グリッドコンピューティング

SGIの取り組みと事例紹介

日本SGI株式会社

説明概要

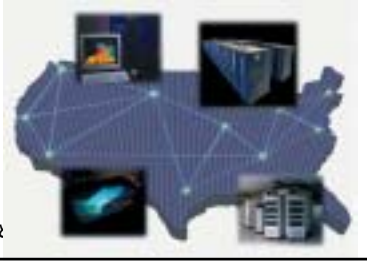
- コンピュータ利用技術の新しい動き
- グリッドコンピューティングの概要
- グリッドコンピューティングの事例
- SGIでの取り組みについて
- まとめ

コンピュータ利用技術の新しい動き

- 背景
  - ネットワークと計算機の高速化
  - Internetの普及
  - 大規模計算の需要の拡大
  - 計算機リソースに対する継続的な要求
  - システム負荷の問題：“over-load” と “under-utilized”
  - 情報システム総合コスト; TCO（total cost of ownership）の低減要求
- より柔軟で、スケーラブルなソリューションへの要求

広域分散コンピューティング

- PCによる広域分散コンピューティング
  - 遊休計算機資源の活用
  - ピア・ツー・ピア技術
- グリッドコンピューティング
  - 広域ネットワーク上に配置された複数の計算資源を利用して分散/並列計算を行うシステムを構築



PCによる広域分散コンピューティング

- ピア・ツー・ピア技術で、PCの余剰時間を集め、大規模な計算が可能であることを示す事例
  - Intel-United Devices CANCER Research Project
  - SETI@home
- 最大級のスーパーコンピュータに匹敵する計算能力
- インターネットやプライベート・ネットワーク上の数百、数千のコンピューターにコンピューティング・タスクを下請けさせるというビジネスモデル

分散コンピューティングプロジェクト

- SETI@home
  - インターネットにつながっているコンピュータを使って地球外知的生命体の探査(SETI：the Search for Extra-Terrestrial Intelligence)を行なう科学実験
  - 無料のプログラムをダウンロードして電波望遠鏡のデータを分析することで、参加することが誰でも参加可能
- Folding@home
  - スタンフォード大学化学部主催
  - タンパク質の折りたたみがどのような課程をたどっているかをシミュレート
  - タンパク質の折りたたみメカニズムの解明を目指す

分散コンピューティングプロジェクト

- United Devices
  - Intel-United Devices CANCER Research Project：**数種類の病気との関連で重要なたんぱく質に、膨大な数の色々な分子がどう作用するか調査し、新しい抗ガン剤・白血病への治療方法等への開発に繋がる物を発見する。**（Intel,**アメリカ国立癌研究財団**(National Foundation for Cancer Research)，**全米癌学会**(American Cancer Society)，**オックスフォード大学との共同**）
  - Intel Philanthropic Peer-to-Peer Program

ピアツーピアのネットワークの利点

- 演算タスクが極めて多くのコンピューターに分散
- 放っておけば空き時間になってしまうコンピューターの時間でピアツーピアの作業処理
- 高価なスーパーコンピューターがなくても研究者が複雑な計算を行なったりプログラムを実行できる方法として人気

ピアツーピアのネットワークの問題点

- 一般的な計算機システムの構築は困難
  - 実際の計算を行うコンピューターとの間の情報伝達のための帯域幅が足りない
  - 計算に参加する無数のコンピューターに、計算ソフトをインストールしなければならない
- “人々にソフトをインストールさせるだけでも十分困難だが、企業が、たとえば計算内容をチップの設計からデリバティブ市場の分析に変えたいとき、人々にインストール済みソフトを新しいソフトと入れ替えてもらうのは更に難しい”

Cell computing ™


NEWS RELEA

分散コンピューティング技術により超大なCPUパフォーマンスを提供する
「cell computing™（セル・コンピューティング）」の事業化に向けた検討開始
ーインテル、NTT東日本、日本IBMの協力を得ー
国内初の大型分散コンピューティングの検討ー

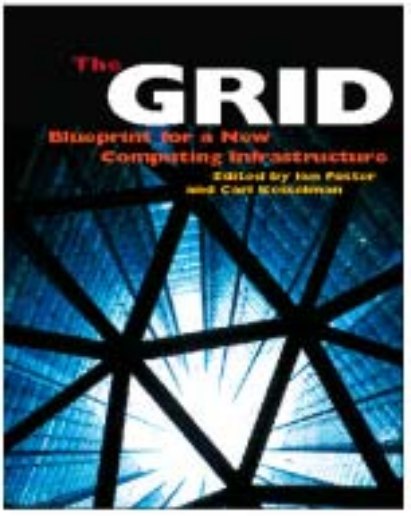
グリッドコンピューティング概要

一般的な定義

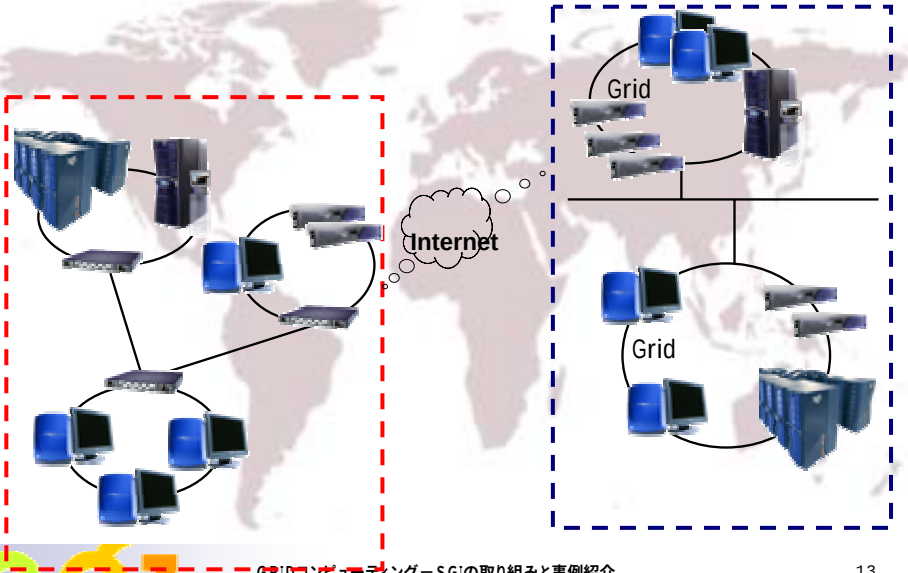
ネットワーク上に点在する異機種の計算機システム、ストレージ、グラフィックスシステムをより効果的、効率的に使用することを目的としたIT技術

グリッドコンピューティングの概要

- 1995年頃から、活発な活動
- ネットワークの上の計算機資源（リソース）を共有することを可能とすることを目的とする
- セキュリティ、リソースマネージメント、異機種コンピュータ環境の構築などの分野の研究・開発が要求される
- 多くの大学、研究所、官庁で採用



グリッドコンピューティング

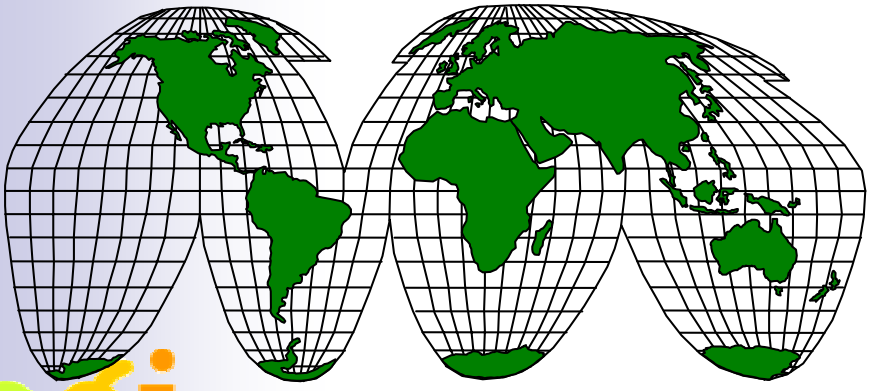


グリッドコンピューティングとは?

- その発展の背景は…
  - “Distributed Computing” – 分散コンピューティング技術の発展
    - 大規模アプリケーションを多くのシステムに分散させて処理
    - システムの‘遊休’リソースを最大限に利用して、ジョブの処理
- ネットワークを介した広域分散コンピューティングの可能性の拡大

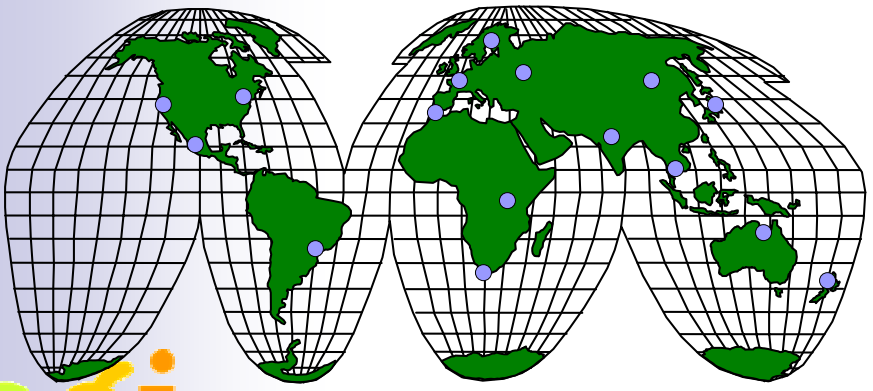
市場の要求

- どのような問題を解く必要があるのか？
  - 科学技術計算分野で、最大規模の問題を解く



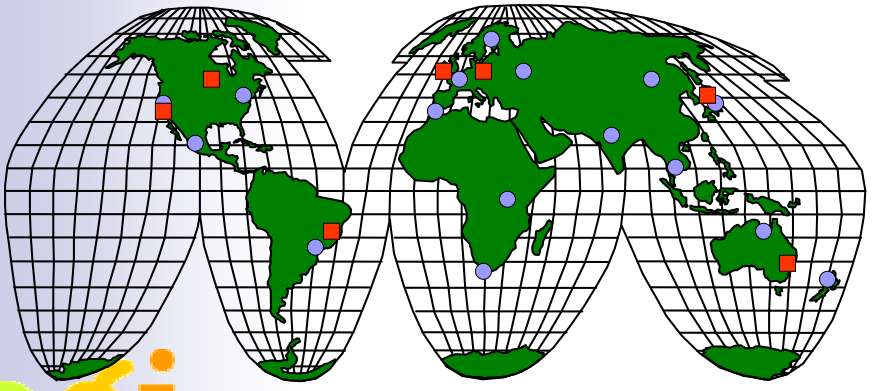
ユーザの分散

- …しかし、そのためには、それに関わる人々がより多く必要であり、その人たちは、広範囲に分散し、また移動している…



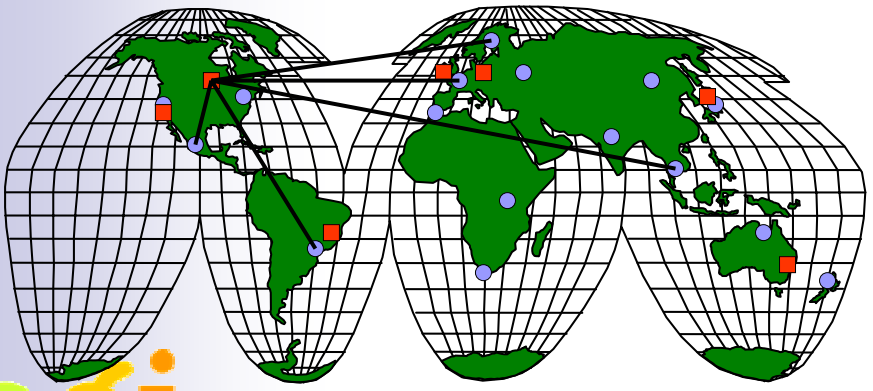
しかし…

- 一方、先端的なコンピュータシステムは限定される



したがって…

- 現在では、ユーザは、リモートにあるシステムにログインして、そのシステムを使用することになる…



グリッドコンピューティングとは?

- ネットワーク結合型コンピュータ利用技術の新たな動き
  - コンピュータリソースをプールし、必要な時に必要なユーザに提供
  - ユーザは、透過的なアクセスが可能で、必要な時に自由に使用可能

➔ 電力グリッドが、何処で発電がなされているかを意識することなく、電力を使用できることを可能としていることに対応して、(Grid Computing) グリッドコンピューティングと呼ばれている

グリッドコンピューティングとは?

- 何が可能となるのか?
  - 広域ネットワーク上に分散配置された計算資源を仮想的な高性能計算機 (メタコンピュータ) と見た立てて分散・並列計算が可能となる
  - ユーザが、コミュニティを形成し、お互いの計算機リソース (計算能力、データなど) を共有化することで、研究や開発の効率化を図ることを可能となる
- ネットワークを介したコンピュータの利用技術であり、計算能力自体の向上を図るものは一般にない

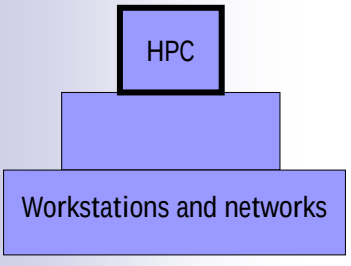
グリッドコンピューティングの各社の見解

- 「ユーザーが膨大な量のコンピューティングパワーを利用できるようにしてくれる非常に高い潜在能力があることから，グリッドが企業の環境で一般的なものとなるのは時間の問題だ」
- 「われわれは，これを避けることのできないものだと考えている。移行が避けられないほどに得られるものは非常に多く，企業にとってのメリットは計り知れない。あとは，その時期と所要時間だけの問題だ」
  - サンのラージアカウントシステム製品グループでグループマーケティングマネジャーのジェフコック氏
- 「私は“インターネットの次の大きな動きは何か”という質問をよく受ける。これまで私にはその答えがなかった。だが今では，グリッドコンピューティングが次の大きな動きになると強く確信している」
  - IBMのインターネット戦略担当副社長，ジョン・パトリック氏

計算機システムレイヤーとマーケット

Volume　　Business　　Grid Application

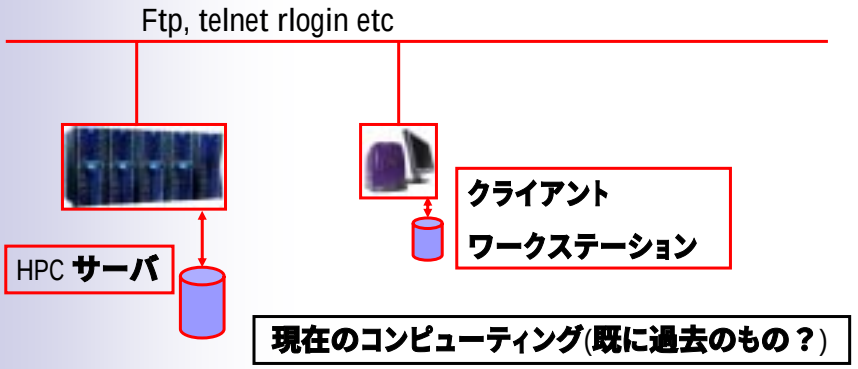
Core Competency Computing :企業などで、最も他に比べて優れていると考えられる技術・能力を活用したコンピューティングによって、研究、製品開発、ビジュアルコラボレーションなどを実現



Core Competency Computing
スケーラビリティ
新しい計算機環境の構築
問題解析のための新たなアプローチ

IT infrastructure
遊休サイクル
アクセス
セキュリティ

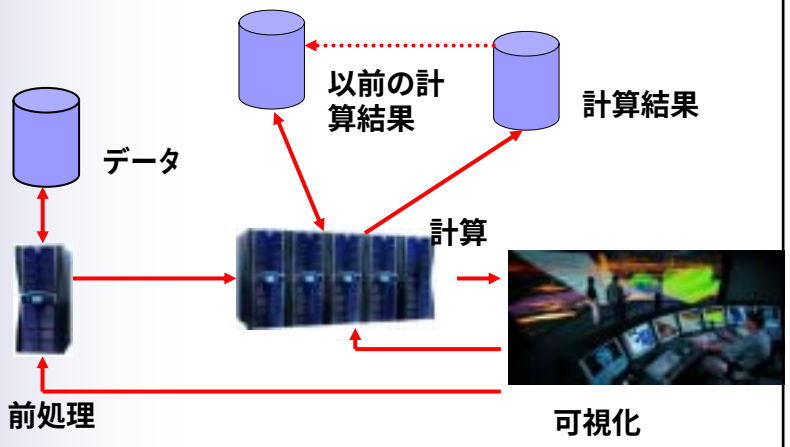
グリッドコンピューティングの価値



現在のコンピューティング(既に過去のもの？)

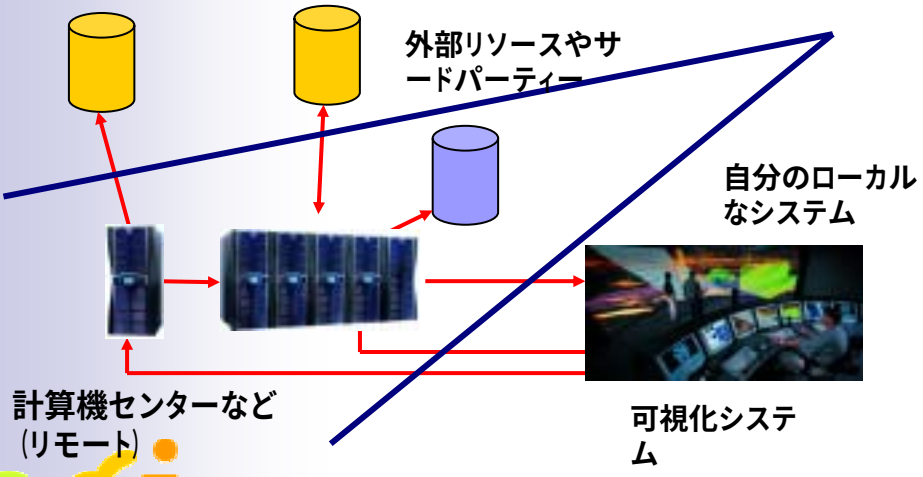
グリッドコンピューティング

- 計算･シュミレーションでのデータの動き



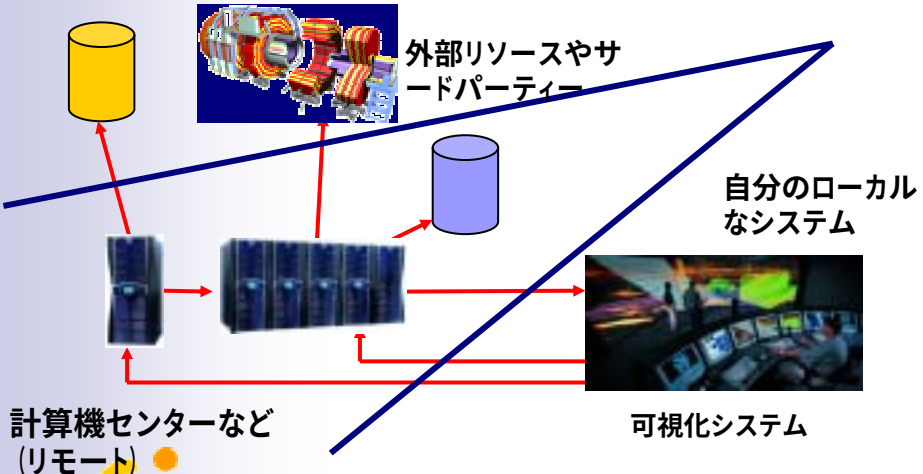
グリッドコンピューティング

- ドメイン間でのデータのやり取り



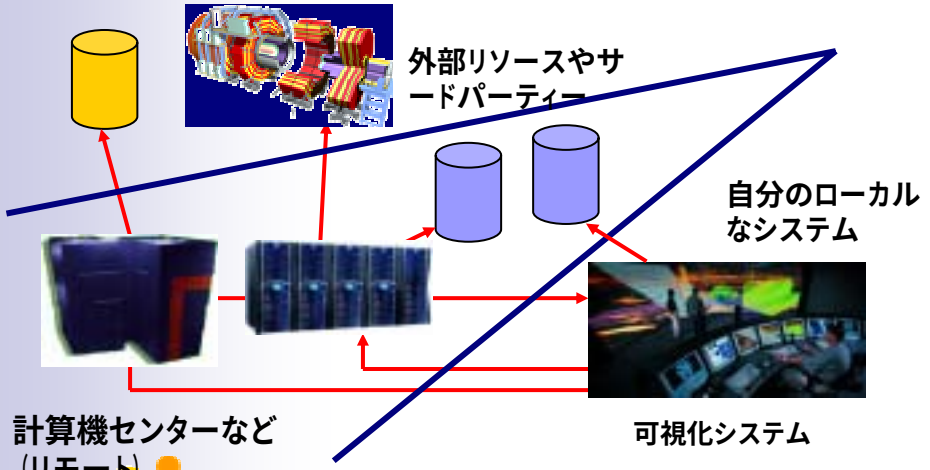
グリッドコンピューティング

- 実験装置などからのデータの取り込み



グリッドコンピューティング

- 他のスーパーコンピュータとの連携



グリッド関連プロジェクト

- The Globus Project (Globus ツールキットの開発と提供)
  - www.globus.org
  - オープンアーキテクチャ/オープンソース
  - サービス
    - セキュリティ
    - リソース管理
    - データアクセス
  - 主に、大学、研究所の方々が参加

グリッド関連プロジェクト

- IEEE Task Force on Clustered Computing (TFCC)
  - “the CCGRID2002 International Symposium on Clustered and Grid Computing”のスポンサー
- New Productivity Initiative
  - www.newproductivity.org
  - システムベンダー、ソフトベンダー、サービスプロバイダーによって構成
    - SGI
    - HP/Compaq
    - Platform
    - Terraport
    - その他、多くのベンダー

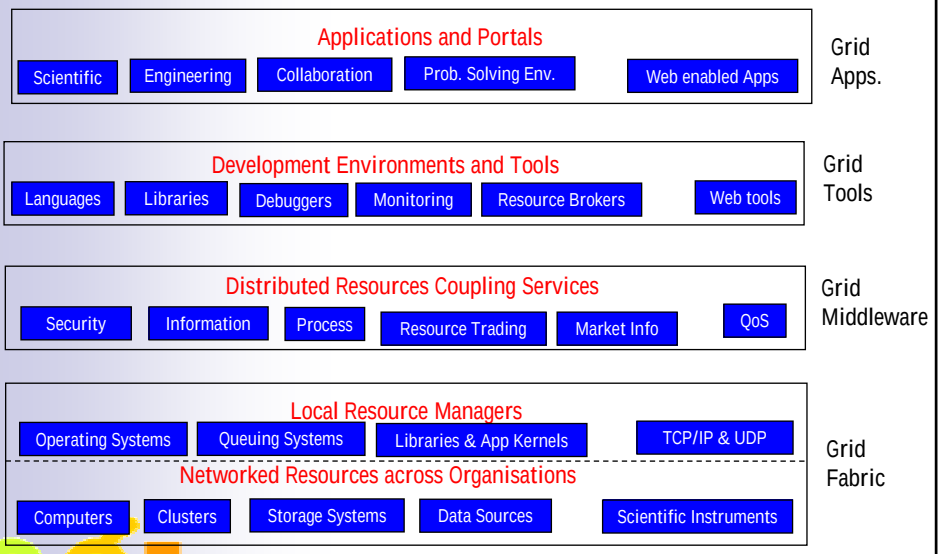
グリッドコンピューティングの実績

- NCSA/Max Planc/SDSC/Argonne – Cactus の実行 (relativistic astrophysics simulation)
  - 3台のOrigin 2000’s (one 256p, two 128p)とIBMの128 nodes (1020 p) をATM/OC12ネットワーク で接続
    - Cactus toolkit/Globus toolkit/MIPCH-G2
    - ピークの70%のスケーラビリティ
- European DataGrid Project
  - http://www.eu-datagrid.org/
  - CERN, CNRS, ESRIN,INFN,NIKHEF,PPARCとその他15 のパートナー

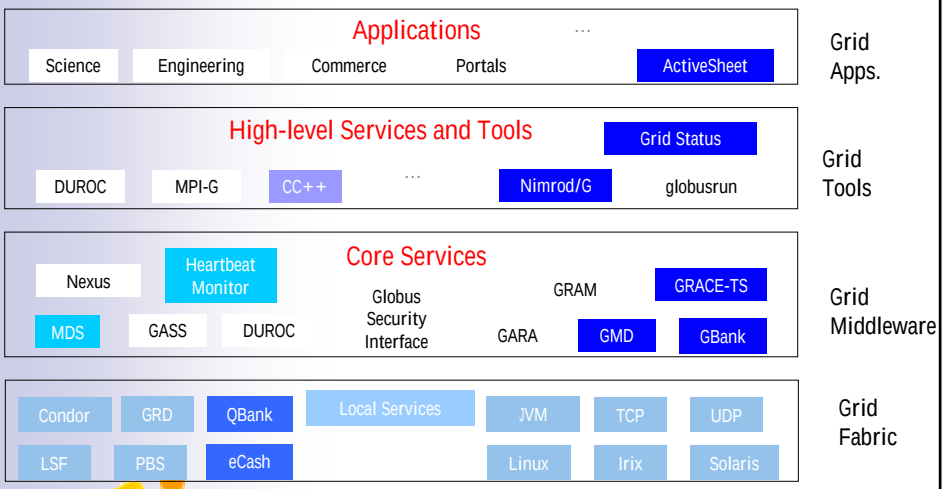
グリッドコンピューティングの実績

- NASA PowerGrid
  - Information Power Grid (IPG)
  - **政府機関、大学、民間企業の共同プロジェクト**
- ASCI Distributed Resource Management Testbed
  - **ローレンスリバモア国立研究所とサンディア国立研究所**
  - **計算リソースの統合と運用管理のための試験台**

グリッドコンポーネント／機能概要



グリッドコンポーネント／ツール概要



ソフトウエアツール(フリー/オープンソース)

- Globus toolkit (アルゴンヌ国立研究所)
  - www.globus.org
  - グリッドアプリケーションのためのサービスツールキット
    - Globusリソースアロケーションマネージャ (GRAM)
    - グリッドでのセキュリティ機能
    - メタコンピューティング用ディレクトリサービス
    - 2次ストレージに対するグローバルなアクセス
    - Neuxとグローバルなコミュニケーションサービス
    - システム間でのハートビートモニター
    - 各サービスでのAPIの提供

ソフトウエアツール(フリー/オープンソース)

- MPICH-G2
  - http://www.niu.edu/mpi/
  - グリッドコンピューティングに対応したMPI v1.1 仕様のMPI
  - Globusのサービスを使用
  - “globus2” デバイスを使用したMPICH
  - アルゴンヌ国立研究所から提供

ソフトウエアツール(フリー/オープンソース)

- Sun Grid Engine ソフトウエア
  - www.sun.com/gridware
  - Sunのホームページからダウンロード可能(OpenSource)
    - 基本プロダクト (Sun Grid Engine Software) は、無料
    - Global Resources Director は製品化
  - クロスプラットフォームのサポート
    - Sparc/Solarisと IA-32/Linux
- Milan Project (Metacomputing in Large Asynchronous Networks)
  - http://www.cs.nyu.edu/milan/milan/
  - ニューヨーク大学とアリゾナ州立大学

ソフトウエアツール(フリー/オープンソース)

- GRACE (Grid Architecture for Computational Economy)
  - http://www.csse.monash.edu.au/~rajkumar/ecogrid/
  - Globus上で実行
  - プラットフォームサポート:
    - Alpha/Tru64
    - Intel IA32/Linux 2.2
    - MIPS/Irix 6.5
    - SPARC/Solaris 7 or 8
    - RS/6000/AIX
    - Alpha/Digintal Unix 4.0
- Condor Project - http://www.cs.wisc.edu/condor
  - Condorプロジェクトは、ワークステーションの計算パワーのより効果的な使用を目的としたもの
  - Condorでは、ワークステーションクラスタの運用管理が可能
  - 分散したワークステーションの遊休CPUを使用して、計算の実行を行う
  - Solaris, IA-32 Linux, IRIX, NT, Digital Unix および HP-UXをサポート

Entropia – 商用ソフトウエア

- 3つの主要コンポーネント
  - ネットワークマネージャ
    - クライアントのリソースのコントロールとどのクライアント上でどのアプリケーションを実行するかを管理
    - GUIベース
  - ジョブマネージャ
    - ジョブの投入とそのジョブの実行状況のモニター
  - クライアントソフトウエア
    - クライアントマシン上でアプリケーションを実行

グリッドコンピューティング事例

グリッドコンピューティング事例

- IPG：the NASA Information Power Grid
  - データベース、計算リソース、実験装置、可視化システムを統合
  - 異種計算機による分散コンピューティングシステムを構築
  - 計算リソースとして、大規模なOriginシステムを使用

NASA Information Power Grid (IPG)

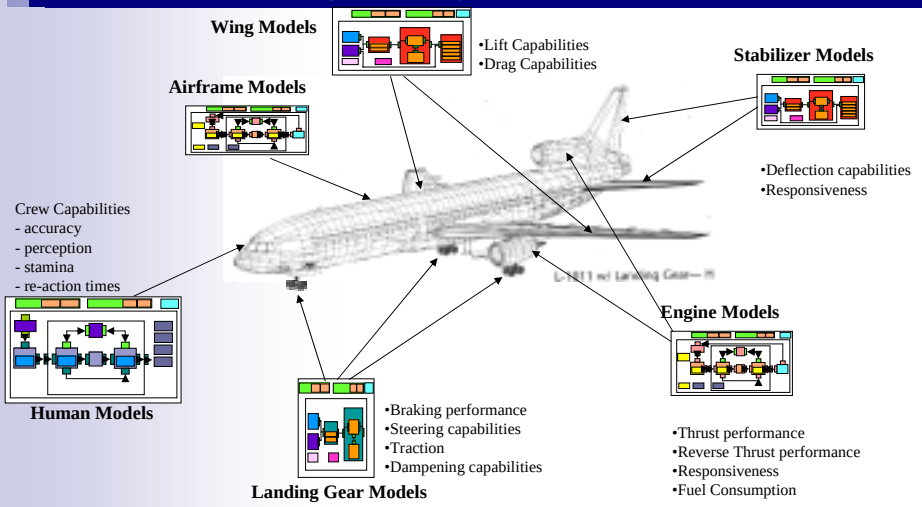
http://www.ipg.nasa.gov

IPG：the NASA Information Power Grid

- NASAの実績と成果
  - 完全な航空機のシュミレーション
    - “*Multi-Disciplinary Simulation*”：より安全性の向上を図る
  - 完全な米国領空のシュミレーション

Multi-Disciplinary Simulation：連成解析や他領域解析などと呼ばれる手法で、一つの解析だけでなく、複数の設計要因や構造を同時に検討するようなシュミレーション。近年、航空機設計や自動車設計などの分野において、盛んに用いられている。

Multi-Disciplinary Simulation

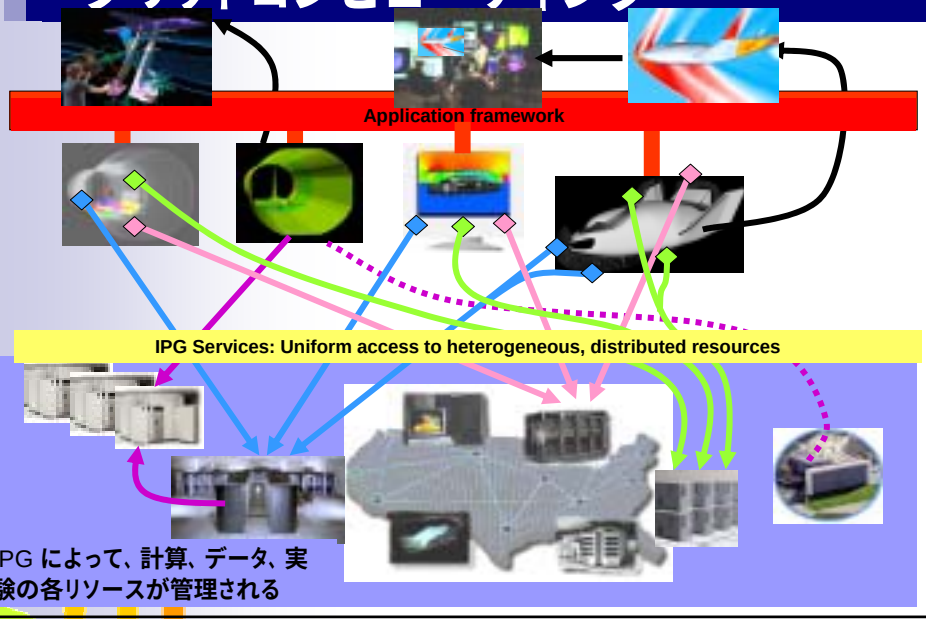


システム全体のシュミレーションは、個々のサブシステムのシュミレーションを組み合わせて実行される

グリッドコンピューティングの効果

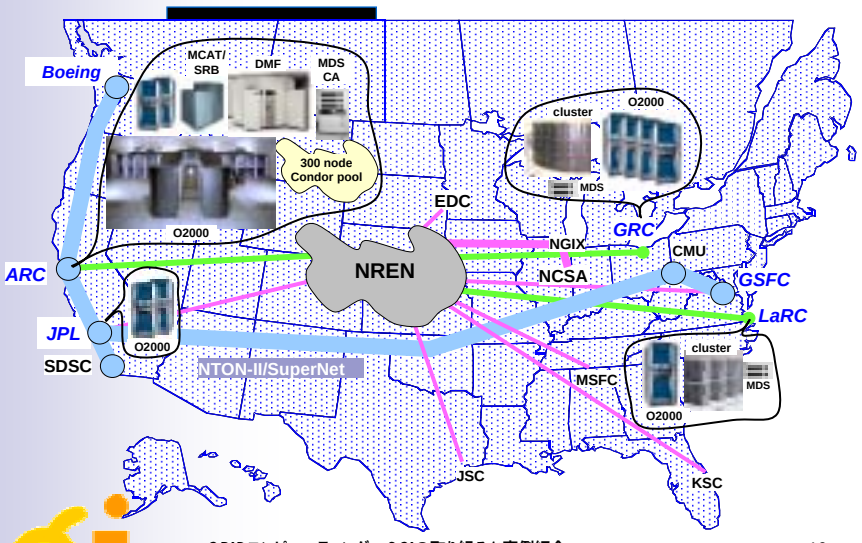
- NASAの実績と成果
  - リアルタイムでの実験データの収集とそのシュミレーション
    - 空洞実験データの取得
    - 衛星データの取得
  - より効率的なリソースの利用
  - データマイニング

グリッドコンピューティング



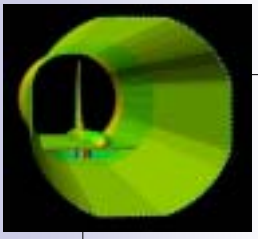
IPG によって、計算、データ、実験の各リソースが管理される

IPG：the NASA Information Power Grid

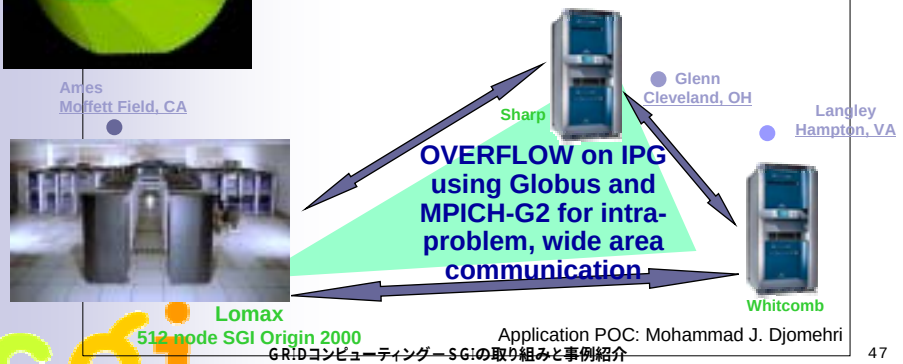


大規模分散コンピューティング:

high-lift subsonic wind tunnel model



The research branch of NAS is investigating algorithms that are suitable for a Grid computing "meta-platform." One candidate is overset grid codes that can tolerate timestep mis-matches on the intra-object boundaries. A version of the OVERFLOW, Navier-Stokes, CFD simulation code is being modified for this approach. It has been demonstrated operating across systems at ARC, GRC, and LaRC, solving for flow about large test objects mounted in a wind tunnel.

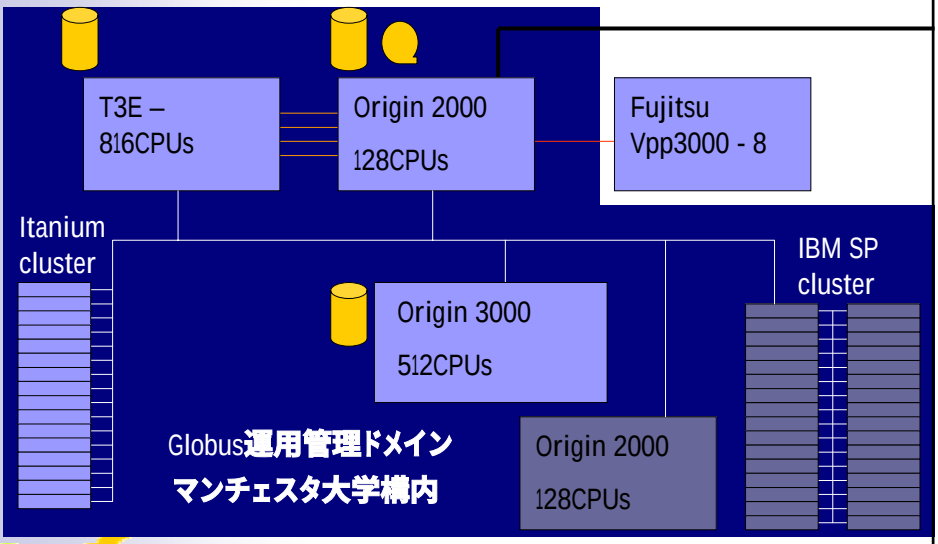


Application POC: Mohammad J. Djomehri

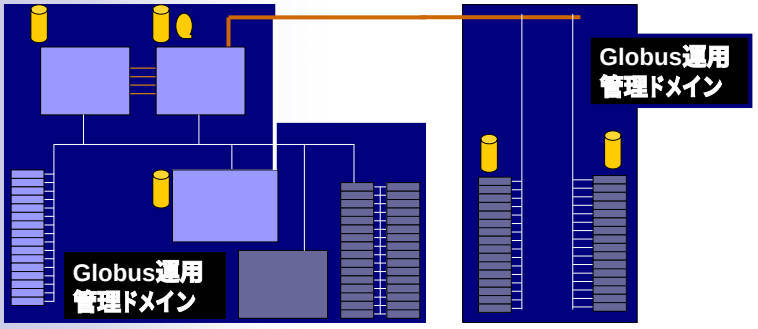
グリッドコンピューティング事例

- CSAR
  - マンチェスタ大学に設置された、英国国内の大学に対するHPCサービスを行う機関
  - Cray T3E, Origin 2000, Origin 3000, Itaniumクラスタおよび Fujitsu システムから構成される
    - 約1500のスカラプロセッサと8台のベクトルプロセッサから構成
  - SGIとCrayが提供するCSA (Comprehensive System Accounting) によって、アカウンティングと課金処理を行っている
  - ディスク、CPU、テープなど個々のリソース毎に異なった料金体系

CSARシステム構成図



CSARシステム構成図



マンチェスタ大学　　Daresbury Labs - CCLRC

- 英国の研究者が使用出来る計算機システムプロトタイプを構築
  (EPSRC:The Engineering and Physical Sciences Research Council)
- ユーザがジョブの管理を行える手法の開発
- Globusによるグリッドコンピューティングの試験台

DoD HPCMP

- **国防総省** (DOD) **のスパコン･センター更新計画** HPCMP (High Performance Computing Modernization Program)
- DoD HPCMP**のイニシアチブを取る**3**本柱**
  - **高性能コンピューティング**(HPC)**センター**
  - **ネットワーキング**
  - Common High Performance Computing Software Support Initiative (CHSSI)
- **プラットフォームコンピューティング社 の技術資料**: http://www.platform.com/PDFs/whitepapers/buildinggrids.pdf

DoD HPCMP**の目標**

- **市販されている最高の高性能**HPC**システムを継続的に確保する**
- **共用ソフトウェアツールとプログラミング環境の確保と開発**
- DoD**高性能コンピューティングのユーザベースの拡張と訓練**
- **高速ネットワークを介したユーザとコンピュータのリンクと協力的な作業環境の作成の促進**
- **全国的**HPC**基盤から生まれる最高のアイデア、アルゴリズム、およびソフトウェアツールの活用**

DoD HPCMP 計算機センター

	Site	Location	Type
AEDC	Arnold Engineering Development Center	Arnold Air Force Base, TN	DC
ARL	Army Research Lab	Aberdeen Proving Ground, MD	MSRC
AFFTC	Air Force Flight Test Center	Edwards Air Force Base, CA	DC
NAVO	Naval Oceanographic Office	Stennis Space Center, MS	MSRC
NRL	Naval Research Laboratory	Washington, DC	DC
RTTC	Redstone Technical Test Center	Redstone Arsenal, AL	DC
SMDC	Space and Missile Defencse Command	Huntsville, AL	DC
SSCSD	Space and Naval Warfare Systems Center, SD	San Diego, CA	DC
TARDEC	Tank-Automotive Research, Development and Engineering Center	Warren, MI	DC
WSMR	White Sands Missile Range	White Sands Missile Range, NM	DC

- 4つのMajor Shared Resource Centers (MSRC:主要共有リソースセンター)
- 17のDistributed Centers (DC:分散センター)

DoD HPCMP 計算機センター

Site	Status	LSF Cluster Name	System Name	System Configuration
ARL	Active	ari-sgi	Gargole	8CPU Origin2000
NAVO	Active	navo-sgi	odyssey	128 CPU Origin2000
NRL	Active	nrl-metag	Neo	128 CPU Origin3000
RTTC	Active	rttc-sgi	lucy, snoopy	2 x 32 CPU Origin2000
SMDC	Active	arc-sgi	arc220	80 CPU Origin2000
TARDEC	Active	tacom-sgi	dracos orion	32 CPU Origin2000 28 CPU Origin2000
WSMR	Active	wsmr-sgi	Deep-kimchi Deep-purple	4 CPU Origin2000 64 CPU Origin2000
AEDC	Inactive			
AFFTC	Inactive			
SSCSD	Inactive			

Confidential

Major Shared Resource Centers (MSRCs)
導入システムとアプリケーション

Organization	Systems	Applications
US Army Research Lab	10 x O2000/O3000 Total of 1256p *(includes 3 x 3800/256, total of 768)*	Structural Mechanics, Electromagnetics
Wright-Patterson AFB Aeronautical Systems Ctr	6 x O2000 Total of 512p	CFD, Electromagnetics
US Army Engineer and Development Center	3 x O2000/O3000 Total of 776p *(includes 3800/512)*	Structural Mechanics Climate Modeling
Naval Oceanographic Office, Stennis	2 x O2000 Total of 136p	CFD, Climate Modeling
Total	2680p	

Confidential

Distributed Centers (DCs)
導入システムとアプリケーション

Organization	Systems	Total CPUs	Applications
AAC (MS)	2 x Onyx2 O2000	68 20	CFD
AEDC (TN)	O2000	64	CFD
JNTF (CO)	Onyx2 O2000	88 64	FEA
NAWCAD (MD)	Onyx2	64	FEA, CEM, Acoustics
NAWCWD (CA)	2 x Onyx2	64	FEA, CEM, Acoustics
NRL (DC)	O2000	128	CFD, Climate, Chemistry
RTTC (AL)	2 x O2000	64	FEA
SMDC (AL)	8 x O2000	320	CFD, FEA
TARDEC (MI)	2 x Onyx2	88	FEA
WSMR (NM)	O2000	128	FEA
MSRC & DC	*48 systems*	*3840 cpus*	

リファレンス

- Foster, I. & C. Kesselman (eds.), The Grid: Blueprint for a New Computing Infrastructure, Morgan Kauffman Publishers, Inc.,
- The GLOBUS Project, http://www.globus.org
- Platform Computing, http://www.platform.com
- Global Grid Forum, http://www.gridforum.org
- New Productivity Initiative, http://www.newproductivity.org
- Department of Defense, United States of America, High Performance Computing Modernization Program, http://www.hpcmp.hpc.mil
- NASA Information Power Grid (IPG), http://www.ipg.nasa.gov

e-HPC.com

"必要な時にスーパーコンピューティング"

- SGI と CSCに共同出資
- グリッドによって、プロダクション環境でのスーパーコンピューティングを提供
- 512P Origin3800と 6 pipe Onyx などのシステム構成
- 現在の顧客: BNFL, BAE Systems, Pratt and Whitney, Merrill Lynch など
- 民間でのGridコンピューティングの展開

e-HPC.com ホームページ

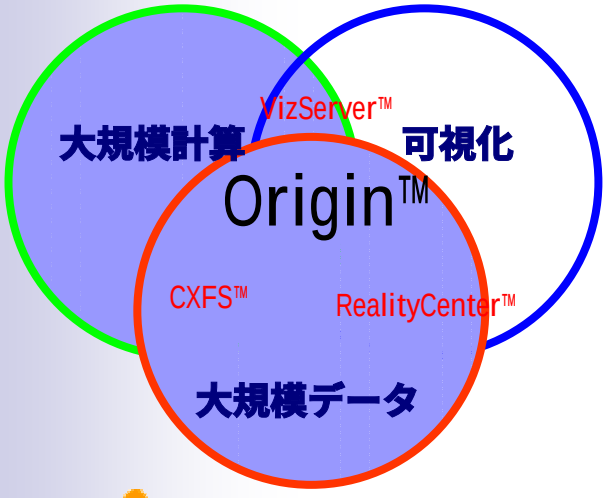
URL:
http://www.e-hpc.com/

SGIのグリッドコンピューティングへの取り組み

グリッドコンピューティング

- SGIの基盤技術
  - NUMAアーキテクチャ (大規模共有メモリ)
  - IRIX ACE (Advanced Cluster Environment)
  - Performance Co-Pilot
  - NUMAflex
  - VizServing
  - CXFS
  - …..

SGI™ プロダクトの特徴



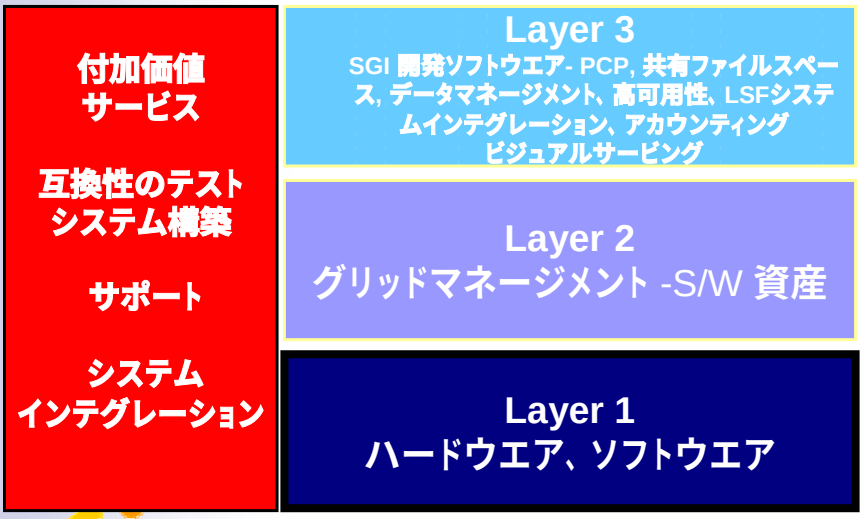
SGI – グリッドへの対応

- SGIが現在所有している知的所有権(IP)による技術は、グリッドコンピューティングに最適
- グリッドコンピューティングでは、SGIが目指すHPC、データマネージメント、可視化のリンクが重要
- オープンソースに欠けた部分を自社で補完
- システムのインテグレーション

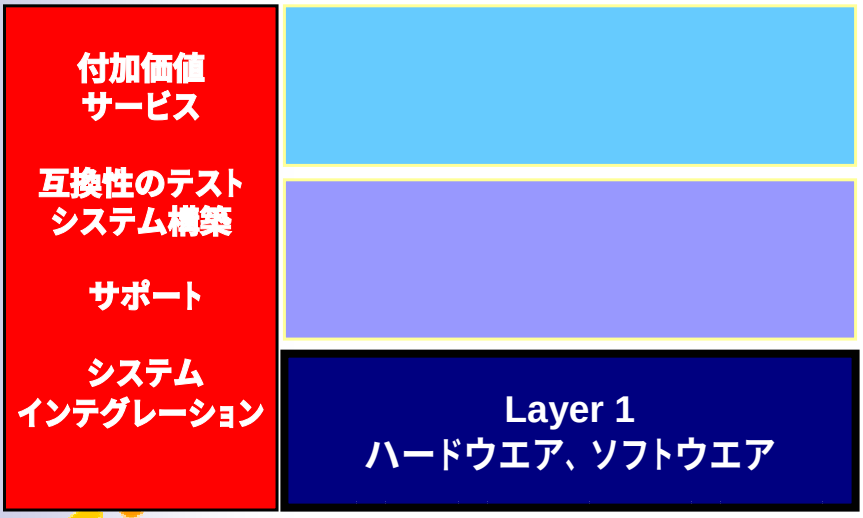
- ビジュアルエリアネットワークの展開

ビジュアルエリアネットワーキング

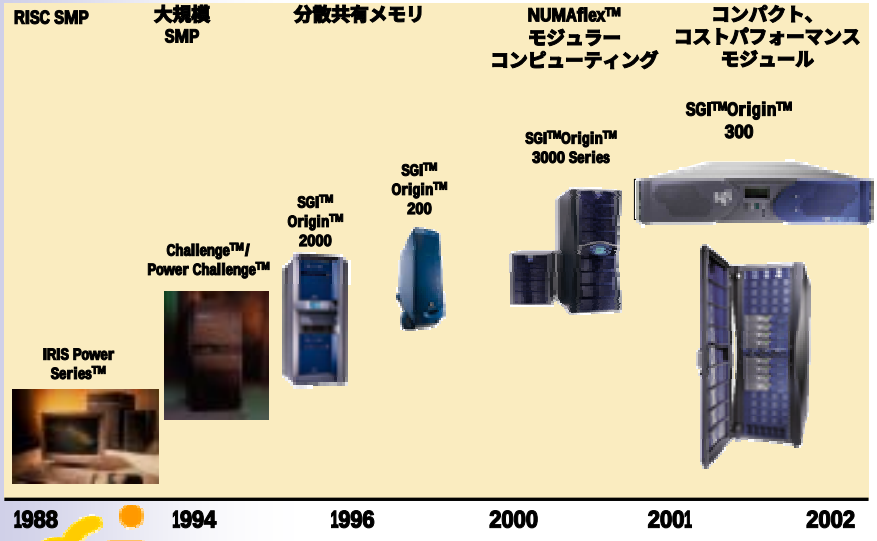
SGI – グリッドへの対応



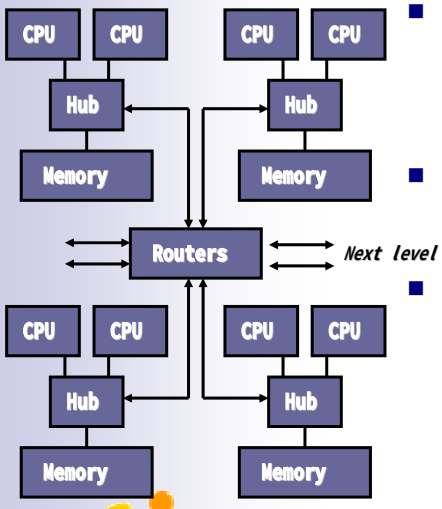
Layer 1　テクノロジー



SGIサーバーの歴史

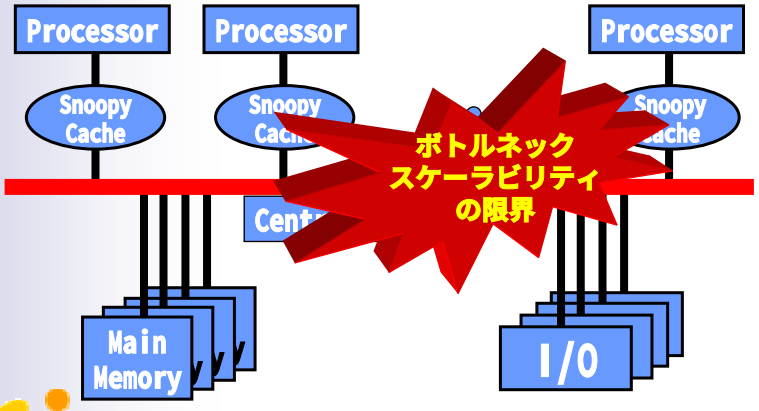


スケーラブルシステムアーキテクチャ

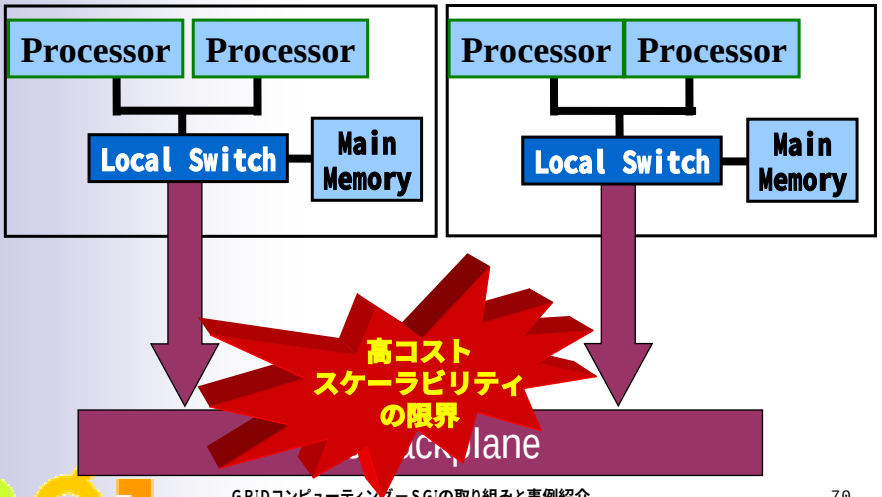


- ccNUMA アーキテクチャ
  - ローカルメモリは、cache-coherent Non-Uniform Memory アーキテクチャによって、システム全体で共有される
- Single-system image (SSI).
  - 単一の共有アドレス空間.
  - 単一のオペレーティングシステム.
- 非常に多くのプロセッサ構成までの拡張性
  - 高いネットワークバンド幅.
  - 低いリモートメモリアクセス遅延
  - 既に多くの512プロセッサシステムの稼動実績（日本でも2台）

典型的なバスベースSMP

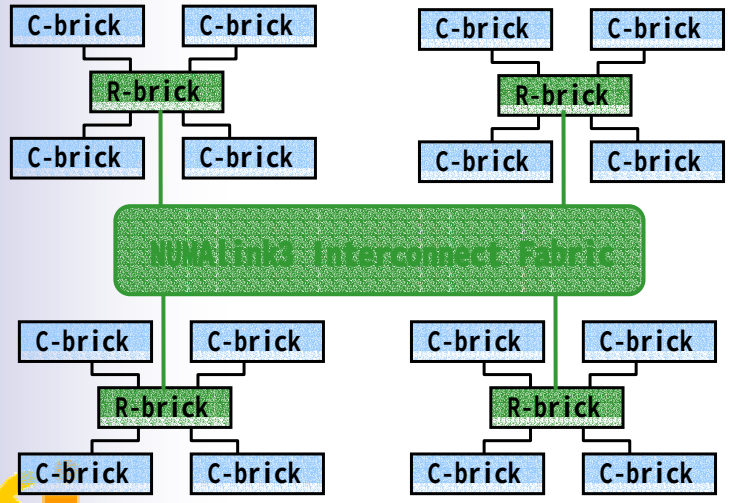


Bus & Switch の複合アーキテクチャ

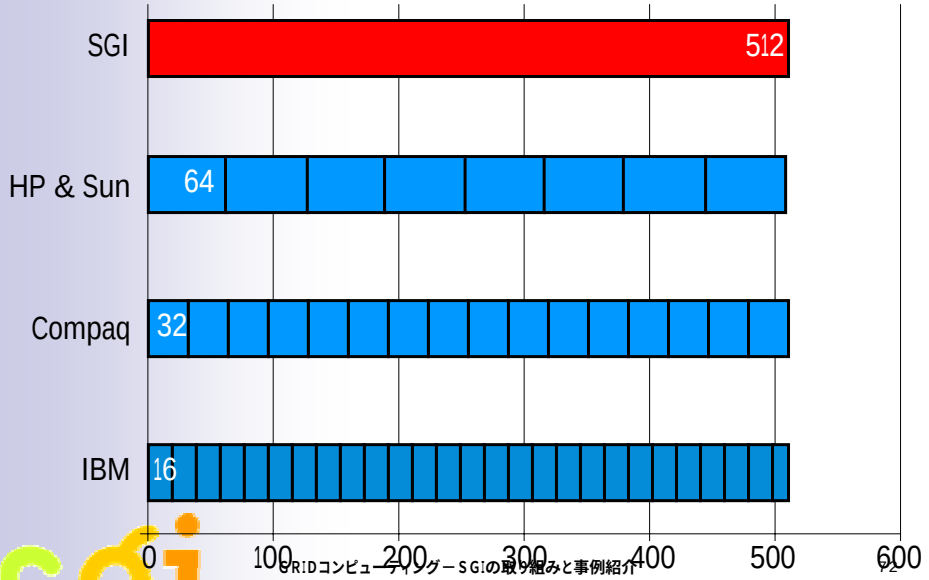


ccNUMA アーキテクチャ

SGI Origin 3000



共有メモリシステムのスケーリング



NUMAflex TM

モジューラコンピューティン

モジュール化の利点

- システム構成の柔軟性
  - 最適なシステム構成が可能
  - 必要に応じてシステムを拡張可能
  - SSI、パーティショニング、クラスタ
  - プロセッサの選択 (MIPS と IPF)
- システムの堅牢性 – 可用性
  - 冗長性をシステム内に実現
  - 高可用性を実現するシステム構成
- システムへの投資効果
  - システム内の構成を個々に技術革新可能
    - CPU, インターコネクト, IO, グラフィックスは単独にアップグレード可能
- システムの高い性能
  - スケーラビリティの拡張
    - Origin 3000では512CPUまでのスケーラビリティ
    - 1000 CPUまで拡張可能なアーキテクチャ的な自由度
  - 最先端のコンパイラとツール

システム構成の柔軟性
システムの堅牢性
システムへの投資効率
システムの高い性能

NASA/Ames: 1024 CPU Origin 3000

CPUs

- 1024 (MIPS R12000)
- 400 MHz CPUs
- 800 MFLOP/s per CPU
- 819 GFLOPS total
- 8 MByte cache per CPU
- 8 GByte total Cache

Memory

- 256 GB main memory

Disk

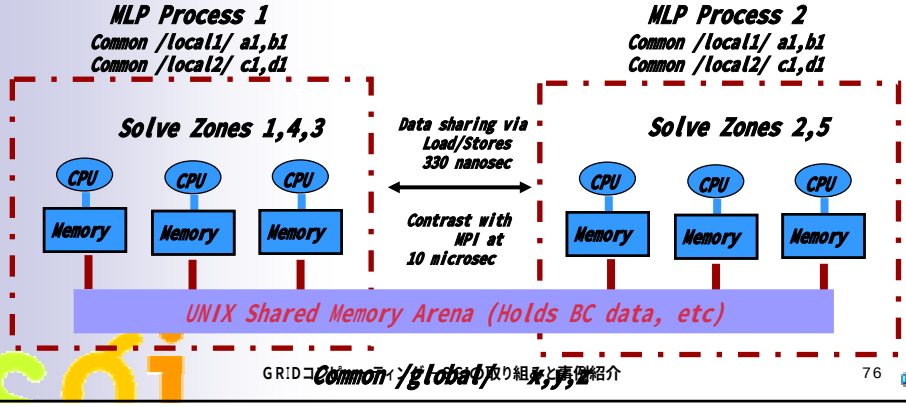
- 4 TB FC Raid disks

System Software

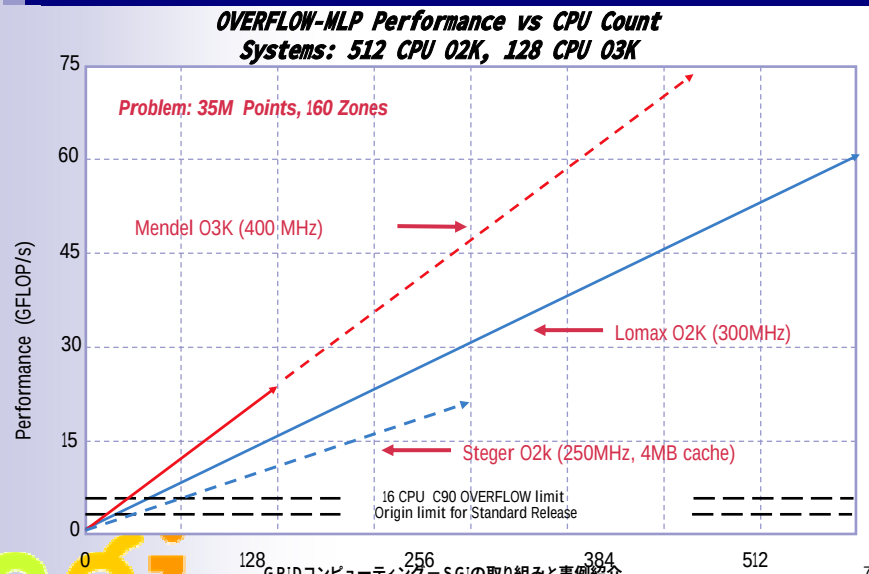
- OS single system image
- Single XFS File System
- Compiler parallel 1024 CPUs wide

共有メモリ MLP

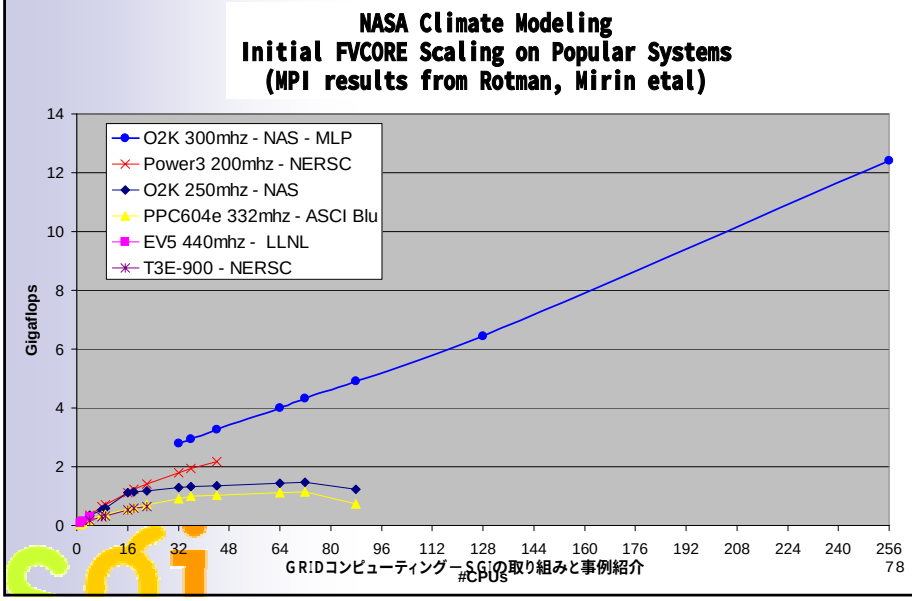
MLPプロセスがどのようにOrigin上で実行を行うかを示す。各プロセスは、細粒度での並列処理（ループレベルでの並列処理）を行うために少ないCPUを使用する。各MLPプロセスが使用するプロセッサ数はロードバランスを調整するために、可変である。このような動的なロードバランスの実装が可能なことは、共有メモリの利点の一つである。MLPで使用する共有メモリアリーナは、他のMLPプロセスで共有され、調節的なアクセスが可能となる。



Origin3000スケーラビリティ



Origin3000スケーラビリティ



大規模共有メモリシステムの利点

- 運用管理
  - 複数の計算機で構成されるクラスタの場合、すべてのノードに、リソースマネージャなどの資源管理ソフトの導入が必要で、その負担は大きい
  - 共有メモリでは、シングルノードにインストールすれば、システム全体のリソースの管理が行える
- 負荷分散
  - 動的にジョブの使用するプロセッサを変動することも可能
  - ロードバランスのマネージメントが容易
- 通信性能
  - 共有メモリでは、通信時のレイテンシを低く抑えることが可能であり、プログラミングを容易にする

新しいモジュール製品

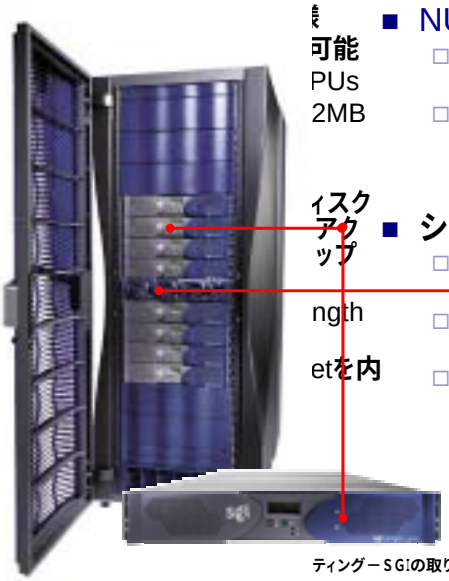
SGI™ Origin™ 300：よりコンパクトな筐体

- SGI Origin 300 ベースサーバ
  - 2 又は 4 MIPS® CPUs, 最大 4GB, を 3.5インチの筐体に実装
  - NUMAlinkによる高いスケーラビリティの実現
- PCIモジュール
  - 12 の64-bit, 66MHz PCI スロットを追加
  - NUMAlink ポートに接続可能
- SGI™ Total Performance 900 (TP900) ストレージシステム
  - 8台のUltra160 SCSI JBOD ドライ
  - 最大 584GB を 3.5インチの筐体に
- NUMAlink モジュール
  - 8つのNUMAlink ポート3．5インチの筐体に
  - 最大 32 CPUs 又は 56 PCI スロットの構成をSSIで実現

Tall rack (39U)
最大 76 CPUs
をシングルラックに搭載

Short rack (17U)
最大 32 CPUs

SGI™ Origin™ 300



可能
PUs
2MB

ィスク
アク
ップ

ngth

etを内

- NUMAlinkモジュール
  - NUMAlink : 高バンド幅/低レイテンシを実現
  - Origin3000のRouter(R-Brick)を使用し、最大32プロセッサまでの共有メモリを実現
- システム構成
  - Short(17U) rack : 2-16P 構成
  - Tall (39U) rack : 2-32P 構成
  - 高い価格/性能比と共有メモリの利点の両立

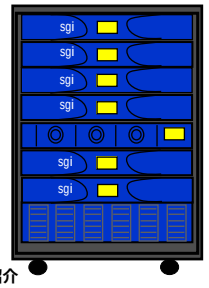
SGI™ Origin™ 300

32p Origin300
39U Origin3000 Rack
16 pci slots...8 busses
16 internal disk slots
8 external SCSI busses
32GB MAX Memory
23U config

Origin30(
Origin30(
Origin30(
Origin30(
Rou
Origin30(
Origin30(
Origin30(
Origin30(
Power I

24p Origin300
17U Origin3000 Rack
2 pci slots...6 PCI busses
12 internal disk slots
6 ext scsi bus
24GB MAX Memory
13U config

Router
ay



Origin300 4P
Origin300 4P
Origin300 4P
Origin300 4P
Origin300 4P
Origin300 4P

Layer 2 テクノロジー

付加価値
サービス

互換性のテスト
システム構築

サポート

システム
インテグレーション

Layer 2
グリッドマネージメント -S/W 資産

Layer 2 - グリッドマネージメント -S/W 資産

- システム構築のための要素技術とサービス
  - ユーザ認証
  - 計算資源管理と情報サービス
  - システムパフォーマンスモニター
  - グローバルコンピューティング用MPI (MPICHとベンダー提供MPIライブラリの拡張)
  - ファイル転送
  - データ管理
  - インターフェイス

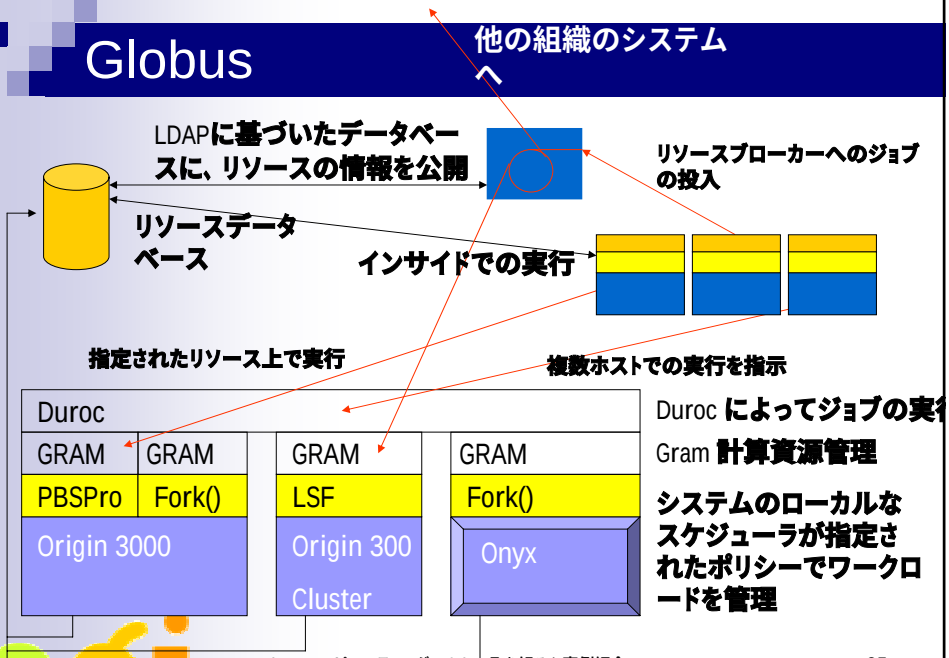
➔Globus (Globus Metacomputing Toolkit)

Globusとは？

- グローバルコンピューティングのためのサービスの集まり (ツールキット)

サービス	名称	機能
Resource Management	GRAM	リソースの割り当ておよびプロセス生成
Communication	Nexus	Unicast/Multicast 通信
Information	MDS	システムの構造および状態に関する情報へのアクセス
Security	GSI	Authenticationなどのセキュリティサービス
Health and status	HBM	システムの状況サービス
Remote data access	GASS	データへのリモートアクセスサービス
Executable management	GEM	実行ファイルの構築、キャッシングおよび配置

Globus



IRIXでのGlobusサポート

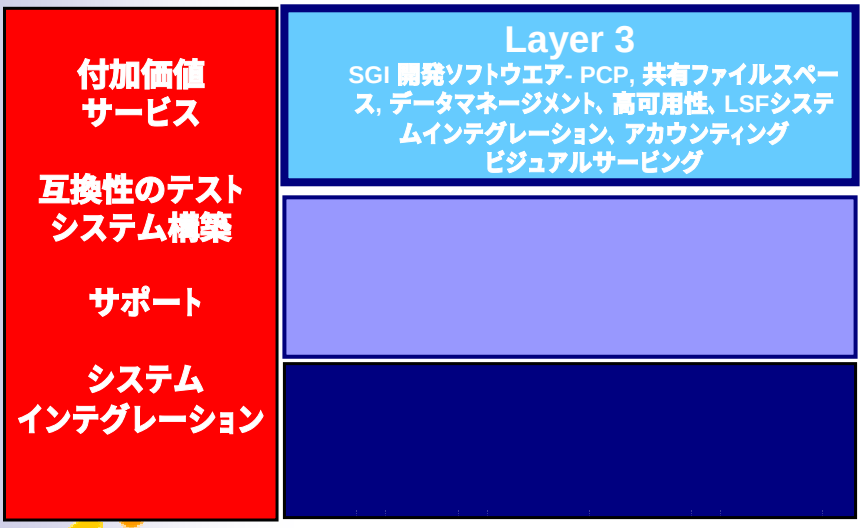
- IRIX上でのGlobusサポートの現状
  - Origin2000, Origin3000, Origin300, Octane…
  - LSF, PBS-Pro, OpenPBS…
  - IRIX message passing toolkit, MPICH-G2
  - OpenLDAP, Netscape server, Oracle...
  - ニュースリリース:
  - http://www.sgi.com/newsroom/3rd_party/111201_globus.html

今後の課題は?

- 現在のGlobusは、以下の機能に関してサポートがない
  - アカウンティング、課金処理
  - World-Wide でのスケジューラ
  - 商用アプリケーションの対応
  - 非グリッドノード (DHCPやWWW)からのアクセス

➡ Layer 3　テクノロジーがこの部分を補完

Layer 3　テクノロジー



SGIの持つ技術とIPの適用

- SGI ‘Layer 3’ 主要テクノロジー
  - Performance Co-Pilot
    - Globusが提供するシステムモニターよりも強力
  - チェックポイント･リスターと機能
    - ワークロードマネージメントとシステム予約などに効果的
  - IRIXシステムアカウンティング
  - システムマネージメントツール
  - CXFS – 高性能共有ファイルシステム
  - DMF - HSM ソリューションの提供
  - VizServer – ビジュアルサービング

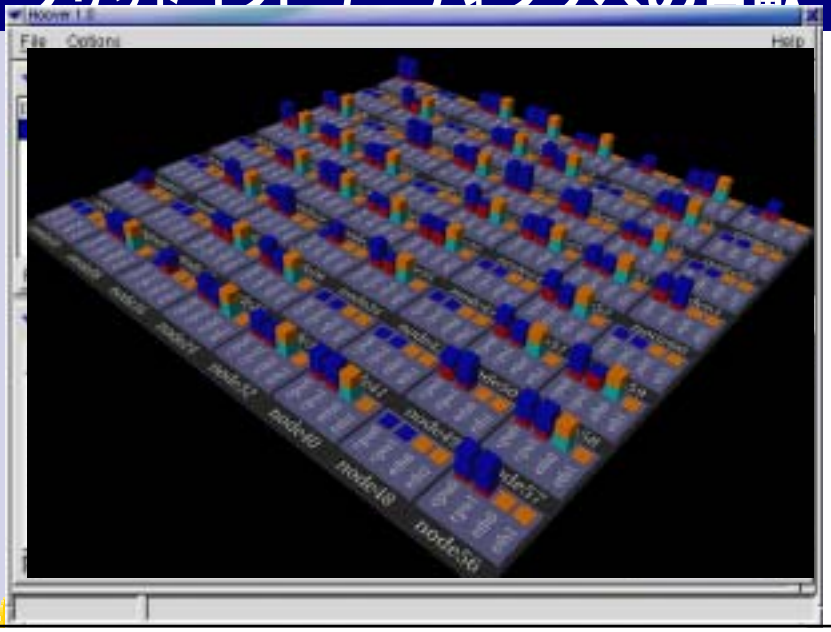
グリッドコンピューティングへの貢献

要求される機能	SGI が提供するツールと開発プラン
アカウント管理	• UID マッピング、ファイル共有など
アプリケーションと試験台	• メタコンピューティングに対応したグリッドを意識したアプリケーションとコラボレーション可視化ツール、リモートコンピューティング
グリッド情報サービス	• 負荷分散、クラスタ利用状況、コンソール管理機能など • PCP(Clusterviz)、SGIConsole(Hoover)など
リモートデータアクセス	• CXFS - クラスタ共有ファイルシステム • AFS - グローバルファイルシステム • リモート RAID モニターリング
セキュリティ	• IRIX が提供するアドバンスドセキュリティ機能、UID マッピング機能など
アドバンスドコラボレーション環境	• VizServer によるネットワーク経由での可視化の実現、Viz ツールによるリモートコラボレーション • ディスクトップビデオ会議ツール、メタコンピューティング環境

グリッドコンピューティングへの貢献

要求される機能	SGI が提供するツールと開発プラン
グリッドプロトコル	• Global Grid Forum の提案
アドバンスドプログラミング環境	• NASA AMES が提案しているマルチレベル並列化手法（MLP)など • Global Grid Forum が提供するモデルなど
グリッドパフォーマンスモニター	• クラスタパフォーマンスモニターツール • Infiniband などの新しいネットワーク技術の導入
スケジューリング	• PBS，LSF，NQE などが使用可能 • 負荷分散や分散リソース間での公平なリソースの分配
ユーザサービス	• アカウンティング、複数サイト間でのリソースの取引、公平なジョブの割り当ての確保

グリッドコンピューティングへの貢献



スケーラブルなデータと情報管理

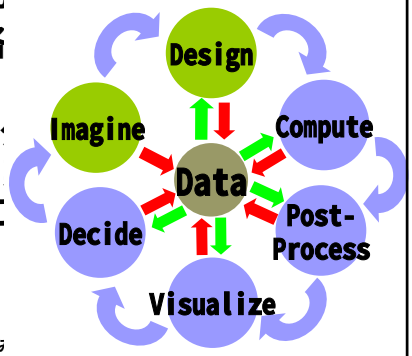
- データ生成　(計算ホスト) :
  - シングルシステムイメージ (SSI)サイズ
  - スループットとメッセージパッシングでのクラスタ
  - 大規模並列アプリケーションシステム
- データ移動:
  - 複数のギガバイト/秒のデータ転送パス
  - ギガバイト/秒の実効転送性能
- 情報管理:
  - 膨大なデータセット
  - データ間の相互依存
  - デザイナー間でのデータ共有

複合データマネージメント

- ハイパフォーマンスコンピューティング (HPC) とビジュアリゼーション (可視化) では、より多くのストレージが必要
- 膨大なビジュアリゼーションデータには、データの資産管理ツールが必要
- HPCシステムと可視化システムは、一体化している場合も、別システムとなる場合もある

理想的なデータの流れの提案

- 一般的な業務の流れを考えてみると…..:
  - コンセプトの段階から、最終的な製品化の段階まで、データは、業務の流れの中心に位置する
  - 情報は、多くのグループで共有され、データは、ホストコンピュータ間で移動
  - データセットのサイズは、各
    くなる
  - もし、各ステップ毎で、デー
    る必要がないのであれば、
    せ、本質的な問題の理解に



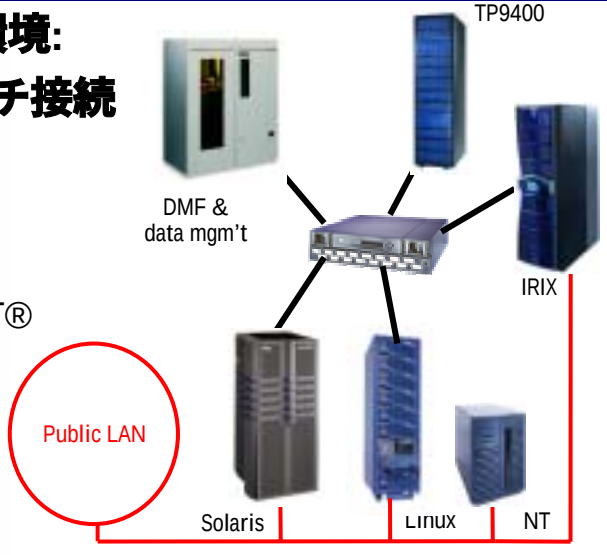
ストレージエリアネットワーク

- **異機種混在環境:**

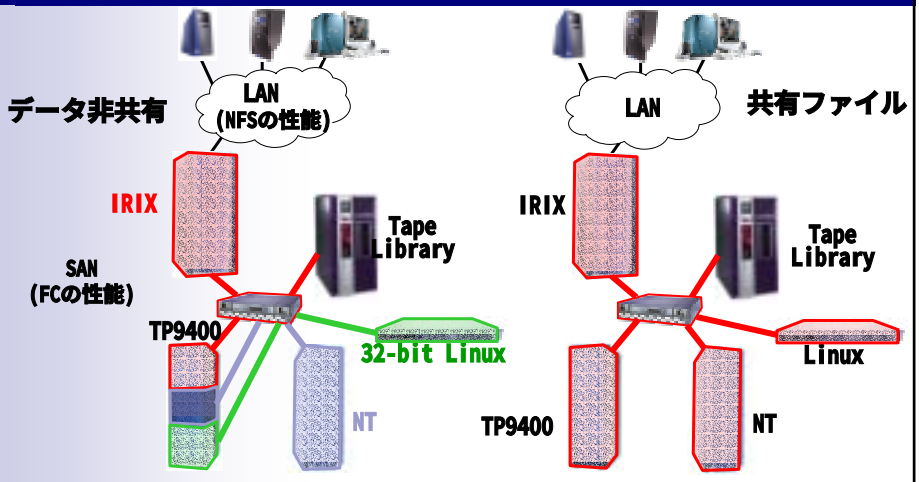
直接又はスイッチ接続

- IRIX®
- Linux®
- Solaris™
- Windows NT®

CXFSによる
ファイル共有
の実現



ストレージの共有

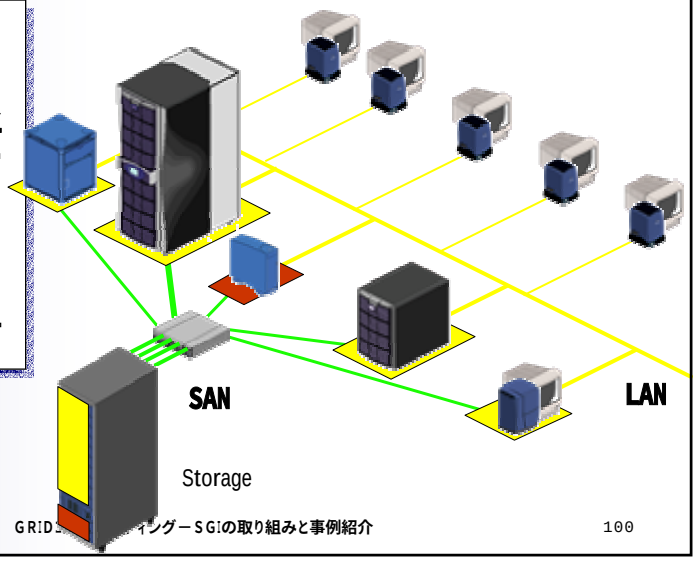


- データは複数ボリュームの複数のファイルシステムに分散
- データの共有には、ホスト間でのレプリケーションが必要
- データは、システム全体で共有
- レプリケーションは不要
- SGI XFSTM の性能)

CXFS™によるデータ共有の実現

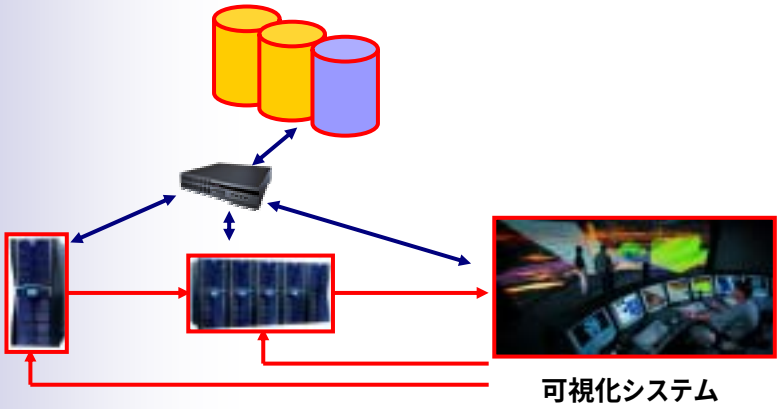
ユニークな特徴

- 全てのホストは、一つ以上のディスクアレイ上で、一つ以上のボリュームを共有
- 非常に高いモジュール化
- 高性能
- 真のファイル共有

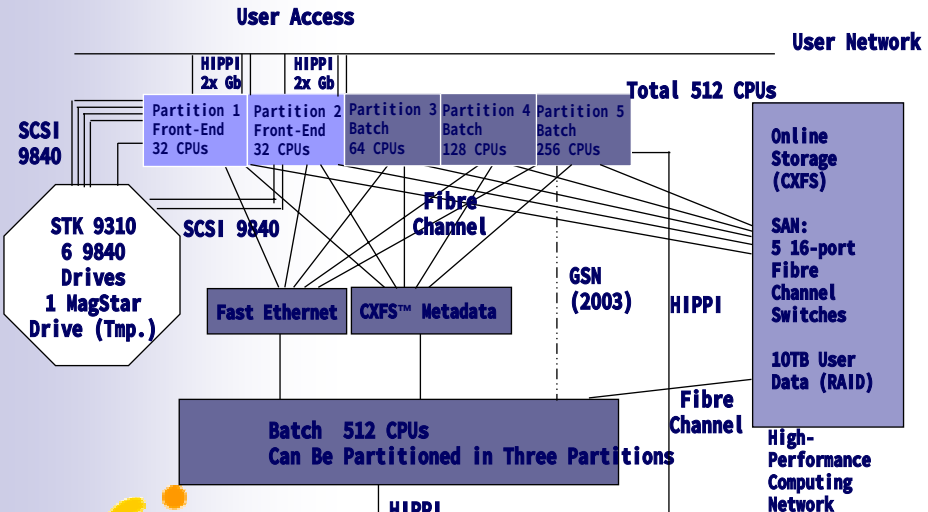


グリッドコンピューティング

- 全てのシステムは、SAN上で、同時に全てのデータに直接アクセス可能



大規模システムでの活用事例



まとめと今後の展開

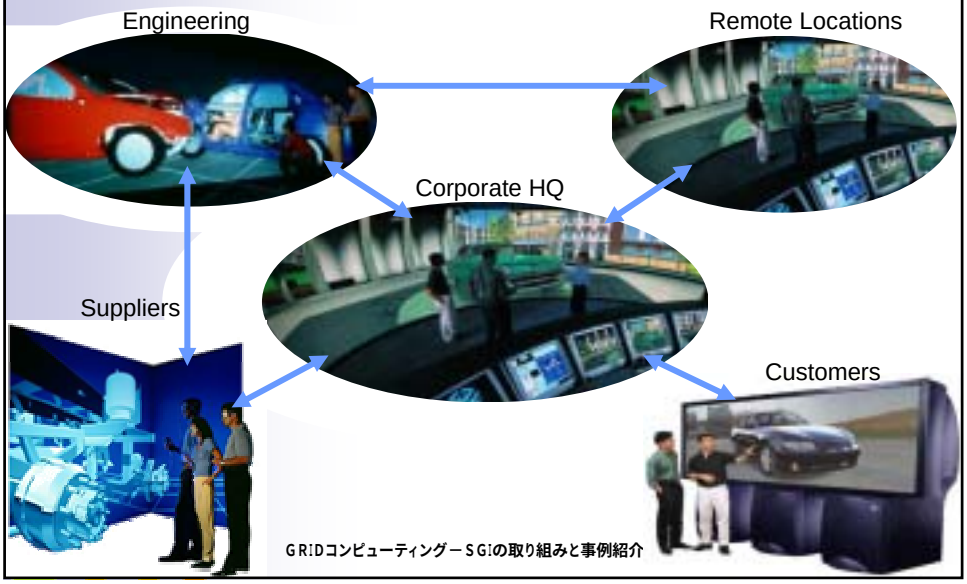
- グリッドコンピューティングは今後のハイパフォーマンスコンピューティングの重要な技術
- SGIの技術の活用で、より効率的なグリッドコンピューティングの展開が可能

- 今後の展開の可能性....
  - ビジュアルエリアネットワーキングへの展開
  - MDOなどの新技術

ビジュアルエリアネットワークング

- リアルタイムシュミレーションの可視化
  - リモートノード上で、可視化の実現
  - コントロールとAPIの提供
  - パラレル可視化ツールの提供

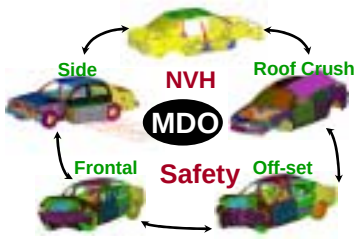
ビジュアルエリアネットワーキング



Ford と SGI による設計最適化プロジェクト

- 自動車車体のNVHと衝突解析における設計最適化プロジェクト
  - Ford社の390K自由度のbody-in-prime モデル
  - NVH：MSC.Nastran、衝突解析：RADIOSS
  - NVHと4つの衝突モードでの最軽量化
  - Response surface と Kriging 最適化
  - Ford社で6ヶ月かかったプロジェクトを4週間で終了

SGIのOrigin3800/256プロセッサで1.5日間で、妥当な設計計算を終了 – これは、PCでは約3年分のCPU時間に相当



Ford Motor
Scientific Research Labs

SGI
Industry HPC Development

SGI ……グリッドコンピューティング

SGI関連 情報リンク

- SGI™ Server Systems:
  - http://www.sgi.co.jp/servers/
- OpenGL VizServer:
  - http://www.sgi.com/software/vizserver/index.html
- CXFS:
  - http://www.sgi.co.jp/products/storage/software.html
- SGI™ Advanced Cluster Environment：
  - http://www.sgi.com/software/ace/irix.html
- SGI™ Onyx Systems:
  - http://www.sgi.co.jp/onyx3000/
- Visualization Systems：
  - http://www.sgi.com/visualization/