HPCシステムの歴史

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の歴史

The Phantom Menace

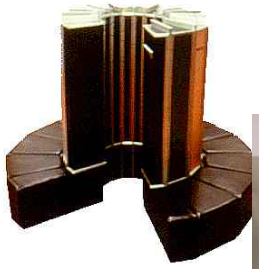
HPC システムズ


PN-07-12/13
PN-08-08/09
PN-09-14/15

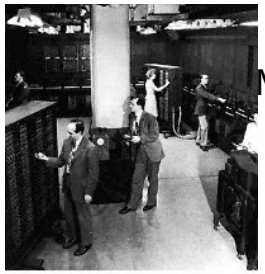

ASCI White – 2000
Lawrence Livermore National Laboratory

過去60年間の進化







1948 Manchr Baby

1949 Edsac

1959 IBM 7094

1976 Cray 1

1991 Intel Delta

1996 T3E

2003 Cray X1

2005 IBM BG/L

10^3 10^6 10^9 10^{12} 10^{15}

One OPS — KiloOPS — MegaOPS — GigaOPS — TeraOPS — PetaOPS

1943 Harvard Mark 1

1951 Pilot Ace

1964 CDC 6600

1982 Cray XMP

1988 Cray YMP

1997 ASCI Red

2001 Earth Simulator





Scalar to super scalar to vector to SMP to DMP to massively parallel to hybrid designs

様々なアーキテクチャのシステム

- Parallel Vector Processors (PVP)
  - NEC Earth Simulator, SX-6
  - Cray- 1, 2, XMP, YMP, C90, T90, X1
  - Fujitsu 5000 series
- Massively Parallel Processors (MPP)
  - Intel Touchstone Delta & Paragon
  - TMC CM-5
  - IBM SP-2 & 3, Blue Gene/Light
  - Cray T3D, T3E, Red Storm/Strider
- Distributed Shared Memory (DSM)
  - SGI Origin
  - HP Superdome
- Single Instruction stream Single Data stream (SIMD)
  - Goodyear MPP, MasPar 1 & 2, TMC CM-2
- Commodity Clusters
  - Beowulf-class PC/Linux clusters
  - Constellations
  - HP Compaq SC, Linux NetworX MCR

コンピュータシステムの分類

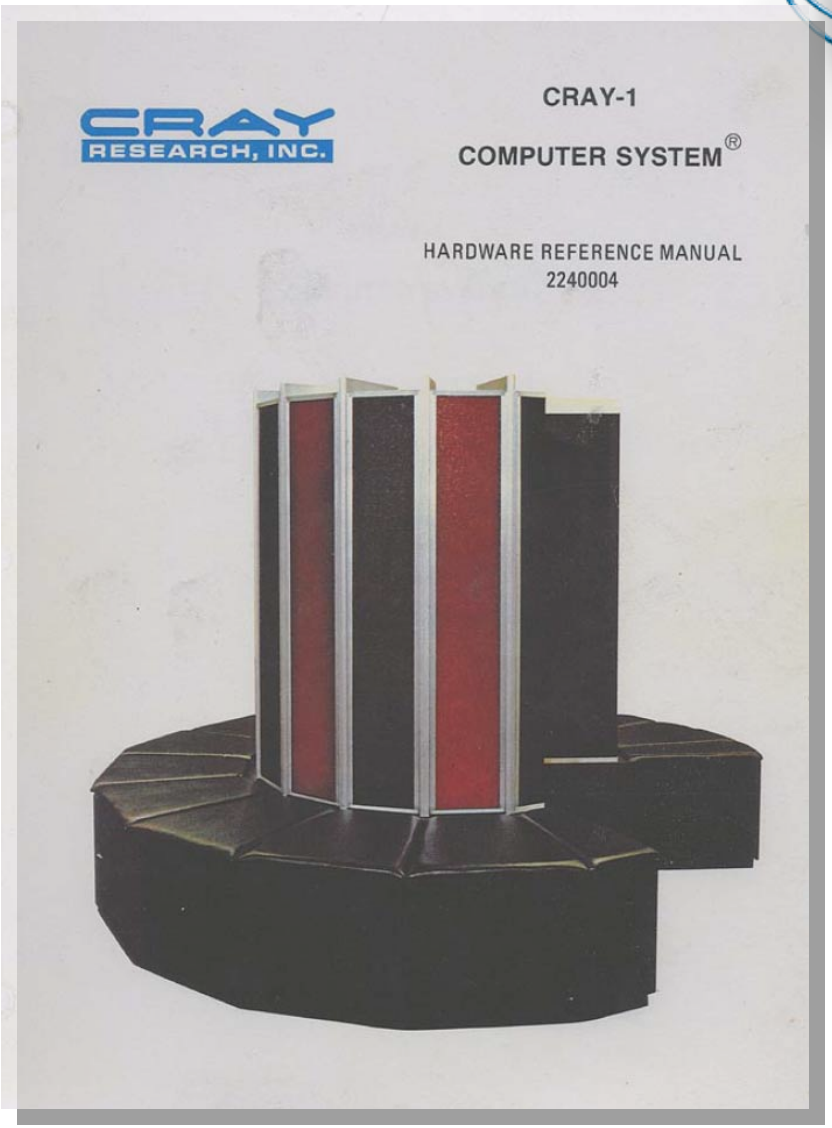
システムアーキテクチャ

	WAN/LAN	SAN	DSM	SM
vector	Supercomputer GRID	Fujitsu VPP	Hitachi SR	NEC SX Cray X…T
Scalar	Legion Condor Beowulf NT clusters	CRAY T3E IBM SP2 NOW	SGI DSM Clusters & SGI DSM	Mainframes Multicore WSs PCs

Clusters

温故知新

- 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

ベクトル計算機

- CRAYシステムの歴史

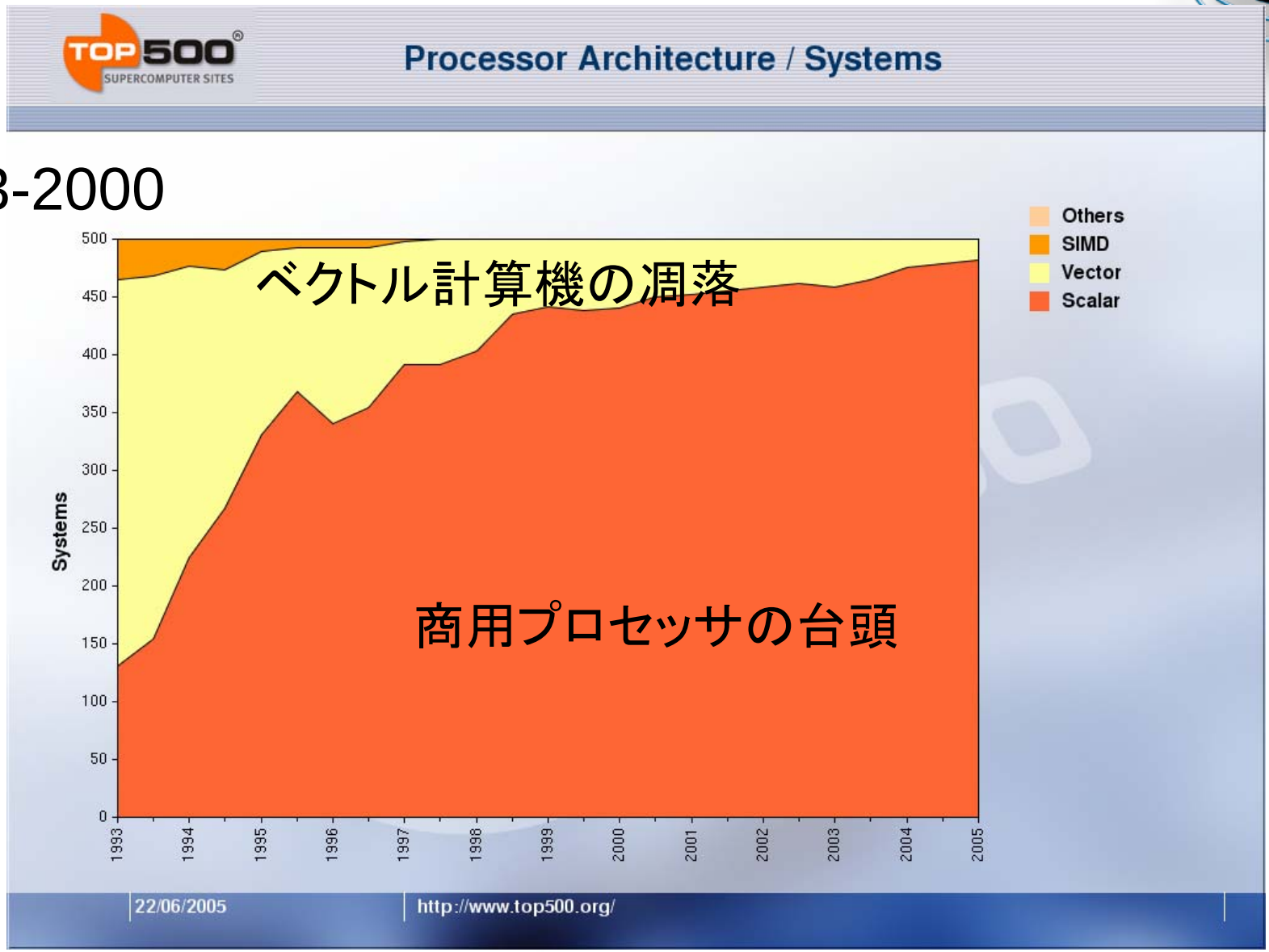
12年間で2倍の動作クロックの向上

- Los Alamos National Laboratory
- 1976 Cray-1　160 MFLOPS, 8 MB, 80 MHz
- 1982 Cray X-MP　500 MFLOPS, 118 MHz
- 1985 Cray-2　1900 MFLOPS, 2048 MB, 244 MHz
- 1988 Cray Y-MP　x*333 MFLOPS, 167 MHz

栄枯盛衰

- ACRI
- Alliant
- American Supercomputer
- Ametek
- Applied Dynamics
- Astronautics
- BBN
- CDC
- Cogent
- Convex > HP
- Cray Computer
- Cray Research > SGI > Cray
- Culler-Harris
- Culler Scientific
- Cydrome
- Dana/Ardent/Stellar/Stardent
- Denelcor
- Encore
- Elexsi
- ETA Systems
- Evans and Sutherland Computer
- Exa
- Flexible
- Floating Point Systems
- Galaxy YH-1
- Goodyear Aerospace MPP
- Gould NPL
- Guiltech
- Intel Scientific Computers
- International Parallel Machines
- Kendall Square Research
- Key Computer Laboratories searching again
- MasPar
- Meiko
- Multiflow
- Myrias
- Numerix
- Pixar
- Parsytec
- nCube
- Prisma
- Pyramid
- Ridge
- Saxpy
- Scientific Computer Systems (SCS)
- Soviet Supercomputers
- Supertek
- Supercomputer Systems
- Suprenum
- Tera > Cray Company
- Thinking Machines
- Vitesse Electronics
- Wavetracer

The Pahntom Menace

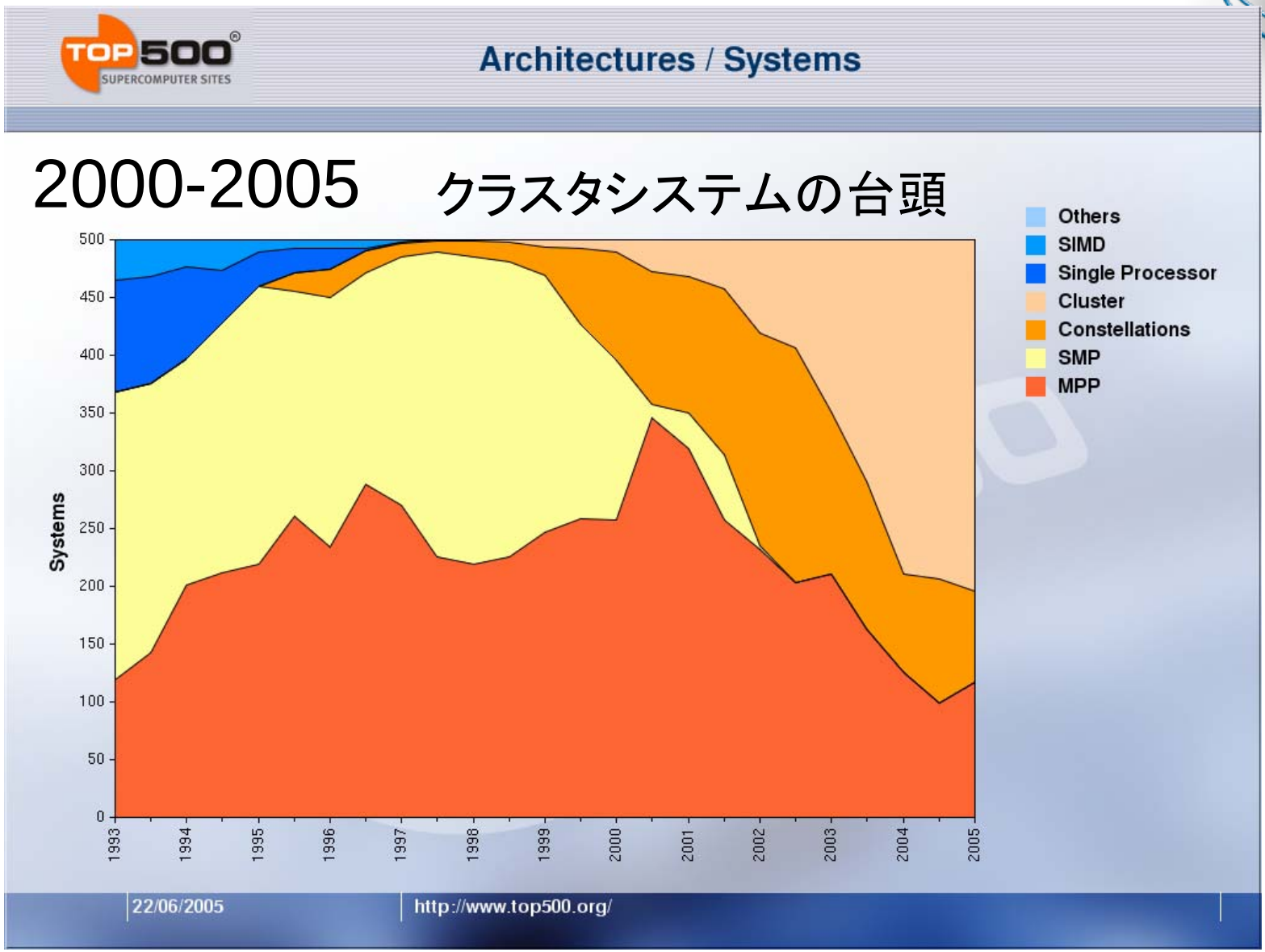


1993-2000

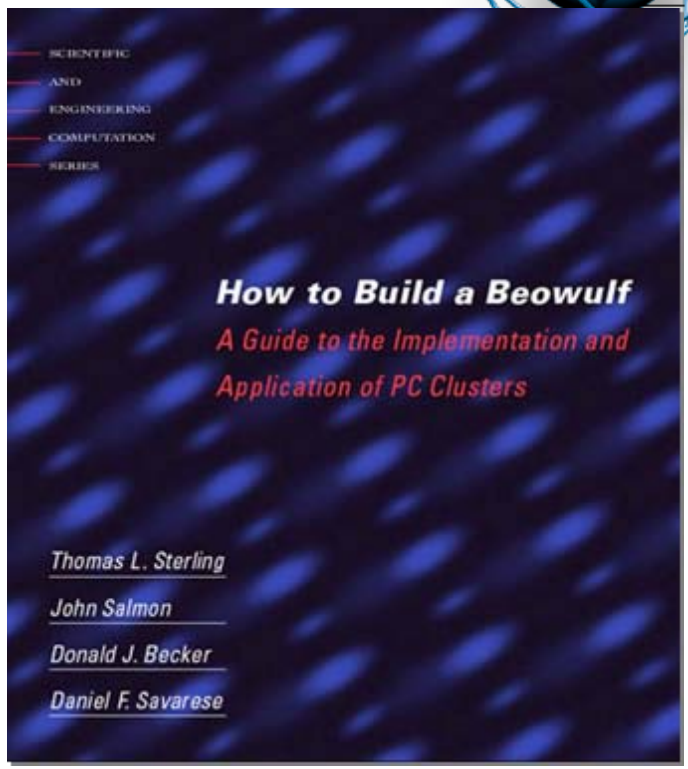
HPCの歴史

Attack of the Clones

Attack of the Clones



Beowulf プロジェクト



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

ベクトル計算機の逆襲

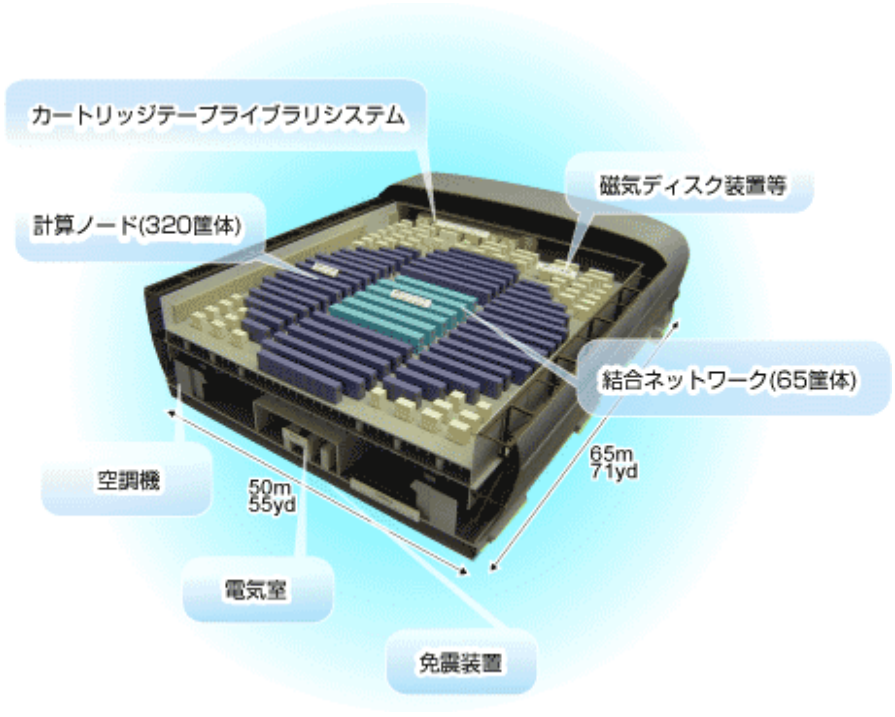
The Empire Strikes Back

Sputnik: October 4, 1957

The Empire Strikes Back

- 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

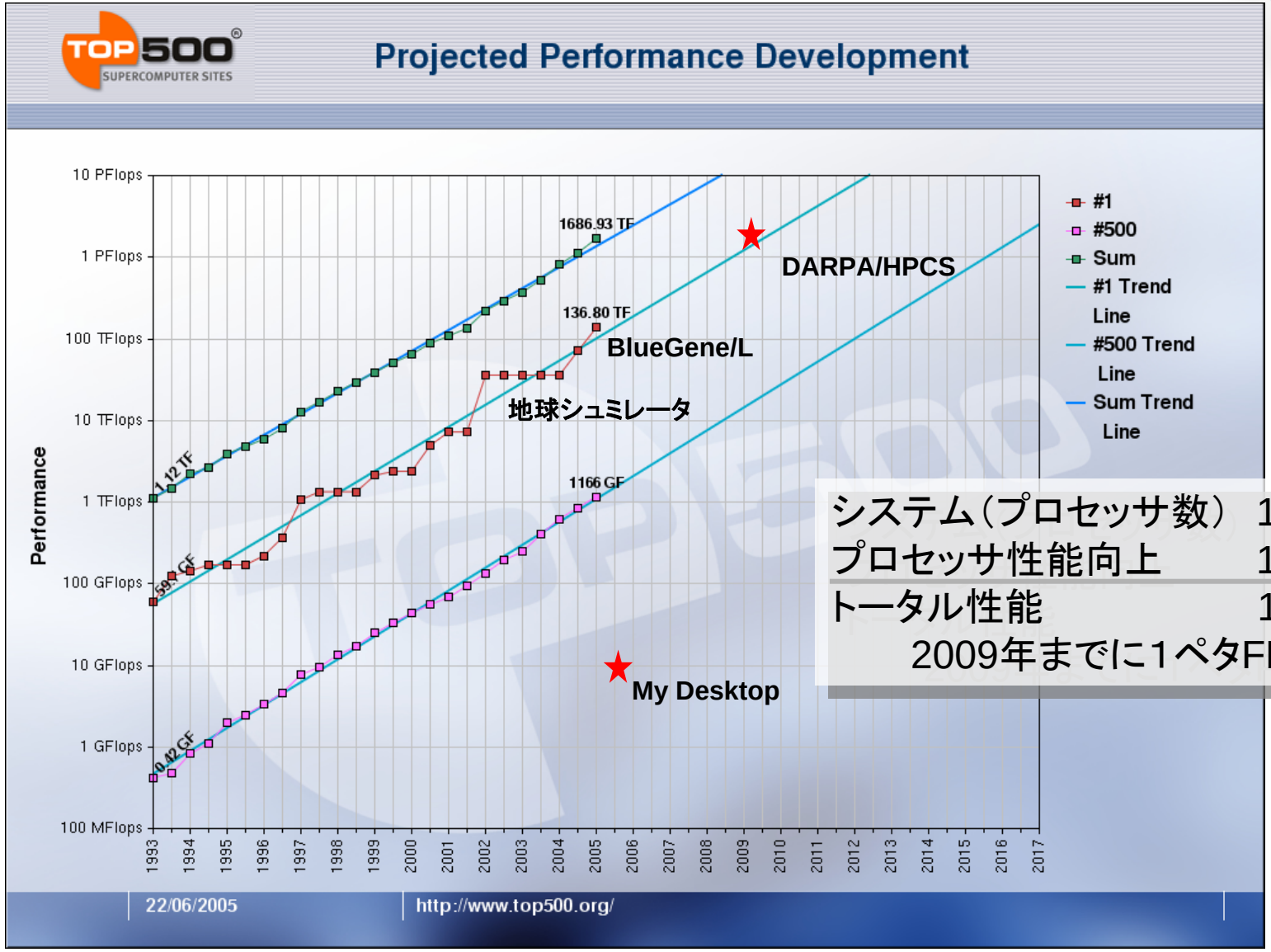
- 5,120 (640 8-way nodes) 500 MHz NEC
- 8 GFLOPS per CPU (41 TFLOPS total)
- 2 GB Memory per CPU (10 TB total)
- 20 kVA power consumption per node



今後の発展と課題

Revenge of the sith

HPC 1993-2005

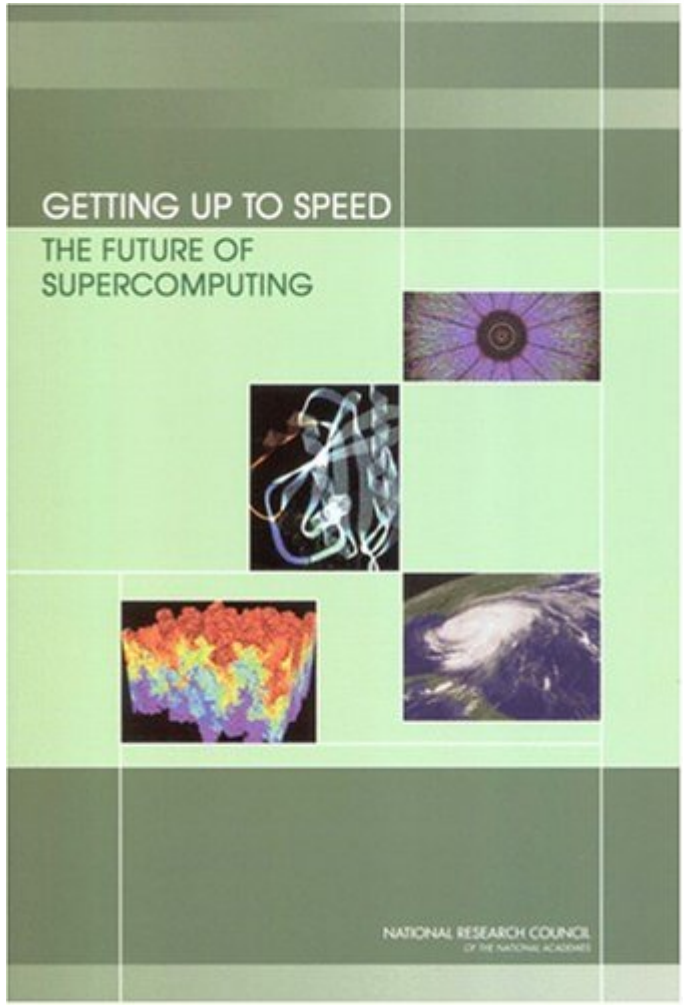


システム（プロセッサ数）	1.3倍/年
プロセッサ性能向上	1.4倍/年
トータル性能	1.8倍/年

2009年までに1ペタFLOPS

USでの研究調査

- USでの多くの調査・研究とプロジェクト
  - (2002) DARPA:HPCS
  - (2003) DoD:IHEC
  - (2004) NCO/NITRD: HECRTF
  - (2004) DOE: HEC Revitalization Act
- 多くのプロジェクトと調査は、HPCの活性化とその再生のために米国政府が中心となって推進
- IBM BlueGeneやNASA Columbia プロジェクトなどでの成果
- 国家プロジェクトとしての ‘スーパーコンピューティング’



High Productivity Computing Systems

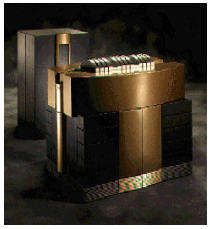
Goal:

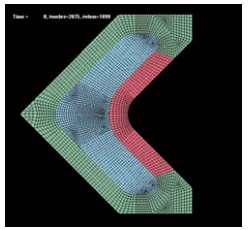
Provide a new generation of economically viable high productivity computing systems for the national security and industrial user community (2009 – 2010)

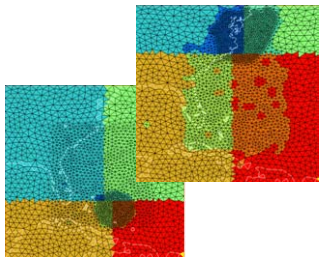
Impact:

- **Performance** (time-to-solution): speedup critical national security applications by a factor of 10X to 40X
- **Programmability** (idea-to-first-solution): reduce cost and time of developing application solutions
- **Portability** (transparency): insulate research and operational application software from system
- **Robustness** (reliability): apply all known techniques to **protect against outside attacks**, hardware faults, & programming errors

HPCS Program Focus Areas







Applications:

- Intelligence/surveillance, reconnaissance, cryptanalysis, weapons analysis, airborne contaminant modeling and biotechnology

Fill the Critical Technology and Capability Gap

Today (late 80's HPC technology).....*to*.....Future (Quantum/Bio Computing)

商用プロセッサの性能予測

- メモリ階層の問題が深刻化

	Annual increase	Typical value in 2005	Typical value in 2010	Typical value in 2020
Single-chip floating-point performance	59%	4 GFLOP/s	32 GFLOP/s	3300 GFLOP/s
Front-side bus bandwidth	23%	1 GWord/s = 0.25 word/flop	3.5 GWord/s = 0.11 word/flop	27 GWord/s = 0.008 word/flop
DRAM latency	(5.5%)	70 ns = 280 FP ops = 70 loads	50 ns = 1600 FP ops = 170 loads	28 ns = 94,000 FP ops = 780 loads

Source: *Getting Up to Speed: The Future of Supercomputing*, National Research Council, 222 pages, 2004, National Academies Press, Washington DC, ISBN 0-309-09502-6.

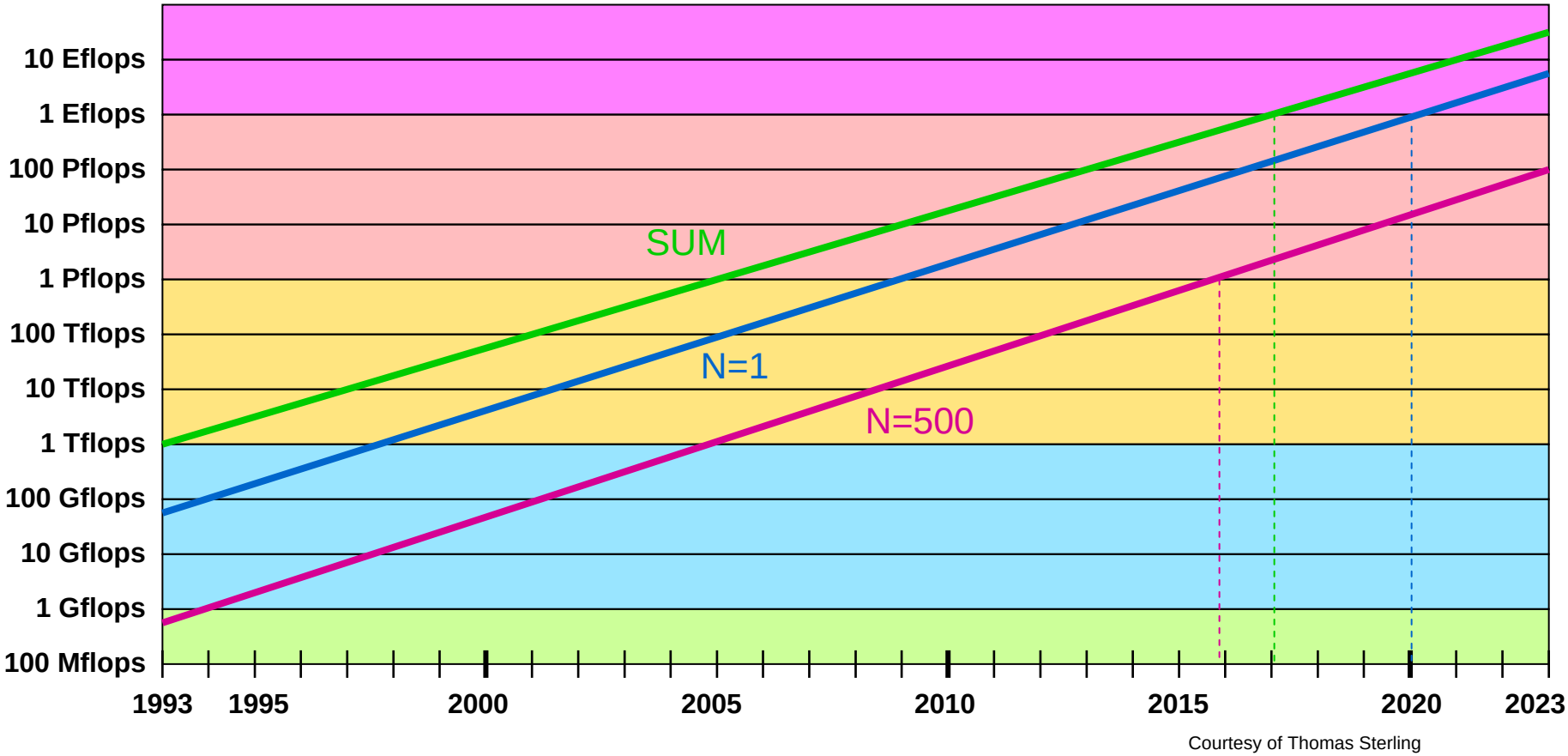
並列計算機予測

	Annual increase(%)	Typical value in 2005	Typical value in 2010	Typical value in 2020
No. of processors	20	4,000	12,000	74,000
General bandwidth (Mword/sec)	26	65（=0.03 word/flops)	260（=0.008 word/flops)	2,600（=0.008 word/flops)
General latency (nsec)	(28)	2,000 (=4,000 flops)	280 (=9,000 flops)	200 (=670,000 flops)
MPI bandwidth (Mwork/sec)	26	65	260	2,600
MPI latency (nsec)	(28)	3,000	420	300

Source: *Getting Up to Speed: The Future of Supercomputing*, National Research Council, 222 pages, 2004, National Academies Press, Washington DC, ISBN 0-309-09502-6.

1 ペタFLOPS時代へ

TOP500リストでの性能を現在のペースを維持するという仮定で予測した場合の性能を示したものです。N=1は、最速システムの性能を、SUMでは、TOP500リストの全システムの総計を示しています。



Courtesy of Thomas Sterling

システムの性能評価の問題

General purpose architecture capable of:

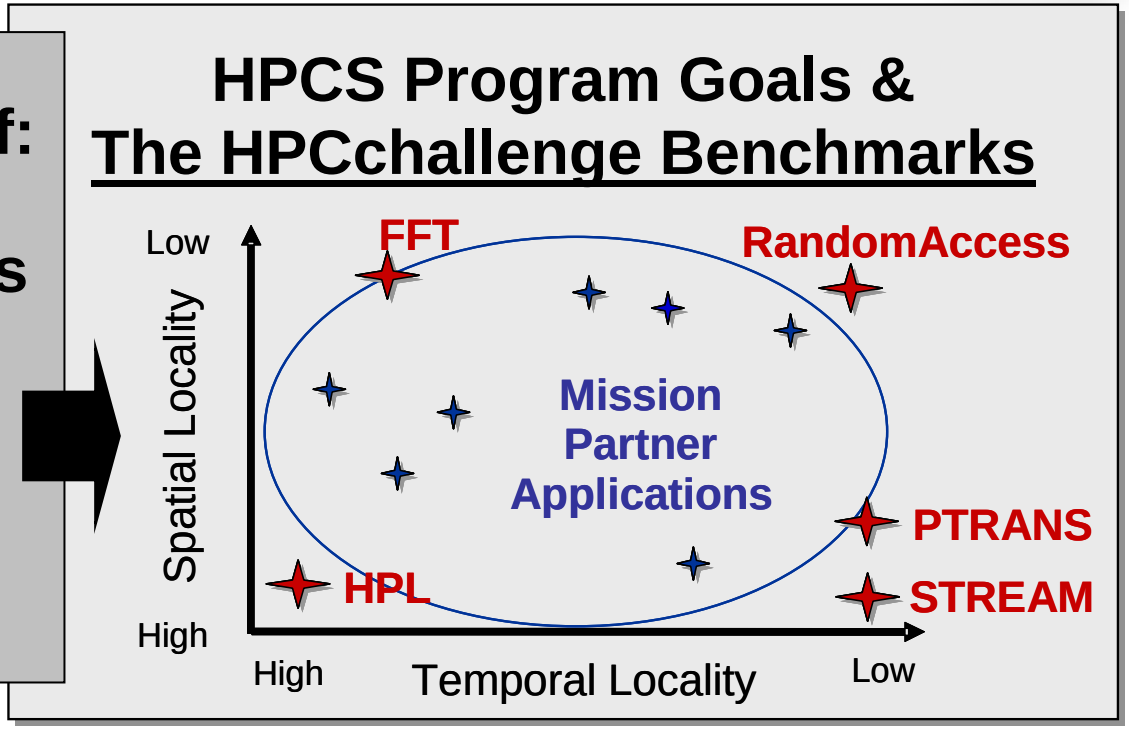
Subsystem Performance Indicators

1)2+ PF/s LINPACK

2)6.5 PB/sec data STREAM bandwidth

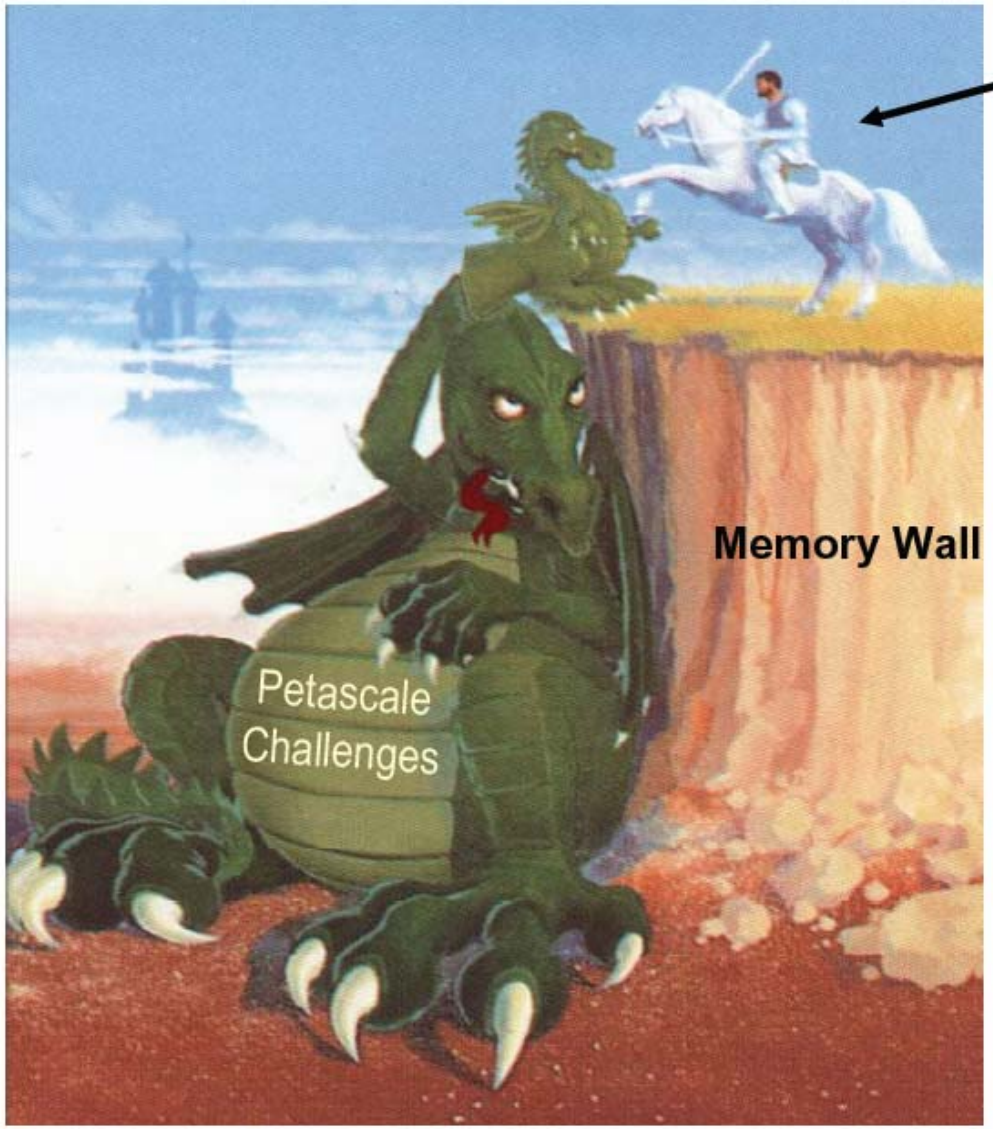
3)3.2 PB/sec bisection bandwidth

4)64,000 GUPS



現在の一般的な標準ベンチマークは、アプリケーション性能の評価に有用なのか？

Petascaleへの最大の課題

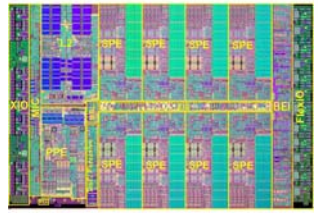


HPCシステムユーザ

- Terascaleの問題の克服は、可能でも、Petascaleの問題の克服には、‘Memory Wall’の問題が立ちはだかる
- より多くのプロセッサを利用して、生産性を現時点よりも高める方法の模索も必要

並列処理の規模的拡大

Japanese ES

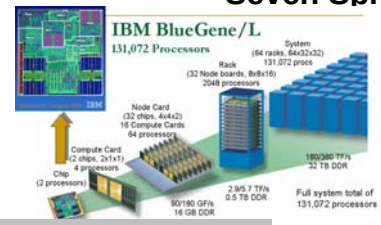


Sony, Toshiba, IBM Cell
256 Gflop/s (26 Gflop/s DP)

64 bit architectures

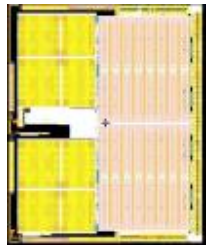
2000 2001 2002 2003 2004 2005

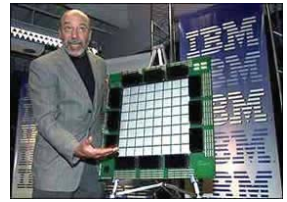
Seven Springs XVI

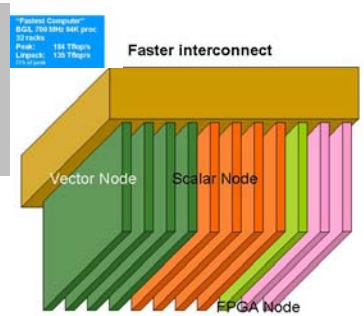


IBM BG/L
131,072 processors
So far 64K procs
135 Tflop/s Linpack

Multi-core chips







Tightly-Coupled Heterogeneous
System 2006?

hypre high performance preconditioners

Hypre

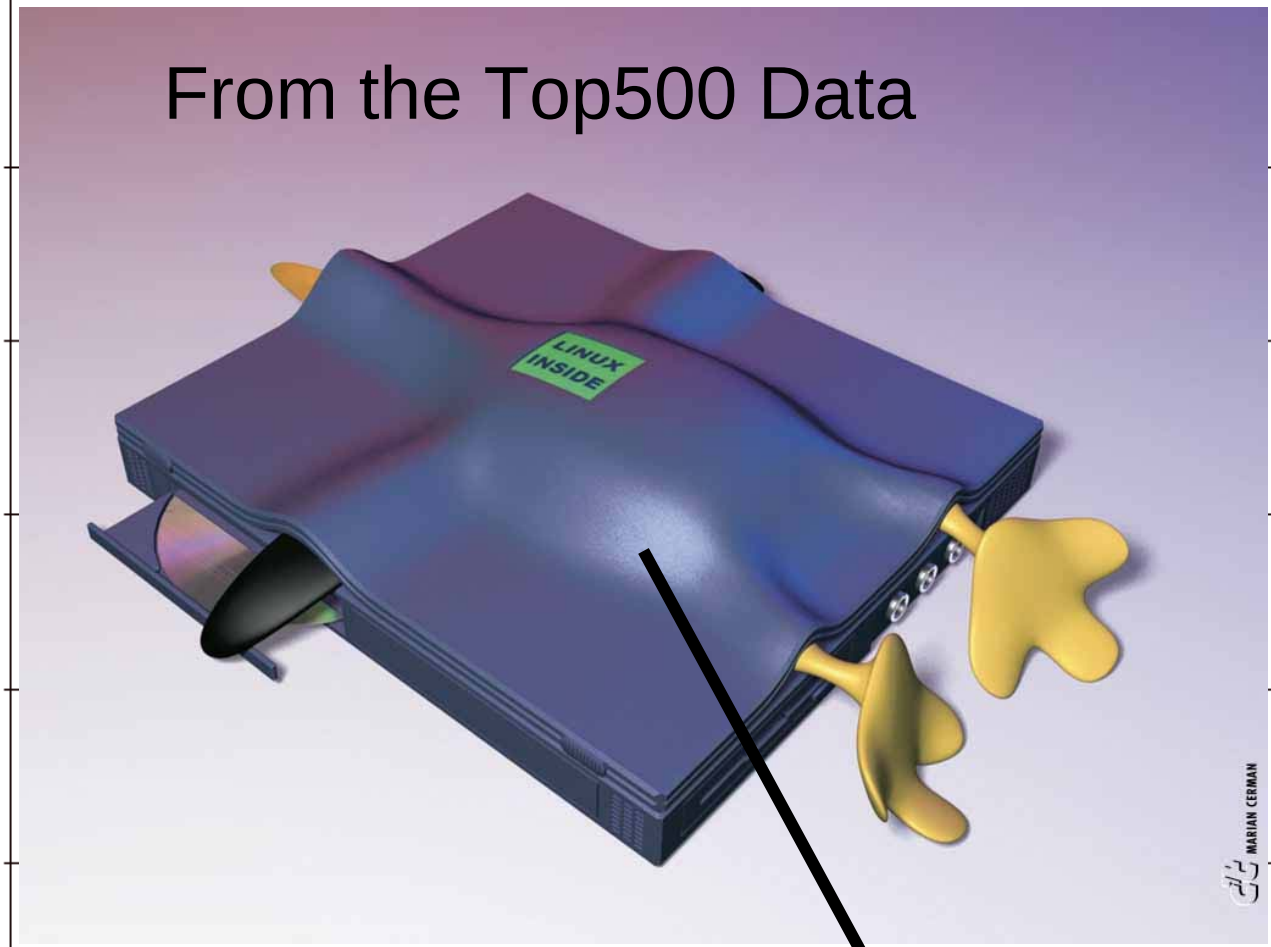
DARPA HPCS

DARPA HPCS
Program begins

今後の発展と課題

A New Hope

Linux Adoption in HPC



Jun-97 Nov-97 Jun-98 Nov-98 Jun-99 Nov-99 Jun-00 Nov-00 Jun-01 Nov-01 Jun-02 Nov-02 Jun-03 Nov-03 Jun-04

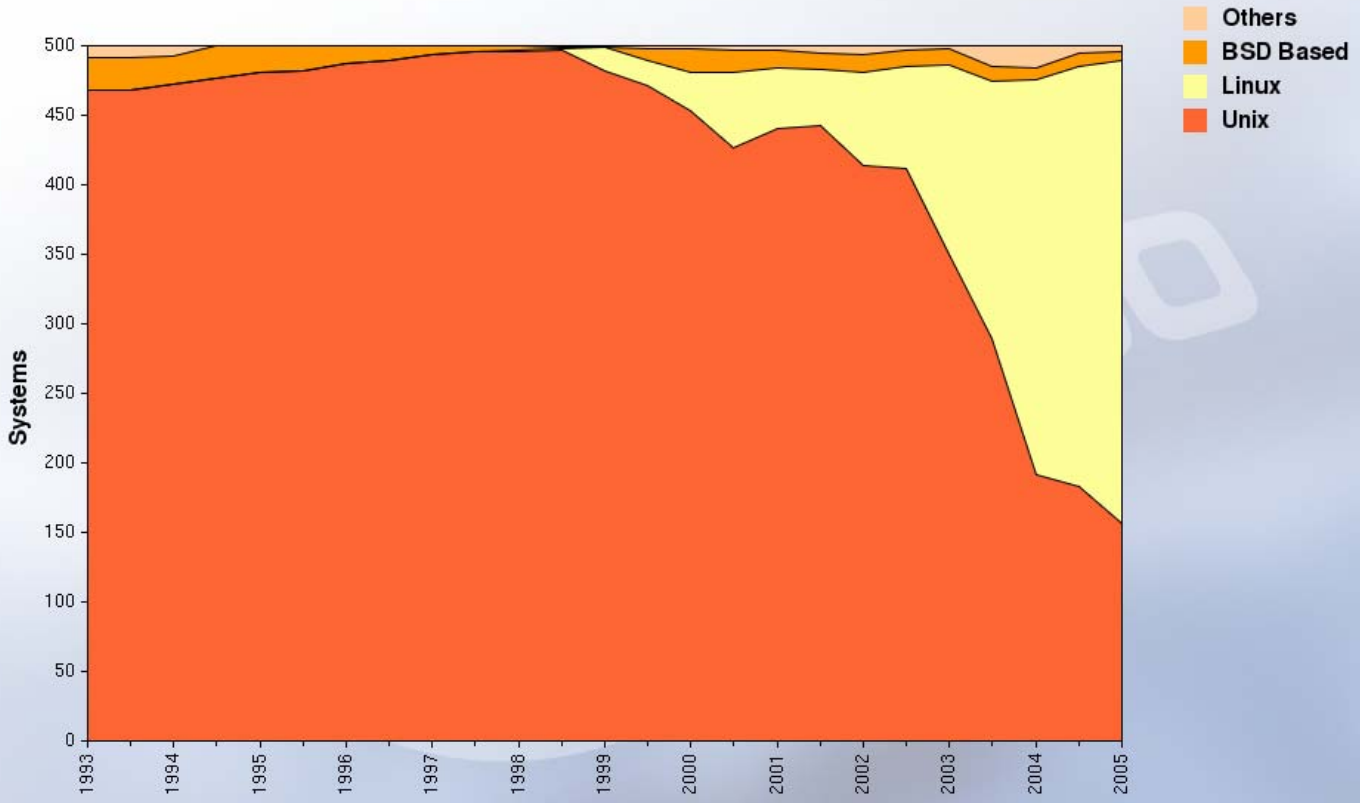
Percent

70 60 50 40 30 20 10 0

Top500 Extreme Linux Report Jun04-A
Pete Beckman <beckman@mcs.anl.gov>

%Linux Machine %Linux Performance

Operating System / Systems



22/06/2005 | http://www.top500.org/

ペタスケールシステムの構築

- 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
  - 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

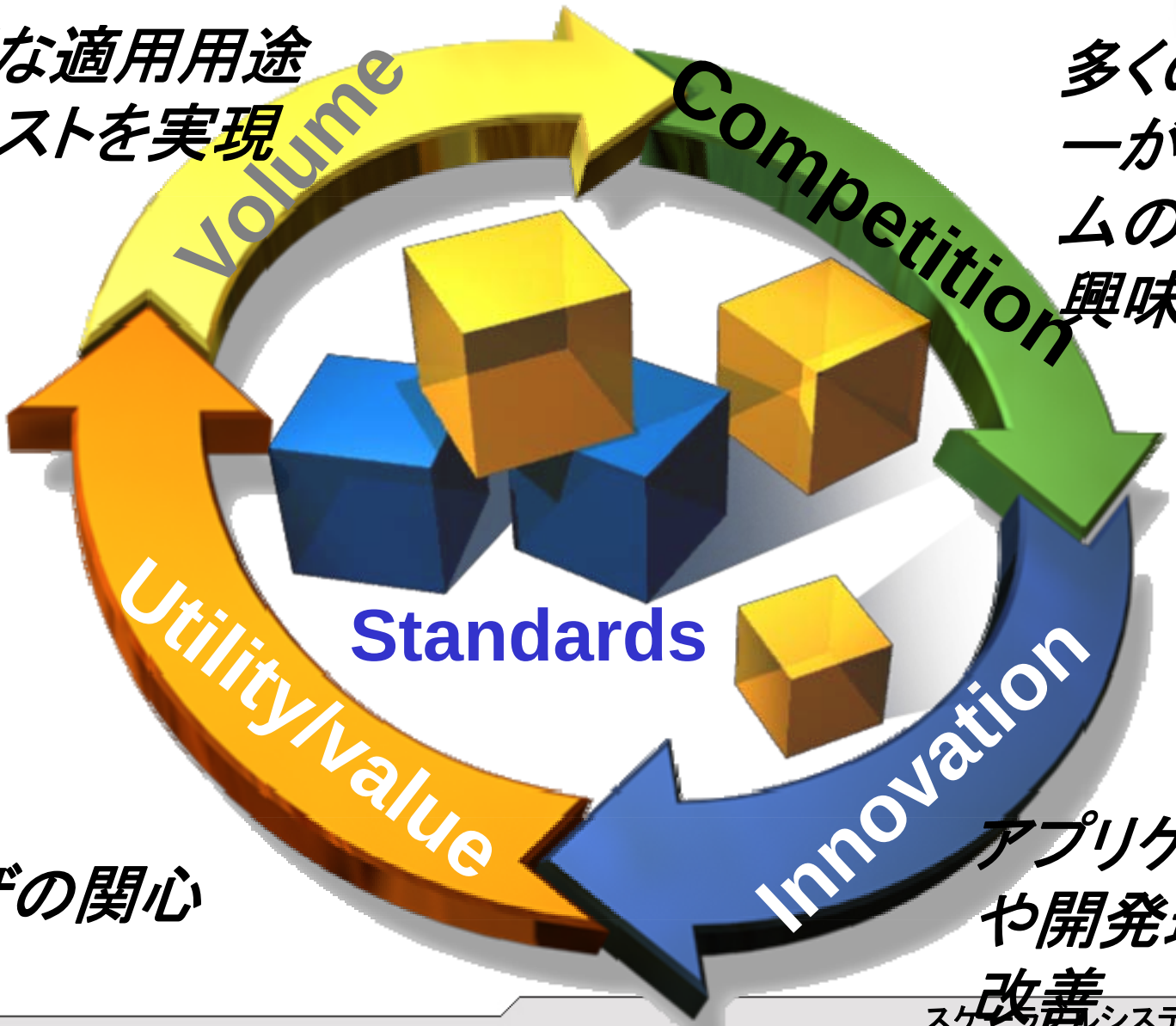
IBM Blue Gene システム

- LLNL BG/L
  - 360 teraflops
  - 64 racks
    - 65,536 nodes
    - 131,072 processors
- Node
  - Two 2.8 Gflops processors
    - System-on-a-Chip design
    - 700 MHz
    - Two fused multiply-adds per cycle
  - Up to 512 Mbytes of memory
  - 27 Watts

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

HPCシステムのサイクル



広範囲な適用用途と低コストを実現

多くのベンダーがシステムの販売に興味を持つ

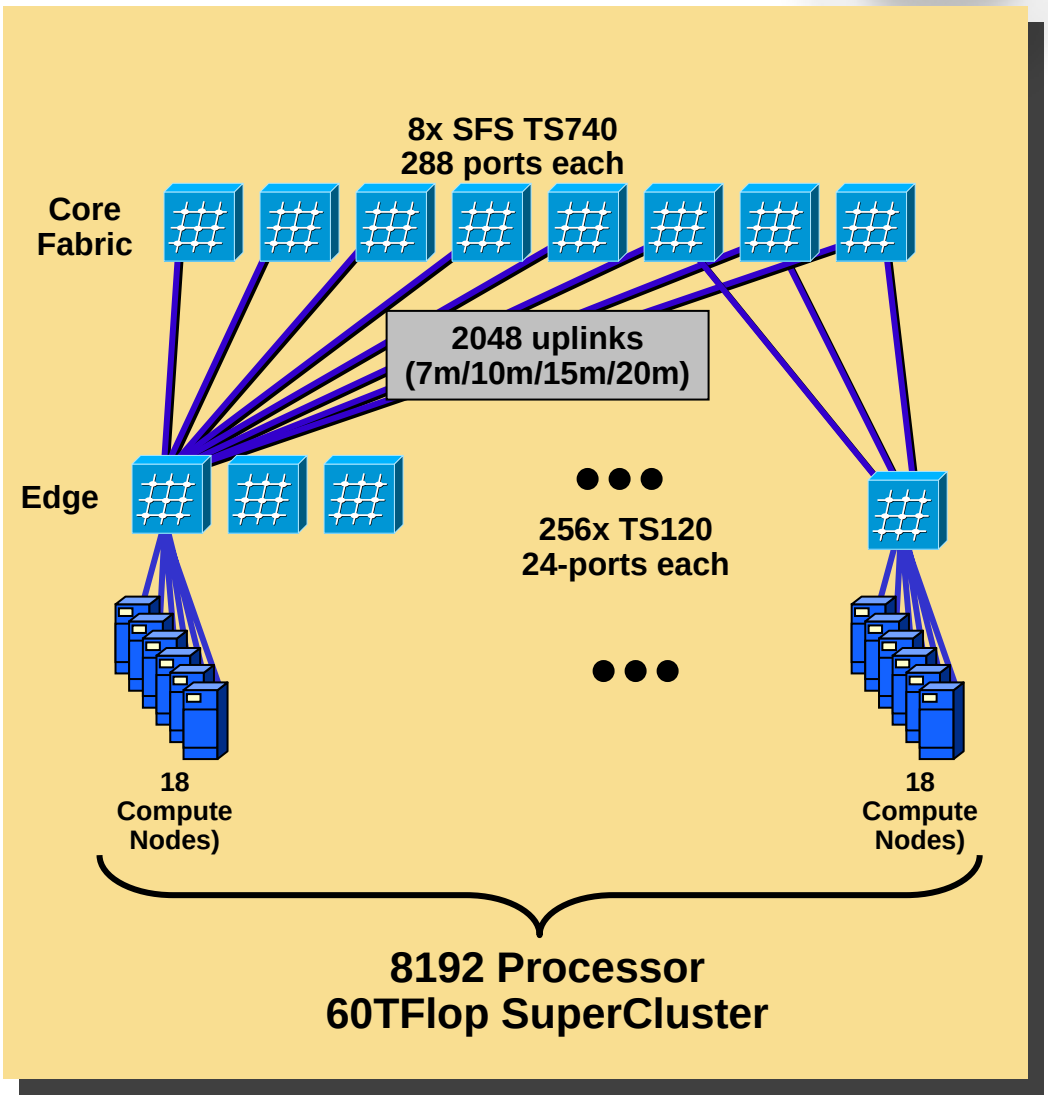
ユーザの関心

アプリケーションや開発環境の改善

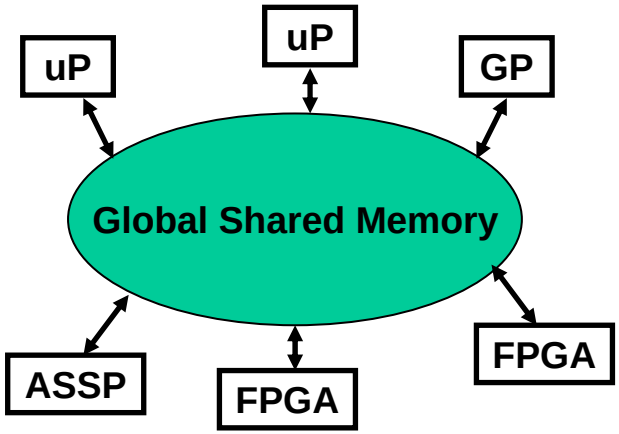
Large Government Lab

Worlds Largest Commodity Server Cluster – 4096 nodes

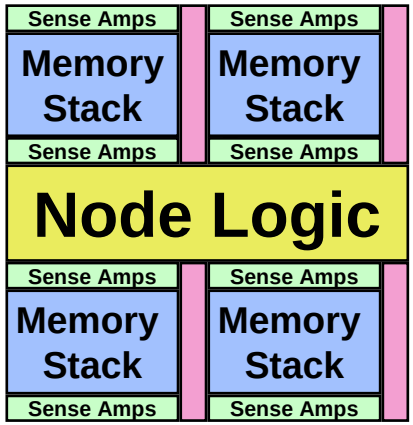
- Application:
  - High Performance Super Computing Cluster
- Environment:
  - 4096 Dell Servers
  - 50% Blocking Ratio
  - 8 TS-740s
  - 256 TS-120s
- Benefits:
  - Compelling Price/Performance
  - Largest Cluster Ever Built (by approx. 2X)
  - Expected to be 2nd Largest Supercomputer in the world by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



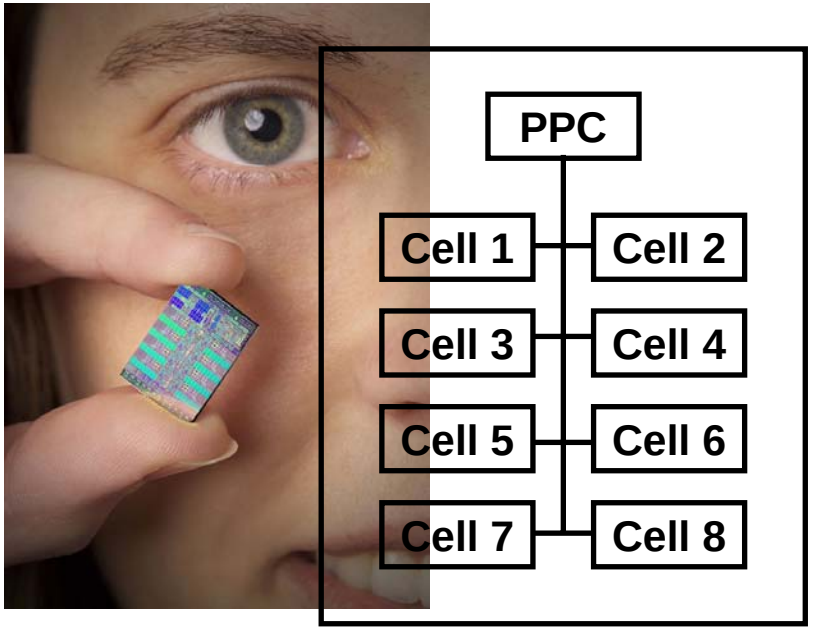
今後の計算機アーキテクチャ



SGI – Heterogeneous Architecture



Microprocessor with PIMs



Cell Processor – IBM, Sony, Toshiba

Heterogeneous Computing on Silicon

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