

# GSIC

Global Scientific Information  
and Computing Center

# TSUBAME

## Grid Cluster

**Tokyo-tech Supercomputer and UBiquitously Accessible  
Mass-storage Environment**



東京工業大学

Tokyo Institute of Technology



## TSUBAMEご紹介

2006年4月に東京工業大学(東工大)が導入したスーパーコンピューティング・グリッドは、Linpackベンチマークで38.18テラフロップスのパフォーマンスを達成し、2006年6月時点で、世界のスーパーコンピュータのトップ500ランキングにおいて、米国外では世界最高速の計算能力を有するスーパーコンピュータとなりました。本学のシンボルにちなんでTSUBAMEと命名されたスーパーコンピュータは、日本電気(株)(以下「NEC」)のシステムインテグレーション技術により、AMD Opteron™ プロセッサコアを10,480個搭載したSun Fire™ X4600サーバ、サン・マイクロシステムズ社とNECのストレージ技術、Voltaire社のInfinibandネットワーク、ClusterFS社のLustre並列ファイルシステム、およびClearSpeed社のSIMDアクセラレータボードを組み合わせて構築されました。

TSUBAMEの理論演算性能は85テラフロップスとなり、21テラバイト以上のメモリと、1.1ペタバイトのハードディスクストレージを有し、米国以外のすべてのシステムに勝る性能を誇ります。本システムの莫大な規模は、単に少数の限られたグランド・チャレン

ジアプリケーションへの利用を想定するに止まりません。豊富な計算資源を、大学内での利用のみならず、スーパーコンピュータをPCのように使いこなす未来の計算科学者の育成、異なる組織に所属する研究者間での産学・国際連携の促進といったスーパーコンピュータの新たな利用モデルの確立まで視野に入れています。本システムの導入により、本学がアジアにおけるスーパーコンピューティング技術・グリッド技術の盟主として認知され、リーダーシップを発揮することができます。さらに文部科学省にて国家基幹プロジェクトとして検討が進められている10ペタフロップス超級スーパーコンピュータの実現に向け、ソフトウェア、アプリケーション、アーキテクチャ等の研究面でも、積極的に貢献していく所存です。



## TSUBAME Introduction

In April 2006, the Tokyo Institute of Technology (Tokyo Tech) has deployed a "Supercomputing Grid" centered around a supercomputing cluster that has become the highest performance supercomputer outside the United States, according to the Top500, by achieving 38.18 Teraflops in the Linpack Benchmark, as of June, 2006. Tokyo Tech's supercomputer, called TSUBAME after the University's symbol bird swallow, is integrated by NEC, featuring a fleet of Sun Fire™ X4600 servers with 10,480 AMD Opteron™ processor cores, Sun and NEC storage technologies, Voltaire's Infiniband network, ClusterFS's Lustre parallel file system software, as well as ClearSpeed SIMD acceleration boards. Overall, the peak speed of TSUBAME is 85 Teraflops, and facilitates over 21 Terabytes of memory, as well as 1.1 Petabytes of

hard disk storage, again besting all other machines outside of the United States.

The immense size of the machine is not merely intended to serve a few, grand-challenge applications. In fact, the expectation is to use the abundant resources to establish a new usage model of super-

computers that will not only be internal to an institution, but rather, will allow incubation of future computational scientists using supercomputers as if they are PCs, as well as fostering of international / industry-academia collaborations among the scientists in different organizations.

As such, TokyoTech hopes to establish recognition as the leading site in the Asia-Pacific region with respect to supercomputing and grid technologies with the new supercomputing grid, paving the way for active contributions in fundamental software, application and architectural research towards the 10 Petaflops-class supercomputing project undergoing planning as the core national infrastructural project in Japan by the Ministry of Education, Culture, Sports, Science and Technology (MEXT).




左記の紹介内容は、2006年6月初旬のパンフレット作成時の以下の公開データで判明した範囲で記載してあります。その後、6月28日に正式な第27回のThe Top500ランキングの発表があり、その中でTSUBAMEはLinpack性能(RMax)では全体で7位、米国外では2位、ピーク性能(RPeak)では米国外で1位となっています。

- 第26回 The Top500 リスト (2005年11月版) <http://www.top500.org/>
- Performance of Various Computers Using Standard Linear Equations Software, (Linpack Benchmark Report), Jack J. Dongarra, University of Tennessee Computer Science Technical Report, CS-89-85, June 17 2006. <http://www.netlib.org/benchmark/performance.ps>

The statement regarding the TSUBAME performances and its relative ranking is based on the following public data available to us as of mid-June, 2006 when this pamphlet was created. On June 28th, 2006 the official 27th Top500 ranking was announced, where TSUBAME was ranked 7th overall and 2nd for a machine outside the US in Linpack performance (Rmax), and 1st in peak performance (Rpeak) for a machine outside the US.

- The 26th Top500 List (November 2005) <http://www.top500.org/>
- Performance of Various Computers Using Standard Linear Equations Software, (Linpack Benchmark Report), Jack J. Dongarra, University of Tennessee Computer Science Technical Report, CS-89-85, June 17 2006. <http://www.netlib.org/benchmark/performance.ps>





**東京工業大学 スーパーコンピュータ**  
**TSUBAME Grid Cluster : みんなのスパコン**  
**<披露式・祝賀会次第>**

国立大学法人 東京工業大学  
平成 18 年 7 月 3 日 (月)  
15 時 30 分~18 時 30 分

**披 露 式**

(百年記念館フェライト会議室) 15:30~16:30

開会挨拶 来賓祝辞	東京工業大学長 文部科学省 研究振興局長	相澤 益男 清水 潔
開発事業社祝辞	日本電気株式会社 執行役員常務 サン・マイクロシステムズ株式会社 代表取締役社長 日本AMD株式会社 代表取締役社長	塩路 洋一郎 末次 朝彦 デービッド・ユーゼ
「TSUBAME」命名者の表彰	命名者 大学院理工学研究科	学術国際情報センター長 酒井 善則 教授 原子核工学専攻 小川 慧
TSUBAME の概要説明	東工大 IC カードによる TSUBAME ポータルサイトへのアクセスデモ	学術国際情報センター 松岡 聡 教授 西川 武志 特任助教授
スーパーコンピュータを使った最新の研究事例		学術国際情報センター 青木 尊之 教授
TSUBAME 紹介ビデオ上映「みんなのスパコン TSUBAME Grid Cluster」		

**見 学 会**

(学術国際情報センター情報棟) 16:30~17:00

**祝 賀 会**

(百年記念館フェライト会議室) 17:15~18:30

来賓祝辞	蔵前工業会理事長 東北大学 情報シナジーセンター長	田中 實 川添 良幸
乾杯	東京工業大学 副学長 (研究担当)	下河邊 明
閉会挨拶	学術国際情報センター長	酒井 善則





## メッセージ message



国立大学法人東京工業大学

学長 相澤益男

国内最高速の計算能力を持つスーパーコンピューティング・グリッドシステム(TSUBAME)を披露するにあたり、一言ご挨拶申し上げます。

東京工業大学は、120余年の輝かしい伝統と歴史を継承しつつ、21世紀の科学技術をリードする“世界最高の理工系総合大学”へと進化を続けています。国内だけではなく国際的にもさらに高い水準に向かって進化を果たすことが私達の目指すところでもあります。

本年4月3日より稼働した「TSUBAME」は、世界のスーパーコンピュータのランキング表であるTop500でもアジアでは最高性能のマシンとして上位にランクインされ、高い処理能力を発揮することが期待されています。

この「TSUBAME」を積極的に教育・研究に活用し、新分野の創出や世界の学術研究の進展に貢献していくことを祈念しています。

I would like to proudly announce that Tokyo Institute of Technology (Tokyo Tech) has deployed TSUBAME, the highest performance supercomputer grid system in Japan.

Following the footsteps of 120 years of brilliant history and tradition Tokyo Tech continues to evolve as one of the world's leading science and technology universities of the 21st century.

Our mission is to establish increased recognition as the world leading institute and to evolve ourselves onto achieve even higher levels of excellence, both domestically and internationally.

TSUBAME is expected to become the highest-performing supercomputer within the Asia-Pacific area, and will be ranked among the top 10 computers in the world in the June 2006 Top500 list, demonstrating its utmost ability and performance.

We hope to utilize the outstanding computing power of TSUBAME for research and education activities, as well as actively contributing to open up new academic areas and to promote research activities.

Professor/Dr. Masuo AIZAWA

President

Tokyo Institute of Technology



メッセージ  
message学術国際情報センター  
センター長 酒井善則

本日、スーパーコンピューティング・グリッドシステム (TSUBAME) 披露式を挙げるにあたり、一言ご挨拶申し上げます。

平成18年4月3日よりスーパーコンピューティング・グリッドシステム (TSUBAME) が稼働開始を致しました。このシステムは、Linpackベンチマークで38.18テラフロップスのパフォーマンスを達成し、平成18年6月現在、スーパーコンピュータのトップ500ランキングにおいてアジアで世界最高速となっております。さらに、理論演算性能で85テラフロップス、総合メモリ容量21.4テラバイト、ディスク総合容量1.1ペタバイトと、スパコンの他の重要な指標でも我が国No.1となります。

このシステムは「みんなのスパコン」として、学部生を含む全学への公開のみならず、産学連携を含む様々な外部との共同研究のために活用されます。また、我が国の将来のペタフロップス級のスーパーコンピュータ構築にも様々な技術的貢献をしていく所存です。

また、この場をお借りしまして「TSUBAME」の仕様策定及び、調達等さまざまな御協力に感謝し、お礼の言葉を申し上げて挨拶に代えさせていただきます。

Today, I am pleased to have an opportunity to celebrate the development and deployment of "TSUBAME," our new "Supercomputing Grid" system.

TSUBAME has been deployed since April 3<sup>rd</sup> 2006 and become the highest performance supercomputer within the Asia-Pacific area, according to the Top500, by achieving 38.18 Teraflops in the Linpack Benchmark, as of June, 2006.

Overall, the peak speed of TSUBAME is 85 Teraflops, and facilitates 21.4 Terabytes of memory, as well as 1.1 Petabytes of hard disk storage, again besting all other machines outside of the United States.

As such, TSUBAME will be available for public use by all constituents of our institutions, including all the undergraduate students, as "everybody's supercomputer." It will also be extensively used for external collaborations with outside institutions, both academic and industry. Moreover, we hope to actively contribute to various software, application and architectural research towards the 10 Petaflops-class national supercomputing project.

Finally I would like to express my highest gratitude to all those who helped in the development and the deployment of TSUBAME.

Professor/Dr. Yoshinori SAKAI  
Director  
Global Scientific Information and Computing Center



## センターの特長

東京工業大学では21世紀にふさわしい新しい情報センターとして、2001年4月に学術国際情報センターが発足いたしました。

このセンターは、従来の総合情報処理センターと理工学国際交流センターを統合して設立されたものです。本センターは、最先端の情報技術を駆使して研究・教育の支援を行い、またその成果を国内外の研究機関、教育機関等に発信して交流・連携を深め、研究・教育の活性化、国際交流の発展に寄与することが期待されています。

最近の情報技術の発展は極めて急速であり、それに伴い研究・教育も、ますます高度化・複雑化してきており、これらに対応したサービスも多彩となるなど、センターとして常に新しい情報技術を

とり入れた研究・教育支援を心がけることが重要です。

本センターでは、常に最先端の情報技術レベルを維持して利用者からの要望に応えるため、情報系2部門（情報基盤、研究・教育基盤）に14名の専任教員・客員教員等を配置し、効率的に研究・開発を進める体制としております。

一方、学術国際交流部門においては情報基盤を活用した国際交流および大規模な国際共同研究を行うことにあります。そのため、この部門では6名の専任教員・客員教員等が情報メディアを駆使した国際交流を実施するために情報系教員と密接に協力し、早期に優れた成果を上げ世界のリーダーシップをとることを目指します。

## 沿革

### 1971年 情報処理センター設置

計算機システムHITAC 8700導入

### 1976年 総合情報処理センター設置

計算機システムをHITAC M-180に更新

- 1977年 一般的情報処理教育を開始 (HITAC M-180)
- 1988年 スーパーコンピュータCDC ETA10 (買取) 導入
- 1994年 キャンパス情報ネットワーク (Titanet) の運用開始 (Titanet 運用センター設置)
- 1995年 スーパーコンピュータを CRAY C916/12256 (レンタル) に更新
- 1997年 Titanet 運用センターを統合
- 1998年 一年次からの情報教育開始 (SGI Origin2000)
- 2000年 スーパーコンピュータをSX-5、Origin2000に更新
- 2001年 SuperTitanet導入  
研究用計算機システムをCOMPAQ GS320に更新

### 1979年 理工学国際交流センター設置

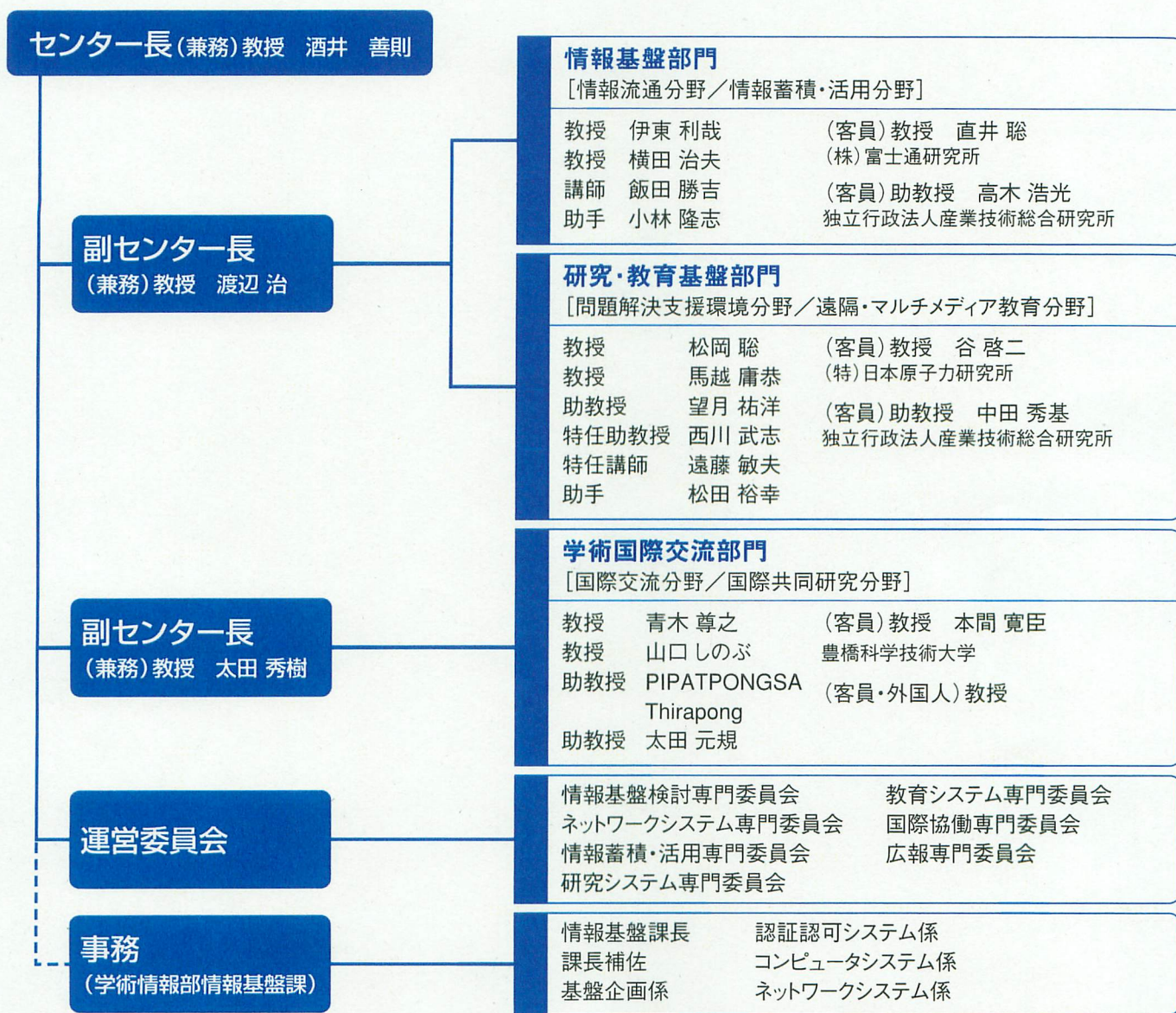
- 1980年 インドネシア大学との拠点大学交流事業の開始
- 1984年 インドネシア科学院との拠点大学交流の開始
- 1984年 バンドン工科大学との拠点大学交流の開始
- 1986年 フィリピン大学との拠点大学交流の開始
- タイ・キングモンクット工科大学との拠点大学交流の開始
- 1991年 JSPS、拠点大学交流をプロジェクト方式に変更

### 2001年 学術国際情報センターに改組

- 2002年 グリッドPC NEC Express5800導入
- 2004年 光ファイバによる全キャンパス間ギガビット接続
- 2005年 キャンパス無線ネットワーク運用開始
- 2006年 TSUBAME Grid Cluster運用開始  
キャンパス共通認証・認可システム (PKI認証) 導入



## 組織



## 教職員

	教授	助教授	講師	助手	事務職員	合計
センター長	(1)					(1)
副センター長	(2)					(2)
情報基盤部門	2 ①	①	1	1		4 ②
研究・教育基盤部門	2 ①	1 ①<1>	<1>	1		4 ②<2>
学術国際交流部門	2 ②	2				4 ②
事務(学術情報部情報基盤課)					13<2> [9]	13 ② [9]
合計	6(3)④	3②<1>	1<1>	2	13<2> [9]	25(3)⑥<2> ② [9]

※ ( )は兼務、○は客員、&lt; &gt;は特任、&lt; &gt;は技術職員、[ ]は非常勤でいずれも外数。



## 研究開発テーマ

学術国際情報センターの使命は、最先端の情報技術を駆使して研究・教育の支援を行い、その成果を国内外の研究機関、教

育機関等に発信して交流・連携を深め、研究・教育の活性化、国際交流の発展に寄与することです。

### 学術国際交流

#### 国際交流分野

- 戦略的国際協力プロジェクトの企画・立案・運営

#### 国際共同研究分野

- 大規模計算力学シミュレーション
- バイオインフォマティクス
- 世界文化遺産地域開発におけるITアプリケーション

### 情報基盤

#### 情報流通分野

- 情報ネットワークの高信頼技術
- 情報ネットワークの高機能技術
- 情報ネットワークの高品質技術

#### 情報蓄積・活用分野

- マルチメディア蓄積技術
- マルチメディア高速検索技術
- 大規模情報蓄積システム構成
- 蓄積コンテンツ統合技術

### 研究・教育基盤

#### 問題解決支援環境分野

- 高性能計算
- グリッド(超広域高性能)計算
- コモディティクラスタ計算
- 高信頼計算
- 超並列計算

#### 遠隔・マルチメディア教育分野

- 教育の情報化
- 教育媒体のマルチメディア化
- 簡便なオーサリングツールの提供
- 授業コンテンツのWeb搭載支援
- ITを活用した遠隔教育の推進

## 研究開発



Main Compute nodes room



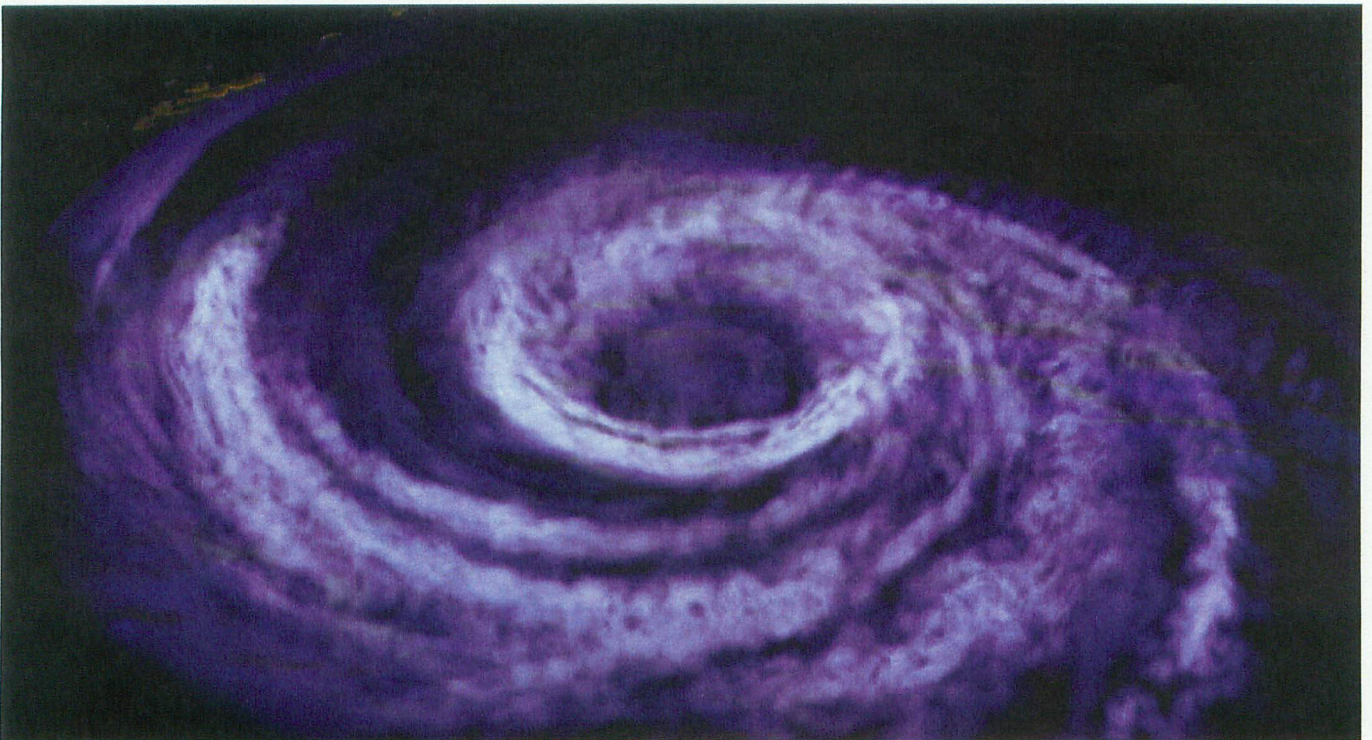
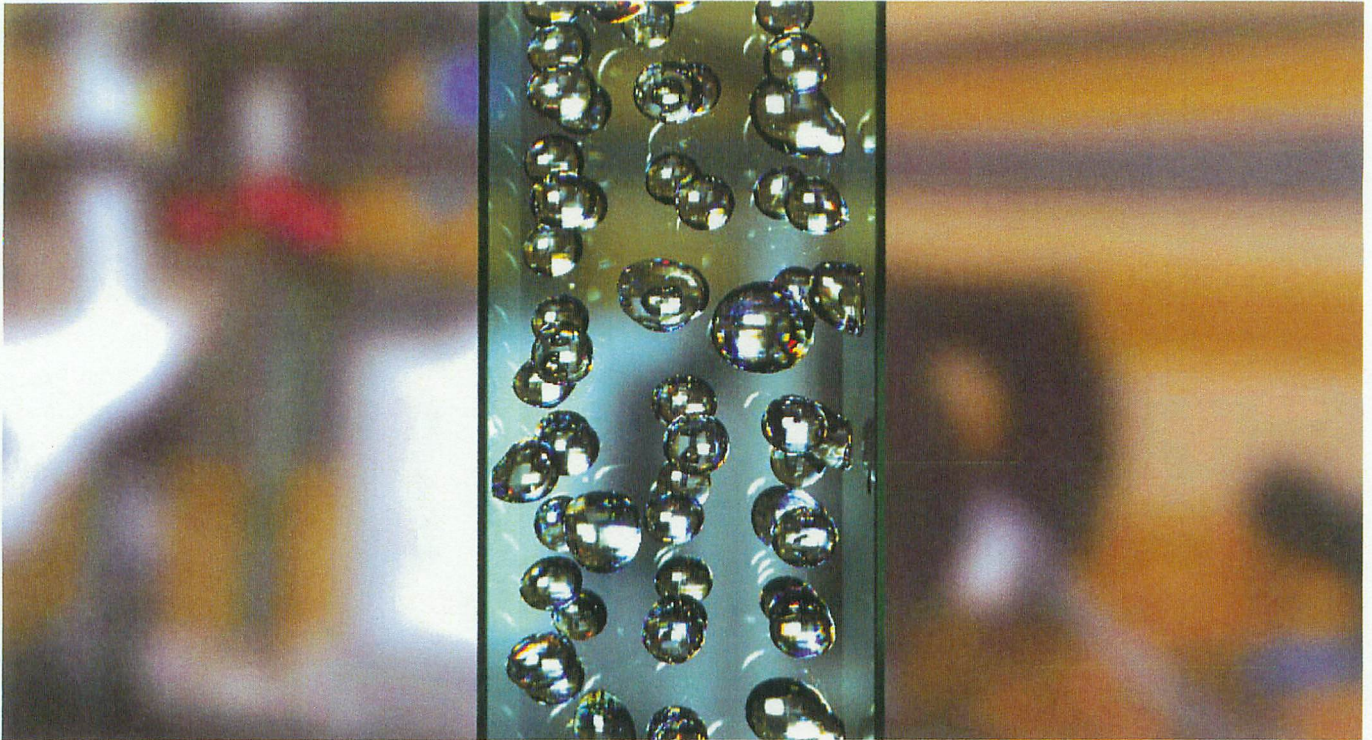
1.1PB storage







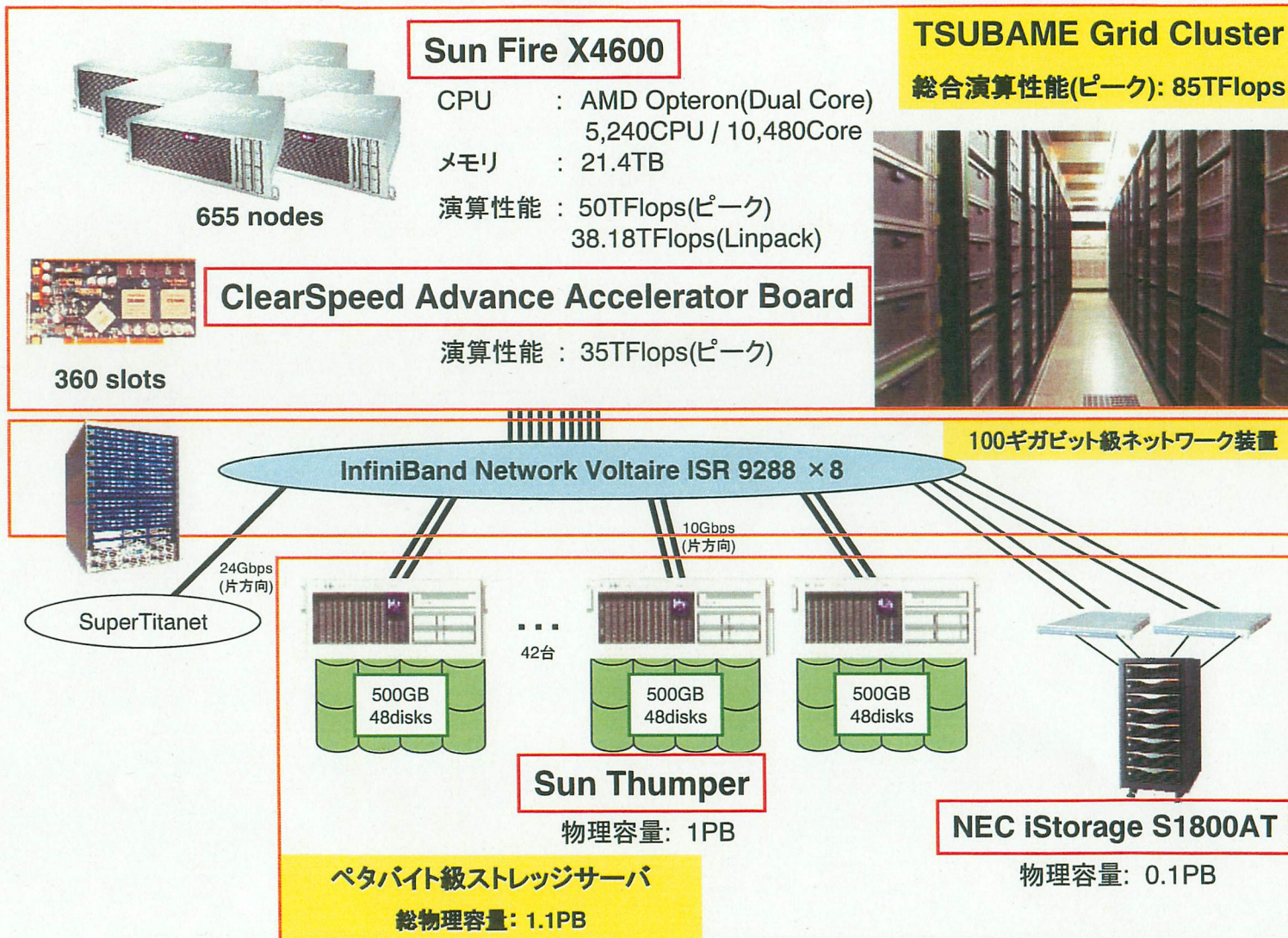
上昇する気泡流の直接シミュレーション  
Direct two-phase flow simulation of rising bubbles



平成16年の台風18号に対するメソスケール高解像度シミュレーション  
Meso-scale cloud-resolved simulation for 2004 #18 Typhoon



## 東京工業大学 学術国際情報センター TSUBAME Grid Cluster



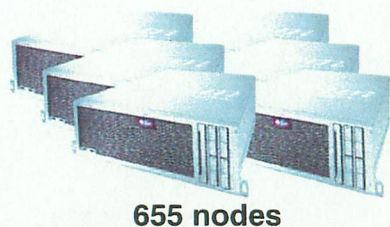


# TSUBAME Grid Cluster

Tokyo-tech Supercomputer and Ubiquitously Accessible Mass-storage Environment

Tokyo Institute of Technology

## Global Scientific Information and Computing Center (GSIC) TSUBAME Grid Cluster Tokyo Institute of Technology



655 nodes

### Sun Fire X4600

CPU : AMD Opteron (Dual Core)  
5,240CPU / 10,480Core  
Memory : 21.4TB  
Performance : 50TFlops (Peak)  
38.18TFlops (Linpack)



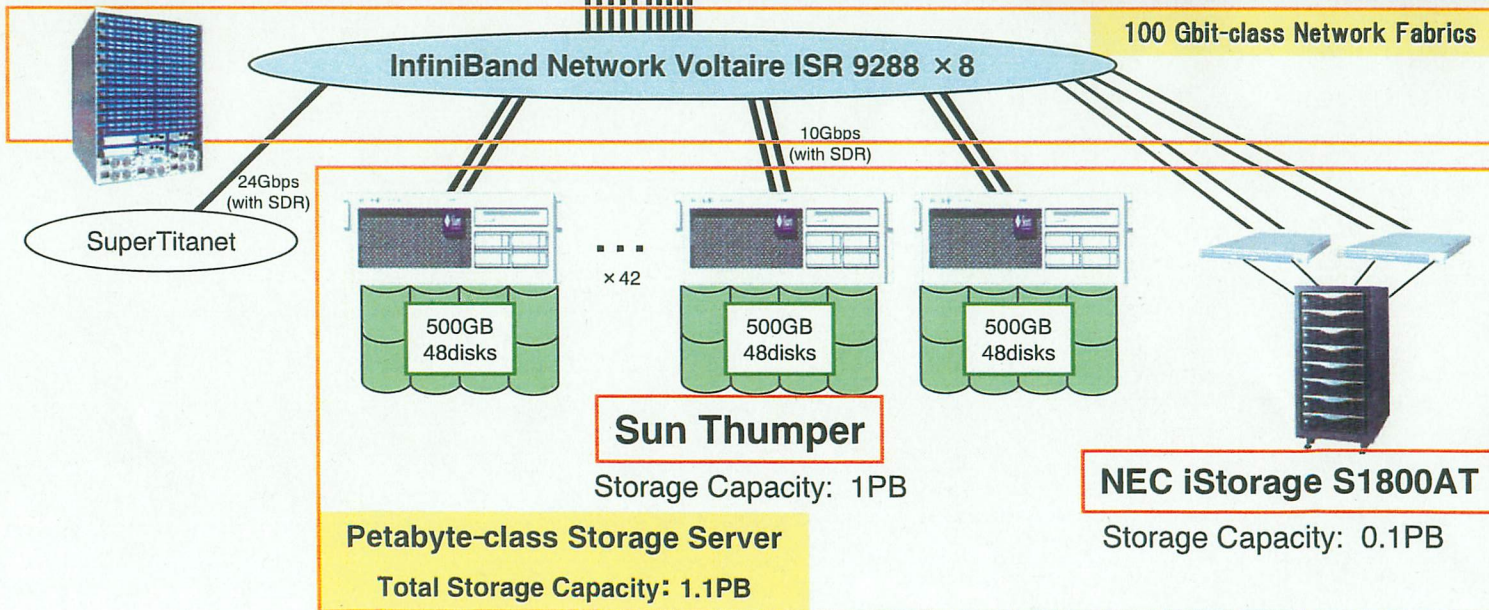
360 slots

### ClearSpeed Advance Accelerator Board

Performance : 35TFlops (Peak)

### TSUBAME Grid Cluster

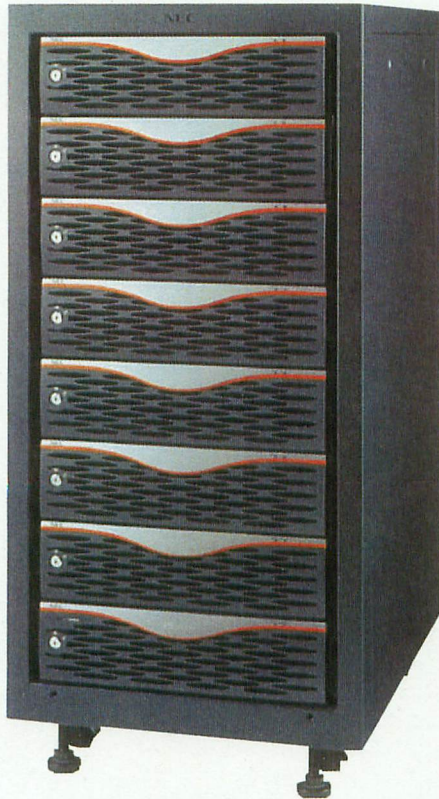
Total Peak Performance: 85TFlops





## NEC

### *iStorage S1800AT*



#### 構成

筐体:1コントローラ+16ディスクエンクロージャ  
(1エンクロージャあたり3U/15HDD)

メモリ:16GB

HDD:400GB SATA ディスクドライブ× 240台

物理ディスク容量:96テラバイト

RAID:RAID6 (2HDD故障でも業務継続)

#### 機器仕様

最大HDD:SATA400GB (7200rpm) × 240HDD搭載

標準:8ホストポート (FC 2Gbps/1Gbps)

#### Constitution

Housing : 1controller+16Disk enclosure  
(1 enclosure area 3U/15HDD)

Memory : 16GB

HDD : 400GB SATA disk drive× 240stand

Physics disk space : 96Terabyte

RAID : RAID6

(Even if 2HDD breaks down,  
duties continuation is possible.)

#### Machinery specifications

HDD (maximum) : SATA400GB (7200rpm) × 240HDD

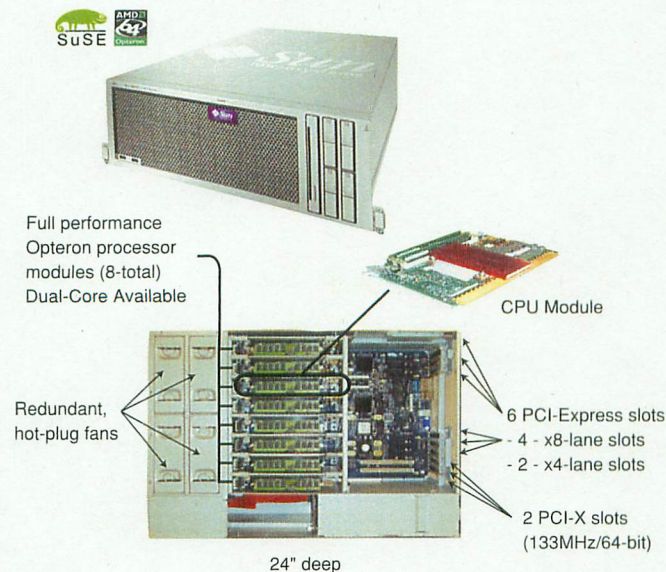
Standard : 8host port (FC 2Gbps/1Gbps)





## Sun Fire X4600

Enterprise-Class Data Center Compute Engine



大規模分散並列型演算サーバ  
高性能密結合分散並列型演算サーバ  
(合計：639台)

CPU：8ソケット デュアルコア800 シリーズAMD Opteron 2.4GHz  
メモリ：32GB メモリ (32x DIMM スロット)  
OS：SuSE Enterprise Linux 9 SP3

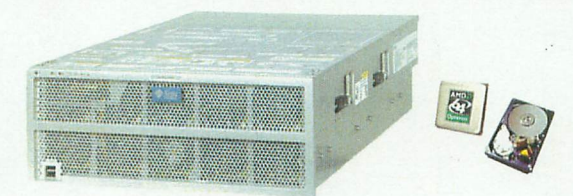
メモリ共有型演算サーバ群  
(合計：16台)

CPU：8ソケット800 シリーズAMD Opteron 2.6GHz  
メモリ：64GB メモリ (32x DIMM スロット)  
OS: SuSE Enterprise Linux 9 SP3

合算ピーク性能：50.4 Tera Flops  
合算メモリ容量：21.4 Tera Bytes

## Thumper

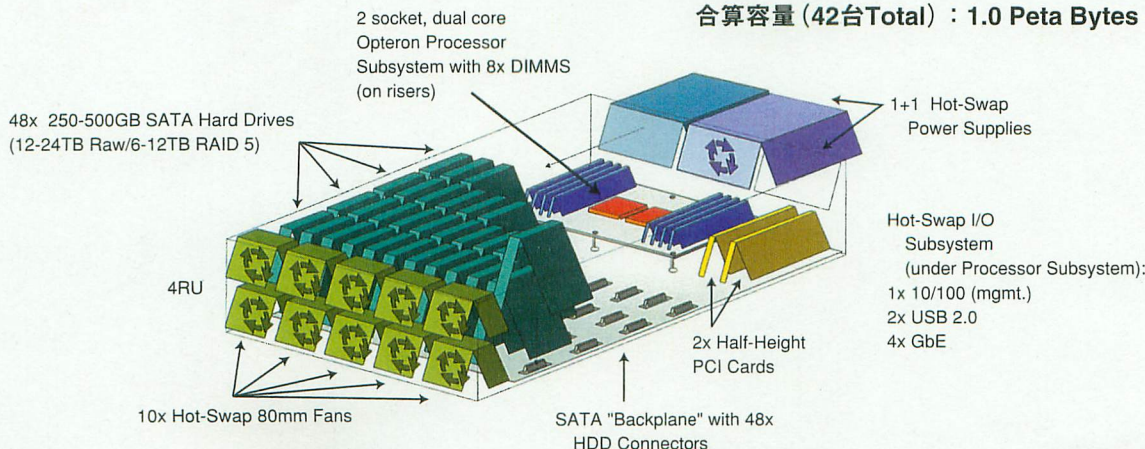
High performance Server with a lot of local data



高性能ストレージサーバ  
(合計：42台)

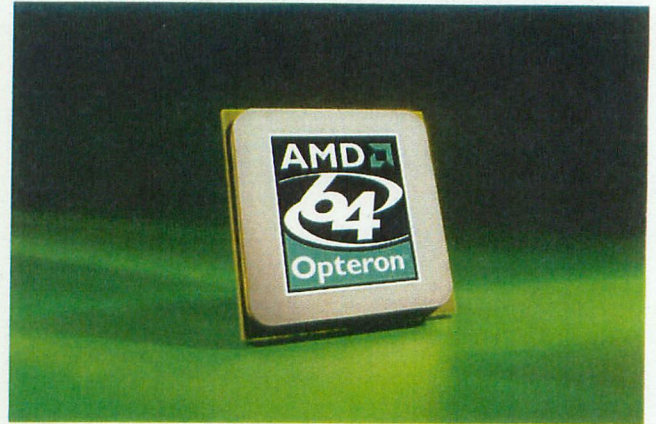
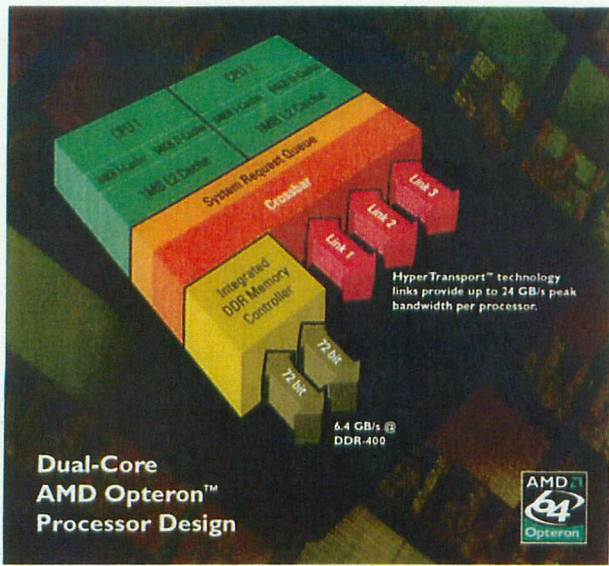
CPU：2ソケット デュアルコア AMD Opteron  
280プロセッサ 2.4GHz  
メモリ：8GB メモリ (8x DIMM スロット)  
OS：Redhat Enterprise Linux 4  
HDD：500GB SATA ディスクドライブx 48

合算容量 (42台Total)：1.0 Peta Bytes





# 64ビット・テクノロジーをリードする AMD Opteron™ プロセッサ



## AMD64は、当初よりデュアルコアを 想定した設計

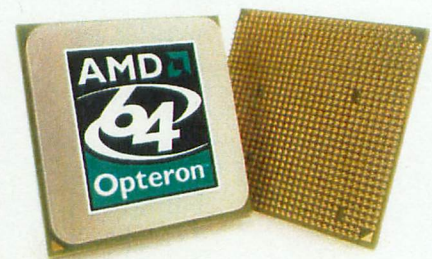
- ひとつのダイに2つのコアが搭載された真のデュアルコア プロセッサ
- コア間の通信はプロセッサスピード

## ダイレクトコネクト アーキテクチャ

- メモリコントローラは、プロセッサに内蔵
- ダイレクトコネクトアーキテクチャにより、クロスバースイッチを介してメモリコントローラ及び、HyperTransport™ technology linkと接続。

## 既存の940pinを踏襲して設計

- BIOS改版のみ
- TDPは変更無し



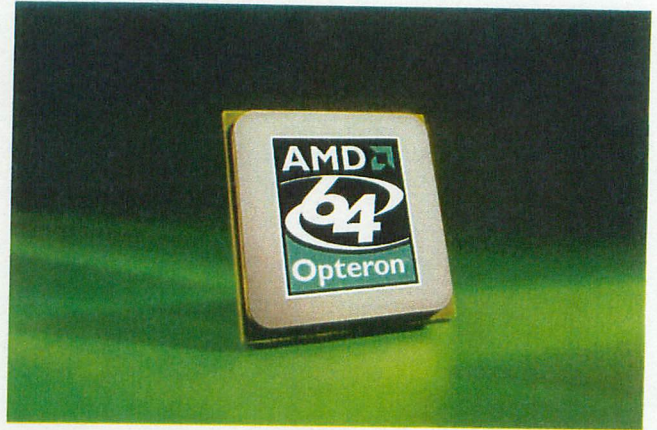
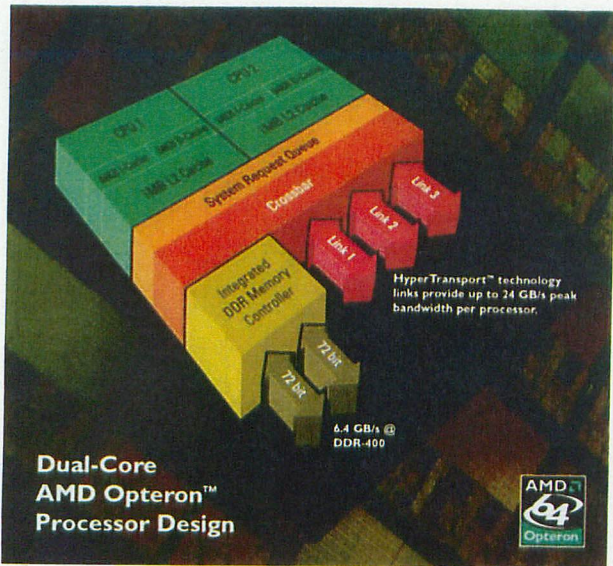
AMD

日本AMD株式会社

〒163-1334 東京都新宿区西新宿6-5-1, 新宿アイランドタワー 34F (私書箱1601)



# AMD Opteron™ Processor & 64-bit Technology



AMD64 designed from the ground up for multiple cores

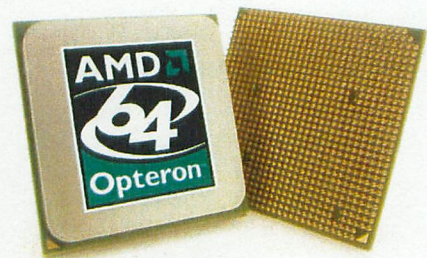
- True dual core, with two CPU cores on one die
- Inter-core communication at processor speed

Direct Connect Architecture

- Inter-core communication at processor speed
- Access via crossbar to memory controller and HyperTransport™ technology link

940-Pin Socket Compatible

- Requires only BIOS update
- No change in TDP



AMD

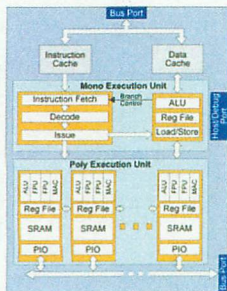
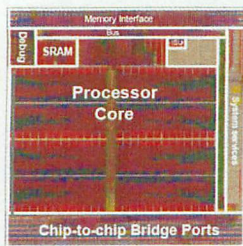
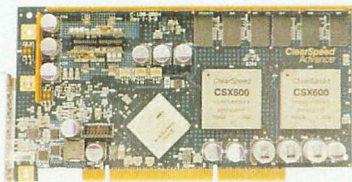
AMD Japan Ltd.

Shinjuku i-LAND Tower 34F, 6-5-1 Nishi Shinjuku, Shinjuku-ku, Tokyo163-1334 Japan



## ClearSpeed™

### Dual CSX600 PCI-X Board



#### ハードウェア

最大消費電力25W

CSX600 processor 2個実装 (合計理論性能96GFLOPS)

IEEE形式64bit浮動小数点型対応

133MHz PCI-X ホストインターフェイス

オンボードメモリ搭載量:最大 4GB

内部メモリ帯域幅: 200Gbytes/s

オンボードメモリ帯域幅: 6.4Gbyte/s

#### ソフトウェア

スタンダードソフトウェアライブラリの形で提供

標準ライブラリ未対応のアプリケーションソフトウェアでも

ClearSpeed's Software Development Kit (SDK) に

よってCSX600ボード対応にさせることが可能

#### アプリケーション

線形代数 - BLAS, LAPACK

バイオインフォマティクス - AMBER, CHARMM, GAUSSIAN

シグナルプロセッシング - FFT (1D, 2D, 3D), FIR, Wavelet

シミュレーション - CFD, FEA, N-body

イメージプロセッシング - filtering, image recognition, DCTs

オイル/ガス - Kirchhoff Time/Wave Migration





## Voltaire® Grid Director™ ISR 9288

強力なSAN/LANマルチプロトコル接続能力を  
持つ数十から数千ノードまでのHPC Clusters  
およびGridに対する一番理想的な選択

Ideal for HPC Clusters and Grids Ranging in  
Size From Tens to Thousands of Nodes With  
Powerful Multiprotocol Capabilities for  
SAN/LAN Connectivity

### Grid Director ISR 9288を紹介

業界で一番大きいInfiniBandスイッチソリューションGrid Director ISR 9288はHPCクラスターおよびグリッドに比べられないレベルのパフォーマンスと拡張性を提供する。

### 最高のパフォーマンス

Grid Director ISR 9288はその最大なスイッチング容量でInfiniBand業界のトップに立っている。そのFAT Tree (Clos) トポロジーはすべてのポートに全二重のバンド幅を提供する。Grid Director ISR 9288は288個のInfiniBand 4X (20 Gbps) ポート (若しくはSDR 10 Gbpsポート) 或いは96個のInfiniBand 12X (30 Gbps) ポートを提供する。ワイヤスピード、ノンブロッキングスイッチング容量は420ns以下の遅延と組み合わせてGrid Director ISR 9288に最高のパフォーマンスを与える。

### マルチプロトコル接続性

InfiniBandスイッチングのほかにGrid Director ISR 9288は強力なマルチプロトコル接続性をこの柔軟性に満ちた同じシステムで提供できる。同じ筐体でVoltaireのルータブレードを搭載することによりサーバクラスター、FC SAN、NAS、IP SANおよびTCP/IPネットワーク (LAN) の間のシームレスな接続性を提供できる。

### 大規模クラスターとグリッドへの低コストソリューション

ISR 9288を用いて大規模クラスターとグリッドを構築する場合最も低コスト効果が達成できる。その一体化するネットワークとストレージ接続性は少ないデバイスで高性能を実現しクラスターそのものの構築の簡単化及びシステムの高信頼性を実現できた。

### Introducing the Voltaire Grid Director ISR 9288

As the industry's largest InfiniBand switching solution, the Grid Director ISR 9288 provides unprecedented levels of performance and scalability for large high performance computing (HPC) clusters and grids.

### Unmatched Performance

The Grid Director ISR 9288 leads the InfiniBand industry with the highest switching capacity. It features a FAT Tree (Clos) topology that provides full bisectional bandwidth for each port. The Grid Director ISR 9288 supports up to 288 InfiniBand 4X (20 Gbps) ports (or 10 Gbps with SDR) or 96 InfiniBand 12X (30 Gbps) ports. Wire-speed non-blocking switch capacity combined with latency of less than 420 nanoseconds make the Grid Director ISR 9288 the highest performing switch available.

### Multi-protocol

Connectivity In addition to scalable InfiniBand switching, the Grid Director ISR 9288 offers powerful multi-protocol connectivity in a single, flexible system. The chassis hosts the Voltaire router blades, providing seamless connectivity between server clusters, FC SANs, NAS appliances, IP SANs and TCP/IP networks (LANs).

### Cost-Effective Solution for Large Clusters and Grids

Large clusters and grids are the most cost-effective when built using the largest available switch ISR 9288. Its integrated networking and storage connectivity provides high performance and requires fewer devices, thereby making the cluster simple to build and more reliable.





## Voltaire® Grid Director™ ISR 9288

### 豊富な管理

Grid Director ISR 9288は豊富かつ強力な管理能力をGridVision™ InfiniBandファブリック管理ソフトウェアで実現できている。この管理能力は外のすべての管理プラットフォームに独立し、スイッチ内部にあってCLI、GUI或いはSNMPマネージャなどによってアクセスできる。GridVisionはリアルタイムの下記のプロアクティブな管理機能を提供できる。総合的なファブリック及びリソースビュー、ファブリック及びスイッチの診断機能へのアクセス、すべてのレベルでのFail-over管理機能、InfiniBandファブリック及びそれに接続されたサーバ、ネットワーク及びストレージリソースの配置機能。

### 高可用性:High-Availability

Grid Director ISR 9288製品のすべての構成部品は最高の可用性を保つためHot Swap可能となっている。パワーサプライ及びファンはシステムの高可用性及びサービスの柔軟性を提供する。冗長管理ブレードは全システムの同期を維持しすべての故障を管理情報の欠損及びポート間のデータ伝送への影響なしにリカバーできる。

### Comprehensive Management

The Grid Director ISR 9288 provides comprehensive and powerful management capabilities through GridVision™ InfiniBand Fabric Management Software. The management capabilities are enabled in the switch independent of any external management platform and can be accessed via CLI, GUI or SNMP managers. GridVision delivers real-time proactive management by providing: aggregated fabric and resource views, access to a suite of fabric and switch diagnostics, the ability to manage fail-over on all levels, and provisioning of InfiniBand fabrics and the attached server, networking and storage resources.

### High-Availability

All of the Grid Director ISR 9288 product components are hot-swappable to allow for the highest availability. Power supplies, as well as fans provide system high availability and serviceability. Redundant management blades maintain synchronization so that a failure can be recovered without the loss of management information or any disruption in port-to-port communication.



### KEY FEATURES

- **Unparalleled scalability:** object-based storage architecture that scales to tens of thousands of clients and petabytes of data—a file system without limits.
- **Reliable:** the Lustre file system is now deployed in production on many clusters, large and small, meeting the uptime requirements of business and national security applications.
- **Proven performance:** dramatic increase in throughput and I/O by intelligent serialization and separation of metadata operations from data manipulation.
- **Open Source, open standards:** developed and maintained as Open Source software with an open networking protocol and POSIX file system semantics—ensuring broad support for industry-standard platforms and heterogeneous networking environments.
- **Innovative file system protocol optimizations:** prevent bottlenecks and increase overall data throughput.
- **Cost effective:** support for industry-standard platforms and heterogeneous networking environments significantly reduces deployment and support costs.

### A TECHNICAL AND ARCHITECTURAL FACT SHEET

Lustre™ redefines I/O performance and scalability standards for the world's largest and most complex computing environments. Ideally suited for data-intensive applications requiring the highest possible I/O performance, Lustre is an object-based cluster file system that scales to tens of thousands of nodes and petabytes of storage with groundbreaking I/O and metadata throughput.

#### UNIQUE STORAGE ARCHITECTURE

Lustre is a highly scalable distributed file system that combines open standards, the Linux operating system, an open networking API, and innovative protocols. Together, these elements create the world's largest "network-neutral" data storage and retrieval system.

Applying intelligence throughout its architecture, Lustre turns commodity hardware into smart storage devices which manage data objects. The objects are dynamically distributed horizontally across the servers. This shatters performance limitations of traditional storage systems.

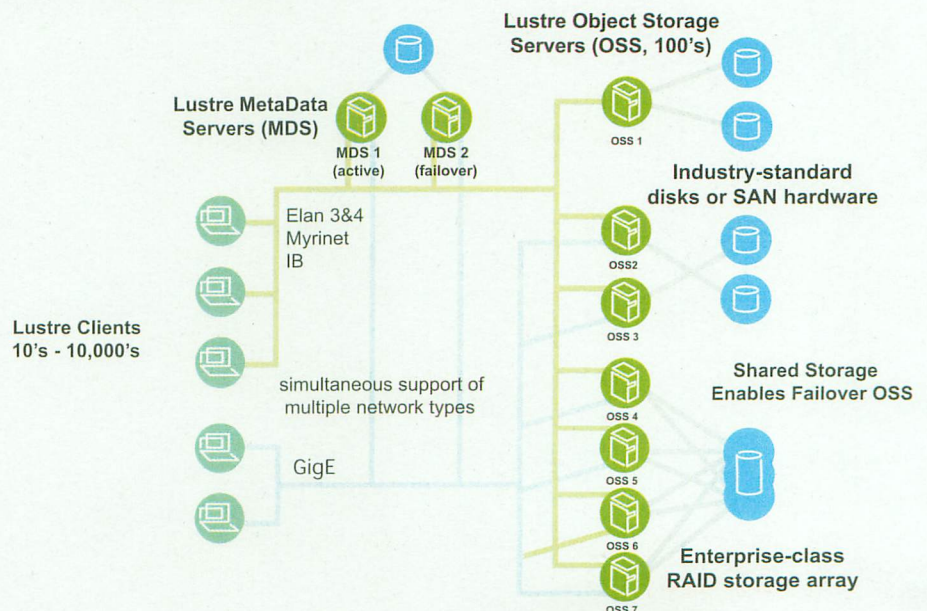
Building a Lustre cluster requires a Lustre MetaData Server (MDS) and Lustre Object Storage Servers (OSSs), each with disk storage. A pool of client systems access these servers through one of many supported networks.

Lustre file operations bypass the MetaData Server completely and fully utilize the parallel data paths to all OSSs in the cluster. This unique approach—separating metadata operations from data operations—results in significantly enhanced performance.

Like other Unix and Linux file systems, Lustre files are represented by inodes. But a key difference in Lustre is that its inodes simply contain references to the objects storing the file data.

#### OPEN SOURCE

Lustre has been developed and maintained as Open Source software under the GNU General Public License (GPL) enabling broad support for industry-standard platforms.





## " Our Lustre-based system is opening up a revolutionary capability for scientific applications."

- Mark Seager, deputy department head for terascale systems, Lawrence Livermore National Laboratory (LLNL runs Lustre on the worlds number one supercomputer, BG/L, and on many other clusters.)

### HETEROGENEOUS NETWORKING

Lustre's network architecture provides flexible support for a wide variety of networks and high performance features. Lustre interoperates with network vendor supplied libraries through Lustre Network Drivers (LND) which utilize advanced features such as Remote Direct Memory Access (RDMA), OS-bypass for parallel I/O, and vector I/O for efficient bulk data movement. LNDs exist for many networks including TCP, Quadrics Elan, many flavors of InfiniBand, and Myrinet GM, each with performance exceptionally close to the raw device throughput.

### RUGGED HIGH AVAILABILITY

Lustre organizes all servers in active-active failover pairs. Together with protocol interoperability between versions, live cluster upgrades are now routine.

### POSIX COMPLIANCE

Lustre provides a tested and fully compliant file system interface in accordance with the POSIX standard.

### CONFIGURATION? JUST MOUNT

Lustre (from version 1.6) revolutionizes configuration simplicity. Routine formatting and mounting of server devices aggregates them into a global high availability cluster file system.

### INNOVATIVE PROTOCOLS

Lustre employs a distributed lock manager to handle access to files and directories and synchronize updates, improving on the metadata journaling approach used by most modern file systems.

#### Intent-based locking

To dramatically reduce bottlenecks and to increase overall data throughput, Lustre uses an intent-based locking mechanism, where file and directory lock requests also provide information about the reason the lock is being requested.

For example, if a directory lock is being requested to create a new, unique file, Lustre handles this as a single request. In other file systems, this action requires multiple network requests for lookup, creation, opening, and locking.

#### Extreme Parallel Computing

The Lustre lock manager automatically adapts its policies to minimize overhead for the current application. Files being used by a single node are covered by a single lock, eliminating additional lock overhead. Nodes sharing files get the largest possible locks which still allow all nodes to write at full speed.

### ABOUT CLUSTER FILE SYSTEMS, INC

Founded in 2001, Cluster File Systems™ is the development and support organization for the Lustre™ File System. The company's premier object-based file system has demonstrated acceptance and capability on the world's fastest cluster supercomputers. Partnered with leading HPC storage, server, and software vendors, Cluster File Systems is poised for continued profitable growth as cluster customers worldwide realize the benefits of scalable, reliable storage with Lustre. The company is headquartered in Boulder, Colorado, with operations in North America, Europe, and Asia.

Lustre is Open Source software, distributed by CFS under the GNU General Public License.

### V1.4 REQUIREMENTS

Lustre 1.4 is now available to the general public under the GNU GPL.

#### Operating systems

- Red Hat® Enterprise Linux 3+ • SuSE® Linux ES 9
- Linux® 2.4 and 2.6

#### Hardware platforms

- IA-32 • IA-64 • X86-64 • PPC

#### Networking

- Quadrics® Elan 3 • Quadrics Elan 4
- TCP/IP • InfiniBand™ • Myrinet®

### PROVEN PERFORMANCE

Lustre powers most of the world's largest Linux supercomputers and is the first production tested object-based Linux cluster file system.

#### Recent results\*

File I/O % of raw bandwidth	>90%
Achieved single OSS I/O	>2.5 GB/sec
Achieved single client I/O	>2.0 GB/sec
Single GigE end-to-end throughput	118 MB/sec
Achieved aggregate I/O	22 GB/sec
Metadata transaction rate	8973 ops/sec
Maximum clients supported	11,500

\* Performance measurements made in 2004-2005 on production and test clusters at Bull, LLNL, and Sandia National Laboratories.

### PROFESSIONAL SUPPORT

Cluster File Systems, Inc. provides technical support, training, and engineering services for Lustre. In addition to tuning and supporting commodity storage components, we work closely with storage and cluster vendors to develop the next generation of intelligent storage solutions.

Please contact us for a free evaluation of the latest version, on-site training for you or your customers, and assistance with your file system deployment.

### FEATURED PARTNERS

- \* Bull
- \* DataDirect Networks®
- \* Hitachi Data Systems®
- \* Novell®
- \* Scali®
- \* Cray®
- \* HP®
- \* Linux NetworX®
- \* Quadrics®
- \* Sun Microsystems®



Copyright © 2005 Cluster File Systems, Inc.  
All rights reserved.

Lustre, the Lustre logo, Cluster File Systems, and CFS are trademarks of Cluster File Systems, Inc in the United States. All other product names mentioned herein may be trademarks or registered trademarks of their respective companies.

CFS#11.05 -- 1.4.6

For more information about Lustre, go to:  
<http://www.clusterfs.com>





# サイエンスグリッドNAREGIプログラム

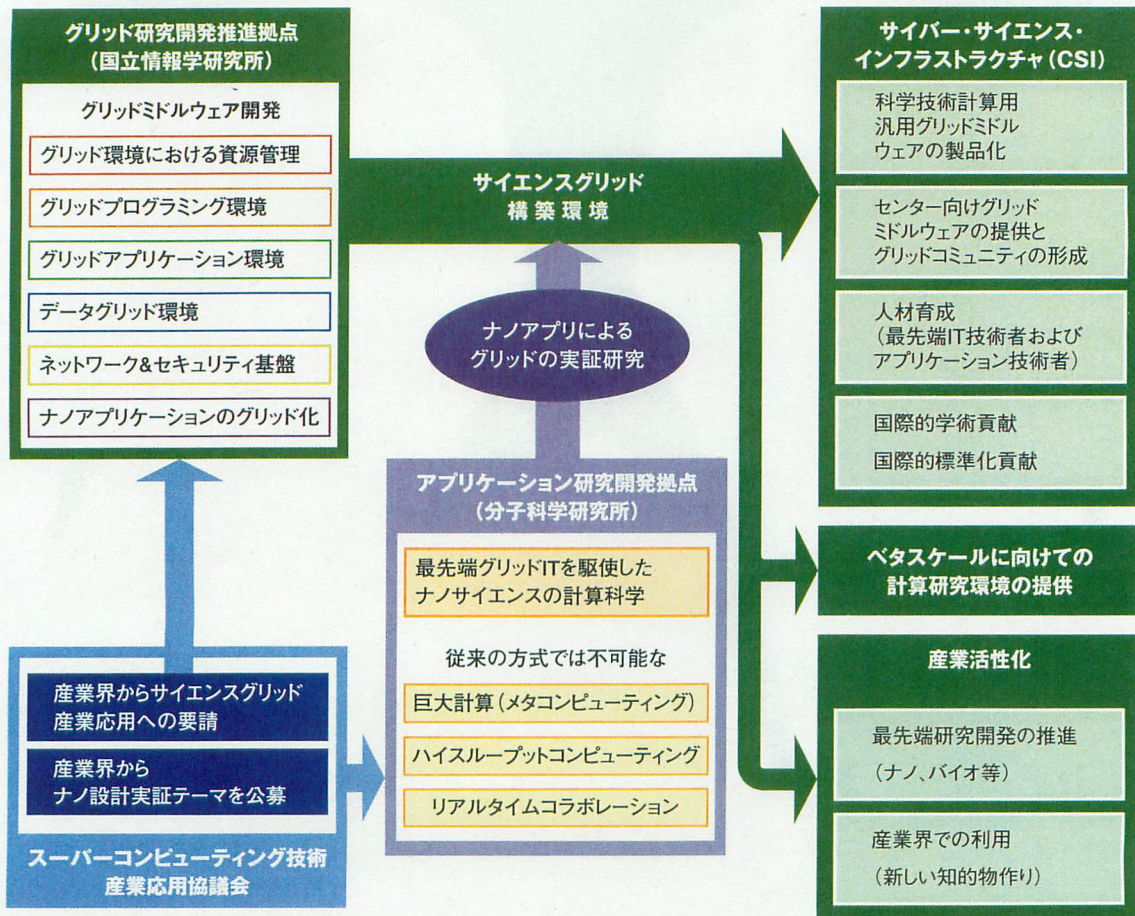
## NAREGIプログラムの目指すもの

NAREGI (National Research Grid Initiative) プログラムでは、「最先端・高性能汎用スーパーコンピュータの開発利用」プロジェクトの一環として、ペタスケール時代の計算研究環境の構築を視野に入れたグリッドミドルウェアの研究開発を行っています。

本プログラムは、国立情報学研究所及び分子科学研究所を中核として、各協力機関と強力な連携を図るために共同研究開発体制を取り、産業界とも連携を取っています。

国立情報学研究所においては、グリッドミドルウェアの研究開発及びグリッド環境の構築・運用に必要なツールの提供を行うとともに、諸外国のグリッド環境との連携を目指しています。ここで得られた成果は、最先端学術情報基盤(サイバー・サイエンス・インフラストラクチャ:CSI)の実現に大いに貢献するものと期待されています。

さらに分子科学研究所においては、グリッドミドルウェアを実証研究拠点として、ナノ分野をターゲットとし、従来コンピュータシステムでは実現不可能であった大規模シミュレーションソフトウェアの研究開発を実施しています。







# Science Grid NAREGI Program

## Objectives of NAREGI Program

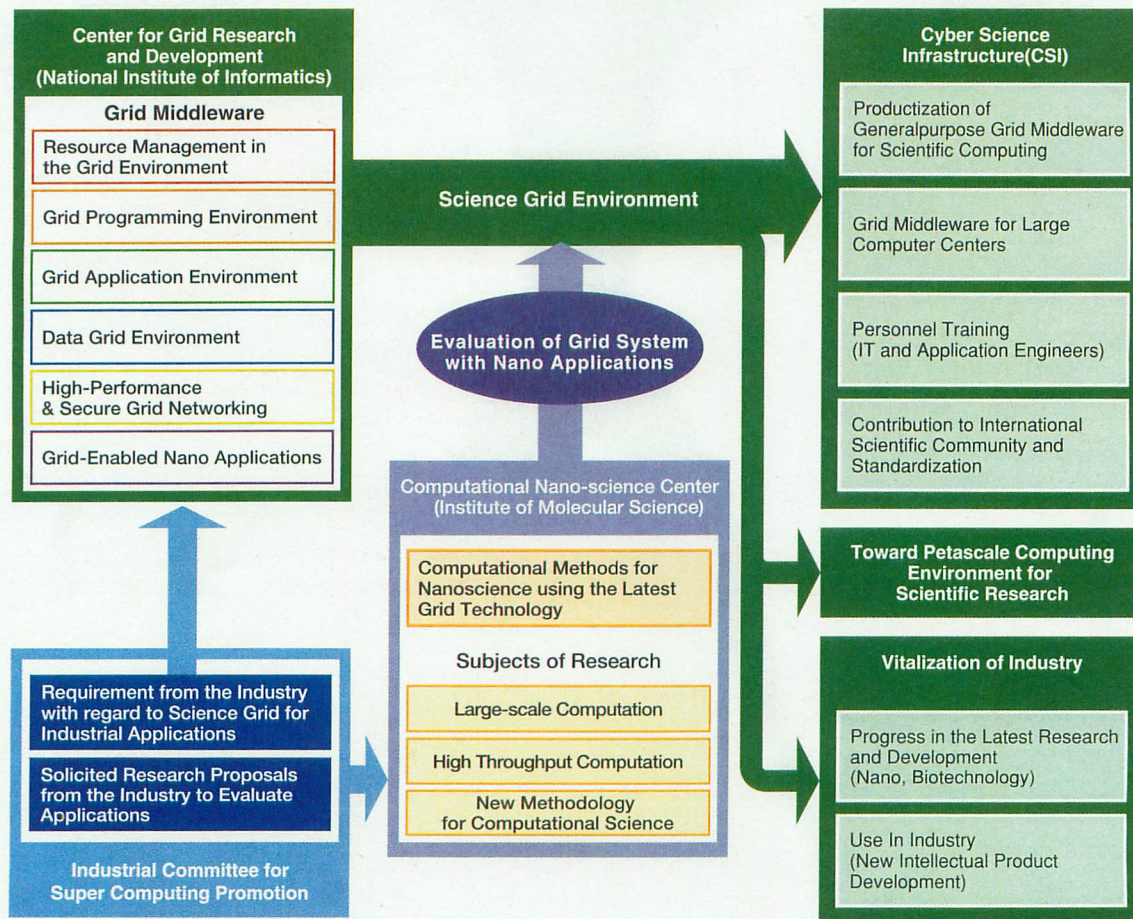
The NAREGI (National Research Grid Initiative) program aims to research and develop grid middleware that will put the construction of the computer research environment in the petascale era in view, as part of the "Development and Application of Advanced High-performance Supercomputer Project."

This program is a system of joint research development that designates the National Institute of Informatics (NII) and the Institute for Molecular Science as its core, in order to cooperate strongly with the cooperating organizations. The program also involves cooperation with the industrial world.

The NII offers the tools that are necessary for the research and development of the grid middleware and the construction and utilization of the resulting grid environments. In addition, the Institute aims at cooperation of the grid environment with those of various countries.

The achieved result is expected to contribute significantly to the achievement of the Cyber Science Infrastructure (CSI).

In addition, the Institute for Molecular Science as the proof research base of the grid middleware is targeting the nano-field by researching and developing large-scale simulation software.





## グリッド研究開発推進拠点における研究テーマの紹介

### グリッド環境における資源管理

- テーマ ●スーパースケジューラ ●グリッドVM ●分散情報サービス
- 研究内容 CPU数、緊急度、費用等ユーザからの要求を取り入れて計算資源の確保を行うブローカ機能を持つスケジューリング全体を司るスーパースケジューラ、下位の個々の計算資源において資源制御・資源保護並びにローカルスケジューリングを行うグリッドVM、さらにはグリッドにおける計算資源・ネットワーク・ソフトウェア並びにユーザ等のアカウントングを行い、それを統合的に管理する分散情報サービス等の研究開発を行います。

### グリッドプログラミング環境

- テーマ ●グリッドRPCシステム ●グリッドMPIシステム
- 研究内容 グリッドRPCシステムでは、遠隔計算機上でライブラリ関数を呼び出すモデルに基づき、数十から数百CPU規模の複数のクラスタを利用するグリッドアプリケーションの容易な開発と高い実行効率を可能とするシステムの研究と開発を行います。グリッドMPIシステムでは、グリッド上での通信遅延を考慮した高性能かつインターオペラブルな通信を実現するためにTCP/IPレベル及びMPIライブラリレベルでの通信ライブラリの研究開発を行います。

### グリッドアプリケーション環境

- テーマ ●グリッドワークフロー ●グリッドPSE ●グリッド可視化システム
- 研究内容 グリッド環境上でアプリケーションを簡単かつ効率的に動かす仕組みが重要となります。そのためにジョブの実行制御や操作性の良いGUIで簡単に記述できるワークフローツール、研究者が開発したアプリケーションをグリッド環境上へ配置、研究コミュニティでの共有のための登録を支援するグリッドPSE (Problem Solving Environment) 及び計算結果をグリッド上で視覚化するためのグリッド可視化システムの研究開発を行います。

### データグリッド環境

- テーマ ●データグリッド基盤技術 ●データベース検索制御技術 ●メタデータ構成技術
- 研究内容 インターネット上に散在する多数のデータベースを、グリッド環境下で仮想的にまとめて利用可能にする技術について研究・開発を行います。WSRFベースのOGSAを基盤としてデータ資源の管理や探索の機構を開発するデータグリッド基盤技術、多数のデータベース検索により引き起こされる組合せ爆発を抑えるための検索制御技術、異分野のデータベース間を意味的に関連付けるためのメタデータ構成技術などの研究開発を行います。

### ネットワーク&セキュリティ基盤

- テーマ ●ネットワーク通信基盤技術 ●セキュリティ・認証基盤技術
- 研究内容 NAREGIが目指すグリッドに必要な、ネットワーク通信基盤技術とセキュリティ及び認証基盤の研究開発を行います。具体的にはグリッド計算のためのネットワーク機能基盤技術に関して、ネットワークトラフィックの計測に基づく最適な経路・バックアップ用多重化経路の制御技術、グリッド上でのファイル転送に最適化された通信プロトコル、複数の組織をまたがる認証基盤の研究開発を行います。



## Research Themes

### Resource management in the grid environment

Super Scheduler Grid VM Distributed Information Services

NAREGI is currently conducting research and development of the Super Scheduler, which administers all the scheduling operations in the grid, including the "resource broker" functions. This function takes into account the requests from users, such as the number of CPUs, degree of urgency, and cost. Furthermore, efforts are also focused on securing computer resources through the Grid VM (Virtual Machine). This machine carries out resource control, resource protection, and scheduling at the local level of the computer resources. Efforts to secure computer resources are also being conducted through the Distributed Information Service, which is used for management and assessment of such aspects in the grid as computer resources, networks, software, and users.

### Grid programming environment

Grid RPC System Grid MPI System

As for the Grid RPC, the NAREGI project has been developing a system enabling easy development and high execution efficiency within the grid application software, with several clusters of a few dozen to hundreds of CPUs; this system is based on a model that allows the library functions to be called from a remote computer. As for the Grid MPI, NAREGI is carrying out research and development on TCP/IP-level or MPI-level communication libraries to realize high-performance, interoperable communication that takes into account the variable communication delay on the network. Both of these projects are expected to contribute to international standardization through the Global Grid Forum.

### Grid applications environment

Grid Workflow Grid PSE Grid Visualization System

Applications being simple and the movements of mechanisms being efficient are important within the grid environment. To this end, NAREGI is conducting research and development of a Grid Workflow and a Grid PSE (Problem-solving Environment). Grid Workflow is meant for easy control of job flow in Grid programming, either in terms of user-friendly GUIs or in terms of comprehensible external interface with the script languages. The research on PSE aims at developing an application development and execution environment that includes the deployment and registration, within the grid environment, of application software developed by researchers. Further efforts are focused on the execution and coordination of and collaboration among, distributed application software, computational modules, and data. Finally, research and development is underway on a Grid Visualization software tool, which visualizes the results of computations.

### Data Grid environment

Data Grid fundamental technology Search control technology for database federation

Metadata-based information integration for heterogeneous data resources

Technologies are under research and development for the federation of numerous databases spread throughout the Internet on the grid environment. The technologies include the Data Grid fundamental technology for managing and querying data resources using the WSRF-based OGSA infrastructure, search control technology (preventing combinatorial explosion caused by searching across many databases), and information integration technology with metadata that mediates heterogeneous data resources.

### High-Performance & Secure Grid Networking

Network communication infrastructure Security authentication infrastructure

With regard to network function infrastructure for grid computing, NAREGI is conducting research and development on the control technology, enabling determination of the optimal route based on the measurement of network traffic as well as establishing multiple alternative routes as backup. Work is also done on the communication protocol infrastructure, that is, optimization of the communication protocol for large-sized file transfer on the grid. As for the security infrastructure, the goals are to develop a security model based on PKI and to implement authentication infrastructure across multiple organizations.



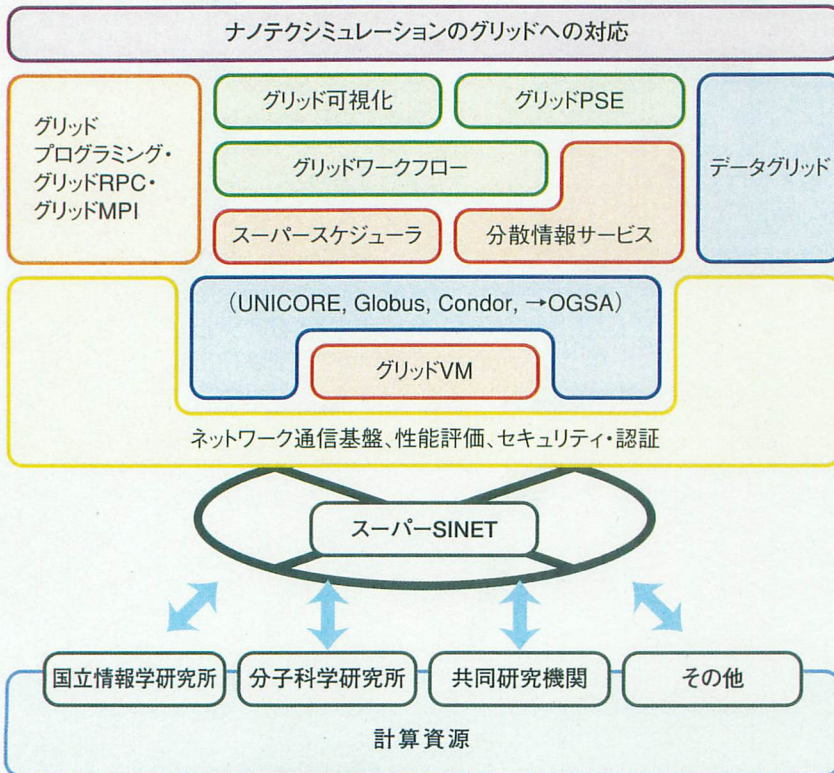
## グリッド研究開発推進拠点における研究テーマの紹介

### ナノアプリケーションのグリッドへの対応

テーマ ●ナノアプリの並列化・分散化等によるグリッド対応  
 研究内容 ナノサイエンス実証研究拠点が開発しているナノ分野アプリケーションのグリッド対応化を行うとともに、ナノ分野連成ミドルウェアの開発とグリッド環境における応用研究を行います。また、グリッドアプリケーション環境の研究開発と連携協力し、グリッド上におけるナノ分野アプリケーションの実行環境を整備します。

### グリッドミドルの利活用技術の研究

テーマ ●APIの研究開発 ●異種グリッド相互利用技術の研究開発  
 研究内容 ITBL (IT-Based Laboratory) プロジェクトにおいて構築された実運用グリッド環境の資産を、次世代スパコンを頂点とする次世代の研究グリッドインフラに円滑に継承するため、アプリケーションプログラミングインターフェースの研究開発、および異なる複数のグリッドミドルウェアの利活用が円滑にできるようにするための相互利用技術の研究開発を行います。



### NAREGIプログラム共同研究機関

- 産：富士通株式会社  
株式会社日立製作所  
日本電気株式会社  
スーパーコンピューティング技術  
産業応用協議会  
(製薬、化学、金属・材料企業等)
- 学：グリッド研究開発推進拠点  
東京工業大学、大阪大学、  
九州大学、九州工業大学等  
アプリケーション研究開発拠点  
東京大学、京都大学、東北大学、  
KEK物質構造科学研究所等  
産業技術総合研究所  
日本原子力研究開発機構等

問合せ先/グリッド研究開発推進拠点(リサーチグリッド研究開発センター)  
 TEL 03-4212-2857 FAX 03-4212-2803 URL <http://www.naregi.org/>



## Research Themes

### Grid-enabled Nano-applications

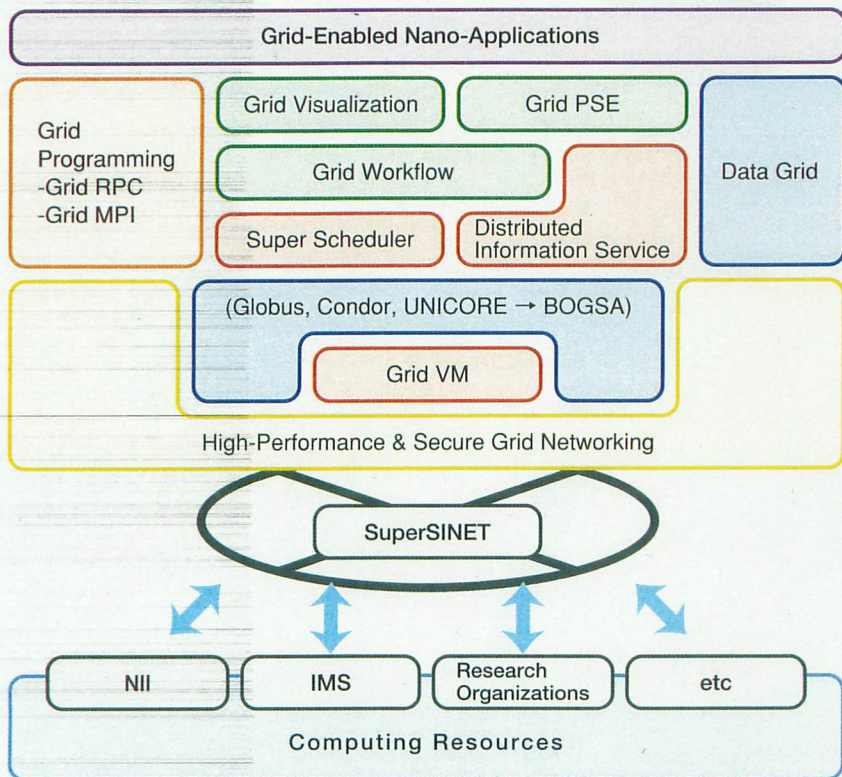
Parallelized and decentralized nano-applications for the grid

The NAREGI project aims at making the nano-application software-which has been developed by researchers at the Computational Nano-science Center at IMS-grid-ready. The NAREGI project is also working on development of middleware for coupled simulations in the nano-science / nano-technology areas, to conduct applied research in the grid environment, and generally to create a grid environment suitable for nano-applications.

### Research on Utilization of Grid Middleware

Application Programming Interfaces (API) Heterogeneous mutual grid utilization technology

Research and development of API and the interoperability technologies, in order to realize a smooth transition from the operational computing environment constructed in the ITBL (IT-based Laboratory) project to the next generation science grid infrastructure, centered around the next generation supercomputing system



### NAREGI Cooperating Research Institutes

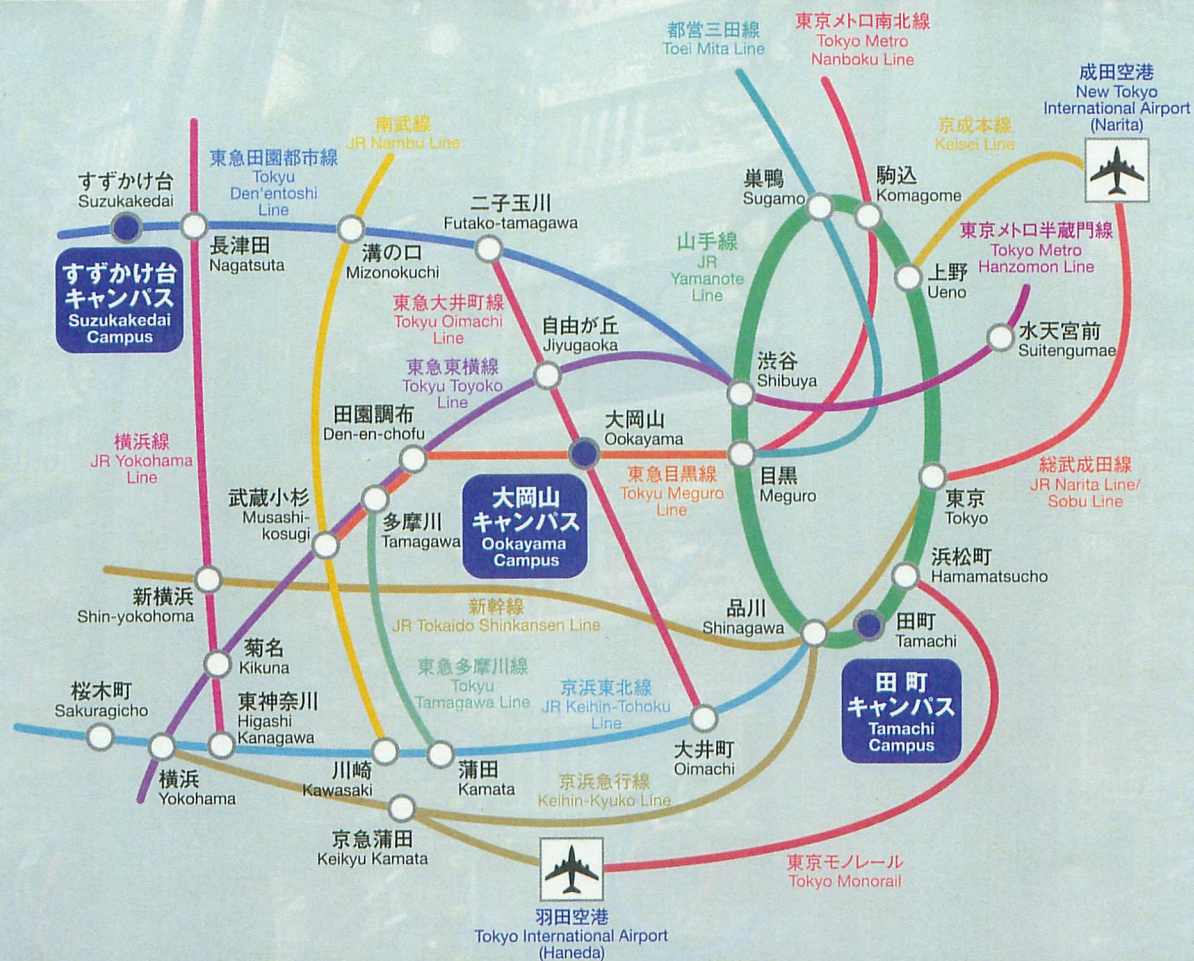
Industrial sector  
 Fujitsu, Hitachi, NEC,  
 Industrial Committee for Super Computing  
 Promotion  
 (including pharmacy, chemical, metal, material  
 companies)

Academic sector  
 Center for Grid Research and Development  
 Tokyo Institute of Technology, Osaka University,  
 Kyushu University, Kyushu Institute of Technology  
 Computational Nano-science Center  
 The University of Tokyo,  
 Kyoto University, Tohoku University  
 High Energy Accelerator Research Organization,

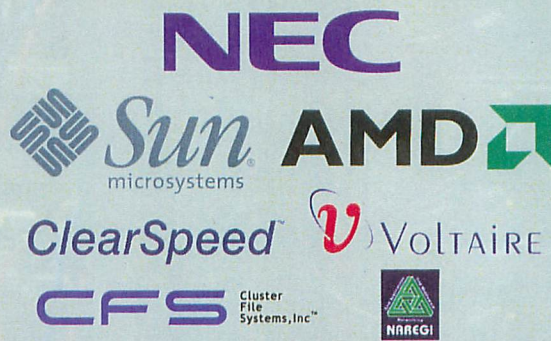
Research laboratories  
 National Institute of Advanced Industrial Science  
 and Technology,  
 Japan Atomic Energy Agency

Contact: Center for Grid Research and Development (Collaborative Center for Research Grid)  
 Tel: +81-3-4212-2857 Fax: +81-3-4212-2803 URL: <http://www.naregi.org/>





大岡山キャンパス  
Ookayama Campus



交通案内

大岡山キャンパス ● 東京急行大井町線 / 目黒線 ■ 大岡山駅下車徒歩1分  
 すずかけ台キャンパス ● 東京急行田園都市線 ■ すずかけ台駅下車徒歩5分  
 田町キャンパス ● JR山手線 / 京浜東北線 ■ 田町駅下車徒歩2分

Ookayama Campus ● Ookayama Station of Tokyu Oimachi Line / Tokyu Meguro Line  
 Suzukakedai Campus ● Suzukakedai Station of Tokyu Den-en-toshi Line  
 Tamachi Campus ● Tamachi Station of JR Yamanote Line / Keihin-Tohoku Line

国立大学法人 東京工業大学  
学術国際情報センター

〒152-8550 東京都目黒区大岡山2-12-1  
 Phone: 03-5734-2087 Fax: 03-5734-3198  
 E-mail: office@gsic.titech.ac.jp  
 URL: http://www.gsic.titech.ac.jp/

Tokyo Institute of Technology  
Global Scientific Information and Computing Center

2-12-1 O-okayama, Meguro-ku, Tokyo 152-8550 JAPAN  
 Phone: +81-3-5734-2087 Fax: +81-3-5734-3198  
 E-mail: office@gsic.titech.ac.jp  
 URL: http://www.gsic.titech.ac.jp/