

3. 情報基盤推進活動

3-1 Titech Grid Cluster





TSUBAMEのH19年度現状と将来計画

研究システム専門委員会報告資料 2008/03/24

東京工業大学 学術国際情報センター(GSIC) /
国立情報学研究所(客員)

教授 松岡 聡

matsu@is.titech.ac.jp

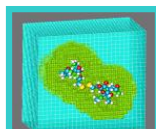




TSUBAME構築の目標---単なるスパコンではない

1. 東工大のシンボル:世界トップレベルの情報インフラ保持

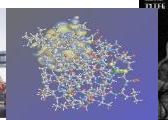
東京工業大学
Tokyo Institute of Technology



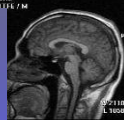
大規模ナノ物性シミュレーション



大規模災害・都市環境・社会・シミュレーション



蛋白質構造解析・ゲノム創薬



脳の血流シミュレーション(医療)

Rank	Country	Organization	Processor	Memory	Nodes	Performance
1	USA	Los Alamos National Laboratory	IBM eServer 8400	16GB	12800	12.5 TFLOPS
2	USA	IBM Thomas J. Watson Research Center	IBM eServer 8400	16GB	12800	11.4 TFLOPS
3	USA	Los Alamos National Laboratory	IBM eServer 8400	16GB	12800	10.5 TFLOPS
4	USA	Los Alamos National Laboratory	IBM eServer 8400	16GB	12800	10.4 TFLOPS
5	USA	Los Alamos National Laboratory	IBM eServer 8400	16GB	12800	10.3 TFLOPS

わが国・アジア一般研究機関で圧倒的1位: (地球シミュレータの2-3倍弱)

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

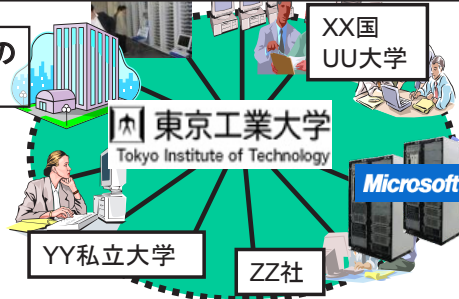
3. 産学連携等の推進、大型プロジェクトへの呼び水、アライアンスを組む他大学計算ニーズホスティング

Top500リスト:世界でトップ10へ

4. 学内の分散した情報基盤の集約化・ホスティング



シミュレーション
科学の飛躍的な
研究推進と教育



2006年4月東工大スパコン "TSUBAME" (スパコンとしての調達分)

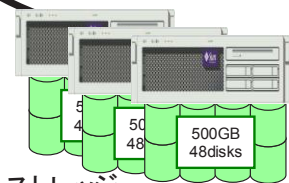
Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

10Gbps+外部ネットワーク

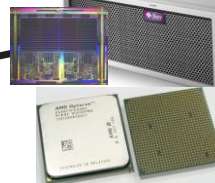


2006年6月現在
アジア No.1, 世界No.7
38.18TeraFlops
(Top500計測値)

Sun/AMD高性能計算クラスター
(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



ストレージ
1 Petabyte (Sun "Thumper")
0.1Petabyte (NEC iStore)
Lustre ファイルシステム
>400Gbps

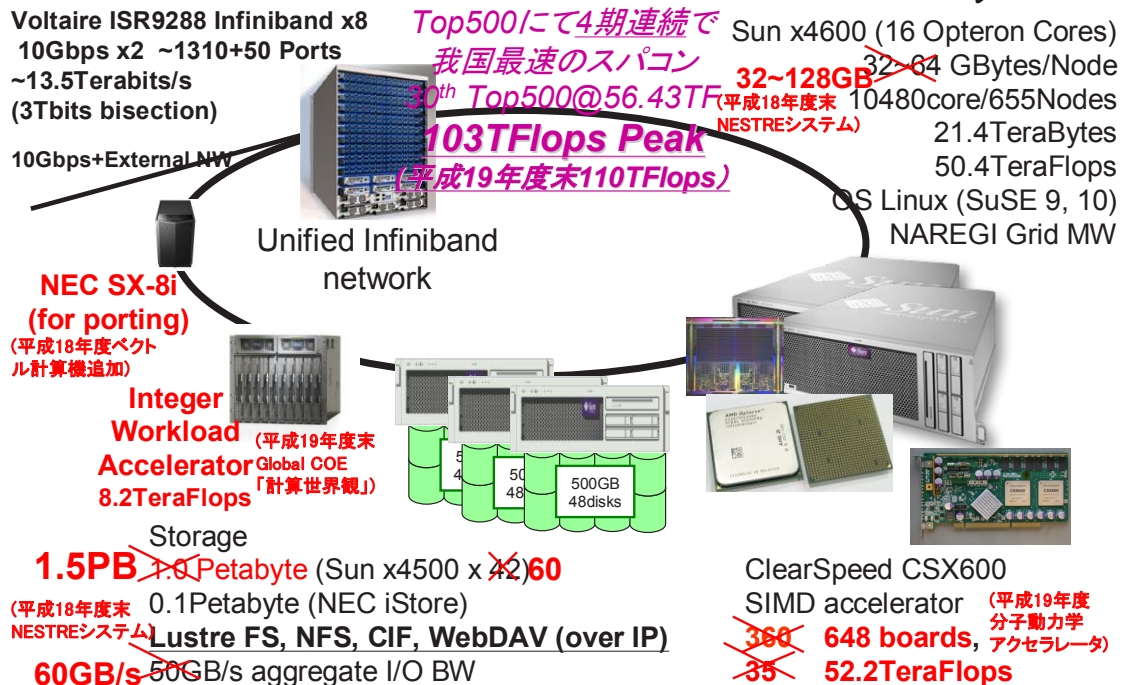


ClearSpeed CSX600
SIMD accelerator
360 boards,
35 TeraFlops





TSUBAMEの'06-07年度機能拡張 (GSICセンター全体の複合システム)



TSUBAMEの一年半の運用成果: 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



4期連続日本一
4期連続性能向上
(世界初)

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

『みんなのスパコン』
TSUBAMEによるペタ
スケールへの飛翔

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】



Global COE
「計算世界観」

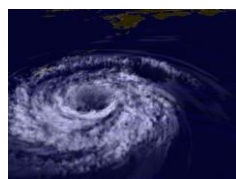


Microsoft 包括契約
HPC Institute/TCI



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 β 2連携

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



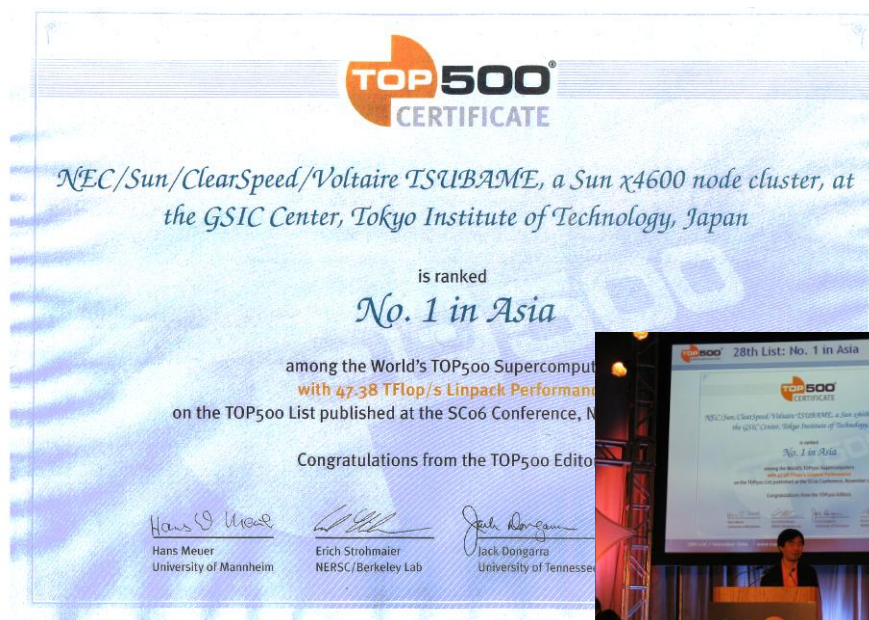
4. 学内の分散した情報基盤の
集約化・ホスティング

・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



27-29th TOP500アジア最速スパコン (4期連続日本最速)



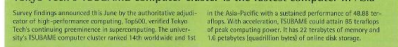
TSUBAMEのテレビ出演



他、NHKニュース
(2007年11月)など



The Wall Street Journal,
日経、毎日、朝日など



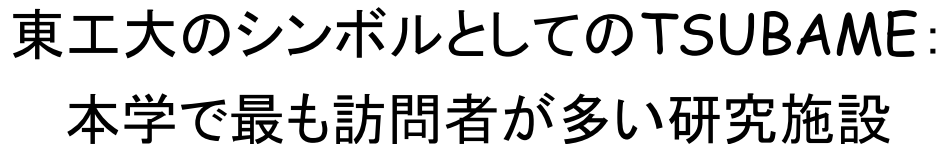
Updell is a joint effort by the University of Manitoba, the University of Toronto, and the National Science Foundation Scientific Computing Center at Lawrence Berkeley National Laboratory. It ranks supercomputers twice annually on the basis of the LINPACK benchmarks. Those benchmarks measure how fast computers handle a regimen of linear equations. TSUBAME, the world's fastest supercomputer, ranked first in the world through all three Top500 rankings conducted during its life.

"We're proud to be a world leader in computing performance," acknowledges Tokyo Tech's Professor Satoshi Matsumoto (pictured above), who directs the TSUBAME project. "But we're even more proud to have another milestone. This system is a huge step toward our goal of making supercomputing available to nongovernment users. TSUBAME's user population has grown to 1,200 super-computer-class users. And any of the more than 10,000 members of the Tokyo Tech community, including undergraduate students, can use the system. That's a huge step. The entire Japanese population of a few score users that a conventional supercomputer would have served."

Mutsaers, of Tokyo Tech's Global Scientific Information and Computing Center, earned his doctorate in computer science from the University of Tokyo in 1993. A pioneer in grid computing, he has been at Tokyo Tech since 1996 and also serves as a subleader of Japan's National Grid Initiative (NAGRI) for creating middleware for next-generation e-business infrastructure. Recognized for his work in grid computing, Mutsaers has received excellence from the Information Processing Society of Japan in 1999 and the SPSS Prize from the Japan Society for the Promotion of Science in 2000, the latter presented by Prince Akishinohima. Mutsaers is a familiar face at the podium at international supercomputing symposia and conferences. He has been invited to give respected international journals, including *Concurrency: Practice and Experience* and the *Journal of Grid Computing*.

TSUBAME's unprecedented accessibility to researchers in diverse fields is partly attributable to the system's PC-based architecture. It is open to all in Matsuyama, and its use is not limited to supercomputing for everyone. That commitment has yielded results through exagical outreach to potential users inside and outside

Nature 特集広告



- 大臣から外国の北欧の高校生まで
- DoE, DoD, NSF, NASA, CNRS(フランス), 等のセンター長クラス



TSUBAMEの認知度 on Google (2008/2/6, 日本語のページ)



- ・ **“TSUBAME” 2006稼動 - 合計20,000件程度**

c.f.

- ・ 「東工大」「東京工業大学」 - 332,000件
- ・ 「東工大ポータル」 2006稼動 - 643件
- ・ 「TITANET」 1994稼動 - 404件
- ・ 「地球シミュレータ」 2002稼動 - 66,400件
- ・ 「白川英樹」 - 28,400件



TSUBAMEの「経済効果」

-TSUBAMEに依存した学内研究の合計資金 -

- ・ **東工大GSIC等合計 24億円 (2007~2012)**
(総額、ただし研究費で運用経費ではない)
- ・ 科研費・受託研究費・COEなどで、TSUBAMEの有料サービスを利用している研究費
 - 以下は除く
 - ・ 無料サービス利用
 - ・ 先端イノベーションなどの企業利用
 - ・ GSICの大規模実験
 - **現状の集計: 平成19年度では合計12.47億円**
 - ・ 間接経費2億円以上(集計中)



2006年度末NESTREシステムによる ストレージ拡張



総合 1.1PB => 1.6PBに
高信頼大容量サービス



TSUBAME拡張ストレージ利用による賞

Argonne, Virginia Tech Win Storage Challenge Competition - Microsoft Internet Explorer - [F5で再読み込み]

SC07 Storage Challenge

HPC wire

The Leading Source for Global News and Information Covering the Ecosystem of High Productivity Computing / November 26, 2007

Home Page | Free Subscription | Advertising | About HPCwire

Breaking News:

Argonne, Virginia Tech Win Storage Challenge Competition

BLACKSBURG, Va., Nov. 26 -- A team of researchers led by Pavan Balaji of Argonne National Laboratory and Wu Feng of Virginia Tech won an international competition for the most effective approach in using large-scale storage for high performance computing. The award was presented Nov. 15 at SC07, the world's premier conference on high performance computing and networking.

Using a novel software framework for distributed I/O called ParaMEDIC, the team of researchers from Argonne National Laboratory, Virginia Tech, and North Carolina State University searched the sequences of all completed microbial genomes against each other. The aim was twofold: to discover missing genes and to speed future searches by generating a complete genome similarity tree. The ParaMEDIC software framework used a semantics-based approach to create a metadata representation that was four orders of magnitude smaller than the actual output data.

"Using ParaMEDIC, the entire genome similarity tree, corresponding to a petabyte of data, can fit into a 4-gigabyte iPod nano," said Balaji.

This entire task required many millions of CPU-hours of computational capability and generated a petabyte of uncompressed output. Since not many supercomputer centers provide both the computational and storage resources required for this task simultaneously, the research team relied on a worldwide supercomputer that aggregated the compute resources from various locations within the U.S. and the TSUBAME storage resources at the Tokyo Institute of Technology in Japan, with technical support from Sun Microsystems. The largest portion of the compute cycles were provided by Virginia Tech's System X



分子動力学用ClearSpeed拡張 (2007年11月)

Advance™ Dual CSX600 PCI-X追加

- PCI-X accelerator boards with peak performance of 96GFlops
 - CSX600 SIMD proc x 2 + 1GB DRAM on board
 - 0.5GFlops PE x 96 per processor
 - 8" single-slot PCI-X
 - 25W / board

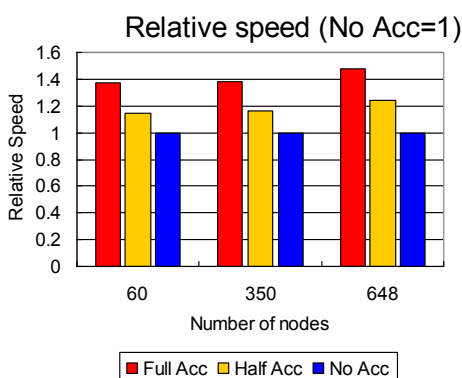
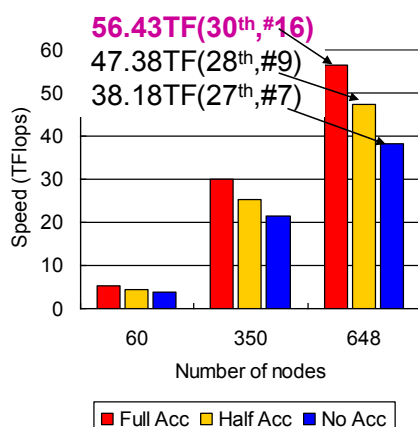


- すでにTSUBAMEでは655
ノード中360枚装備していたが、
今回ほぼ全ノードまで拡張配備、
主に分子動力学用
(他のアプリも間接的に利益が)

両者とも~80GFlops / node、
消費電力
1300W (x4600 Opteron)
vs.
25W (ClearSpeed)



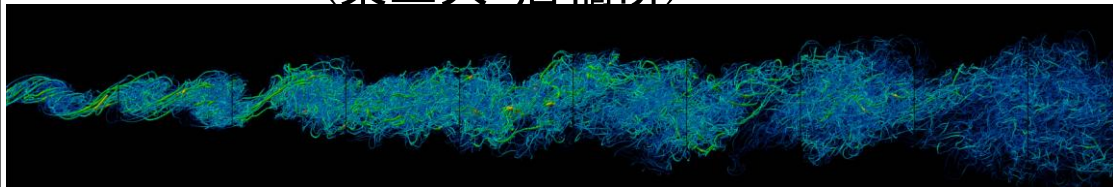
ClearSpeed拡張による性能向上



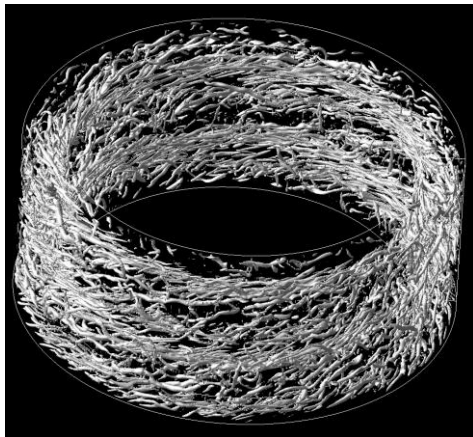
- 47.38TF with 648 nodes and 360 Accelerators for the 28th Top500 (48.88TF for the 29th Top500, #14)
 - +24 % improvement over No Acc (38.18TF)
 - +25.5GFlops per accelerator
 - Matrix size N=1148160 (It was 1334160 in No Acc)
 - 5.9hours
- With new DGEMM&648 cards, 56.43TF (16th on the 30th Top500)



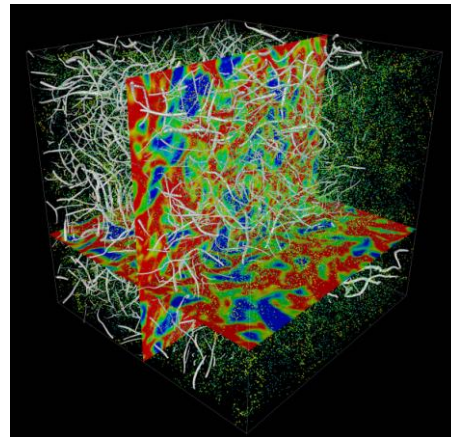
アプリ例: 様々な乱流場のDNS (東工大・店橋研)



空間発展乱流混合層



Taylor-Couette 流



液滴を伴う一様等方性乱流

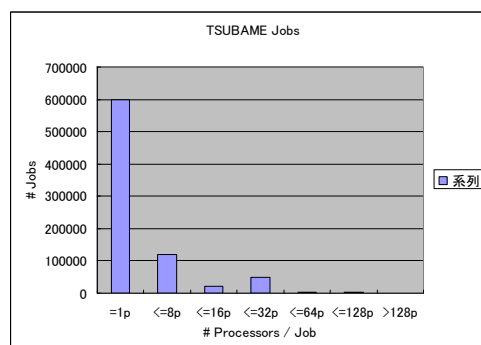


TSUBAME ジョブ統計

Dec. 2006-Aug.2007 (#Jobs)

1400人のユーザ

- 797,886 Jobs (~3270 daily)
- 597,438 serial jobs (74.8%)
- 121,108 $\leq 8p$ jobs (15.2%) **90%**
- 129,398 ISV Application Jobs (16.2%)
- *However, $>32p$ jobs account for 2/3 of cumulative CPU usage*

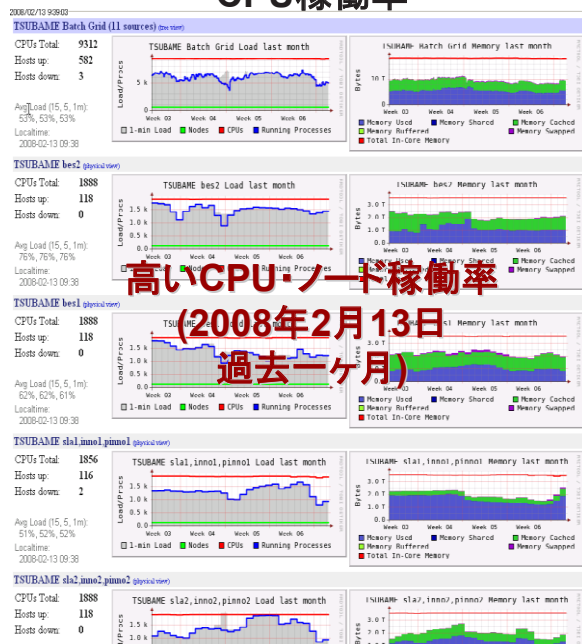


Coexistence of ease-of-use in both
- short duration parameter survey
- large scale MPI
予想通りのジョブ配分



TSUBAMEの危機的な混雑度

CPU稼働率



高いCPU・ノード稼働率
(2008年2月13日
過去一ヶ月)

2008/02/13 9:27:26
The Network Weather Service

Tsubame

待ち時間予測システム (UCSBと共同研究)

	0-2	3-4	5-8	9-256	
RAM128GB	●	●			
RAM4GB	●	●	●	●	
bes1	●	●	●	●	
bes2	●	●	●	●	
inno1	●	●	●	●	
inno2	●	●	●	●	
inno3	●	●	●	●	
mopac	●	●	●	●	
novice	●	●	●	●	
sla1	●	●	●	●	
sla2	●	●	●	●	
sla3	●	●	●	●	

Legend

- 0-17 hours
- 17-34 hours
- 34-71 hours
- 2-5 days
- >5 days
- X No Such Node Range for Queue



みんなのスパコン

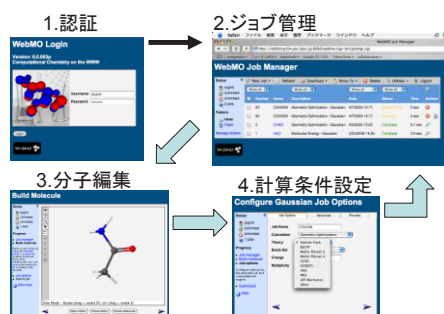
スパコンを全学教育研究で活用

「みんなの全学アカウント」

- スパコンを用いた高度教育
 - スパコンではじめて可能になる高度シミュレーションを学部の学生演習・実験で
- グリッド技術を用いた研究室のIT環境とスパコンとのシームレス化
- ポータルによる高度アプリ利用

アプリポータル事例: WebMO

量子化学プログラム(Gaussian, NWChem, GAMESS, MOPAC, Molpro等)をWebブラウザから簡単に利用できるインターフェース



デスクトップがそのまま数千プロセスになったみたい😊



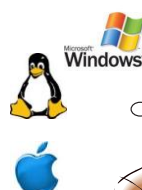


みんなのスパコン



(4) スパコンによるIT統合の新潮流

- All University AAA and E-mail system
 - 50TB of Thumper, 9 Galaxy 1 nodes from TSUBAME
- 全学ストレージサービス(試行中・H20より本格運用)
 - 全学全員数十GB (パソコンでマウント可)
 - 研究結果のレポジトリ化 (シミュレーション結果など)
- 高度研究用アプリの教育利用(Open CourseWare)
- 図書館業務のホスティング
- 事務系のホスティング(VMによる)



これでラップ
トップの全デー
タをバックアップ
できるな◎



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】

- 企業のパソコン・クラスタ上の所有ソフトがそのまま、数十倍の速度や規模で動作
 - 単にCPUパワーだけでなく、メモリ・ストレージ・ネットワークも
 - 「アキバのプログラムは動くがアキバでは買えないスパコン」
- アジアーのスパコンを、既存シミュレーション利用者以外(ICT、金融、流通、サービス業界等)にも開放
- 本学への産学連携・共同研究の呼び水



Ruby
A Programmer's Best Friend



The Apache
Software Foundation
<http://www.apache.org/>

APACHE
PORTALS

python

PYTHON
Powered

mod_perl



tapestry

php

Perl

Java

OpenEJB

Struts





採用課題一覧 (2008/1) 11件

- 戦略分野利用推進
 - 巨大生体分子の非経験的分子軌道法による設計指針構築 三菱化学科学技術研究センター
 - タンパク質一次構造の網羅的解析による創薬技術の開発 (株)ライフフィクス
 - コンピュータ支援によるポリアミン誘導体医薬品の開発 (株)アミンファーマ研究所
 - CONFLEXを用いた配座探索および結晶多形解析 (株)コンフレックス
 - 電磁場中大規模粉体挙動シミュレーションによる電子写真設計プロセス革新 (株)リコー
- 新規利用拡大
 - 銀行業・保険業におけるALM (Asset Liability Management) システムの開発 (株)ニューメリカルテクノロジーズ
 - 大規模分散検索エンジン製品の開発 (株)ビジネスサーチテクノロジー
 - ワイドギャップナノ構造体精密加工のシミュレーション (株)日本電気
 - 機能性無機材料の光学的電子的物性と構造設計の研究 (株)住友化学
 - 混相流シミュレーションコードの並列拡張性能の評価 (株)計算流体力学研究所
 - 高層ビルの大規模耐震構造解析 (株)アライドエンジニアリング
- その他、東芝・三菱住友UFJ・ニッセイなどが応募予定



三井住友系の金融ベンチャー Numerical Technologies



ニューメリカル テクノロジーズ 株式会社
代表取締役社長

鳥居秀行



「移植は一日」「3000CPUのモンテカルロで画期的な成果」
「自分で用意したら数十億円」



GlobalCOE「計算世界観」

科学の対象を
計算を中心に見直そう！

新視点の改革



HPCSプロジェクト

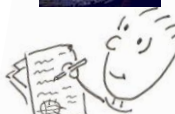
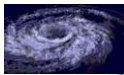
by 米国 DARPA

目標: スパコン生産性向上

手法: 巨大システム開発

超大規模予算

背景



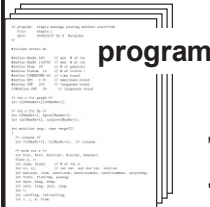
科学におけるコンピュータ
シミュレーション

複雑・大規模

改革が必要！



⇒ スパコン



program



TSUBAME@東工大
アジア第一のスパコン

例)・物質分析→物質合成
・社会システムの分析



さらなる要望

数々の成果

従来型改革



GlobalCOE による HPC人材育成

実践ができる数理計算科学者の育成

目指す人物像

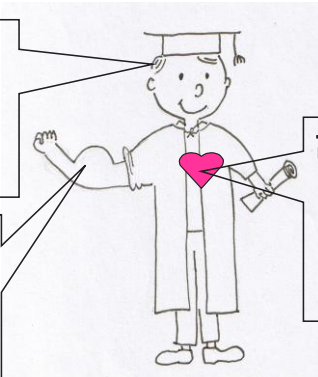
「計算世界観」を実践できる人材

解析・設計の理論

- ・アルゴリズム
- ・離散構造解析
- ・確率構造解析

HPC 実践技術

- ・モデル化技法
- ・プログラミング
- ・システム実現



自らの専門+

計算世界観

- ・異分野に
乗り込む
- ・交流力
- ・吸収力



対象分野の研究者

人材不足！

解消へ！

1. 魅力的なテーマ

2. 組織的な指導体制

3. 多くの分野へ

実践もできる博士

計算世界観

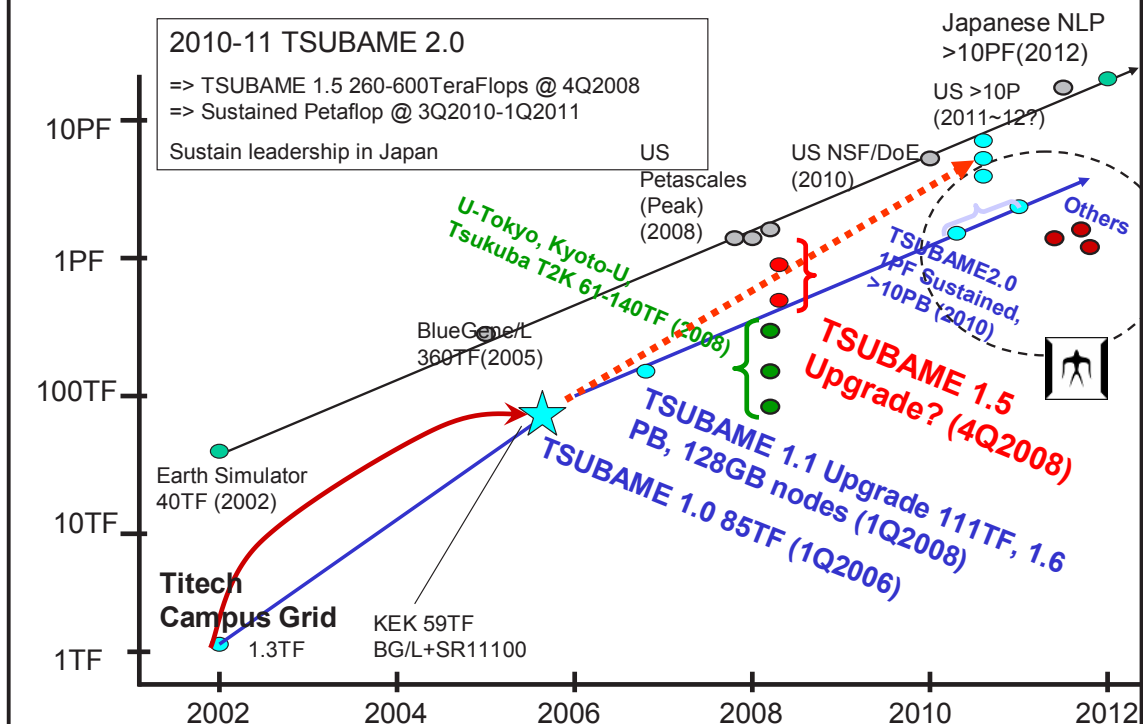


H20年度へ向けて

- 更なる性能向上 => TSUBAME 1.5
 - Top500日本一維持または奪還
- Global COEマシンなどの資源バスターによる一体化による増強
- 教育・事務系を含むホスティングの更なる強化
 - 教育でのさらなる活用も → システム統合へ
- 本格的な大規模データ基盤・データグリッドサービスの開始
- 全国基盤センター化へ向けた国立情報学研究所と連携したCSI(サイバーサイエンスインフラ)グリッド運用
 - NAREGIグリッドミドル v.1.0の運用開始
- 特任教員(連携研究員含む)やSEの拡充
 - 特任教員 => ネットワーク・認証を合わせて4名追加
 - TSUBAME SEも実質1名増加(データ・ホスティングなど)
- 課金体制の見直し
 - 公平性の確保 => より高いレベルのQoSを妥当な対価で
- TSUBAME2.0へ向けたアクティビティ開始



TSUBAME2.0へ向けた今後の進展

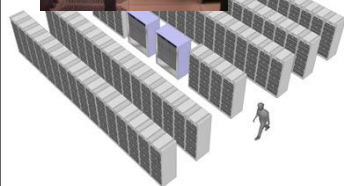




In the Supercomputing Landscape, Petaflops class is already here... in early 2008



Other Petaflops 2008/2009
 - LANL/IBM "Roadrunner"
 - JICS/Cray(?) (NSF Track 2)
 - ORNL/Cray
 - ANL/IBM BG/P
 - EU Machines (Julich...)
 ...



2008 LLNL/IBM "BlueGene/P"
 ~300,000 PPC Cores, ~1PFlops
 ~72 racks, ~400m2 floorspace
 ~3MW Power, *copper* cabling

2008Q1 TACC/Sun "Ranger"
 ~52,600 "Barcelona" Opteron
 CPU Cores, ~500TFlops
 ~100 racks, ~300m2 floorspace
 2.4MW Power, 1.4km IB cx4
copper cabling
 2 Petabytes HDD

> 10 Petaflops
 > million cores
 > 10s Petabytes
 planned for 2011-2012
 in the US, Japan, (EU),
 (other APAC)



2010年TSUBAME 2.0の概要

- 我が国初のペタフロップスマシン(?)
 - 世界的なリーダーシップ、単なるマシンでなくレファレンス
 - わが国初のペタフロップスマシン
- 性能・消費電力等様々な技術革新⇒世界基準スパコンへ
 - 低調達コスト・低消費電力・高バンド幅・最適設計
- 現状のTSUBAMEアーキテクチャとの連続性
 - 現TSUBAMEからも、研究室や企業のパソコンからも簡単に
- 大規模大容量データ
 - 10ペタバイト級・長期間(10年単位)データアーカイブのサポート
- 研究室などインフラとよりシームレスな運用・ホスティング
 - OSや環境の選択・仮想マシンによる運用も
 - NII-サイバーサイエンス ナショナルグリッドのリソースセンター
- 高信頼なITサービスのホスティング
 - 高セキュリティ・高信頼・高アベイラビリティ
 - 学内の更に多くのサービスをホスティングへ(特に教育システム統合)

すでに概要設計、さらに各メ
 ーカーとの技術協議は開始



TSUBAMEの電力性能は優れているが。。。
今後10倍以上の向上が必要

Machine	CPU Cores	Watts	Peak GFLOPS	Peak MFLOPS /Watt	Watts/CPU
TSUBAME(Opteron)	10480	800,000	50,400	63	76.336
TSUBAME(w/ClearSpeed)	11,200	810,000	103,000	127.16	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	2000000	2500	20