サイエンティフィック・システム研究会 2005 **年** 10 **月** 26 **日**
2005 **年度 科学技術計算分科会 資料**

マルチベンダー環境の運用事例　－北陸先端大を例として－

北陸先端科学技術大学院大学
情報科学センター
松澤　照男

[アブストラクト]
学内ネットワークおよびそれに接続された情報機器を情報環境と呼ぶ。情報環境の主システムではなく、一部として計算サーバ Cray XT-3(360 CPU, 2.88TB)、SGI Altix3700(128 CPU, 768GB)、Sun Fire 15K(32CPU, 32GB)、Fujitsu VPP 5000(2CPU,28GB)などが導入されている。それぞれ特徴のあるシステムであり、本学での導入のコンセプトおよびそれに伴う運用の考え方と実績などを報告したい。
[キーワード] 情報環境、計算サーバ、並列計算機群、導入、運用

1　はじめに

今年は本学の開学 15 周年記念にあたる。ネットワークとそれに接続された情報機器からなる情報環境は開学時からマルチベンダーで構築されており、情報科学センターはマルチベンダー環境を常態として運用をしている。今報告は、マルチベンダー環境の運用事例として、マルチベンダーで構成されている本学の情報環境の概要を紹介し、主として計算サーバの導入および運用の考え方を報告する。

2　情報環境

本学の情報環境は、本学の構成員（教職員、研究員、学生など）のすべてが世界最高水準の研究を組織的に推進するために必要となる、高度かつ先端的な情報環境を提供することを目的としている。そのために、情報科学センターが情報環境を一体的に運用し、全構成員に対して「一様かつレベルの高いサービス」を推進している。

情報環境の特徴は、

① 全構成員に一人一台のワークステーション、パーソナルコンピュータあるいはシン端末の提供。
② 利用者は情報環境を 24 時間、365 日利用可能。
③ 各種サーバ（ファイルサーバ、メールサーバなど）による情報（ファイル、メールなど）を集中管理。

④　1500 台以上のワークステーションや各種サーバなどを超高速ローカルエリアネットワークによる結合
⑤　各種超並列計算機群の提供
⑥　遠隔会議や遠隔教育のサポート
⑦　無線ネットワーク等様々な先端的ネットワーク
⑧　高速インターネットへの接続、日本におけるインターネットの重要な研究拠点
などがあげられる。

利用者にとって使い勝手の良い情報環境にすることはもちろんとして、利用者が作成したファイルを保全することが、情報科学センターの大きなミッションの一つである。そのために、
①　高速、大容量、高可用性のファイルサーバ
②　高速、高可用性、利便性の満たしたローカルエリアネットワーク（LAN）
を整備している。特に LAN は、基幹系を常用系と待機系の２系統（それぞれ 2Gbps）を設けており、フロアーネットワークからどちらかの系統を通りファイルサーバにアクセスできる。今年度中に基幹系を 10Gbps、フロアーネットワークを 1Gbps の広帯域ネットワークにする予定である。また、研究室、講義室、ゼミ室などに無線 LAN が整備され、利用者が移動してもローミングされ、ファイルサーバなどのアクセスが可能である。さらに、学外とは WIDE 10Gbps、SuperSINET 4Gbps、JGNII 10GBps で接続している。

３　計算サーバの変遷

本学には複数台の超並列計算機を導入していることが特徴の一つである。これらの計算サーバは、
①　情報科学研究科における超並列アルゴリズムやデータバースの研究用
②　他研究科におけるシミュレーションの研究用
の２つの側面がある。

本学の計算サーバの開学からの変遷は以下の通りである。
①　超並列研究用システム
CM5 (Thinking Machine) → T3E (Cray) → T3E-1200 (Cray) → XT3 (Cray)
②　高度データベース処理研究用システム
nCUBE2 (nCUBE) → nCUBE3 (nCUBE) → RS/6000 SP (IBM) → Altix3700 (SGI)
③　超並列ソフトウェア研究用システム
GC/MPC-128(Parsytec) → GP7000 (Fujitsu) → SunFire 15K → SunFire V890
④　小規模計算サーバ（ベクトル計算機）
VP1100 (Fujitsu) → VX-E (Fujitsu) → VPP5000 (Fujitsu) → SX8 (NEC)
C3440 (Convex) → J90 (Cray) → SV1 (Cray) → 後継なし
⑤　クラスター計算機

HPC2000 (BestSystems) → HyperBlade (APPRO)

4　計算サーバ導入の基本的な考え方

新設大学（1990 年：設立、1992 年：学生受け入れ開始、本年度(2005 年度)：開学 15 周年）のため、情報環境予算（借料）が年次進行で配分された。したがって、毎年政府調達を行い、全情報環境の約４分の１ずつが更新されている。計算サーバを導入する際の基本的な考え方は以下の通りである。

① 導入のタイミングで最新の機器を選定（多少のリスクが伴う）。
② 教育的効果から可能な限りアーキテクチャの異なる計算機を導入する。
　⇒ ユーザが最適な計算機を選択する。
③ 情報環境の一部として導入 ⇒ 最終責任の所在を明確にする。
④ 更新に際には、前計算機の資産にとらわれない。

情報環境はもともとマルチベンダー指向であり、計算サーバも上で紹介したようにマルチベンダーである。ただし、計算サーバやファイルサーバなどを情報環境一式として調達しており、落札業者（富士通(株)）が最終的に責任をもつ体制を作っている。しかし、主要なマシン（計算サーバやファイルサーバ）などは、ベンダーと定期的に保守に関する打ち合わせをもっている。

5　主たる計算サーバの導入の経緯

主たる計算サーバとして、超並列処理研究用システム(Cray XT3) と高度データベース処理研究用システム (SGI Altix 3700) を導入した経緯を説明する。

5－1　超並列処理研究用システム

このシステムは以下の用途、目的で導入した。

① 超並列アルゴリズムの開発・検証用
② 並列化を陽に用いて、高い性能を引き出すことが目的
③ ユーザは基本的に MPI, SHMEM などが使えることを仮定
④ 流体力学、分子動力学など、数値シミュレーション系の研究が中心

そのために、以下の要求要件とした。

① 分散メモリマシンであること
② アルゴリズムの挙動が分かりやすいこと → 共有メモリシステムだと分析が難しい
③ MPI を前提としたシステムであること
④ 浮動小数点演算性能に優れること

この要件を満たすシステムとして、Cray XT3 を導入した。導入したマシンの基本的な構成は以下の通りである。

① 並列処理計算機

- 90 計算クラスタノード （360CPU）
- 4 CPU/計算クラスタノード
- 2.88TB 主記憶容量（32GB/計算クラスタノード、８GB/CPU）

② フロントエンド計算機
- 12 サービス・IO ノード
- 96GB　主記憶容量（8GB/ノード）

③ 二次記憶装置
- 4.7TB RAID 5 ファイバチャネルディスクアレイ装置

④ システム相互結合ネットワーク
- 4 x 12 x 8 3D トーラスネットワークトポロジー
- 491.52GB/s バイセクションバンド幅

５－２　高度データベース処理研究用システム

このシステムは、以下の用途、目的で導入した。

① 幅広い研究プラットホームとして導入
② 必ずしも並列処理が目的ではない → 処理を高速化するための手段として並列処理を用いる
③ ユーザは並列処理の専門家とは限らない
④ データベース処理を始め画像処理や暗号処理などに利用

そのために以下の要求要件とした。

① 複雑なプログラミングをせずに、大きなメモリ空間が利用できること
- 利用者は、計算能力は CPU 性能ではなく、メモリ容量でマシンを選択
- SMP または cc-NUMA だと有難い

② MPI 以外でも簡易な並列化が可能なこと →　性能は出ないかもしれないが･･･
③ 整数演算性能が優れていること

この要件を満たすものとして SGI Altix 3700 を導入した。導入したシステムの特徴は以下の通りである。

① C－ブリック 32 台を NUMAlink3(3.2GB/秒)で結合させた共有メモリ型の並列計算機
② C－ブリック（4 個の 64 ビット Intel(R) Itanium(R) 2 プロセッサ 1.3GHz、24GB のメモリ
- プライマリキャッシュ（L1）＝ 32KB　（レイテンシ 1 クロック）
- セカンダリキャッシュ（L2）＝ 256KB　（レイテンシ 5 クロック）

③ システム合計 128 個の CPU と 768GB のメモリ
④ 単一の Linux オペレーティング・システム
- 64Bit Linux（SGI Linux Environment with SGI ProPack)

⑤ 36GB x 4 = 144 GB の内臓ディスク　⇒　OS 用

ここでユーザのディスク領域は Altix 3700 の DAS として確保するのではなく、高速ネットワーク (1Gbps) 経由で高速ファイルサーバ (PRIMEPOWER 450 & ETERNUS3000, ２８TB,RAID5, Fujitsu)に NFS 接続した。

６　運用の基本的な方針

これらの計算サーバは以下のポリシーのもとで運用している。

① センターはユーザ利用に関して、出来る限り干渉しない

② ユーザグループの育成

- メーリングリストで情報交換
- 利用法についての質問 ⇒ ユーザで解決
- 処理時間などの計測で占有するときは他ユーザの了承を得る
- ディーラーへの質問はセンター経由で行う

③ 障害はセンターで対応

７　おわりに

JAIST の情報環境および計算サーバ（Cray XT3、SGI Altix 3700）の概要を紹介した。さらに、計算サーバの導入の基本的な考え方および導入の経緯、計算サーバの運用の基本的な考え方を紹介した。情報環境はもともとマルチベンダー環境であり、一括導入することにより最終責任を明確にすることと、個々のベンダーとは保守に関して定期的に打ち合わせ会をもつことが重要と考えている。また、利用者優先を運用の方針としている。

マルチベンダー環境の運用事例

北陸先端科学技術大学院大学
情報科学センター　松澤　照男

2005年10月26日

アウトライン

- 情報環境の概要
- 計算サーバの変遷
- 計算サーバの導入の基本的な考え方と経緯
- 代表的な計算サーバ
  - Cray XT3
  - SGI Altix3700　他
- 運用の基本的な考え方と稼動統計
- 姫野ベンチマーク
- おわりに

JAIST情報環境の目的

- 本学構成員（教職員や学生など）が世界最高水準の研究を組織的に推進するために必要となる、高度かつ先端的な情報環境を提供する。

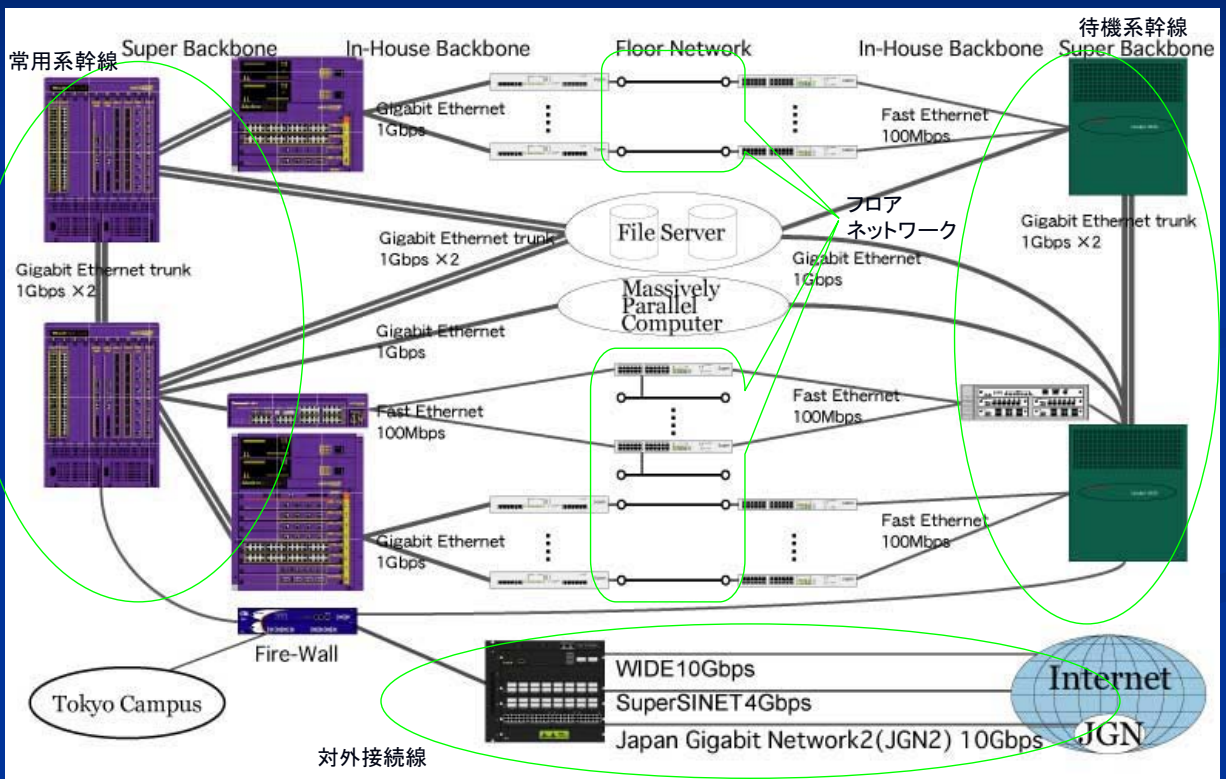
JAIST情報環境の特徴

- 日本におけるインターネットの重要な研究拠点
- 教職員、学生 各自一台（材料の学生はおよそ3人に一台）のワークステーション 24時間利用可能
- 1500台以上のワークステーションの超高速ローカルエリアネットワークによる結合
- 超並列計算機群へのアクセス
- 各種サーバによる、データや情報の集中管理
- コラボレーションルームにおける遠隔会議や遠隔教育の実施
- インターネット、ギガビットネットワーク への接続
- 無線ネットワーク等様々な先端的ネットワーク

本学のネットワーク

- Frontnet
  - 幹線、対外接続線、無線ネットワーク、フロアネットワーク（各研究室）から構成
  - 高速性、高可用性、利便性を満たすための機構
  - 情報科学センターによる集中管理

Frontnet（全体図）

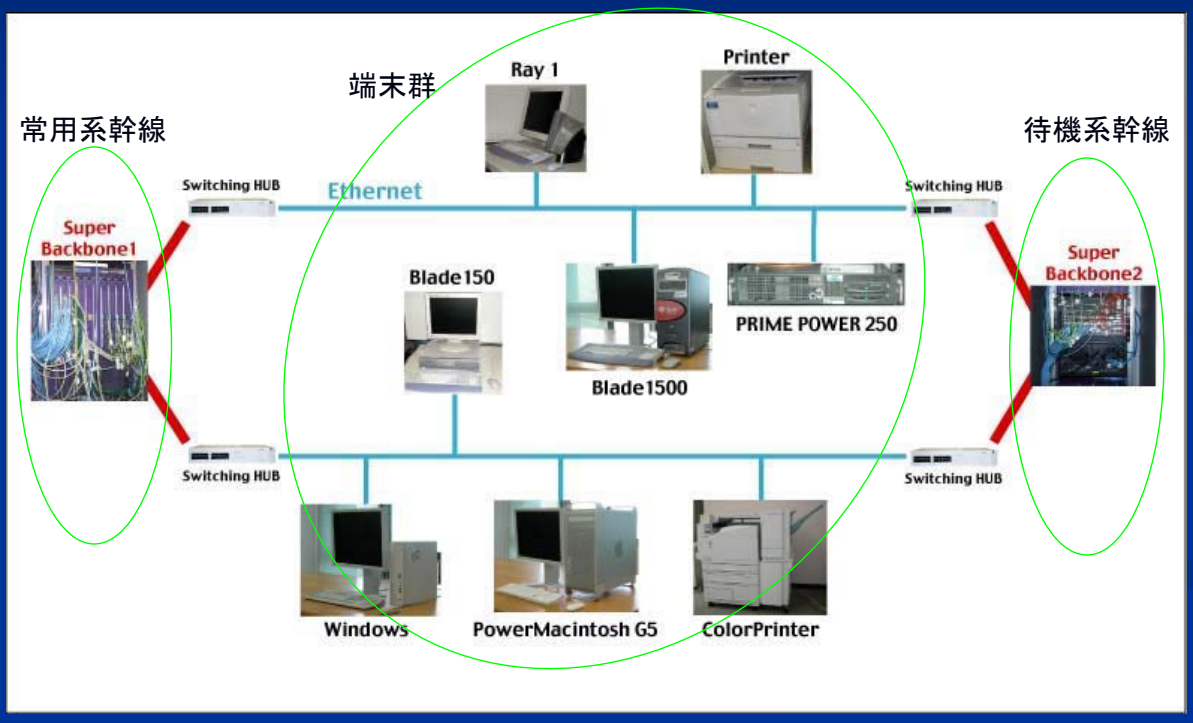


高速なネットワーク

- 幹線
  - 2Gbpsイーサネットによる高速光ファイバ接続
  - 常用系・待機系の2系統
  - ファイルサーバへ高速アクセス

  本年度中に10GbE化の予定
- 対外接続線
  - WIDE
  - SuperSINET
  - JGN2
- 無線ネットワーク
  - ノートパソコン等の機動力を活用

Floor Network



フロアネットワーク

- 100Mbpsイーサネットによってユーザの端末を接続
- 各フロアネットワークは常用系・待機系両方に接続され、トラブル時には自動で待機系の利用を開始

本年度中に1GbE化の予定

JAIST情報環境の特徴
超並列計算機群

- 情報科学での高効率な超並列処理研究
- 材料設計やゲノム情報処理、並列処理などに現在、活用されている
- シミュレーションオリエンテッド・エンジニアリングに対応するコンピューティング環境の整備が計画されつつある
  - ナノテクノロジー
  - ゲノムデータベースを対象とした検索、知識発見
  - 第1原理分子動力学シミュレーション
  - 金属の薄膜成長および相変態のシミュレーション

アウトライン

- 情報環境の概要
- 計算サーバの変遷
- 計算サーバの導入の基本的な考え方と経緯
- 代表的な計算サーバ
  - Cray XT3
  - SGI Altix3700　他
- 運用の基本的考え方と稼動統計
- 姫野ベンチマーク
- おわりに

超並列計算機

Cray
XT3

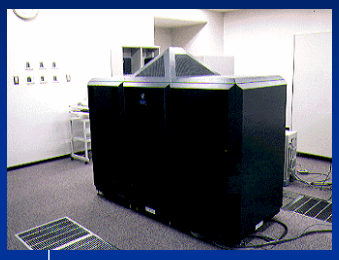
Cray
T3E-1200

Cray
T3E

Thinking Machines
CM-5

1992　1996　2000　2004　2008　2012

高度データベース処理計算機



SGI

Altix3700

IBM

RS/6000 SP

nCUBE

nCUBE-3

nCUBE

nCUBE2

1992 1996 2000 2004 2008 2012

超並列ソフトウェア研究用システム / PCクラスタシステム

APPRO

HyperBlade

Best Systems

HPC2000

Sun

SunFireV890

Sun

SunFire 15K

Fujitsu

GP7000S

Parsytec

GC/MPC-128

1992 1996 2000 2004 2008 2012

小規模計算サーバシステム

アウトライン

- 情報環境の概要
- 計算サーバの変遷
- 計算サーバの導入の基本的な考え方と経緯
- 代表的な計算サーバ
  - Cray XT3
  - SGI Altix3700　他
- 運用の基本的な考え方と稼動統計
- 姫野ベンチマーク
- おわりに

導入の基本的な考え方

- 導入のタイミングで最新の機器を選定（多少のリスクが伴う）。
- 教育的効果からなるべくアーキテクチャーの異なる計算機を導入する。

  ⇒　ユーザが最適な計算機が選択する。
- 情報環境の一部として導入　⇒　最終責任の所在を明確にする。
- 更新に際には、前計算機の資産にとらわれない。

主たる計算サーバの導入の経緯

- JAISTの超並列システム
  - 超並列処理研究システム
    - 超並列アルゴリズムの研究用
  - 高度データベース処理研究システム
    - 幅広い研究のプラットホームとして利用

超並列処理研究システム

- 超並列処理研究システム
  - 超並列アルゴリズムの開発・検証用
  - 並列化を陽に用いて，高い性能を引き出すことが目的
  - ユーザーは，基本的にMPI, SHMEMなどが使えることを仮定
  - 流体力学，分子動力学など，数値シミュレーション系の研究が主
- 要求要件
  - 分散メモリマシンであること
    - アルゴリズムの挙動が分かりやすいこと
      - 共有メモリシステムだと，分析が難しい
  - MPIを前提としたシステムであること.
  - 浮動小数点演算性能に優れること

→ **Cray XT3**

高度データベース処理研究システム

- 幅広い研究プラットホームとして導入
- 必ずしも並列処理が目的ではない.
  - 処理の高速化するための手段として，並列処理を用いる
- ユーザーは，並列処理の専門家とは限らない
- データベース処理を始め，画像処理や暗号処理などに利用

- 要求要件
  - 複雑なプログラムをせずに，大きなメモリ空間が利用できること.
    - 現代では，計算能力はCPU性能ではなく，メモリ容量で規定される
    - SMPまたはcc-NUMAだと有難い
  - MPI以外でも，簡易な並列化が可能なこと
    - 性能は出ないかもしれないが…
  - 整数演算性能が優れていること

→ **SGI Altix3700**

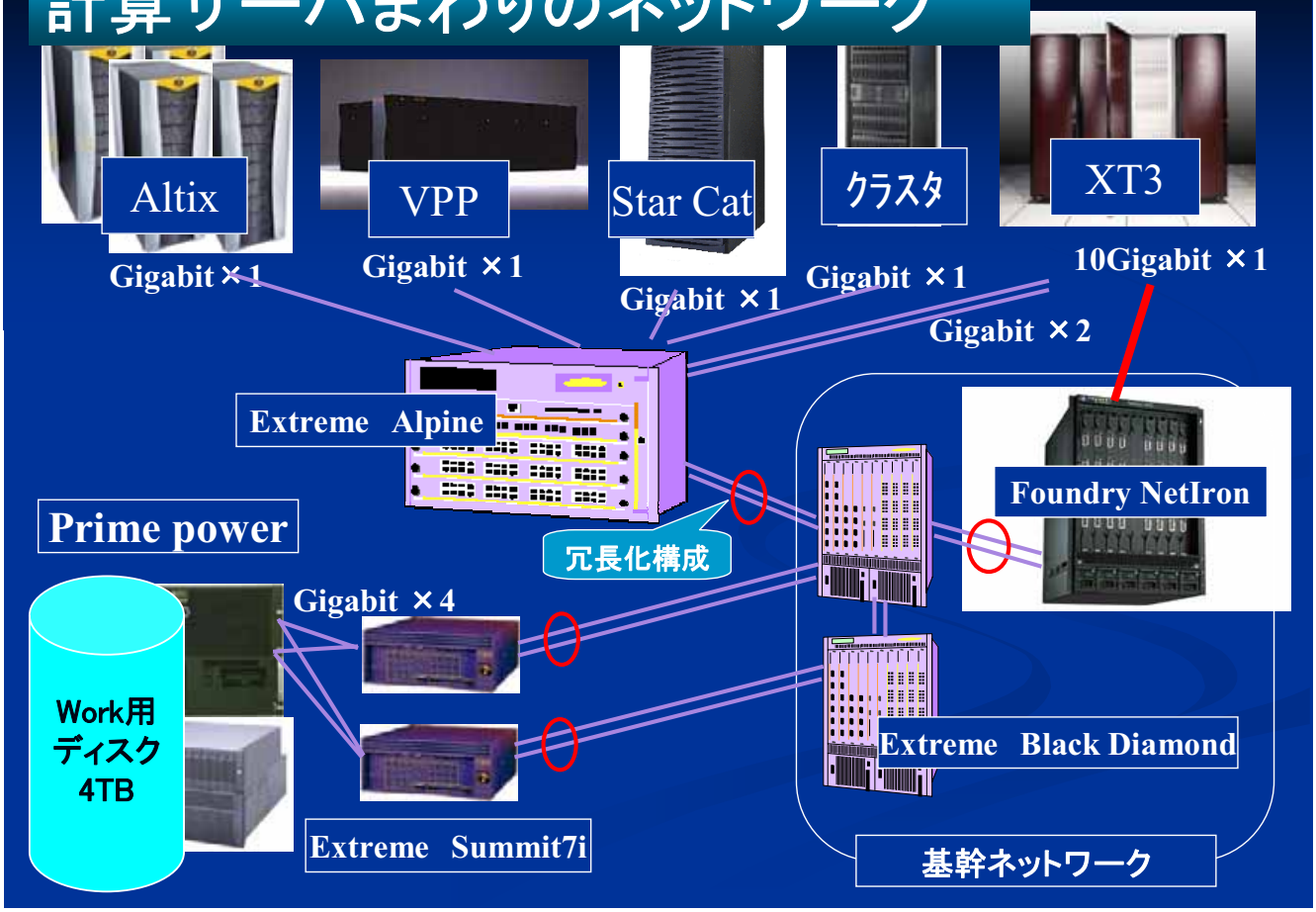
アウトライン

- 情報環境の概要
- 計算サーバの変遷
- 計算サーバの導入の基本的な考え方と経緯
- 代表的な計算サーバ
  - Cray XT3
  - SGI Altix3700　他
- 運用の基本的な考え方と稼動統計
- 姫野ベンチマーク
- おわりに

現在稼動中の主な計算サーバ

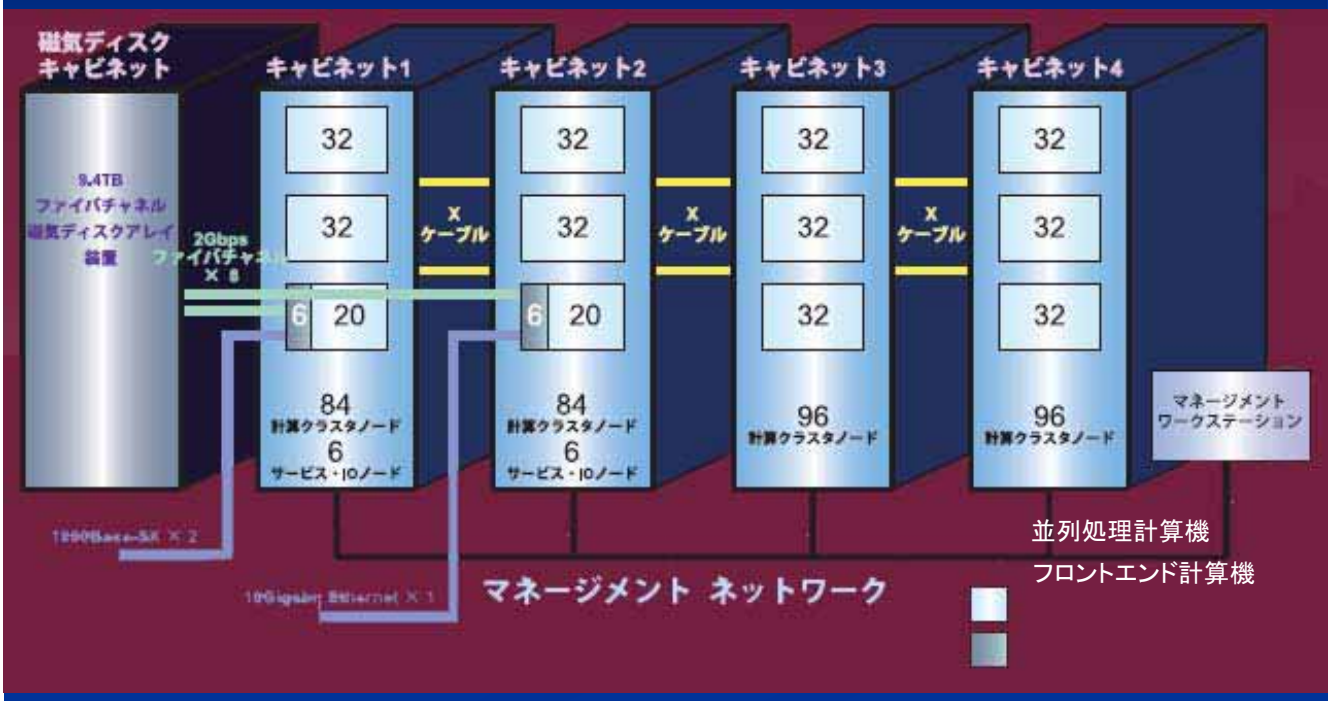
- Massively Parallel Computers
  - **SGI Altix3700**（ 128 cpus）
  - **CRAY　XT3**　（ 360 cpus）
  - **SUN Fire15K**（ 32 cpus）
- Computing Servers
  - **Fujitsu VPP5000**（ 2 cpus）
- PC cluster
  - **BestSystems HPC2000**（ 32cpus）

計算サーバまわりのネットワーク



Cray XT3

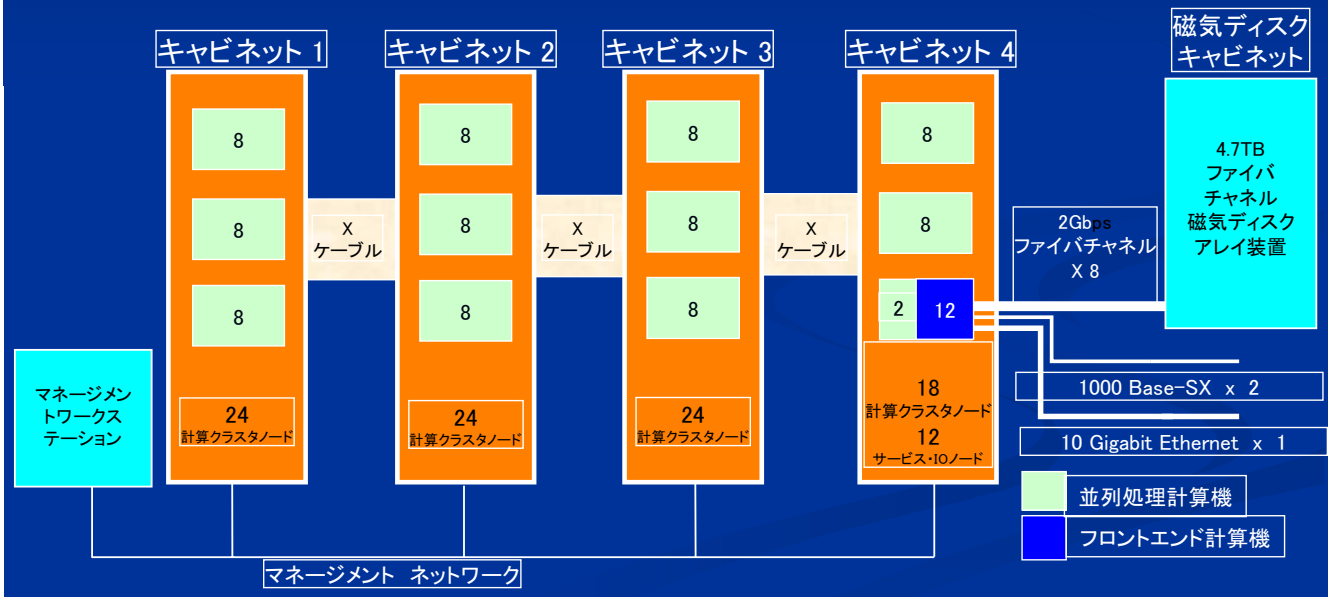
Cray XT3 システム構成



Cray XT3 システム構成

- 並列処理計算機
  - 90 計算クラスタノード（360CPU）
  - 4 CPU/計算クラスタノード
  - 2.88TB 主記憶容量（32GB/計算クラスタノード、8GB/CPU）
- フロントエンド計算機
  - 12 サービス・IOノード
  - 96GB　主記憶容量（8GB/ノード）
- 二次記憶装置
  - 4.7TB RAID 5 ファイバチャネルディスクアレイ装置
- システム相互結合ネットワーク
  - 4 x 12 x 8 3Dトーラスネットワークトポロジー
  - 491.52GB/sバイセクションバンド幅

Cray XT3 システム構成



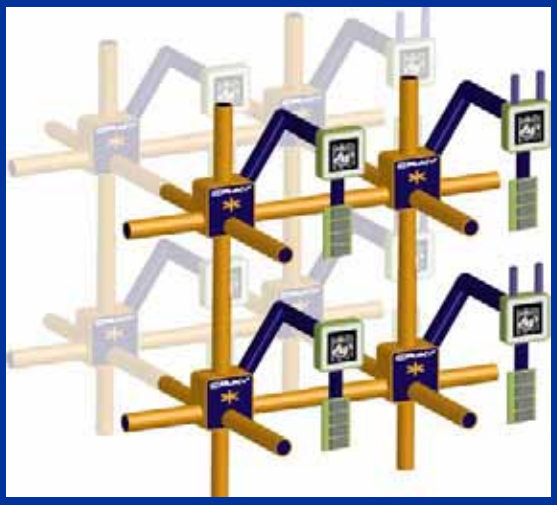
Cray XT3 プロセッサノード

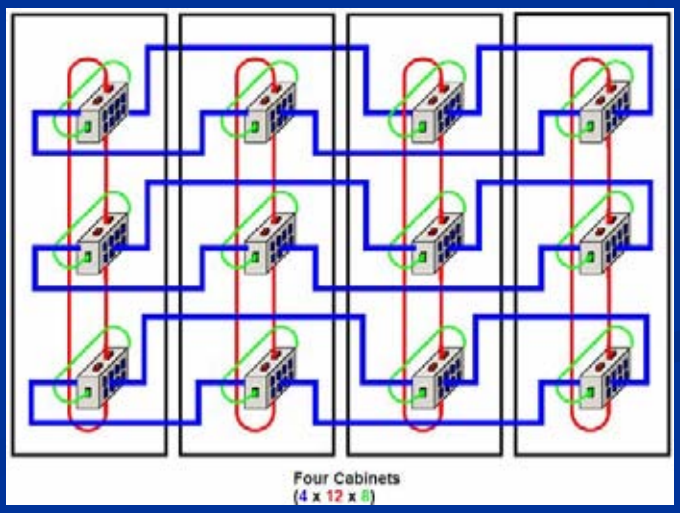
- 計算プロセッサ
  - 64ビット高性能プロセッサ（AMD社Opteron、4.8 Gflops）

    プライマリキャッシュ（L1）＝ 64KB （命令用）
    プライマリキャッシュ（L1）＝ 64KB （データ用）
    セカンダリキャッシュ（L2）＝ 1MB
- 内部ネットワーク・通信ハードウエア
  - 専用の高速ルータ（Cray SeaStar）
  - 高帯域（6方向計45.6 GB/秒）

Cray XT3 内部ネットワーク

- X,Y,Z全ての方向に環状接続を成す（3次元トーラス）
- どのプロセッサノード間でも最も効率の良い経路で
  データ通信を実現





SGI　Altix 3700

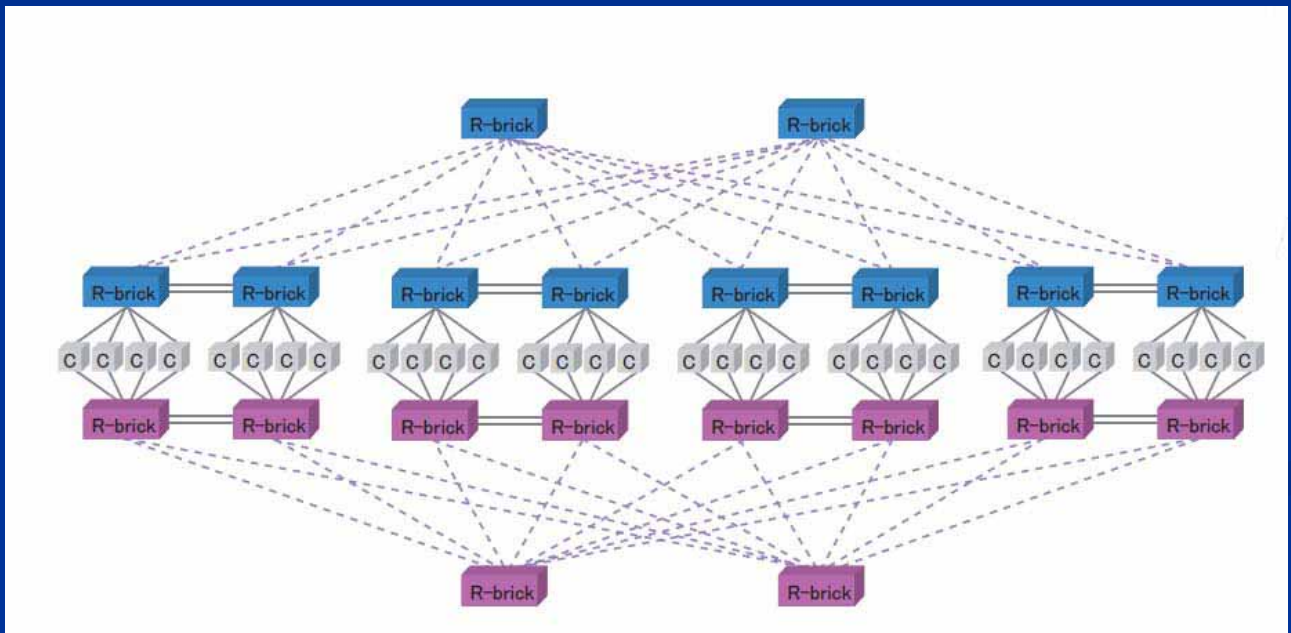
SGI Altix3700 の特徴

- Cーブリック32台をNUMAlink3(3.2GB/秒)で結合させた共有メモリ型の並列計算機
- Cーブリック（4個の64ビット Intel(R) Itanium(R) 2プロセッサ 1.3GHz、24GBのメモリ

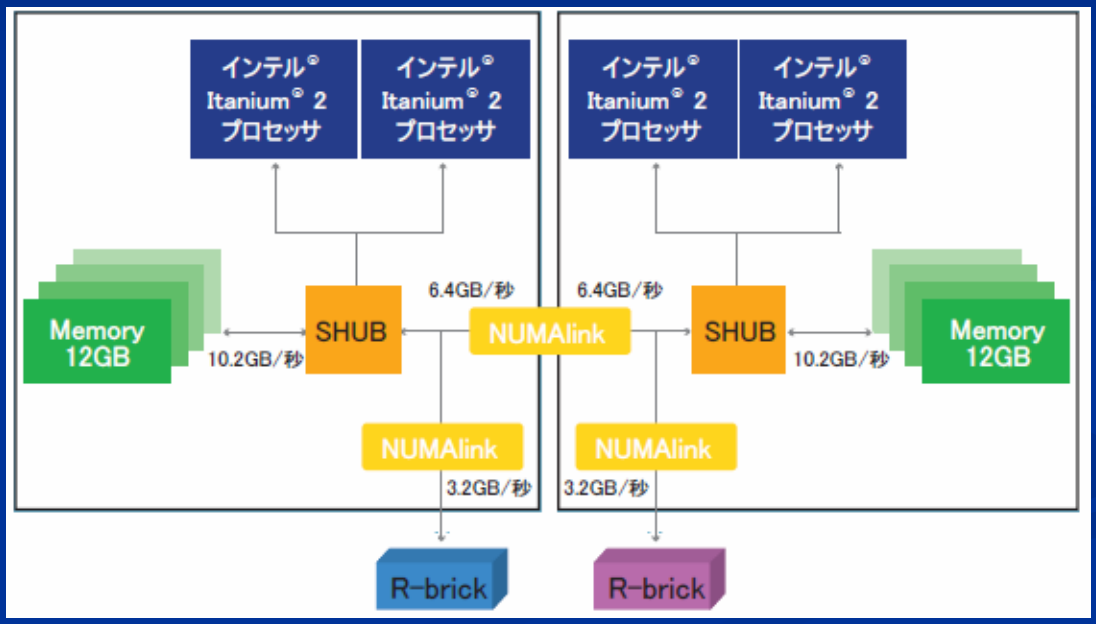
  プライマリキャッシュ（L1）＝　32KB　（レイテンシ1クロック）
  セカンダリキャッシュ（L2）＝ 256KB　（レイテンシ5クロック）
- システム合計128個のCPUと768GBのメモリ
- 単一のLinux オペレーティング・システム

  64Bit Linux（SGI Linux Environment with SGI ProPack)
- **36GB x 4 = 144 GB の内臓ディスク　⇒　OS用**

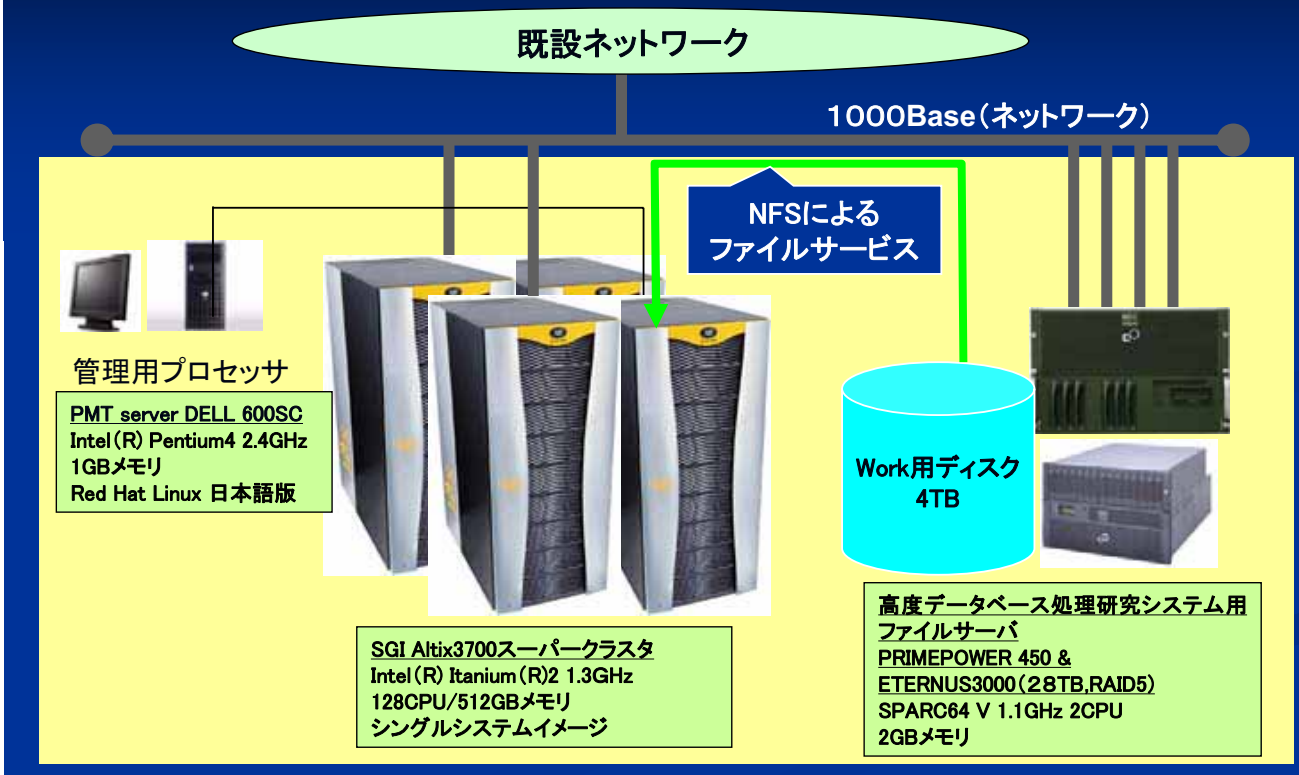
SGI Altix3700トポロジー



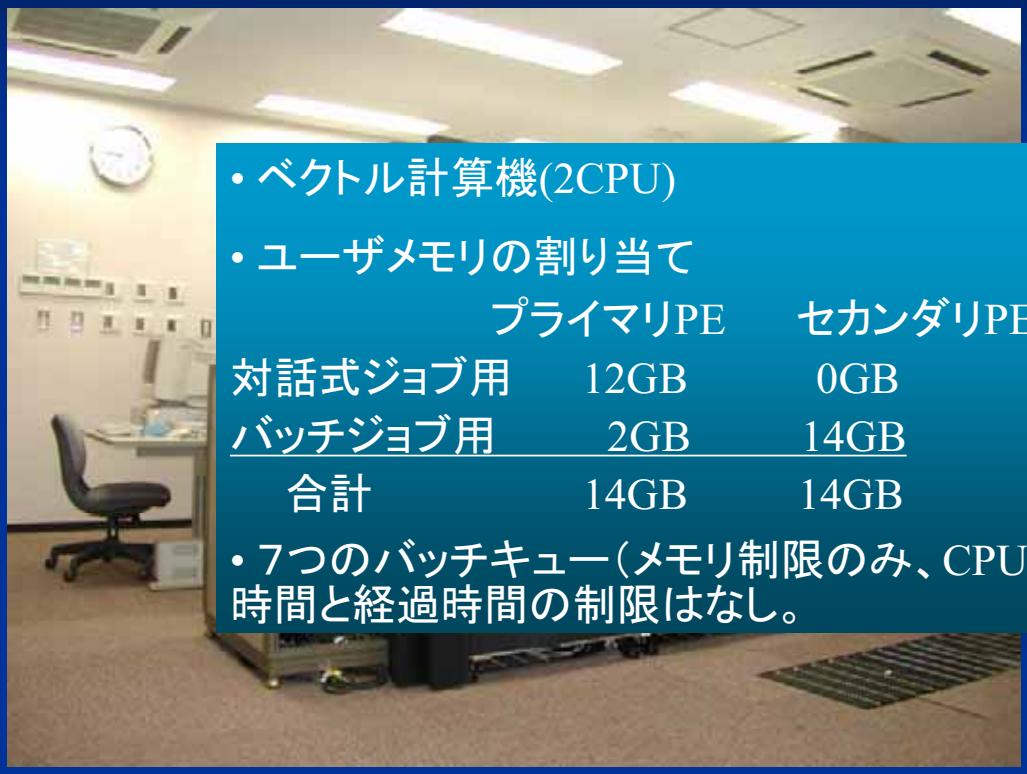
SGI Altix3700 Cブリック構成



SGI Altix3700 構成



計算サーバ(Fujitsu VPP5000)



- ベクトル計算機(2CPU)
- ユーザメモリの割り当て

	プライマリPE	セカンダリPE
対話式ジョブ用	12GB	0GB
バッチジョブ用	2GB	14GB
合計	14GB	14GB

- 7つのバッチキュー(メモリ制限のみ、CPU時間と経過時間の制限はなし。

計算サーバ(SunFire 15K)

- メモリー共有型計算機
- Ultra5バイナリ互換な超並列システム
- UltraSparc III Cu(900MHz)×32CPU, 32GB memory

PCクラスター(Best Systems HPC2000)

- 32CPU
- 各CPU
  Pentium3 1GHz
  512MB memory
  40GB disk,
- Myrinet2000 と Gigabit Ethernet

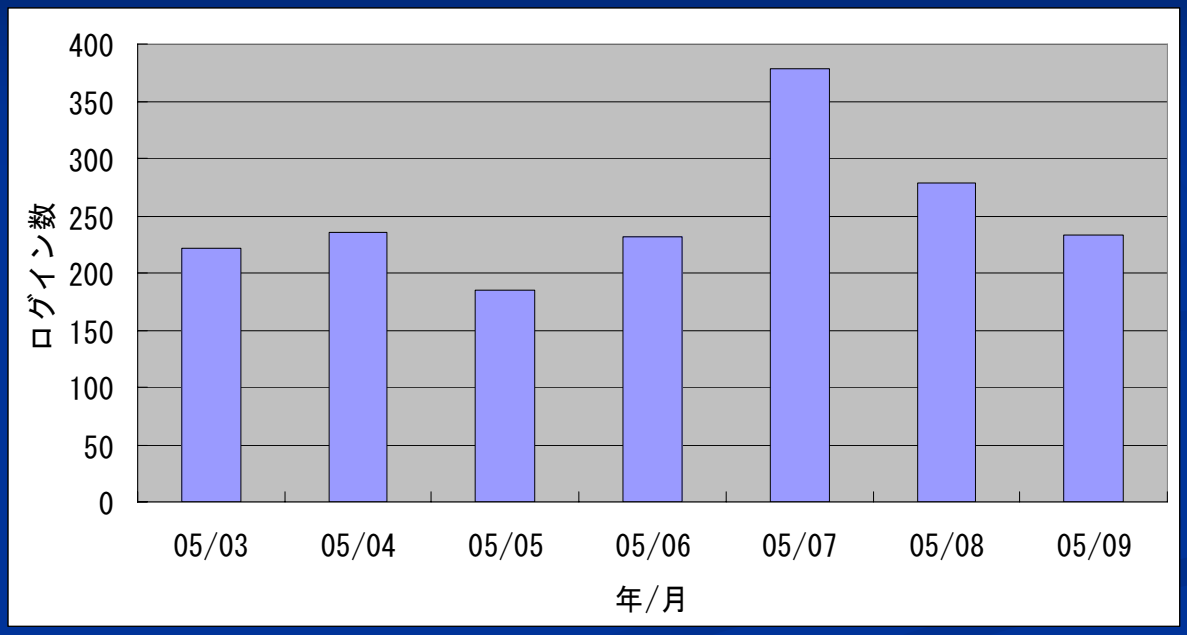
アウトライン

- 情報環境の概要
- 計算サーバの変遷
- 計算サーバの導入の基本的な考え方と経緯
- 代表的な計算サーバ
  - Cray XT3
  - SGI Altix3700
- 運用の基本的考え方および稼動統計
- 姫野ベンチマーク
- おわりに

運用の基本的な考え方

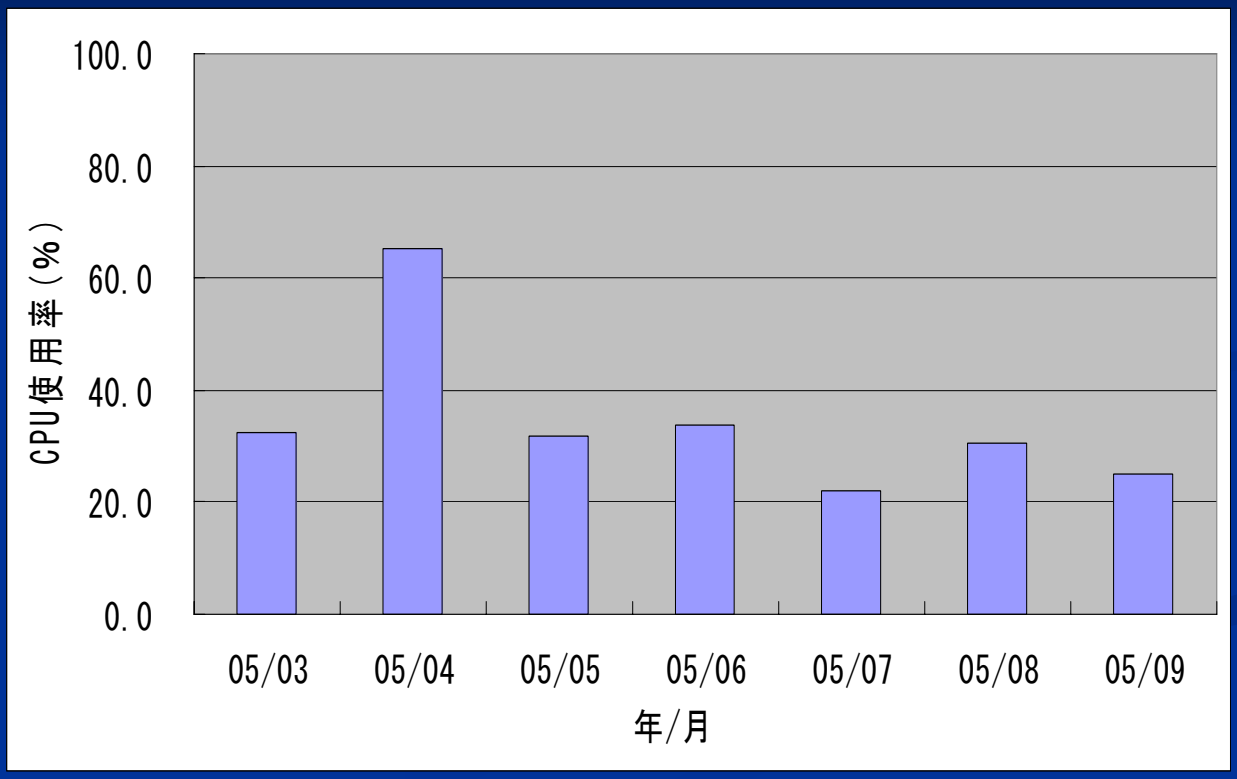
- センターはユーザ利用に関してなるべく干渉しない
- ユーザグループの育成
  - メーリングリストで情報交換
  - 利用法についての質問 ⇒ ユーザで解決
  - 処理時間などの計測で占有するときは他ユーザの了承を得る
  - ディーラーへの質問はセンター経由で行う
- 障害はセンターで対応

SGI Altix3700 の月別ログイン数

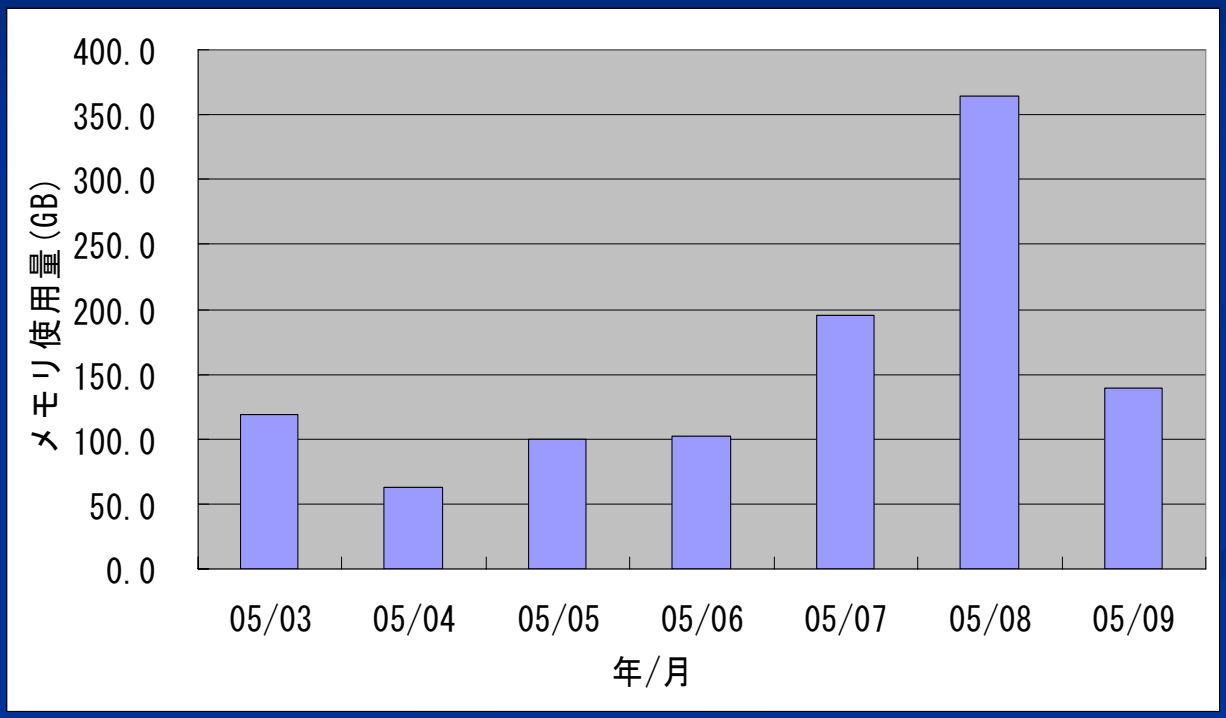


（1利用者のログインを1日1回のみカウント）

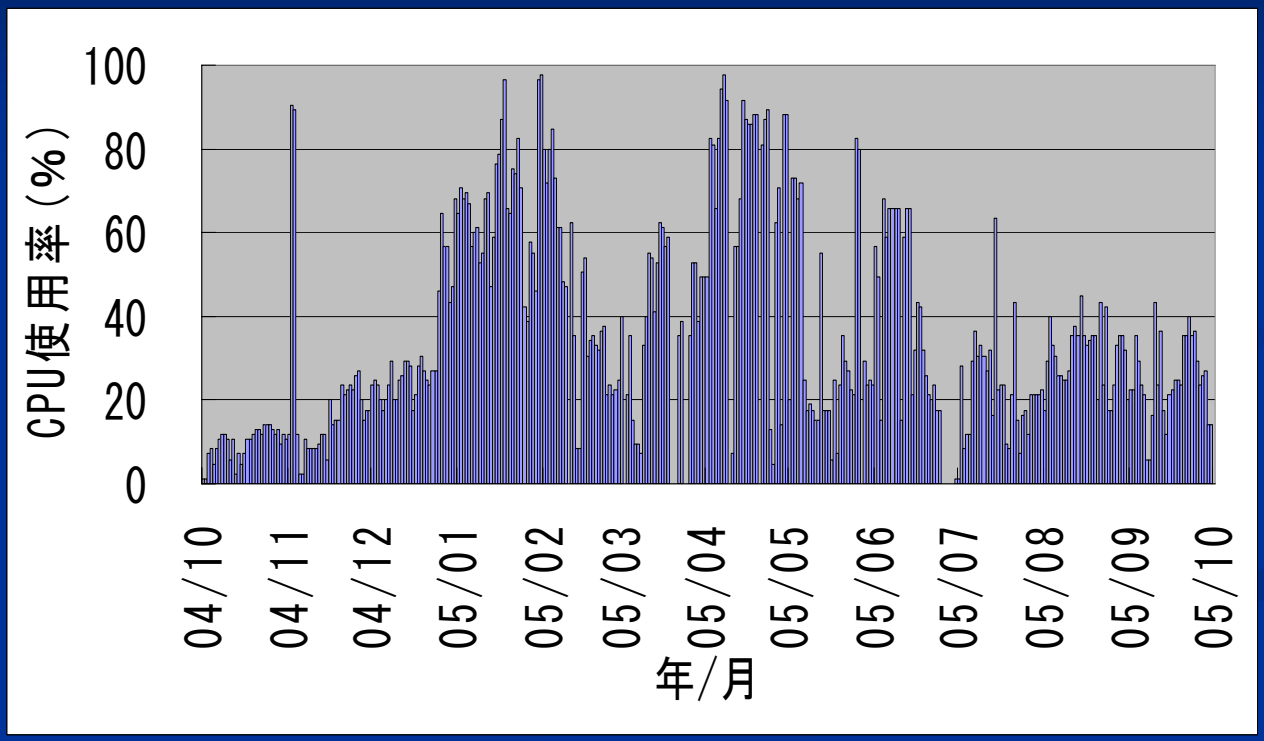
SGI Altix3700 の月別平均CPU使用率



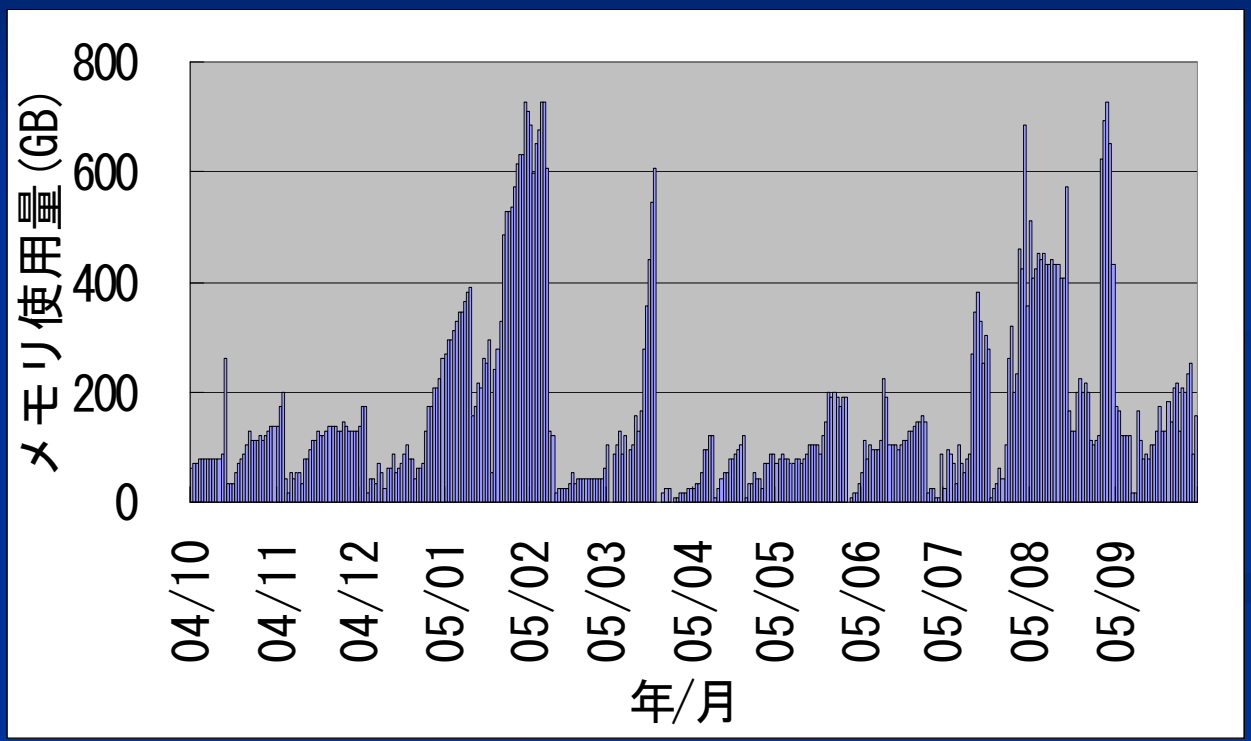
SGI Altix3700 の月別平均メモリ使用量



SGI Altix3700 の日別平均CPU使用率



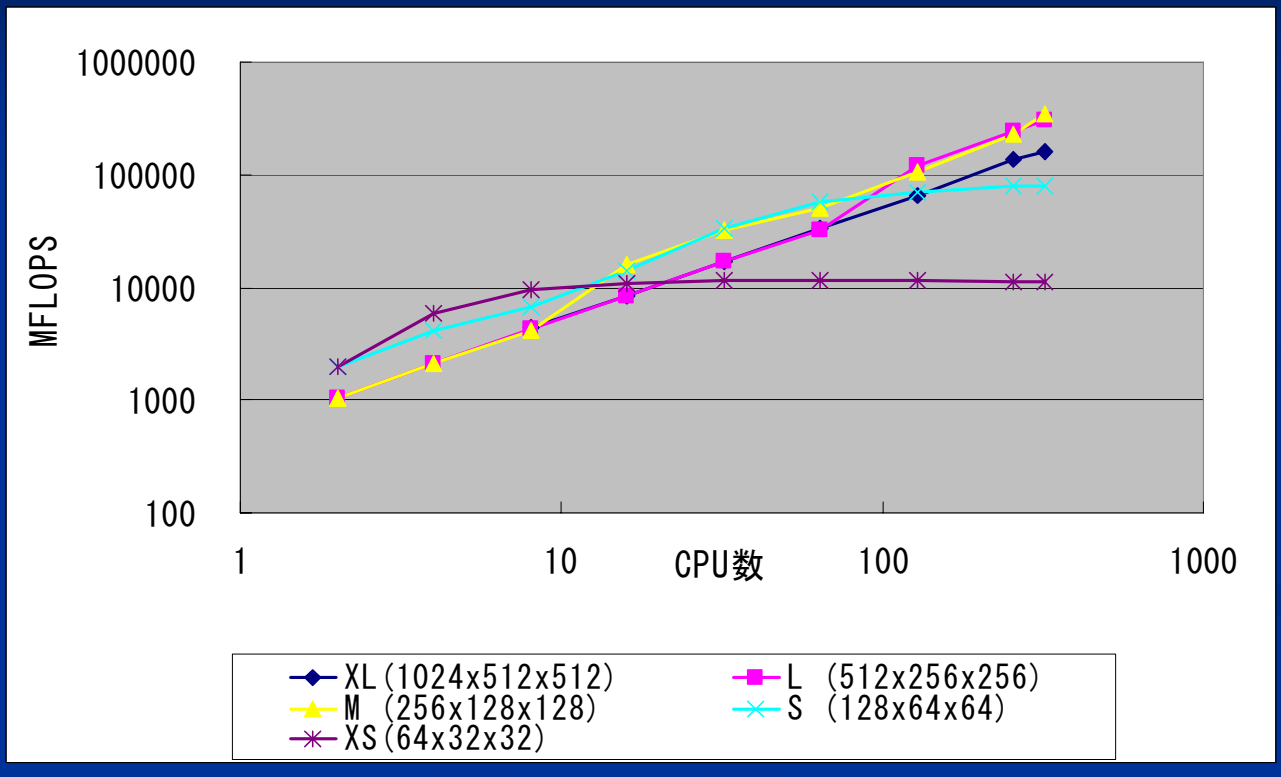
SGI Altix3700 の日別平均メモリ使用量



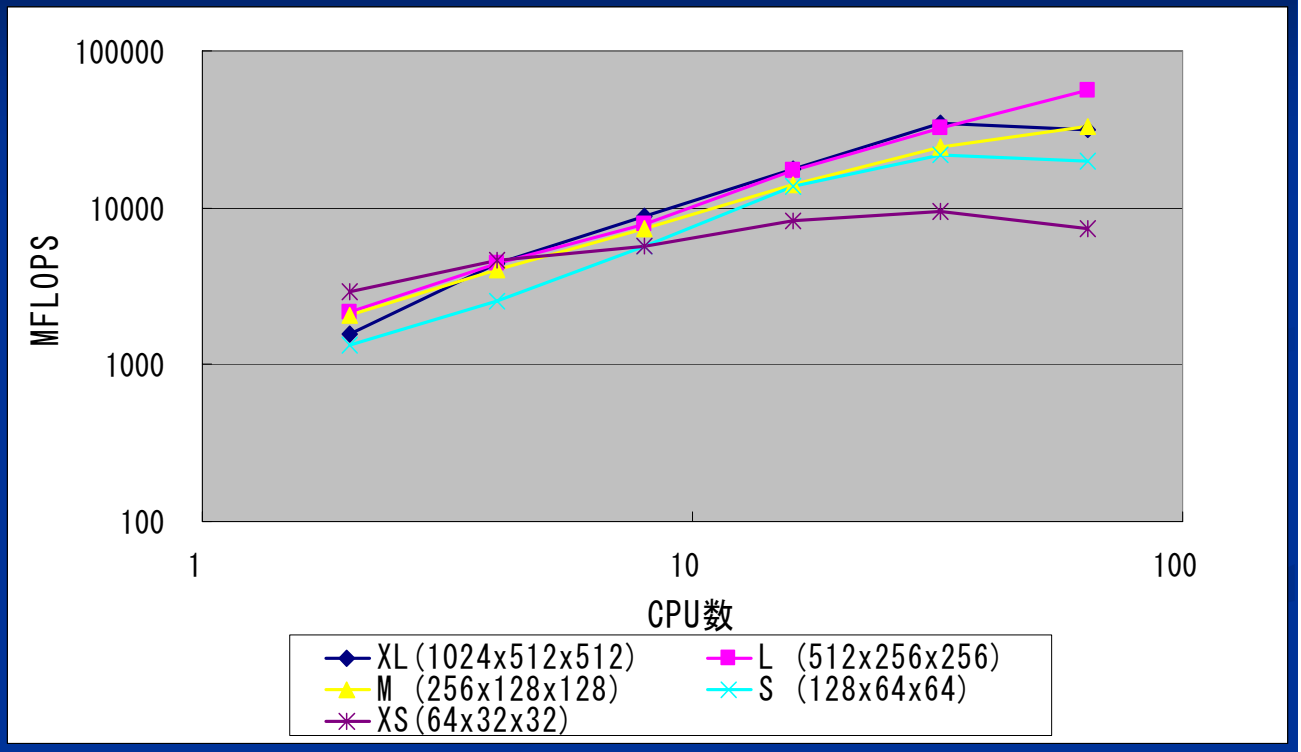
アウトライン

- 情報環境の概要
- 計算サーバの変遷
- 計算サーバの導入の基本的な考え方と経緯
- 代表的な計算サーバ
  - Cray XT3
  - SGI Altix3700　他
- 運用の基本的考え方および稼動統計
- 姫野ベンチマーク
- おわりに

姫野ベンチ (Cray XT3　MPI)



姫野ベンチ (SGI Altix3700, MPI)



おわりに

- JAIST の情報環境および計算サーバの紹介
  - Cray XT3
  - SGI Altix3700
  - ベクトル計算機、SMP計算機、PCクラスターなど
- 計算サーバの導入の基本的な考え方および導入の経緯

  ⇒ 情報環境の1部(マルチベンダー)、最新の計算機
- 計算サーバ運用の基本的な考え方

  ⇒ センターの関与をなるべく少なく、ユーザが主体
- 稼動統計と姫野ベンチマークテストの紹介