将来（２０１０年前後を想定）の ペタフロップス超級スパコンセンターとの連携について

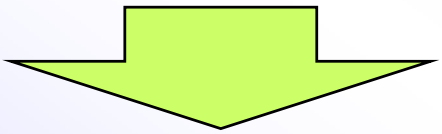
平成17年3月8日

大阪大学サイバーメディアセンター
下條真司、東田学

大阪大学サイバーメディアセンター

平成12年4月に学内外の情報基盤を支える組織として新たに設置された全国共同利用施設

旧大型計算機センター、旧情報処理教育センター、図書館（一部）などを再構成



各部局との連携により、**先端的研究成果を追求**し、**最先端の情報処理技術基盤の教育と普及**を行う中核的拠点を目指す。

大学の情報処理技術基盤の課題

情報関連ソフトウェア
利用技術の立遅れ

情報化された社会生活
実現への要求

情報コンテンツ発信
基盤整備の要望

大型計算機センター
- 演算サーバー機能
- 文献情報データベースの提供
- コミュニケーション支援機能
- 運用管理・課金機能
- 映像コンテンツ作成機能

- ソフトウェア資源の効果的な管理
- 市販ソフトの一元化管理・運営

図書館
- 書誌情報の提供（OPAC）
- 電子ジャーナル、文献情報データベースの提供

- 貴重資料の電子化・配信
- 電子ジャーナル等の情報提供サービスの充実
- マルチメディアアーカイブ機能の充実
- アクセス環境の充実、地域との連携

情報処理教育センター
- 教育用計算機システムの提供
- 学生用アカウントの管理

- アクセス環境の充実
- 教材作成・蓄積・配信環境の充実

情報処理技術基盤を整備するための拠点形成が必須

- 情報基盤整備に関する全学的議論の場
- 恒常的、長期的に計画を立案，実行していくための組織
- 大規模高速計算、情報ネットワークを用いた先端研究の支援、情報基盤に関する評価手法の確立などの研究開発情報基盤に支えられた高度な教育の実践
- 知的資源の電子的管理および提供
  - 貴重図書の電子化、マルチメディア教材

大阪大学サイバーメディアセンターの役割

マルチメディアを利用した科学教育
- 高度なコンピュータ利用の教育
- コンピュータ関連科学と自然科学の方法論との緊密な融合

情報ネットワーク整備
- キャンパスネットワークODINSの運用支援
- 広帯域ネットワーク、移動計算環境などの新しいネットワーク技術の導入

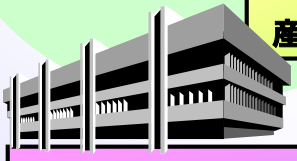
マルチメディア教室を基盤とした遠隔講義
- SCS遠隔教育の企画と運用支援
- ネットワーク利用遠隔講義の促進
- マルチメディア遠隔講義システム

国際化・言語教育
- マルチメディアを活用した外国語教育
- マルチメディア教材の開発

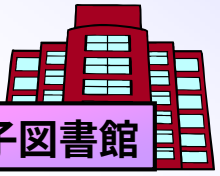
サイバーメディアセンター

大阪大学IT拠点
国際貢献
産官学連携推進

情報メディア教育支援
- 「読み書き算盤」としてのコンピュータ利用
- インターネットを利用した情報収集と発信
- マルチメディア教材の作成



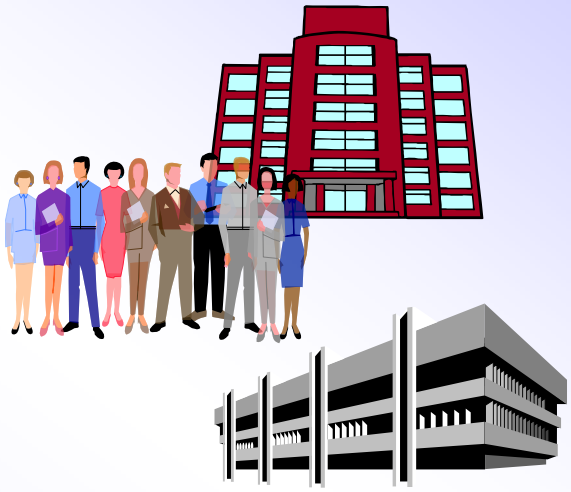
スーパーコンピューティング
- スーパーコンピュータによる計算サービス
- 新しい計算パラダイムに基づく計算機利用
- 大規模計算科学におけるシミュレーション



電子図書館
- 貴重なコンテンツのディジタル化
- 各種データベースの運用
- マルチメディアコンテンツの高次処理

サイバーメディアセンターの7研究部門

- 情報メディア教育研究部門
- マルチメディア言語教育研究部門
- 大規模計算科学研究部門
- コンピュータ実験科学研究部門
- サイバーコミュニティ研究部門
- 先端ネットワーク環境研究部門
- 応用情報システム研究部門



地域との連携
・公開講座など
・研究所・地域との共同研究

情報メディア教育研究部門

- **高度な情報処理教育環境の構築**
- **情報処理教育と情報倫理教育の実施**
- **情報処理教育担当者へのファカルティデベロップメント**

マルチメディア言語教育研究部門

CALL教室

- **マルチメディア言語教育環境の構築**
- **マルチメディア言語教育教材開発の支援**
- **ネットワークを用いた国際化教育の実施**
- **全学共通教育科目における外国語教育の実施**

コンピュータ実験科学研究部門

- **汎用コンピュータ・システムの運用支援**
- **科学問題設定・解決のための計算機応用に関するファカルティデベロップメント**
- **科学問題設定・解決のための過程習得に関連する科目の教育の実施**

大規模計算科学研究部門

- スーパーコンピュータ･システムの運用支援
- 計算結果可視化技術の普及
- 大規模計算システムの高度利用技術の啓蒙
- 計算科学及び関連する科目教育の実施

Top 19th (Jun 2002)

- 8ノードシステム
- 主記憶：1024GB
- 処理能力：1280GFLOPS
- ファイルシステム：20TB

NEC SX-5/128M8

サイバーコミュニティ研究部門

- SCS遠隔教育の企画と運用支援
- 社会との連携による先端技術にかかわる遠隔研修の企画と運用
- サイバーコミュニティ計画推進に関わる企画と運用

現行機SX-5/128M8の特徴

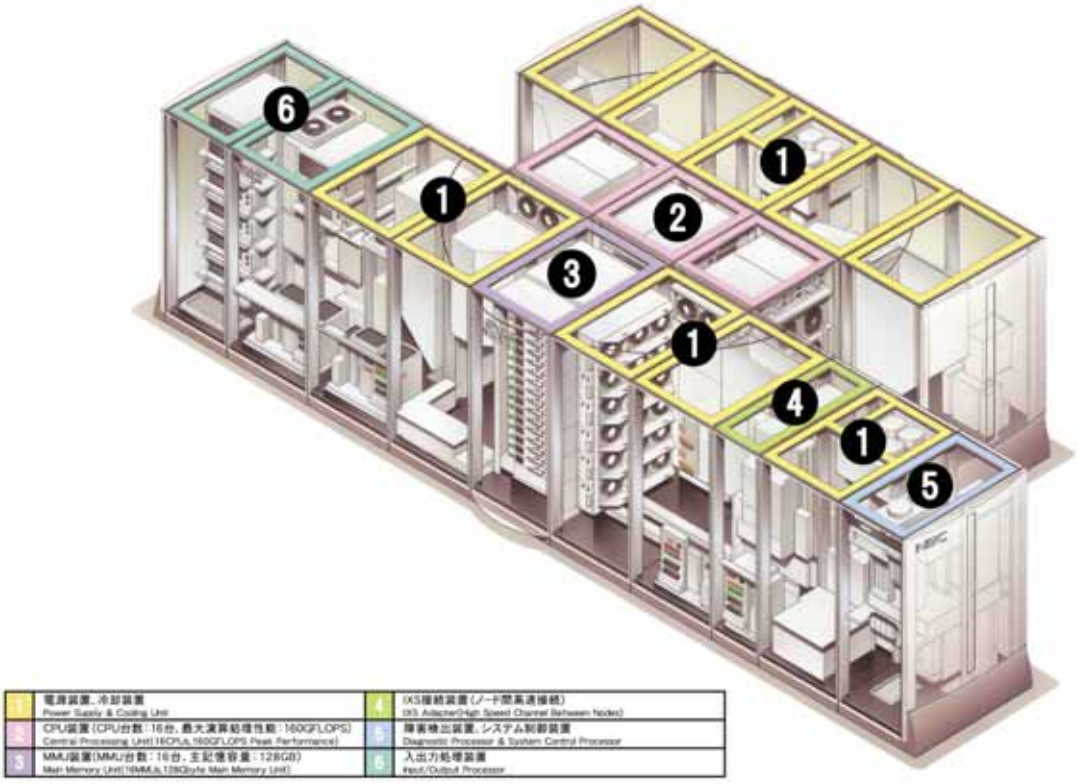
大型計算機センターのあるべき姿

- 「基礎研究」の支援
  - 計算科学研究者
    - 新規テーマに新規プログラミングで挑戦
      - NOT! 確立されたソフトウェアによるパラメータスタディ
      - NOT! 市販ソフトウェアによる設計・開発
    - 大規模シミュレーションによる求解から理論構築へ
    - 学生教育も行うが
      - プログラマーを養成しているのではない
  - 性能を最大限に引き出すプログラミングが容易な計算機が必要

現行機種SX-5/128M8の特徴

- 2001年1月導入、6年借料 ⇒ 2007年1月更新予定 (2年後)
- 高効率・高スループット
  - WSやPC、PCクラスタとは一線を画す性能を自動ベクトル化・並列化コンパイラによって容易に達成できる
  - ベクトル型
    - CPUあたり10GFLOPS
    - 自動ベクトル化コンパイラ (Fortran90, C/C++)
  - 共有メモリ型並列演算
    - 16並列: 160GFLOPS、128GB
    - 自動並列化コンパイラ (Fortran90, C/C++)
- MPIで使っても早い
  - ノード内は共有メモリを介した通信
  - ノード間はIXS (16GB/s) を介した通信
  - 自動並列との混在も可能

SX-5/16Af 透視図



大規模 Vector-Parallel システム
“地球シミュレータ” との比較

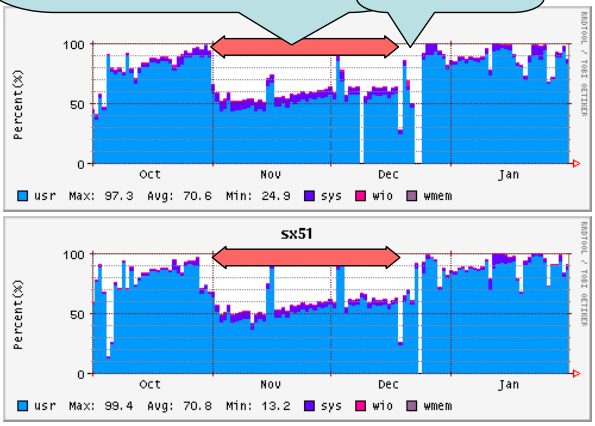
機種名称		NEC SX-5/128M8	地球シミュレータ SX-6/5120M640	性能 (容量) 比
ノード数		8	640	80 倍
プロセッサ数		128	5,120	40 倍
駆動周波数		312.5 MHz	500 MHz	1.6 倍
ピーク演算性能	総合	1,280 GFLOPS	40.960 GFLOPS	32 倍
	ノード毎	160 GFLOPS	64 GFLOPS	1 / 2.5 倍
	単体プロセッサ	10 GFLOPS	8 GFLOPS	1 / 1.25 倍
Linpack HPC		1,192 GFLOPS	35,860 GFLOPS	30.08 倍
	ピーク演算性能比	93.1 %	87.5 %	
	並列実効率	**99.9 %**	**99.997 %**	
主記憶容量	総合	1,024 GB	10,240 GB	10 倍
	ノード毎	128 GB	16 GB	1 / 8 倍
プロセッサへのデータ供給能力	総合	5,120 GB/s	163,840 GB/s	32 倍
	ノード毎	640 GB/s	256 GB/s	1 / 2.5 倍
	単体プロセッサ	40 GB/s	32 GB/s	1 / 1.25 倍

MPIプログラミングはしないのか？

- ノード内
  - PCクラスタからのスケールアップとして頻繁に利用される
- ノード間
  - ほとんど行われていない
    - システムが常に高負荷でマルチノード･ジョブが実行されにくい
    - 32並列は期待ほど性能が向上しない
      - 並列化率98% (本センターにおけるおおよその平均値) では、16並列よりも1.6倍高速になるだけ
      - それなら1.6倍待ってもらった方がシステム全体のTATが向上する
    - それでも単一ノードのメモリ容量を超えるジョブはチャレンジする
      - 2ヶ月延々と実行することも
      - 障害による再実行を予防するため定期チェックポイント (自動)
        - » さらに高負荷…

一月を超える長期実行に伴って
逐次結果を出力し
演算の妥当性を解析する必要があり、
並列化率は97～98%に留まる
そのため実行効率は60%弱

ジョブの終了を待って保守

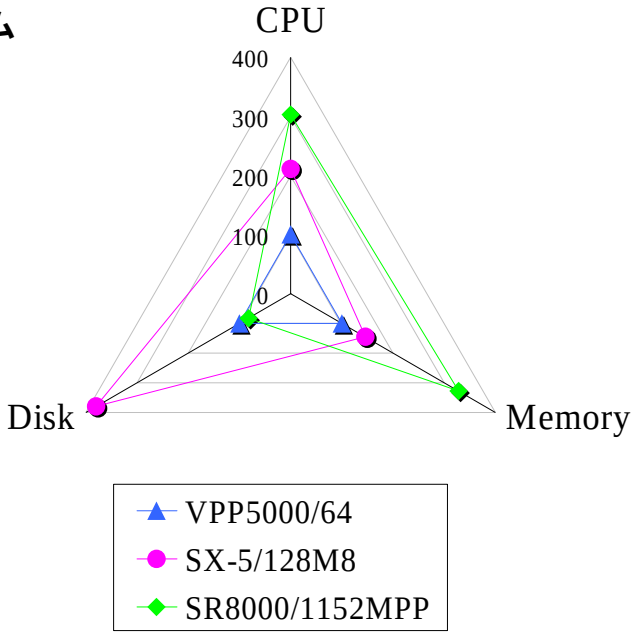


アプリケーション

プログラミング言語	Fortran 90, HPF, C, C++
数値演算ライブラリ	ASL/SX, MathKeisan (BLAS, LAPACK, ScaLAPACK,...), IMSL, NAG, CERNLIB
ジョブ･キューイング･システム	NQS, LSF
統合開発環境	PSUITE, Cygnus Code Fusion
並列化支援ツール	Etnus TotalView
可視化支援ツール	AVS/Express Developer, RVSLIB, SpaceFinder
流体解析アプリケーション	STREAM
衝撃解析アプリケーション	LS-DYNA
構造解析アプリケーション	MSC.Nastran
分子化学アプリケーション	Gaussian98

同時期に導入されたシステムの比較
九大VPP5000/64 を基準として

- **九大** VPP5000/64
  - **旧大型計算機センター導入システムの典型的構成例**
- **東大** SR8000/1152MPP
  - CPU **と** Memory **に重点**
  - Memory 2TB **に対して** Disk 4 TB
- **阪大** SX-5/128M8
  - Memory 1TB **に対して** Disk 20 TB
  - **それでもまだ足りない**…



SX-5/128M8**の稼働状況**

システムの稼働率と利用率

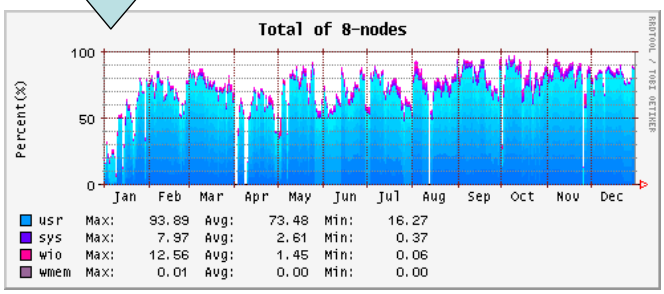
	稼働率 (通電時間)		利用率 (ユーザ時間)		稼働率×利用率	
平成13年度	93.0%		76.8%		**71.4%**	
平成14年度	95.7%	(1.03)	78.6%	(1.02)	**75.2%**	(1.05)
平成15年度	96.7%	(1.01)	75.3%	(0.96)	**72.8%**	(0.97)
平成16年度 (2005/01/31まで)	93.2%	(0.96)	86.6%	(1.15)	**80.7%**	(1.11)

括弧内は前年度比

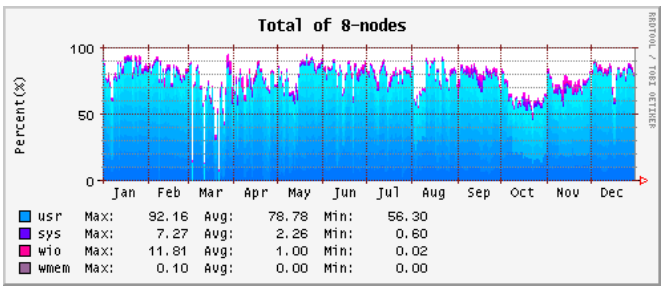
- 稼働率 = 通電しOSによって統計情報の取得が可能だった時間の割合
- 利用率 = 通電時間に対するユーザがCPUを利用した時間の割合
  - システム時間やアイドル時間を除く
- 稼働率×利用率 = 実利用率

利用率の推移

2001/01/05
借料開始

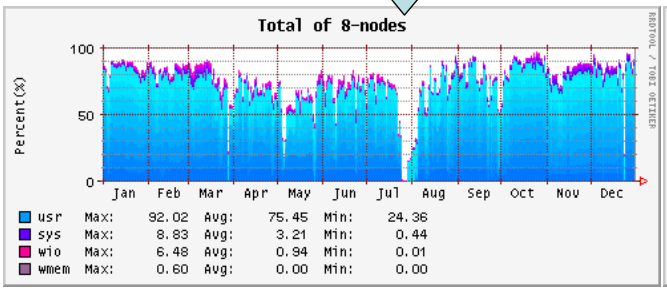


2001年



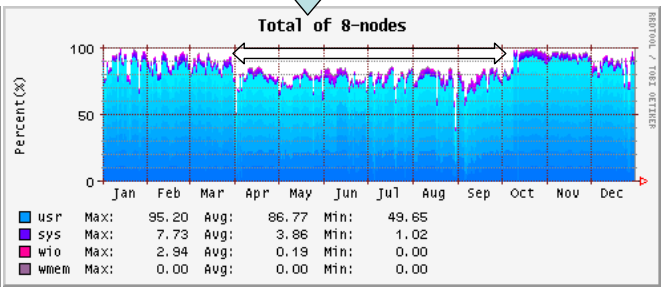
2002年

2003/07-08
OSバージョンアップ



2003年

2004/04-09
1ノード節電停止



2004年

安定稼働と節電要請の背反

- 安定した利用の立ち上がりと持続
  - 自律的なプロセスマイグレーションによってマルチノード間の負荷バランスを保つシステムの安定稼働例として稀有の存在
- 節電しなくちゃ・・・
  - 運営交付金、負担金見込額がショート
    - 平成15年度は、2ノードを約20日 (約100万円の節約)
    - 平成16年度は、1ノードを約160日 (約400万円の節約)
      - ただしOSバージョンアップに伴うシステム効率の改善によって、実利用率は8-pt近く向上
      - 後に示すように、ピーク性能に対する実演算率も向上

演算効率

	MFLOPS値		ベクトル化率		平均ベクトル長	
平成13年度	288,354.4		94.8%		302.9	
平成14年度	231,502.4	(0.80)	96.8%	(1.02)	323.0	(1.07)
平成15年度	193,220.6	(0.83)	93.4%	(0.96)	298.1	(0.92)
平成16年度 (2005/01/31まで)	213,340.8	(1.10)	97.4%	(1.04)	338.9	(1.14)

括弧内は前年度比

- MFLOPS値: 単位時間あたりの浮動小数点演算数
  - SX-5/128M8では最大1,280,000MFLOPS (= 1,280GFLOPS)
  - 年平均15%から23%の実演算性能を研究者に提供
- ベクトル化率: 演算のうちベクトル化された割合
- 平均ベクトル長: ベクトル命令が処理した要素数 (ループ長) の平均値
  - SX-5では最大512

ピーク性能に対する実演算効率

	稼働率×利用率		MFLOPS値		稼働率×利用率 ×MFLOPS値		ピーク 性能比
平成13年度	71.4%		288,354.4		205,885.0		**16.1%**
平成14年度	75.2%	(1.05)	231,502.4	(0.80)	174,089.8	(0.85)	**13.6%**
平成15年度	72.8%	(0.97)	193,220.6	(0.83)	140,664.6	(0.81)	**11.0%**
平成16年度 (2005/01/31**まで**)	80.7%	(1.11)	213,340.8	(1.10)	172,166.0	(1.22)	**13.5%**

括弧内は前年度比

- SX-5/128M8**のピーク性能値**1,280GFLOPS**に対して、**
  通年平均して11%**から**16%**の実効率で演算能力を研究者に提供している**

実効性能値の漸減

- **借料期間も半ばを過ぎ、ロースキル･ユーザによる大規模利用が増えてきた**
  - **ケアレスミスによる効率低下**
    - e.g. 16**並列キューに**4**並列指定のままジョブ投入**
  - **未熟なチューニング**
    - **現行機は前機種よりも高いチューニング･スキルを要求**
      - SSRAM ⇒ SDR-SDRAM
      - **メモリバンク数が半減**
- **歯止めがかからないかに思われたが･･･本年度なんとか回復**
  - **ジョブクラス構成の変更**
    - **より並列度の低い (効率の高い) キューへの誘導**
  - **講習会の見直し**
    - **初級、中級、上級編**

現在のセンターにおける課題

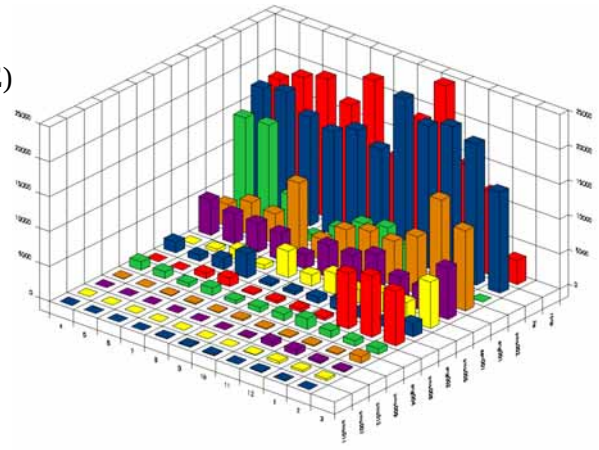
演算効率向上と研究者の負担

- 演算効率向上
  - 自動ベクトル化･自動並列化
    - チューニングノウハウは十二分に浸透
  - MPIライブラリによる手動並列化
    - プログラミングも含めて正直面倒くさい
    - チューニングしても性能が上がる保証がない
      - 並列化率による制限、ボトルネックの解消、機器障害による中断への対処
- 効率向上を全面に打ち出しすぎると研究者は逃げる
  - プログラムをチューニングしたいのではなくて理論を実証したい
    - プログラムチューニングに費やす時間も惜しい
    - 出力データを次々に解析したい
    - 実験システムにオンザフライで反映したい
  - むしろ研究現場のワークフローを記述し、構成に反映できるデータフロー記述言語が必要
    - 今はNQSバッチスクリプト
    - Grid技術に期待

学内センター間連携による マルチレベル、マルチフィジックス シミュレーションへのアプローチ

- **三センター連携による共同運用**: CMC, ILE, RCNP
  - **高エネルギー物理分野**
    - **レーザーエネルギー学研究センター** (ILE)
    - **核物理研究センター** (RCNP)
  - **運営コストの分散**
    - SX-5/128M8**の分散設置**
      - CMC**主機室の設置要件 (フロア面積、電源容量) の制約から**
    - **高熱水料費の相互負担**
  - **階層型フェアシェアによる 負担額に応じた資源配分**
    - **シェア率の割当**: CMC: 4/8、ILE: 2/8、RCNP: 2/8
    - **利用実績値を随時積算し、割り当てたシェア率が達成・維持されるようジョブの実行優先度を自動制御**
    - **利用実績値は設定した半減期間に応じて漸減**
      - **閑散期に投機的なジョブ実行を促す**

平成16年度のグループ毎の利用実績の月別推移



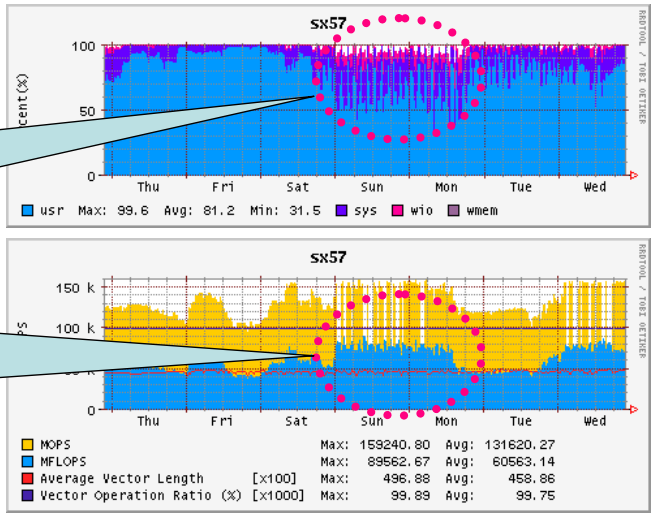
それぞれのセンターが所有する “汎用計算機”との連携

- **分散設置したSX-5ノード間**
  - **複数のHiPPIやGigabit Ethernetインターフェイスにて占有接続**
- **キャンパスネットワークODINSによる各センターの所内LANとの相互接続**
  - **高エネルギー物理学分野特有のセキュリティの要請からDMZ、F/Wを介した接続**
  - ILE
    - SX-6/4 (32GFLOPS, 24GB)
    - SX-8/6A (96GFLOPS, 64GB)
  - RCNP
    - IBM RS/6000 SP
      - 56-procs. of POWER4
      - 40TB Disk + 20TB Tape

- 数百GB〜数TBのデータが出力されるのに、データ転送速度が1GB/min.
  - データをホーム・センターに引き上げるのがまさに一日仕事
  - IOPを介したファームウェア処理によるI/Oの問題(SX8では解決)
  - データ転送中はスパコンにシステムコールを発行し続け、演算を妨げる
    - ファイルサービスを集中して行うサービスプロセッサを設定し回避
- 処理のワークフローを記述しジョブ進行を同期させる枠組みが必要
  - 既存のNQSバッチスクリプト
  - 今こそGrid技術の適用

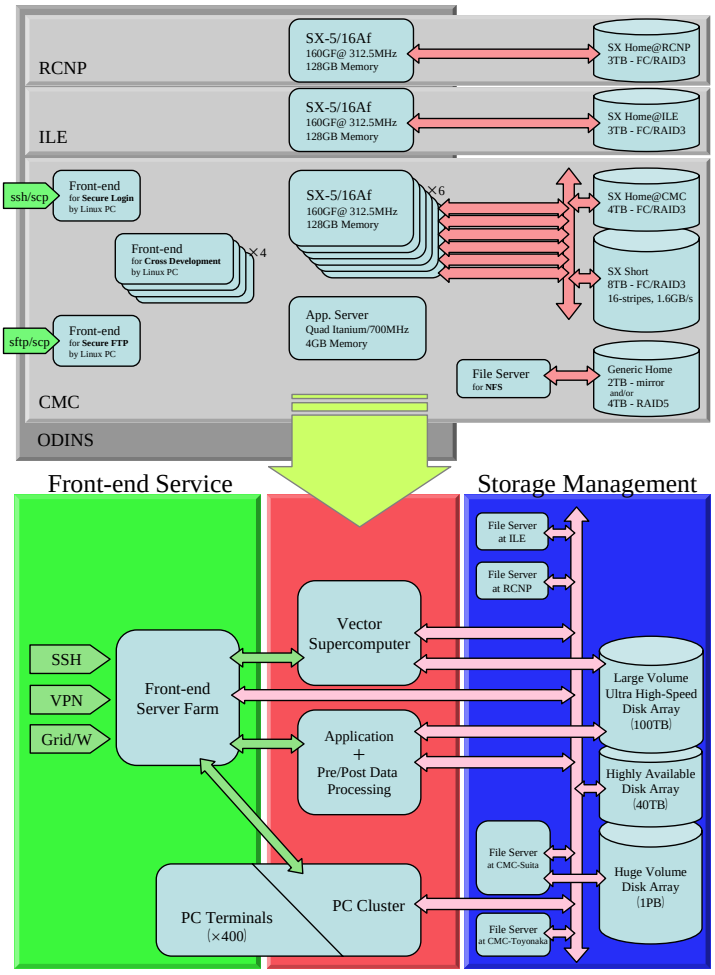
I/O処理を伴う
多数のジョブの平行実行
によるシステム負荷

ジョブのFLOPS値は
十分出ている
I/Oが高効率化すれば
もっとTATが向上するはず…



次期システム構想

- 現行システム: 横割り
  - ODINSというキャンパスネットワークを土台に
    3センターへ資源分散
- 次期システム: 縦割り
  - サービス層を明確に
    - スーパーコンピューティング・サービスの提供に特化したフロントエンド
    - バックエンドにストレージ共有に特化した占有システム・エリア・ネットワークを敷設
    - 演算資源はその双方の層と密接に接続



現状の大型計算機センターでは
個別に克服できない課題

商用機のトレンドに右往左往

- 幾度のネガティブキャンペーン:「ベクトル機は死んだ」
  - 「ベクトル機にはスケーラビリティがないですよ」
    - MPP (Massively Parallel Processors) はいかがですか？
  - 「ベクトル機はコストパフォーマンスが悪いですよ」
    - PCクラスタはいかがですか?
- そうこうするうちにベクトル機メーカーが衰退
  - NEC SXとCray X以外に選択の余地がない

TOP500 Rank				# of procs.	R_{max}	R_{peak}	R_{max}/R_{peak}
29	Cray	X1		504	5,895	6,451	91.4%
32	Fujitsu	PRIMEPOWER HPC2500	1.3GHz	1,472	5,406	11,980	45.1%
59	Hitachi	SR1100-H1/56	1.7GHz	56×16	3,319	6,093	54.5%
88	NEC	SX-6/248M31	563MHz	248	2,155	2,232	96.6%

R_{max}/R_{peak}値がGigabit Ethernetで相互接続したPCクラスタに劣る効率でいいのか？

GFLOPS/kW値にみる PCクラスタのコスト感覚

	駆動周波数	プロセッサ当りのピーク演算性能	製造プロセス	測定ホスト	GFLOPS/kW
Pentium 4	2.4GHz	4.8GFLOPS	0.13um	DELL PowerEdge650	34.2
NEC SX-5	312.5MHz	10GFLOPS	0.25um	SX-5/Af	3.53
NEC SX-6	500MHz	8GFLOPS	0.15um		9.93
NEC SX-8	1GHz	16GFLOPS	90nm		15-20 (推定値)

データ提供: 大阪ガス 河本 薫氏

- PCクラスタは必ずしも低コストではない
  - 消費電力
  - 保守性
    - 障害発生に伴うジョブの再実行
  - プログラミング・コスト

GFLOPS/kWからTOP500リストを検証

- SX-5 (マルチチップCMOS, 0.25um) は確かに消費電力が大きい…

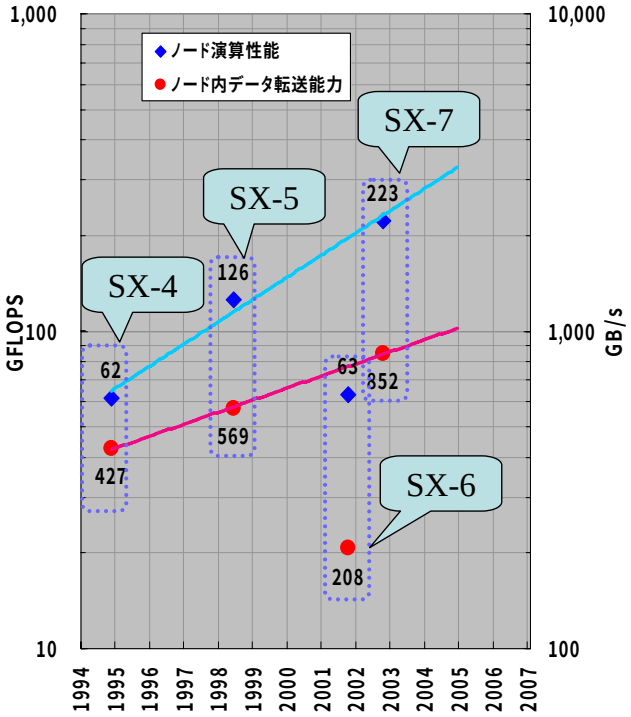
TOP500 Rank			# of procs.	R_{max}	R_{peak}	R_{max}/R_{peak}	推定消費電力
291	SX-5/128M8	IXS	128	1,192	1,280	93.1%	362.6kW
292	Xeon 2.4GHz	GbE	354	1,182	1,982	60.0%	58.0kW

- SX-6 (シングルチップCMOS, 0.15um) では同規模のシステムと大差はない
  - そもそもGigabit Ethernetで結合したPCクラスタで60%を超える実効性能がでるのか？

TOP500 Rank			# of procs.	R_{max}	R_{peak}	R_{max}/R_{peak}	推定消費電力
83	Xeon 3.4GHz	Myrinet	1028	2,200	6,990	31.5%	204.4kW
88	SX-6/248M31	IXS	248	2,155	2,232	96.6%	224.8kW
91	Xeon 3.06GHz	GbE	568	2,140	4,160	61.4%	101.6kW

ノード当たりの演算性能と データ供給能力の推移 (実効性能)

- ノード当たりの演算性能の向上率に比べると、データ供給能力の性能向上率は緩やか
  - プログラムの実行効率向上によってこの限られた帯域を有効に活用するのは必是
- 民間需要だけでこの性能向上率を維持するための開発コストは捻出できるだろうか？



ペタフロップス超スパコンセンターとの 計算機センターのあるべき姿

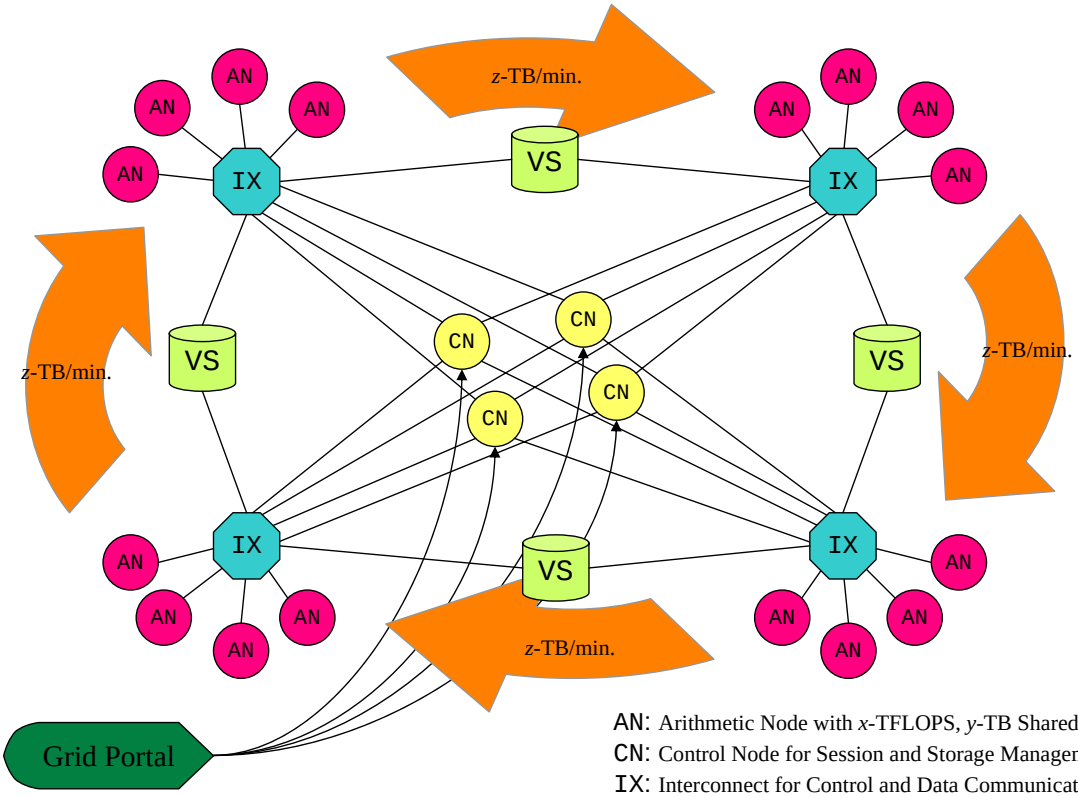
「ペタフロップス超スパコンセンター」への資源集約は是か非か？

- マルチレベル、マルチフィジックス・シミュレーションを実現するためには、高効率を維持できる規模のサブシステムを多段に密結合した高スループットなシステムが必要
  - 高効率を維持できるサブシステムの規模
    - 2010年では演算性能が数TFLOPS、主記憶が数TB
  - 多段結合による高スループットの実現
    - サブシステム間で数TB/min (数十GB/sec.) のパイプライン処理を組む
- メタスパコンセンターによって実現できるか？
  - グリッド技術に期待
    - 研究現場のワークフローを落とし込む
      - データフローを記述
      - ジョブ投入管理スクリプトと同期

ペタフロップス超級スパコンセンターとの連携

- 大型計算機センターの役割
  - 多様な利用者と運用者が切磋琢磨
  - 利用技術の蓄積
  - 地域密着型
  - 大学と連携した人材育成
  - 高スループットを目指した運用
- ペタフロップス超級スパコンセンター
  - 少数の利用者によるgrand challenge
  - プロによる利用
  - 全国選抜
  - 高性能利用を目指した運用

Proposed Configuration of A Peta-scale System



AN: Arithmetic Node with *x*-TFLOPS, *y*-TB Shared Memory
CN: Control Node for Session and Storage Management
IX: Interconnect for Control and Data Communication
VS: Storage for Intra-Node Virtual Stream Buffer