

ペタに向けた センタースパコングリッドインフラ

- ペタスケールのインフォマティクスへ -

計算科学の戦略と次世代スーパーコンピュータ
シンポジウム

2006年4月5日(金)@筑波大学

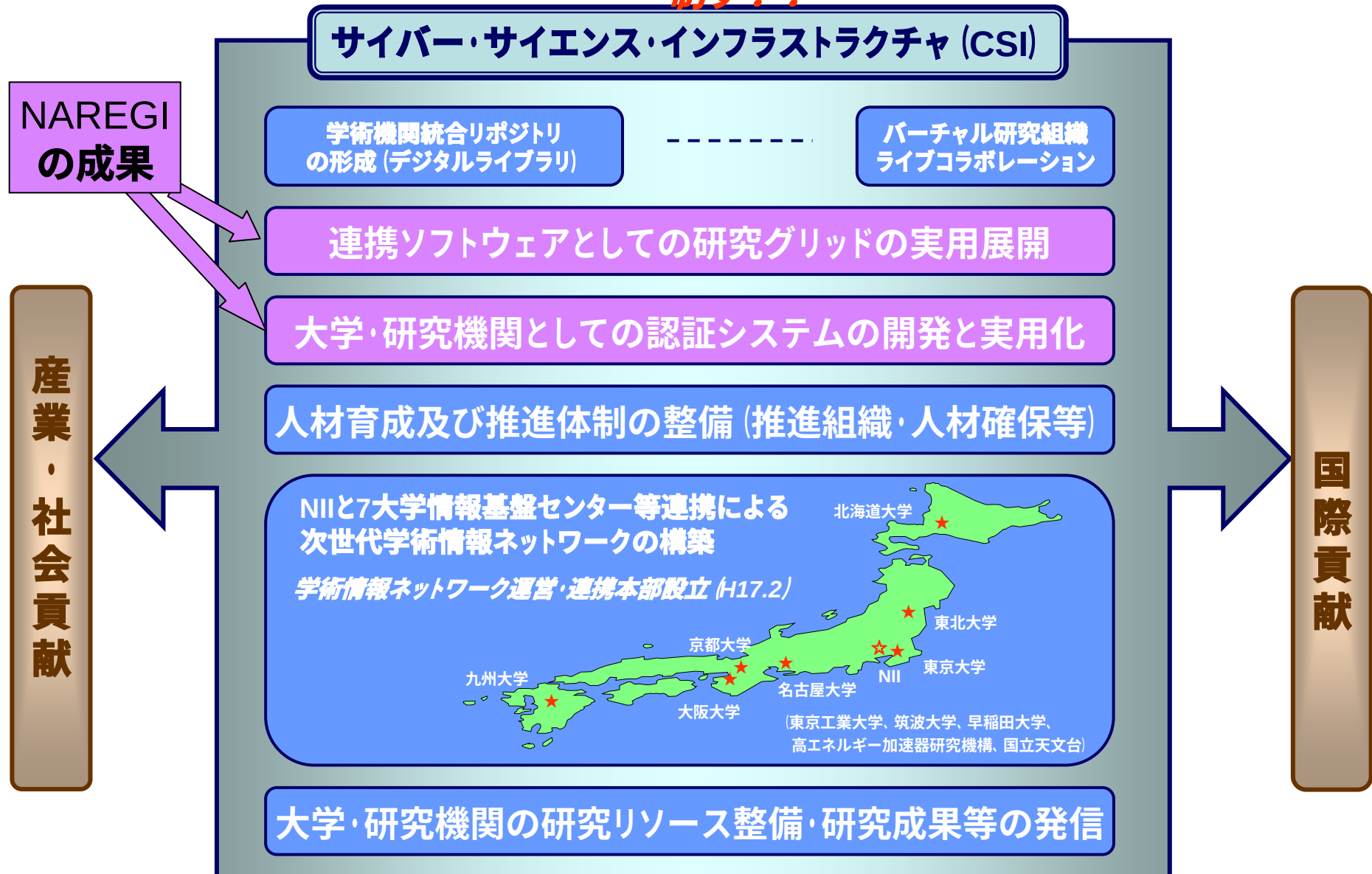
東京工業大学
学術国際情報センター
& 国立情報学研究所 NAREGIプロジェクト

松岡 聡

日本の最先端学術情報基盤の構築に向けて

2

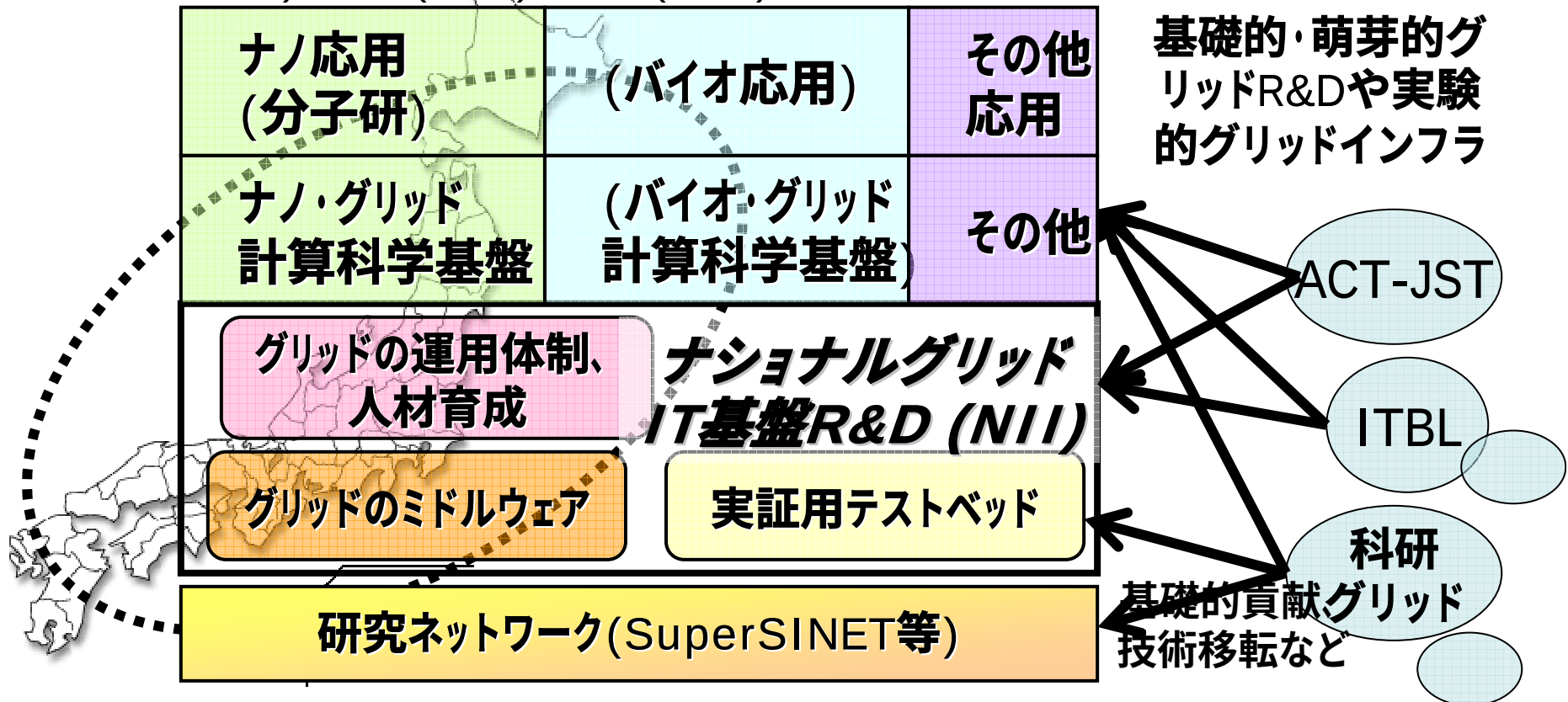
最先端の学術研究基盤が、今後の学術分野・産業分野での国際協調・競争の死命を制す！！



NAREGIプロジェクト(2003-2011)



- ナショナル研究グリッドの基盤ソフトウェアおよびアプリケーション技術のR&D(特にセンターグリッド)
 - 2003年4月、文部科学省、国立情報学研究所、分子研、にて開始
(National Research Grid Initiative)
 - プロジェクトリーダー: 三浦健一(NII)、サブリーダー: 関口(産総研), 松岡(東工大/NII), 下條(阪大)、青柳(九大)



その中で、大学センタースパコンに対する 要求: バイオ(インフォマティックス)の例

- バイオインフォの中だけでも様々なアプリケーションの種類・規模・用途

- FP Intensive
 - 第一原理計算(MO): Gaussian, MOPAC, UT-Chem, CHARMM...
 - MD (粒子系): AMBER, Molpro, GROMACS...
- (Almost) Embarrassingly Parallel FP
 - FMO
 - SA-Replica Exchange
- Integer Intensive / Search
 - Homology Search/Sequence Matching (BLAST, Smith Waterman, ...)
 - Multiple Alignments
 - Graph-Theoretic Metabolic Network
 - Phylogenetics
- Pattern Matching / Vision processing
 - HEM Structural analysis, various Cardiology
- Data Intensive
 - Genomic Data Mining
- Database / Data Archiving
 - PDB, Genbank, ...
- Visualization
 - GaussView, ...

単なるFPの計算エンジンだけでなく、全体として計算機の「総合力」が求められる

- これらが複雑にからむワークフロー・連成⇒個々のユーザごとに異なる
- 従って、ユーザの要求は千差万別、それぞれハイエンド、時には(いい意味で)我侔

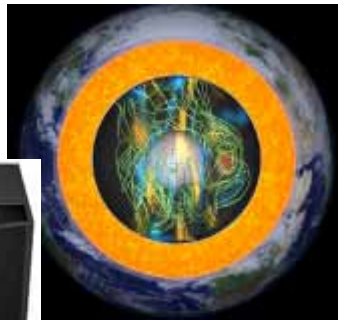
PF級スーパーコンピューティングの 必要性と将来

莫大な計算機パワーの有効な維持管理と活用・計画的な増強

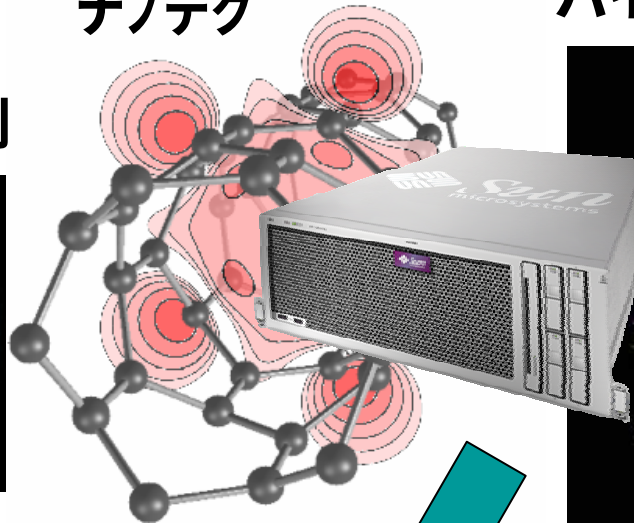
流体 (CIP法)



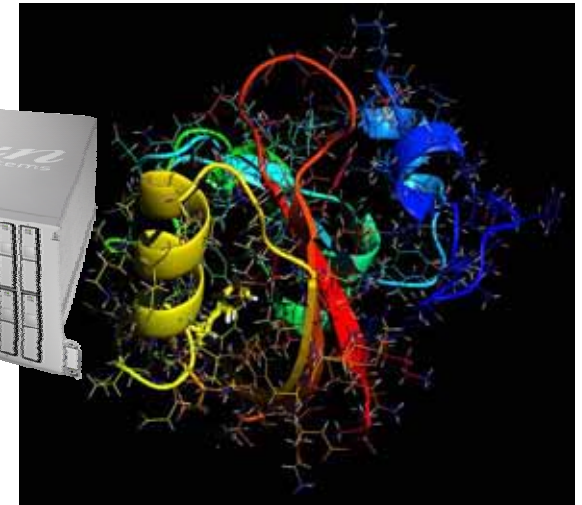
地球磁場予測



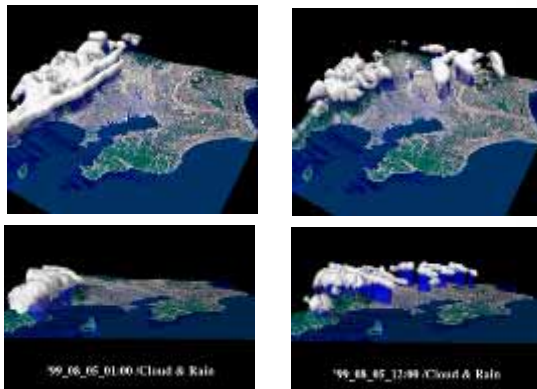
ナノテク



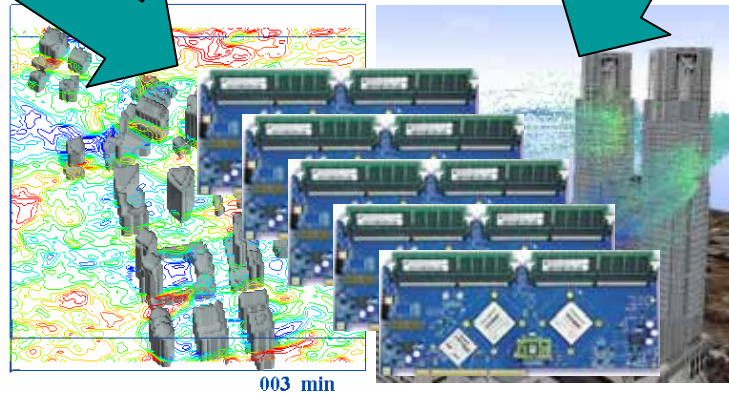
バイオシミュレーション



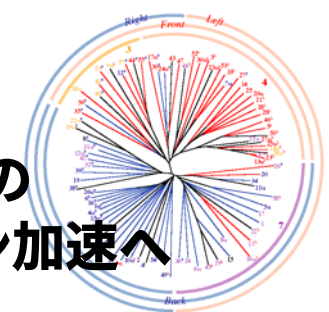
高精度気象予測



都市環境予測



バイオシミュレーションと
バイオインフォマティクス
の融合



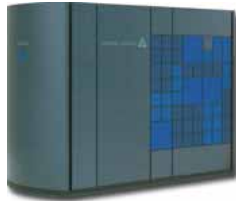
高 C/P の
スパコン加速へ

みんなのスパコン



東工大全体へのスーパーコンピューティングパワーの解放

「孤高」な
ハイエンド



多大な利用環境のギャップ

- Window, Linux等と異なるOSやミドル、コマンド
- ファイルなどのパソコンとスパコンでの非共有
- 特殊なマシンでの重要なアプリの非対応
- プログラム言語・ツール(Visual Studio)などの欠如
- 時代遅れのバッチ環境

速いけど
難しい。
パソコンで
いいか。



現状のキャンパスグリッド: 数百ユーザ
→1500ユーザに増加(2006)

東工大に入学すると**新入生**はいきなり
日本一のスーパコンピュータユーザー
になれる。



シームレス・ユビキタスな研究・教育用
キャンパススパコングリッド環境の構築

=> **スーパーコンピューティングの普遍化**
による**ブレークスルーへの期待**

みんなのスパコン



キャパシティ: スパコンを研究・教育・事務のあらゆる側面で活用

- 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なシミュレーションを学部学生演習・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
- 高度なISPサービス
 - 例: 一人あたり数十GBの超高信頼ストレージサービス
 - 研究結果のレポジトリ化



サイバーサイエンスにおける基盤センター役割 「スパコンセンター」からCSI基盤センターへ



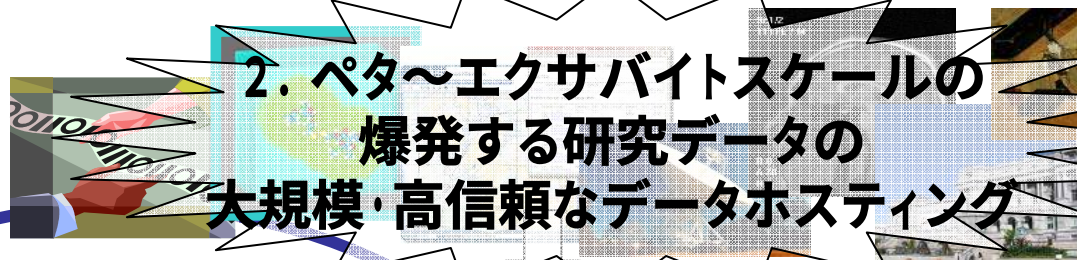
- ケーパビリティ・キャパシティ・グリッド・データ・サービスのホスティング-

1. HPCセンターとしての京速へのコミット・キャパシティとケーパビリティの両立



京速
ケーパビリティ
10PFlops以上

2. ペタ～エクサバイトスケールの爆発する研究データの
大規模・高信頼なデータホスティング



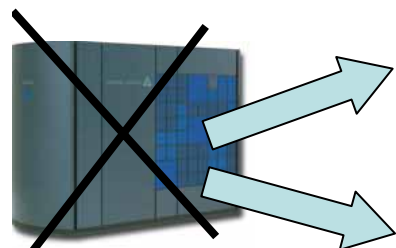
スケールアップ

4. 学内外高性能
ネットワークおよびその上でのグリッド・汎用サービスのホスティング

マルチ・ペタバイトストレージ



3. サイバーサイエンス基盤
全国および国際
グリッド・認証などの
ホスティング



現状：貧困な特殊スパコン数～10数TFlops
欠乏ストレージ
貧弱ネットワーク

サブ地球
ケーパビリティ
数100T
～1PFlops

広範利用
キャパシティ
数百TFlops
デスクトップ・小クラス
タとの連続性



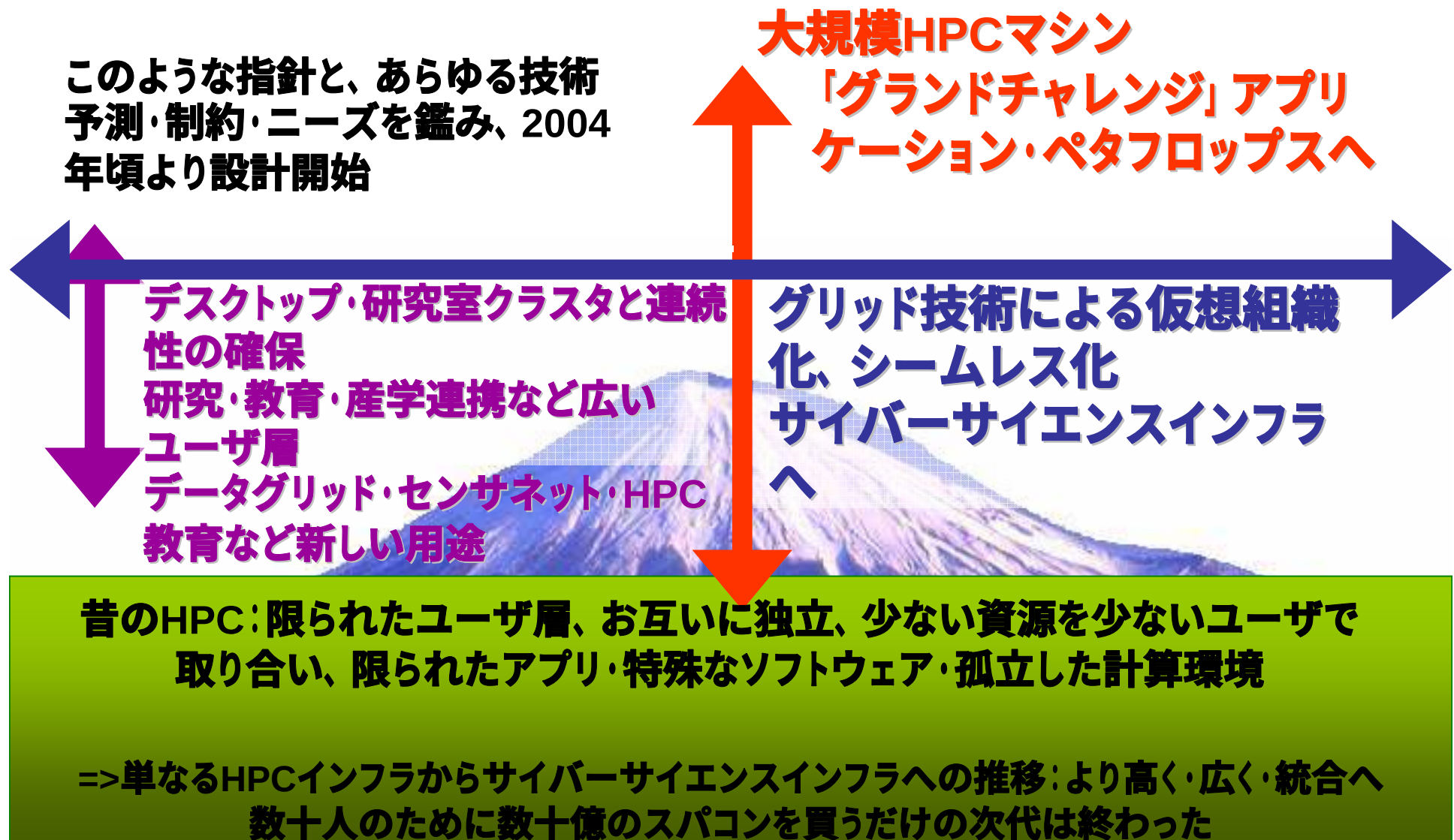
マシン種別	ベクトル SX, X1	BlueGene	RISC/Itanium	x86 (サーバ級)	ClearSpeed Grape等
CPU全体スカラー性能	低	低	高	高	一般はNA 特殊アプリで高
CPUクロックあたりFPU性能	中高	中高	中	中	特殊アプリで 極高
FLOPS相対メモリバンド幅	高	高	中高	中	低～極低
メモリ空間 (小 数GB 中 数10GB 大 それ以上)	中～大	小	中～大	小～中 (将来は大)	小～極小
FLOPS相対マシン内ネットワーク性能	高	高	中～高	小～高 (様々)	NA
汎用I/O性能 (HDD)	中～高	中～高	高	中～高	NA
汎用ネットワーク性能(TCP/IP)	低	中	高	高	NA

マシン種別	ベクトルSX, X1	BlueGene	RISC/Itanium	x86 (サーバ級)	ClearSpeed Grape等
ソフト・アプリ 汎用性・移植性	中	中	高 ただしx86 には負ける	高	極低
FLOPS相対コスト パフォーマンス	極低	中	低	中	高(適用アプリ)
マシン・ソフトの 継続性	中～低	中～低	中高	高	低
汎用資源としての グリッド上の availability	低	低	中	高	低
FLOPS相対消費 電力	極高	低	高	中	低～極低
今後の個々のア プリや、バイオ VOへの適合性	?	?	?	?	?

センターにおけるマシンデザイン

- 種々の技術選択肢および制約事項
 - 今まで述べた以外に
 - コンポーネンツのロードマップ・availability
 - 予算
 - 設置面積
 - 電力消費量
 - 冷房
 - SE/スタッフ
 - 学内政治
 - 〇 〇 〇 〇 〇 〇 〇 〇 〇

スパコンを含むHPCインフラに対する要求と指針



2006年東工大スパコン "TSUBAME"

ハイブリッドアーキテクチャによる高コストパフォーマンス・ 汎用性とペタスケールに向けた加速の両立

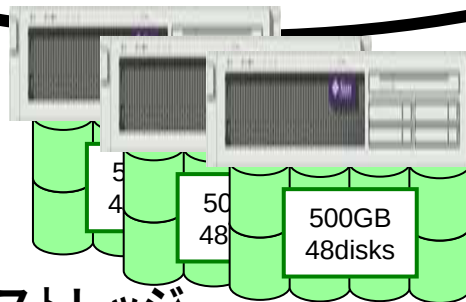
Voltaire Infiniband
10Gbps x 2 (x 2)
x ~700ポート

10Gbps+外部
ネットワーク

NEC SX-8
ベクトル計算機
(レガシー・移植試験等)
(計画中)



2006年4月現在、ピーク
性能では我が国 No.1



ストレージ

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

Sun/AMD高性能計算クラスター
(Opteron Dual core 8-Way)

10480core/655ノード

50TeraFlops

OS(現状) Linux

(検討中) Solaris, Windows

NAREGIグリッドミドル



ClearSpeed
CSX600
SIMDアクセラレータ
35TeraFlops
(各ノード1ボード)

東工大新スパコンの設計 (1)

- 「トロフィー・ハイエンド」と「みんなのスパコン」の両立
 - **高性能** 64bit x86 **互換マルチコアCPUの採用**
 - 圧倒的な汎用性・INT/FP高性能・高コストパフォーマンス・高信頼
 - 最先端の設計とプロセス技術: 高性能かつ低消費電力
 - Best applications & software availability: OS (Linux/Solaris/Windows), languages/compilers/programming tools, libraries, Grid tools, all ISV Applications
 - FAT Node Architecture
 - Multicore SMP – most flexible parallel programming
 - Facilitate multithreaded acceleration
 - High memory capacity per node (32/64GB)
 - Large total memory – 21.4 Terabytes
 - Low node count – improved fault tolerance, easen network design
 - High Bandwidth Infiniband Network
 - **今回は** (Restricted) two-staged fat tree (3-D Torus**も可能**)
 - High bandwidth (10-20Gbps/link), multi-lane, low latency (< 10microsec), reliable/redundant (dual-lane)
 - Very large switch (288 ports) => low switch count, low latency
 - Resilient to all types of communications; nearest neighbor, scatter/gather collectives, embedding multi-dimensional networks

東工大新スパコンの設計 (2)

- Hybrid Architecture: General-Purpose Scalar + SIMD Vector Acceleration w/ ClearSpeed CSX600
 - 35 Teraflops peak @ 90 KW (~ 1 rack of TSUBAME)
 - General purpose programmable SIMD Vector architecture
- **PB級大規模・超高速・超高信頼ストレージ**
 - All Disk Storage Architecture (no tapes), 1.1Petabyte
 - Ultra reliable SAN/NFS storage for /home (NEC iStore), 100GB
 - Fast NAS/Lustre PFS for /work (Sun Thumper), 1PB
 - Low cost / high performance SATA2 (500GB/unit)
 - High Density packaging (Sun Thumper), 24TeraBytes/4U
 - Reliability thru RAID6, disk rotation, SAN redundancy (iStore)
 - Overall HW data loss: once / 1000 years
 - High bandwidth NAS I/O: ~60GBytes/s Livermore Benchmark
 - Unified Storage and Cluster interconnect: low cost, high bandwidth, unified storage view from all nodes w/o special I/O nodes or SW

各パートナーの技術



Empowered by Innovation

0.1ペタバイト超高信頼ストレージ

地球シミュレータおよび東工大キャンパスグリッドの構築・運用技術



高性能・低消費電力サーバCPU
(2.4/2.6GHz, 10480 コア)



ノード計算機: 8-Way/16-Core 655ノード50TFlops
1ペタバイト大容量・高密度・高速ストレージ



96 GigaFlops SIMD アクセラレータボード
(360 枚、35TeraFlops)



Infiniband 高性能10Gbps ネットワーク
288ポート高性能スイッチ



Cluster
File
Systems, Inc.

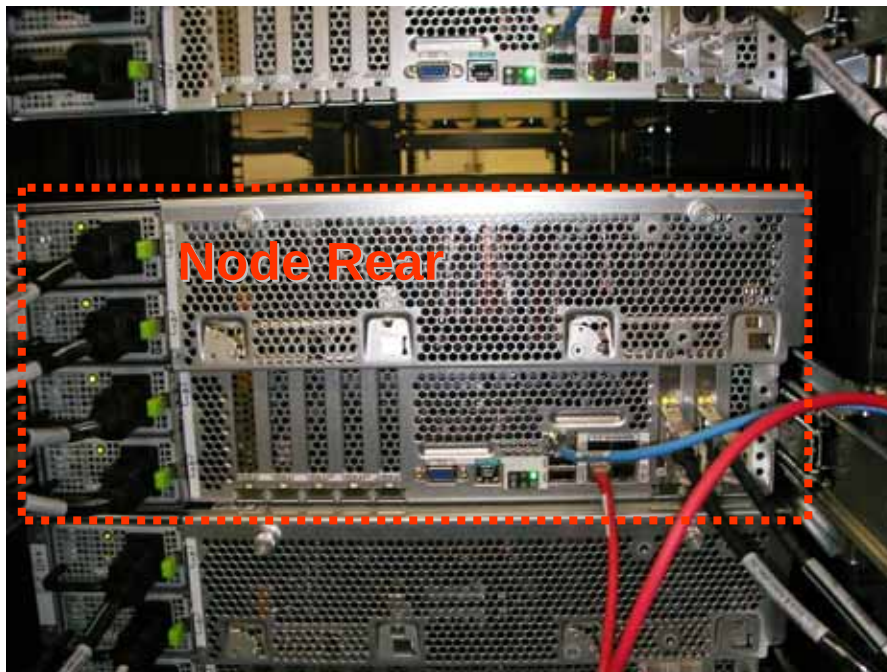


大規模並列ファイルシステム



研究グリッド国家プロジェクト・グリッドおよびアカウント管理ミドル





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 (~20 units)

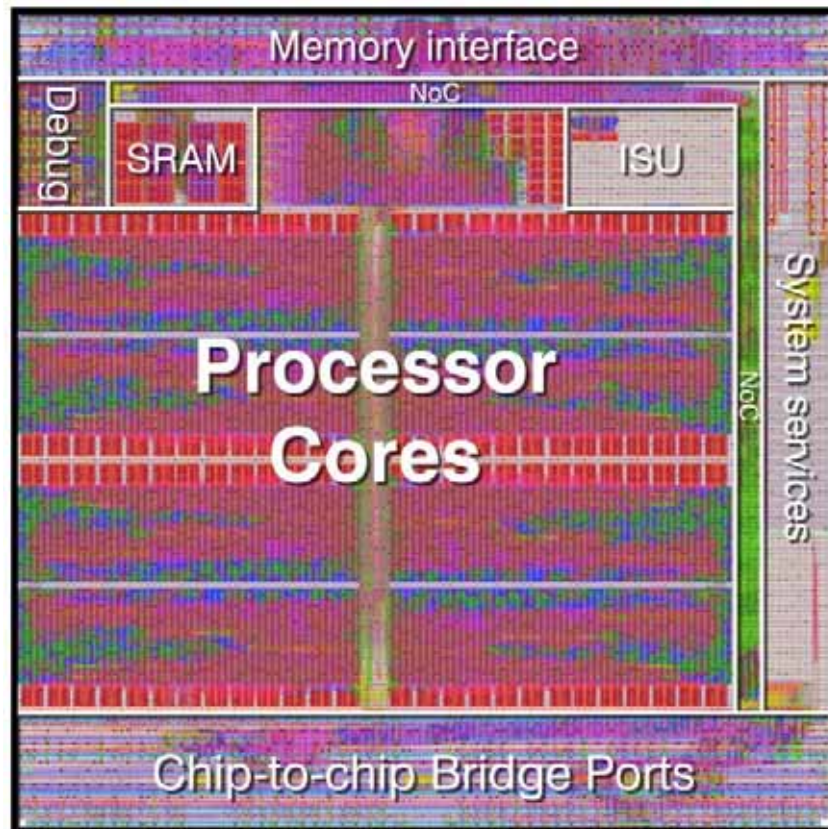
Advance™ Dual CSX600 PCI-X accelerator board



**Currently 360 boards in
TSUBAME -> Plans to increase to
~600 (>50TeraFlops)**

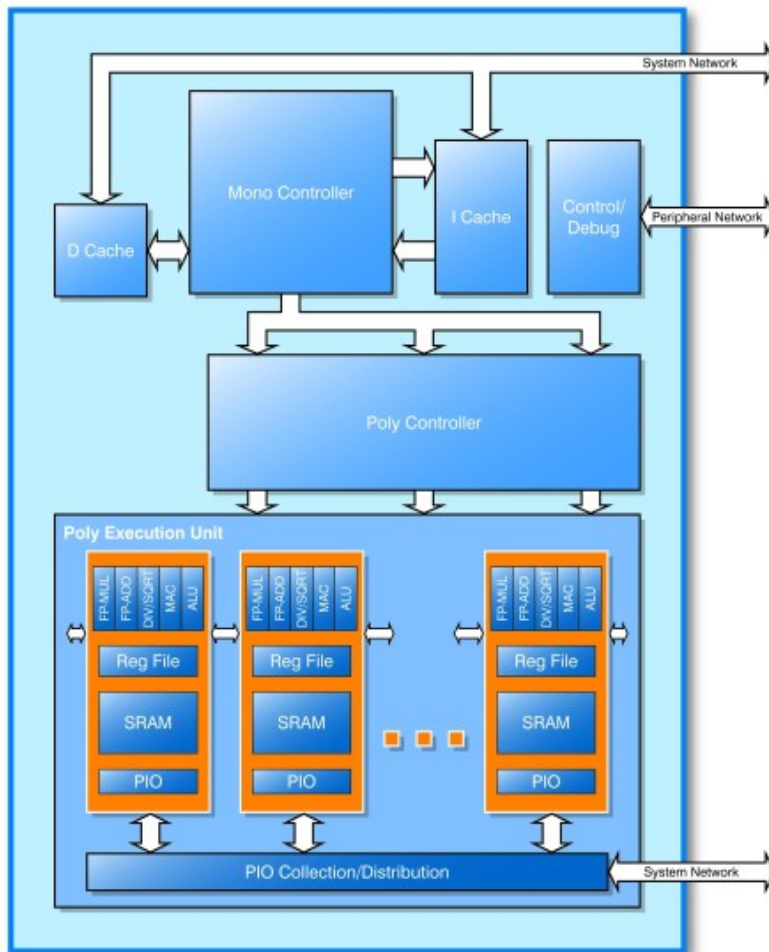
- 50 DGEMM GFLOPS sustained
- 0.4 M 1K complex single precision FFTs/s (20 GFLOPS)
- ~200 Gbytes/s aggregate B/W to on-chip memories
- 6.4 Gbytes/s aggregate B/W to local ECC DDR2-DRAM
- 1 Gbyte of local DRAM (512 Mbytes per CSX600)
- ~1 Gbyte/s to/from board via PCI-X @133 MHz
 - 4 Gbytes/s PCI-E RSN
- < 25 watts for entire card (8" single-slot PCI-X)

CSX600 coprocessor layout



- Array of 96 Processor Elements
- 250 MHz, 48 GFlops (Double FP c.f. Grape SFP)
- IBM 0.13 μ m FSG process, 8-layer metal (copper)
- 47% logic, 53% memory
 - More logic than most processors!
 - About 50% of the logic is FPUs
 - Hence around one quarter of the chip is floating point hardware
- 15 mm x 15 mm die size
- 128 million transistors
- Approx. 10 Watts

CSX600 processor core



Multi-Threaded Array Processing

- Programmed in high-level languages
- Hardware multi-threading for latency tolerance
- Asynchronous, overlapped I/O
- Run-time extensible instruction set
- Bi-endian (compatible with host CPU)

Array of 96 Processor Elements (PEs)

- **Each is a Very Long Instruction Word (VLIW) core, not just an ALU**
- **Flexible data parallel processing**
- Built-in PE fault tolerance, resiliency

High performance, low power dissipation

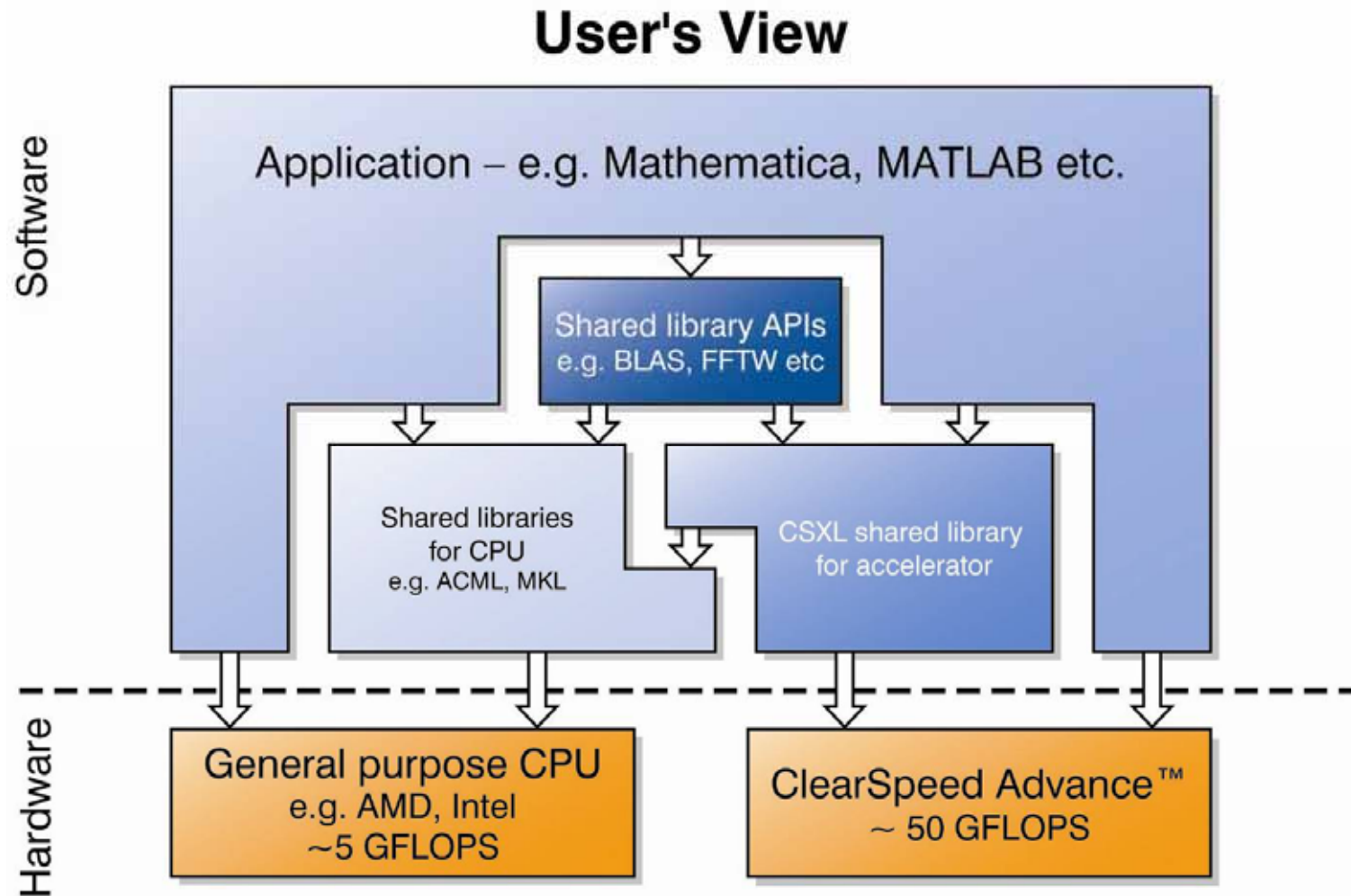
ClearSpeed Mode-of Use

- 1. User Application Acceleration
 - Matlab, Mathematica, Amber, ...
 - Transparent to users
- 2. Acceleration of Standard Libraries
 - BLAS/DGEMM, IMSL, FFTW...
 - Transparent to users (Fortran/C bindings)
- 3. User Applications
 - Arbitrary User Applications
 - Need MPI-like programming with C-dialect

BLAS/LAPACK/FFTW uses

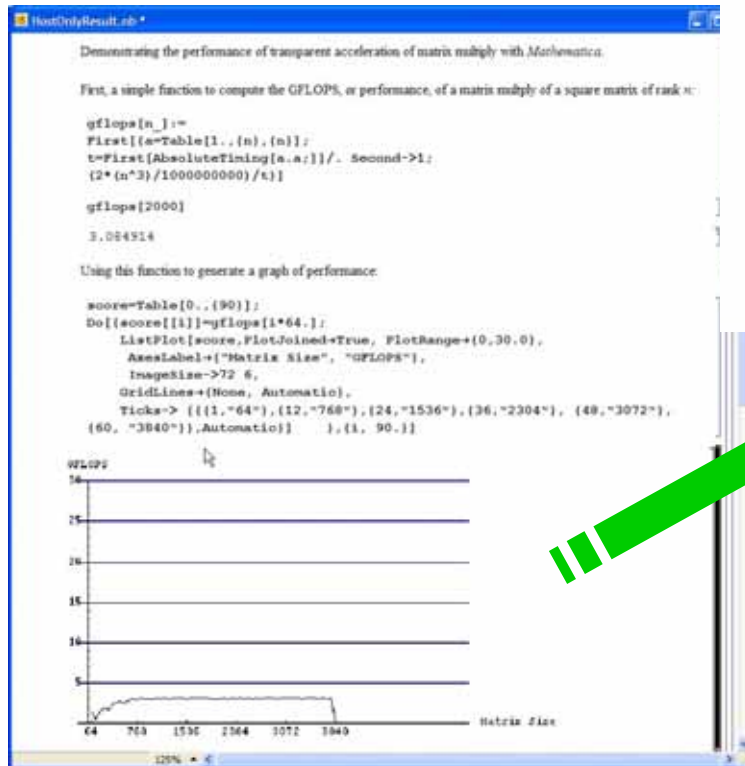
- Software known to use BLAS, LAPACK, FFTW...
 - MATLAB, Mathematica, Maple, Octave, ...
 - LINPACK, HPCC
 - IMSL, BCSLIB-EXT, SuperLU, NAG
 - FEA, CFD, Finance codes
 - ABAQUS, ANSYS, MSC (Nastran, Marc, ADAMS), ...
 - LS-DYNA parallel implicit (uses BCSLIB-EXT)
 - CPMD, Molpro, NWChem, GAMESS, Gaussian, ...
 - Some silicon design (EDA) tools
 - Numerous Oil & Gas in-house codes
 - Many, many more!
- ClearSpeed has a profiler for analysing an application's use of standard libraries (ClearTrace)

Application acceleration structure

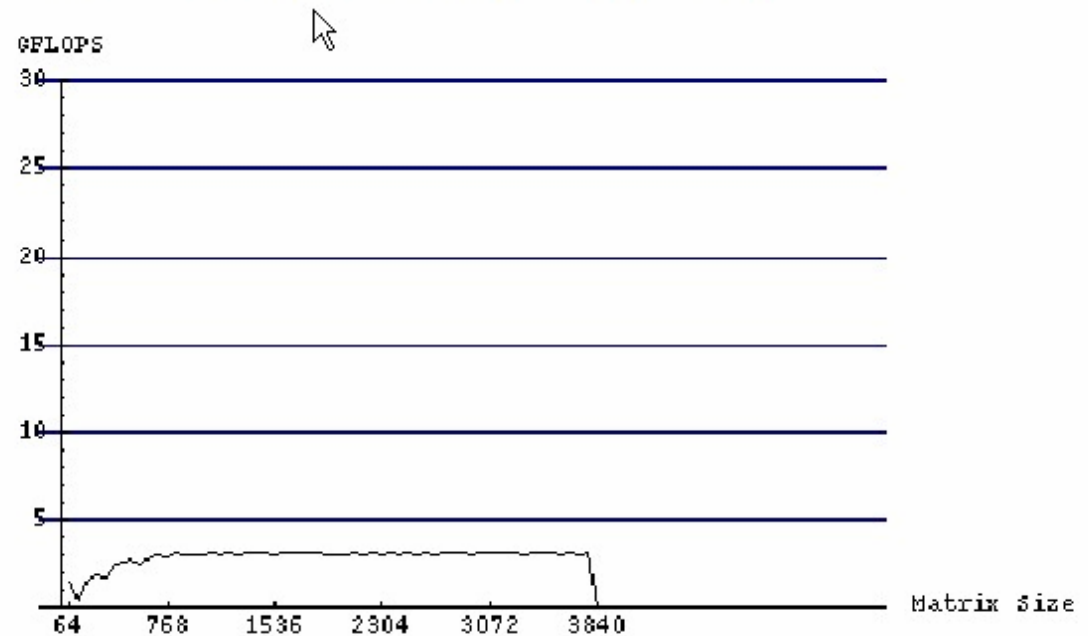


Mathematica Performance Data

Mathematica Notebook
Data for system with NO
CSX600 Advance card



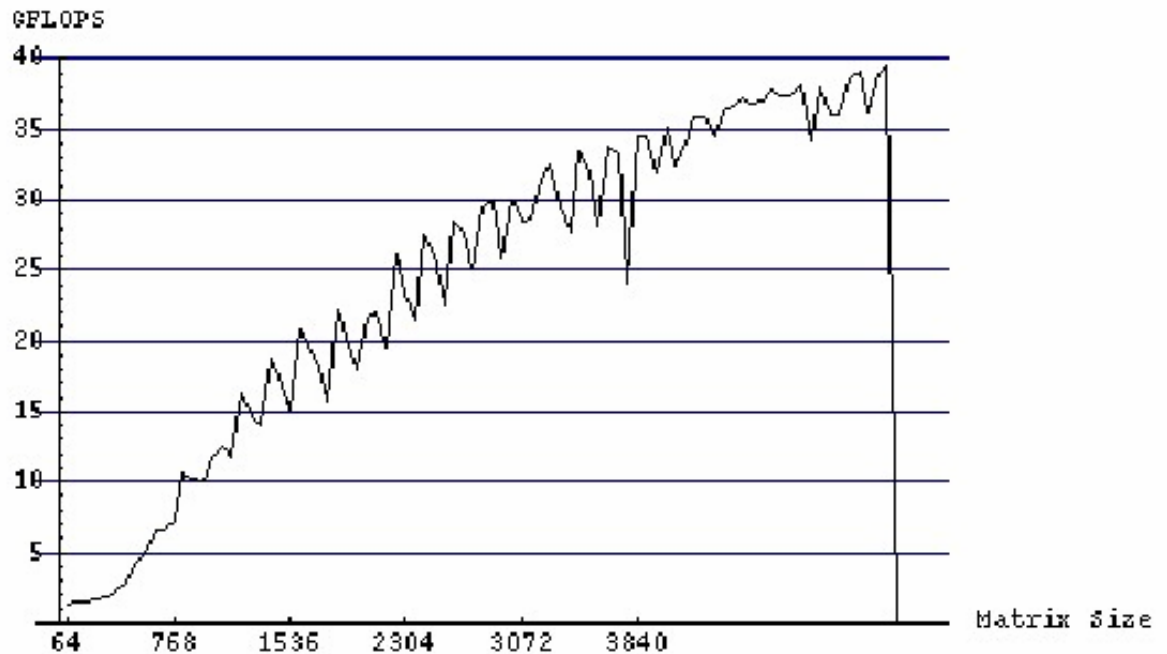
W



Maximum ~ 3 GFLOPS DGEMM

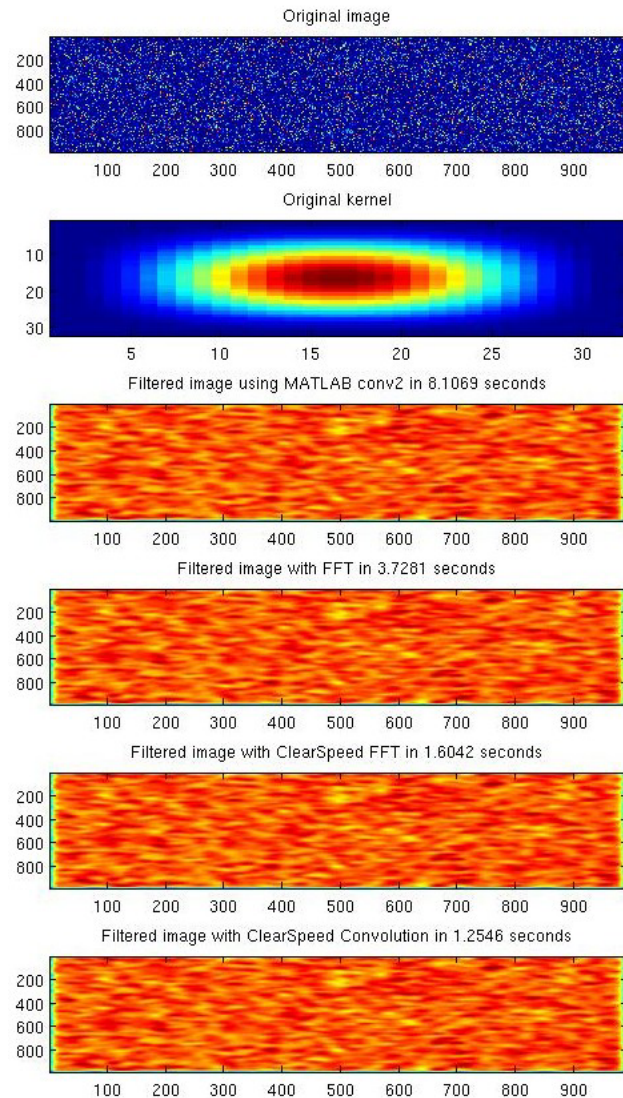
Mathematica Performance Data

Mathematica Notebook
Data for system with
CSX600 Advance card



Maximum ~ 40 GFLOPS DGEMM

ClearSpeed MATLAB FFT/BLAS acceleration



Plug-and-play MATLAB acceleration via standard libraries (BLAS, FFTW)

- 1,024x1,024 double precision convolution
- Original time on 3.2 GHz Xeon:
 - 8.1 seconds
- Time with ClearSpeed FFTW acceleration:
 - 1.6 seconds
- Time with ClearSpeed convolution acceleration:
 - **1.2 seconds**
- **6X acceleration!**
- Performance does not deteriorate with larger FFTs as conventional cache-optimized processors.
 - 3.8GFlops sustained per chip ~= SX-5
 - At only 10 Watts!

東工大 GSICのキャンパスグリッド Titech (Campus) Grid (2002-)

Titech Gridは、キャンパス内に次世代E-Scienceアプリのためのグリッドテストベッドを構築する、世界にも類を見ない大規模実験(平成13年度末補正予算で開始)

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(仮想分散高性能計算環境)を構築、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アプリの実行環境・実際の実現

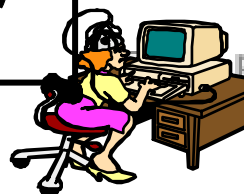
設置場所は2キャンパスにまたがり15箇所

専攻設置
サテライトPC群
(24プロセッサ)

× キャンパス内10台以上
(大岡山、すすかけ台)

GSICメインPC群
- 256プロセッサ × 2

ユーザはキャンパス内から仮想的に単一の計算機としてアクセス・利用が可能



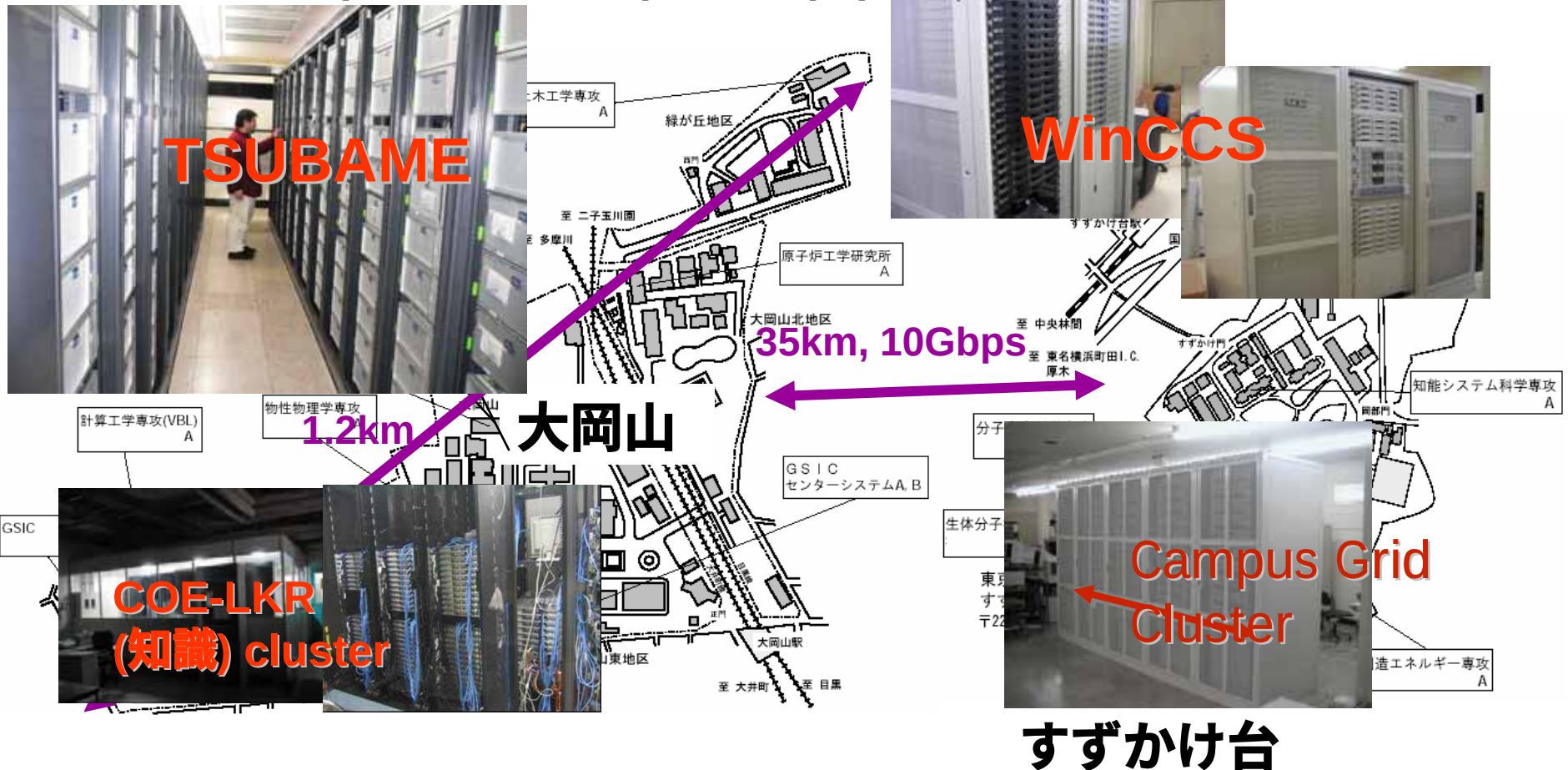
Super
TITANET
(1-4Gbps)

Super SINET
(マルチGigabits)
学外グリッドとの接続

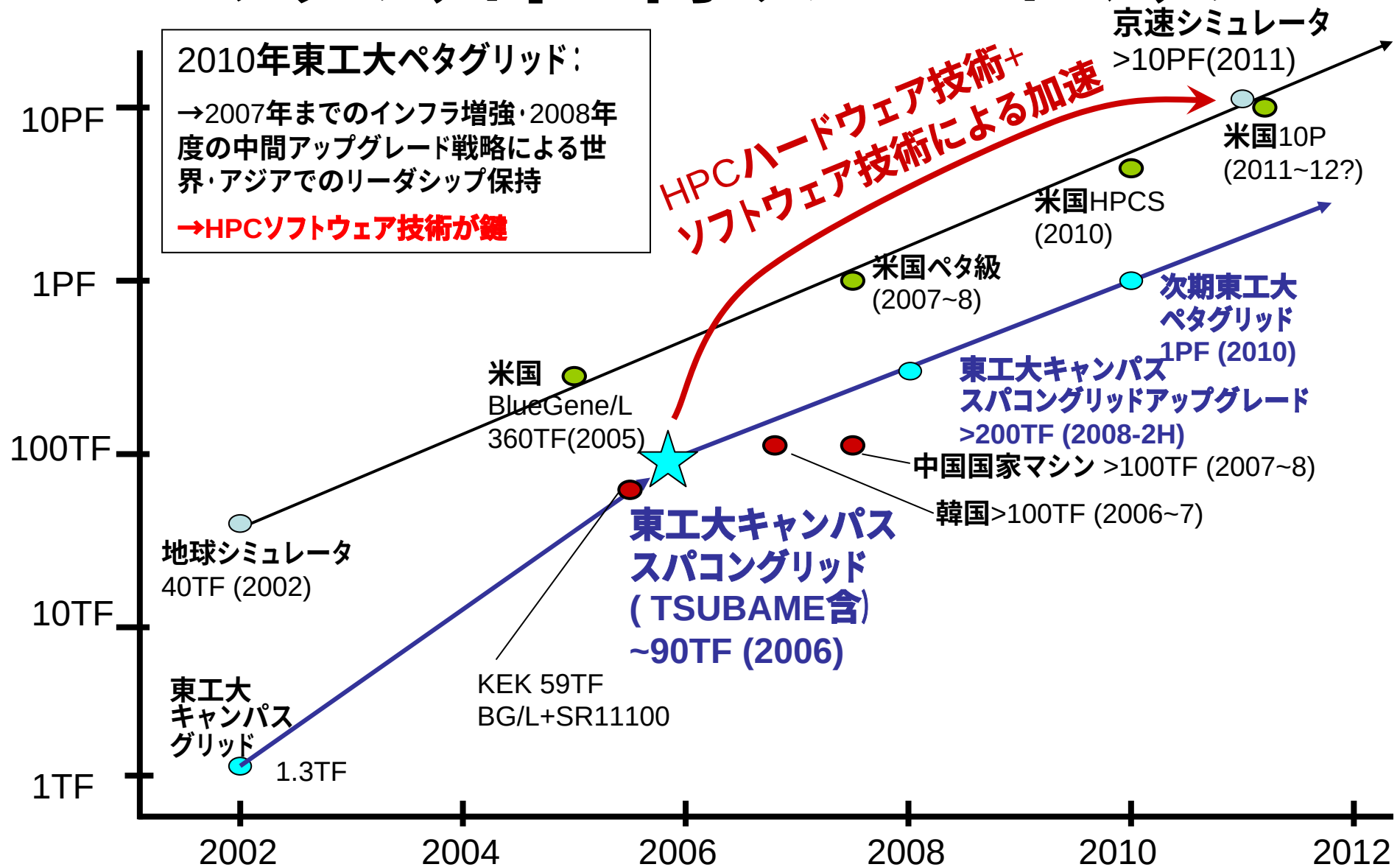
800プロセッサ以上、1.3TeraFlops、25TeraByteストレージの学内グリッドの構築

Titech Supercomputing Grid 2006

- **~13,000 CPUs, 90 TeraFlops, ~26 TeraBytes Mem, ~1.1 Petabytes Disk**
- **CPU Cores: x86: TSUBAME (~10600), Campus Grid Cluster (~1000), COE-LKR cluster (~260), WinCCS (~300) + ClearSpeed CSX600 (720 Chips)**



ペタスケール・インフォーマティクスによる 「ペタグリッド」に向けたロードマップ



2007-8年のペタフロップスマシン構築

	SX-8/9	64bitRISC (Power6)	64bit PC Cluster (Intel Tigerton, AMD K8L/HT2)	IBM BlueGene/P
GFLOPS/CPU	22	20	12	4
CPU CORE/Chip	1	2	4	4
CPU Chips/Cabinet	8	80	144	1000
TFLOPS/Cabinet	0.176	3.2	6.912	16
Memory BW/Chip (GB/s)	64	51.2	25.6	12.8
Memory BW/Cabinet (GB/s)	256	4096	3686.4	12800
Memory Bytes/s/Flop	1.4545455	1.28	0.533333333	0.8
Network BW/Chip(GB/s)	NA	4	2	5
Network BW/Cabinet(GB/s)	32	320	288	5000
Network Bytes/s/Flop	0.1818182	0.1	0.041666667	0.3125
#Cabinets for 1PetaFlops	5681	312	144	62
Physical size relative to ES	7.40	0.41	0.19	0.08
Power/Cabinet (KW)	9	30	20	25
Total Power (30% cooling) (MW)	66.47	12.17	3.74	2.02
Power relative to ES (11MW)	6.04	1.11	0.34	0.18
Cost/Cabinet (\$Million US)	1	1	0.6	1.5
Total Cost (\$Billion US)	5.68	0.31	0.09	0.09
Cost relative to ES (\$400 mil)	14.20	0.78	0.22	0.23

この進化の速度では、2011-12年には x86で 10PFのマシンが米国は十分構築可能
しかも、莫大なアプリ・ソフトベースにて⇒我が国がそれに勝つには？

ペタスケール・インフォマティックスへ

