'*Ts' for HPC - インテル・テクノロジのHPCにおける価値の考察

スケーラブルシステムズ株式会社
代表取締役　戸室　隆彦

スケーラブルシステムズ株式会社

1986　日本クレイ株式会社入社
　　SE、セールスサポート、マーケティングサポートなどの活動と技術面で会社をリードしています
1996　日本SGI株式会社（SGIのCray買収により）
　　SEディレクター、製品技術本部長など
2003　執行役員チーフテクノロジーオフィサー
　　SGI製品はもちろん、広範囲な技術動向についてお客様へのご紹介や各社とのアライアンスの活動を行いました。
2005　スケーラブルシステムズ株式会社設立

SSTC High Performance and Productivity スケーラブルシステムズ株式会社　2005

Scalable Systems

スケーラブルシステムズは、CRAYとSGIでの豊富なHPC関連の経験を生かして、新たなソリューションをご提供します。

2000　2005

1985　1990　1995　1996

CRAY Research Inc.

ベクトル計算機、MPPシステム、スーパーサーバ（SUN互換機）などの様々なアーキテクチャのシステムでのHPCソリューションの提供のための活動を行ってきました。ベクトル処理、並列処理での最先端技術の日本への紹介も行っています。

Silicon Graphics

初めての商用DSM（分散共有メモリシステム）や大規模NUMAシステムでのHPCソリューションの提供をおこなってきました。

Linuxとインテルプロセッサによるスケーラブルシステムの製品化とそのシステムの導入支援を行っています。

温故知新

故きを温ねて新しきを知れば、以て師と為るべし

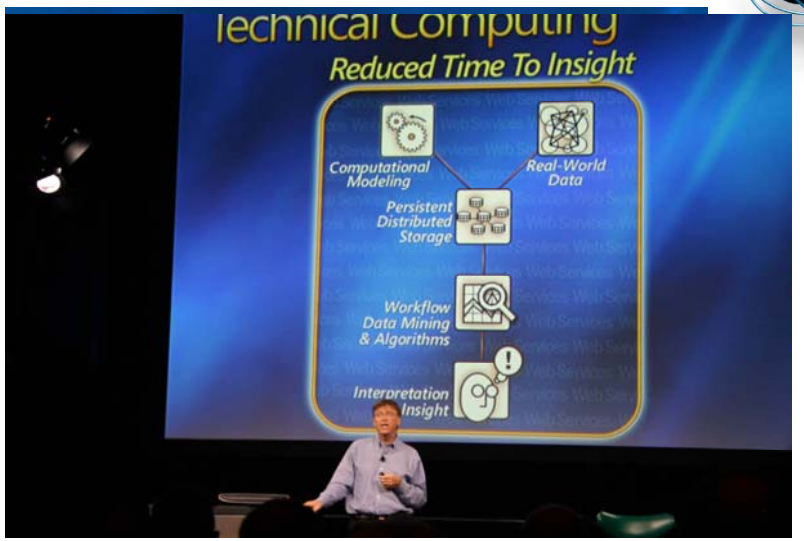
温故知新

- はじめに
- HPCシステムの歴史
- HPCシステムの課題
  - ソフトウエア
  - ハードウエア
  - マイクロプロセッサ
- 4.　'*Ts' for HPC - インテル・テクノロジのHPCにおける価値の考察

HPCシステム

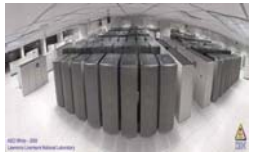
- HPCからHPMS (High-Performance Modeling and Simulation)
  - 計算システム＋ストレージ＋可視化の統合システム
  - High Performance と High Productivity
- Capability（単一ジョブの高速処理）.vs. Capacity（複数ジョブの多重処理）
- ハイエンドコンピューティングに関する課題
  - プログラミングモデル(Programming Productivity - Safety, Portability, Performance, Integrationなど）
  - 仮想化、IO、OS、APIなど様々か課題
- マイクロプロセッサの動向の変化

このスライドは誰が?



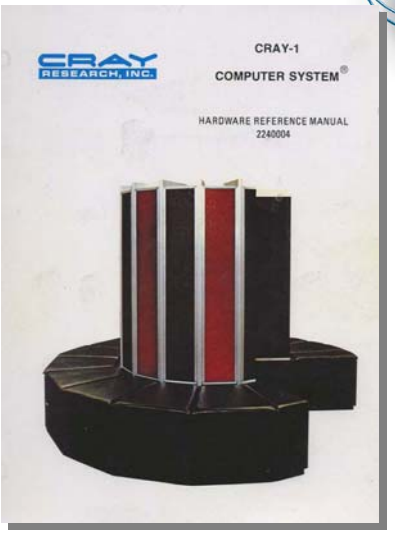
HPCの歴史

Episode I
The Phantom Menace



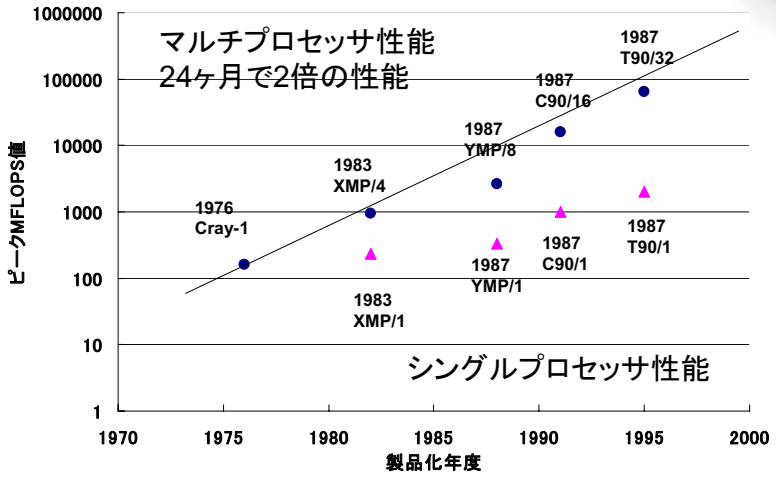
Cray システム

- Cray-1 (1977)
  - 250 MFLOPS
  - 80 MHz
  - 1 MWord (64-bit)
- PC 8088 (1979)
  - 5 MHz
  - 1 MB RAM
- Modern PC (Pentium 4)
  - 3.2 GHz （Dual Core)
  - 12.8 GFLOPS
  - 4 GB RAM

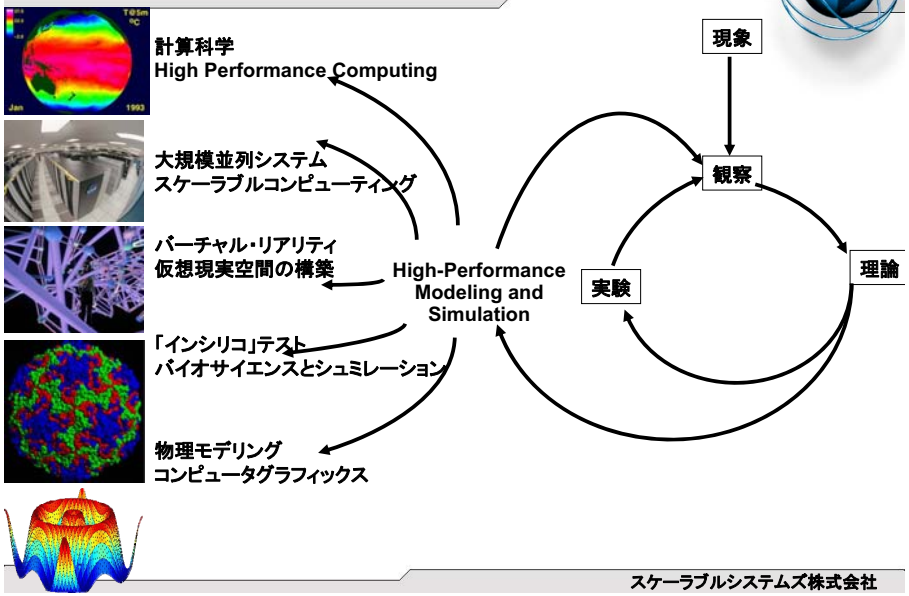


http://ed-thelen.org/comp-hist/CRAY-1-HardRefMan/CRAY-1-HRM.html

Crayシステム:ピーク性能



HPMS (High-Performance Modeling and Simulation)



シングルプロセッサ性能:Linpack

Linpack MFLOPS

UNIT = 10**6 TIME/(1/3 100**3 + 100**2)

Facility	TIME N=100 secs.	UNIT micro- secs.	Computer	Type	Compiler
NCAR	14.8 .049	0.14	CRAY-1	S	CFT, Assembly BLAS
LASL	4,64 .148	0.43	CDC 7600	S	FTN, Assembly BLAS
NCAR	3.5 .192	0.56	CRAY-1	S	CFT
LASL	3.27 .210	0.61	CDC 7600	S	FTN
Argonne	2.31 .297	0.86	IBM 370/195	D	H
NCAR	1.91 .359	1.05	CDC 7600	S	Local
Argonne	1.77 .388	1.33	IBM 3033	D	H
NASA Langley	1.40 .489	1.42	CDC Cyber 175	S	FTN
U. Ill. Urbana	1.36 .506	1.47	CDC Cyber 175	S	Ext. 4.6
LLL	1.24 .554	1.61	CDC 7600	S	CHAT, No optimize
SLAC	1.19 .579	1.69	IBM 370/168	D	H Ext., Fast mult.
Michigan	1.09 .631	1.84	Amdahl 470/V6	D	H
Toronto	.772 .890	2.59	IBM 370/165	D	H Ext., Fast mult.
Northwestern	.477 1.44	4.20	CDC 6600	S	FTN
Texas	.356 1.93*	5.63	CDC 6600	S	RUN
China Lake	.352 1.95*	5.69	Univac 1110	S	V
Yale	.265 2.59	7.53	DEC KL-20	S	F20
Bell Labs	.197 3.46	10.1	Honeywell 6080	S	Y
Wisconsin	.197 3.49	10.1	Univac 1110	S	V
Iowa State	.194 3.54	10.2	Itel AS/5 mod3	D	H
U. Ill. Chicago	.144.10	11.9	IBM 370/158	D	G1
Purdue	.094 5.69	16.6	CDC 6500	S	FUN
U. C. San Diego	.062 13.1	38.2	Burroughs 6700	S	H
Yale	.0[illegible] 17.1*	49.9	DEC KA-10	S	F40

* TIME(100) = (100/75)**3 SGEFA(75) + (100/75)**2 SGESL(75)

ベクトル計算機の性能

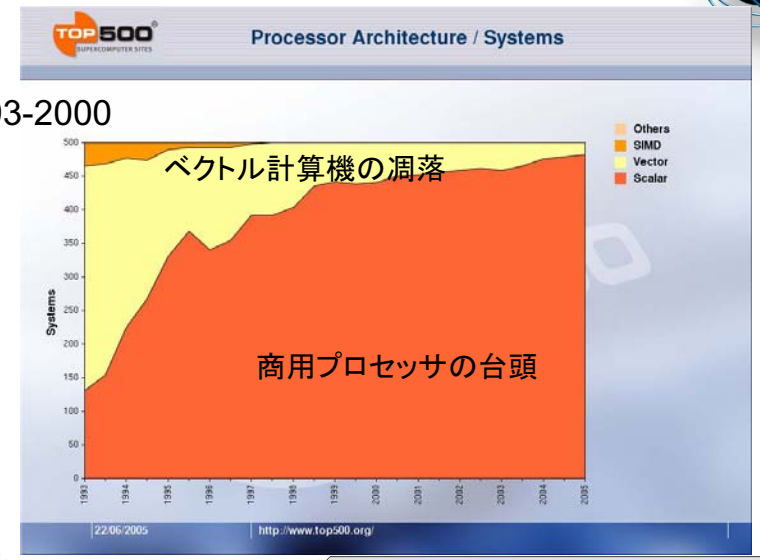
Q: なぜ、ベクトル計算機の性能が、マイクロプロセッサの性能のように向上しなかったのでしょうか?

A: ベクトル計算機は、グローバル共有メモリに対する高い接続性能にその性能が依存していたために、このメモリ間接続の性能向上がボトルネックとなってしまいました。

例:DRAMメモリの性能と仕様

1979:標準DRAM	1999:200 MHz SDRAM	1979→1999
16K bit	256 Mbit	X 16000
1-bit wide interface	16-bit wide interface	X 640
5 Mb/s uniform access BW	3200 Mb/s uniform access BW	X 500
2 Mb/s random access BW	1000 Mb/s random access BW	X 25

The Pahntom Menace



1993-2000

ベクトル計算機の凋落

商用プロセッサの台頭

ベクトル計算機の逆襲

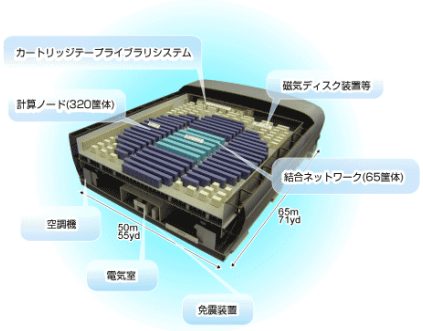
Episode V

The Empire Strikes Back

Sputnik: October 4, 1957

ベクトル計算機の逆襲

- 2002
- 地球シュミレータ
- コンピュータにおける
  スプートニックショック



HPCの歴史

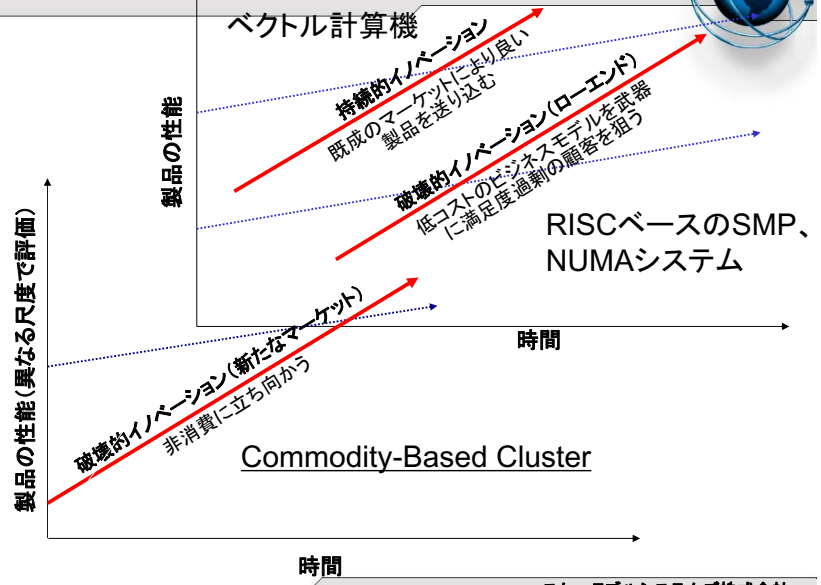
Episode II

Attack of the Clones

イノベーションのジレンマ

- クレイトン・クリステンセンの「イノベーションのジレンマ」
- 持続的イノベーションと破壊的イノベーションによるマーケットの動向を分析
- 持続的イノベーション
  - 技術革新が顧客の求める性能向上軸に沿っている
- 破壊的イノベーション
  - 既存顧客が求める性能とは異なる軸の性能（特性）

破壊的イノベーション



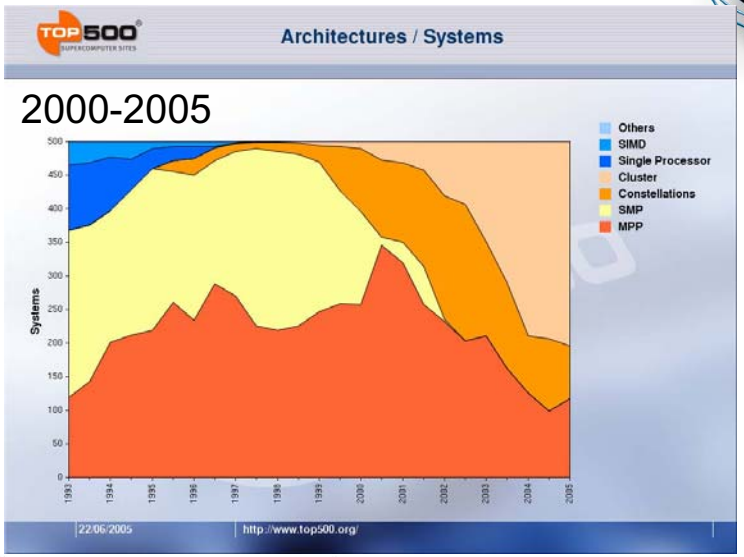
Beowulf プロジェクト

- Wiglaf - 1994
- 16 Intel 80486 100 MHz
- VESA Local bus
- 256 Mbytes memory
- 6.4 Gbytes of disk
- Dual 10 base-T Ethernet
- 72 Mflops sustained
- $40K

- Hrothgar - 1995
- 16 Intel Pentium100 MHz
- PCI
- 1 Gbyte memory
- 6.4 Gbytes of disk
- 100 base-T Fast Ethernet (hub)
- 240 Mflops sustained
- $46K

- Hyglac-1996 (Caltech)
- 16 Pentium Pro 200 MHz
- PCI
- 2 Gbytes memory
- 49.6 Gbytes of disk
- 100 base-T Fast Ethernet (switch)
- 1.25 Gflops sustained
- $50K

クラスタシステムの台頭

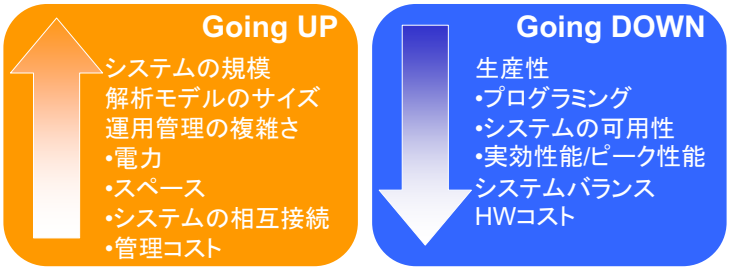


課題

Episode III

Revenge of the sith

HPCの現状



HPCマーケット

- HPCマーケットでのHPCシステム構築及び製品は、次の3つのセグメントに分かれている
  - 一般商用システム（ Commodity-based systems）
    - 一般のクラスタシステム（Dell HPCCなど）
  - 付加価値システム（ Value-based systems）
    - 多くのSMPやNUMAシステム（SGI Altixなど）
  - 特定目的システム（ Purpose-built systems）
    - アプリケーションと解析対象に合わせたシステム設計（IBM BlueGene/Lなど）
- IDCなどのレポートでも、一般商用システムのHPCマーケットでの導入がもっともその成長が大きい
  - 付加価値システムの課題（一般商用システムとの競合に対する対応、もしくは、新たな分野の開拓→ペタスケールコンピューティング）
  - HPCSプログラムは、この付加価値システムのベンダーにとっても、生き残りを賭けた戦い？ （2006、July)

HPCシステムの現状分析

- Good News！

  “HPCシステムにおける問題は、たった2つだけである”

ソフトウエアとハードウエア

- ソフトウエア：The Law of More…..
  - システム規模とその複雑さの急速な増加・拡大
  - ソフトウエアの準備が出来た時点でハードウエアは既に陳腐化し、次のシステムの導入の検討が進む・・
- ハードウエア：Moore's Law (ムーアの法則)
  - 消費電力の問題のため、プロセッサの動作クロックを今までのペースで上げることは困難
  - プロセッサとメモリの性能差の拡大によるCPUサイクルとのギャップ
  - ピーク性能と実効性能のギャップの拡大

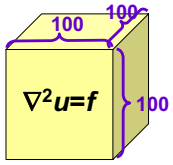
ソフトウエア：The Law of More…

- 研究者は、より多くの時間（More Time)をソフトウエアの開発のために必要としている
- 問題はより複雑（More Complex)になり、そして、より多くのプロセッサ(More Processors)を利用して処理を行うには、より多くの困難(More Difficult)が伴います

アルゴリズムの最適化

- 計算機自身の進化と共に計算アルゴリズムも最適化されている
- 例：編微分方程式の解法
  - N=106の場合、ガウスの消去法で線形方程式を解く場合とMGでの計算では、108倍の計算量が違う
  - これは、1Mflops/sの計算機で、100Tflops/sの計算機に相当する計算を行ったことになる

アルゴリズム	計算オペレーション数（概数）
Banded Gauss Elimination	$O(N^{7/3})$
Gauss Seidel	$O(N^{5/3}\ log(N))$
Optimal SOR	$O(N^{4/3}\ log(N))$
CG/MILU	$O(N^{7/6}\ log(N))$
F-cycle MG	$O(N)$



ソフトウエア：The Law of More…

- 一般の商用製品を活用したクラスタソリューションでは、「Capacity」の実現は容易であるが、「Capability」の実現については依然として課題が多い
  - コストパフォーマンスの高いシステムの構築は可能だとしても、コストプロダクティビティの高いシステムの構築も課題
- 数百～数千プロセッサ構成のシステムの利用技術と解析対象の検討
  - 小規模、中規模問題の高速処理への対応
  - ソフトウエア開発の生産性
- 数プロセッサ～数十プロセッサをより簡便に、容易に利用できる技術
  - シングルプロセッサ、シングルスレッドを利用するのと同じように…..

ソフトウエアとハードウエア

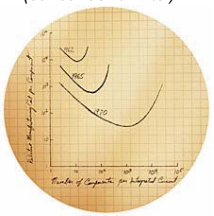
- ソフトウエア：The Law of More…..
  - システム規模とその複雑さの急速な増加・拡大
  - ソフトウエアの準備が出来た時点でハードウエアは既に陳腐化し、次のシステムの導入の検討が進む・・
- ハードウエア：Moore's Law (ムーアの法則)
  - 消費電力の問題のため、プロセッサの動作クロックを今までのペースで上げることは困難
  - プロセッサとメモリの性能差の拡大によるCPUサイクルとのギャップ
  - ピーク性能と実効性能のギャップの拡大

計算機の性能向上

- 動作周波数（クロック）の向上
  - 過去12年間で、Pentiumプロセッサの動作周波数は、60 MHz から 3,800 MHz にまでアップ
  - 現在までの高性能化の約80% はクロック周波数の向上によるもの

ハードウエアの問題
Moore's Law：ムーアの法則

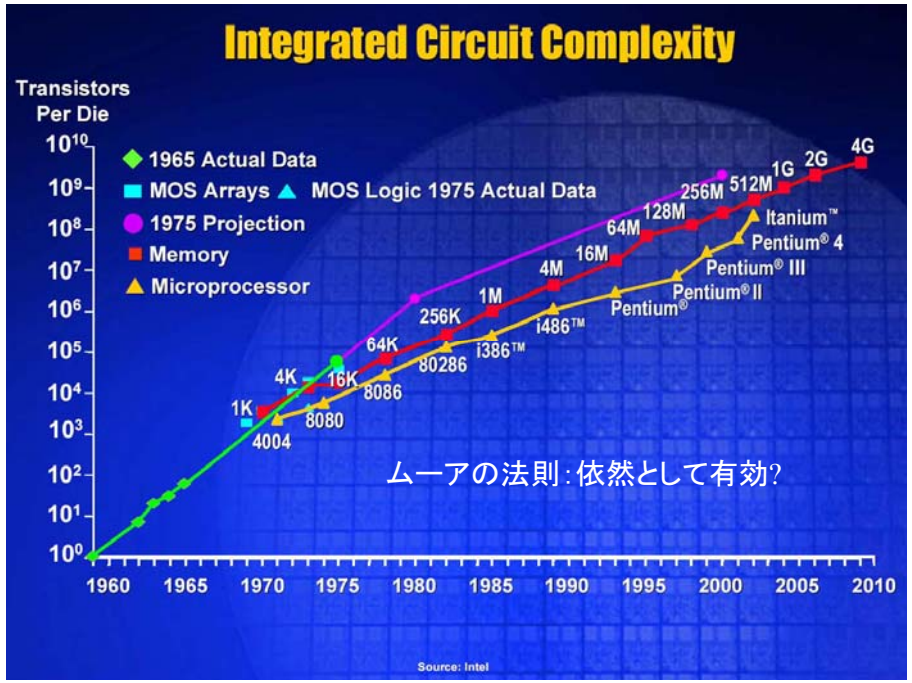
Dr. Gordon Moore
(co-founder of Intel)



- インテルの共同設立者の1人である Gordon Moore 博士が、1965年4月19日号の**「Electronics」**誌に投稿した、「一定面積に集積されるトランジスタの数は12か月で倍増し、それに伴いトランジスタの動作速度が向上する」という予測 (その後、1975年に Moore 博士はチップの複雑化を考慮してトランジスタ数の倍増ペースを24か月に修正)
- また、一般にはあまり知られていないがテクノロジの進歩とともに製造コストが劇的に下落することも予測（左図）

指数関数的成長は永遠には続かない。しかしその永遠を先延ばしにすることはできる

http://www.intel.co.jp/jp/developer/technology/silicon/mooreslaw/index.htm



性能向上の源泉は？

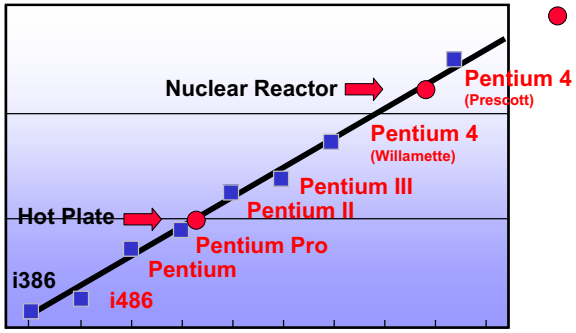
ハードウエアデバイス技術の進歩
- ロジック回路のスイッチング速度の向上とデバイス密度
- メモリサイズの拡大とアクセス速度の向上
- 通信性能（バンド幅とレイテンシの向上）

コンピュータ・アーキテクチャ
- 命令発行・実行速度の向上
  - パイプライン化
  - 分岐予測
  - キャッシュ
  - Out-of-order など
- 並列性
  - 1サイクルでの命令実行数
    - 命令レベルでの並列性（ILP）
    - ベクトル処理
  - プロセッサあたりコア数
  - ノードあたりのプロセッサ数
  - システムあたりのノード数

GHz競争

- 2000年に開催されたIEEE国際電子デバイス会議2000（2000 IEEE International Electron Devices Meeting：IEDM）において、インテル社は4億個以上のトランジスタを集積した、10GHz駆動のプロセッサが2005年までに実現可能だと発表しました。
  - 実際には、インテル社の最速プロセッサは、6ヶ月前に発表された3.8GHz（Intel Pentium 4)となっています。
- Prescottプロセッサの6xxシリーズ発表に際して、インテル社は、“adding value beyond GHz” のコメントを出しています。それ以降、インテル社の多くのドキュメントやプレスリリースは、この“adding value beyond GHz” についての内容を含んでいます。

発熱の問題が深刻化

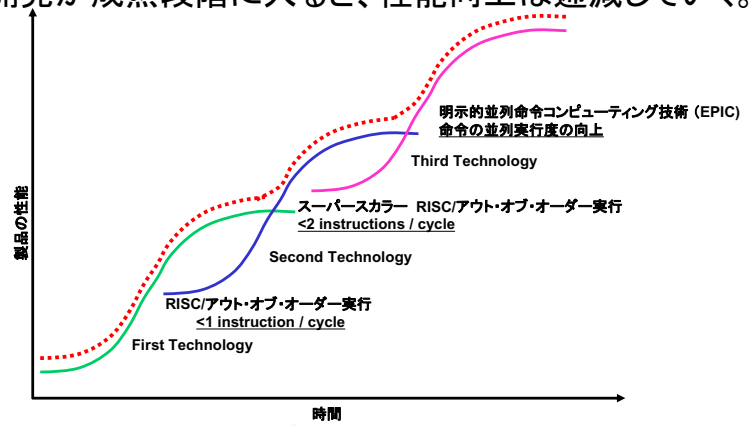


計算機の性能向上

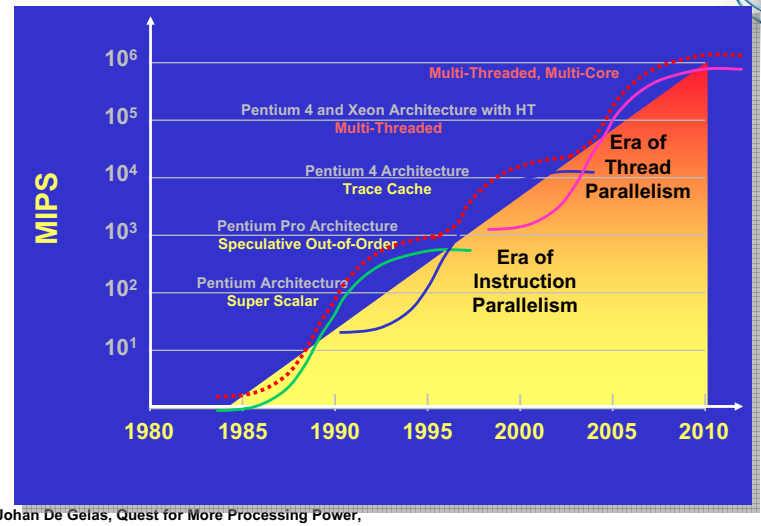
- 動作周波数（クロック）の向上
  - 過去12年間で、Pentiumプロセッサの動作周波数は、60 MHz から 3,800 MHz にまでアップ
  - 現在までの高性能化の約80% はクロック周波数の向上によるもの
- 命令実行の強化と最適化
  - より強力なインストラクションセット
  - 命令実行の最適化（パイプライン化、分岐予測、複数命令の同時実行、命令実行順序の変更など）

技術のSカーブ
技術開発の進展と製品性能の成長の関係

- 技術開発の初期は製品性能はゆっくりと向上するが、しだいに性能の向上の幅が大きくなる。しかし次第に技術開発が成熟段階に入ると、性能向上は逓減していく。



マイクロアーキテクチャのSカーブ



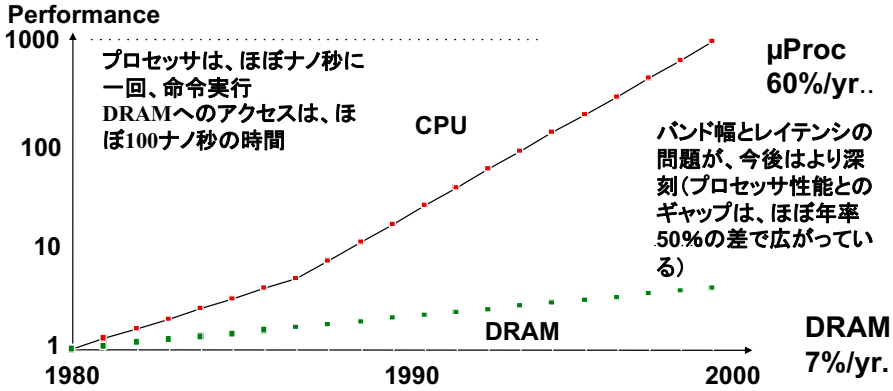
Johan De Gelas, Quest for More Processing Power,
AnandTech, Feb. 8, 2005.
http://www.anandtech.com/cpuchipsets/showdoc.aspx?i=2343

計算機の性能向上

- 動作周波数（クロック）の向上
  - 過去12年間で、Pentiumプロセッサの動作周波数は、60 MHz から 3,800 MHz にまでアップ
  - 現在までの高性能化の約80% はクロック周波数の向上によるもの
- 命令実行の強化と最適化
  - より強力なインストラクションセット
  - 命令実行の最適化（パイプライン化、分岐予測、複数命令の同時実行、命令実行順序の変更など）
- 大容量キャッシュ
  - プロセッサの速度とメモリレイテンシ（待ち時間）とバンド幅のギャップの拡大に対する対策・対応としての容量の拡張

性能ギャップの問題

- プロセッサ速度とメモリアクセスの速度差によって、プロセッサがより高速になったとしても、プロセッサはその演算能力を完全に使い切ることが出来ない



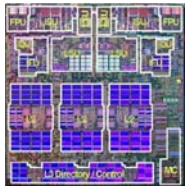
今後の発展と課題

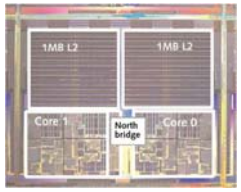
Episode IV
A New Hope

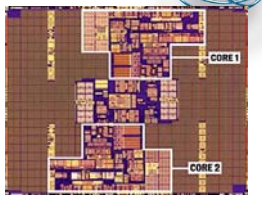
計算機の性能向上

- 動作周波数(ク
  - 過去12年 entiumプロセッサ 作周波数は、60 MHz か 00 MHz にまでアッ
  - 現在ま 高性能化の約80% ック 数の向上に よるも
- 命令実 の強化と最適
  - より強 なインストラ ンセット
  - 命令実 の最適 イプライン化、分岐 、複数命令の同時 、実行順序の変更など)
- 大容量キャ
  - プロセッサの (モリレイテ っ時間)とバンド幅のギャップの拡大 応としての容量の拡張

デュアルコアプロセッサ



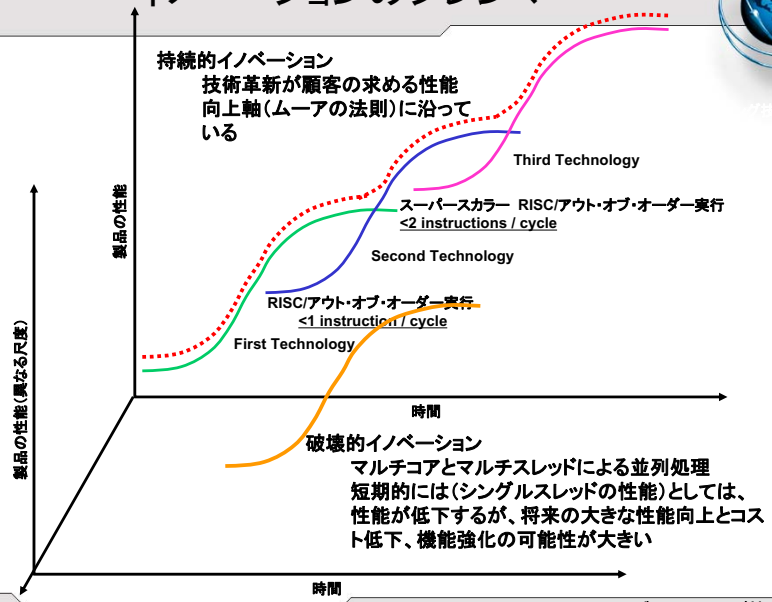




- チップ上のトランジスタのより有効活用が可能
- スレッドレベルでの並列処理を活用
- よりシンプルなプロセッサの設計が可能

- 将来のマイクロプロセッサはより多くのコアを実装可能
- 将来のマイクロプロセッサはより大容量のキャッシュの実装が可能

イノベーションのジレンマ



マルチコアの利点?

ワークロードの処理効率の向上

- マルチスレッドアプリケーション
  - 現在、多くのアプリケーション(データベース、WEB、科学技術計算)はマルチスレッド化
  - マルチコアプロセッサでは、これらのアプリケーションのマルチスレッドでの実行が容易に可能
- 複数ジョブの処理
  - システムでは、複数のワークロード同時に処理することが必要
  - マルチコアでは、これらのワークロードへの処理が可能

マルチコアの利点?

消費電力あたりの性能を最大にし、高性能で低消費電力のシステム構築が可能

- OS自身のマルチスレッド対応
  - OSのサービスもマルチスレッドで処理することで、より効率よく処理することが可能
- 仮想化
  - サーバのセキュリティや管理の強化
  - 管理するノード数を減らし、運用コストの削減を図る
- 最新のソフトウエア・テクノロジの活用

大きな変革・・しかし、容易ではない

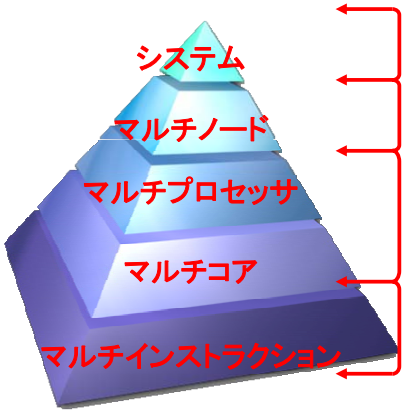
マルチコアプロセッシング(または、汎用もしくは専用プロセッサをソケットに複数搭載可能なこと)は、Ethernetの誕生以来、ITインフラに対しての大きなインパクトをもたらします。

— Multicore Processing: Disruption or Distraction for the IT Infrastructure?, Vernon Turner, IDC, November 18, 2004.

デュアルプロセッサは、386プロセッサの発表以来、性能に関して最大の向上を実現します。しかし、このような性能向上には、ソフトウエアの最適化がプロセッサの性能をフルに発揮するためには必要です。

— Readying Applications for New Server Technologies, Martin Reynolds, Gartner Research, April 12, 2005.

並列性(Parallelism)の利用



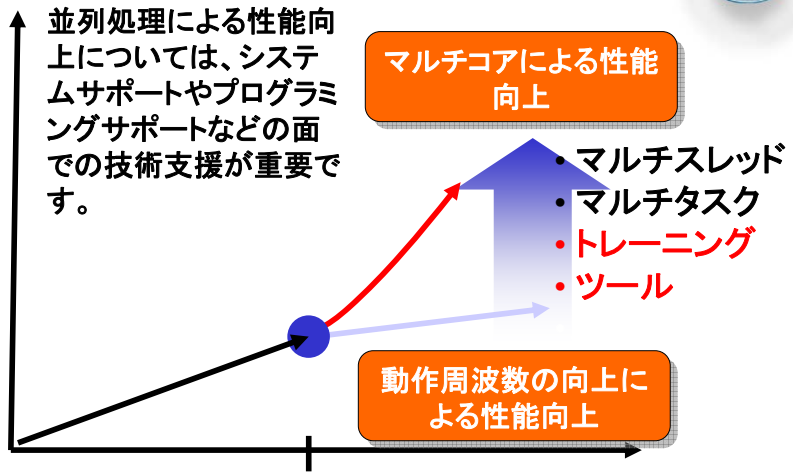
マルチタスク処理

MPIなどによる複数のノード間での並列処理

ユーザ及びコンパイラによるスレッドレベルでの並列処理(TLP)

コンパイラによる命列実行レベルでの並列処理(ILP)

ムーアの法則（GHz から MC へ）



Episode VI

Return of the Jedi

将来予測の難しさ

- “I think there is a world market for maybe five computers.”
  - Thomas Watson, chairman of IBM, 1943.
- “There is no reason for any individual to have a computer in their home”
  - Ken Olson, president and founder of digital equipment corporation, 1977.
- “There are only about 100 potential customers worldwide for a Cray-1”
  - Seymour Cray, 1977.
- “640K [of memory] ought to be enough for anybody.”
  - Bill Gates, chairman of Microsoft,1981.

「未来を予測する最良の方法は、それを創造してしまうことである」

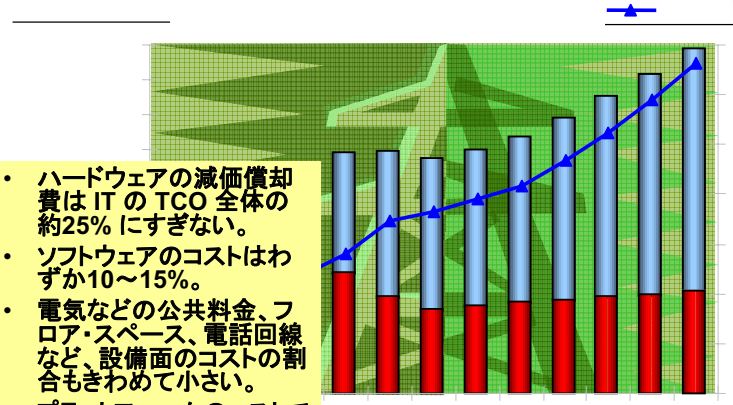
"The best way to predict future is to invent it."

Dr. Alan Kay, President of Viewpoints Research Institute, Inc.,

ITマネージメントの課題

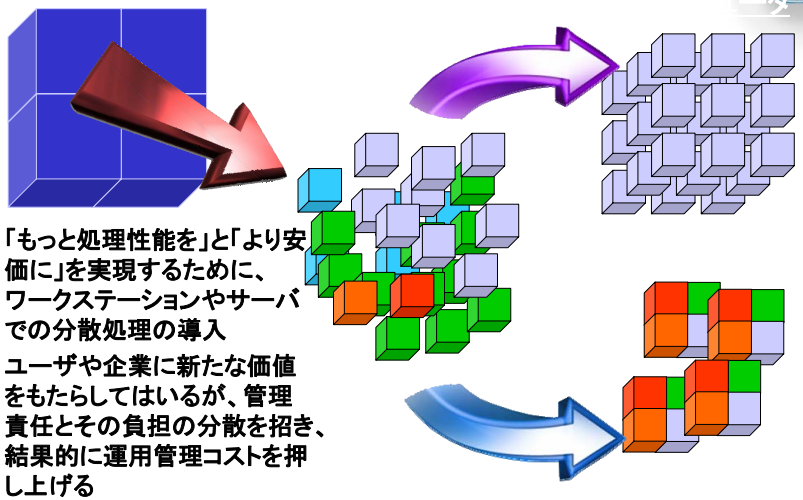
- プラットフォームの内部からの保護:
  - ウイルスやワームなど悪意あるソフトウェアからの保護
- 資産管理:
  - 多くの IT 部門では、特定できない資産が問題
- オンラインおよびリモート管理・診断機能:
  - アップグレード、診断、復旧のための作業の効率化
- アプリケーション統合の困難さ:
  - アプリケーションの高度化と複雑化によって、複数のアプリケーションを組み合わせるての動作に問題
- 動的なリソース割り当て:
  - 組織内で未使用のCPUやメモリの活用

マーケットトレンド



- ハードウェアの減価償却費は IT の TCO 全体の約25% にすぎない。
- ソフトウェアのコストはわずか10～15%。
- 電気などの公共料金、フロア・スペース、電話回線など、設備面のコストの割合もきわめて小さい。
- プラットフォームのコストではなく、TCOの大きな比率を占めるのは人件費となっている。

運用管理コストの低減

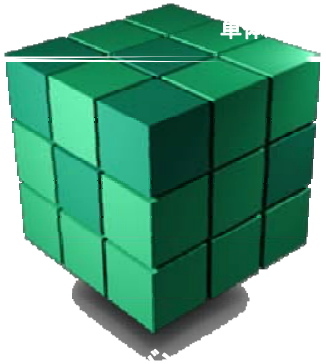


- 「もっと処理性能を」と「より安価に」を実現するために、ワークステーションやサーバでの分散処理の導入
- ユーザや企業に新たな価値をもたらしてはいるが、管理責任とその負担の分散を招き、結果的に運用管理コストを押し上げる

次世代HPCインフラ

- コアとスレッド
  - より多くのスレッドを効率よく利用可能
  - マルチスレッド向け最適化
- 電力管理
  - 省電力
  - データセンター運用管理機能
- 仮想化
  - 柔軟性と優れた運用管理
  - 仮想的なシステムパーティション
- RAS
  - ハードウエアベースの自己監視/自己管理
  - ファームウエアベースのエラー履歴管理
- システム管理
  - より低いTCOを実現するための一般・標準化されたマネージメント機能

システムの‘バランス’



インテルにおけるプラットフォーム

- 「特定の**利用モデル**を実現するために**さまざまな構成要素を統合**したものであり、これによって既存市場の成長と新市場の創出を図り、**構成要素個々の合計を超えた利点をエンドユーザに提供**するもの」
- インテル社ホームページより抜粋
  - http://www.intel.co.jp/jp/platforms/whatis.htm

インテル・プラットフォームの構成要素

- ハードウェア:
  - プロセッサ、チップセット、通信モジュール、メモリ、ボード、システムなど
- ソフトウェア:
  - オペレーティング・システム (OS)、アプリケーション、ファームウェア、コンパイラなど
- **テクノロジ: *Ts（スター・ティーズ）**
  - **ハイパー・スレッディング (HT) テクノロジ、インテル® バーチャライゼーション・テクノロジ、インテル® I/O アクセラレーション・テクノロジ (インテル® I/OAT)、インテル® アクティブ・マネジメント・テクノロジ (インテル® AMT) など**
- 構想と標準規格:
  - Wi-Fi*、WiMAX*、ワイヤレス・アクセス検証プログラム (Wireless Verification プログラム) など
- サービス:
  - デジタル・メディア配信、通信サービス、システム・マネジメント・サービスなど

インテルテクノロジ

テクノロジ	説明
64-bit computing / Intel EM64T	x86アーキテクチャをベースに64bit拡張を行ったアーキテクチャ
Demand Based Switching (DBS) with EIST	ノートPC向けプロセッサで採用されている省電力技術「拡張版Intel SpeedStep Technology」をサーバ向けプロセッサに応用した技術。プロセッサ負荷に応じて、動作クロックと電圧を制御することで、プロセッサの消費電力を低減する。インテルによれば、最大30%程度の省電力効果があるという。
FBD (Fully Buffered DIMM) Memory	FB-DIMMは、DIMM基板上にAMB(Advanced Memory Buffer)と呼ぶチップを実装することで、DDR2などの既存のDRAMチップを利用しながら、DIMMとメモリ・コントローラ間をPCI Expressをベースにしたシリアル・インターフェイスで接続可能にするものだ。メモリ・コントローラ側のインターフェイスとメモリ・デバイス側のインターフェイスが分離されるため、AMBチップを交換することにより、1つのチップセットで複数世代のDRAMに対応可能になる。例えば、当初のFB-DIMMはDDR2 DRAMが採用されるが、FB-DIMMがDDR3 DRAMなどに移行しても、メモリ・コントローラ（チップセット）を変更せずに対応することもできる。
*Ts（スター・ティーズ）	Hyper-Threading Technology（HT）、Intel Extended Memory 64 Technology（EM64T）、Intel Virtualization Technology（VT）、Intel Active Management Technology（iAMT）、Intel I/O Acceleration Technology（I/OAT）、LaGrande Technology（LT）など、インテルがプロセッサやプラットフォームに投入済み／予定の技術の総称。
Intel Virtualization Technology (VT)	システムの仮想化を支援するプロセッサに実装される技術。最終的には完全に独立した、すべての仮想マシンが等しく扱われる完璧な仮想化を実現する見込みだが、その前段階としてフルスペックの仮想マシンと一部機能が制限された仮想マシンが並立する2段階での実装が予定されている。

http://www.atmarkit.co.jp/fsys/keyword/016server_keyword2006/016server_keyword2006.htm

インテルテクノロジ

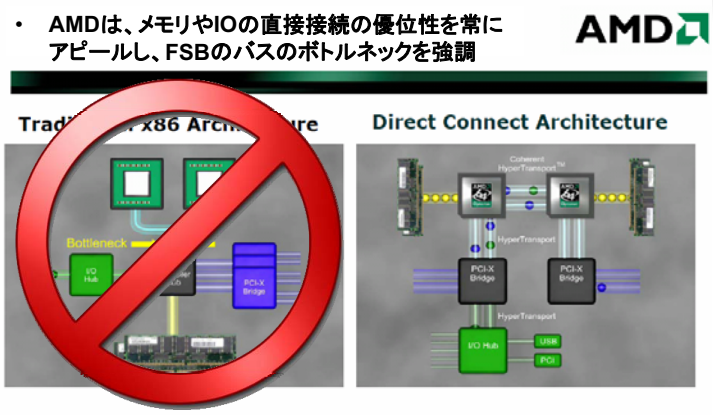
テクノロジ	説明
Intel I/O Acceleration Technology (I/OAT)	ネットワーク・コントローラ、デバイス・ドライバ、アプリケーション間の通信を、プロセッサ、ネットアーク・コントローラ、デバイス・ドライバ、メモリ・コントローラ、ソフトウェアといったすべてをコンポーネント・レベルで最適化することで、I/O処理に伴うプロセッサの負荷を軽減しようという技術。インテルによれば、ネットワーク接続されたクライアントとサーバ・アプリケーション間の通信を最大30%向上させることが可能であるという。
Intel Active Management Technology(iAMT)	さまざまなプラットフォームを対象にした遠隔地からのシステム管理を容易にする技術。iAMTにより、管理者は遠隔操作でプラットフォームにアクセスし、診断やリカバリ、インベントリ管理などが実行できるようになる。OSやシステムの電源状態に関係なく、アウトバンド通信を利用してアクセスできるハードウェア／ファームウェア・ソリューションである。
LaGrande Technology(ラグランデ・テクノロジ)	MicrosoftのNext-Generation Secure Computing Base(NGSCB)に対応するセキュリティ機能。プラットフォーム全体でセキュリティ機能を実現するには、チップセット、周辺I/Oコントローラ、OSなどトータルでのサポートが必要となるため、現時点では有効化されていない。2006年登場予定のWindows Vistaからサポートされる予定。
Pellston	エラーの発生したキャッシュ・ラインに再書き込みテストを行うことで、エラーがハードウェアによるものか否かを診断する。ハードウェア・エラーと判断された場合は、当該のキャッシュ・ラインを無効にすることで、大容量の3次キャッシュを安全に利用可能にするという技術である。
Foxton	プロセッサの消費電力(温度)と動作クロックを動的に変更することで、性能と信頼性の向上を図る。消費電力に余裕がある場合、プロセッサの動作周波数を定格以上に高めることが可能な一方で、一定の消費電力枠内で最大の性能となるよう最適化することもできる

サーバ・プラットフォーム

- 基本的に、プラットフォームの柱となるのはプロセッサ、I/O、メモリという3つの要素
  - 完全にバランスのとれたプラットフォームを実現するには、これら3つの要素がすべて同じレベルのパフォーマンスを備えることが理想
- プロセッサ
  - マルチコア、64ビット・コンピューティングなど、最近の革命的な新技術の導入によってパフォーマンスは飛躍的に向上
- ローカル I/O ＆インターコネクト
  - 業界標準規格として PCI Expressテクノロジが導入され、、I/O サブシステムは大幅にパフォーマンスが向上
- メモリ・サブシステム
  - エンタープラ イズ・プラットフォームに対応するための技術が、FB-DIMM テクノロジ

AMDプレゼンテーション

- AMDは、メモリやIOの直接接続の優位性を常にアピールし、FSBのバスのボトルネックを強調

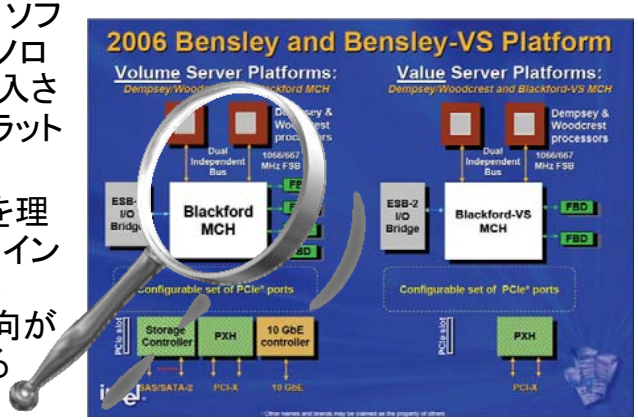


Direct Connect Architecture revolutionizes the system architecture by eliminating the bottlenecks of the front-side bus

Source：AMD

Bensley プラットフォーム

- インテルプラットフォームとして、ハードウエア、ソフトウエア、テクノロジ(＊T)が投入された新しいプラットフォーム
- このシステムを理解することで、インテルのプラットフォームの動向がより理解できる



Next Steps

HPCの二極分化

Going UP
‘Peta-Scale’
コンピューティング
•複雑なシステム構成
•新しいプログラミング
APIの提案
•アプリケーション開発

Going DOWN
‘Commodity’
コンピューティング
•商用HW/SW
•オープンソース
•パーソナルクラスタ
•商用アプリケーション
•マルチスレッド

システムとユーザの尺度

システムの尺度		ユーザの尺度
Flop/s	⇔	計算終了までの時間
メモリサイズ(GB)	⇔	モデルのサイズと計算結果
プロセッサ数	⇔	ワークロードでの試行
データ長	⇔	計算精度
システム構成(クラスタ)	⇔	導入コストと運用コスト
スケーラビリティ	⇔	ベンチマーク

- ユーザの尺度での性能(Performance)は、時間当たりにどれだけの仕事を処理出来るか(仕事量 / 時間)
- Flopsでの評価は実際には意味がない。また、問題の規模(small, medium, large) という評価も難しい。
- “スケーラビリティ”は、対象を明確に規定する必要がある

HPCシステムの動向
国家プロジェクトと商用製品のギャップの拡大

Going UP
‘Peta-Scale’
コンピューティング
•複雑なシステム構成
•新しいプログラミング
APIの提案
•アプリケーション開発

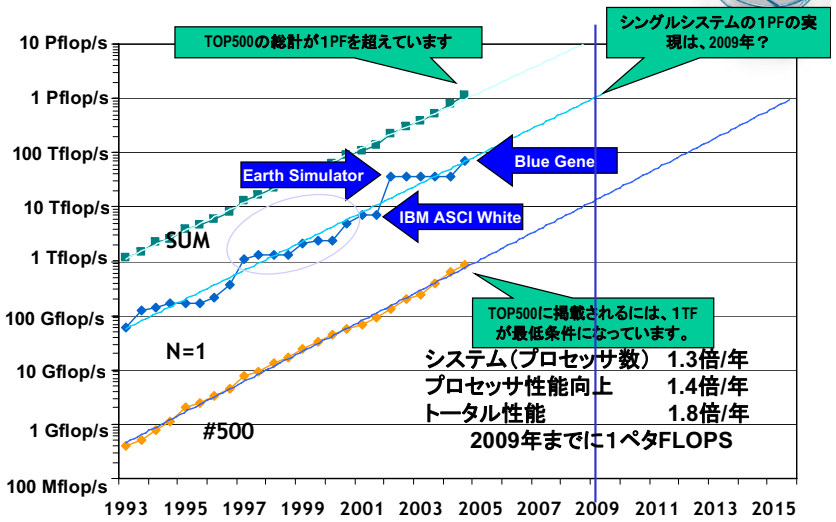
Going DOWN
‘Commodity’
コンピューティング
•商用HW/SW
•オープンソース
•パーソナルクラスタ
•商用アプリケーション
•マルチスレッド

HPCシステムの動向
国家プロジェクト

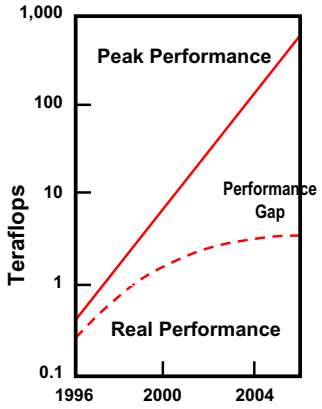
Going UP

'Peta-Scale' コンピューティング
- 複雑なシステム構成
- 新しいプログラミングAPIの提案
- アプリケーション開発

TOP500性能予測



性能ギャップの拡大



NERSC User Group Meeting June 24-25, 2004
Osni Marques and Tony Drummond
Lawrence Berkeley National Laboratory

- ピーク性能の大幅な向上
  - 1990年台は、性能の向上は、10^2のオーダーでしたが、2000年台になると10^3のオーダーで性能は向上しています。

  しかし…
  - 多くの科学技術計算用途のアプリケーションのピーク性能に対する実効性能の比率は、5-10%となっています。（1990年代のベクトル計算機は、40-50%の対ピーク性能を示していました。）
- 今、必要なのは
  - より高い実効性能を発揮することが可能な計算アルゴリズムと手法の開発とスケーラビリティの向上
  - プログラミングモデルなども含めて、スケーラブルな計算機環境の構築

ペタスケールシステムの構築

- ソフトウエア（アプリケーション、OS、プログラミングAPIなど）の課題の克服が課題
- システムの複雑さと生産性

例：
Linpack Benchmark
- オリジナルベンチマークプログラム　～100ライン
- HPL ベンチマークプログラム　～10,000ライン（x100より複雑？）

HPCシステムの動向
商用製品

標準コンポーネントの進化

- プロセッサの性能向上
  - 'マルチコア' による省電力での性能向上が可能
  - HPCアプリケーションは、容易に 'マルチコア' の利点を活用可能 （OpenMPやMPI)
- ファイルシステム
  - 高性能なスケーラブルファイルシステム（オープンソース）
- インターコネクト
  - PCI-Express (メモリ←→インターコネクト）
  - 高速の商用製品やオープンソースでの強力（OpenIBなど）

標準コンポーネントの利点

- 特定のベンダーからのシステムを組み合わせるのではなく、他社のシステムも含めてベストなシステムの選択が可能
  - スケーラブルSMP、ベクトル計算機、クラスタの幅広い選択肢
  - 64ビット、マルチコアマイクロプロセッサの性能向上を最大限に活用
- 標準コンポーネントの技術革新の活用
  - PCI-Expressや、FB-DIMMの利用技術

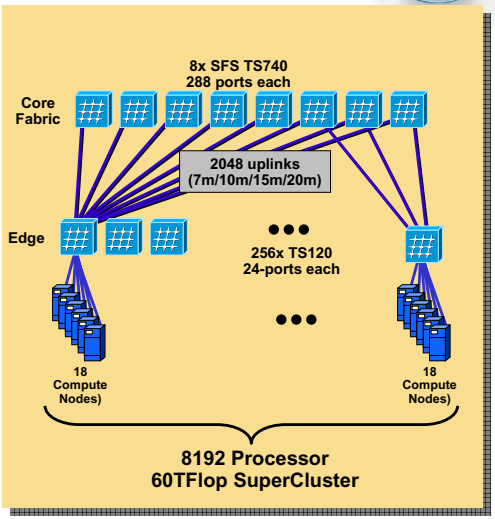
Breaking the 1-2K nodes Barrier !

- **音の障壁，サウンド・バリヤー（sound barrier）**
  飛行機の速度が音速近くになると，衝撃波の発生によって，抵抗の増大，境界層の剥離など，設計・運用上のさまざまな障害（壁）に出合って，超音速飛行は不可能かと思われた時代があった（1947年ごろまで）ので，音の障壁といわれていた。

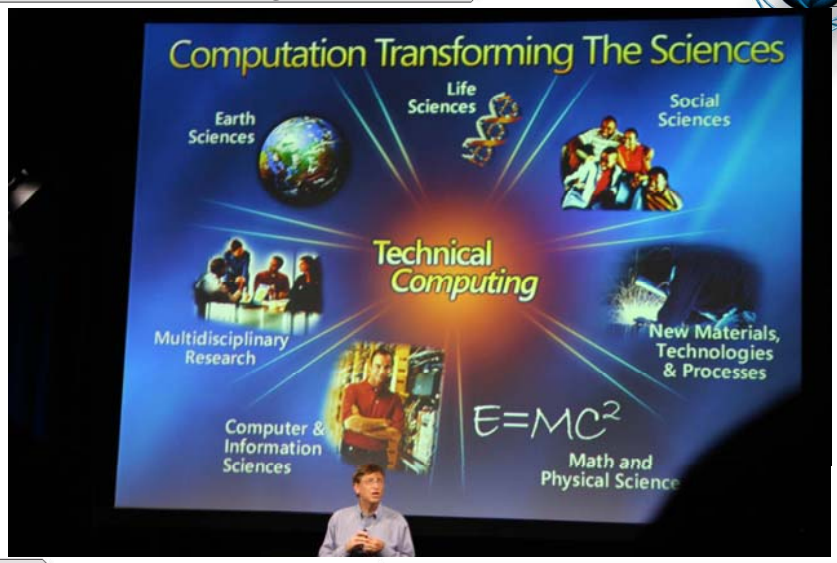
クラスタのノード数が、ある規模に近くなると、その構築や運用において、負担の増大、システムの安定稼動、スケーラビリティなど、設計・運用上のさまざまな障害（壁）に出合って，クラスタ構築は不可能と思われた時代があった（？）

米国エネルギー省 サンディア国立研究所

- システム:
  - 4096 Dell Servers
  - 50% Blocking Ratio
  - 8 TS-740s
  - 256 TS-120s
- TOP500 (Nov.5th)
  - No.5
- 用途:
  - 'Capability' クラスタ
  - 標準コンポーネントでのシステム構築



ビル・ゲイツ氏の基調講演 HPC goes mainstream



「Fast」「Good」「Cheap」のパズル

Fast + Cheap Inferior
高い性能を廉価なシステムで構築することも可能です。ただ、そのようなシステムの場合、システムの構築や利用は、必ずしも容易ではありません。

Good + Fast Expensive
付加価値の高い、性能の高いシステムは一般には、高価です。その付加価値がユーザにとって、メリットが無ければ、コスト・パフォーマンスの悪いシステムになるだけです。

Good + Cheap Slow
比較的小規模なシステムであれば、廉価で使い勝手の良いものを探すことは可能です。しかし、そのようなシステムでは、拡張性やより大規模なシステム構築が出来ません。

まとめとして

- 「テクノロジー」をどのようにとらえるか？
  - 企業経営基盤のコア要素
  - ユーザの本質的な課題を解決する戦略的な武器
- マーケットを牽引する「テクノロジ」に求められること
  - テクノロジとHPCにおけるITインフラの関係を明確にすること
    - ユーザに何らかのメリットをもたらさない「テクノロジー」は、意味を成さない
  - テクノロジーを最適に組み合わせることで、問題解決のため のソリューションの提供が可能

まとめとして

- '*Ts' for HPC - インテル・テクノロジのHPCにおける価値
  - インテル・テクノロジは、HPCにおいて、重要な構成要素となっている
  - それらの構成要素を統合することで、より高い価値の提供が可能となる
  - 二分化しつつあるHPCシステムにおいて、「標準コンポーネント」としてのプラットフォームの動向として、今後もその動向には注目する必要がある

さらに詳しい情報は.....

www.sstc.co.jp/biz

- 弊社のコンサルテーションに関するご提案資料もダウンロード可能です。(非公開WEBページ)別途、弊社に内容等については、お尋ねください。

お問い合わせ先:

〒102-0083
東京都千代田区麹町3-5-2
BUREX麹町 8F
電話:03-5875-4718
FAX:03-3237-7612
E-mail:biz@sstc.co.jp
http://www.sstc.co.jp

スケーラブルシステムズ株式会社

ハイエンドコンピューティングに関するコンサルテーションとして、幅広いサービスをご提供致します。

このサービスを最大限に活用していただくことで、コラボレーションによる「顧客志向」のコンサルテーションサービスをご提供できればと思っております。

2005年11月