HPCシステムの課題

スケーラブルシステムズ株式会社
代表取締役　戸室　隆彦

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” **ハードウエアとソフトウエア**

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）

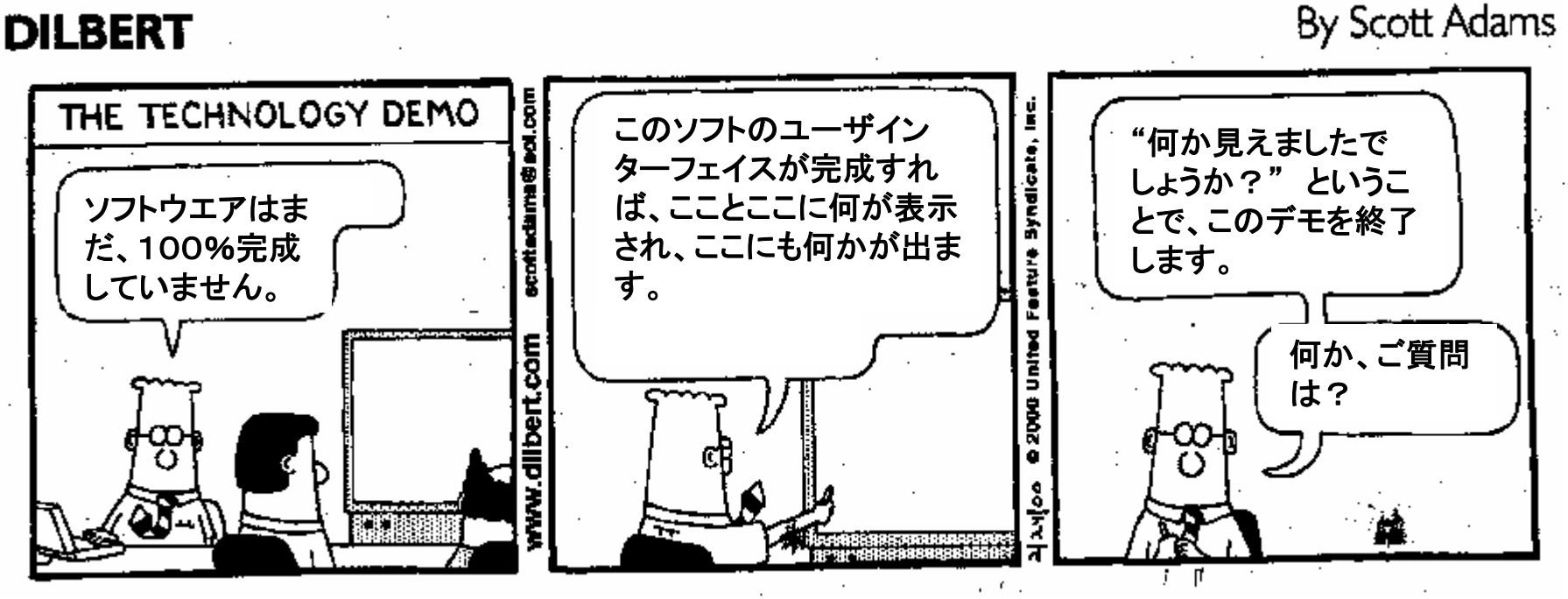
- 消費電力の問題のため、プロセッサの動作クロックを今までのペースで上げることは困難
- プロセッサとメモリの性能差の拡大によるCPUサイクルとのギャップ
- ピーク性能と実効性能のギャップの拡大

ソフトウエア：The Law of More…

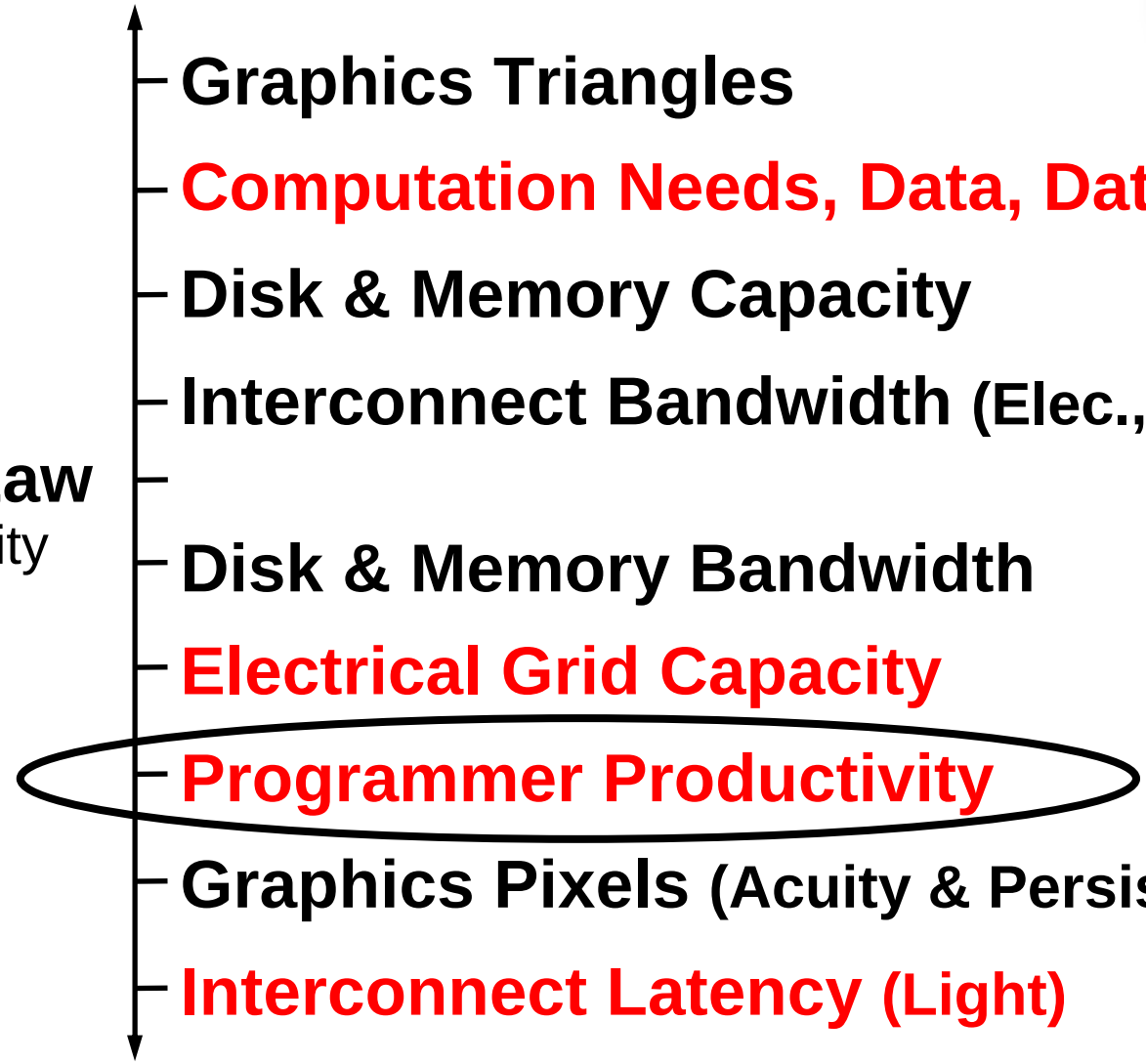
- 研究者は、より多くの時間（More Time)をソフトウエアの開発のために必要としている
- 問題はより複雑（More Complex)になり、そして、より多くのプロセッサ(More Processors)を利用して処理を行うには、より多くの困難(More Difficult)が伴います

ソフトウエアに関する問題については、この資料では、詳しくは解析をしていません。

テクノロジデモ


DILBERT
By Scott Adams
THE TECHNOLOGY DEMO
ソフトウエアはまだ、100%完成していません。
このソフトのユーザインターフェイスが完成すれば、ここととここに何が表示され、ここにも何かが出ます。
“何か見えましたでしょうか？” ということで、このデモを終了します。
何か、ご質問は？
www.dilbert.com scottadams@aol.com
© 2000 United Feature Syndicate, Inc.

コンピューティングに関するトレンド

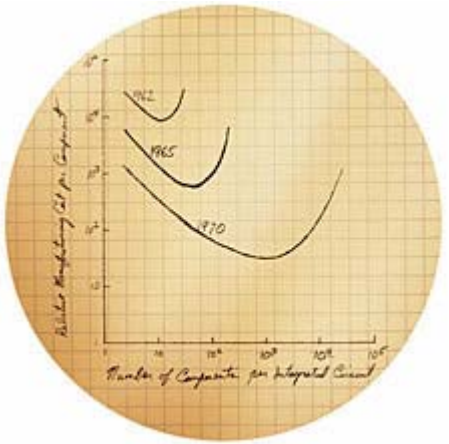


ソフトウエア：The Law of More…

- 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…..

Moore’s Law

Dr. Gordon Moore
(co-founder of Intel)

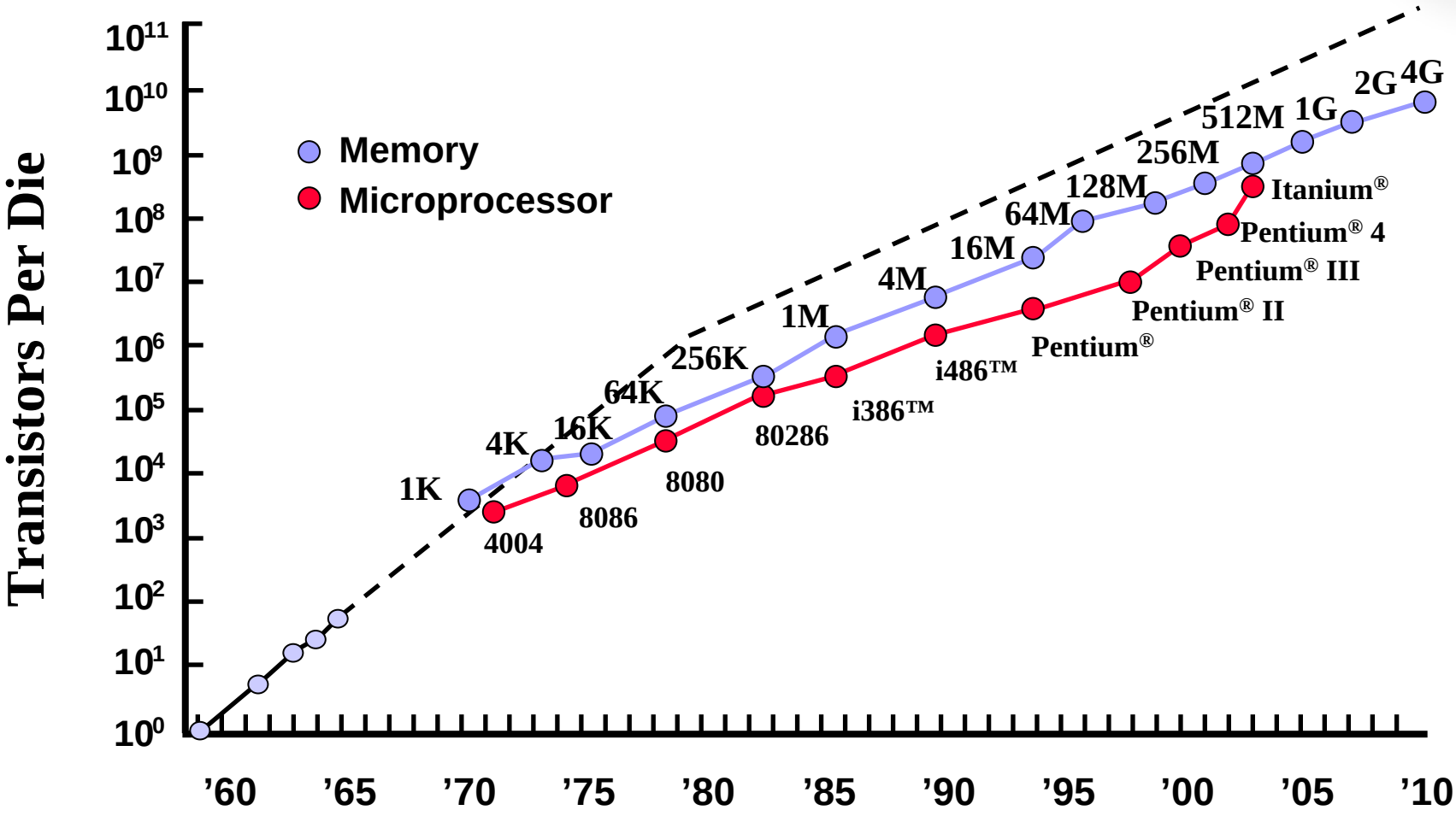


- インテルの共同設立者の1人である Gordon Moore 博士が、1965年4月19日号の「**Electronics**」誌に投稿した、「一定面積に集積されるトランジスタの数は12か月で倍増し、それに伴いトランジスタの動作速度が向上する」という予測 (その後、1975年に Moore 博士はチップの複雑化を考慮してトランジスタ数の倍増ペースを24か月に修正）
- また、一般にはあまり知られていないがテクノロジの進歩とともに製造コストが劇的に下落することも予測（左図）

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 [英語: PDF 形式 2MB]
Gordon E. Moore、2003年2月10日、ISSCC (International Solid State Circuits Conference) でのプレゼンテーション

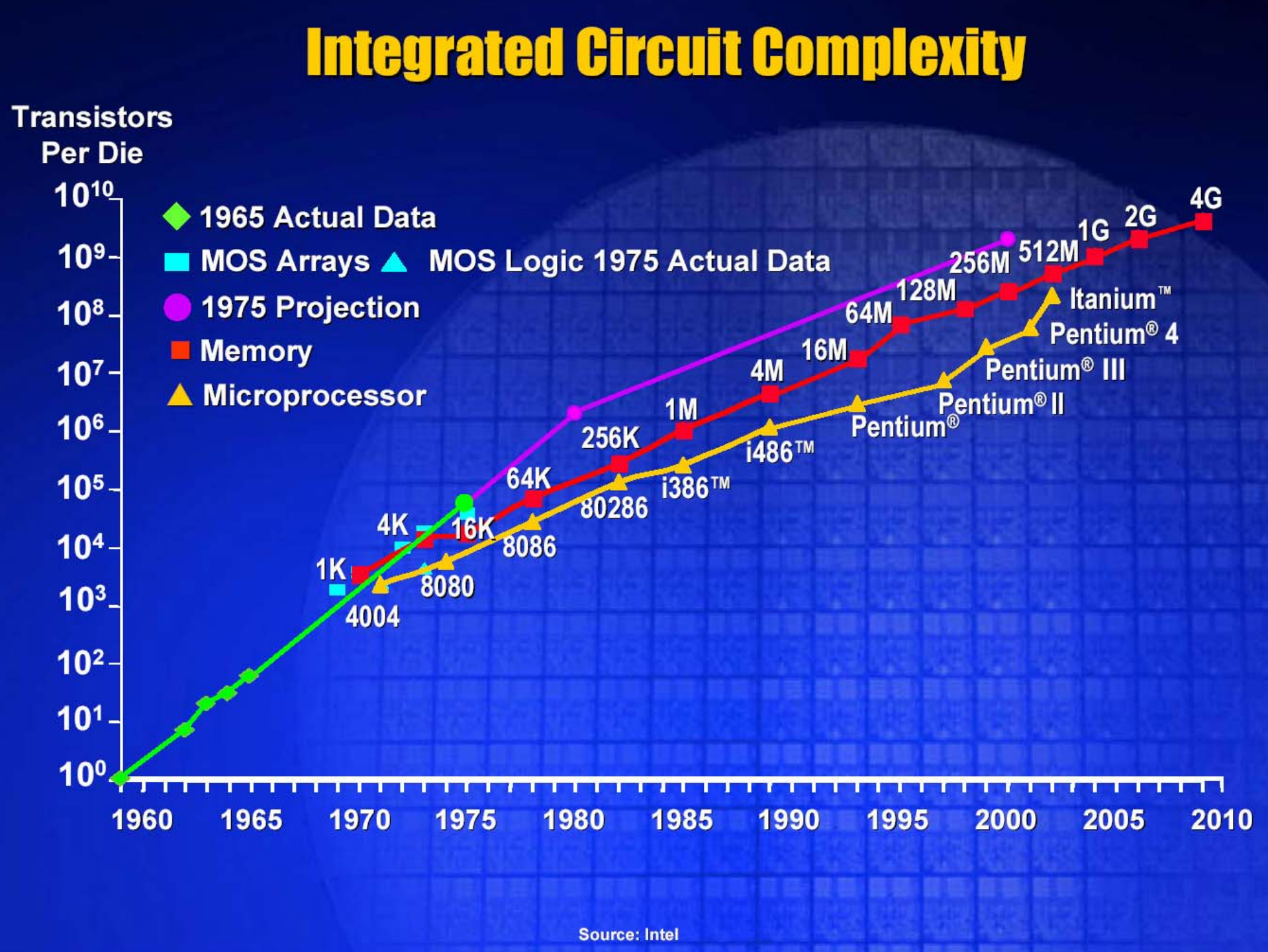
http://www.intel.co.jp/jp/developer/technology/silicon/mooreslaw/index.htm

ムーアの法則：依然として有効?



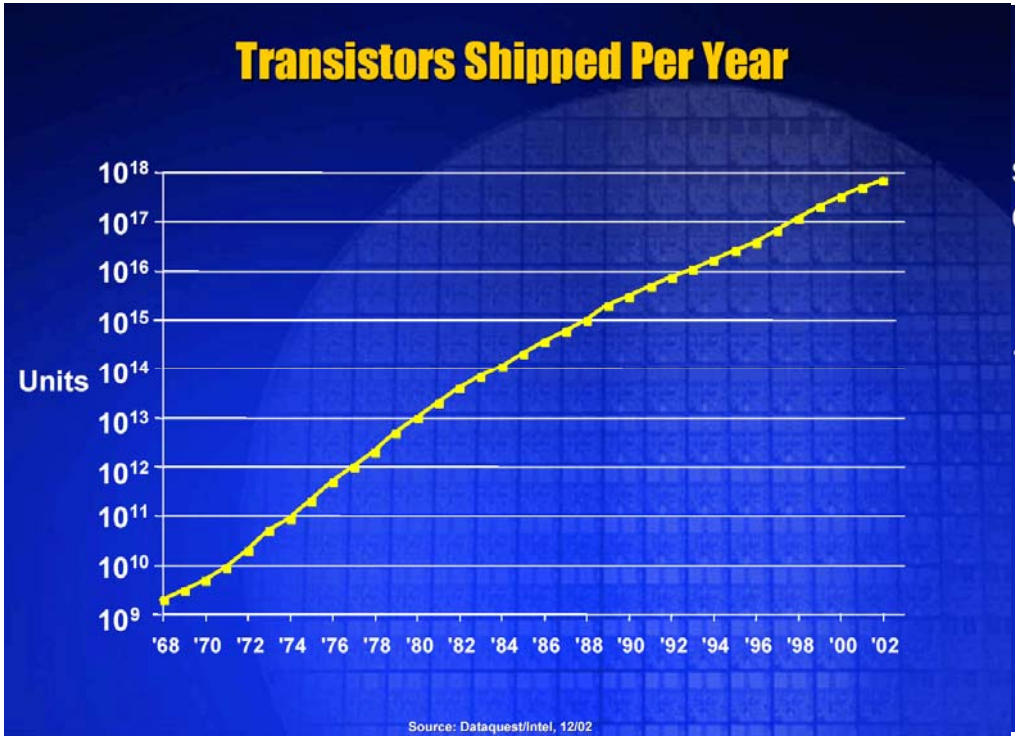
Source: Intel

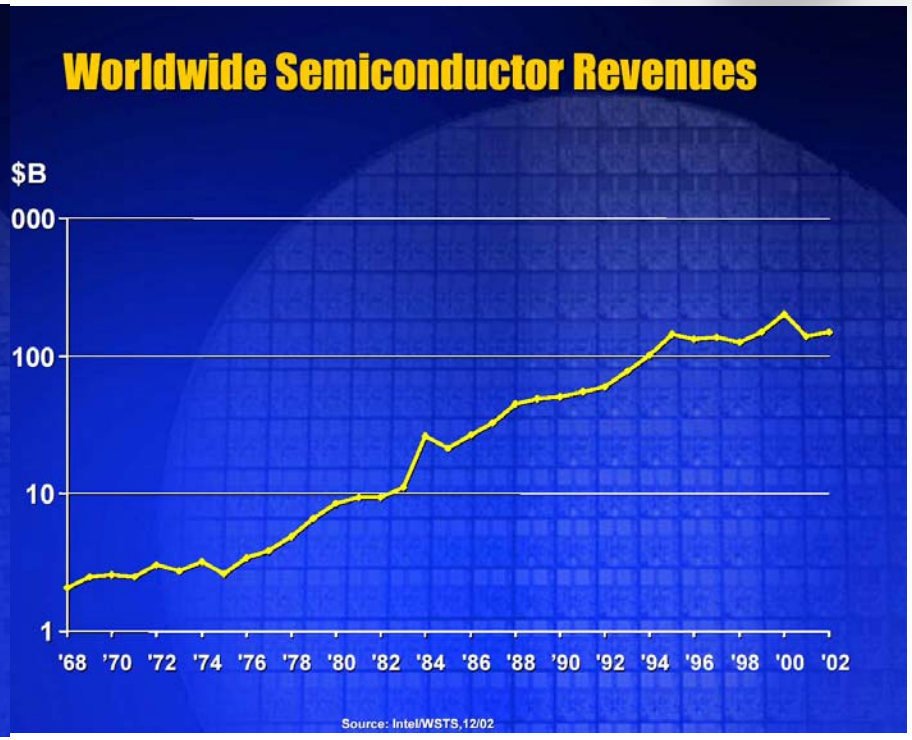
Integrated Circuit Complexity



Source: Intel

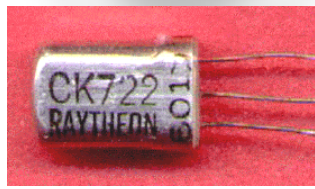
ビジネストレンド

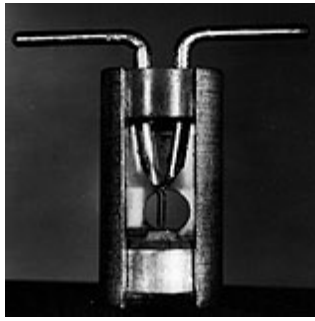

Transistors Shipped Per Year
Units
10^{18}
10^{17}
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
10^{9}
'68 '70 '72 '74 '76 '78 '80 '82 '84 '86 '88 '90 '92 '94 '96 '98 '00 '02
Source: Dataquest/Intel, 12/02

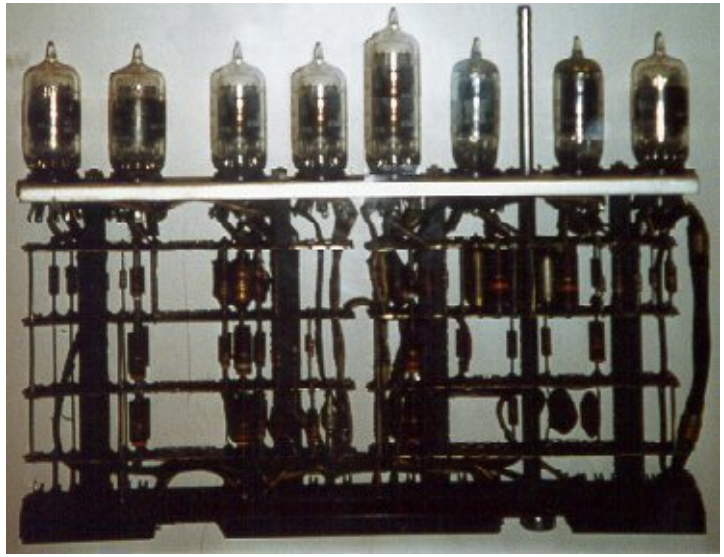

Worldwide Semiconductor Revenues
$B
000
100
10
1
'68 '70 '72 '74 '76 '78 '80 '82 '84 '86 '88 '90 '92 '94 '96 '98 '00 '02
Source: Intel/WSTS,12/02

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
    - プロセッサあたりの1サイクルでの命令実行
      - 命令レベルでの並列性（ILP)
      - ベクトル処理
    - プロセッサあたりのプロセッサコア数
    - 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





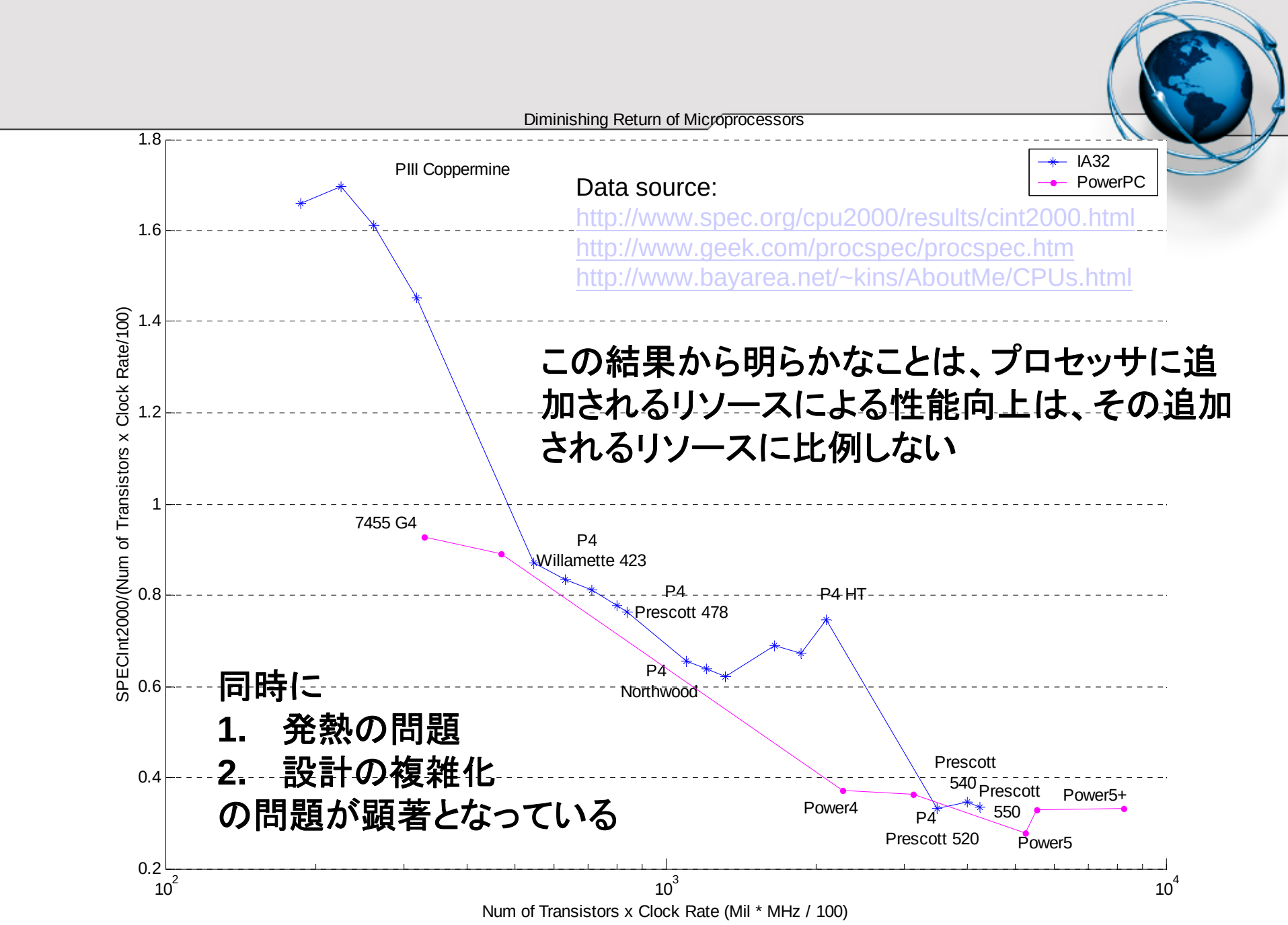


GHz競争

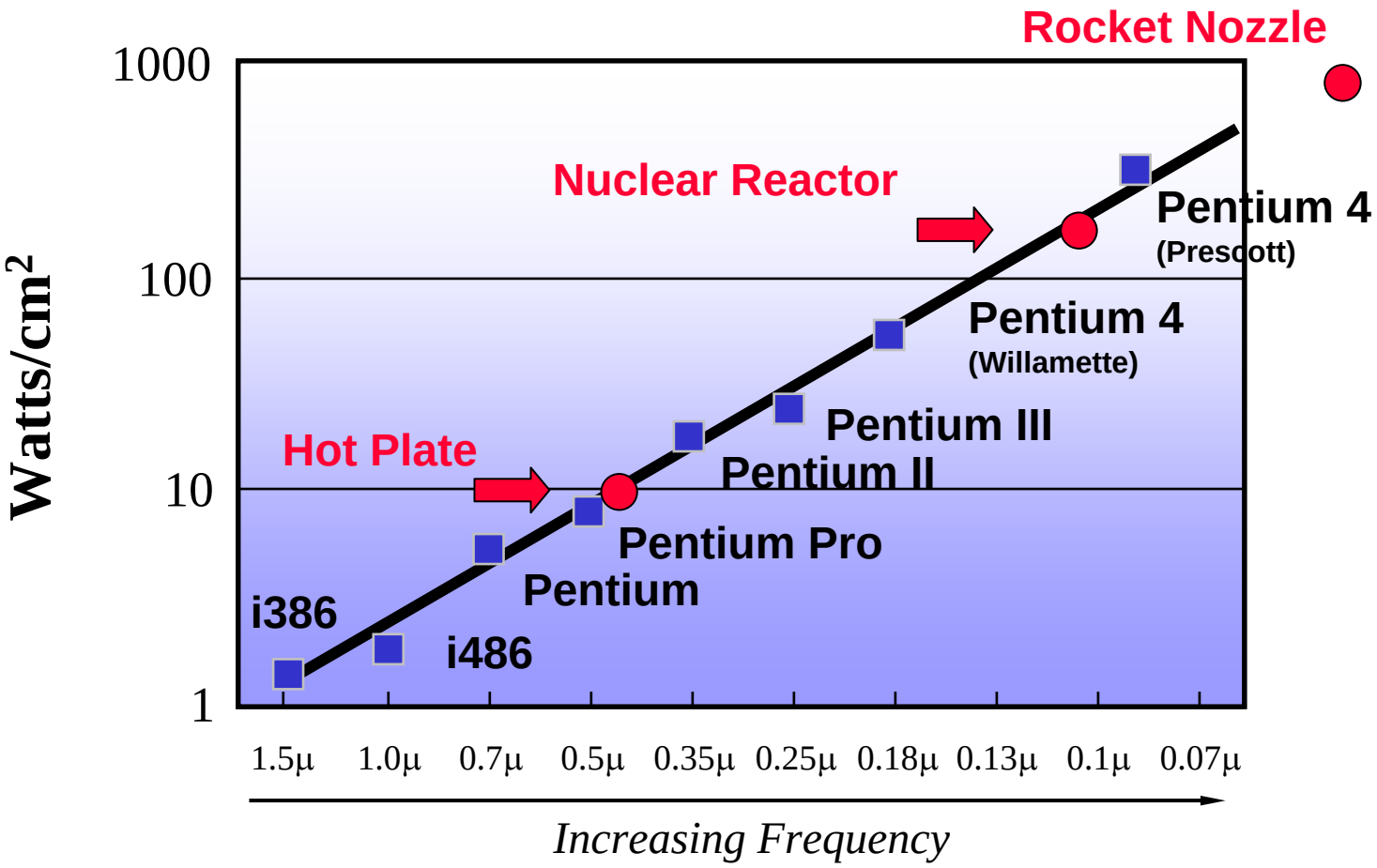
- 2000年に開催されたIEEE国際電子デバイス会議2000（2000 IEEE International Electron Devices Meeting：IEDM）において、インテル社は4億個以上のトランジスタを集積した、10GHz駆動のプロセッサが2005年までに実現可能だと発表しました。
  - 実際には、インテル社の最速プロセッサは、6ヶ月前に発表された3.8GHz（Intel Pentium 4)となっています。
- Prescottプロセッサの6xxシリーズ発表に際して、インテル社は、“adding value beyond GHz” のコメントを出しています。それ以降、イン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

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

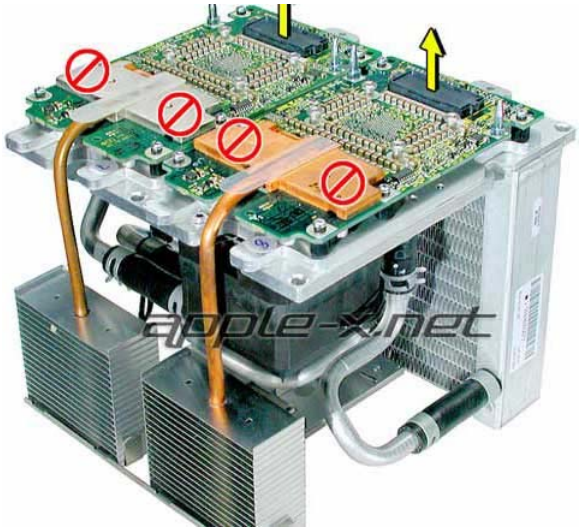

Diminishing Return of Microprocessors
SPECInt2000/(Num of Transistors x Clock Rate/100)
Num of Transistors x Clock Rate (Mil * MHz / 100)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
10^2
10^3
10^4
IA32
PowerPC
PIII Coppermine
Data source:
http://www.spec.org/cpu2000/results/cint2000.html
http://www.geek.com/procspec/procspec.htm
http://www.bayarea.net/~kins/AboutMe/CPUs.html
この結果から明らかなことは、プロセッサに追加されるリソースによる性能向上は、その追加されるリソースに比例しない
7455 G4
P4
Willamette 423
P4
Prescott 478
P4 HT
P4
Northwood
同時に
1. 発熱の問題
2. 設計の複雑化
の問題が顕著となっている
Prescott
540
Prescott
550
Power5+
Power4
P4
Prescott 520
Power5

発熱の問題が深刻化



Bob Colwell氏の資料より抜粋

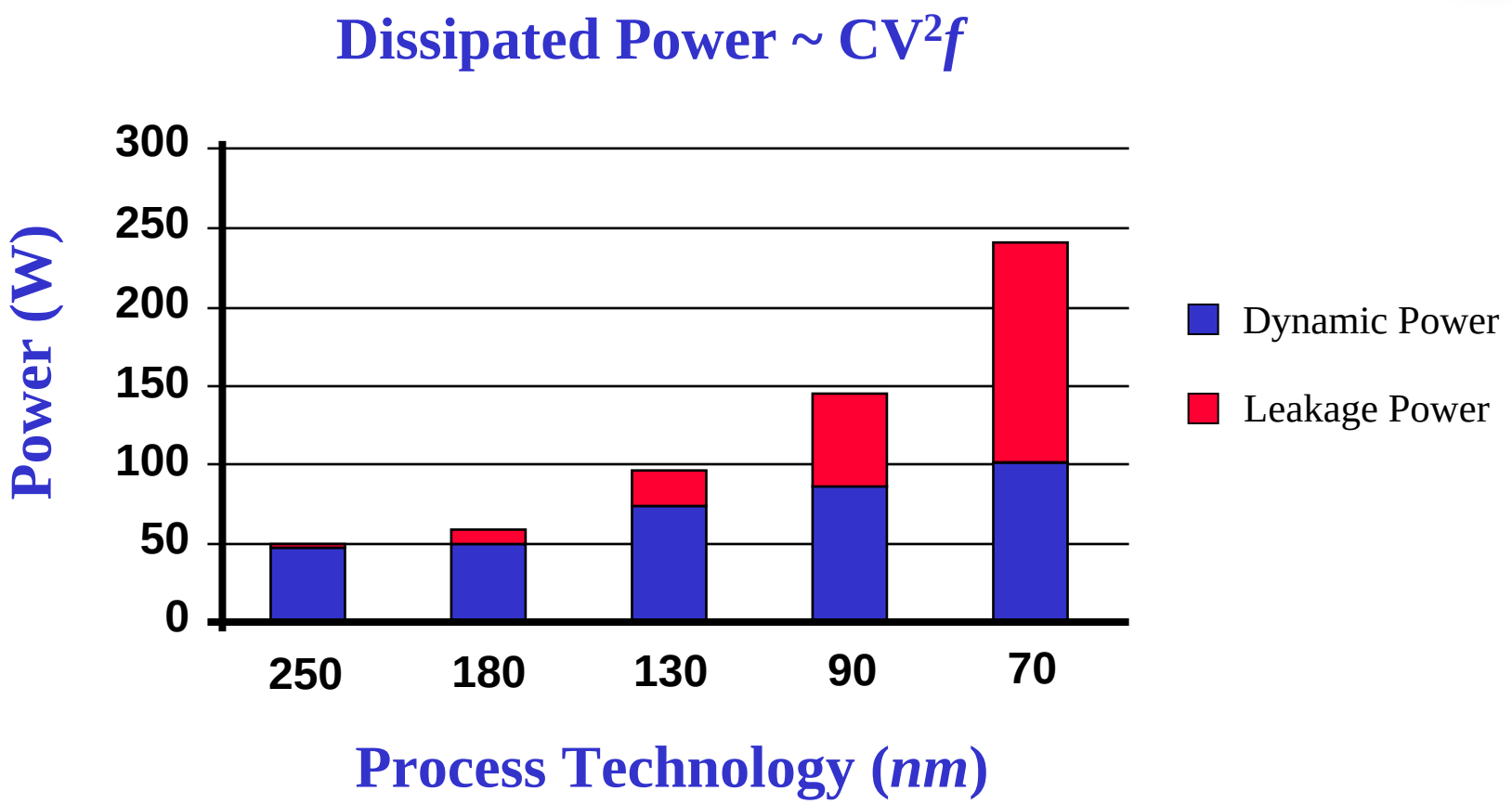
冷却方法の検討



Apple G5での水冷方式

Pentium 4 の6XXシステム
でのヒートシンク

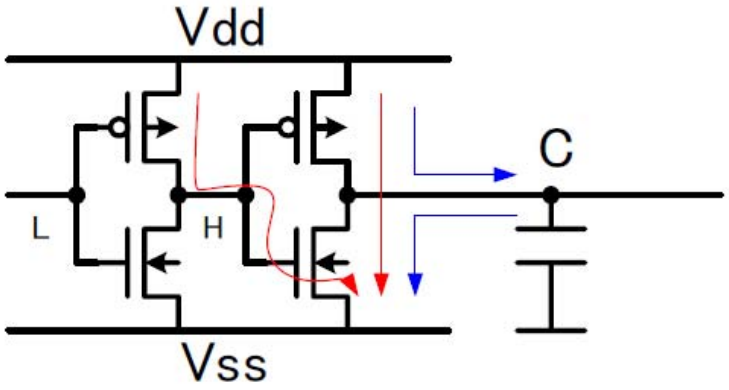
消費電力におけるリーク電流の影響



消費電力におけるリーク電流の影響

CMOSデバイスの特性と回路特性との関係

CMOSゲートの消費電力の要因



$P = C V^2 f + I_L V$

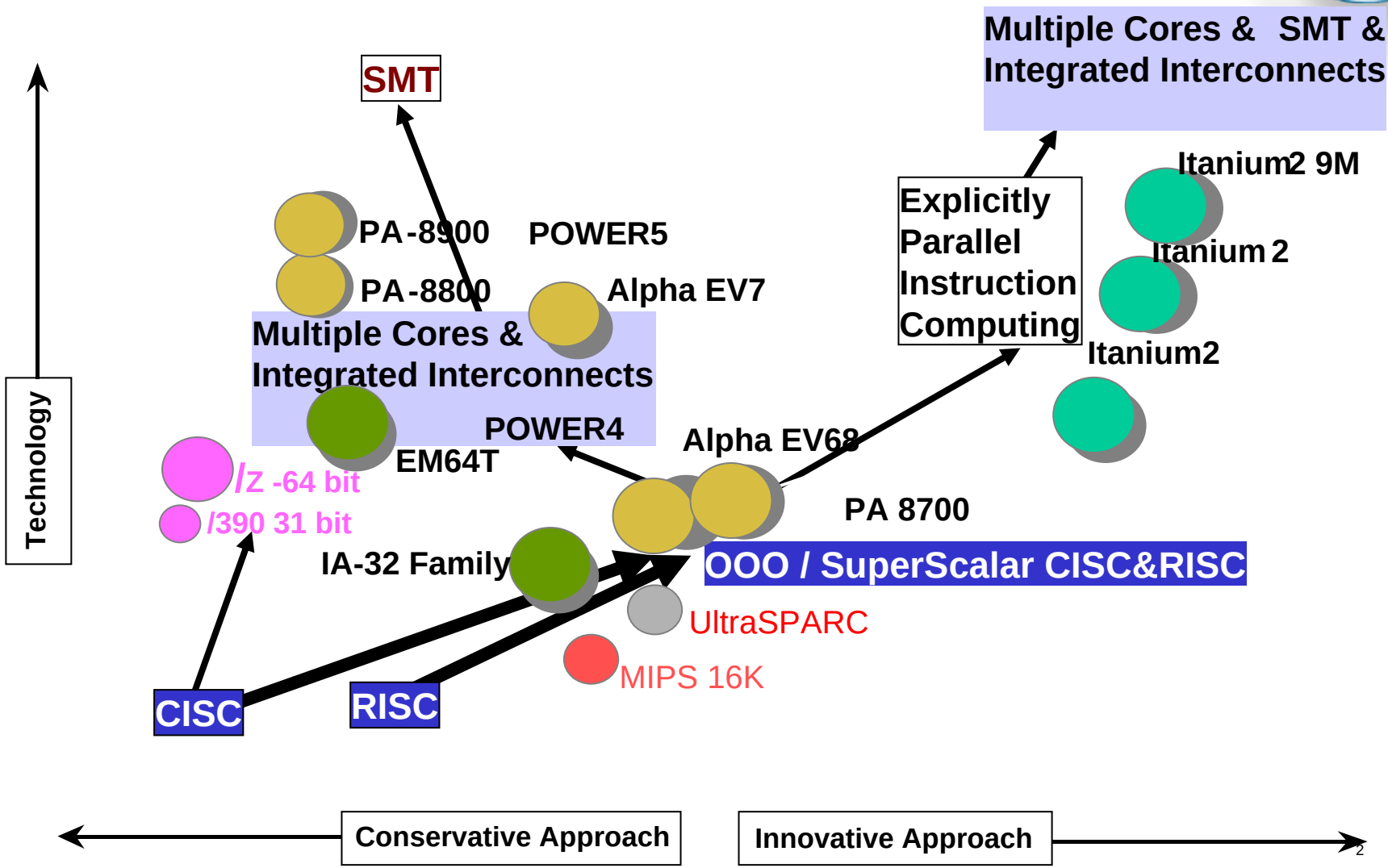
IL：リーク電流

AC成分　　DC成分

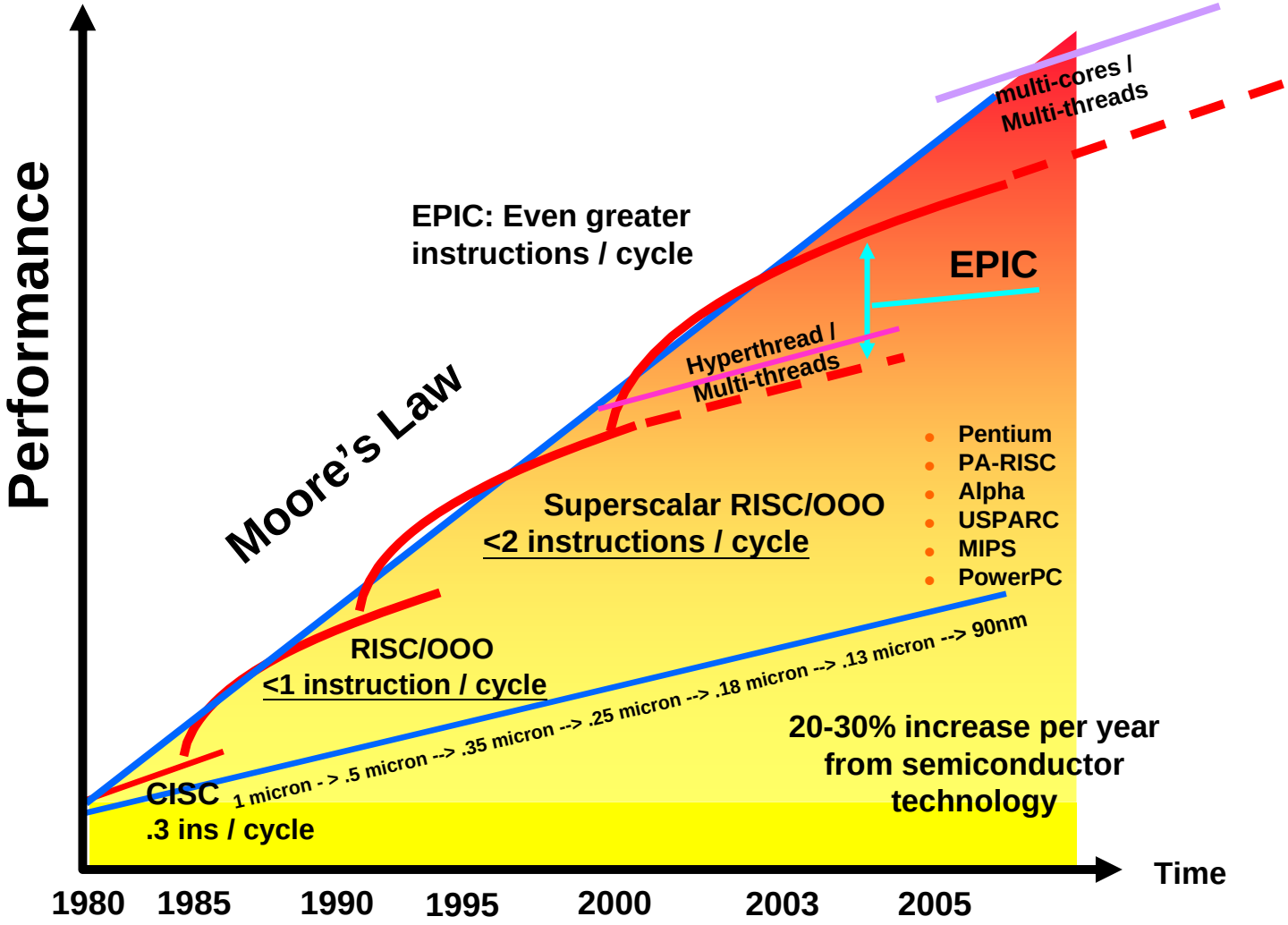
計算機の性能向上

- 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)

プロセッサテクノロジー一覧



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

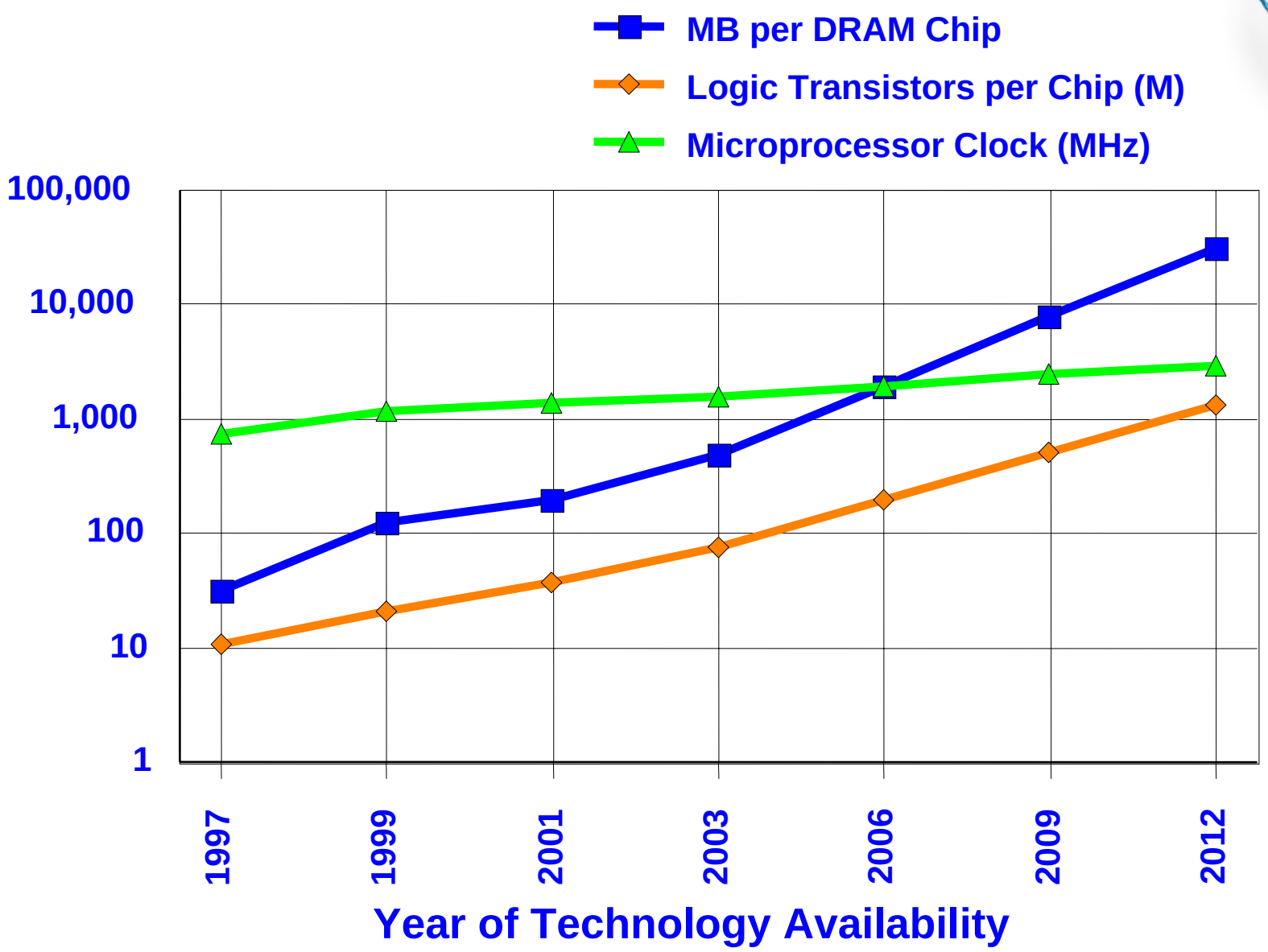


Source : Intel

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

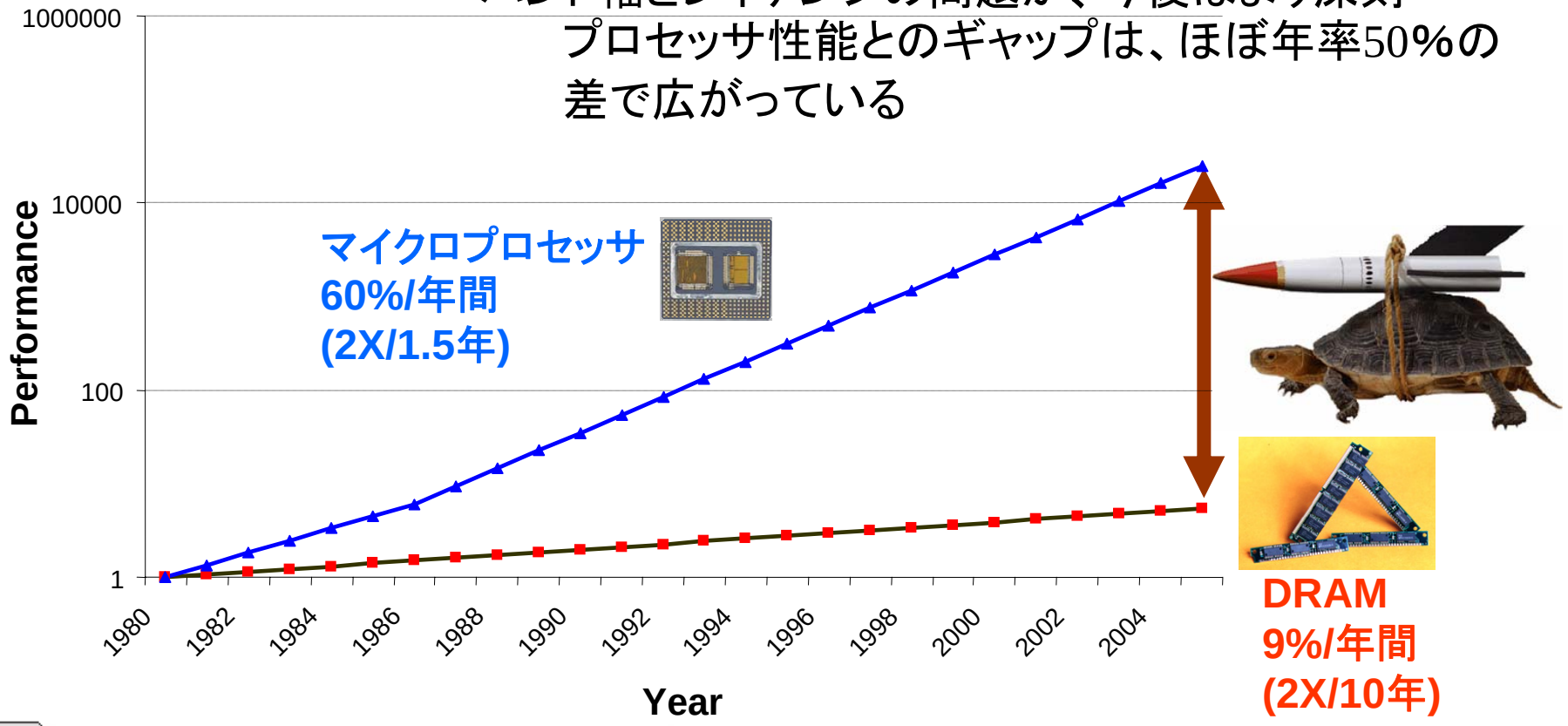
半導体の技術動向予測


MB per DRAM Chip
Logic Transistors per Chip (M)
Microprocessor Clock (MHz)
100,000
10,000
1,000
100
10
1
1997
1999
2001
2003
2006
2009
2012
Year of Technology Availability

性能ギャップの問題

- プロセッサは、ほぼナノ秒に一回、命令実行を行っている
- DRAMへのアクセスには、ほぼ100ナノ秒の時間がかかる
- 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

バンド幅とレイテンシの問題が、今後はより深刻
プロセッサ性能とのギャップは、ほぼ年率50%の差で広がっている



メモリ階層

階層	プロセッサ・クロック
Register	1
L1 Cache	2－3
L2 Cache	6－12
L3 Cache	14－40
Near Memory	100-300
Far Memory	300-900
Remote Memory	$O(10^3)$
Message-Passing	$O(10^4)$

近年のプロセッサは、クロック周波数の向上、命令レベル並列性の活用などにより高性能化が図られていますが、DRAMを利用した主記憶へのアクセス性能はプロセッサほどは改善していません。このような状況のもと、近年のプロセッサの性能はメモリの性能により制限されていると言えます。これが、プロセッサのピーク性能とアプリケーションの実効性能に大きなギャップが生じる理由の一つです。HPCにおいては、下位のメモリ階層へのアクセスの頻発により性能が大きく低下することを如何に低減するかが課題です。

この問題に対処するために、従来からキャッシュメモリが用いられてきましたが、大規模なシュミレーションでは、キャッシュが有効に機能しないことが多い。キャッシュ容量に比べデータセットが非常に大きく、また、データの局所性が限定されるような場合には、キャッシュミスによる性能劣化が顕著になります。

集積度の向上を頼りにプロセッサチップ上へのキャッシュ実装も限界があり、このメモリ階層での性能劣化への対応が重要になっています。

64ビットアーキテクチャ

	64-bit Intel® Xeon™ Processor	64-bit Intel® Itanium2™ Processor
64-bit flat virtual address space	○	○
Physical address space	36 bits (DP)/40 bits (MP)	50 bits
Address space supported in platforms	最大 1TB	最大 1PB
64-bit pointers	○	○
64-bit wide general purpose registers	○	○
64-bit integer support	○	○
64-bit OS support	○	○
64-bit applications	○	○

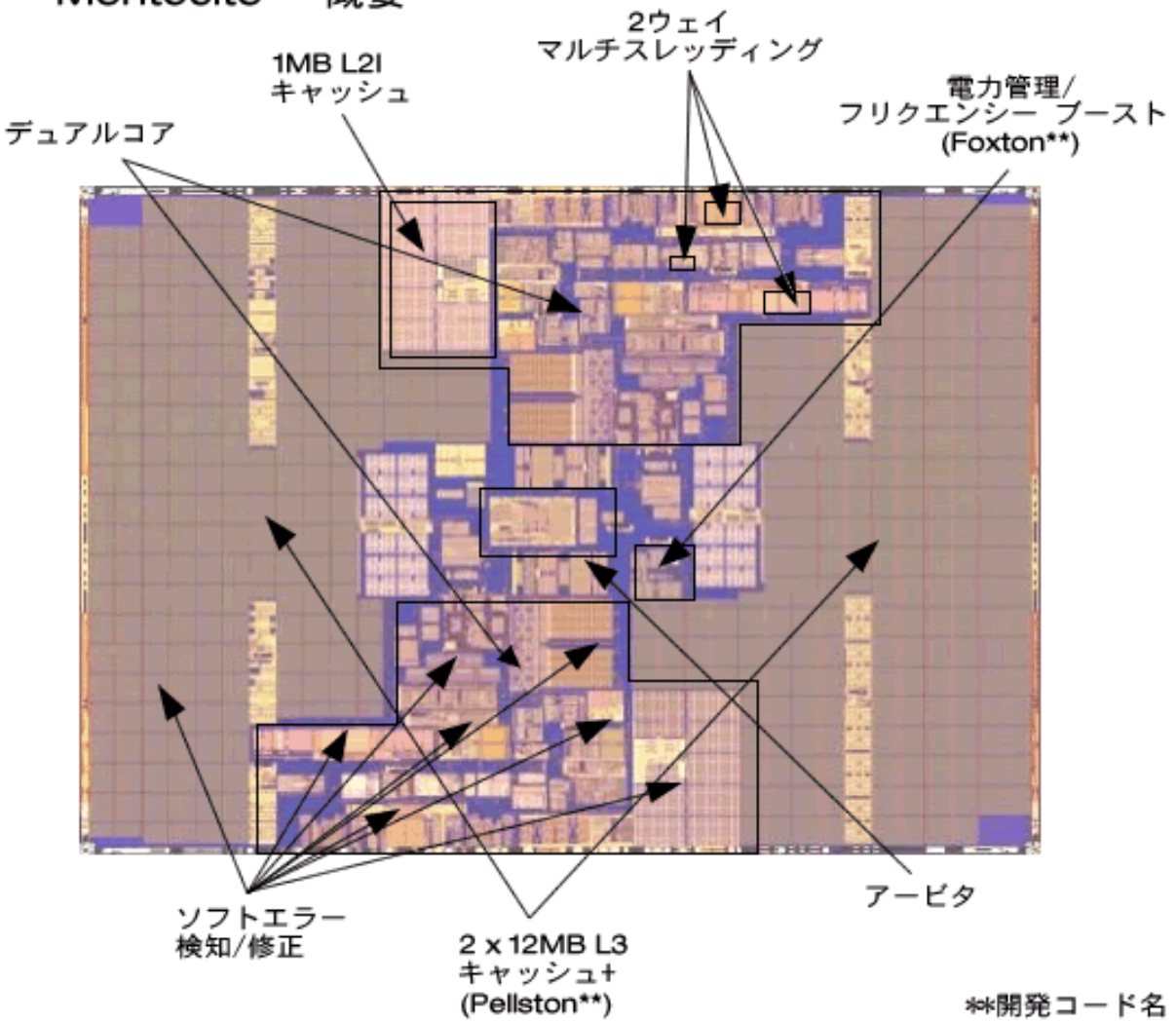
Itanium vs. Opteron (アドレス変換)

	Opteron	Itanium
TLB entries (instructions)	512	128
TLB-I associativity	4	128 (!)
TLB entries (data)	512	128
TLB-D associativity	4	128 (!)
supported page sizes	4 KB, 2 MB	4 KB … 4 GB
resulting address range with no TLB miss/fault	1 GB	512 GB

X86-64では、メモリページのサイズの制限があり、1GB
を超えるメモリでは、TLBミスの発生の可能性が大きくなる

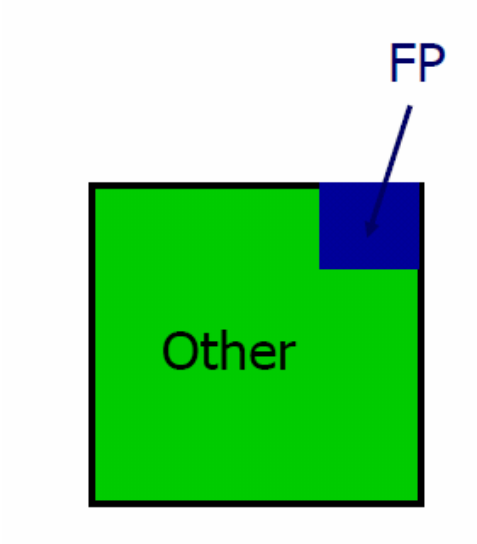
大容量キャッシュ：Montecito

次世代インテル® Itanium® プロセッサ・ファミリ製品
Montecito** 概要



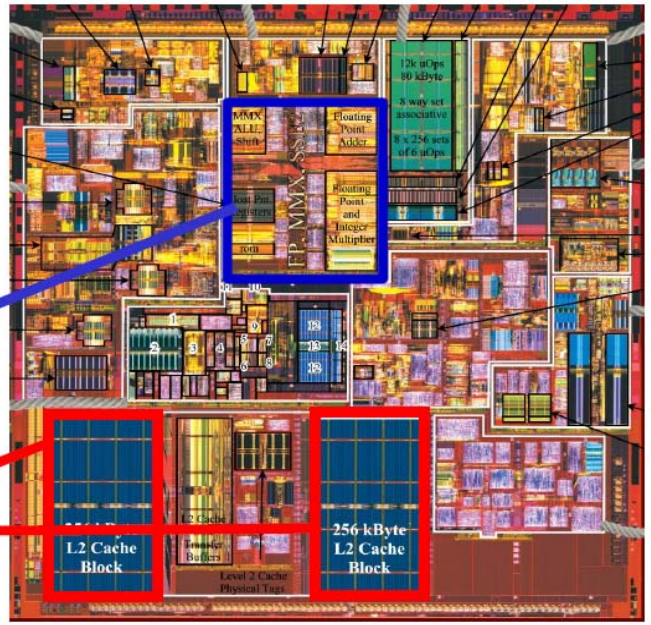
**開発コード名

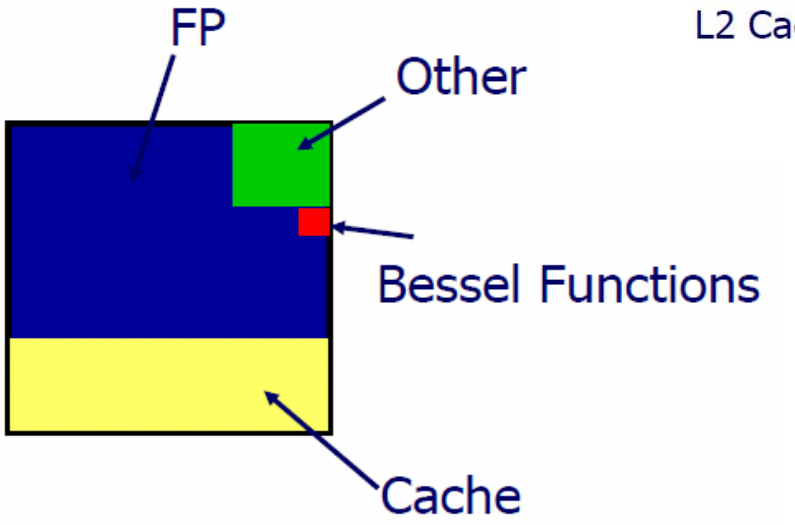
大容量キャッシュ:HPC？



- 90 nm CMOS
- Clock 3.4 GHz

Pentium Prescott





計算機の性能向上

- 動作周波数(ク
  - 過去12年 Pentiumプロセッサ 作周波数は、60 MHz か 300 MHz にまでアッ
  - 現在ま の高性能化の約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 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 、 行順序の変更など)
- 大容量キャ
  - プロセッサの (モリレイテ 待ち時間)とバンド幅 のギャップの拡大 対応としての容量の拡張

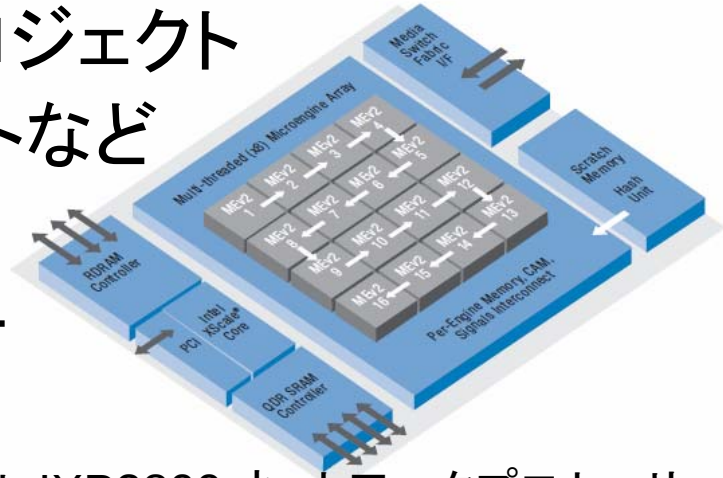
新しいプロセッサ技術の導入

- 省電力プロセッサ
  - 今まで以上のアプリケーションのスケーラビリティ
    - ~100,000プロセッサでのスケーラビリティ(ピーク)
    - ~1,000プロセッサ(通常運用での利用?)
  - プロセッサ障害でのリカバリ(耐障害性やチェックポイント)
- マルチコアプロセッサ

> “Supercomputer efforts can glean more performance from parallel computing and multi-core chips that will help fuel the HPC fight”
> Top500 list co-founder and co-editor Erich Strohmaier

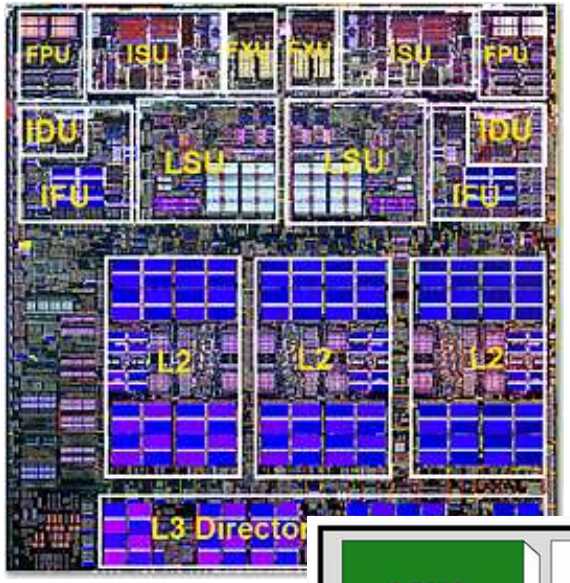
チップ上にマルチコアを実装

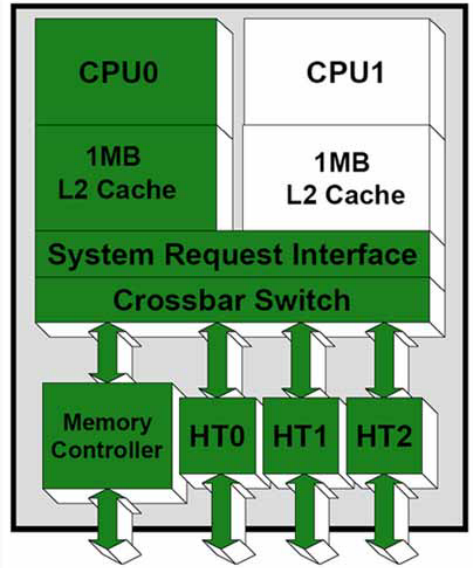
- このような考えは、特に新しいものではない！
- プロセッサの設計・製造の限界の克服のため、このコンセプトの実現が求められている
- 事例:
  - インテルのマルチコアプロジェクト(ディスクトップ、サーバ、モバイルの全ての用途に対して)
  - 組み込みシステム用途のプロジェクト
  - 米国政府のHPCSプロジェクトなど
  - 大学の研究プロジェクト
  - IBMやAMDの商用プロセッサ



インテル IXP2800 ネットワークプロセッサ

デュアルコアプロセッサアーキテクチャ

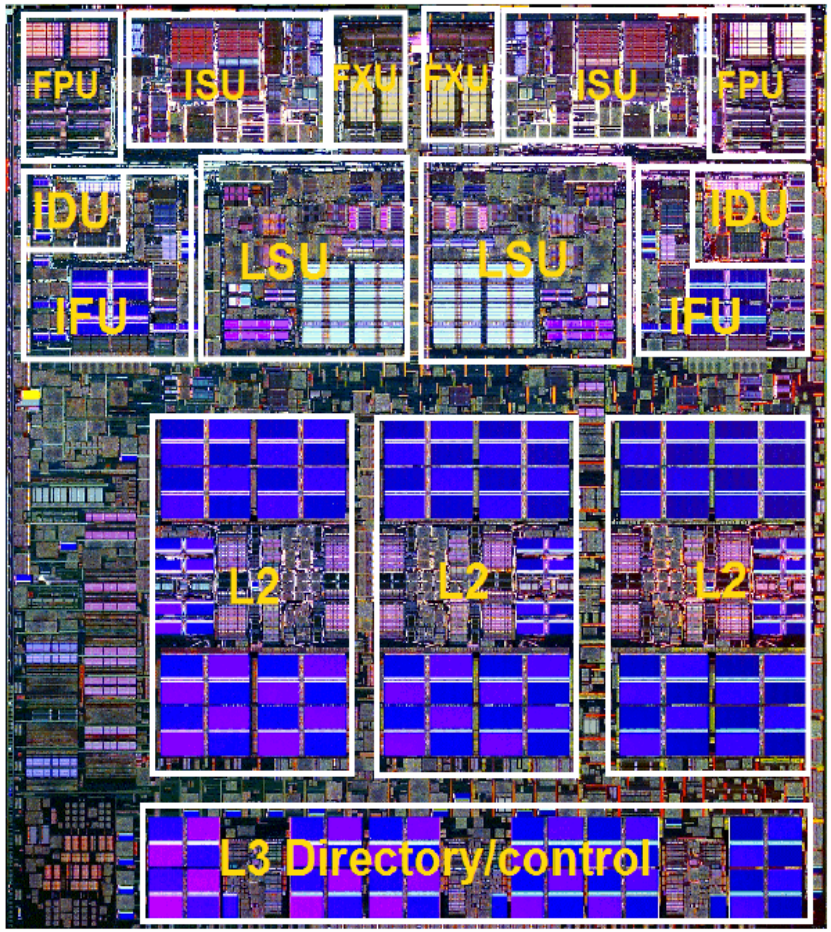




- IBM Power5
  - Shared 1.92 Mbyte L2 cache
- AMD Opteron
  - Separate 1 Mbyte L2 caches
  - CPU0 and CPU1 communicate through the SRQ
- Intel Pentium 4
  - “Glued” two processors together

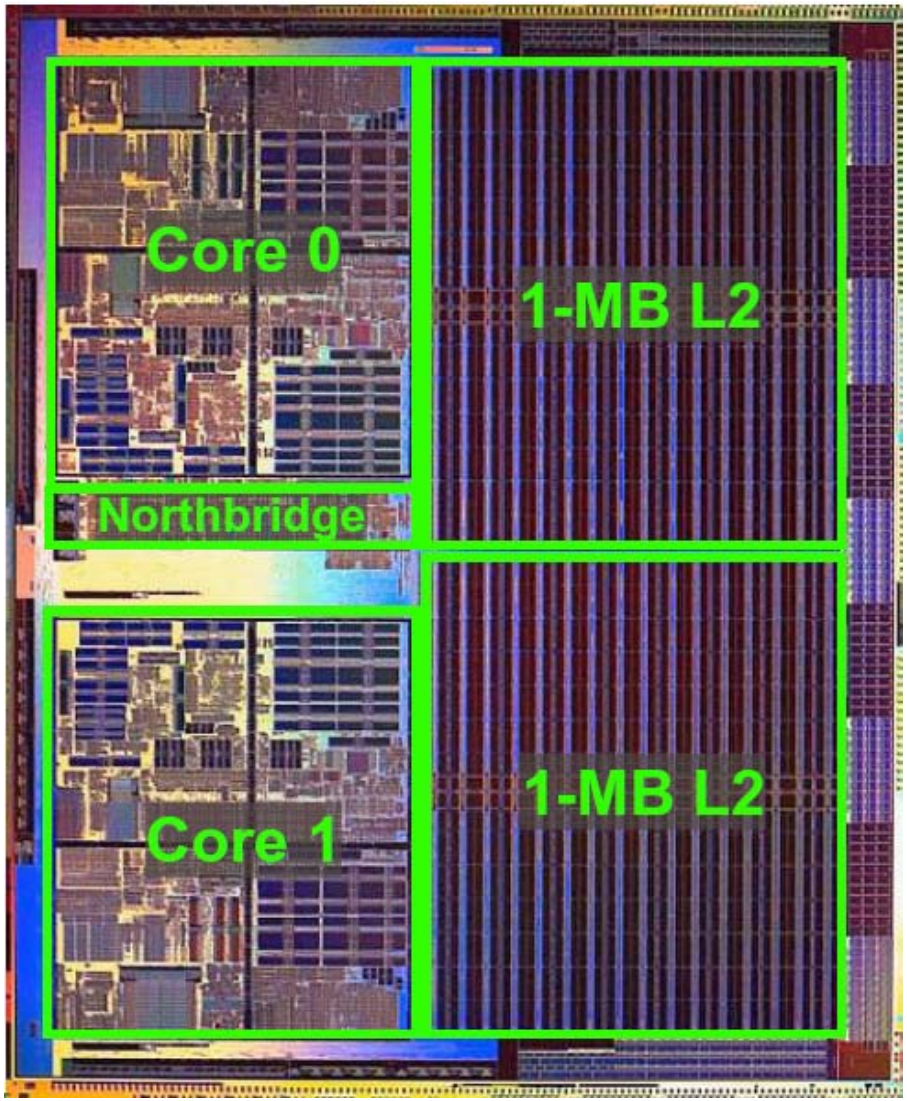
IBM Power5 Multicore Chip

- Technology: 130nm lithography, Cu, SOI
- Dual processor core
- 8-way superscalar
- Simultaneous multithreaded (SMT) core
  - Up to 2 virtual processors per real processor
  - 24% area growth per core for SMT
  - Natural extension to POWER4 design



Courtesy *of "Simultaneous Multi-threading Implementation in POWER5 --IBM's Next Generation POWER Microprocessor" by Ron Kalla, Balaram Sinharoy, and Joel Tendler of IBM Systems Group*

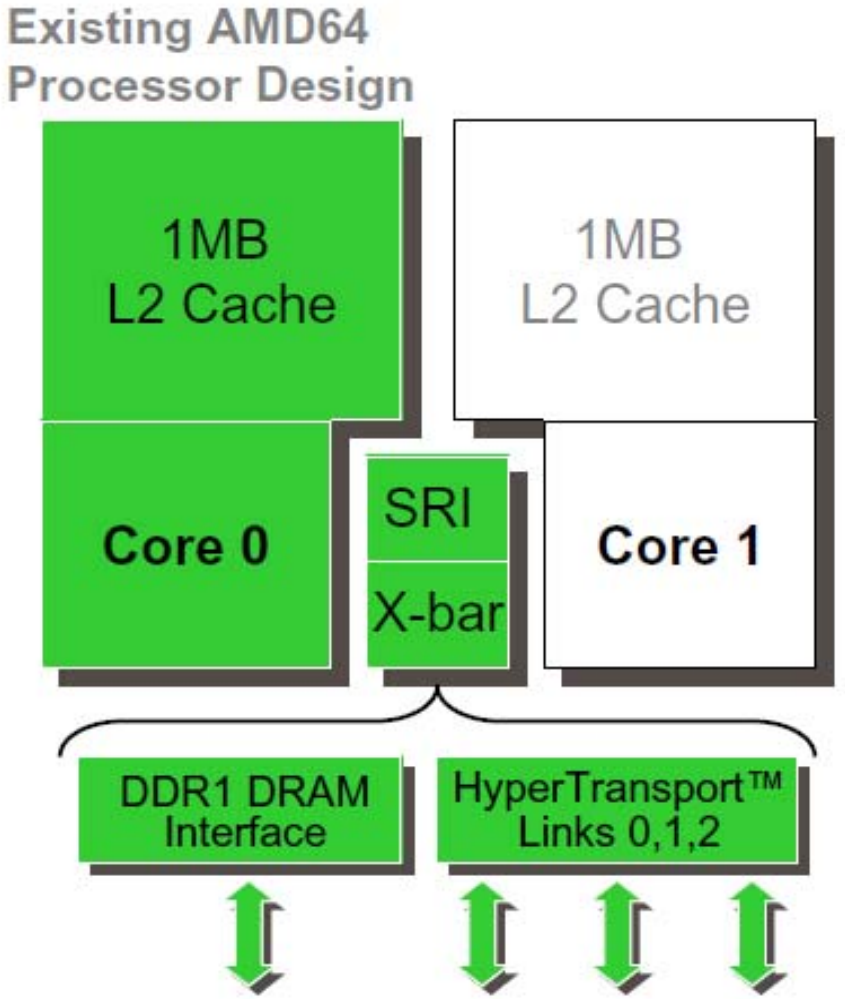
AMD Opteron™ CPU



- Two AMD Opteron™ CPU cores on a single die, each with 1MB L2 cache
- 90nm, ~205 million transistors*
  - Approximately same die size as 130nm single-core AMD Opteron processor*
- 95 watt power envelope fits into 90nm power infrastructure
- Retains compatibility with existing 32-bit and 64-bit x86-base software
- Introduced with “K8” Revision E core in April 2005

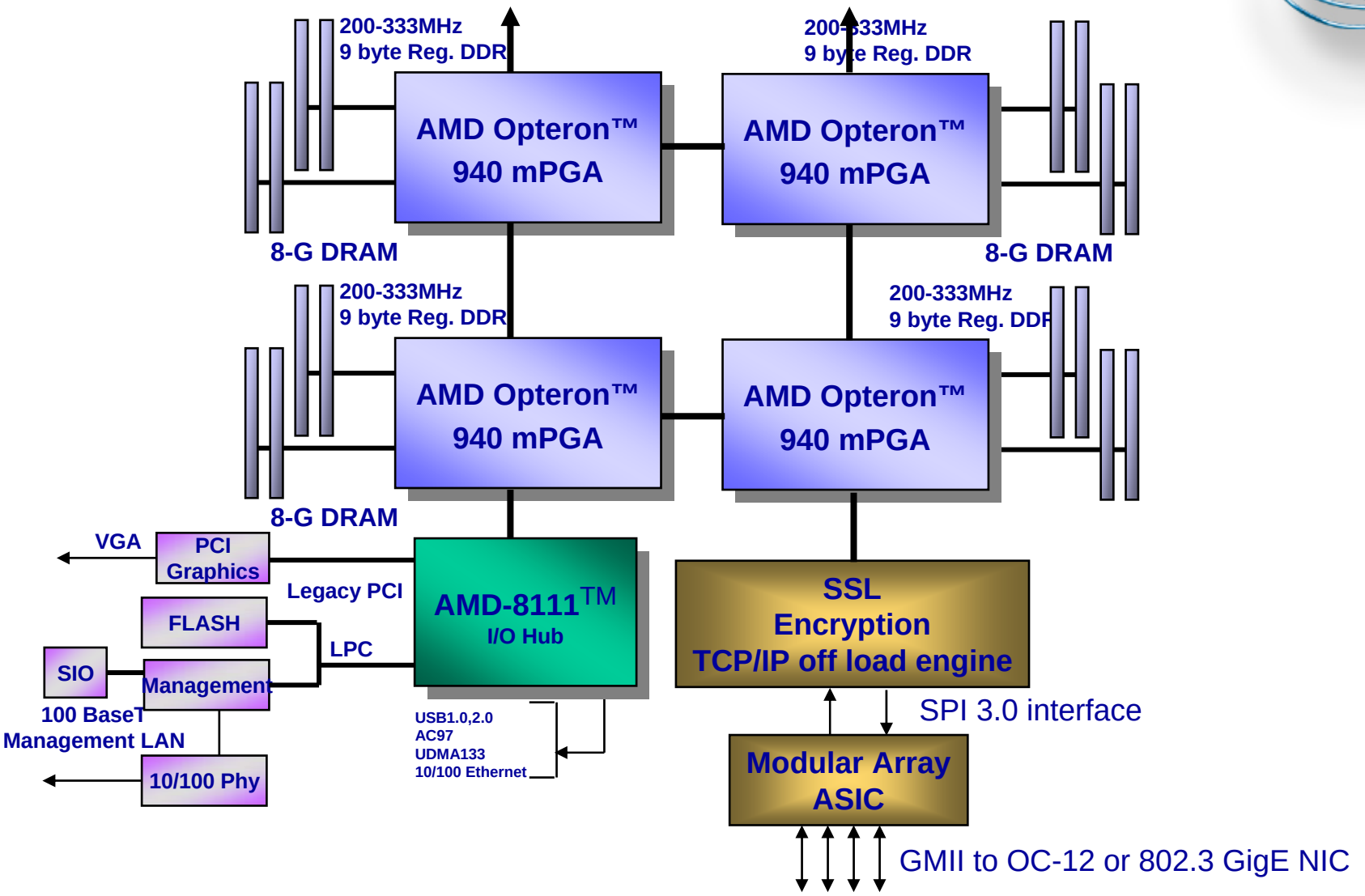
AMD Opteron™ CPU

- Shared Northbridge
  - 3 HyperTransport™ technology links
  - Dual-channel (128 bit) DDR i/f
- AMD Opteron™ CPU with Direct Connect Architecture was designed as CMP from the start
  - Second port on SRI, request management, two APICs
- Two complete CPU cores
  - SMP model
  - Simpler, less-restrictive programming model than “logical core” approach
  - No need to “pause” one core to give other exclusive use of shared resources



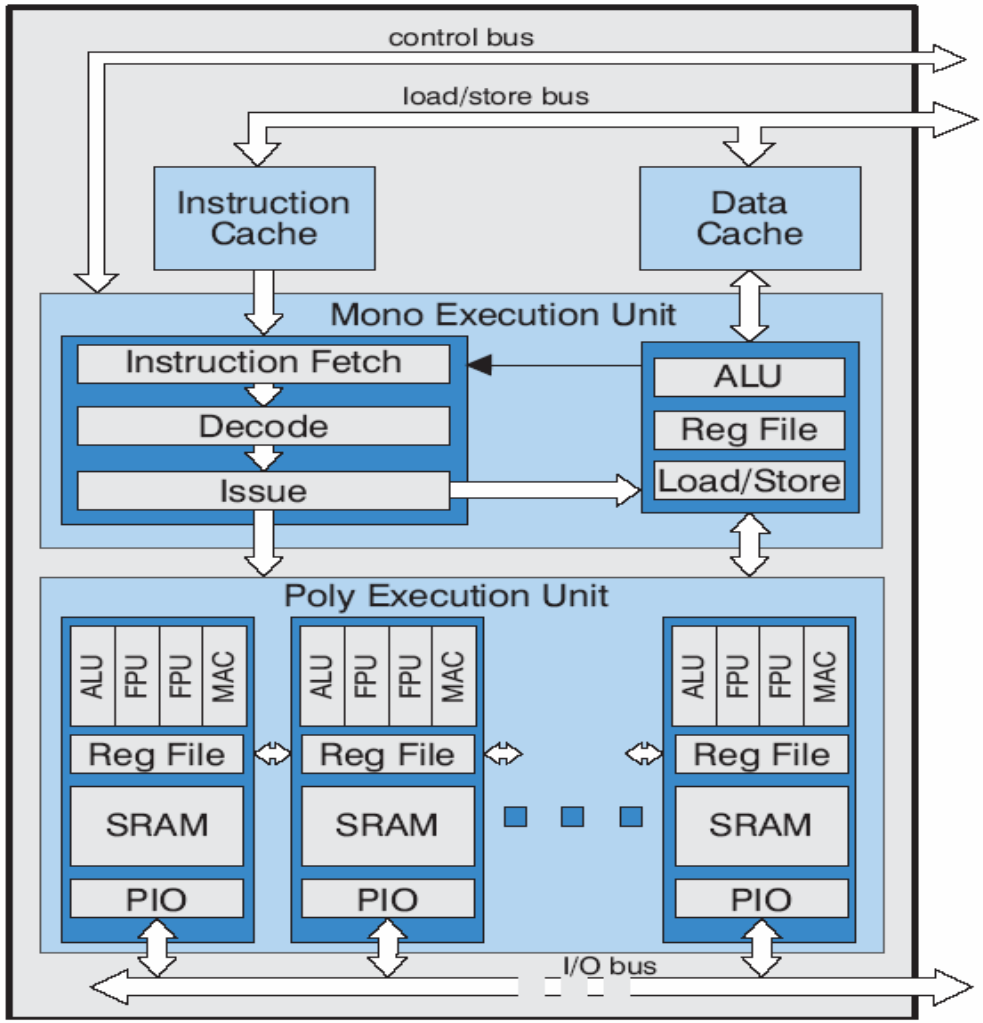
AMD Presentation For Linux Kernel Summit
Richard A. Brunner
AMD Fellow のプレゼンテーションからの抜粋

Quad AMD Opteron™



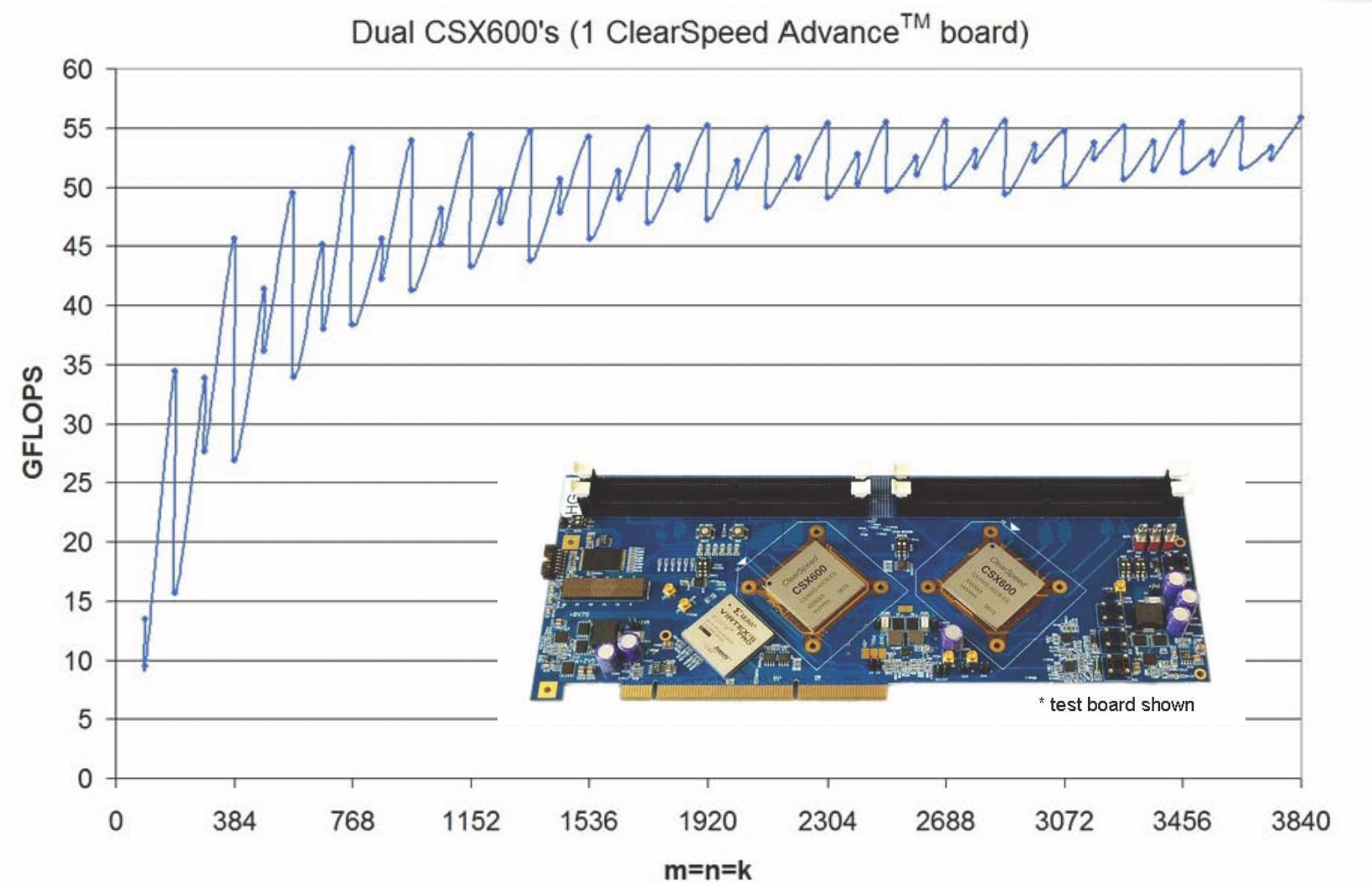
ClearSpeed CSX600

- 250 MHz clock
- **96** high-performance processing elements
- 576 Kbytes PE memory
- 128 Kbytes on-chip scratchpad memory
- 25,000 MIPS
- 50 GFLOPS single or double precision
- 3.2 Gbytes/s external memory bandwidth
- 96 Gbytes/s internal memory bandwidth
- 2 x 4 Gbytes/s chip-to-chip bandwidth



Courtesy of CSX600 Overview on http://www.clearspeed.com/

ClearSpeed CSX600



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.

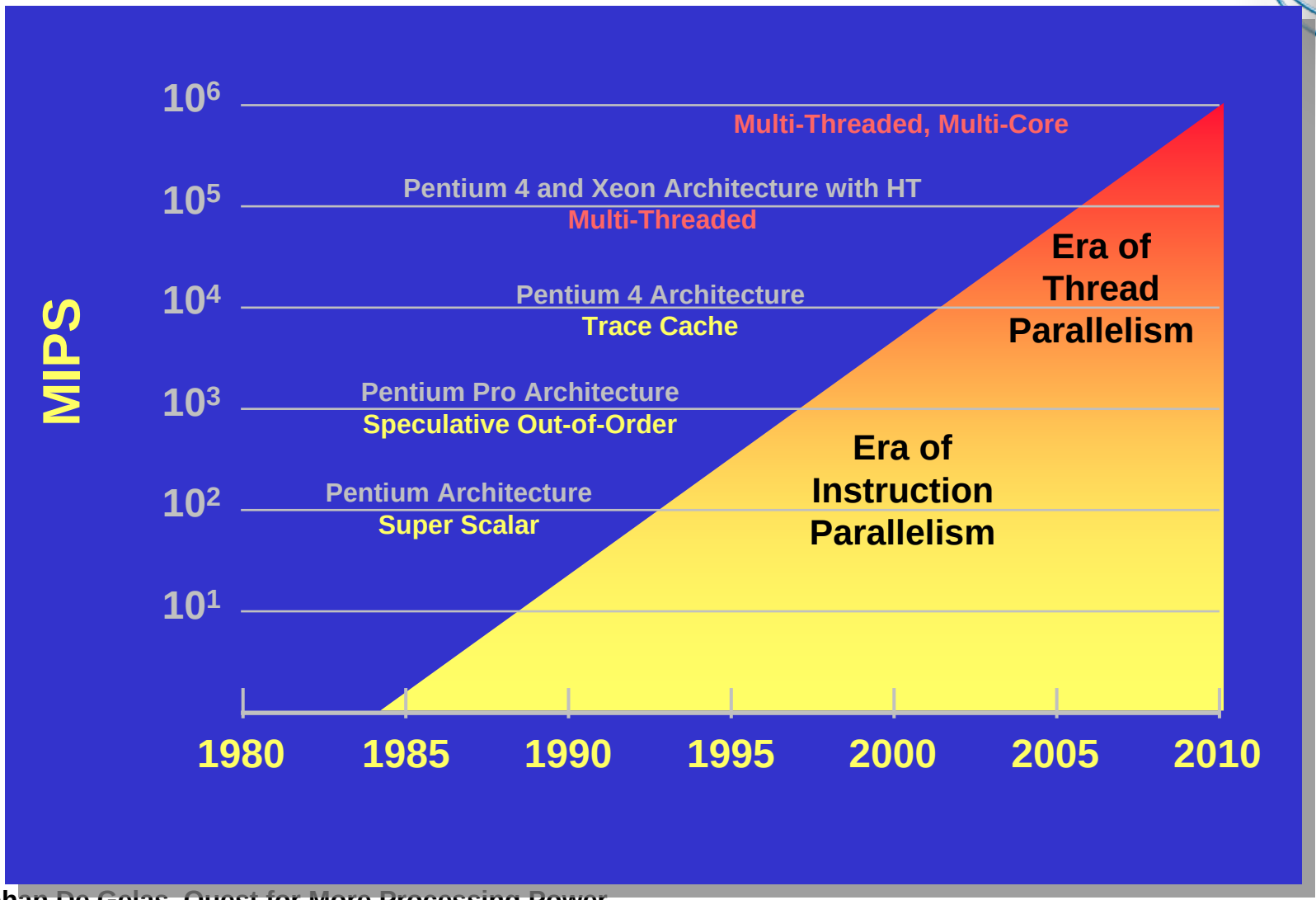
デュアルプロセッサは、386プロセッサの発表以来、性能に関して最大の向上を実現します。しかし、このような性能向上には、ソフトウエアの最適化がプロセッサの性能をフルに発揮するためには必要です。

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

インテルの方向性

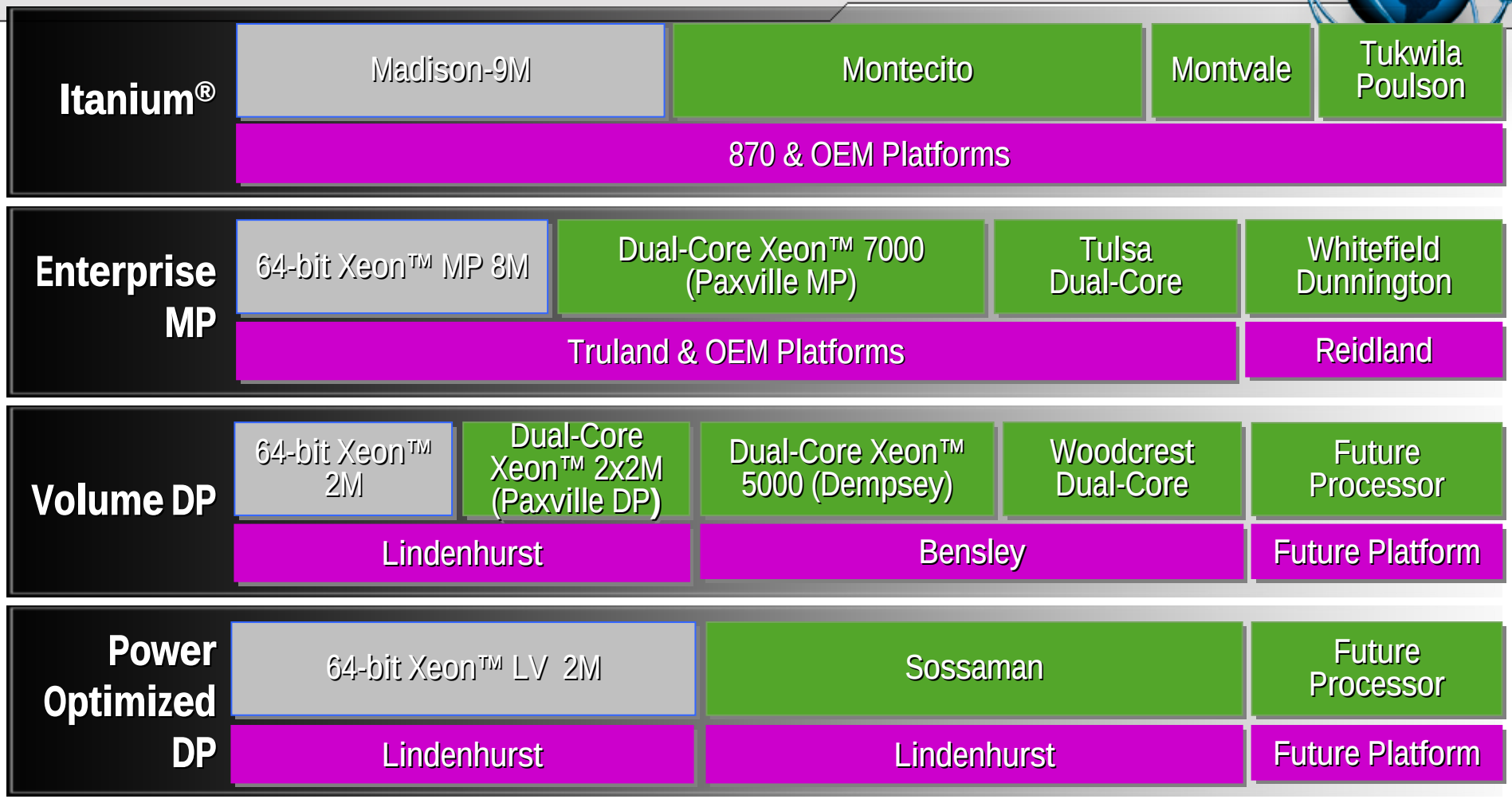
- インテルはマルチコア・プロセッサ用のマルチスレッド・ソフトウェアを目標(2005/03/03)
  - “Intel Targets Multithreaded Software for Multi-core Processors,” EE Times, March 3rd, 2005 issue).
- “アプリケーションに特有の言語による並列処理問題を解決することによって、将来のマルチコアプロセッサーは、高速シングルプロセッサの速度に依存することなしに分断攻略を進めることになる、とインテルは予測している。インテルは、エクストリーム版と呼ばれるデュアルコアペンティアムの最初のサンプル組立ての成功を報告し、2005 年前半に出荷し始めることを計画している。

マイクロアーキテクチャのトレンド



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

Server & Workstation Roadmap



IDF Fall 2005で公開されたサーバ向けプロセッサのロードマップ
2006年には、高密度サーバ向けにノートPC向けを流用した開発コード名「Sossaman」で呼ばれるプロセッサが投入される。チップセットとしては、現行のIntel Xeon向けのIntel E7520／E7320が組み合わされる。

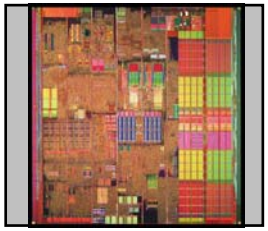
新しいプロセッサ技術の導入

- 省電力プロセッサ
  - 今まで以上のアプリケーションのスケーラビリティ
    - ~100,000プロセッサでのスケーラビリティ(ピーク)
    - ~1,000プロセッサ(通常運用での利用?)
  - プロセッサ障害でのリカバリ(耐障害性やチェックポイント)
- マルチコアプロセッサ
  - アーキテクチャに関する正しい理解と問題点の把握
    - メモリバンド幅
    - キャッシュの競合やコヒレンシの問題など

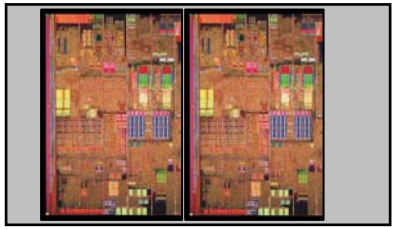
マルチコアテクノロジ

- 将来のマイクロプロセッサ
  - ソケットあたりの性能向上をマルチコアテクノロジで実現
  - 複数プロセッサによる並列処理によって、スケーラブルな性能の実現を目指す
  - マルチコアは、サーバだけでなく、ワークステーション、デスクトップ、モバイルのすべてのコンピュータで利用可能
- これらのマルチコアの性能を最大限に発揮するためのソフトウエアインフラの整備が重要

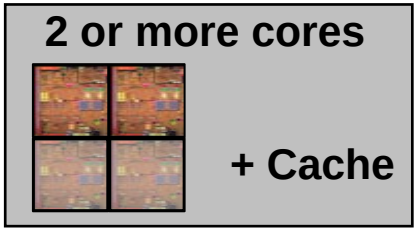
Today
Single Core



2005-2006
Dual Core

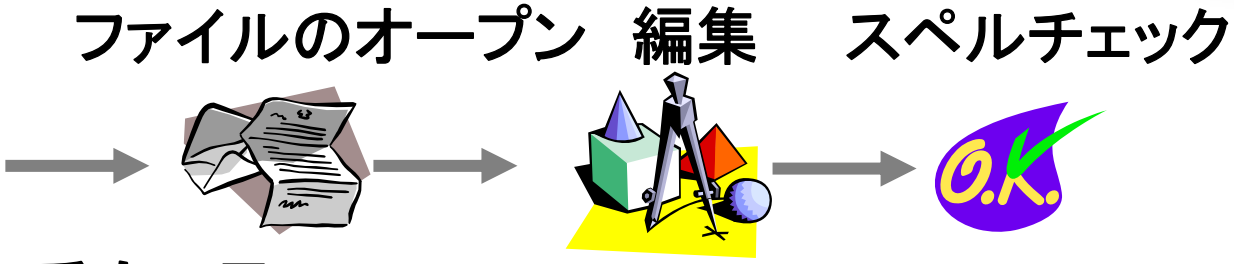


Future
Multi-Core

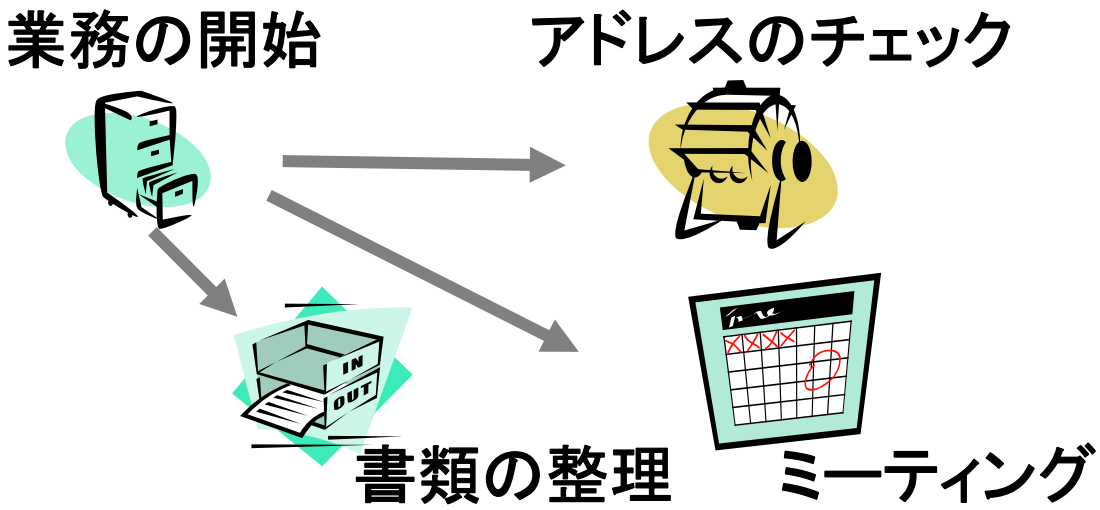


逐次、多重処理

- タスクの逐次処理



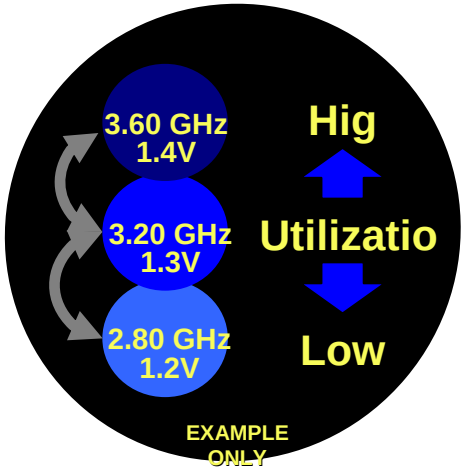
- タスクの多重処理



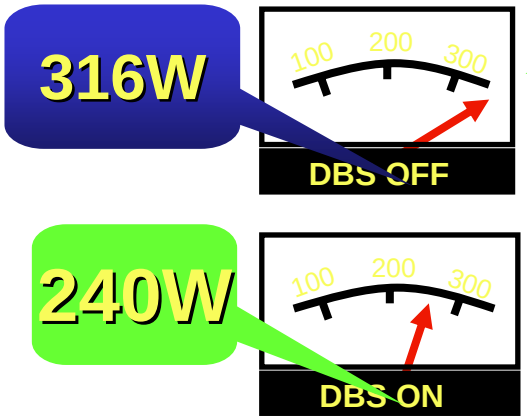
動的な電力管理機能:DBS

1U Intel® Xeon™ processor platform

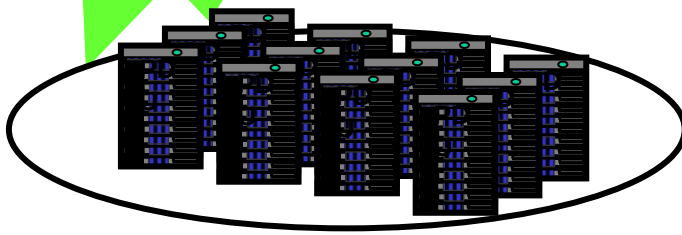
45% CPU utilization running WebBench



Power Meter



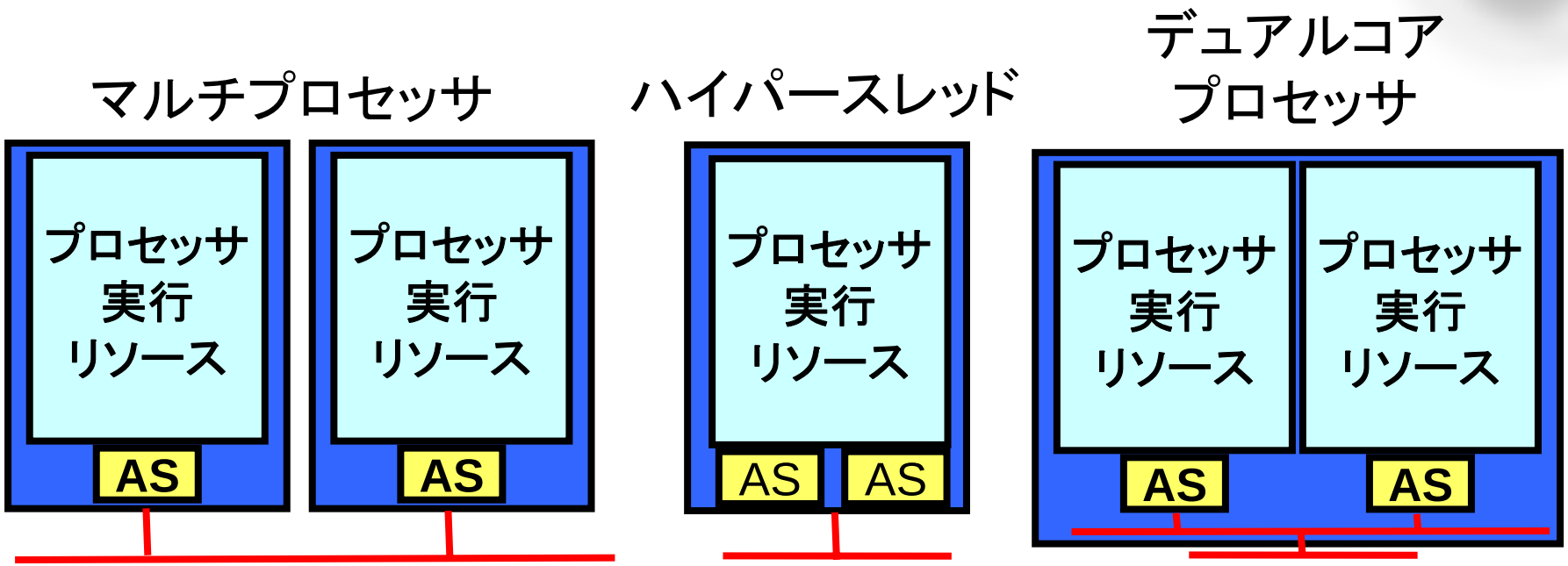
最大28%の電力を節約し、システムの運用コストを大幅に低減する



Demand Based Switching (DBS)
Intel® SpeedStep 技術を拡張

Any difference in systems tested, configurations or assumptions made may affect actual results or performance.

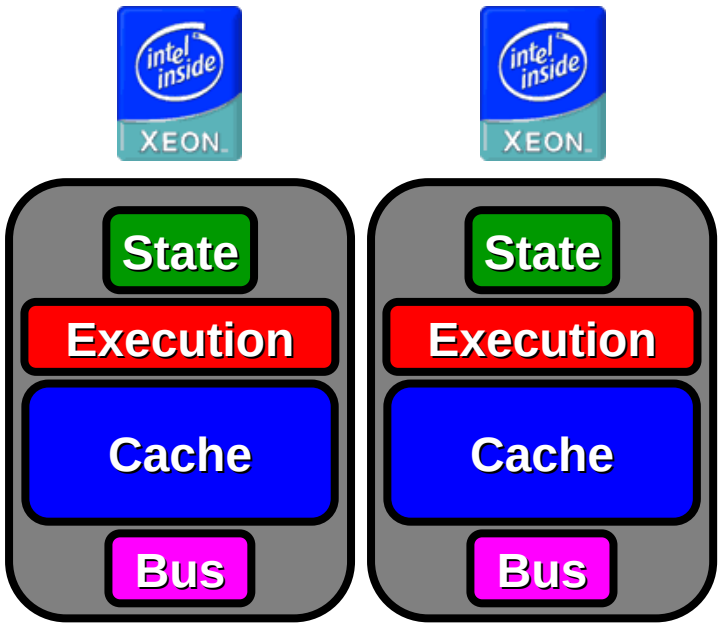
プロセッサアーキテクチャ



AS (Architecture State) は、汎用レジスタや制御レジスタ、APIC (Advanced Programmable Interrupt Controller) レジスタなどプロセッサの状態を保持するものです。

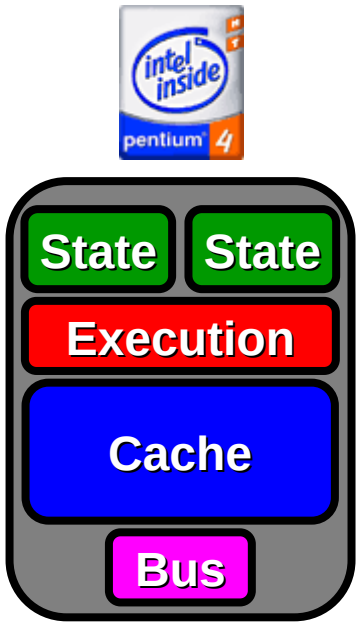
マルチスレッド処理

Dual Xeon Processors



2 Threads
2 Packages

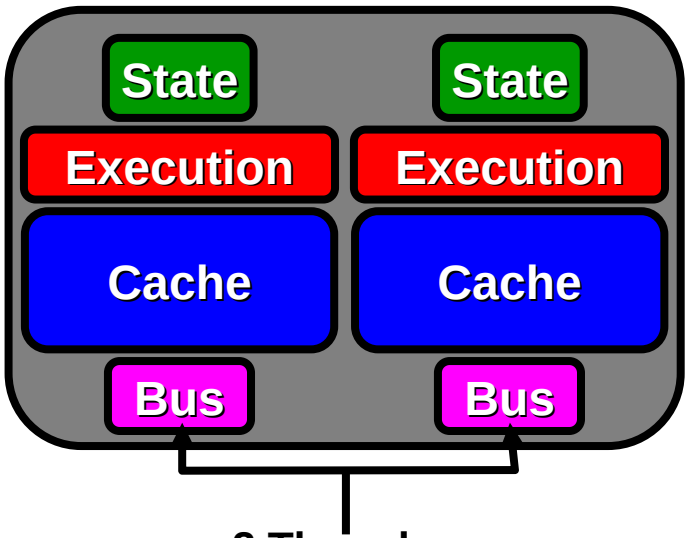
Pentium 4 with HT



2 Threads
1 Package

デュアルコアプロセッサ

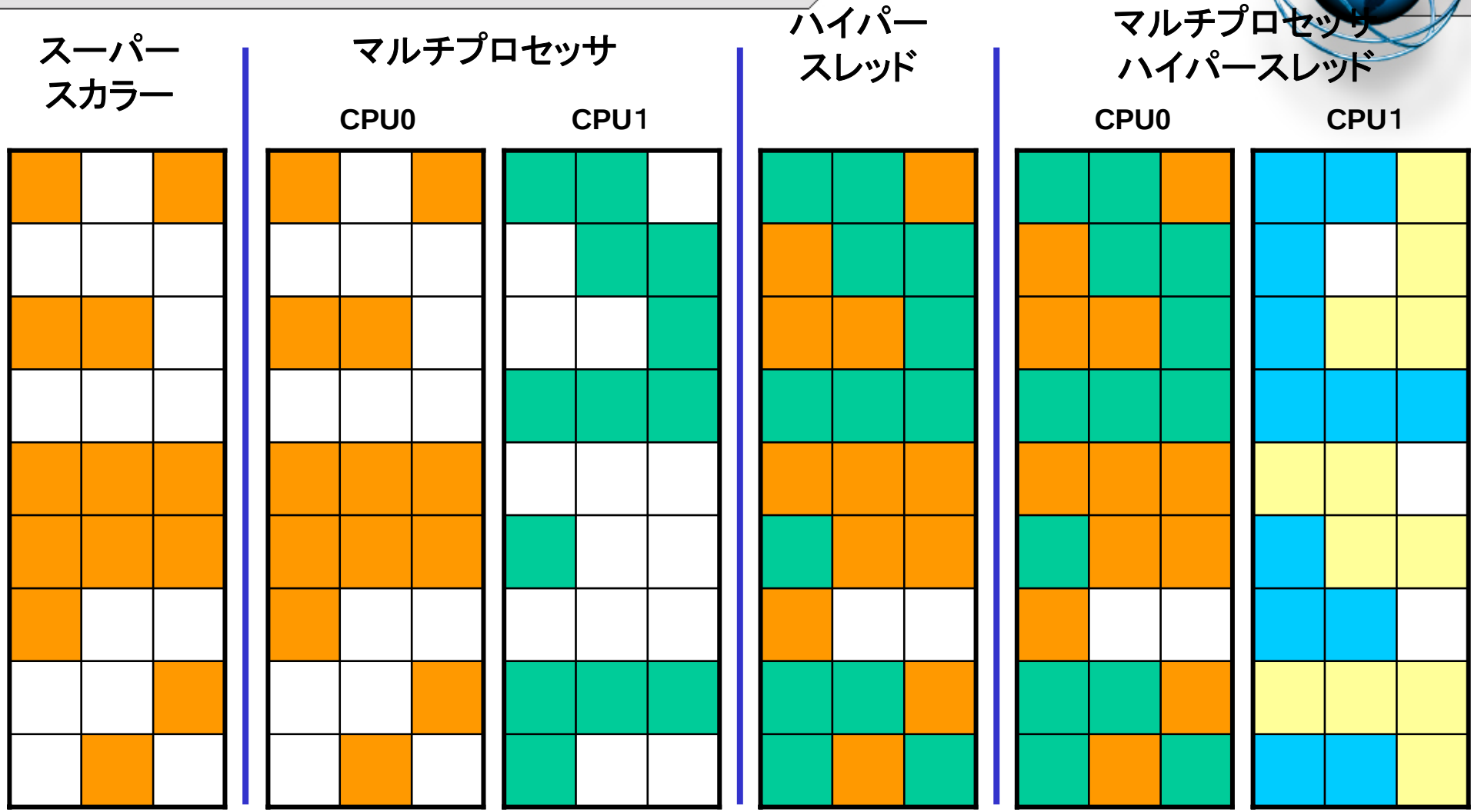
同じプロセッサ内に完全に独立したプロセッサコアを実装



2 Threads
1 Package

シングルパッケージ内でのマルチスレッドサポート

マルチスレッド処理



各ボックスがプロセッサの各演算機を示しています。1クロックあたり、複数の演算器が同時に並列に動作します。

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

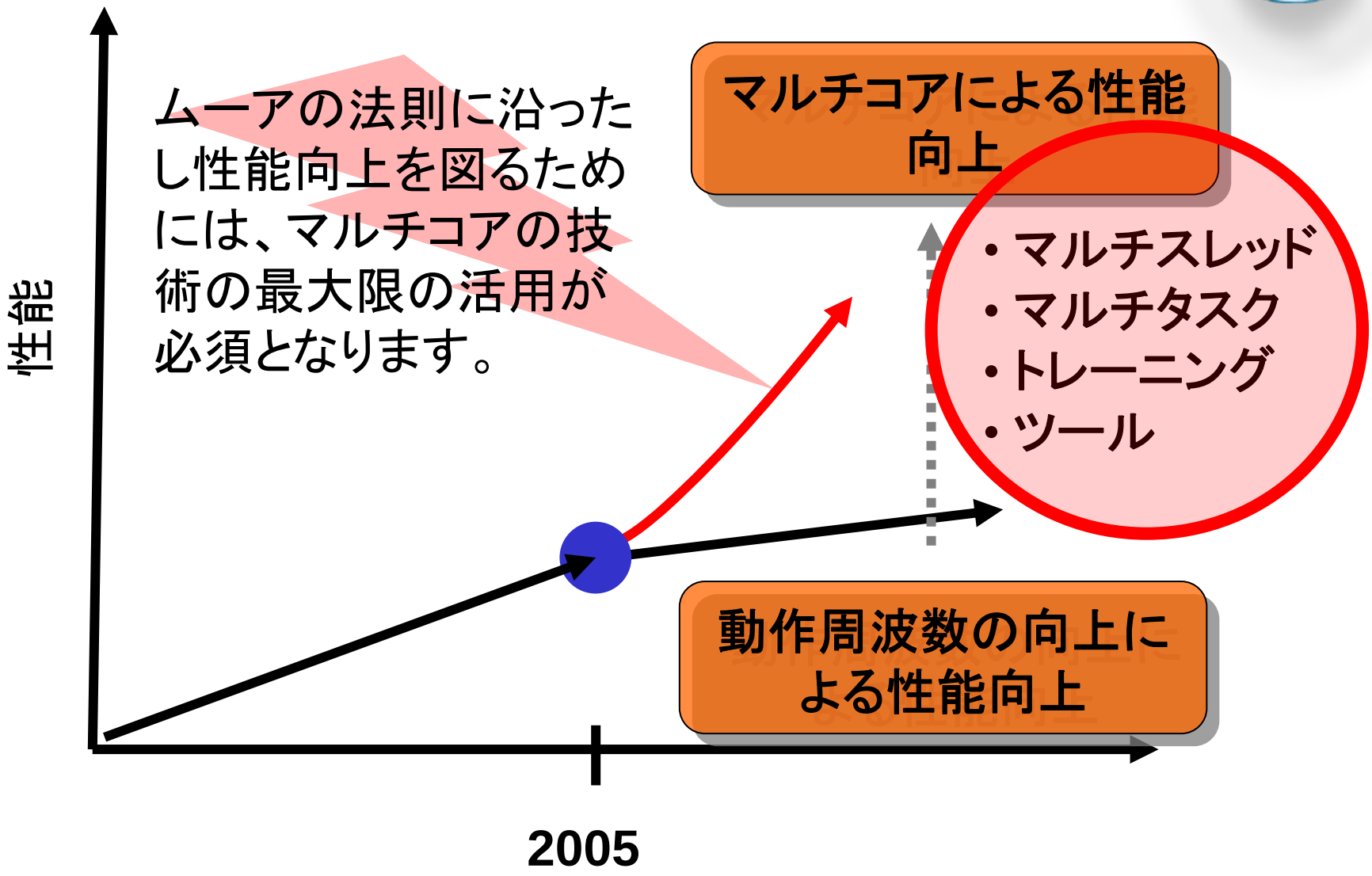
- OS自身のマルチスレッド対応
  - OSのサービスもマルチスレッドで処理することで、より効率よく処理することが可能
- 仮想化
  - サーバのセキュリティや管理の強化
  - 管理するノード数を減らし、運用コストの削減を図る
- 最新のソフトウエア・テクノロジの活用

マルチコアの利点?

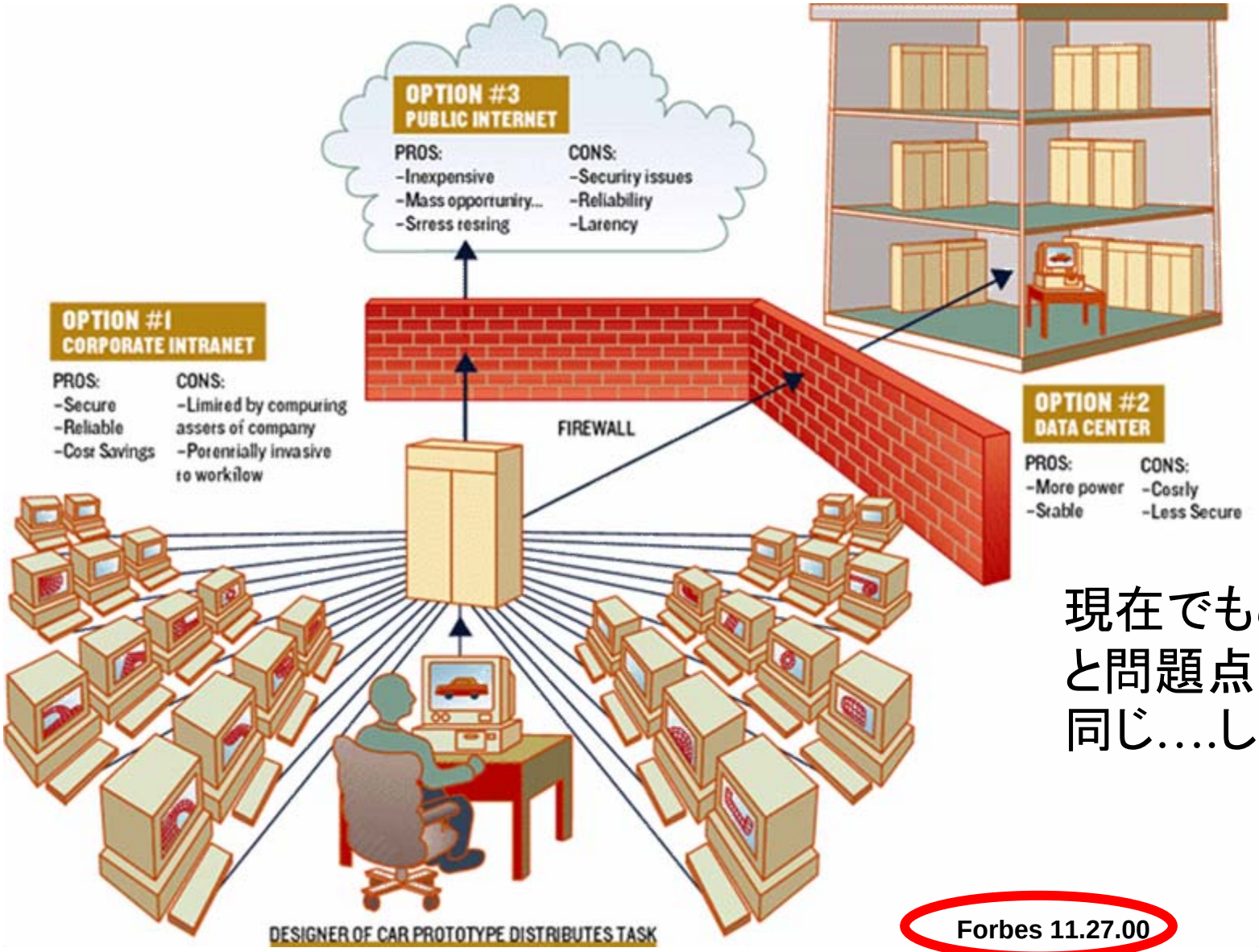
最新のソフトウエア・テクノロジの活用
消費電力あたりの性能を最大にし、高性能で低消費電力のシステム構築が可能

- マルチスレッド
  - 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
  - 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
- 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
- 仮想化
  - サーバのセキュリティや管理の強化

ムーアの法則（GHz から MC へ）



コンピュータ利用形態

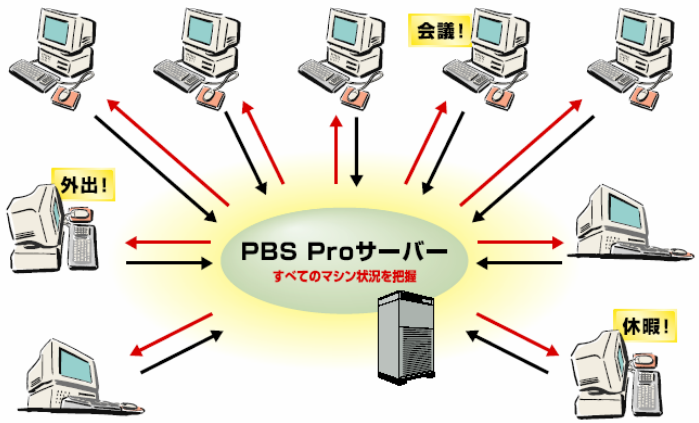


現在でもこの利点と問題点の問題は同じ….しかし

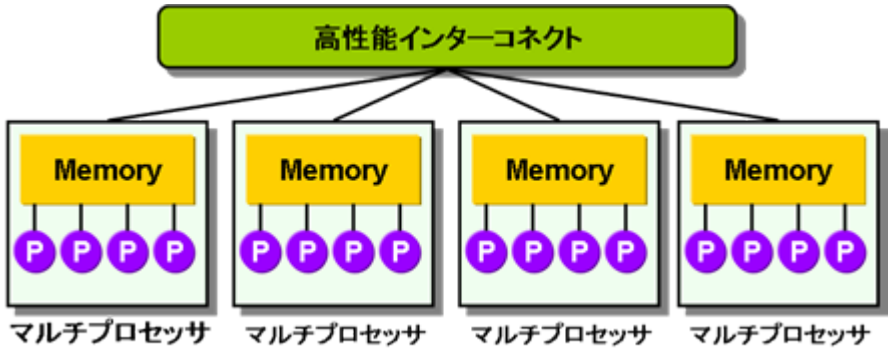
Forbes 11.27.00

計算機利用形態の進化

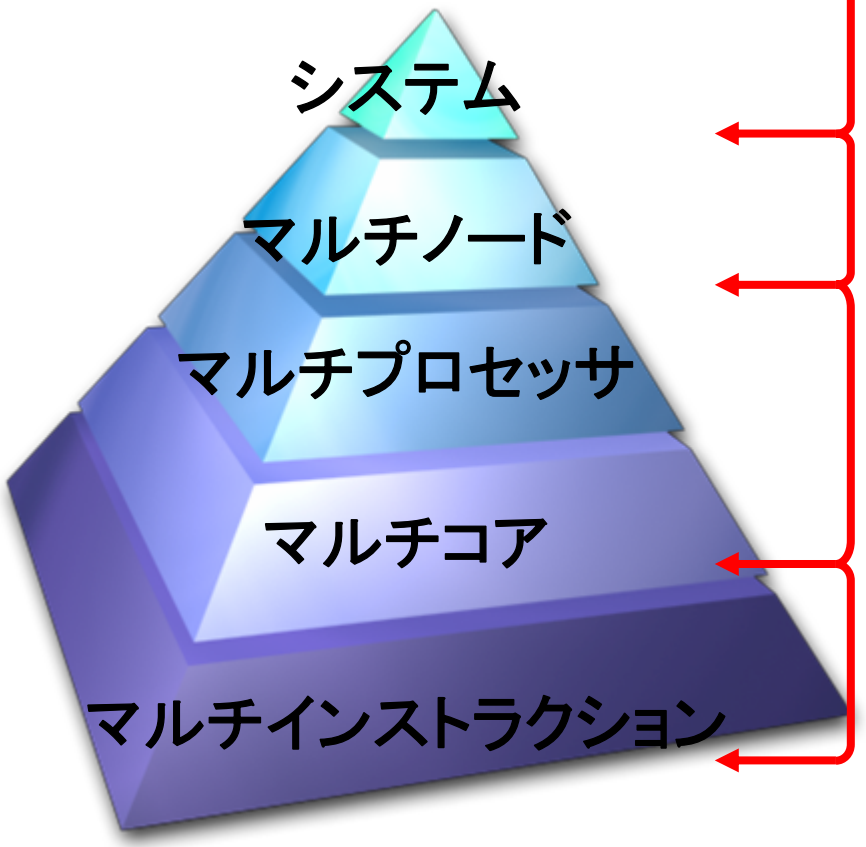
- デスクトップの計算能力の著しい向上（>10GFLOPS級の計算能力）
- 遊休CPUリソースを利用した‘計算クラスタ’の構築



- クラスタノードの性能向上（SMPノード）
- より大規模なクラスタ構成
- ハイブリッド型（SMP+MPI)や新しいプログラムモデルへの対応が必要



並列性(Parallelism)の利用



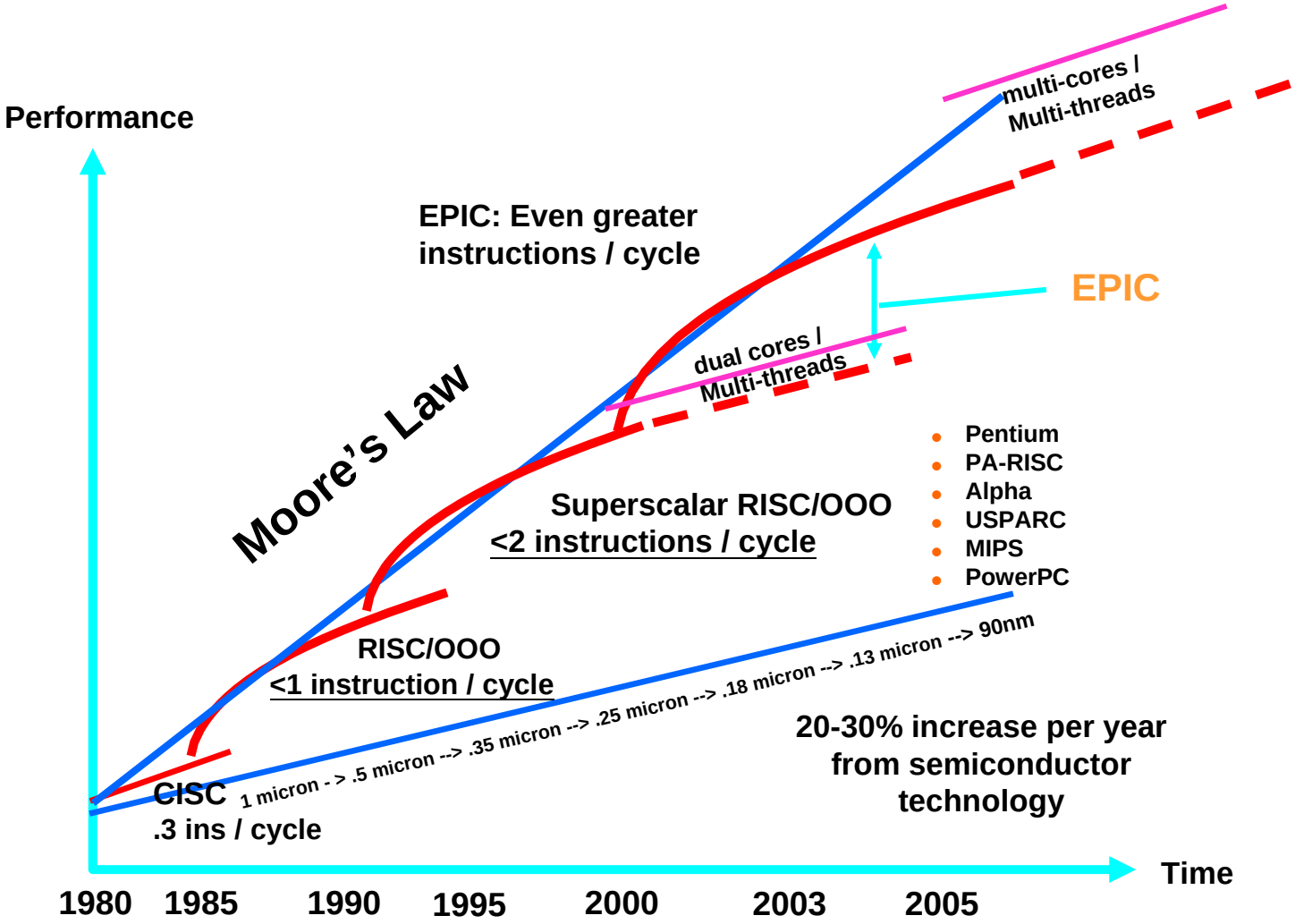
マルチタスク処理

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)

ムーアの法則の維持?


Performance
multi-cores /
Multi-threads
EPIC: Even greater
instructions / cycle
EPIC
dual cores /
Multi-threads
Moore's Law
Pentium
PA-RISC
Alpha
USPARC
MIPS
PowerPC
Superscalar RISC/OOO
<2 instructions / cycle
RISC/OOO
<1 instruction / cycle
1 micron - > .5 micron --> .35 micron --> .25 micron --> .18 micron --> .13 micron --> 90nm
20-30% increase per year
from semiconductor
technology
CISC
.3 ins / cycle
Time
1980
1985
1990
1995
2000
2003
2005

システムの選択

Good, Fast, Cheap -- Pick any two ！

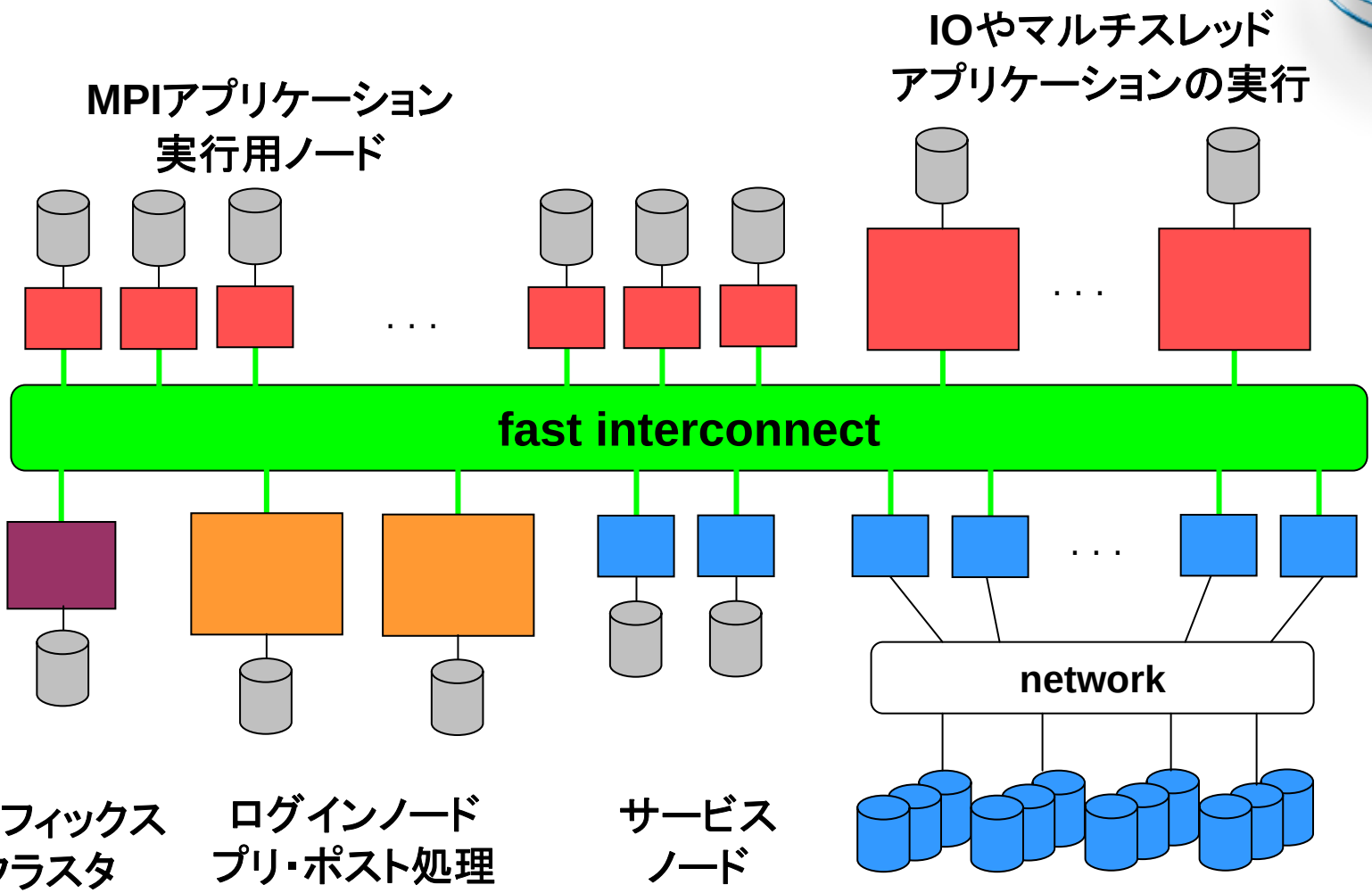
「Fast」「Good」「Cheap」の選択肢

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。

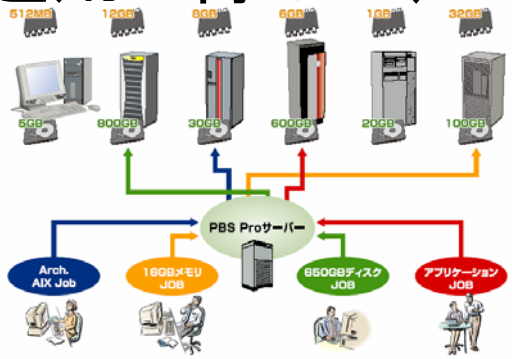
マルチ - 並列アーキテクチャ - システム

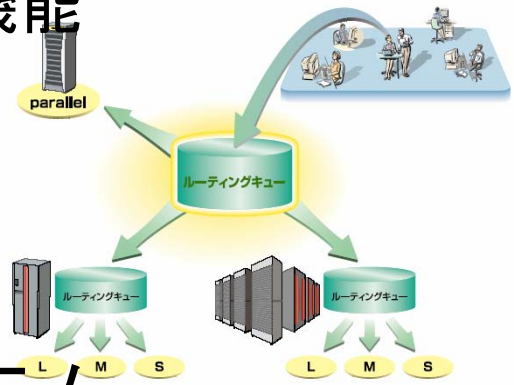


複数の並列アーキテクチャの混在環境

マルチ-並列アーキテクチャ-システム

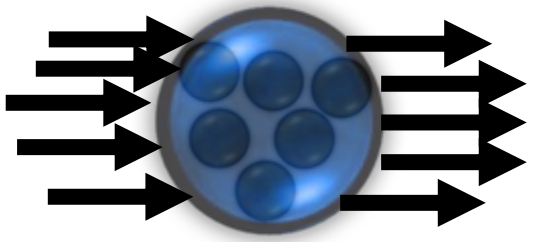
- 単純なバッチシステムでは機能不足
- マルチ‘並列アーキテクチャ’システムの運用に際して、必要とされるバッチシステムの機能
  - マルチレベル・ルーティングキュー機能
  - アクセスコントロールリスト(ACL)機能
  - 動的なリソース割当て機能
  - アーキテクチャを考慮したジョブの割り当て機能
  - リソース要求に応じたジョブの投入機能





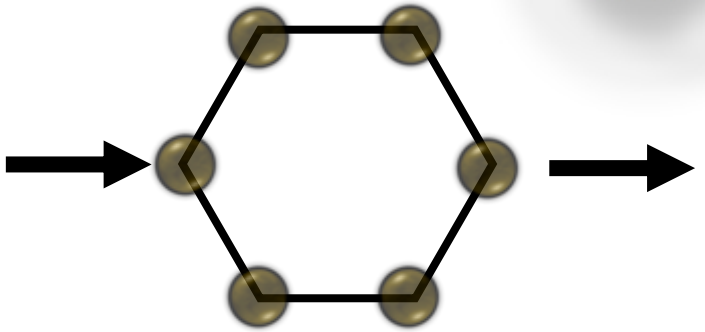
上記の全ての機能を含む、マルチ－並列アーキテクチャ－に最適なバッチシステム

仮想化によるサーバ・コンソリデーション

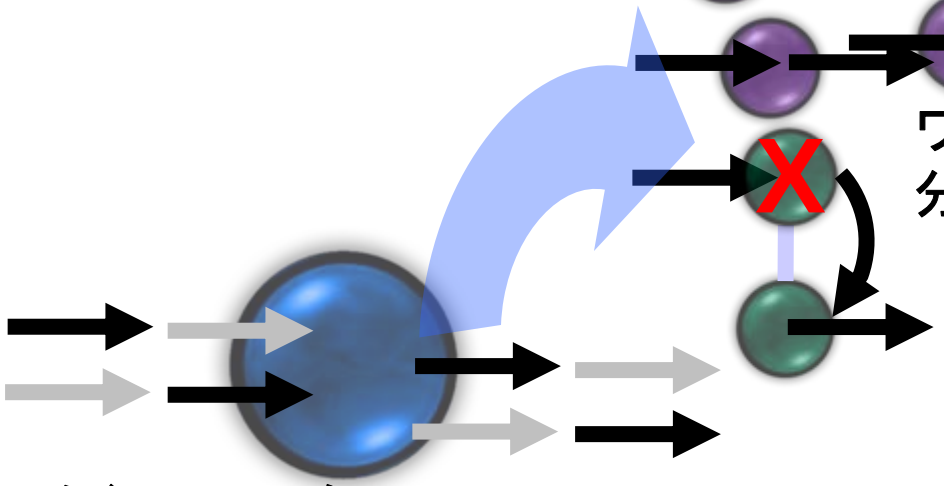


↑コスト・パフォーマンス
↑スケーラビリティ
↑管理コストの削減
↑安定性・信頼性
↓オーバヘッド

クラスタによる並列処理



↑コスト・パフォーマンス
↑スケーラビリティ
↓管理コストの増大



ワークステーション、サーバでの
分散処理

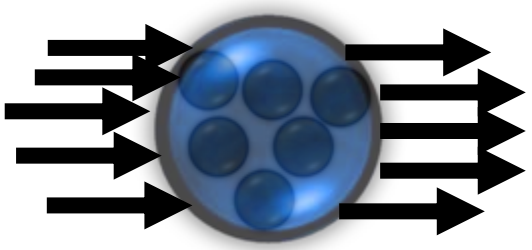
↑コスト・パフォーマンス
↑スケーラビリティ
↑耐障害性
↓管理負担・管理責任の分散
↓管理コストの増大

メインフレーム
スーパーコンピュータ

仮想化によるサーバ・コンソリデーション

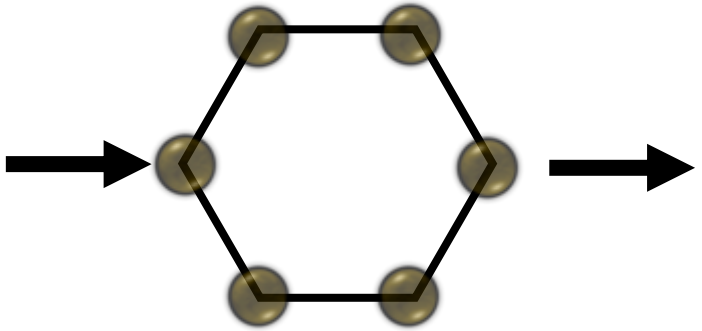
1つのプラットフォーム上で複数のソフトウェア環境を「仮想システム」として混在させ、並列稼動させる仮想化技術を活用し、サーバ群の再統合化し、利用効率、TCOの削減を目指す

仮想化によるサーバ・コンソリデーション



↑コスト・パフォーマンス
↑スケーラビリティ
↑管理コストの削減
↑安定性・信頼性
↓オーバヘッド

クラスタによる並列処理



↑コスト・パフォーマンス
↑スケーラビリティ
↓管理コストの増大

仮想化/バーチャライゼーション

1つのプラットフォーム上で複数のソフトウェア環境を「仮想システム」として混在させ、並列稼動させる

- 仮想化に対する市場の関心は高く、ソフトウェア・ベースで仮想化を実現するソリューションは、ベンダ各社から提供されている
- そうした仕組 みでは仮想化モニタ、つまり、仮想システムをモニタリングし、リソースを割り当てる仕組み自体がシステム・リソースを消費するため、パフォーマンスの劣化 は避けられない

仮想化/バーチャライゼーション

- 現在、各社はこのパフォーマンス劣化に対応するための技術開発に注力
- インテル社：“インテル® バーチャライゼーション・テクノロジ”
  - ハードウェア・レベルで仮想化の仕組みを支える技術で、ソフトウェアだけで実現しようとした場合に見られるボトルネックを最小限に抑え、仮想化本来の メリットを最大化する
- 他社も同様のプロジェクトでの仮想化/バーチャライゼーションに対応（AMDの「Pacifica」など）
  - **“Virtual Leverage: Server Consolidation in Open Source Environments”** Margaret Lewis,Commercial Software Strategist,AMD

サーバ仮想化の利点

柔軟なサーバ・コンソリデーション	サーバ仮想化によって、さまざまなオペレーティング・システムやアプリケーションを短時間で簡単に 2-way ～ 16-way 以上のプラットフォームに統合できる
可用性とセキュリティの向上	ソフトウェア障害やデジタル攻撃を仮想パーティションに隔離したり、フェイルオーバ・パーティションを設置して簡単かつ経済的にニーズに合った可用性を実現できる
OS およびハードウェア移行の簡略化	サーバ仮想化によって、レガシ・アプリケーションや既存 OS のバージョンを変更せずに仮想パーティションに移行できる
テストおよび開発環境の合理化	単一のプラットフォームでソフトウェア・スタックごとに複数のテスト環境をホスティングし、繰り返し利用できる
ビジネスの機敏性の向上	仮想パーティションのプロビジョニングやサイズ変更を簡単に行って、新しいアプリケーションやワークロードの増大、システム保守に対応できる

インテル社ホームページより

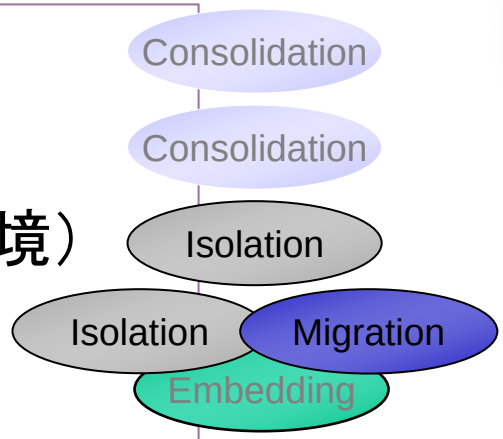
サーバ・コンソリデーション

- サーバ群の再統合化
  - 分割が進み過ぎて、管理コストが増大したサーバ群の再統合
  - ネットワーク・コンピューティングの利点を生かしながら、管理コストの削減を図り、また、手厚く管理すべき部分への適切な対応を図ることを目指す
  - 仮想サーバの一部に問題が発生しても、他のサーバへの影響を排除し、安定性や信頼性を犠牲にすることなく、統合化を図る

仮想化の利用モデル

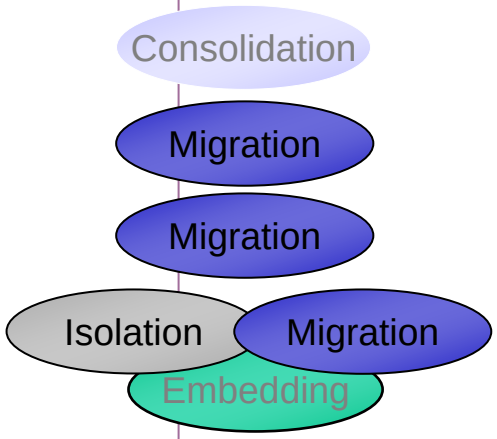
CLIENT

- レガシーSWサポート
- トレーニングとQA
- パーティション（OS環境）
- 管理機能
- …



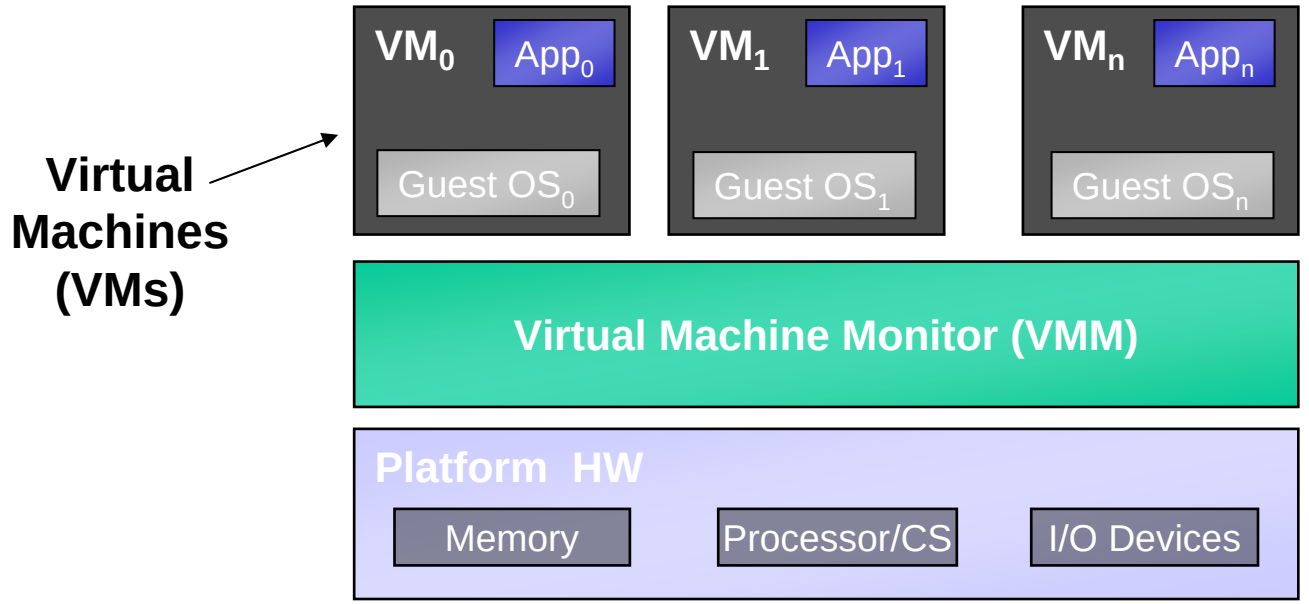
SERVER

- サーバの合併
- フェールオーバ
- データセンター運用
- 管理機能
- …



サーバをはじめ、デスクトップやノートブック PC まで含めたすべての
プラットフォームで利用可能

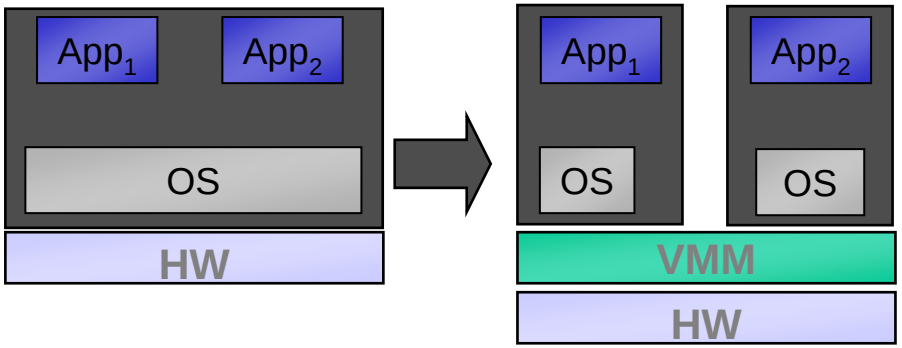
Virtual Machine Monitors (VMMs)



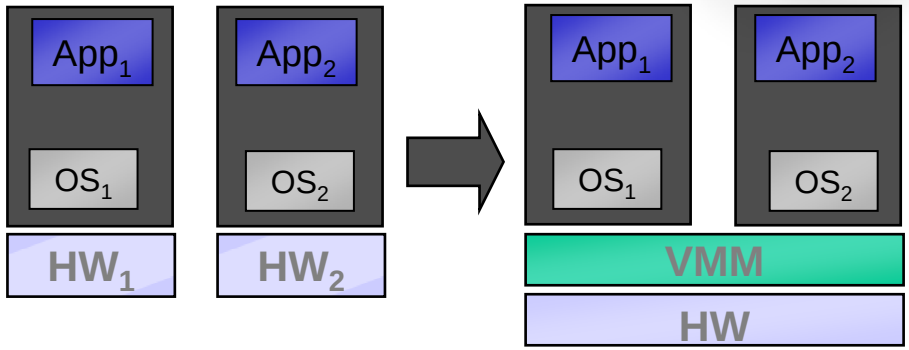
- VMM は、システムソフトウエアの一つのレイア
  - 複数のVMがプラットフォームハードウエアを共有
  - アプリケーションを変更なしで実行可能とする

仮想化の利用方法

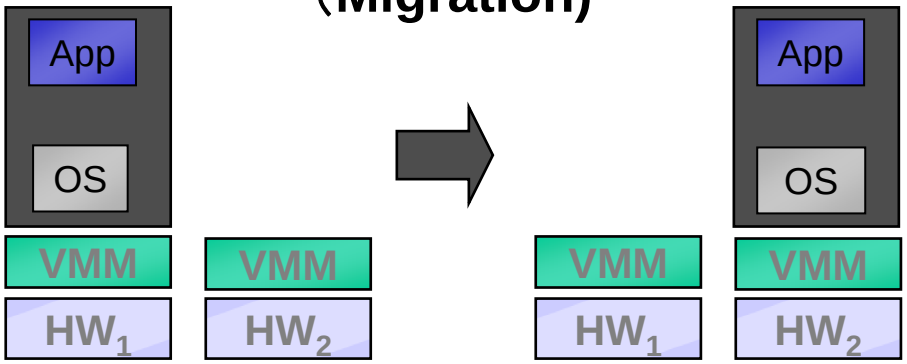
ワークロードの分離（Isolation）



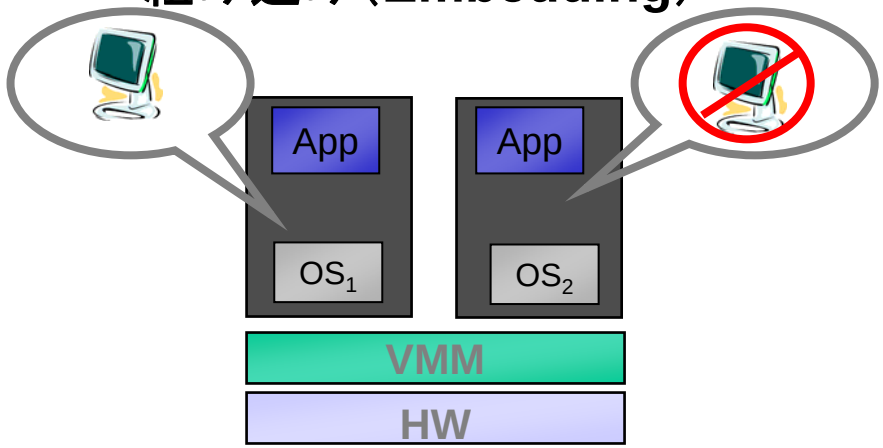
リソースの統合（Consolidation）



ワークロードのマイグレーション（Migration)



組み込み（Embedding）

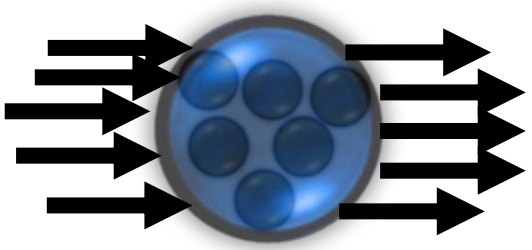


ユーザに対して、柔軟性が高く、運用効率に優れたコンピューティング環境を提供

クラスタによる並列処理

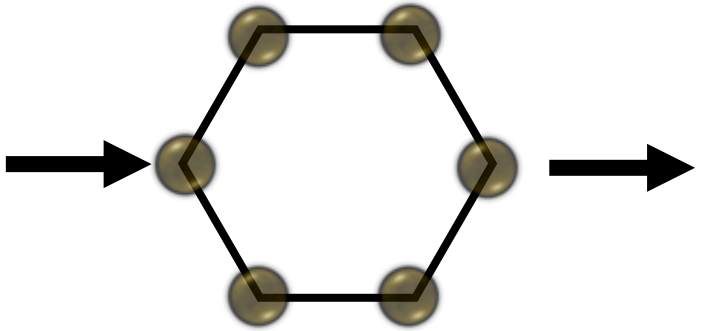
HPCシステムは、基本的にはノードと呼ぶ計算機システムを複数接続したものになります。このノードとなる計算機システムとしては、ベクトル計算機、RISCプロセッサを搭載したSMPシステム、インテル Xeonなどに代表される商用プロセッサを1～2プロセッサなどがあります。

仮想化によるサーバ・コンソリデーション



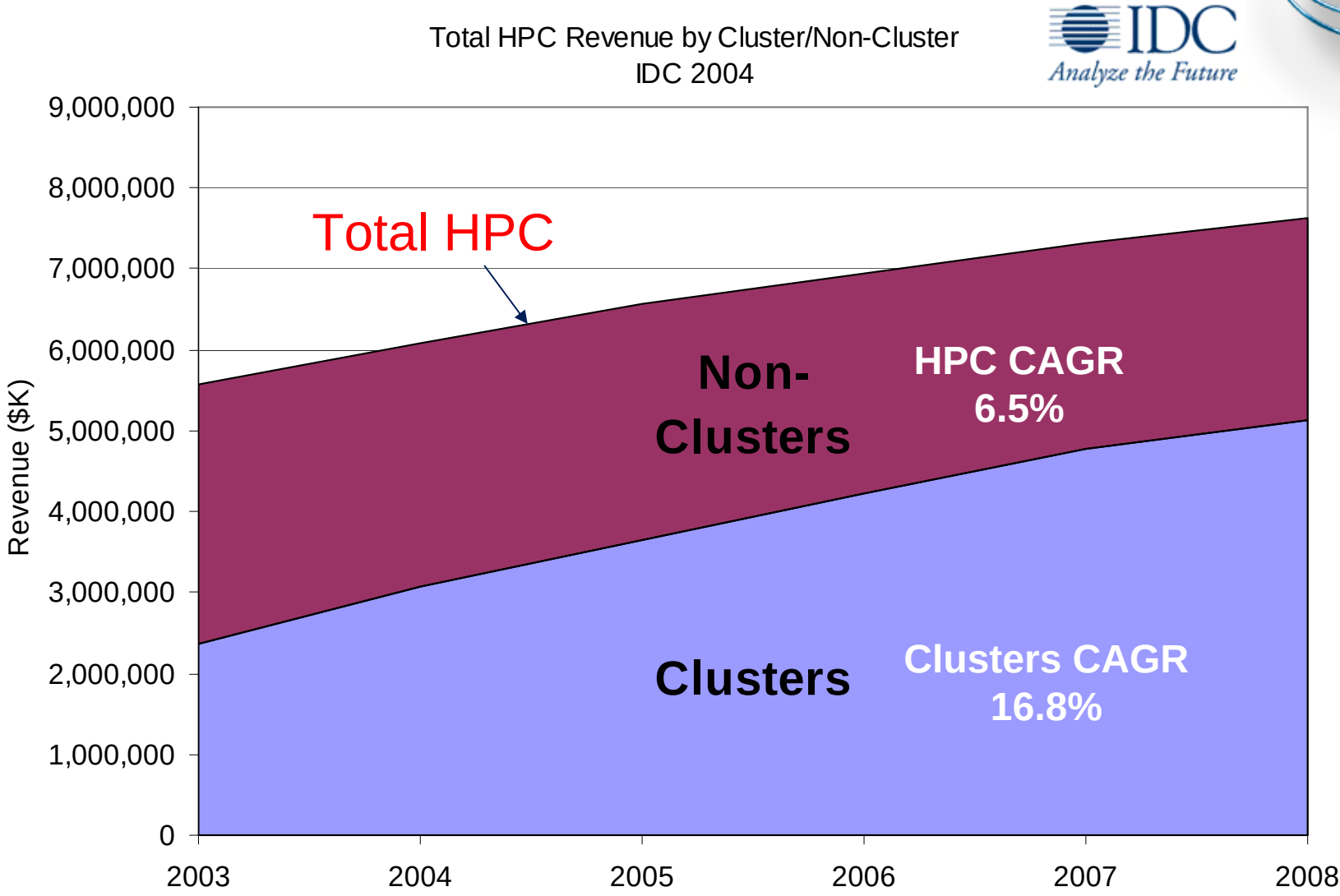
↑コスト・パフォーマンス
↑スケーラビリティ
↑管理コストの削減
↑安定性・信頼性
↓オーバヘッド

クラスタによる並列処理



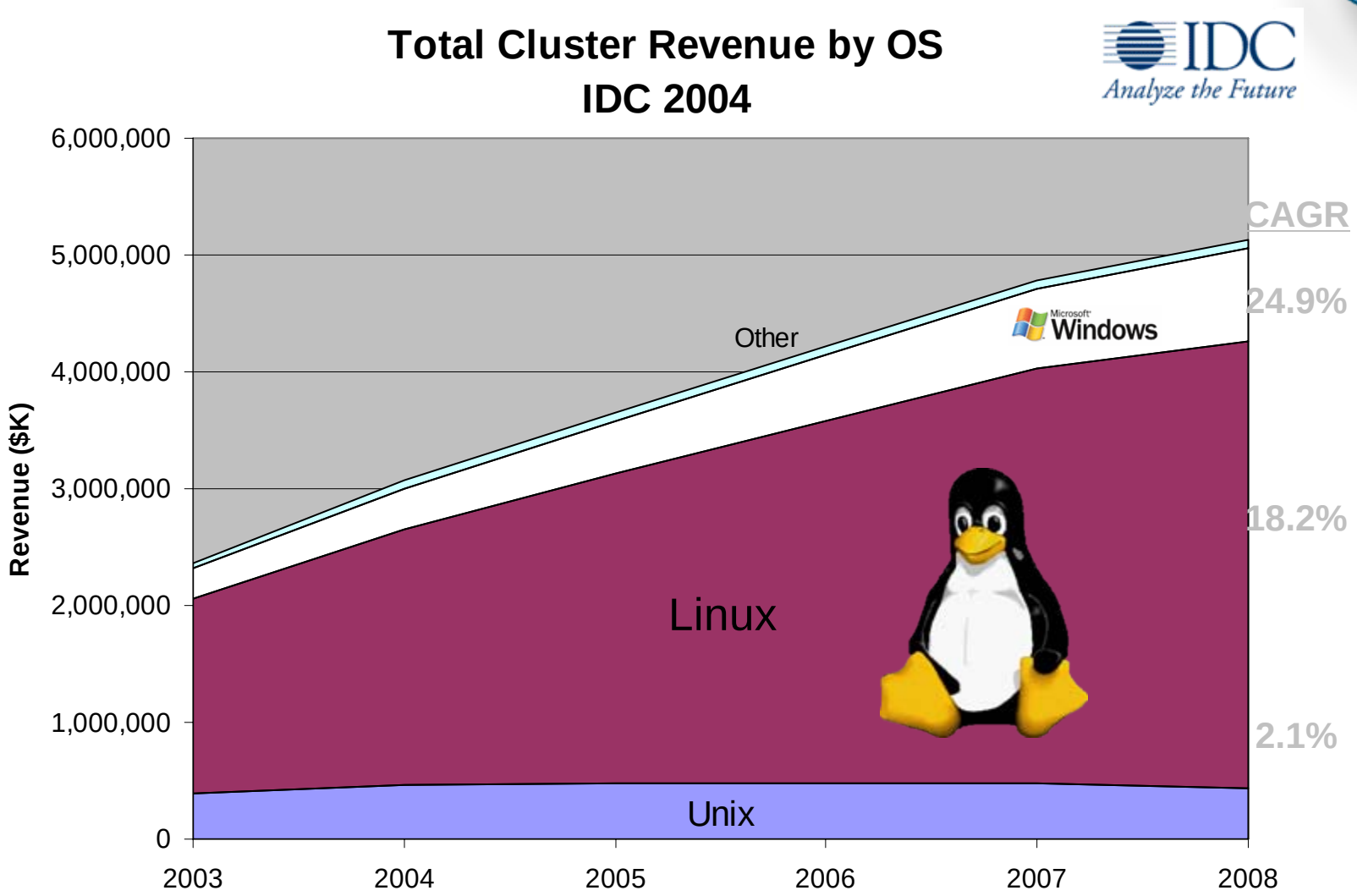
↑コスト・パフォーマンス
↑スケーラビリティ
↓管理コストの増大

HPC関連売り上げ(クラスタ・非クラスタ)

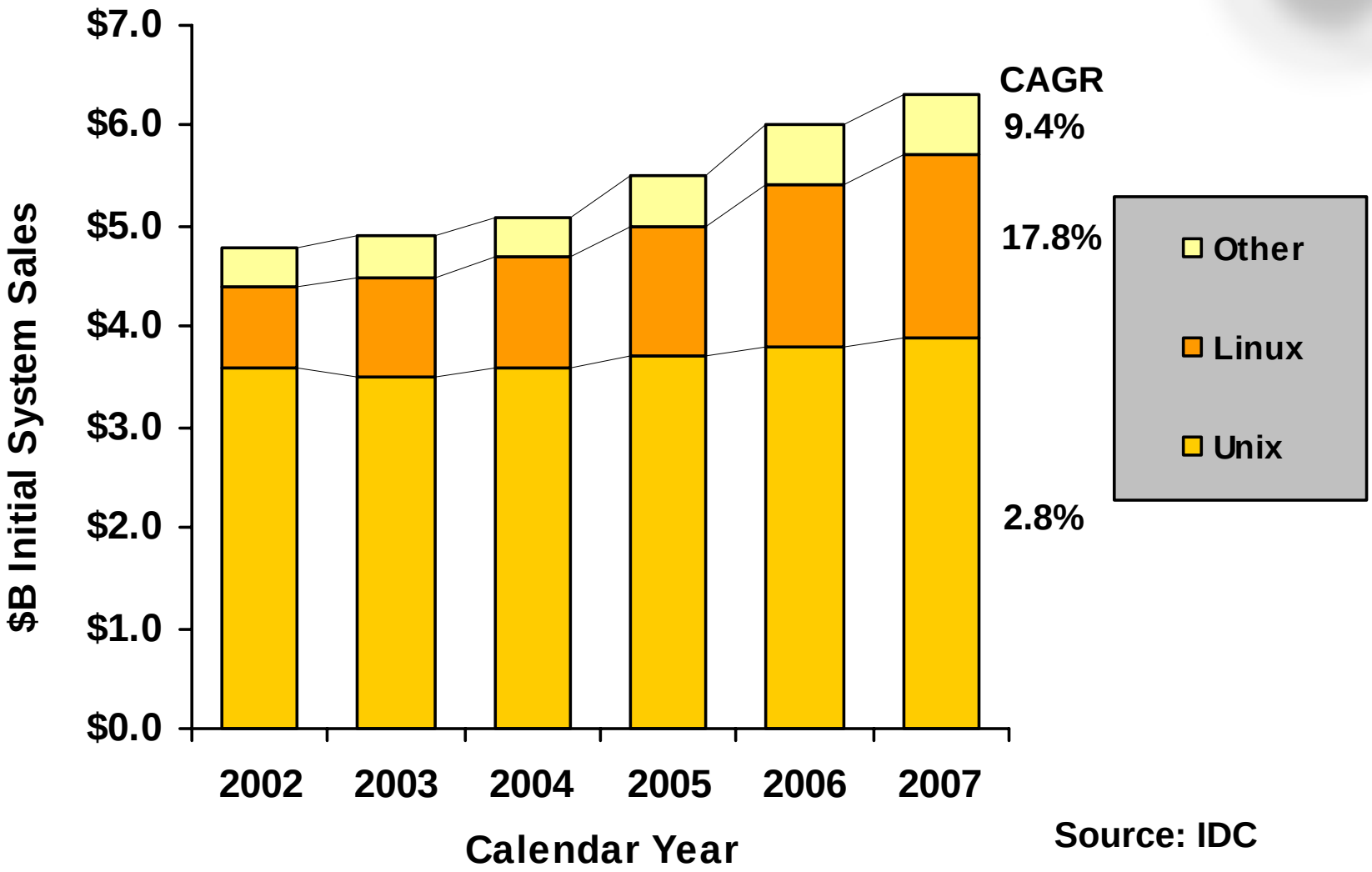


Source: IDC 2004

OS別クラスタ売り上げ

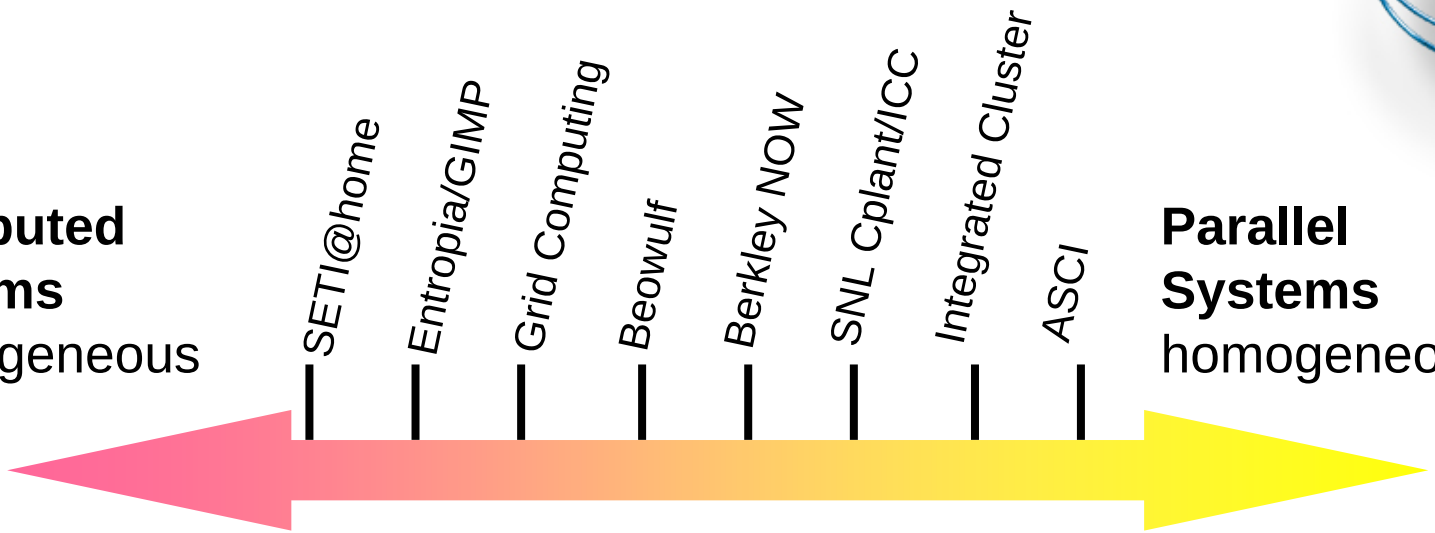


OS別のクラスタシステムの売り上げ



Source: IDC

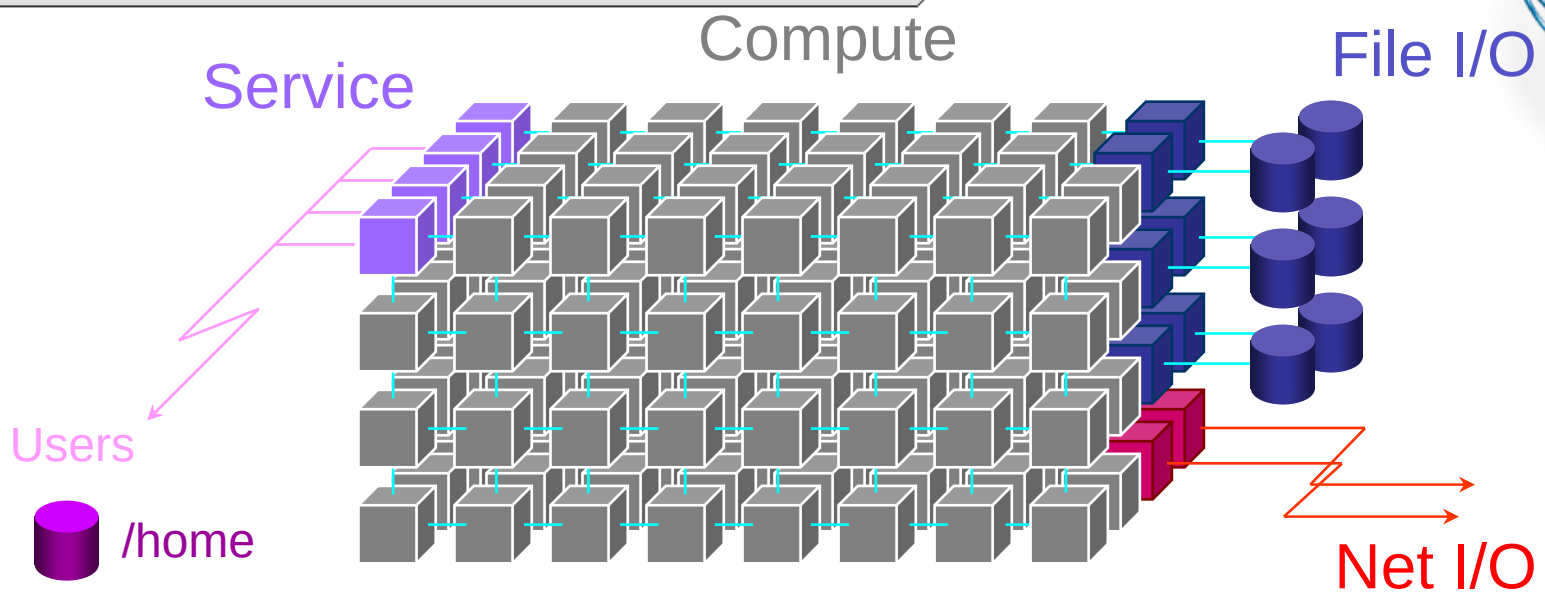
‘Distributed’ と ‘Parallel’ システムの比較



- 広範囲の資源を対象
- 遊休資源の活用
- システムSWによる管理
- システムSWによる付加価値
- 運用上のオーバヘッドには寛容 (>20%?)
- 利用可能な資源に応じた利用アプリケーションの選択
- 各ジョブの終了時間は不定
- 利用資源は基本的には共有

- 管理された運用資源
- アプリケーションは資源全体の利用を想定
- アプリケーションの実行のための資源
- システムSWは、アプリケーションの効率的な実行がメイン
- 可能な限り少ないオーバヘッド
- 利用するアプリケーションに応じた計算機資源の導入
- より短時間でのジョブの処理
- 計算機資源はスペースを共有

SNL Cplant/ICCコンセプト

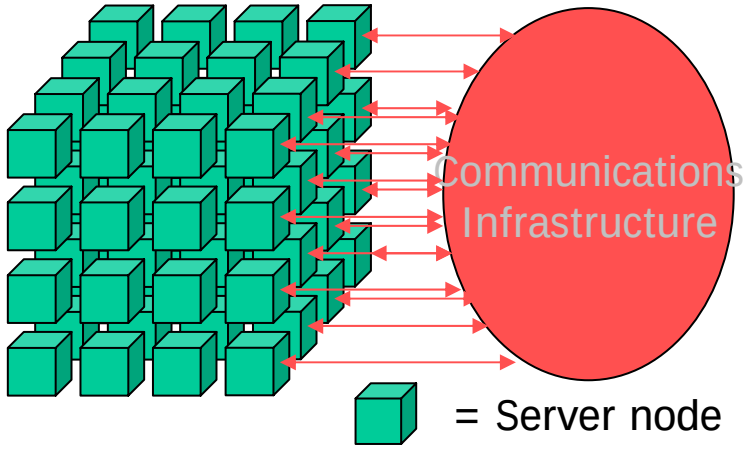


SNL Cplantは、非常に速い時期に計算やファイルI/Oも含んだ統合されたクラスタシステムの提案を求めています。このシステムでは、単にPCを組み合わせてシステムを構築するのではなく、スーパーコンピュータとしての、クラスタの構築を目指しています。

クラスタシステム

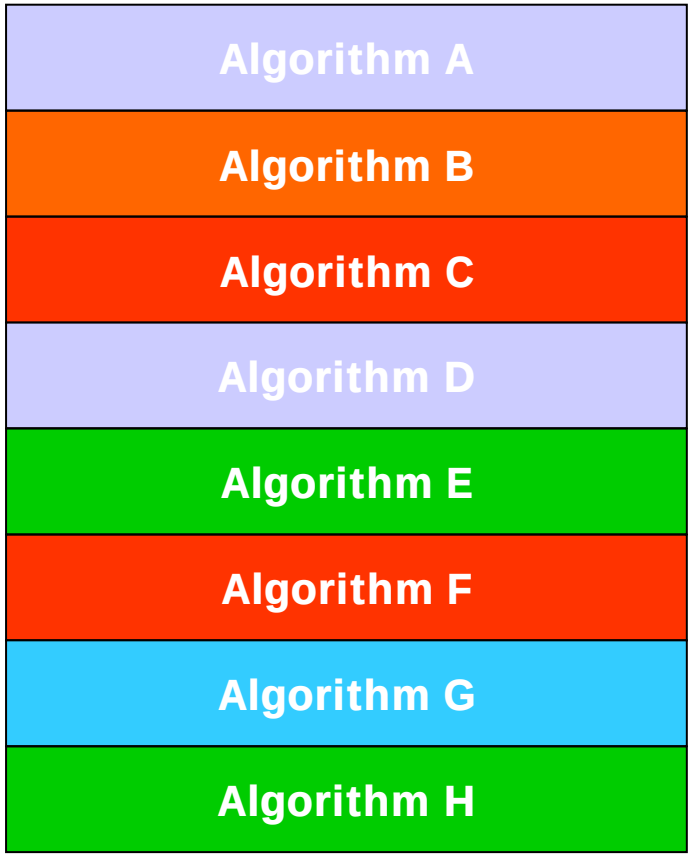
- 複雑なワークロードへの対応
  - スケジューリングの選択肢
  - クラスタ構成トポロジーへの対応
- スケーラビリティ
  - 数プロセッサから数百プロセッサを使用するジョブの処理
- 課題としての可用性の向上
  - Fail-Over
  - 非常に長時間のジョブ実行時間への対応
- 計算機資源の有効利用と計算の生産性の向上（ターンアラウンドの改善）
- 実績

Cluster Parallel Processing



- 一般商用プロセッサを利用した計算ノード
- 商用のインターコネクト用通信インフラによるシステム構築
- システムアーキテクチャに適したアプリケーションの開発

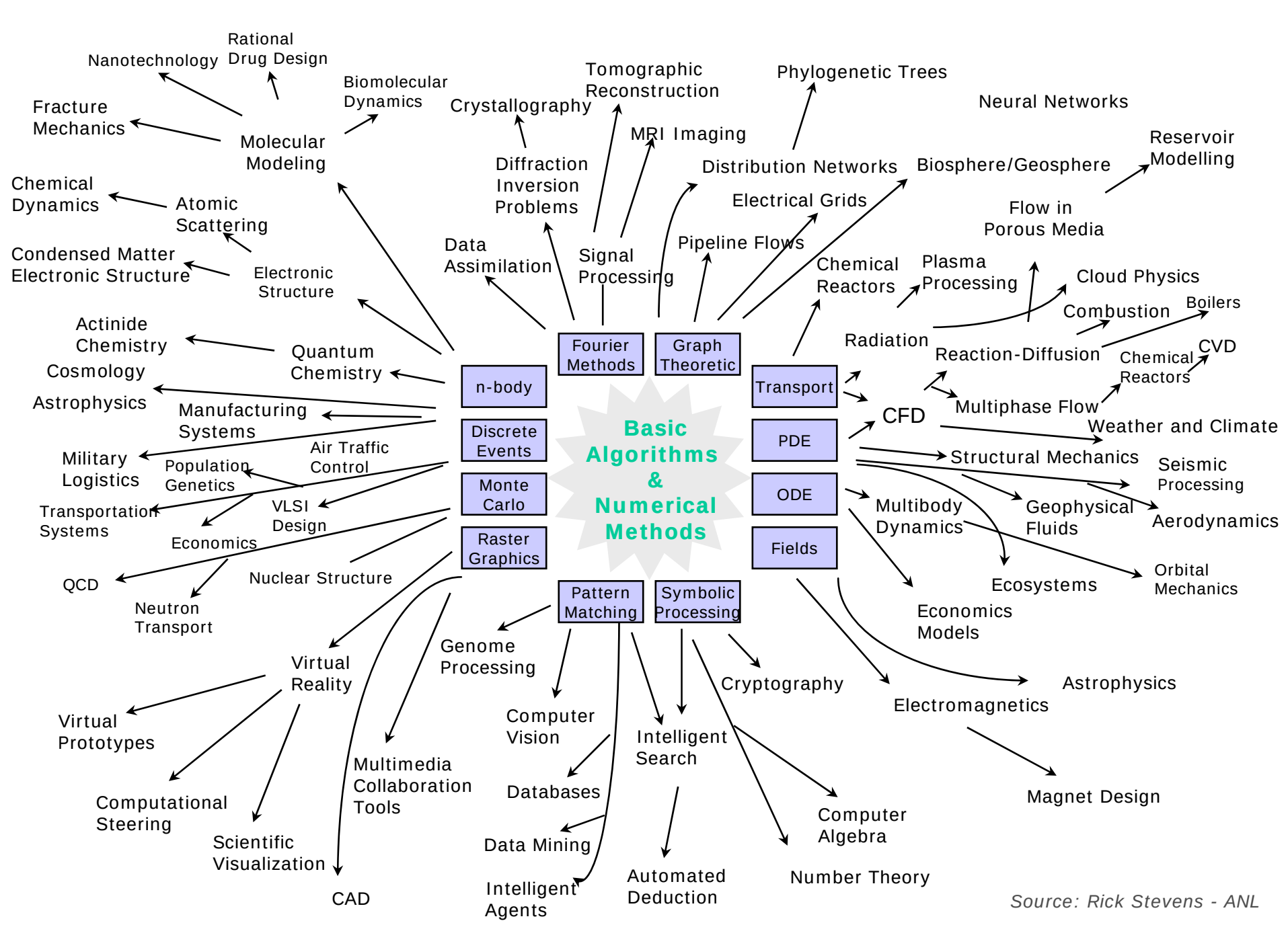
アプリケーションの実装



- 一般にアプリケーション毎に解析アルゴリズムは異なる
- これらの複数のアルゴリズムのアプリケーションをリアルタイムでハードウエアにマッピングする必要がある
- アルゴリズムごとに要求するコンピュータリソースはかなり異なる

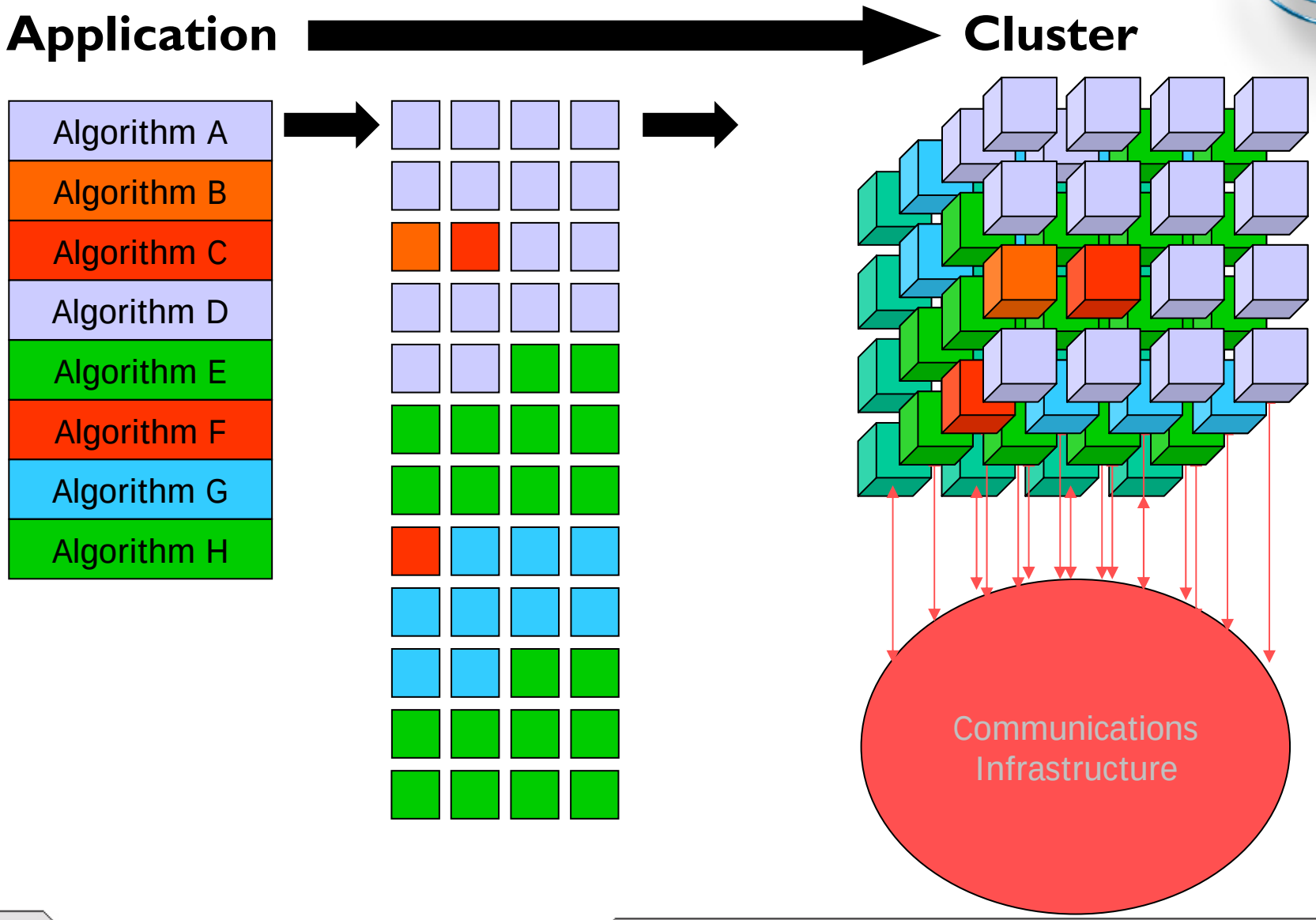
その他の技術動向

- Field Programmable Gate Arrays (FPGAs)
  - 非常に急速にその性能が向上
  - ただし、効率良くソフトウエア開発が可能なツール類の整備が不可欠
- ヘテロな計算機環境の提案
  - シングルシステムでの異なったプロセッサタイプを実装
    - ベクトルプロセッサ、スーパースカラー、FPGAなど
  - それらのプロセッサ要素を高速のインターコネクトで接続
  - 複数の物性、材料、現象の複合的な解析

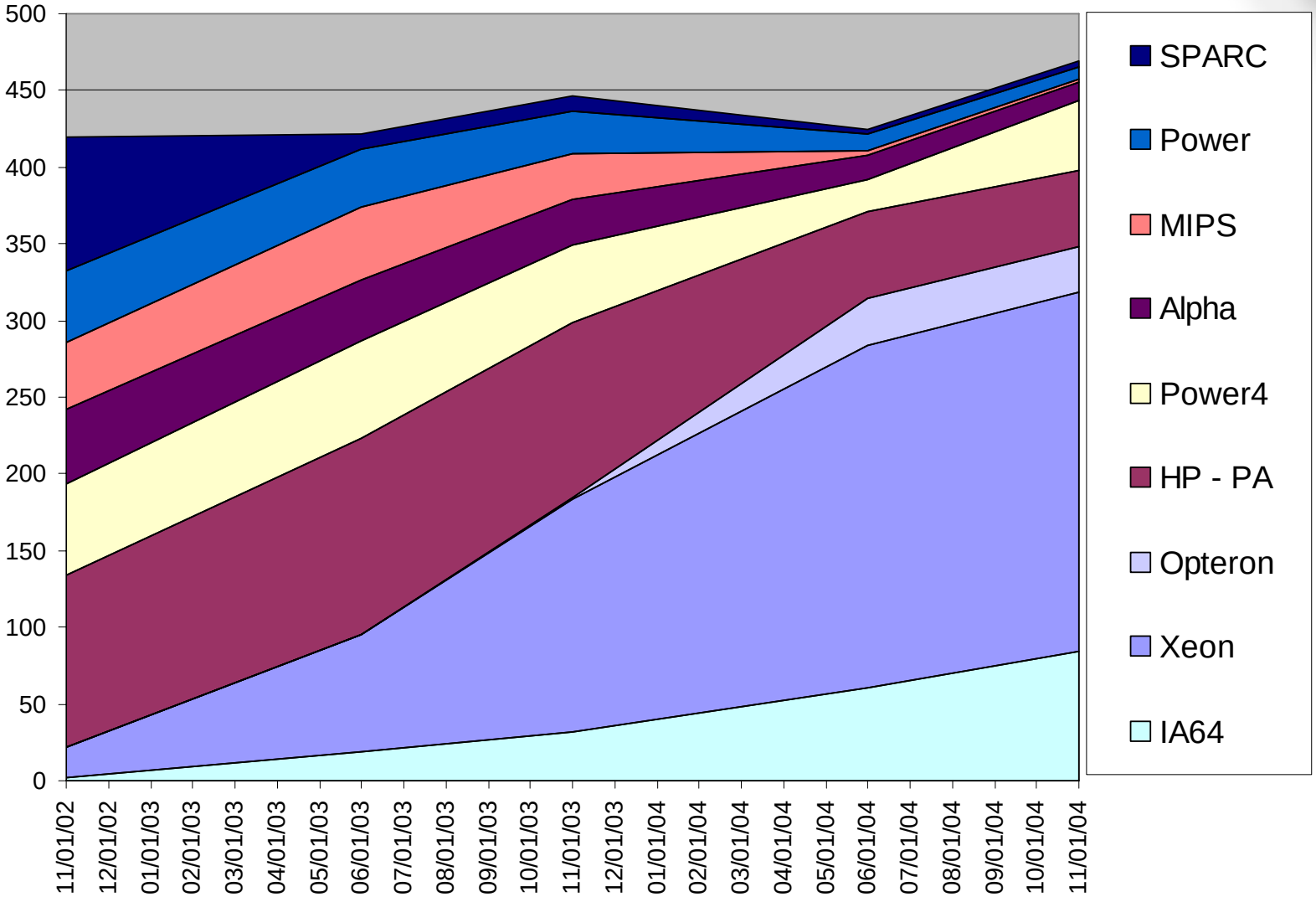


Source: Rick Stevens - ANL

アプリケーションのマッピング

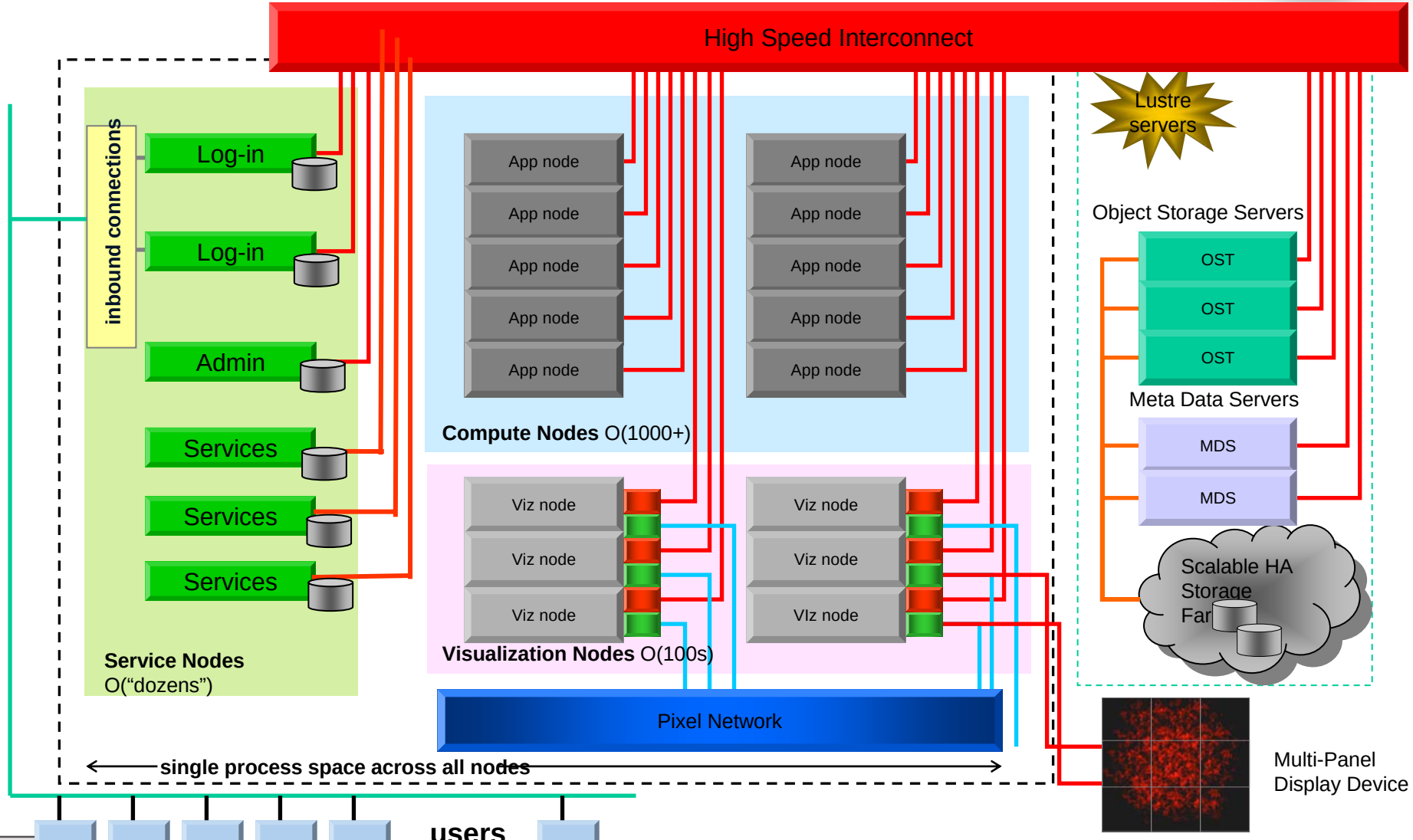


2002年6月から2004年11月でのプロセッサ別TOP500リストの変遷

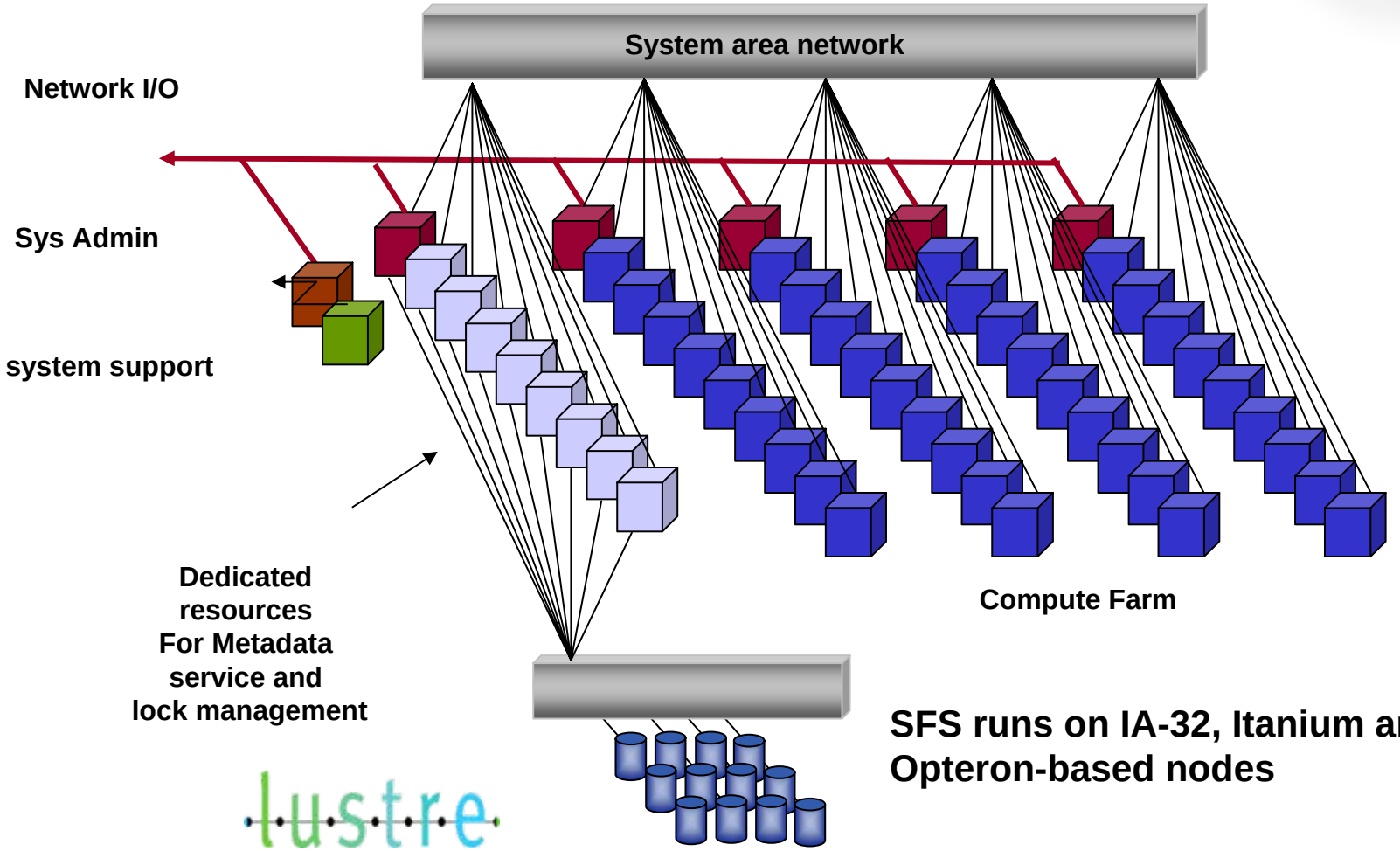


クラスタシステム

標準コンポーネントの活用　－　より高速なインターコネクトを採用

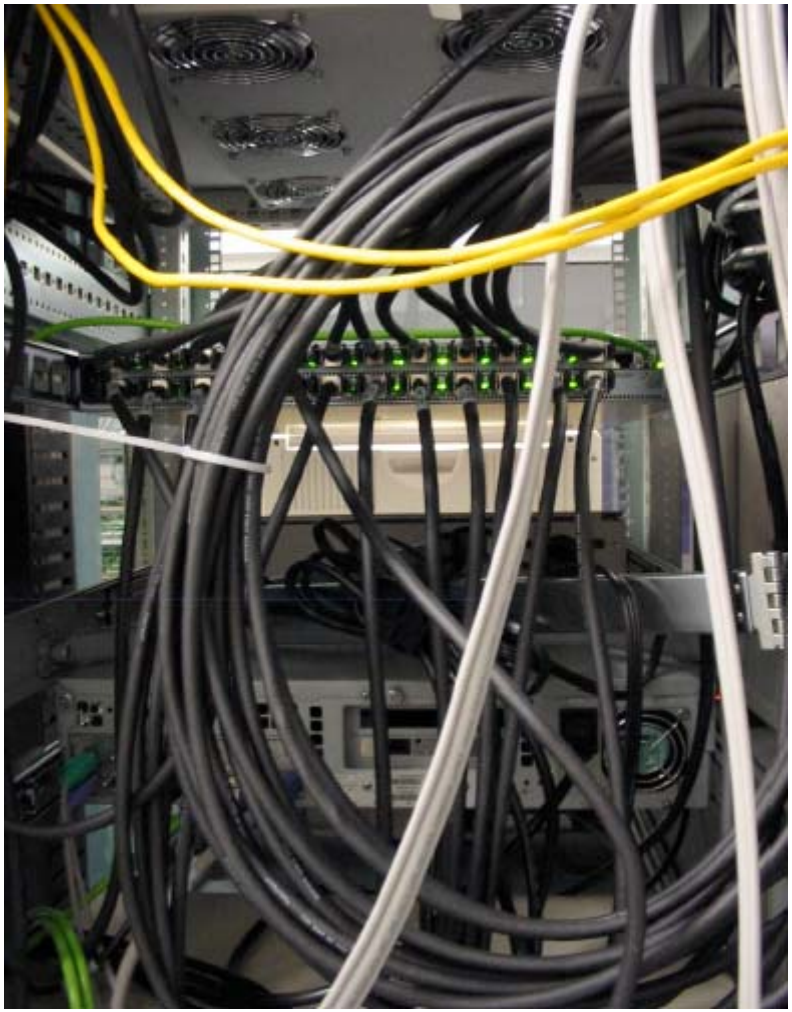


スケーラブルクラスタファイルシステム



SFS runs on IA-32, Itanium and Opteron-based nodes

InfiniBand:可用性・保守性の向上



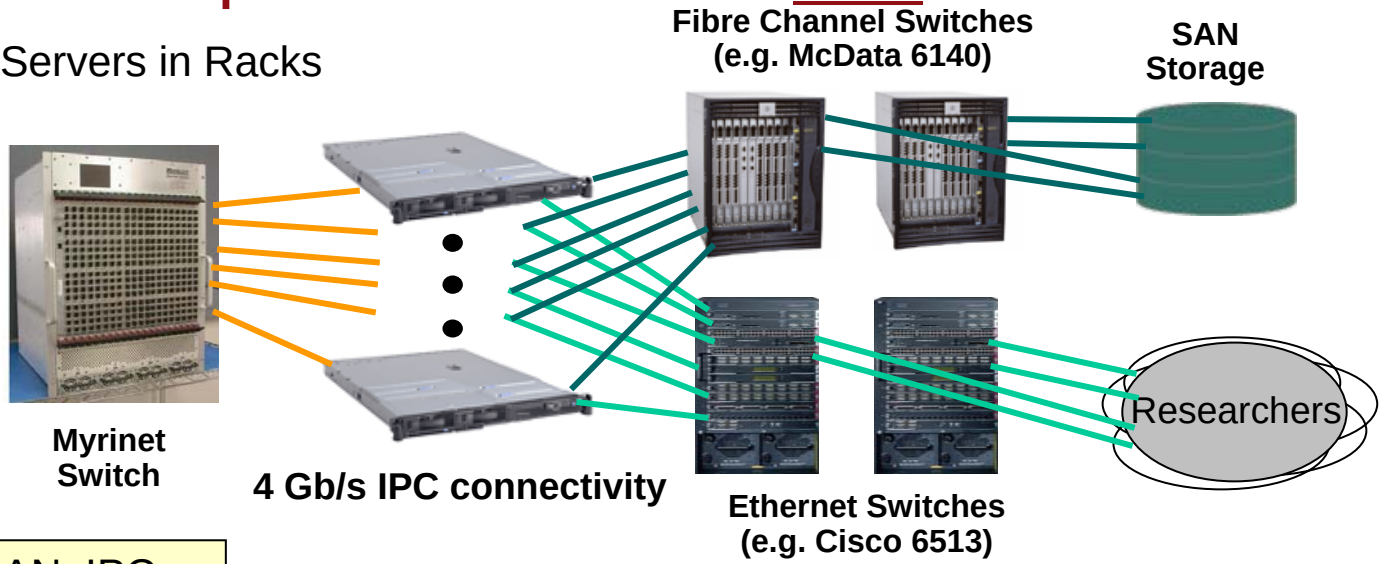
Before

After

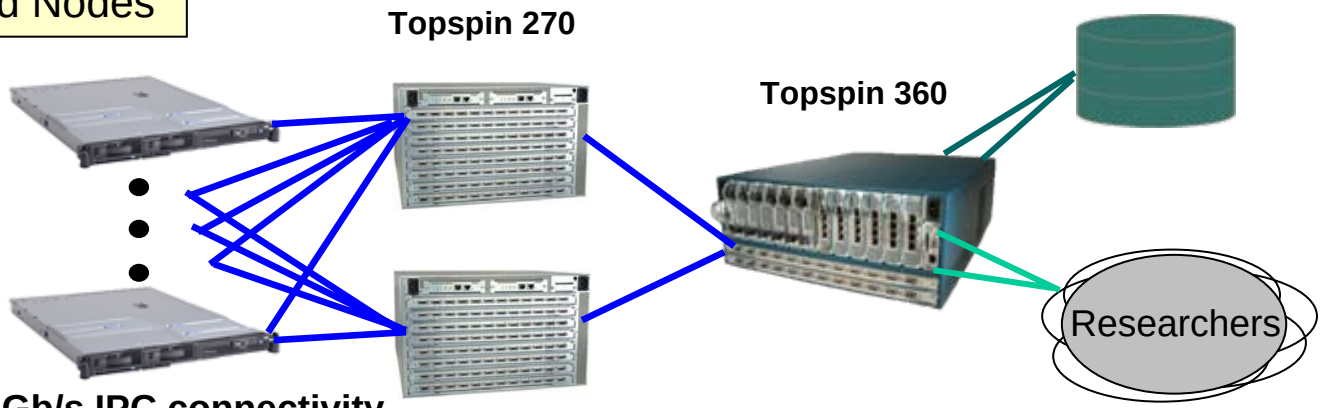
Fabric Consolidation (MFIO)

All compute nodes LAN and SAN attached

512 Servers in Racks

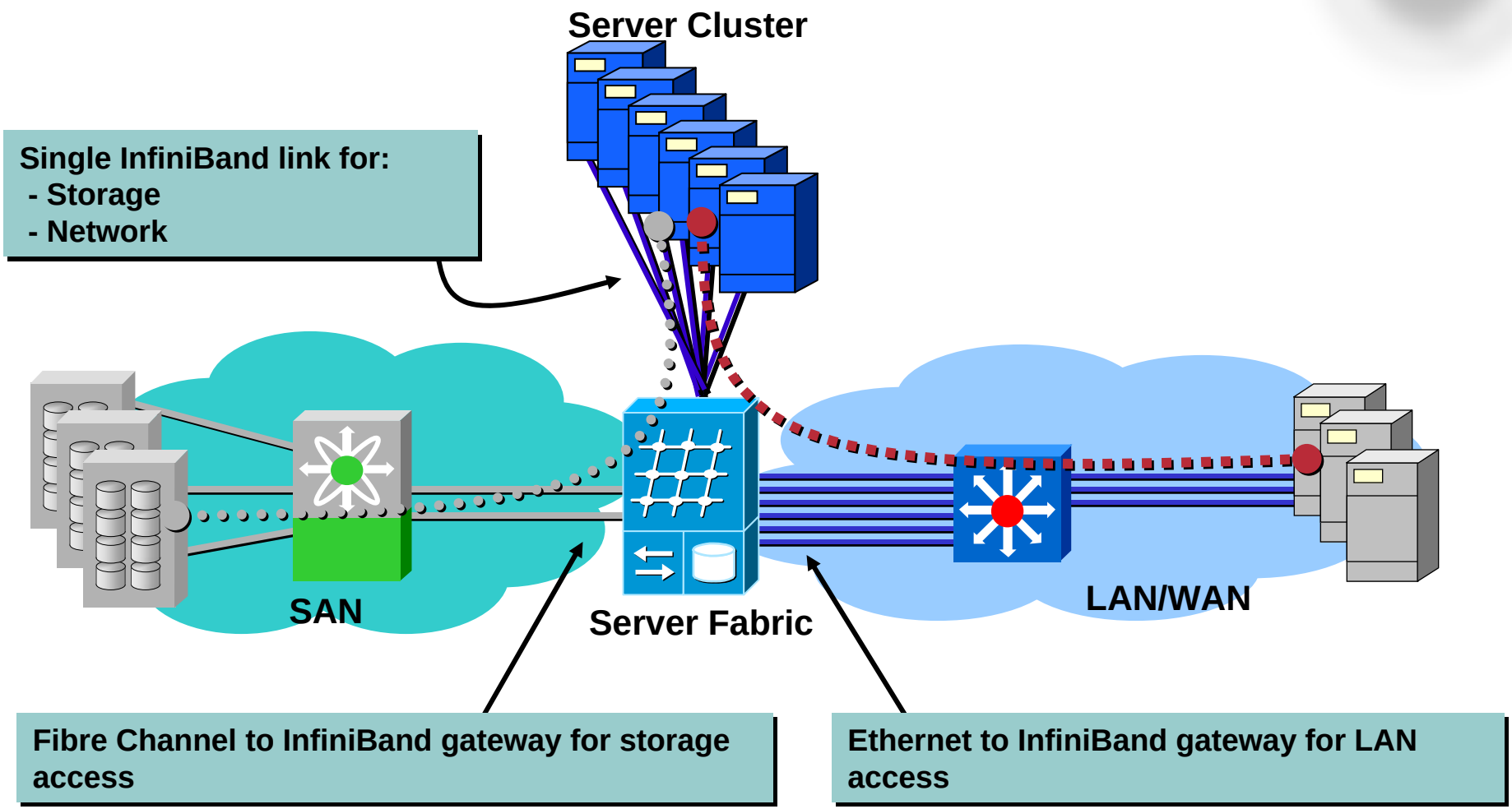


512 LAN, IPC and SAN-attached Nodes



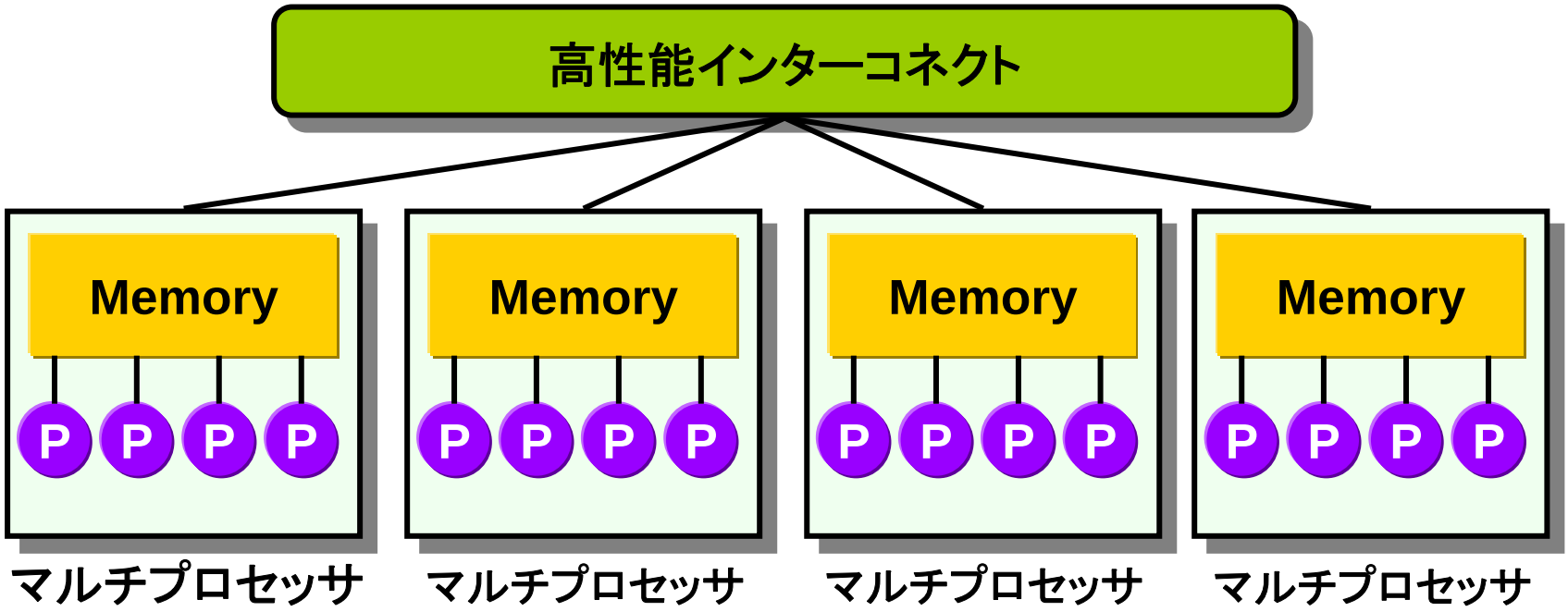
Ethernet and Fibre Channel Gateways

Unified “wire-once” fabric



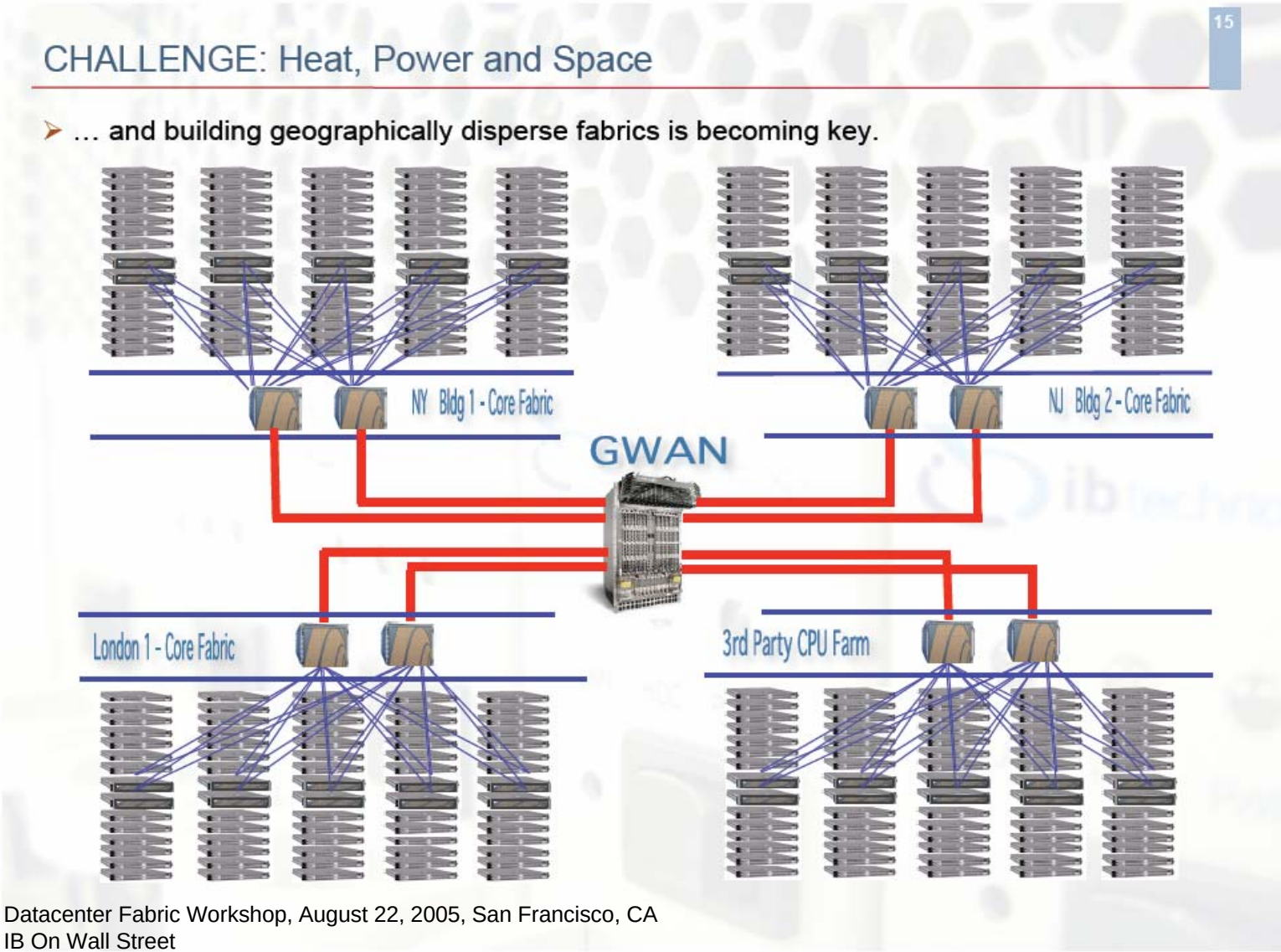
クラスタシステム

クラスタ内の各ノードは高速インターコネクト・テクノロジで接続されます。InfiniBandや PCI Expressテクノロジが登場する前は、独自規格に基づいた高性能で高価なテクノロジと、標準規格に基づいた低コストでやや性能の低いテクノロジのいずれかを選択する 必要がありました。コスト制約の厳しいクラスタの場合は、ネットワーク接続用の Ethernet テクノロジが広く利用されていますが、これは並列アプリケーションのようにノード間の緊密な連携が要求される環境ではボトルネックとなります。InfiniBand ベースのインターコネクトを導入すれば、このようなトレードオフは解消されます。



デュアルコアおよびマルチコア・プロセッサは1つのプロセッサの中に2つまたはそれ以上の完全な実行コアを搭載することによって、複数の処理を同時に実行 可能であり、このようなマルチコア・プロセッサを複数搭載したSMP (Symmetric Multiprocessing) 構成となり、高速のメモリアクセスとノード内でのマルチスレッドプログラミングが可能

CHALLENGE:Heat,Power and Space

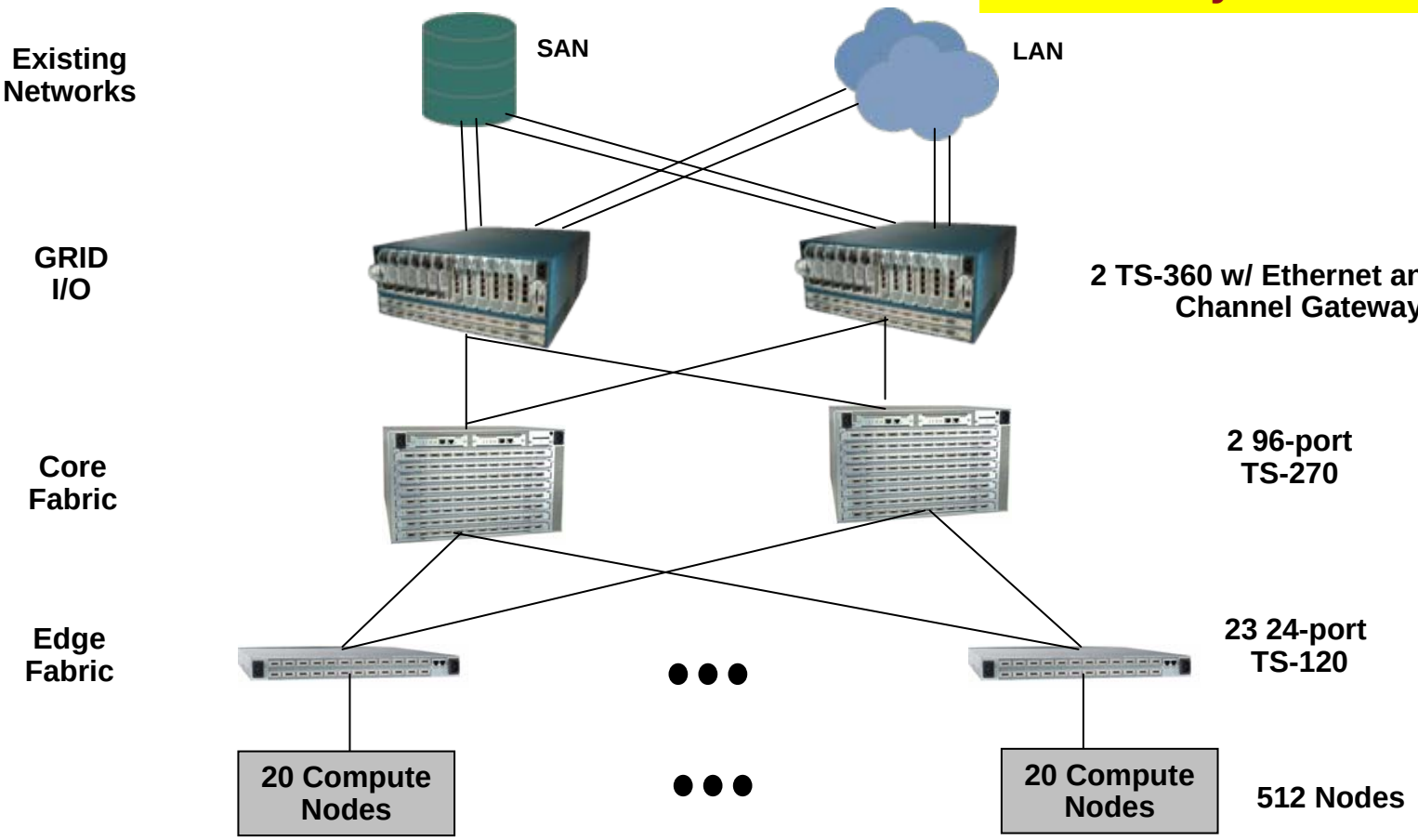


Datacenter Fabric Workshop, August 22, 2005, San Francisco, CA
IB On Wall Street
Speakers: Ty Panagoplos (JPMC), Peter Krey (JPMC)

Large Wall Street Bank

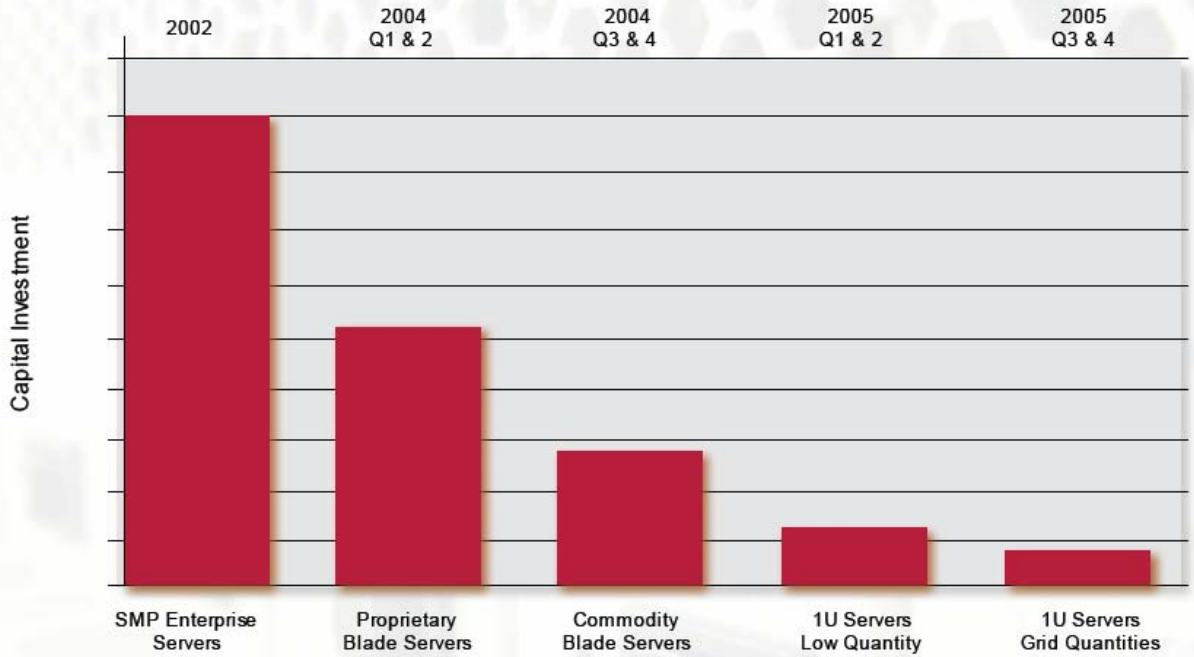
512 Node Commercial LINUX GRID

Fibre Channel and GigE connectivity built seamlessly into the cluster



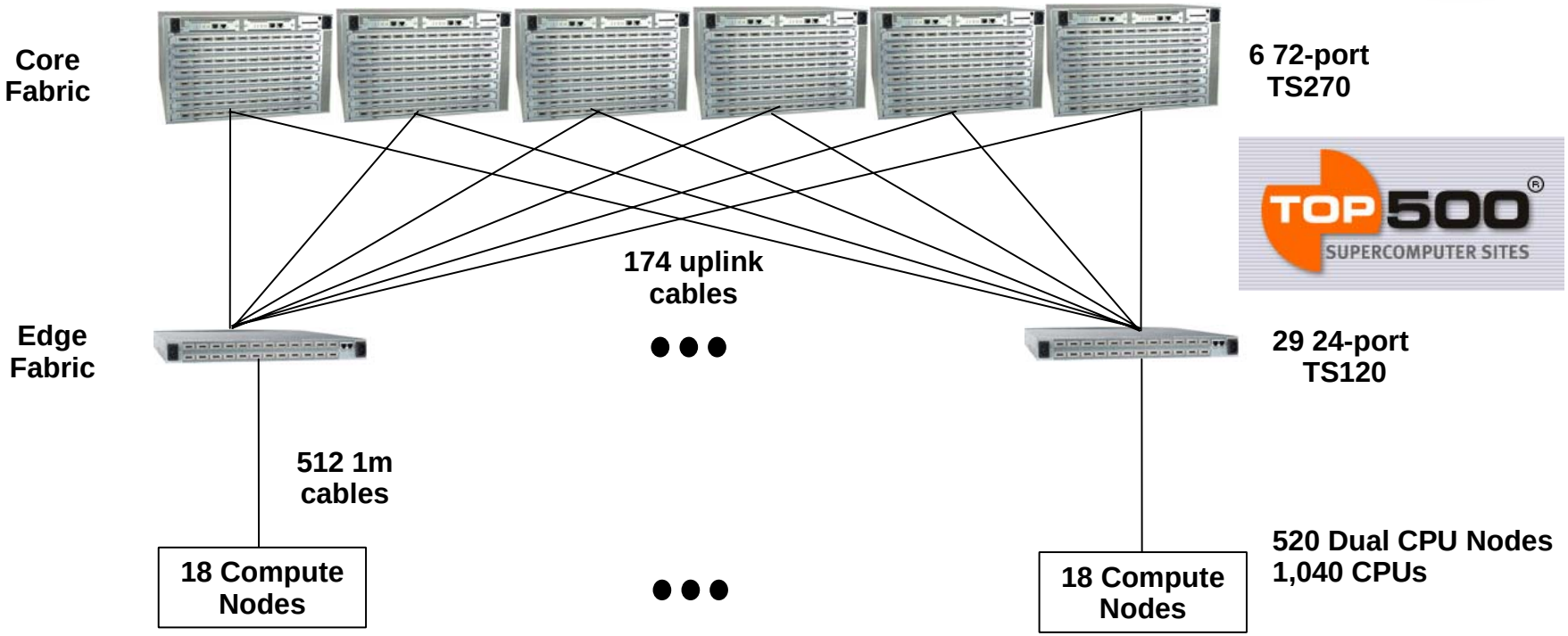
Grid Infrastructure Builds on Commodity Hardware

- Driving Down CPU/Hour Costs From Over $10 Down Below $0.50



OpenIB.org/Intel Developer Forum 2005
JPMorganChase

Tungsten 2: 520 Node Supercomputer



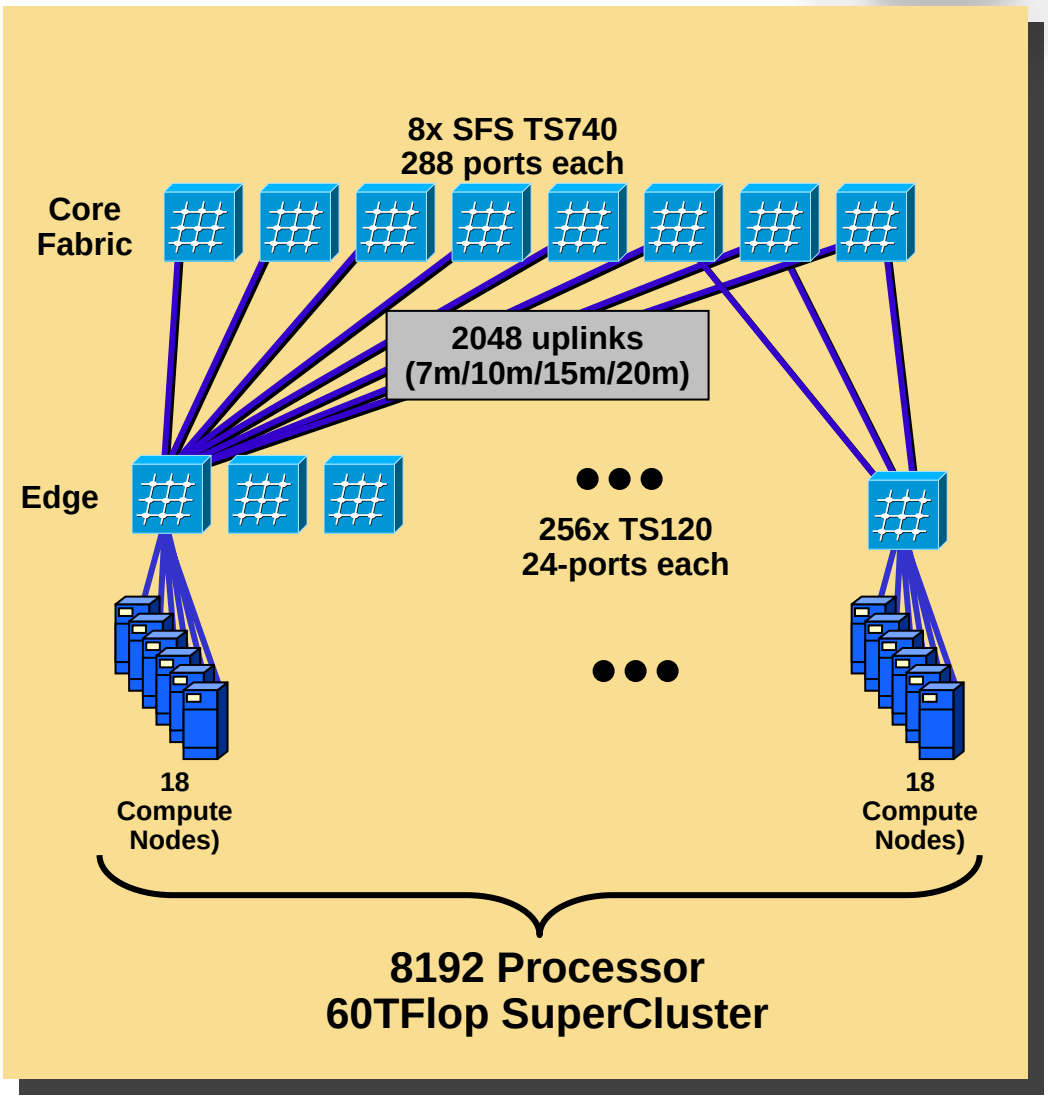
- Parallel MPI codes for commercial clients
- Point to point 5.2us MPI latency

Deployed: November 2004

Large Government Lab

Worlds Largest Commodity Server Cluster – 4096 nodes

- Application:
  - High Performance Super Computing Cluster
- Environment:
  - 4096 Dell Servers
  - 50% Blocking Ratio
  - 8 TS-740s
  - 256 TS-120s
- Benefits:
  - Compelling Price/Performance
  - Largest Cluster Ever Built (by approx. 2X)
  - Expected to be 2nd Largest Supercomputer in the world by node count



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（？）

社名、製品名などは、一般に各社の商標または登録商標です。無断での引用、転載を禁じます。

2005年10月