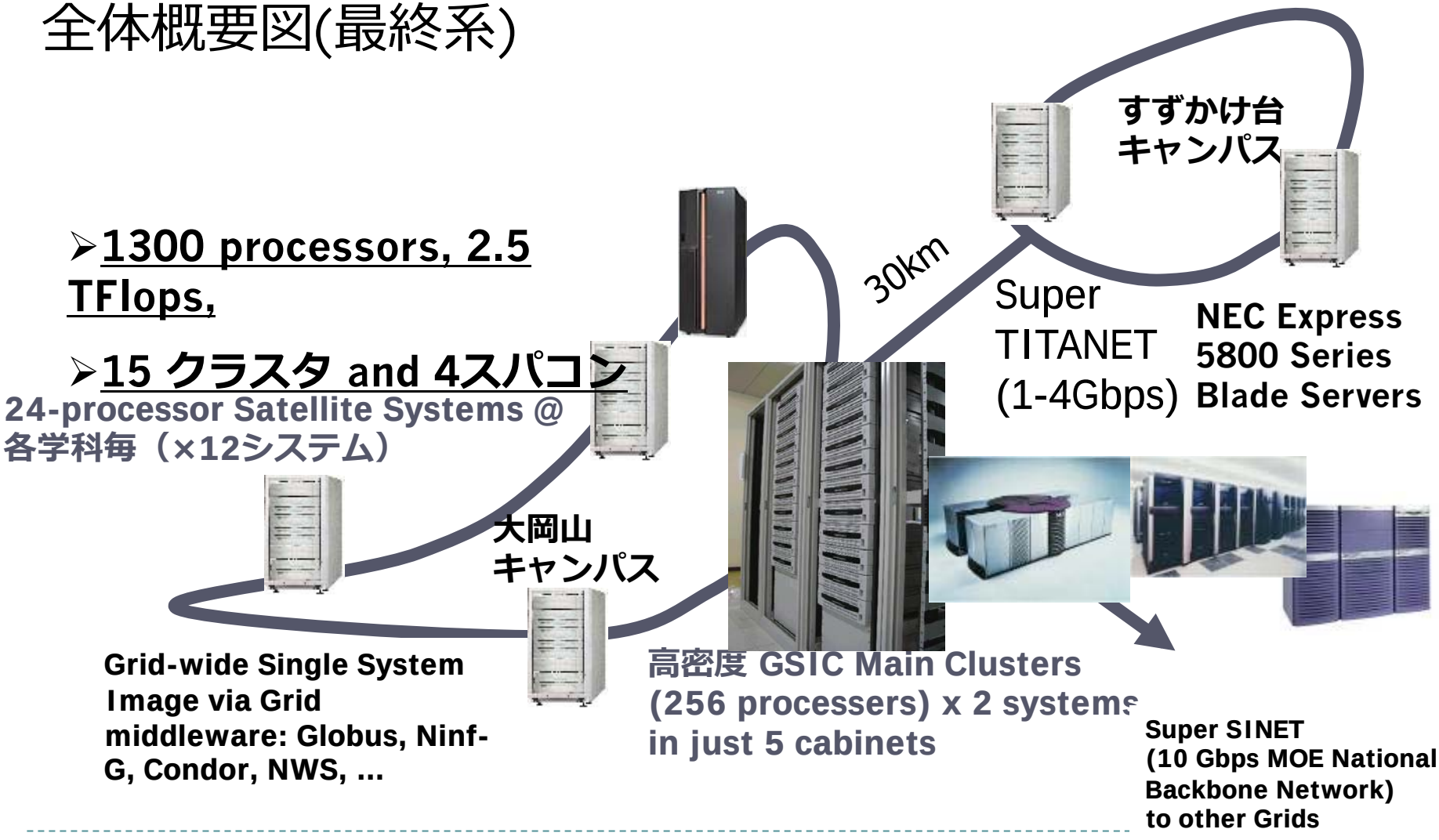
事例　東工大キャンパス・グリッド
Titech Grid(2002/3-2006/3)

- キャンパス内に次世代E-Scienceアプリのためのクラスタ・グリッドを構築する大規模実証実験(平成13年度末補正予算で開始)
  - センターは管理できるのか
  - ベンダーはサポートできるのか
  - ユーザは使えるのか
- 最初は800プロセッサ以上のPC群をキャンパス内で分散配置、ギガビットネットワーク(Super TITANET)で接続
- キャンパス内でグリッドのテストベッド
- グリッドミドルウェアによる実装・運用
- 様々な新世代E-Scienceアプリ
- 設置場所は2キャンパス・十数個所

TITECH GRID

東工大キャンパス・グリッド (cont'd)
Titech Grid(2002/3-2006/3)

全体概要図(最終系)



- **1300 processors, 2.5 TFlops,**
- **15 クラスタ and 4スパコン**

24-processor Satellite Systems @ 各学科毎（×12システム）

グリッドの難しさ

- セキュリティ
  - 異なる管理組織に属する資源を利用するのでセキュリティを担保するのが難しい
  - ポリシーの違い，OSの違い，ミドルウェアの違い

→　煩雑なソフトウェア設定

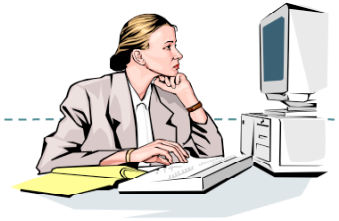
グリッドの難しさ (cont'd)

- 性能
  - 広域のネットワークを介して資源を接続
    - 遅延 (10 ～ 300 msec over)
    - バンド幅 (1 ～ 100 MB/sec over)
  - 技術トレンドの変化
    - 仮想マシンの成熟(2000年代～)
      - VMWare, Xen, KVM
    - x86ベースの64bit CPUの登場(2003年～)
      - たくさんのメモリを使える
    - CPUのマルチコア化(2005年～)
      - 1台のマシン上で複数のVMを動作させるのが合理的

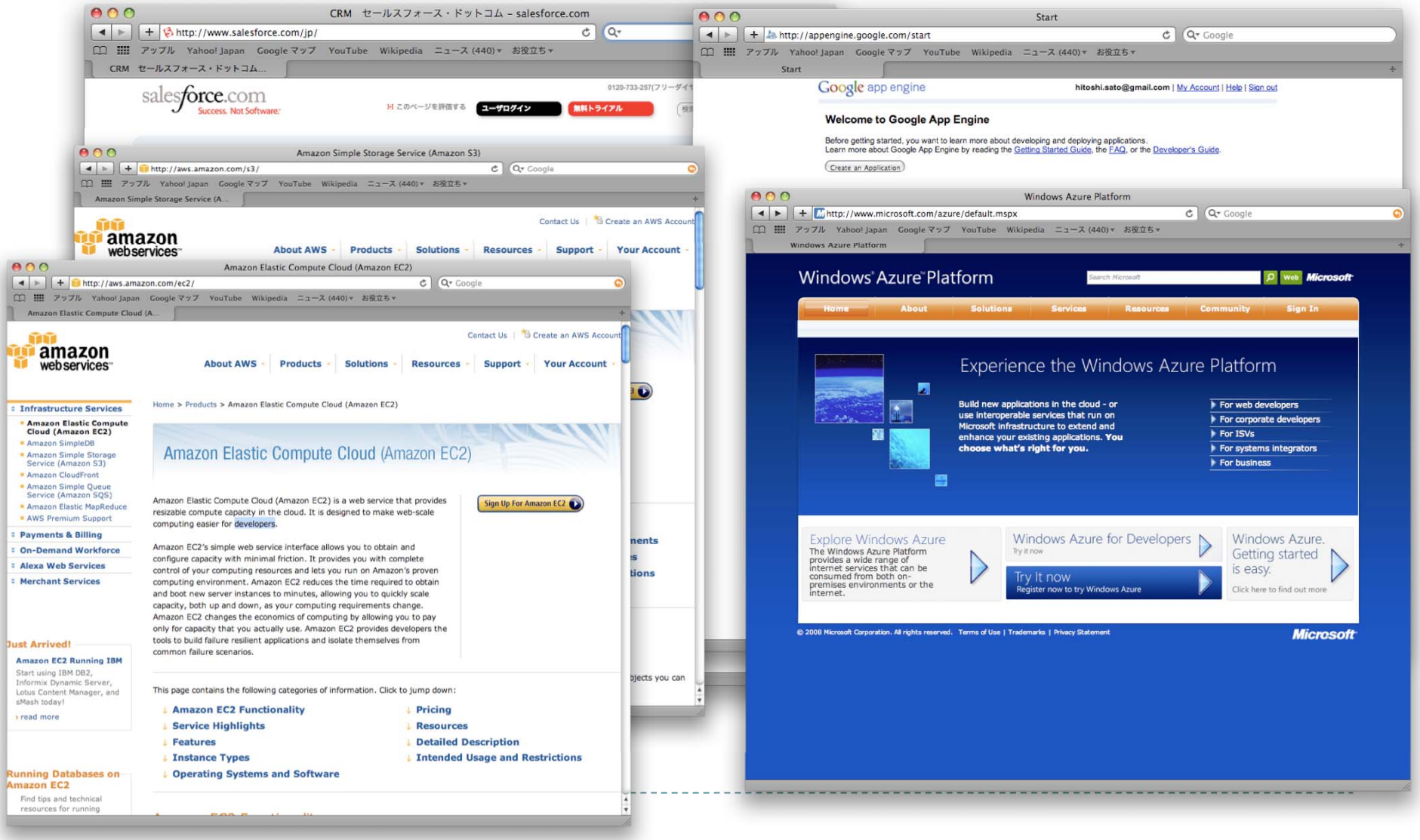
マシンを集めて使いたい!!
できれば，面倒なく，ナイスに使いたい!!

クラウドコンピューティング

- インターネットを介して，資源を利用した何かしらのサービスを提供
  - アプリケーション，計算資源，ストレージ，仮想化環境など
  - 2006年頃から徐々に盛り上がり始める



有名なクラウドベンダー(の一部)



事例：Amazon Elastic Compute Cloud (EC2)

- 拡張性の高い(resizebleな)計算環境を仮想マシンにより提供
- 従量課金
  - CPU, RAMの性能，容量と利用時間に応じて課金
- 構築環境をインターフェースにより操作可
  - Web Service (SORP, REST)
  - AWS Management Console
  - EC2 Command Line Tool
- 数百もの仮想マシンイメージをあらかじめ提供
  - Linux or Windows

定義

- 定義は明確ではない (いろいろある)
  - “.. style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet” - **Wikipedia**
  - “Clouds are hardware-based services offering compute, network and storage capacity where: Hardware management is highly abstracted from the buyer, Buyers incur infrastructure costs as variable OPEX, and Infrastructure capacity is highly elastic” - **McKinsey & Co. Report: ‘‘Clearing the Air on Cloud Computing**

定義 (cont'd)

- Cloud computing has the following characteristics: (1) The illusion of infinite computing resources..., (2) The elimination of an up-front commitment by Cloud users ..., (3) The ability to pay for use ... as needed ..." - **UC Berkeley RAD Labs**

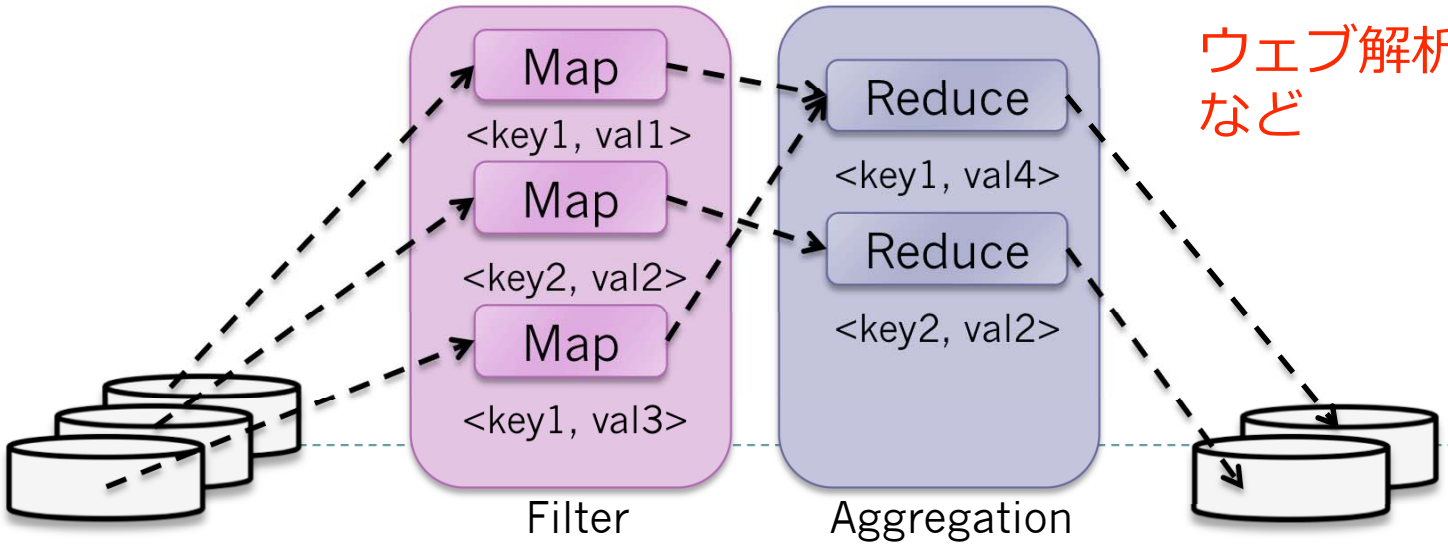
- "... a pay-per-use model for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." - **National Institute of Standards and Technology (NIST)**

利用形態に特徴がみられる

- 従量課金
- 需要に応じた資源利用
- セルフサービス
- 資源の抽象/仮想化

科学技術計算への適用

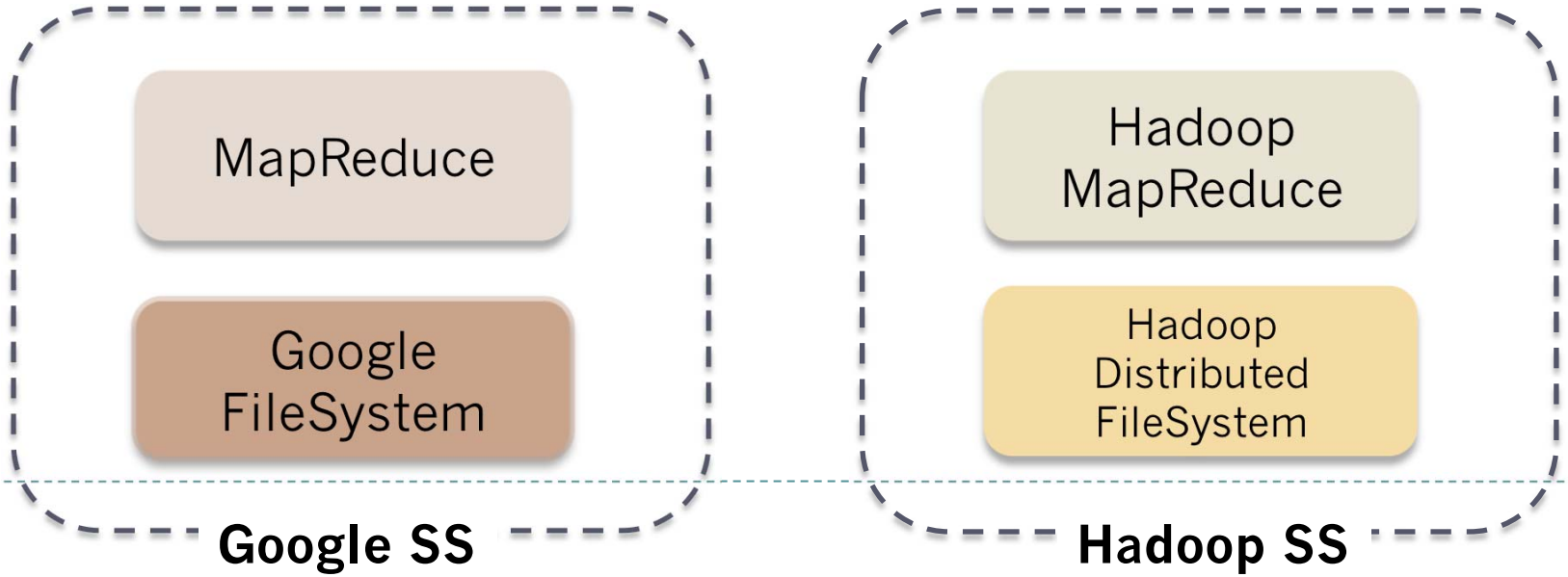
- MapReduce
  - Programming model for data-intensive computing
  - Map
    - processes a key/value pair to generate a set of intermediate key/value pairs
  - Reduce
    - merges all intermediate values associated with the same intermediate key



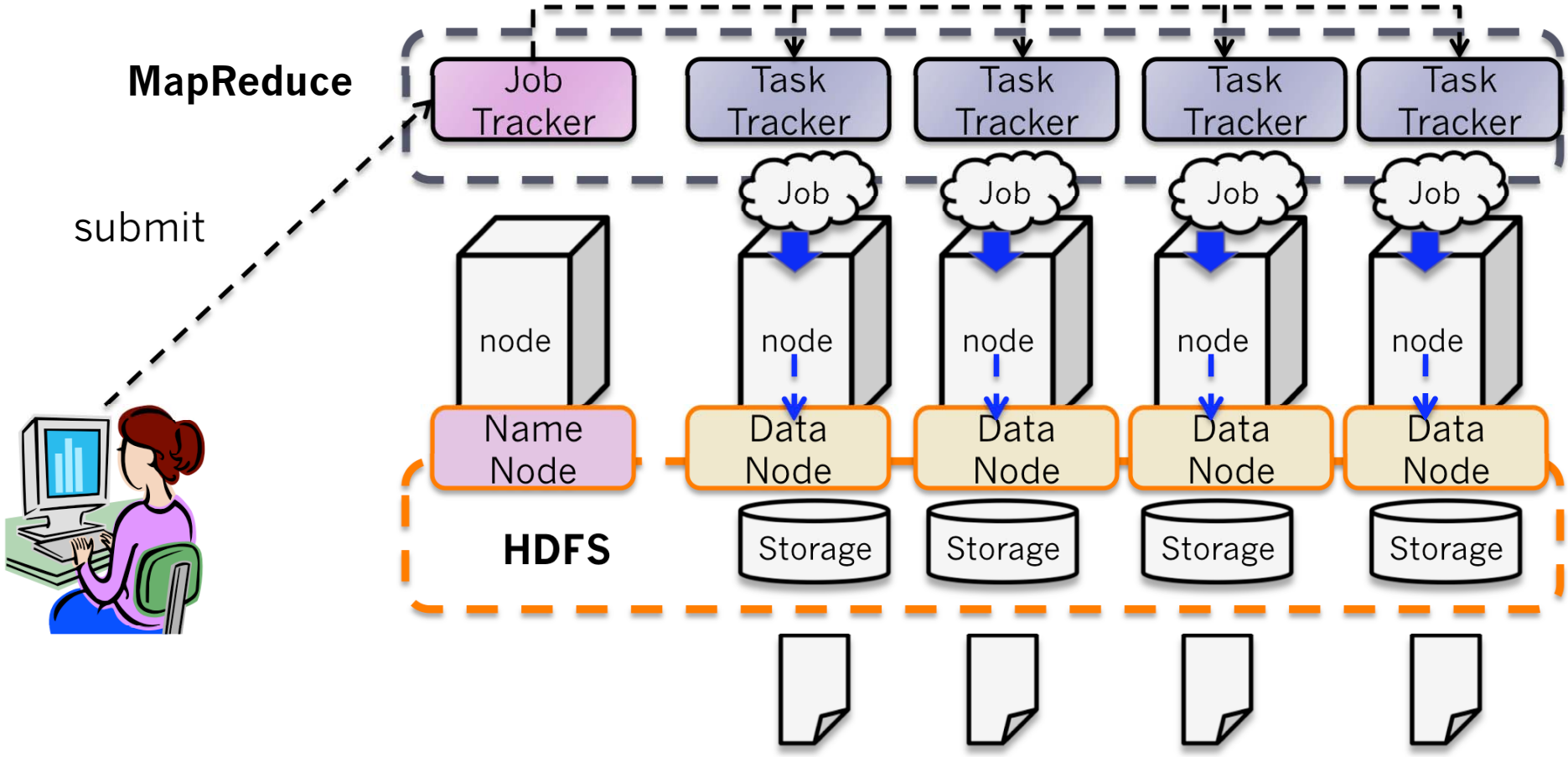
ウェブ解析，自然言語処理など

Hadoop

- OSS inspired by Google projects
  - GFS, MapReduce, etc.
- Software platform for MapReduce-based data-intensive computing
  - Hadoop MapReduce
  - Hadoop Distributed Filesystem (HDFS)



Hadoop Configuration



グリッドとクラウドの違い

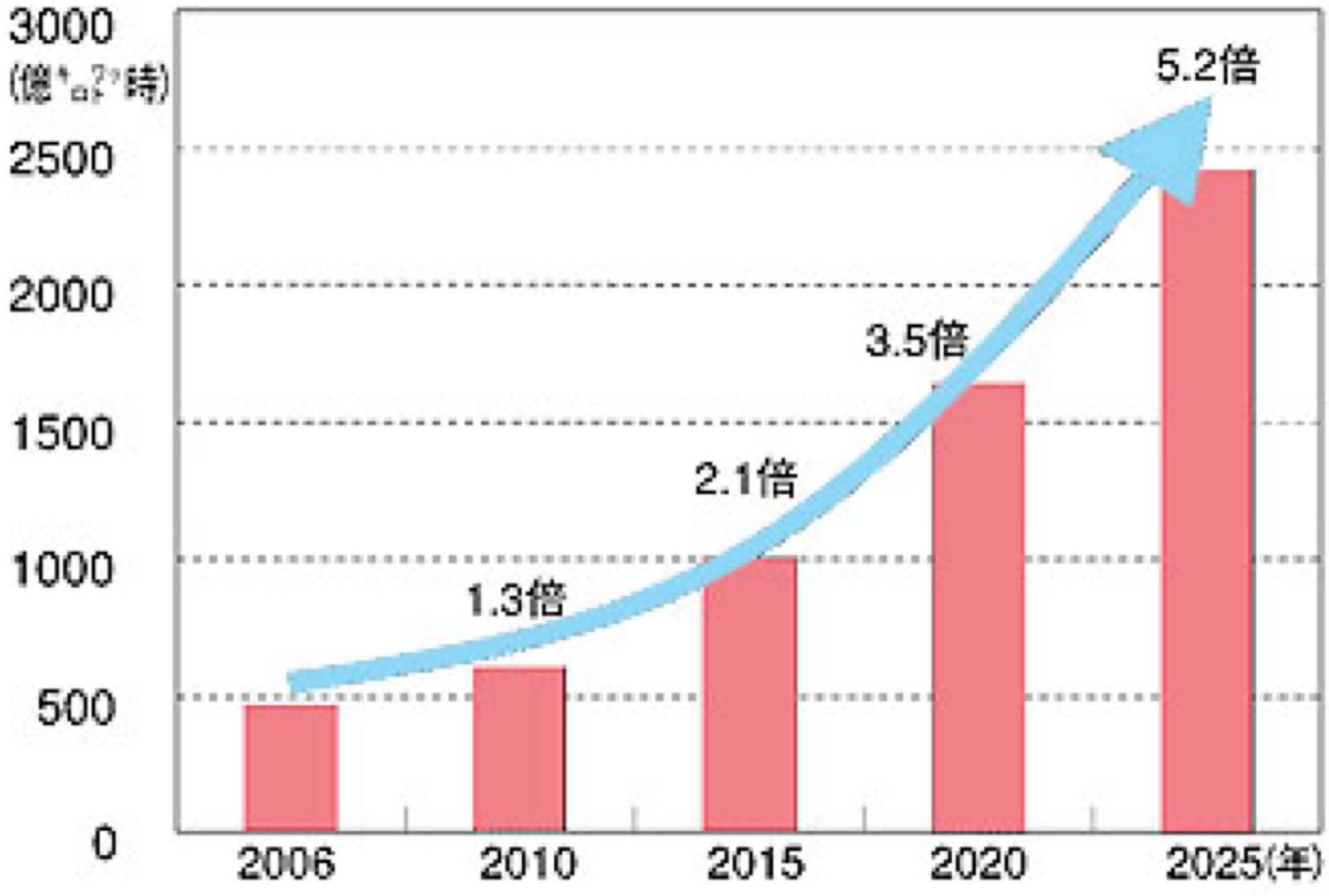
- グリッド
  - いろいろな人や組織がばらばらに所有している異なるコンピュータ資源を統一的に一つの仮想コンピュータとして見せる技術

- クラウド
  - 一つの組織(例えばGoogleとかMicrosoftとか)が所有するデータセンターの比較的均質でかつ莫大な数のコンピュータ資源を，他の人や組織に切り売りする技術

将来への課題

消費電力

- 日本におけるIT機器の消費電力予測



（出所）経済産業省／グリーンIT推進協議会試算(2008)

スーパーコンピュータの消費電力

Machine	CPU Cores	Watts	Peak GFLOPS	Peak MFLOPS /Watt	Watts/ CPU	
TSUBAME(Opteron)	10480	800,000	50,400	63	76.336	
TSUBAME(w/ClearSpee	11,200	810,000	85,000	104.94	72.321	
Earth Simulator	5120	6,000,000	40,000	6.7	1171.9	
ASCI Purple (LLNL)	12240	6,000,000	77,824	12.971	490.2	
AIST Supercluster	3188	522,240	14400	27.574	163.81	
LLNL BG/L (rack)	2048	25,000	5734.4	229.38	12.207	
Next Gen BG/P (rack)	4096	30,000	16384	546.13	7.3242	
TSUBAME Next Gen (2010)	40000	800,000	1000000	1250	20	

ペタフロップスには今後10倍以上の向上が必要

GPGPU
(General Purpose GPU)

- GPUの科学技術計算への応用
  - 例：Nvidia 8800GTX/8800GTS/280GTX, Tesla 10p
- 特徴
  - 高いピーク性能: (1TFlops)
    - 密結合問題に適している: 例: N体問題
  - 高いメモリバンド幅: (> 100GFlops)
    - 流体シミュレーションなど疎結合問題に適
    - 高い三次元FFTの性能もメモリバンド幅に起因

→ 高いエネルギー効率

TSUBAME 1.2への進化=>GPUの試験的追加

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

NEC SX-8i

500GB
48disks

Storage
1.5 Petabyte (Sun x4500 x 60)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
60GB/s aggregate I/O BW

10Gbps+External NW

Unified Infiniband network

10,000 CPU Cores
300,000 SIMD Cores
~900TFlops-SFP,
~170TFlops-DFP
80TB/s Mem BW (x2 ES)

Sun x4600 (16 Opteron Cores)
32~128 GBytes/Node
10480core/655Nodes
21.4TeraBytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

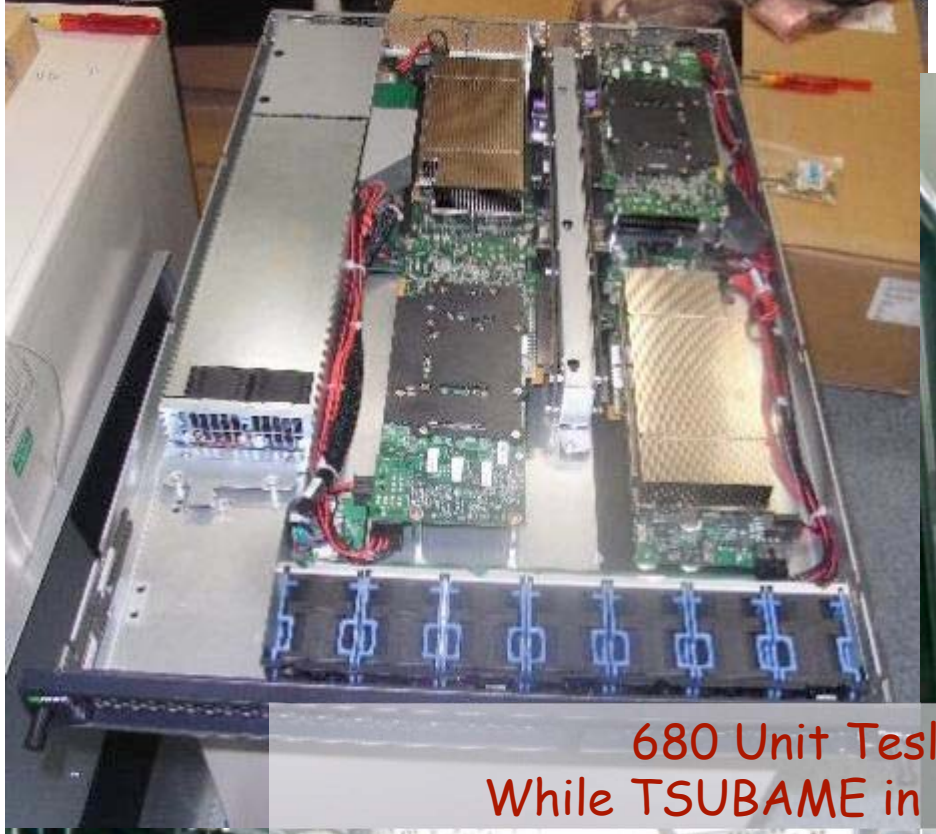
GCOE TSUBASA
Harpertown-Xeon
90Node 720CPU
8.2TeraFlops

NEW: co-TSUBAME
72Node 586CPU (Low Power)
~5TeraFlops

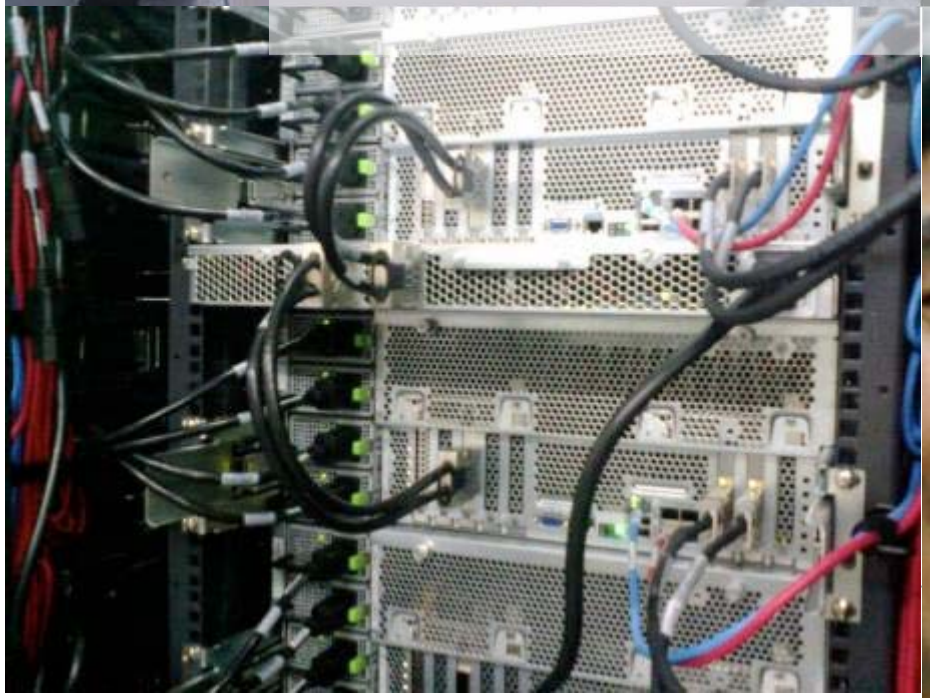
PCI-e

ClearSpeed CSX600
SIMD accelerator
360 648 boards,
35 52.2TeraFlops

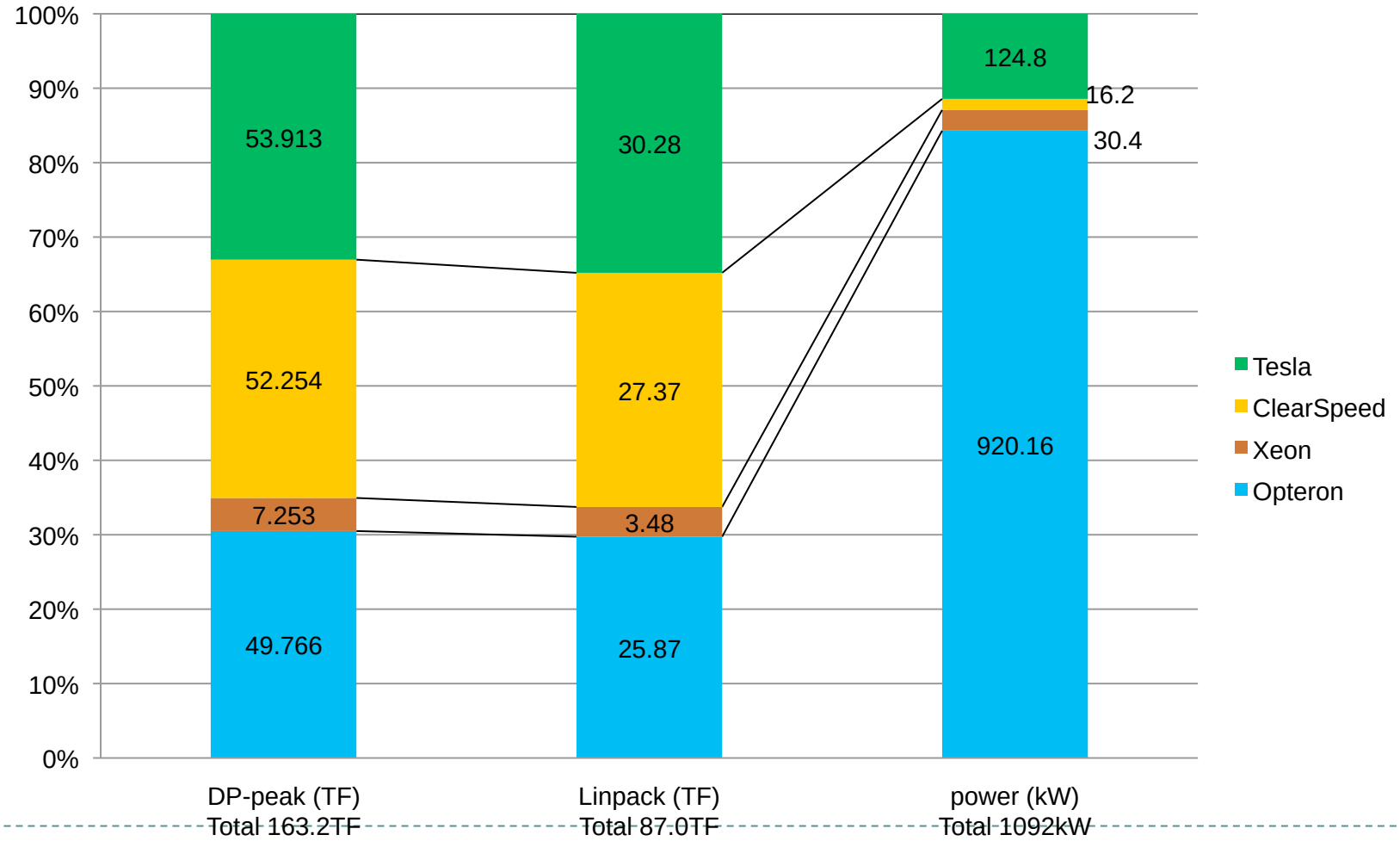
Nvidia Tesla S1070: 170台, 総計 680カード
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0 (130TF SFP / 80TF DFP)



680 Unit Tesla Installation…
While TSUBAME in Production Service (!)



TSUBAME1.2におけるLinpack性能と消費電力



東工大TSUBAME2.0・2010年6月
新時代のGPU-accelerated Super Computer

東京工業大学
Tokyo Institute of Technology

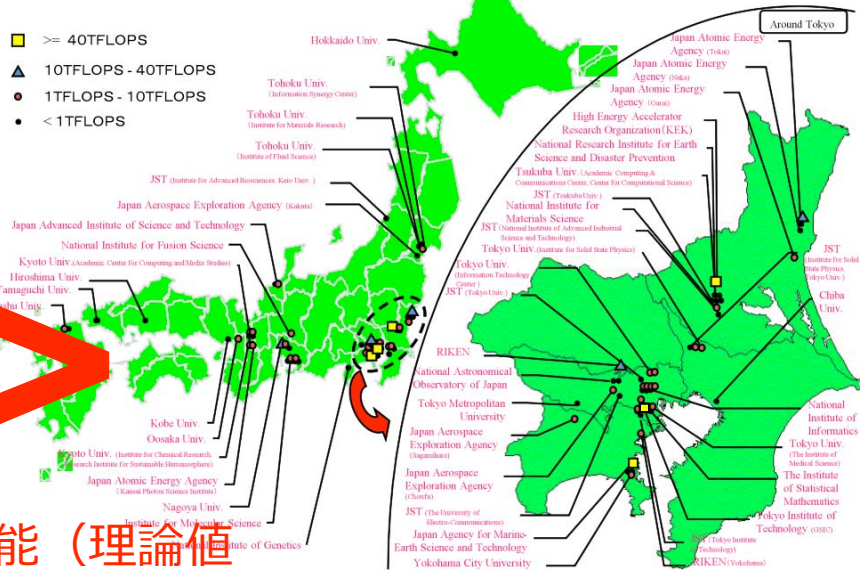
TSUBAME2.0 2010年6月稼働

1~数ペタフロップス
1ペタバイト/秒メモリバンド幅
超高速光結合網-数百テラビット/秒
数十ラック程度 (200m2)
通常の全国共同研究・共同利用情報基盤センターの経費で

1~数倍以上の性能（理論値も実アプリも)

日本全土のスパコン全て

我国のスパコン(2010年合算1ペタフロップス以下)



2010年合算 1ペタフロップス以下
全国60か所
年額300 億円以上

TSUBAME2.0へ向けた性能向上