高性能コンピューティング特論
講義メモ8
「スーパーコンピュータ事例紹介」

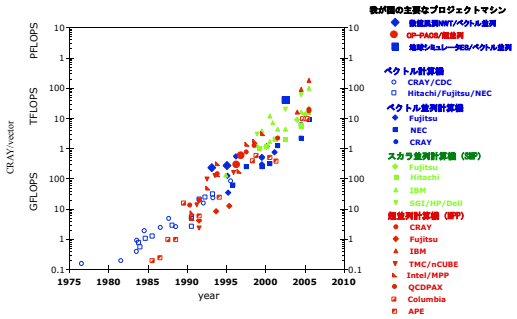
TOP500リスト

http://www.top500.org

スーパーコンピュータの発展

発展を支えた様々な技術：

- 半導体技術の進歩
（ムーアの法則）



- 数々のアーキテクチャ上のイノベーション
  - 1976年　ベクトルプロセッサアーキテクチャの出現
  - 1980年代　ベクトル計算機の隆盛
  - 1990年頃　マイクロプロセッサの発展と並列計算機アーキテクチャ
  - 1990年代以降　並列計算機の隆盛
    ベクトル並列、超並列、スカラ並列（SMP）

三好甫

数値風洞NWT (1993)

- 旧航空技術研究所にて、富士通と協力のもとに、開発・製作
- ベクトル並列アーキテクチャの導入
- 大規模フルクロスバー結合ネットワークの実現
- 1993年11月～1995年11月　トップ500一位
- 流体計算をはじめとする多様な分野で活躍
- ベクトル並列機へと発展

Fujitsu VPP500

Fujitsu VPP5000

NEC SX-4

NEC SX-5

CP-PACS(1996　筑波大学)

- 我が国初めての大規模
  汎用超並列型スーパーコンピュータ
  - スカラープロセッサだが擬似ベクトル機能
  - 柔軟・高性能なネットワーク
- 物理学と計算機工学の共同作業
- 大学とメーカ（日立）の産学連携
  （日立の商用機へと発展）
- 基礎物理（素粒子、宇宙）でブレークスルー
  - モデルではない、第一原理（基本方程式）からの近似なしの計算
  - 場（流体、電磁場、波動関数など）による自然記述一般に通ずる汎用性

超並列計算機PAX(PACS)の開発の歴史

- 1977年に研究開始（星野・川合）
- 1978年に第一号機が完成
- 1996年のCP-PACSはTOP500第一位
- 2006年のPACS-CSは第7号機

1978
第1号機PACS-9

1980
第2号機PAXS-32

1989
第5号機QCDPAX

1996
世界最高速を達成した第6号機CP-PACS

2006
PACS-CS

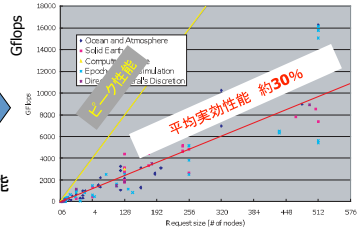
完成年	名称	計算速度
1978年	PACS-9	7千回/秒
1980年	PAXS-32	50万回/秒
1983年	PAX-128	4百万回/秒
1984年	PAX-32J	3百万回/秒
1989年	QCDPAX	14億回/秒
1996年	CP-PACS	614億回/秒
2006年	PACS-CS	14336億回/秒

- 計算科学者＋計算機工学者の共同開発による「実用的スパコン」
- Application-drivenな開発
- 持続的な開発による経験の蓄積

地球シミュレータ（2002）

- ベクトル並列型による超大規模システム
  - one-chip vector processorによるベクトルノード
  - 大量のバンク構成による高いメモリバンド幅
  - 高いネットワークバンド幅の1次元クロスバ結合
- 本格的計算科学としての地球規模の気候変動研究を始めて実現

高い
実効性能



HPCシステム概観（1）

- 100MFLOPS～1GFLOPS時代（’70代後半～’80代前半）
  - Cray-1に代表されるベクトル計算機
  - ベクトルレジスタと高バンド幅メモリによる高性能計算
  - 史上初の over 1GFLOPS 計算機：NEC SX-2
- 10GFLOPS～100GFLOPS時代（’80代後半～’90代前半）
  - ベクトルの高性能化に加え、パイプラインを並列に持つ
  - ベクトル計算機を共有メモリ結合（数台～数十台規模）
  - 超並列計算機の兆し
  - 1996年：国産計算機がTOP500中#1~#3を独占
    - CP-PACS, SR2201: 準ベクトル超並列
    - NWT: 並列ベクトル（共有メモリ）

HPCシステム概観（2）

- 1TFLOPSオーダー時代（’90代後半）
  - 汎用マイクロプロセッサによる超並列：ASCI machines
  - 史上初のover 1TFLOPS machine: SNL ASCI Red
- 10TFLOPS時代（’00代前半）
  - ASCI 終盤マシン
  - 地球シミュレータ (40TFLOPS)：共有／分散メモリ並列ベクトル
- 100TFLOPS時代（’00台中盤）
  - 超並列・省電力・省スペース：IBM GlueGene/L
- 1PFLOPS時代（’08）
  - 超並列・クラスタ・アクセラレータ：LANL RoadRunner
  - 超並列・汎用クラスタ：ORNL Jaguar
  - 史上初のover 1PFLOPS machine: Roadrunner, Jaguar
- この間、クラスタシステムの性能はコンスタントに向上し続けている

HPCシステム概観：全体的な流れ

- 初期スパコンはベクトル計算機
  - 大量のベクトルレジスタと高バンド幅メモリ（バンクメモリ）による物量作戦
  - ベクトルコンパイラ技術の進歩
- ベクトル単体の性能限界⇒並列ベクトル方式
  - 最初はベクトルパイプラインを並列化
  - ベクトル計算機を複数並列に（共有メモリ）
  - 共有メモリベクトル計算機の限界⇒分散メモリ並列ベクトル
- 超並列計算機の台頭
  - 汎用マイクロプロセッサの高性能化
  - 様々な並列ネットワークの登場
  - ベクトル計算機も超並列化
- クラスタ計算機の出現⇒主流へ
  - MPP（Massively Parallel Processor：超並列計算機）より安価
  - 汎用ネットワークの高性能化・大規模化
  - アクセラレータの出現とこれを利用しやすいシステム構成

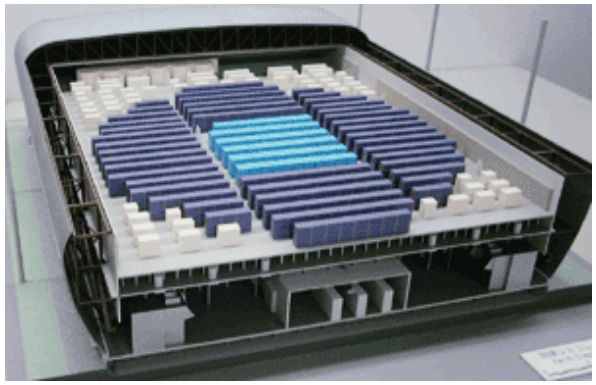
現在の大規模HPCシステムの姿

- 基本的に並列処理システムしかあり得ない
  - 要求される計算性能／データ処理性能に対し、単一プロセッサでは対応不可能
- 2つの大きな流れ
  - 世界最高性能レベルのシステム（数百TFLOPS～PFLOPS）のためには数万以上のプロセッサコアが必要
⇒極めて高密度な実装が必要
⇒超並列専用アーキテクチャ
  - 並列処理は計算センターや大企業だけでなく、研究室レベルや中小企業レベルでも手軽に使えるようになってきた
⇒安価で対価格性能比のよいPCクラスタが主流
⇒超大型システムのためには設置場所や空調等の条件が必要だが小・中規模システムではかなり手軽に導入可能

CP-PACS

- 筑波大学計算物理学研究センター
- 筑波大学＋日立
- 1996年完成
- 大学主導計算機として世界最高速となった貴重な例
- 計算物理学のための計算機
- ソフトウェアベクトル処理のために強化されたプロセッサ
- 2048 CPU 614GFLOPS

地球シミュレータ



- 海洋技術研究所・地球シミュレータセンター
- NEC
- 2002年完成
- 国産ベクトル計算機として世界最高速
- 大規模気象シミュレーション等様々な分野で応用
- 共有メモリ結合されたベクトルプロセッサ
- 5120 CPU
  40 TFLOPS

ASCI Purple

- Lawrence Livermore National Lab.
- IBM
- 2006年完成
- 共有メモリスカラプロセッサをネットワーク結合
- 応用分野は特になし
- 12208 CPU
  92 TFLOPS

Blue Gene/L

- Lawrence Livermore National Lab.
- IBM
- 2005年完成
- 組み込み用低性能プロセッサを非常に多数ネットワーク結合
- 素粒子計算、流体計算等
- 65536 CPU
  360 TFLOPS

PACS-CS

- 筑波大学計算科学研究センター
- 日立＋富士通
- 2006年完成
- メモリバンド幅・ネットワークバンド幅とCPU性能のバランス重視
- 計算科学全般
- 2560 CPU
  14.3 TFLOPS

Roadrunner

- Los Alamos National Lab.
- IBM
- 2008年完成
- 各ノードにOpteronプロセッサとIBM Cell Broadband Engineを搭載したハイブリッド型クラスタ
- 世界初のPFLOPSコンピュータ
- 129600 core
  1.46 PFLOPS
  (Linpack: 1.11 PFLOPS)

Jaguar

- Oak Ridge National Lab.
- Cray (XT-5)
- 2008年完成
- CPUはAMD Opteron (quad-core)だがネットワークは特別開発の3D-Torus
- 計算科学全般
- 18769 node
  (QC x 2 socket / node)
  1.38 PFLOPS
  (Linpack 1.06 PFLOPS)

クラスタ計算機によるスーパーコンピュータ

- クラスタ計算機：今日のHPCへの最大の貢献者
  - 昔のクラスタ："Poor Men's Supercomputer"
    ⇒今は小規模から超並列までクラスタ
  - TOP500リストにおけるランクイン数
  - ベクトル計算機等に比べ格段の対価格性能比（ピーク性能）
  - プロセッサとネットワークの両方にコモディティ技術を投入
- 同じプラットフォーム上で汎用計算と専用計算の両者を吸収可能
  - 64bit化されたIA-32 (x86)プロセッサをLinux環境で利用するのが典型的（近年、Windowsクラスタが出現）
  - 柔軟な拡張性（I/Oを通して）を用いた「加速装置」による高速化が可能
- 超並列化は必須
  - 単体CPUは速くはなっているが、全体の計算機性能要求はそれ以上
  - コモディティネットワークによる大規模化が可能になってきた

TOP500リスト (2010/11)

- アーキテクチャ

Architecture	Count	Share (%)	Rmax (GF)	Rpeak (GF)	Processor
Cluster	414	82.8	25641314	40666452	3602852
Constellations	2	0.4	94970	112947	17648
MPP	84	16.8	17936809	23875912	2851827
Totals	500	100.00	43673093	64655311	**6472327**

- プロセッサファミリ

Processor	Count	Share (%)	Rmax (GF)	Rpeak (GF)	Processor
Intel EM64T +AMD x86_64	449	89.8	37195653	56467398	4983111
Others	51	10.2	6477439	8187912	1489216
Totals	500	100.00	**43673092**	64655310	4672327

コモディティCPU

- COTS (Commodity Of The Shelf) CPUによる高い対価格性能比⇒本来「一種の専用計算機」だったスパコンを汎用技術で実現
  - Commodity – 「研究者（ユーザ）が開発費用を払わなくてよい計算機要素技術」（世界中の一般ユーザが代わりに払ってくれている）

    ⟷ ベクトル計算機 or MPP：ユーザが開発コストを払う
- 過去10年間で、単体CPUの性能は飛躍的に伸びている
  - 周波数の向上
  - マルチコア技術
  - SIMD (SSE-type) 命令によるFLOP/clockの向上

コモディティネットワーク（相互結合網としての）

- 古典的なコモディティネットワーク
  - Ethernet: 10base -> 100base -> 1000base -> 10Gbase
  - 極めて高い対価格性能比がバンド幅において得られている
    ただし、レイテンシ（遅延時間）については今一つ
  - 基本的に「木構造」なため拡張性が低い
- SAN (System Area Network / Storage Area Network)
  - Myrinet, Infiniband, Quadrix, ...
  - バンド幅とレイテンシの両方で高い性能
    ただし、かなり高価
  - Clos網やFat-Tree網が使えるため拡張性が高い
- 近年、SANの価格が急激に低下している
  - SANがコモディティとなってきた
  - On-board Ethernet NICの代わりにOn-board Infiniband NIC等も出てきた

クラスタの持つ高い「性能拡張性」

- 汎用I/Oバスの進化
  - PCI ⇒ PCI-X ⇒ PCI-Express ⇒ PCI-E gen2
  - 並列リンク ⇒ 複数の超高速シリアルリンク
  - CPUからの直接リンク: Hyper Transport, Quick Path
- ハードウェアアクセラレータの装着
  - Clear Speed: 東工大TSUBAME
  - Cell Broadband Engine: LANL Roadrunner
  - GRAPE: 筑波大FIRST Cluster
  - GPGPU: PC clusters全般

日本のTOPマシン（2010/11現在）

Machine	Site	Vendor	Rpeak (GF)	Rmax (GF)	#rank
TSUBAME2.0	Tokyo Inst. Tech.（東工大）	HP+NVIDIA	2287630	1192000	4
Xeon Cluster	JAEA（原研）	Fujitsu	200080	191400	33
Altix	U. Tokyo（東大物性研）	SGI	180019	161800	42
ES2 (SX/9)	JAMSTEC（海洋研）	NEC	131072	122400	54
FX1	JAXA	Fujitsu	121282	110600	59
T2K-Todai	U. Tokyo（東大）	Hitachi	138956	101741	70
RICC	RIKEN（理研）	Fujitsu	106042	97940	74
T2K-Tsukuba	U. Tsukuba（筑波大）	Appro	95385	77280	96

T2K-Tsukuba: 大規模汎用クラスタ

#20 at TOP500 on June 2008 (Linpack: 76.46 TFLOPS)

計算ノードとファイルサーバ

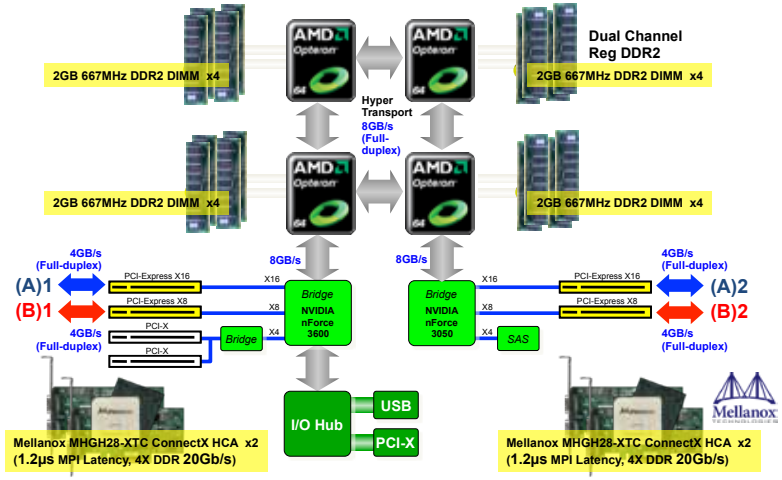
Computation node（70racks）

648 node (quad-core x 4socket / node)
Opteron "Barcelona" B8000 CPU
2.3GHz x 4FLOP/c x 4core x 4socket
= 147.2 GFLOPS / node
= 95.3 TFLOPS / system
20.8 TB memory / system

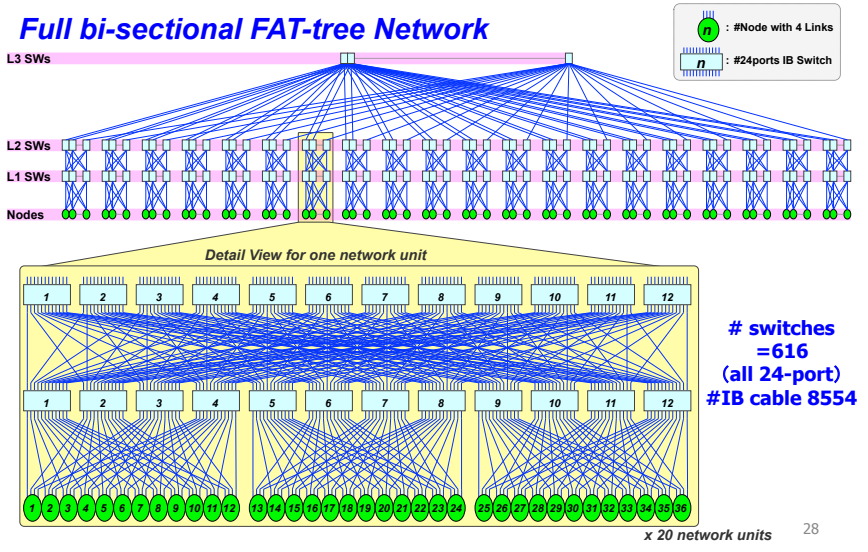
800 TB (physical 1PB) RAID-6
Luster cluster file system
Infiniband x 2
Dual MDS and OSS config.
⇒ high reliability

File server (disk array only)

計算ノードのブロックダイアグラム



Infiniband 4xDDR x 4-rail の Fat-Tree網



大規模HPCクラスタのトレンド

- 飛躍的なCPU性能
  - Quad-core processorの登場⇒さらに6~8 coreへ
  - IA-32 SSE命令⇒1 clock当たりの演算性能向上
  - SSE dual issue⇒ 1 clockで2つのSSE命令実行
- 高性能インターコネクトの低価格化と性能向上
  - Infinibandに代表されるSAN (System Area Network)がHPCクラスタの普及と共に急激に低価格化している
  - 大規模化に対応する大容量スイッチの登場
  - 小容量スイッチの低価格化
  - 光ケーブルの低価格化
  - Infinibandの性能：DDR⇒QDR（40Gbps/direction）⇒EDR…
- 高性能大規模クラスタは全盛時代へ

Heterogeneous Computing Platform

- 計算ノードに汎用プロセッサと専用プロセッサを混在（混載）させた（並列処理）システム
  - 基本的なスタイル：クラスタ型並列計算機の計算ノードに何らかの演算加速装置を搭載
    - ClearSpeed
    - GRAPE
    - GPGPU (General Purpose – Graphic Processing Unit)
    - Cell Broadband Engine
  - 従来より特定アプリケーションにおける高性能化が図られてきたが、2008/06 にLANLにおいてRoadrunnerがOpteron + Cell BEのヘテロ構成でover 1PFLOPS Linpack性能を達成（ピーク性能は1.3PFLOPS）
  - “Hybrid Computing” という言葉は共有メモリと分散メモリのprogramming paradigmを指す場合があるので、ここではHetero-… と呼ぶ

FIRST

- 筑波大学計算科学研究センター
- HP＋浜松メトリクス
- 2005年完成
- 全ノードにBlade-GRAPE重力アクセラレータボードを搭載
- 計算宇宙物理学におけるハイブリッド計算
- 512 CPU
  +256 GRAPE
  3.5TFLOPS (CPU)
  +35TFLOPS (GRAPE)

TSUBAME

- 東京工業大学
- SUN Microsystems + NEC
- 2006年完成
- 各計算ノードをmulti-core CPU (dual-core Opteron) とアクセラレータ（ClearSpeed, GPU) によって構成したヘテロクラスタ
- 計算科学全般
- 655 nodes/10480 cores
- 163TFLOPS（アクセラレータ込み）
  (Linpack 87TFLOPS)

2010/11時点の最高速マシン Tianhe-1A (China)

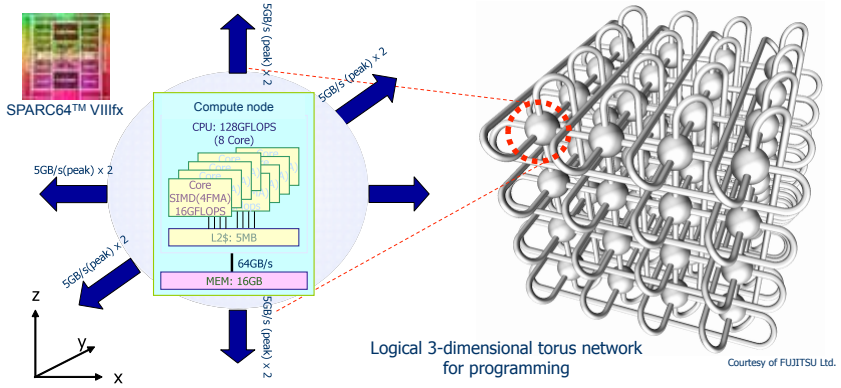
- GPU accelerated PC cluster
  - Intel Xeon + NVIDIA Tesla (Fermi)
  - Xeon (6-core) 14336台 + Tesla 7168台
  - Rpeak=4.7PFLOPS, Rmax=2.57PFLOPS
- Network: original (NUDT, Infiniband QDRx2相当？)
- 中国がTOP500が初めてトップに
- 総電力4MW
  ⇔ 2位のJaguar (Rpeak=2.33PFLOPS) の電力は6.8MW
- 世界5位（日本1位）のTSUBAME2.0もPFLOPSを超えるGPUクラスタだが、Tianhe-1Aは規模が2.5倍ぐらい大きい
- ただし、GPU性能当たりのLinpack効率と電力効率（FLOPS/W）はTSUBAME2.0の方が上回っている
- GPUによる over PFLOPS clusterは実現されたが、今後は大規模並列アプリケーションがどれだけ走るかが問題

「京」コンピュータ (“K Computer”)

- 日本が「次世代スパコン」として2006年ぐらいから計画、2008年から本格的開発開始（当初、「京速」コンピュータと呼ばれた）
- 2012年に10PFLOPS以上の性能（Linpack性能）を目標に理化学研究所が開発
  ⇒ 2010/10にAICS (先端計算科学研究機構、神戸）が運用・研究機関として設立される
- 富士通がシステム構築を請負
  （当初、NEC+日立が担当するベクトル機部分もあったが2009/5に両者が計画を降りた）
- 8-coreの高性能プロセッサコアを持つCPU (Fujitsu SPARC64VIIIfx)を1ノードに1つ搭載、4ノードを1つのマザーボードに実装
- 相互結合網は独自の6次元トーラス⇒これをユーザには3次元トーラスとして見せる： ToFu network

Compute Nodes and network

- Compute nodes (CPUs): > 80,000
  - Number of cores: > 640,000
- Peak performance: > 10PFLOPS
- Memory: > 1PB (16GB/node)
- Logical 3-dimensional torus network
- Peak bandwidth: 5GB/s x 2 for each direction of logical 3-dimensional torus network
- bi-section bandwidth: > 30TB/s

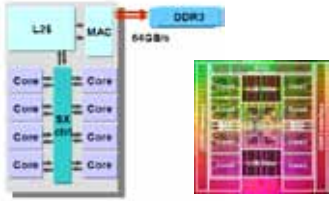


Logical 3-dimensional torus network for programming

Courtesy of FUJITSU Ltd.

CPU Features (Fujitsu SPARC64™ VIIIfx)

- 8 cores
- 2 SIMD operation circuit　16GF/core(2*4*2G)
  - 2 Multiply & add floating-point operations (SP or DP) are executed in one SIMD instruction
- 256 FP registers (double precision)
- Shared 5MB L2 Cache (10way)
  - Hardware barrier
  - Prefetch instruction
  - Software controllable cache
    - Sectored cache
- Performance
  - 16GFLOPS/core, 128GFLOPS/CPU

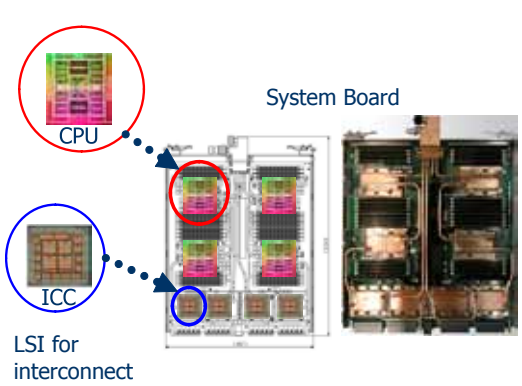


45nm CMOS process, 2GHz
22.7mm x 22.6mm
760 M transisters
58W（at 30℃ by water cooling)

Reference: SPARC64™ VIIIfx Extensions
http://img.jp.fujitsu.com/downloads/jp/jhpc/sparc64viiifx-extensions.pdf

Photo of proto-type system

- Prototype system has been built.
  - Several system boards are compiled and set into a cabinets.

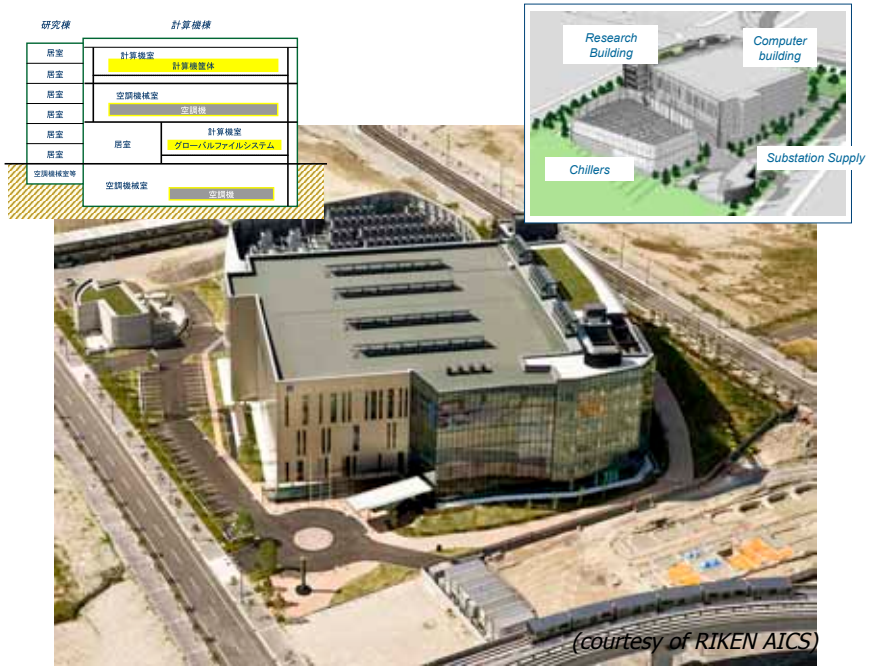


Courtesy of FUJITSU Ltd.

「京」コンピュータ

AICS building at Kobe

(courtesy of RIKEN AICS)



(courtesy of RIKEN AICS)

今日の最先端コモディティCPUの性能上の問題

- メモリとI/Oのバンド幅に比べ、極めて高い浮動小数点演算性能
  - CPU周波数は限界に達しつつあり（消費電力のため）これ以上大幅な向上は見込めない
  - 半導体テクノロジは進歩し続けており (90[nm] ⇒ 65 ⇒ 45 ⇒ 22 ⇒ ...) ダイ上のトランジスタ数は増加を続ける
  - 「ピーク性能維持」のため、マルチコア／メニーコア化していくのは自然な流れ
- メモリバンド幅: “Rich Vector” vs “Poor Scalar”
  - CPU FLOPS とメモリバンド幅のギャップは確実に増え続けている ⇒ 深刻な問題
  - ピーク性能 (Linpack) と実効性能 (non-cache-aware applications)の差が益々大きくなっていく

Balance on CPU : Memory : Network

Systems	C : M : N = GFLOPS : GB/s : GB/s			C : M : N (M = 1.0)		
CP-PACS	0.3	1.2	0.3	0.25	1	0.25
Earth Simulator	64	256	12.5	0.25	1	0.05
PACS-CS	5.6	6.4	0.75	0.90	1	0.1
T2K-Tsukuba	147.2	42.7	8	3.50	1	0.2

Cが小さくNが大きいほど「バンド幅的に」よい

かつてのベクトル計算機＝”4Byte/FLOP”
（上の表では C:M = 0.25:1)

QCD (Quantum Chromo Dynamics) に必要なメモリバンド幅

- “QCD-mult” benchmark core equation

$$\mathrm{Mult}(n,m)^{a,b}_{\alpha,\beta} = \sum_{\mu=1}^{4}\left[(1-\gamma_\mu)_{\alpha,\beta}(U_\mu(n))^{a,b}\delta_{n+\hat{\mu},m} + (1+\gamma_\mu)_{\alpha,\beta}(U^\dagger_\mu(n-\hat{\mu}))^{a,b}\delta_{n-\hat{\mu},m}\right]$$

dim.	computation	load	store	B/flops
t	168(x), 120(+) = 288 flop	21*2 complex = 672B	12 complex = 192B	3.00
z	144(x), 192(+) = 336 flop	21*2+12 complex = 864B	12 complex = 192B	3.14
y	144(x), 192(+) = 336 flop	21*2+12 complex = 864B	12 complex = 192B	3.14
x	144(x), 192(+) = 336 flop	21*2+12 complex = 864B	12 complex = 192B	3.14
clover	288(x), 312(+) = 600 flop	21*2+12 complex = 864B	12 complex = 192B	1.76

(by Prof. Ishikawa @ Hiroshima U.)

⇒ 5088B / 1896flop = 2.68 Byte/flop
⇒ 近年のコモディティCPUの傾向では圧倒的に不足

メモリバンド幅ボトルネック

- メモリバンド幅を求めるHPCアプリケーション
  - 流体計算、気象予報
  - QCD (Quantum Chromo Dynamics)
  - FFT
  - 遠距離相互作用を伴う粒子シミュレーション
  - ...
- これらのアプリケーションでは現在のコモディティCPUのトレンドであるマルチコアアーキテクチャが必ずしも性能向上に結びつかない
  - 必要とされる Byte/FLOP 性能が提供されない
  - 本質的にレジスタ、キャッシュ等のオンチップストレージではまかない切れないデータ容量
- データの localization（局所化）が鍵⇒万能ではない
  - アプリケーションをキャッシュにフィットさせるチューニングが主流 ⇒ cache-awareness

システムアーキテクチャの今後

- クラスタシステム
  - マルチコア／マルチソケット／マルチNIC化
    - メモリ階層の多様化とネットワークインタフェースの複数化によりプログラミングと性能チューニングは益々複雑に
    - 共有メモリ（マルチコア＋マルチソケット）と分散メモリ（インターコネクト）のハイブリッドアーキテクチャ上のプログラミングをどうするか
      ⇒現在はユーザレベルで共有メモリ（ex. OpenMP）＋分散メモリ（ex. MPI）
      を明示的に記述
      ⇒新たな programming paradigm, compiler technologyが求められる
  - 大規模化はまだ可能だが、設置上の制約が厳しい
    - スペース：特に日本
    - 電力：コア当たり電力は下がっているがノード当たりは横ばい
    - 空調：「安いクラスタ」が仇に…

⇒クラスタの大規模化による性能向上はいつかは終わるのではないか？

演算加速装置への期待

- 従来の専用演算装置から（準）汎用演算装置へ
  - 従来型GRAPEのような特殊な演算加速から、SIMD命令による汎用的な演算パイプラインが利用可能
  - GPGPUでは標準的なプログラミングツールが準備されつつある（nVIDIAのCUDA や PGIのCUDA準拠コンパイラ等）
  - 汎用CPUに比べ極めて高い性能／電力比
    - Opteron B8300 ：37GFLOPS / 120 W = 0.3GFLOPS/W
    - TESLA C1060 card : 1TFLOPS / 160W = 6.25GFLOPS/W

約20倍!!

- On-chipでのメモリ／データア・クセスバンド幅
  - GPGPUでは非常に高い内部メモリバンド幅を実現
  - GRAPEではツリー型データパスにより各データ流のスループットは非常に高い
- 大規模計算を行う際の信頼性
  - 現在のGPGPUはECCがない！

現在の演算加速装置の問題点

- メインCPUとの間のデータ移動バンド幅が小さい
  - 何らかの外部バスによる接続⇒現在は標準的にPCI-Expressが用いられる
    ⇒ PCI-E Gen2 x 16 でも理論ピーク性能 10GB/s
    最高性能のDDR3 メモリ等に比べ数分の一
  - 例：1台のGForceGTX280上でCUDAを用いて姫野ベンチマークが70GFLOPS以上で動作する
    ⇒あくまで1台のGPU上での話
    ⇒複数台接続やマザーボード上でCPUと結合すると性能が著しく低下する
- レジスタ、内部メモリ容量、SIMD命令等により適用アプリケーションの制約がまだある

超並列計算システムの今後

- 最大の課題：電力
  - 今後、Peta-FLOPS から Exa-FLOPS まで伸びるにはコア当たり消費電力を飛躍的に小さくする必要がある
    - 半導体プロセス技術の限界
    - リーク電流が相対的に大きく、DVFS (Dynamic Voltage Frequency Scaling) 等では対応不可能
    - 根本的な半導体、回路レベルでの技術革新が必要
- ネットワークの大規模化への対応
  - クラスタで主流となっている fat-tree 構成のインターコネクトは規模の限界へ
    - 近接通信アーキテクチャ：「近距離＝省電力」の魅力
    - アプリケーション／アルゴリズムの改良により遠距離通信を削減
- システム安定化技術
  - 耐故障性の提供：数百万のプロセッサとネットワークデバイスを無故障運転させることは不可能
    ⇒動的な耐故障技術が必須
  - 古典的な check-pointing / restart 技術を越える、超並列対応の failure recovery 技術が求められる