

クラスタコンピューティングとT2Kアライアンス ～過去・現在・将来～

東京大学
石川 裕

The University of Tokyo

2008/04/07

Outline



1. クラスタコンピューティングの歴史
2. T2Kオープンスパコンアライアンス
3. クラスタコンピューティングの将来
4. T2Kアライアンスの将来
5. 東大情報基盤センターの今後

クラスタの歴史: 1994年 クラスタの夜明け

1980

1990

1995

クラスタの夜明け

Cray Y-MP (88) 2.6Gflops/8procs		NEC SX-4 2Gflops/CPU (94)	
	CM-5 (91) Paragon (91)	Cray T3D (93)	Cray T3E (95)
Sparc (87)	SS 2 (90)	SS 10 (92)	SS 20 (94)
		Alpha 21064 200MHz (92)	Alpha 21164 300MHz (94)
		Pentium 60MHz (93)	Pentium Pro 200MHz (95)
10Mbps Ethernet (81)		100Mbps Ethernet (93)	Myrinet (94)
			Gigabit Ethernet (95)
			ATM/LAN(95)
	PVM (2.0/91)	MPI (94)	

• 背景

- 高性能低価格マイクロプロセッサの台頭
- 高速ネットワークの出現
- デファクト スタンダード通信ライブラリの制定
 - MPI
- コストパフォーマンス重視

• クラスタとは？

- 本来単体で使われるコンピュータ群をネットワークで接続し、それらをシステムとして使用する。並列業界では、特にコモディティ部品を使って作られたシステムをクラスタと呼び、超並列コンピュータと区別

クラスタの歴史: 1995年～1996年 黎明期



クラスタの夜明け

1995年8月 Sun SS20 x 5



10Mbps Ethernet (81)

1996年2月 Sun SS20 x 36 connected by Myrinet



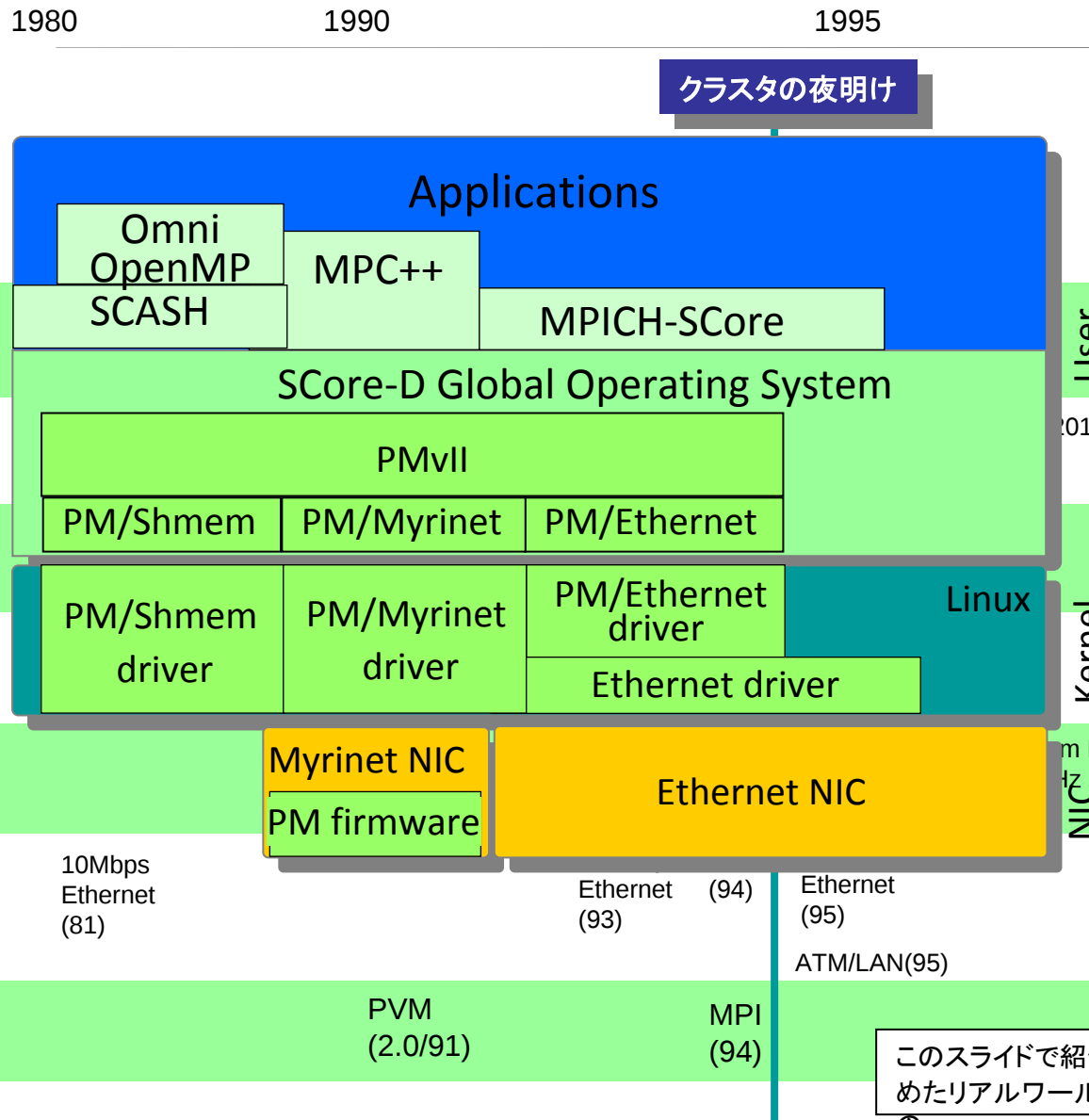
NEC SX-4 2Gflops/CPU (94)			
Cray T3D (93)		Cray T3E (95)	Hitachi (96)
10 (93)	SS 20 (94)	Ultra I 200MHz (95)	
Alpha 21064 300MHz (93)	Alpha 21164 300MHz (94)	Alpha 21164 500MHz (95)	Alpha 21164 500MHz (96)
Pentium 60MHz (93)		Pentium Pro 200MHz (95)	
100Mbps Ethernet (93)	100Mbps Myrinet (94)	Gigabit Ethernet (95)	ATM/LAN(95)

PVM
(2.0/91)

MPI
(94)

このスライドで紹介してるクラスタは、経済産業省が1992年から10年間進めたリアルワールドコンピューティングプロジェクトにおいて、石川が主導して開発したもの

クラスタの歴史: 1995年～1996年 黎明期



- 高性能通信ライブラリPM
 - Myrinet、Ethernet
 - ゼロコピー通信とピンダウンキャッシュ
- グローバルOSSCore-D
 - ギャングスケジューリング
 - チェックポイント・リスタート
 - デッドロック検出
 - リアルタイムモニタ
- マルチスレッド言語MPC++
 - マルチスレッドテンプレートライブラリ
- ソフトウェア分散共有メモリ
SCASH
- クラスタ向けOpenMP処理系
Omni OpenMP

このスライドで紹介してるクラスタソフトウェアは、経済産業省が1992年から10年間進めたリアルワールドコンピューティングプロジェクトにおいて、石川が主導して開発したものの

の

クラスタの歴史: 1995年～1996年 黎明期

1980

1990

1995

クラスタの夜明け

1996年10月 Pentium Pro x 32 connected by Myrinet

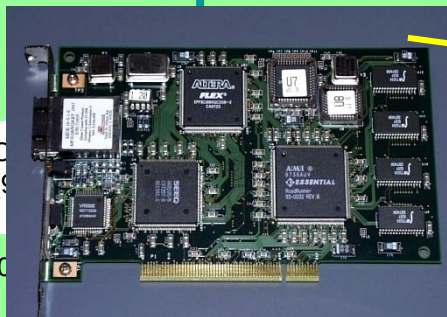
Cray Y-MP
(88)
2.6Gflops/8procs

CM-5
(91) Paragon
(91)

Sparc
(87)

SS 2
(90)

SS 10
(92)



(95)

Alpha 21064
200MHz
(92)

Alpha 21164
300MHz
(94)

Alpha 21164
500MHz
(96)

Pentium
60MHz
(93)

Pentium Pro
200MHz
(95)

Pentium
333MHz
(97)

10Mbps
Ethernet
(81)

100Mbps Myrinet
Ethernet (93)

Gigabit
Ethernet
(95)

ATM/LAN(95)

PVM
(2.0/91)

MPI
(94)



このスライドで紹介してるクラスタは、経済産業省が1992年から10年間進めたリアルワールドコンピューティングプロジェクトにおいて、石川が主導して開発したもの

クラスタの歴史: 1997年～1999年 普及期

1980

1990

1995

2000

クラスタの夜明け

1997年～1998年



NEC SX-4
2Gflops/CPU
(94)

T3D

NEC SX-5
8Gflops/CPU
(98)

Cray T3E
(95)

Hitachi SR2201
(96)

Hitachi SR8000
(99)

SS 20 Ultra I
400MHz

Ultra II
480MHz
(98)

Ultra III
900MHz
(00)

Alpha 21164
500MHz

Alpha 21264
800MHz

1999年

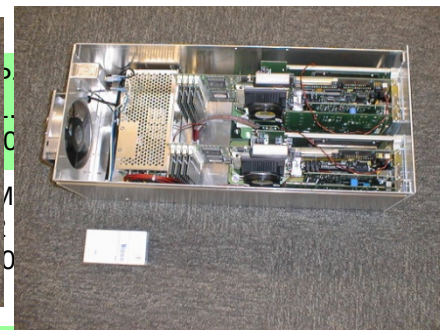


10Mbps
Ethernet
(81)



(2001)

(94)



このスライドで紹介してるクラスタは、経済産業省が1992年から10年間進めたリアルワールドコンピューティングプロジェクトにおいて、石川が主導して開発したもの

クラスタの歴史: 2000年～2001年 大規模実証期

1980

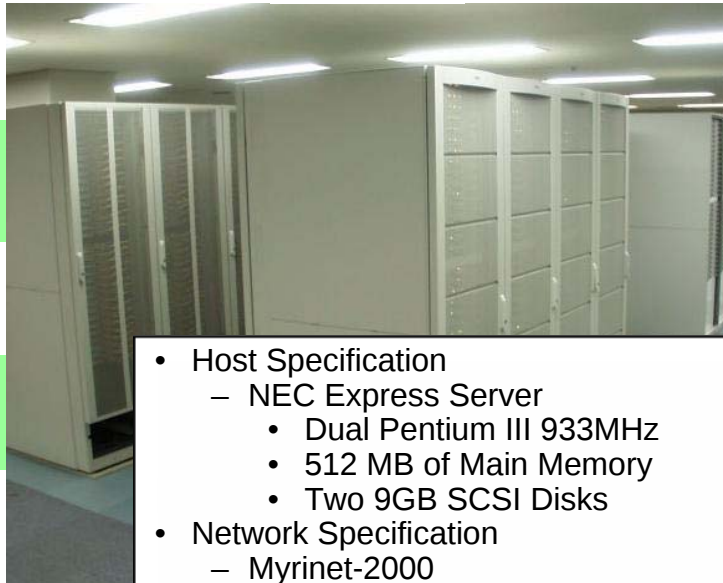
1990

1995

2000

クラスタの夜明け

2001年



10Mbps
Ethernet
(81)

- Host Specification
 - NEC Express Server
 - Dual Pentium III 933MHz
 - 512 MB of Main Memory
 - Two 9GB SCSI Disks
- Network Specification
 - Myrinet-2000
 - Two 100Mbps Ethernet Links
- Number of Hosts
 - 512 (1,024 CPUs)
 - Peak Performance: 955 Gflops
- Software
 - Kernel
 - Linux 2.2.16
 - Cluster Software
 - SCore 3.3.2/TurboLinux
- Linpack Benchmark Result
 - 618.3 Gflops

X-4
CPU
(4)

NEC SX-5
8Gflops/CPU
(98)

NEC
Earth
Simulator
(02)

Cray T3E
(95)

Hitachi SR2201
(96)

Hitachi SR8000
(99)

Ultra II
480MHz
(98)

Ultra III
900MHz
(00)

Alpha 21164
500MHz
(96)

Alpha 21264
800MHz
(98)

Pro
z

Pentium II
333MHz
(97)

P- III
500MHz
(99)

P-IV
1.4GHz
(00)

et

Myrinet-2000
2 Gbps
(00)

N(95)

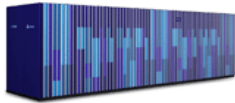
Infiniband
10 Gbps

このスライドで紹介してるクラスタは、経済産業省が1992年から10年間進めたリアルワールドコンピューティングプロジェクトにおいて、石川が主導して開発したもの

クラスタの歴史: 2001年～ 計算センターの転換期

20世紀型

計算センター



- 計算資源の提供
- プログラミング支援

大学・研究機関

ユーザ層

- 大規模計算ユーザ
- 中規模計算ユーザ
- 計算環境を保守したくないユーザ

21世紀初頭型

現在のパソコン

- 数Gflops、2GBメモリ
- 10年前のスパコン

大学・研究機関

ユーザ層

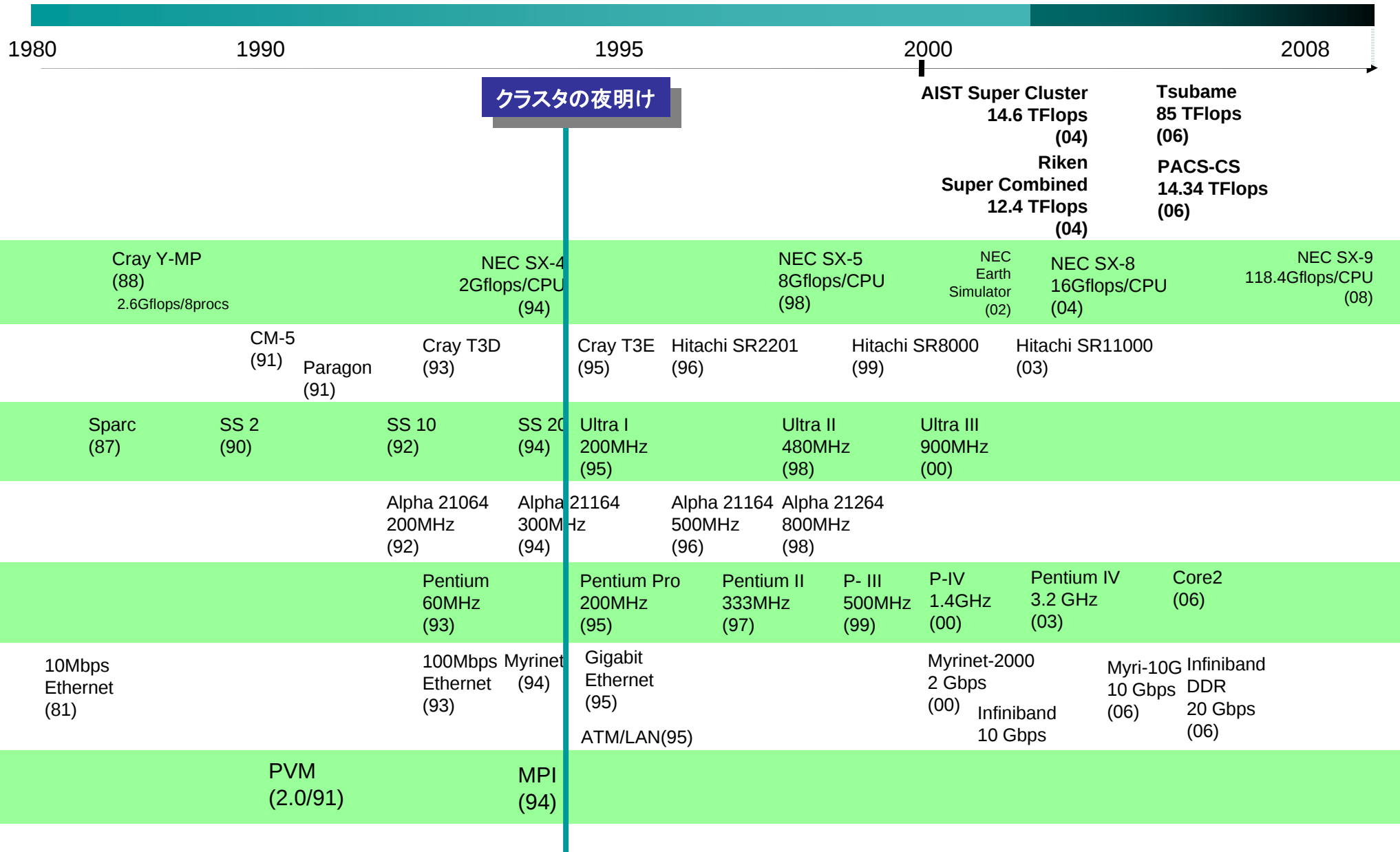
- 大規模計算ユーザ
- 計算環境を保守したくないユーザ

大学・研究機関

- 中規模計算ユーザ
- 計算環境を構築できるユーザPCクラスタユーザ

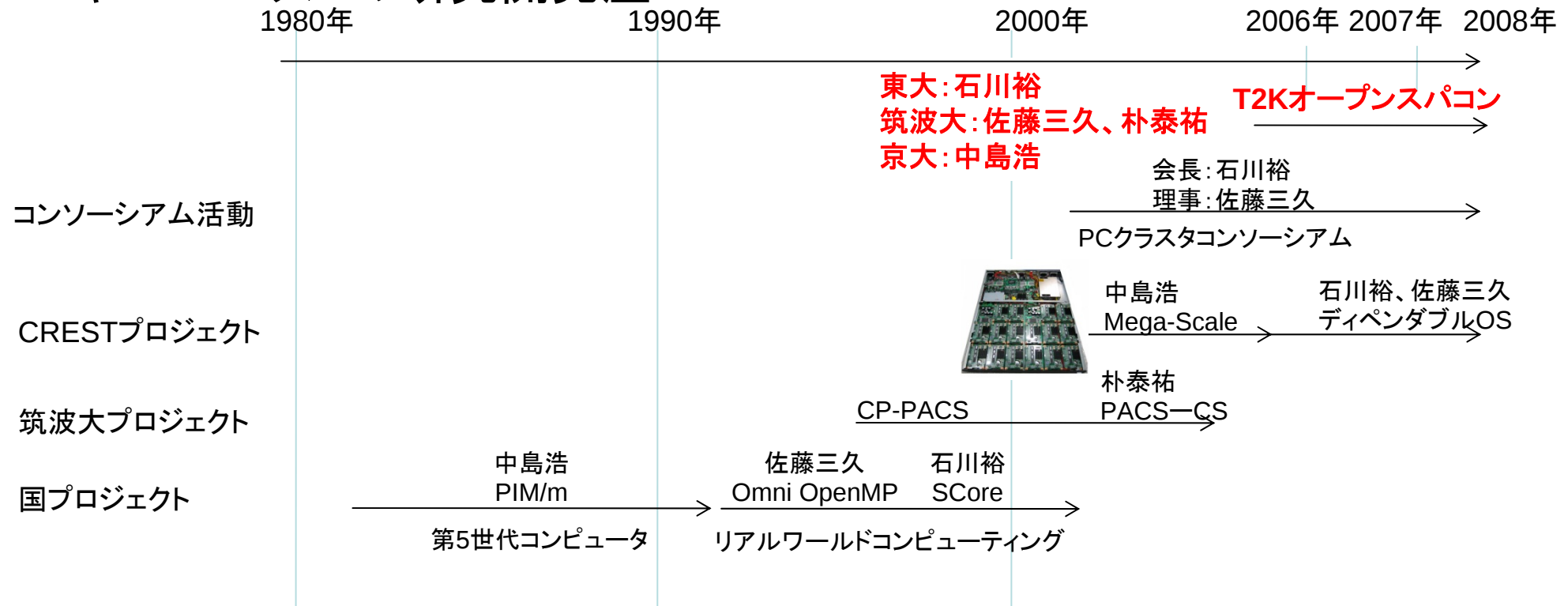


クラスタの歴史：2001年～2006年 計算センター導入



2006年： T2Kオープンスパコンアライアンス

背景：計算機科学者によるボトムアップ的アプローチ キーパーソンの研究開発歴



- ❑ PIM/mとは、中島が主導して研究開発を行った並列推論マシン
- ❑ Omni OpenMPとは、佐藤が主導して研究開発を行ったOpenMP処理系
- ❑ SCoreとは、石川が主導して研究開発を行ったクラスタのためのシステムソフトウェア

SCoreユーザ
(プロジェクト終了後)

理研スーパーコンバ
インドクラスタ

PACS-CS



英国Streamline社による
SCoreクラスタ販売



T2K背景:PCクラスタユーザの現状

- 満足しているユーザ
 - 小中規模科学技術計算ユーザ
 - Embarrassingly Parallel Applicationユーザ
 - モンテカルロ手法など、並列度はあるが通信をほとんどしない計算
 - 大量のジョブを一度に実行する
- 不満を感じているユーザ
 - 性能が出ない
 - ネットワークの問題
 - 保守・維持が大変
 - よく壊れる
 - 熱設計がされていない
 - システムソフトウェアの保守
 - 電気代がかかる

大規模化への道筋

T2K背景: 計算ニーズの多様性

- 従来スパコン応用分野
 - 超規模計算科学、超規模計算工学
 - 100テラフロップス超、数十 Tbytes超主記憶
- 新興応用分野
 - 大規模ゲノム情報処理
 - 超規模アーカイブ検索
 - コンピュータグラフィックス
- PCクラスユーザ
 - 小中規模科学技術計算
 - Embarrassingly Parallel Applications

新しい計算センターはこれらニーズを吸収するマシンが必要

背景：利用形態の多様性

- バッチ・インタラクティブ
 - 従来の利用形態
- WEBユーザインターフェイス
 - ポータル
- グリッド
 - データ共有、ワークフロー

オープンスパコンの理念

- 基本アーキテクチャのオープン性
 - コモディティ高性能プロセッサを基本
 - コンピュータ市場を牽引しているコモディティ高性能プロセッサを使用することにより、最新技術を使用することにより、高性能かつ低消費電力を実現したシステムを導入することが可能
- システムソフトウェアのオープン性
 - オープンソースに基づく先端ソフトウェア技術を基本
 - 多くのユーザが使用するこれら資産をシームレスに利用できる環境を提供することにより、より多くのユーザが大規模並列処理環境へ移行することが促進できます
- ユーザのニーズに対するオープン性
 - 従来の計算センターユーザでないニーズに対して応える
 - 大規模ゲノム情報処理、大規模データマイニング

計算センターから研究室までをカバーするアーキテクチャ

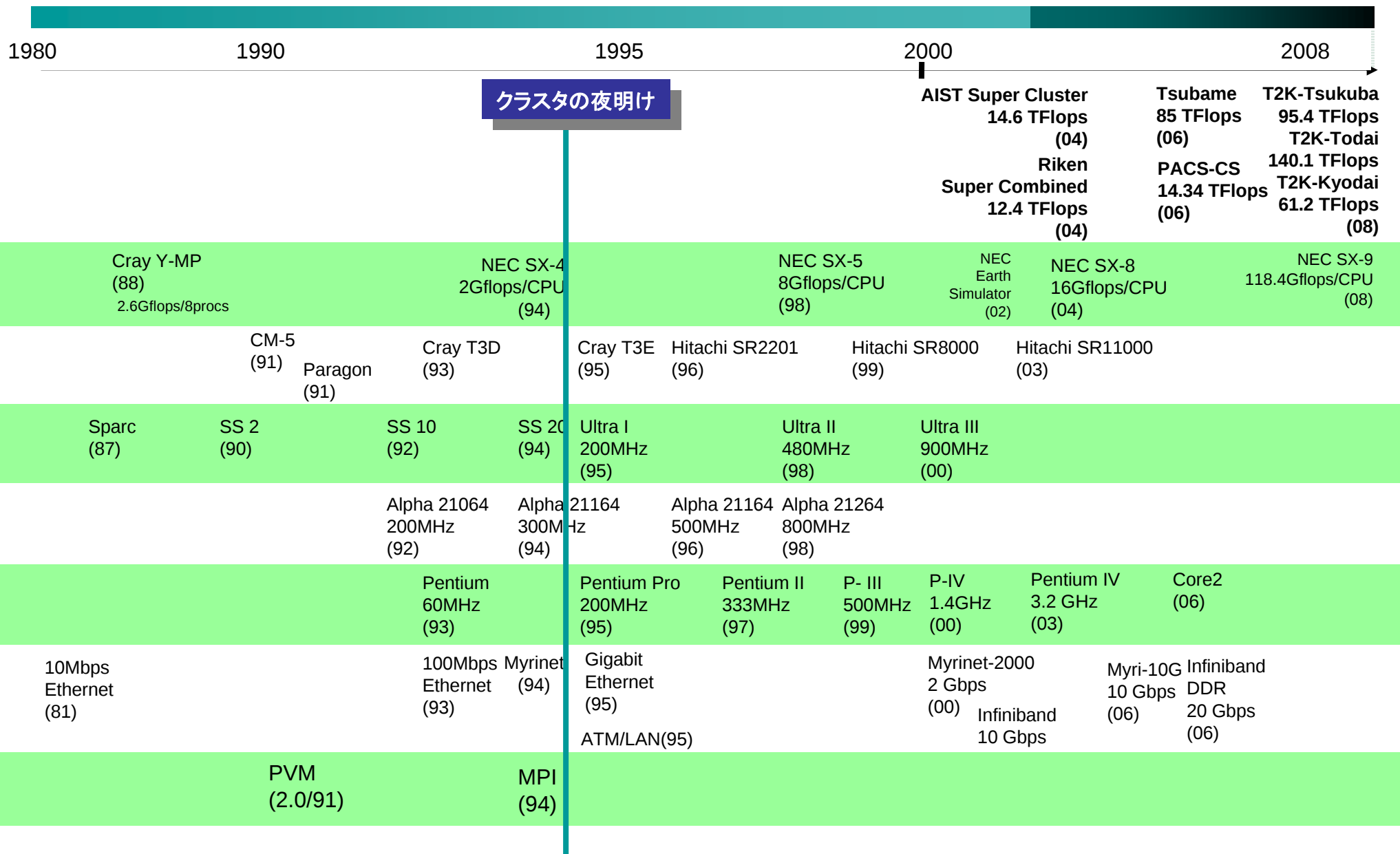
- 目的

- 2007年～2008年に出荷されるCPU技術を想定し、センタ運用できるスパコン仕様を提示
 - ハードウェア仕様
 - ノード構成
 - ネットワーク
 - ソフトウェア仕様
 - オペレーティングシステム
 - コンパイラ&プログラミング環境
 - ライブラリ
 - 商用アプリケーション

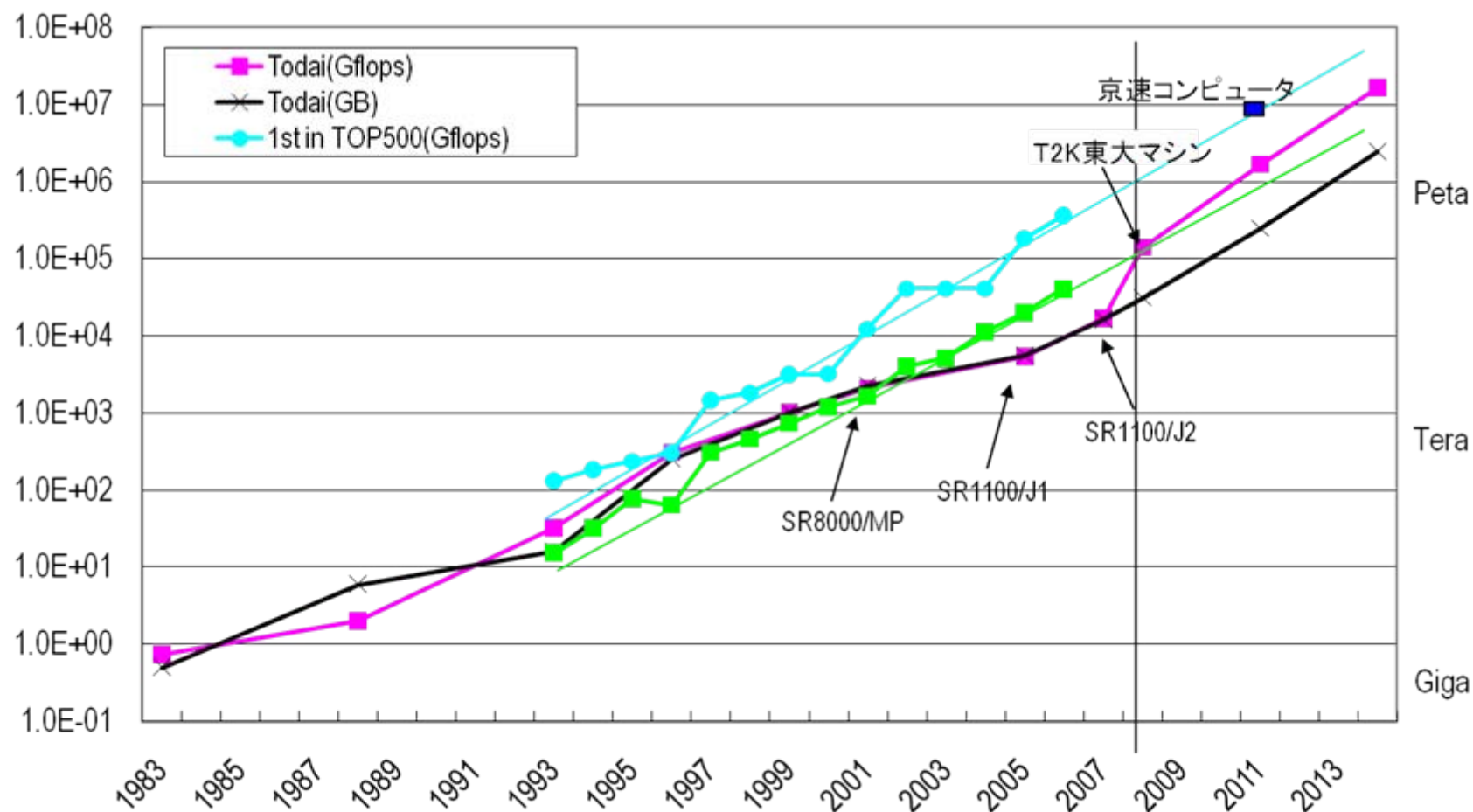
- 手法

- 現状のコモディティCPU技術の検証
- 必要とされる通信性能とその実現性の検証
- 現状のコンパイラ技術、プログラミング環境、オペレーティングシステムの検証
- 新しいタイプのユーザに対する運用方法の検討
- 共通ベンチマーク策定

クラスタの歴史: 現在



東京大学情報基盤センターマシン



東京大学情報基盤センター スパコン

HITACHI SR11000 model J2

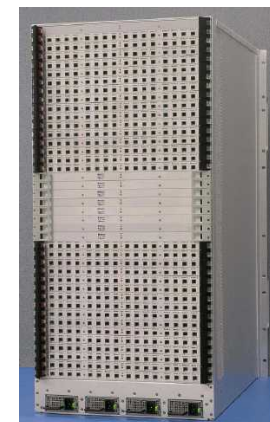
Total Peak performance	: 18.8 TFLOPS
Total number of nodes	: 128
Total memory	: 16384 GB
Peak performance per node	: 147.2 GFLOPS
Main memory per node	: 128 GB
Disk capacity	: 94.2 TB



T2K東大

HA8000-tc/RS425 x 952

Total Peak performance	: 140 TFLOPS
Total number of nodes	: 952
Total memory	: 32000 GB
Peak performance per node	: 147.2 GFLOPS
Main memory per node	: 32 GB, 128 GB
Disk capacity	: 1 PB



クラスタの将来: 計算センターの役割

20世紀型

- 計算資源の提供
- プログラミング支援

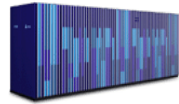
大学・研究機関

ユーザ層

- 大規模計算ユーザ
- 中規模計算ユーザ
- 計算環境を保守したくないユーザ

21世紀初頭型

計算センター



21世紀型

- 計算資源の提供
- プログラミング支援
- 計算科学・工学教育支援
- ユーザのニーズ毎にカスタマイズした利用環境提供
 - 研究室レベルでのPCクラスタ運用雑用の解放
- 社会貢献
 - 民間の計算シミュレーション新規利用支援

大学・研究機関

ユーザ層

- 大規模計算ユーザ
- 計算環境を保守したくないユーザ

大学・研究機関

ユーザ層

- 中規模計算ユーザ
- 計算環境を構築できるユーザ
- PCクラスタユーザ



大学・研究機関

ユーザ層

- 大規模計算ユーザ
- 中規模計算ユーザ

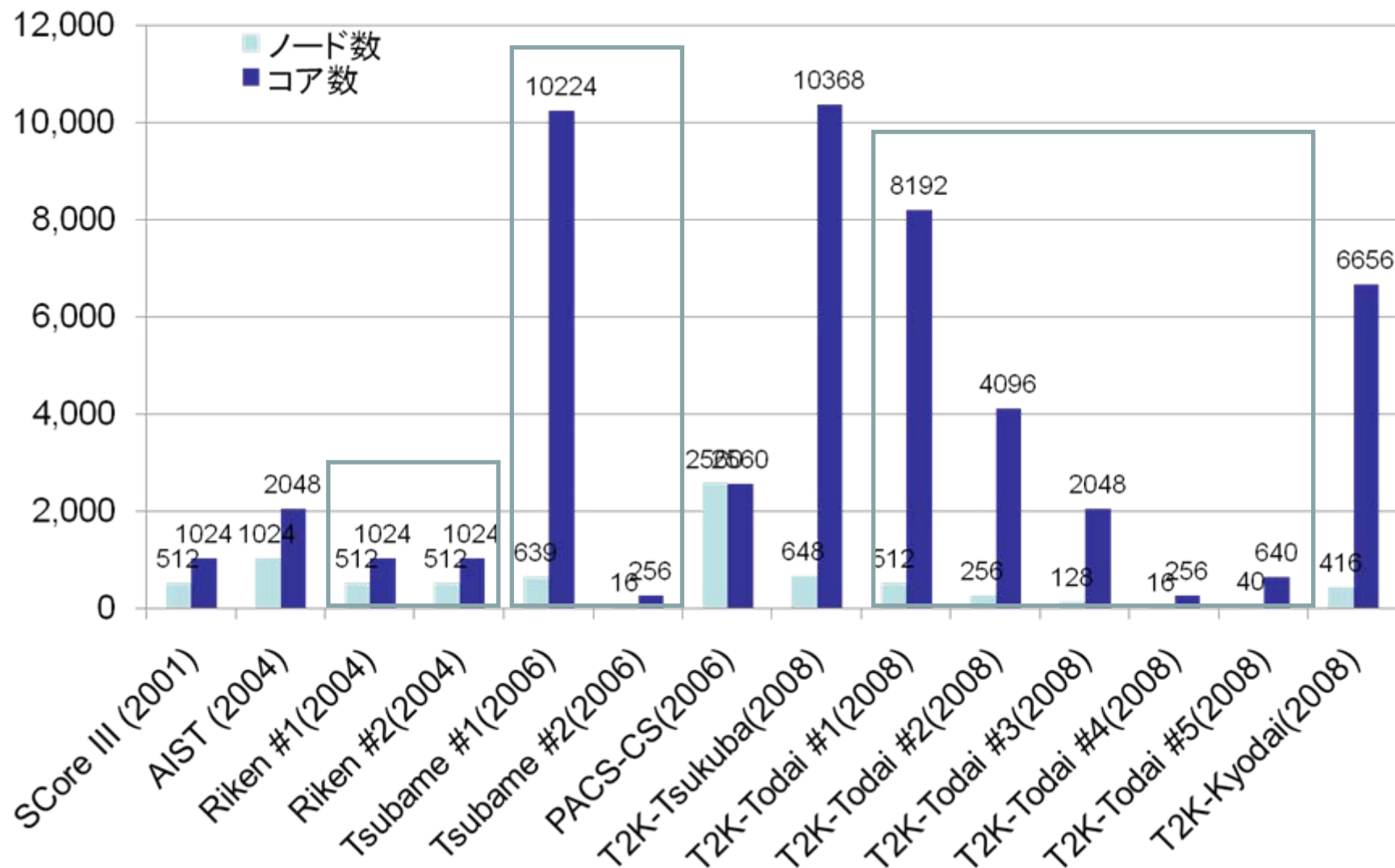
民間企業

クラスタの将来

- 大規模化
 - 一つのジョブで1PFlops必要とするアプリケーション
 - 一つのジョブは1TFlops未満だが、大量のジョブを動かしたいユーザ
 - たとえば、1ケースのジョブは1TFlopsで良いが、パラメータを変えて1000ケース流したいとすれば、1 Pflops必要
- パーソナル化
 - 保守フリーなアプライアンス型クラスタ
 - 4年償却
 - パーソナル製品なら税法上も4年償却可能
- システムソフトウェアの課題
 - 資源管理、ジッタリング、メモリアフィニティ、電力管理、耐故障、ファイルシステム、プログラミング環境、セキュリティ

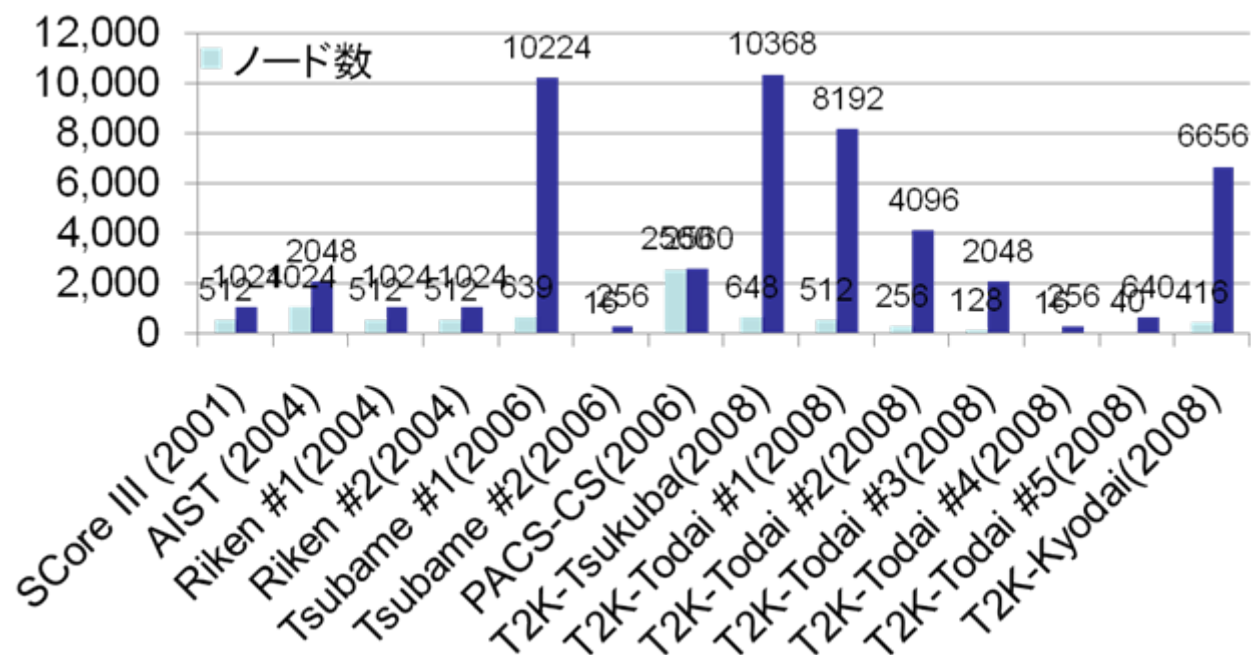
クラスタの将来: 大規模化 ノード数、コア数の変遷

- ネットワークトポロジとしてボトルネックリンクを持たないノードの塊をグループとした時のノード数
- Tsubameの構成は不明



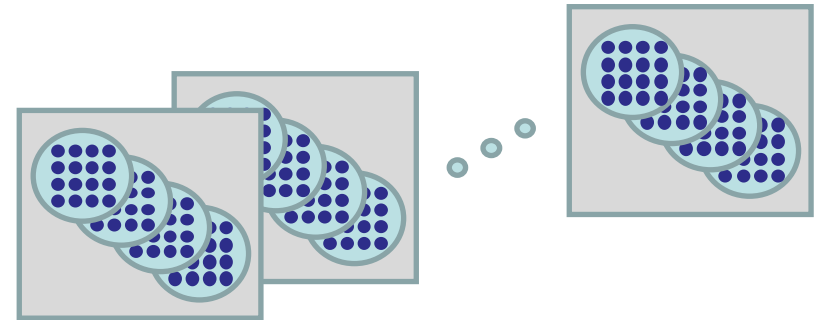
クラスタの将来: 大規模化

- (密結合された)ノード数を制限する要因
 - 設置面積 & 耐加重、空調、消費電力
- 日本の大学計算センター
 - 500平米前後
 - 1 MW前後



クラスタの将来： 大規模化

- ノード構成
 - Multicore
 - CPU+SIMD
 - GPU or accelerator connected by I/O
- ノード数
 - 1 K Nodes with 1 Tflops Node
 - 10K Nodes with 100 Gflops Node
- 総メモリ容量
 - 1 PB or $\frac{1}{4}$ PB ?
- ネットワーク



Node (Gflops)	Network (Gbyte/sec)	Byte/Flop
1000	50	0.050
1000	25	0.025

100 Gb Ethernet = 12.5 GB/sec

	Node (Gflops)	Network (Gbyte/sec)	Byte/Flop
SCore-III	1.866	0.25	0.134
Riken	12.24	1.064	0.087
TSUBAME	76.8	2	0.026
T2K	147.2	8	0.054

クラスタの将来: パーソナル化

- アプライアンス型クラスタによる保守フリー & TCO削減
- アプライアンスとは
 - 本来の英語の意味は、家庭用の設備・器具
 - 特定の機能に特別化したコンピュータを指す
- 中規模計算シミュレーション用に特化
 - 数TFlops程度のクラスタにシミュレーションソフトウェアと可視化ソフトウェア
 - 計算機室を必要としない構造
- 実現性は？
 - 市場規模の拡大が必須

T2Kオープンスパコンアライアンスの今後の連携

計算科学・工学分野の 教育・研究支援

計算科学・工学分野向け教育体系の確立と計算環境支援

- 大規模並列プログラミング
- アプリケーション利用方法
- 研究室レベルPCクラスタ利用者のスパコン利用促進



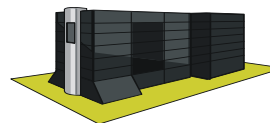
- 大規模マシン資源の提供
- プログラミング支援

先端的スパコンプログラミング環境の研究開発

計算機科学者と計算科学・工学者との連携による計算科学者・計算工学者の知恵の具現化のための道具の開発

- 並列言語処理系
- ライブラリ、チューニングツール
- 可視化ツール
- ミドルウェア
- シェル

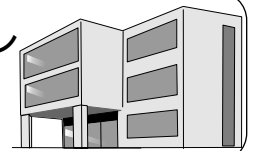
T2Kオープンスパコン



相互利用及び運用支援システム開発

グリッド技術による3大学T2Kスパコンの相互利用体制の構築並びに必要な支援ツールの開発

次世代スパコン
(ペタコン)



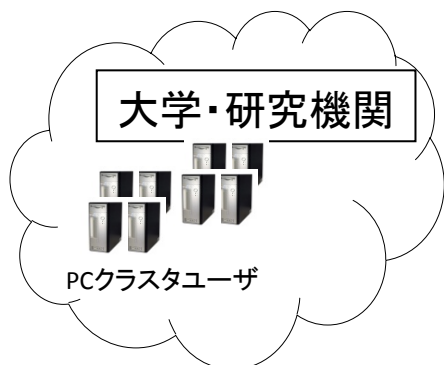
大規模処理
ユーザの橋渡し

東京大学情報基盤センターの今後

計算科学・工学分野の教育・研究支援

計算科学・工学分野向け教育体系の確立と計算環境支援

- 大規模並列プログラミング
- アプリケーション利用方法
- 研究室レベルPCクラスタ利用者のスパコン利用促進



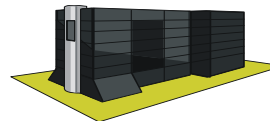
- 大規模マシン資源の提供
- プログラミング支援

先端的スパコンプログラミング環境の研究開発

計算機科学者と計算科学・工学者との連携による計算科学者・計算工学者の知恵の具現化のための道具の開発

- 並列言語処理系
- ライブラリ、チューニングツール
- 可視化ツール
- ミドルウェア
- シェル

T2Kオープンスパコン



大規模処理ユーザの橋渡し

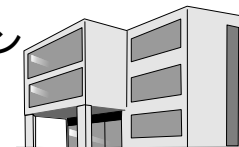
ASP (Application Service Provider) との連携

- 計算環境支援
- コンサルテーション
- 計算受託

相互利用及び運用支援システム開発

グリッド技術による3大学T2Kスパコンの相互利用体制の構築並びに必要な支援ツールの開発

次世代スパコン (ペタコン)



大学の社会貢献

イノベーション創出、大規模並列計算ニーズ拡大に向けて、民間企業支援

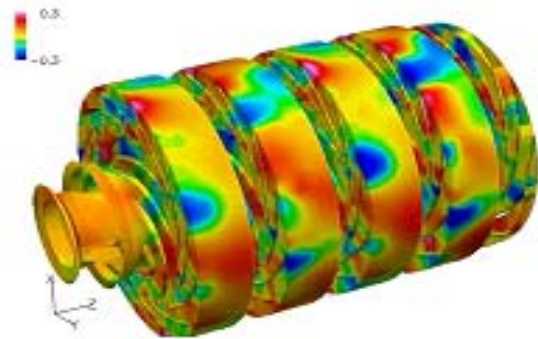
- 計算機環境の貸し出し
- プログラミング支援
- 大学発アプリケーション利用支援

民間企業

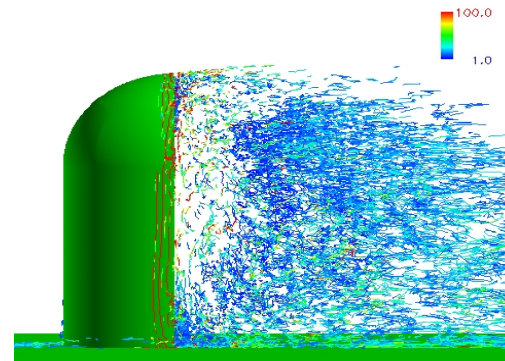
ASP例:みずほ情報総研

大規模流体解析コード FrontFlow/blue

- LESによる非定常流体解析(非圧縮)
- 「革新的ソフト」PJで東大・加藤教授を中心に開発
- ターボ機械内部流れ、乱流音騒音等を解析
- 数値計算スキーム
 - 有限要素法(六面体要素、四面体要素)
 - クランクニコルソン法、Fractional-Step法
- 並列計算機能($\sim 4 \times 10^3$ CPUs)、乱流音解析、マルチフレーム機能、オーバーセット機能、キャビテーション解析、熱輸送解析、DES解析



多段遠心ポンプ内部流れ



ドアミラー付近の流体騒音

出典: 大谷泰昭、「大規模流体解析を対象としたASP事業のフィジビリティスタディ」、
先端的大規模計算シミュレーションプログラム利用シンポジウム、主催東大他、2008年3月28
日

FrontFlow/blueの計算コスト

■ 計算時間

$$\frac{\text{演算回数}[FLOP]}{\text{マシン性能}[FLOPS]} = \frac{(20 \times 10^3) \times \text{グリッド数} \times \text{時間ステップ}}{\text{ピーク性能} \times \text{実行効率}}$$

(e.g. 20 sec/step for 10^6 grids using 1.0GFLOPS)

■ メモリ容量

10^3 words/grid → 4KB/grid
(e.g. 4GB for 10^6 grids)

■ ディスク容量

入力ファイル: 44 Bytes/grid
(e.g. 44MB for 10^6 grids)
出力ファイル: 16 Bytes/grid
(e.g. 16MB for 10^6 grids)

	CPU数	ピーク性能 (TFlops)	実行効率 (%)	計算時間
1.E+06 ケース	128	1	10	5.4時間
1.E+07 ケース	128	1	10	2.3日
1.E+08 ケース	1024	8	5	5.8日

出典: 大谷泰昭、「大規模流体解析を対象としたASP事業のフィジビリティスタディ」、
先端的大規模計算シミュレーションプログラム利用シンポジウム、主催東大他、2008年3月28
日

みずほ情報総研の 超大規模計算ASP事業展開

超大規模計算の潜在ニーズ

解析対象	利用分野	潜在ニーズ
固体	ものづくり全般、エネルギー	大規模構造系、振動、音場、破壊
流体	ものづくり全般、エネルギー、環境	乱流、騒音、混層流、界面変化、燃焼
電磁場	機械、エネルギー、電気・電子	タービン、磁気ヘッド、光学材料（核融合）
原子・分子	素材、化学、半導体、バイオ	ナノ計測、材料物性、創薬、量子素子
データ	気象、バイオ、経済	Bioinformatics、Data Mining、データ同化

出典：大谷泰昭、「大規模流体解析を対象としたASP事業のフィジビリティスタディ」、
先端的大規模計算シミュレーションプログラム利用シンポジウム、主催東大他、2008年3月28
日

まとめ

1. クラスタコンピューティングの歴史
 - 黎明期からセンター運用
2. T2Kオープンスパコンアライアンス
 - 3大学計算センターの計算機科学分野教授陣による推進
3. クラスタコンピューティングの将来
 - 大規模化 & パーソナル化
 - システムソフトウェアの課題
4. T2Kアライアンスの将来
 - 計算科学・工学研究者との連携
 - 教育支援
 - グリッド運用
5. 東大情報基盤センターの今後
 - ASPとの連携