SC08 参加報告

防衛大学校　応用物理学科　萩田 克美

2008 年 11 月 18 日～20 日、アメリカ合衆国テキサス州オースティンで開催された SC08 に参加した。SC は、1988 年以来、毎年開催されている「世界最大のスパコンに関する会議・展示会」である。SC は、ベンダーを中心とした展示会がメインであり、スパコンをユーザーとして利用する私にとっては、まさに、自分の買い物とは関係ない先端的な車や個別の部品が展示される「東京モーターショー」にでも来ている感覚である。展示会と並行して、HPC に関連する先端的な研究を報告する国際会議も催されている。ここでは、理研のレポーターとして、いくつかの研究報告を聴講した結果を中心に報告する。本報告では、HPC に興味があり多様な専門を背景にもつ方々が読者であると想定し、記録的価値を高めるとともに SC08 の雰囲気を伝えるために、写真を多く用いることとした。偶然ではあるが、私は、昨年度の SC07 と欧州で毎年開催されている ISC08 にも参加している。それらの経験を踏まえつつ、本報告を読んでいただける方とって参考になる情報提供ができれば幸いである。

昨年から注目されていた D. E. Shaw リサーチの Anton の動向であったが、事前のアブストラクト Web 上では見当たらなかった。D.E. Shaw の関連した研究の発表によれば、2008 年 10 月に 512 ノードでのプロトタイプが動作したとのことであり、アナウンスは控えめであった。25,000 原子のたんぱく質の MD についてミリ秒の計算を実現するのは、まもなくといったところのようである。

SC では、国際会議に先立ち、HPC 教育セッションとチュートリアルが開催されている。本年度は、参加することができなかったが、価値の高い催しなので、1 年さかのぼって簡単にレポートしておく。SC07 の教育セッション（Educational Program）では、全米における HPC 関連の初等中等教育についての発表がなされると共に、スパコン利用を支援する Web ベースアプリなどのハンズオンでの講習会が催されていた。また、SC07 のチュートリアルでは、CUDA プログラム、MPI と OpenMP のハイブリッドプログラム、OpenMP 3.0 など、最新のトピックが、実例と経験を基に、日本国内の講習会などでは聞けないハイレベルな内容があり、実際にスパコンを使う研究者・技術者にはお勧めである。

さて、今年の SC08 の報告に移ろう。

Registration は、昨年度同様、PC で各自参加登録し、有人ブースでネームプレートを受け取るものである。ネームプレートには、磁気テープ付いており、展示ブースに設置された読み取り機を通じ個人情報が展示者に引き渡される仕組みとなっている。このあたりのシステマティックさは、1 万人の参加者を裁く「The 展示会！」といった感じである。展示会の顔ぶれは、昨年と大体同じである。日本からの参加としては、地球シミュレータを ES2(NEC SX-9)にリプレースする JAMSTEC が復活参加している以外、TOP500 に絡むようなスパコンを有する大手の HPC ファシリティが継続的に参加している。ロスアラモス研究所、NASA、オークリッジ研究所、アルゴンヌ研究所などのペタ flops 級のスパコンが導入されている大型ファシリティのブースは、非常に大きく、目立っていた。相対的に理研のブースが小さく感じたのが率直な印象である。

初日、8:30 より、Welcome セッションと Keynote セッションが開始された。SC20 周年の会議ということもあり、Gordon Bell を始め、スパコン業界の著名人が、これまでのスーパーコンピュータの歴史語る 10 分間のビデオが上映された。その中で、日本の計算機としては地球シミュレータ誕生の衝撃とその成果について絶賛されていた。あらためて地球シミュレータのスパコン史における位置づけの重要さを認識した。最初の Keynote としては、Dell の創始者 Michael Dell が、Dell の HPC への Contribution について講演した。ベクトル、マルチプロセッサと HPC が進化し、現在では、x86 などの標準化された HPC が、TOP500 に占める割合を高めるに至り、Dell が活躍しているという骨子である。

その後、10 時半からは、大注目のロスアラモス研究所の Roadrunner の講演であった。Roadrunner のシステム詳細が紹介され、各アプリケーションでの性能評価の結果が示された（詳細は、聴講報告①参照）。アーキテクチャの説明から受けた印象では、PlayStation3 の MPI プログラミングの延長として対応できるとの感触を得た。次の講演では、NASA が所有する最新計算システムのアプリケーション毎の性能比較についての紹介であった（詳細は、聴講報告②参照）。CPU の単体性能としては Xeon よりも Power5+や Itanium2(Montvale)がよいという当然の結論であるが、SGI ICE では CPU 数を増やして性能を出している。Roadrunner と SGI ICE の両方とも、LINPACK に対する経済性は十分であるが、アプリケーション開発に対する経済性について苦しんでいるのではないかという感覚を覚えた。

午後は、「HPC in Transportation」セッションに参加した。Goodyear と Boeing の発表であった。両者とも、HPC を駆使して開発をしている現状を事例紹介した。産業利用のためには、実験とシミュレーションを確実に連携させている点は学ぶところが大きいと考えた。また、Goodyear は、エネルギー省の CRADA

プログラムの下、Sandia 研究所と長年共同研究を進め、着実に成果をあげている点も注目すべきであると感じた（詳細は、聴講報告③参照）。

休憩を挟み、「HPC in Finance」セッションに参加した。Morgan Stanley のみの発表であった。当初は、Golden Tree Asset Management も発表する予定であったが、キャンセルになっていた。一連の金融危機が関係あるのかもしれません。Morgan Stanley の発表では、リスク計算のモンテカルロ法などでの入力条件に対する分岐により計算が膨大になるので、CPU の数が勝負となる旨強調していた。

1 日目の夕方のイベントは、最新の TOP500 リストの発表とポスター発表である。TOP500 の BOF で配布された「The TOP500 Report」に記載される TOP25 から、ついに、日本勢が姿を消している。復活することを期待するのみである。ちなみに、TOP9 まで USA、10 位が中国である。1 位、2 位が 1 ペタ級、3 位から 6 位が 0.5 ペタ級、7 位～11 位（ドイツ）が、0.2 ペタ級である。今回の注目は、2008 年 6 月(ISC08)で初めて 1 ペタ flops 超えて 1 位となった Roadrunner と前評判の高い Cray XT5 の勝負であった。結果として、ISC08 からさらにアップグレードした Roadrunner が僅差で死守した形である。

TOP500 の発表の後は、ポスター発表に参加した。ポスター発表は、あらかじめ注目していた 3 件を中心に回った（詳細は、聴講報告④～⑥参照）。プログラムのチューニング、GPGPU や CELL のプログラミング、CPU やネットワークシステムの性能評価など、HPC の基盤技術に関する基礎的な検討に関する発表が幅広く行われている。正直に言えば、大半の発表が専門的すぎ理解することができないものであった。このような専門的で地道な研究が結集して、HPC ができているということを改めて感じた。

2 日目、午前は、癌に対するテーラーメード医療のための情報供給システムの構築の話と、UC バークレーにおける新たな並列計算に関する研究センターについての紹介があった。前者については、巨大な情報システムの構築という点では理解できるが、大規模計算という意味での HPC がどこに組み込まれてくるのか良く分からなかった（詳細は、聴講報告⑦参照）。Grid で情報収集し分析を通じて業務に活かしていくことはあらゆる分野で成り立つ手法であり、先駆的な取組みで一般の技術レベルが向上していくことは価値が高いと考えられる。後者の講演では、100Core 以上の並列ソフトウェアの作成を目標に、84 台の FPGA を用いた 1008Core の RISC エミュレーション環境を構築して、新しい並列計算に関する研究を進めているとのことである（詳細は、聴講報告⑧参照）。

その後、Blue Gene の BOF に参加した（詳細は、聴講報告⑨参照）。実質的には、エネルギー省の INCITE プログラムで実施された計算事例紹介であった。

午後、Gordon Bell 賞の Finalists の発表を聴講した。今年度の Gordon Bell の Finalists は、すべて、アメリカの最新スパコンを利用し、数万の CPU を利用した並列計算である。昨年は、日本からも地球シミュレータや MD-GRAPE3 などで Finalist を排出していたが、今年はついにアメリカ以外は手を出せない状況といった感じである。最初の講演は、地震波の進行に関するシミュレーションを Texas の Ranger システムで 62,976 個の Core を用いて並列計算したものであった。論文の PDF は、http://scyourway.nacse.org/conference/view/gb109 を通じ、IEEE のアーカイブか ACM のデジタル図書館から入手可能である（以下同じ）。次の講演は、オークリッジ研究所の高温超伝導の不純物効果に関するシミュレーションであった。この研究では、Cray XT4 の 49,282CPU（2/3 がオークリッジ研究所のスパコン、1/3 はテネシー大学のスパコン）を用いて、ハバード模型の量子モンテカルロ計算を行い、409TFlops の性能を示したものである。論文は、http://scyourway.nacse.org/conference/view/gb118。さらに、Cray XT5 と Cray XT4 を同時利用し、1.3 ペタ Flops の実効性能を出したとのことである。本日最後の講演は、Texas の Ranger を用いたマントル対流の計算であった。論文は、http://scyourway.nacse.org/conference/view/gb132。32,786Core の並列性能はよいが、62,464Core ではパフォーマンスが落ちていた。

夕方は、SGI ICE ユーザーの BOF に参加した（詳細は、聴講報告⑩参照）。1Core あたりのメモリが小さく苦労しているというのが全体を通じての印象であった。

最終日、冒頭で、次回 SC09 が開催されるポートランドの説明があった。富士山のようなきれいな山があるということが印象に残った。さて、午前は、「エネルギーと HPC」というテーマで２つの講演を聴講した。最初の講演では、エネルギー問題全般の話と、ITER、燃焼、材料などの分野で HPC が重要な役割を果たすであろうことを力説していた。次の講演では、CO2 貯蔵技術におけるシミュレーション事例についての紹介がなされた。両講演から、エネルギー問題において HPC の果たす役割は大きいと感じた。

その後、Mathematica の MPI 拡張に関する Exhibiter Events に参加した（詳細は、聴講報告⑪参照）。

引きつづき、Gordon Bell 賞の Finalist の講演を聴講した。本日最初の講演は、Roadrunner を用いたレーザープラズマのシミュレーションであった。論文は、http://scyourway.nacse.org/conference/view/gb121。２番目の講演は、引き続き Roadrunner を用いた MD シミュレーションであった（詳細は、聴講報告⑫参照）。論文は、http://scyourway.nacse.org/conference/view/gb124。Cell 上での MD プログラムのチューニングをしっかりやっている印象を受けた。また、個

人的には、Cell-SPE と Opteron の間の通信が多いような気がしたが、実際にやってみなと判断がつかないと思われる。最後の講演では、ローレンス・バークレー研究所の Cray XT5 などを用いた LS3DF というフラグメント法の電子状態計算であった。論文は、http://scyourway.nacse.org/conference/view/gb115。この計算では、Cray XT4 の 36,864Core 並列で 135Tflops、Blue Gene/P の 163,840Core 並列で 224Tflops、Cray XT5 Jaguar の 147,456Core 並列で 442Tflops の性能を示すとのことである。その後、HPC Advisory Council Initiative の BOF に参加したが、収穫はなかった（詳細は、聴講報告⑬参照）。

午後は、SC Award の発表であった。今年度の Gordon Bell 賞（ピーク性能）は、オークリッジ研究所の「New algorithm to enable 400+ TFlop/s sustained performance in simulations of disorder effect in high-Tc superconductors」であった。この研究では、Cray XT4 の 49,282Core を用いて、ハバード模型の量子モンテカルロ計算を行い、409TFlops の性能を示したものである。Special Award（Algorithmic Innovation）は、ローレンス・バークレー研究所の発表「Linear Scaling Divided-and-conquer Electronic Structure Calculations for Thousand Atom Nanostuctures」が受賞した。

午後の後半は、「Applications: Models and Analysis」というセッションに参加した。このセッションでは、地震のシミュレーション、MD 軌道の並列解析、CELL を用いた並列推定計算の 3 件の講演があった。MD 軌道の並列解析については、D. E. Shaw リサーチの発表であり、Anton の動向が得られるであろうと期待していた。実際、Anton 関連では、プロトタイプが動作したとの情報のみであった。講演の内容は、分散入力の解析ツールを整備したとのことであった（詳細は、聴講報告⑭参照）。

最終日の夕方、会議のディナーがテキサスの牧場で開かれた。バスに乗るまで約 1 時間、そして 30 分の移動後についた牧場は、難民キャンプの様相であった。とにかく、この時期のテキサスは寒い。明らかに 10℃よりも低い。まずはお酒を確保し、列に並ぶこと 30 分、やっと食料にたどりつくという感じであった。井上室長、野村技術参事、南チームリーダーには大変お世話になりました。

最後に、SC08 への参加の機会をいただいた理化学研究所次世代スーパーコンピュータ開発実施本部にお礼申し上げます。特に、出張の支援をしていただいた内田さんにはお世話になりました。また、本出張のために、授業の代理をしていただいた防衛大学校応用物理学科の同僚には、感謝の意は絶えません。そして、同じ理研レポーターの立場で参加された斉藤さん、牧野さんには、お世話になりました。特に、旅程を同じくした斉藤さんには、多くの楽しい時間とドキドキした時間を共有させていただきありがとうございました。

●聴講報告

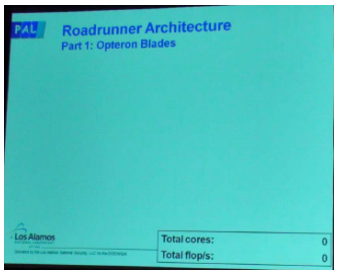
① Papers「Entering the Petaflop Era: The Architecture and Performance of Roadrunner」by Kevin Barker et. al. (Los Alamos National Laboratory)

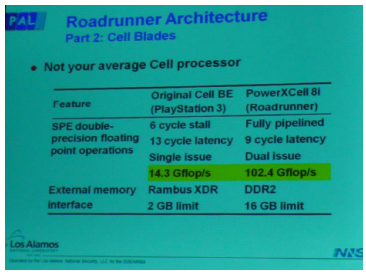
ロスアラモス研究所に導入され、今回の TOP500 で 1 位となった Roadrunner についての紹介であった。アーキテクチャの説明からは、PlayStation3 の MPI ベースのプログラミングスキルがあれば、対応できると感じた。Opteron と Cell の両方を活用してアプリケーション性能を引き出すことに苦労している様子が伺えた。Opteron と Cell の PPE との通信性能がネックになっていると感じた。また、3.6Gflops の Opteron と 102.4Gflops の PowerXCell がアプリレベルの性能が同じオーダーというのは、アプリが Opteron よりに造られているためであり、PPE-Opteron の通信律速に加え、アルゴリズム上の課題も大きいと感じた。

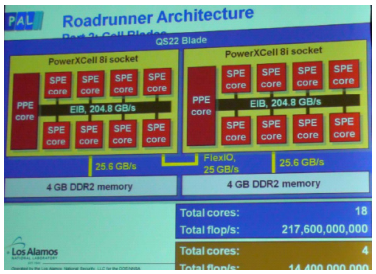
論文の PDF は、http://scyourway.nacse.org/conference/view/pap333。

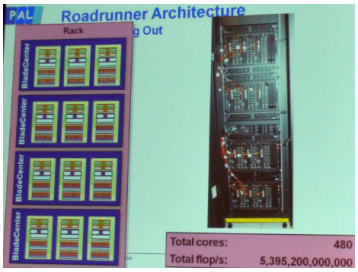
発表のポイントは次の通りである。

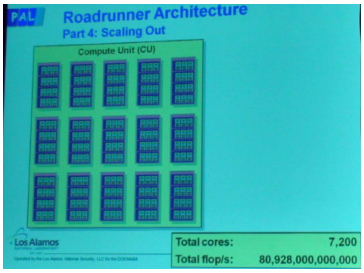
- Roadrunner とは、ロスアラモス研究所に設置された 1.38Pflops@peak、1.026Pflops@LINPACK の性能を持つ計算機である。12240Core の Opteron と 12240Core の CELL のハイブリッド計算機。ピーク性能が倍精度で 1.38 ペタ flops、単精度 2.9 ペタ flops。1Core のメモリは 4GB で総メモリは 98TB。（実際の計算には、28TB が Opteron、52TB が Cell に割り当てられている。）
- PPE と Opteron の間の通信が遅くてもよいとの考えでシステム構築を行っている。これは意外にも思えるが、Playstation3 での MPI 通信を扱うプログラムを考え、Opteron レイヤーで汎用的な計算を行うとすれば、CELL の性能を引き出すために必須の SPE と PPE の通信と Opteron 間の通信が速ければよいということは納得である。
- しかしながら、現行システムでは、PPE と Opteron の通信がボトルネックである。PPE と Opteron の通信が、Opteron 間の通信に比べ、相対的に性能が上回るあたりで、よい性能を示している。
- 7 種類のアプリケーションを Roadrunner の初期アプリと設定し、その評価結果を発表した。CELL を使えば、Opteron のみに比べ 2～6 倍速くなる。

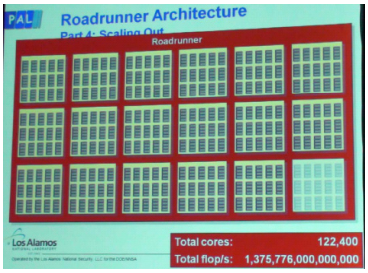


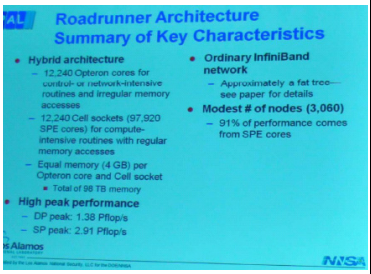


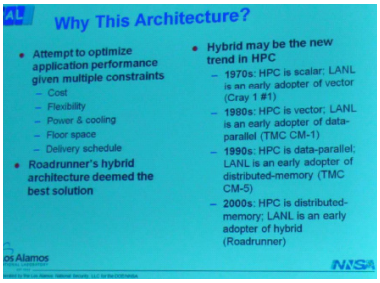


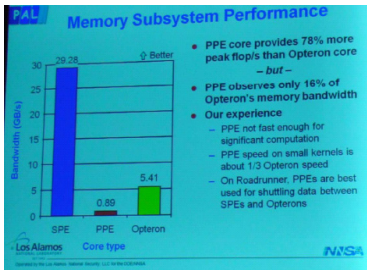

Roadrunner Architecture
Rack
Total cores: 480
Total flop/s: 5,395,200,000,000

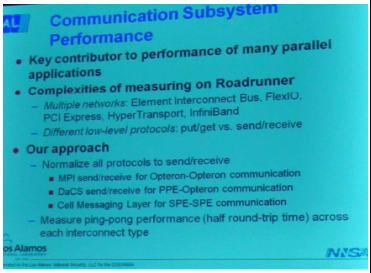

Roadrunner Architecture
Part 4: Scaling Out
Compute Unit (CU)
Total cores: 7,200
Total flop/s: 80,928,000,000,000

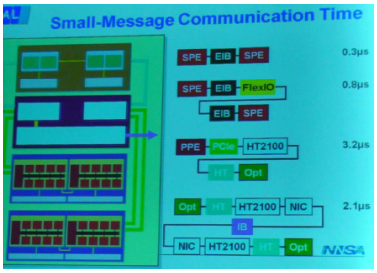

Roadrunner Architecture
Part 4: Scaling Out
Roadrunner
Total cores: 122,400
Total flop/s: 1,375,776,000,000,000

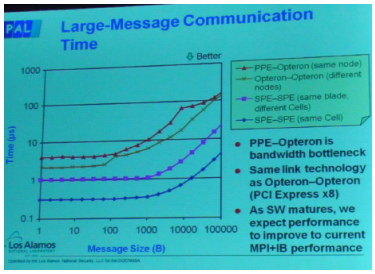

Roadrunner Architecture
Summary of Key Characteristics
Hybrid architecture
12,240 Opteron cores for control- or network-intensive routines and irregular memory accesses
12,240 Cell sockets (97,920 SPE cores) for compute-intensive routines with regular memory accesses
Equal memory (4 GB) per Opteron core and Cell socket
Total of 98 TB memory
High peak performance
DP peak: 1.38 Pflop/s
SP peak: 2.91 Pflop/s
Ordinary InfiniBand network
Approximately a fat tree—see paper for details
Modest # of nodes (3,060)
91% of performance comes from SPE cores

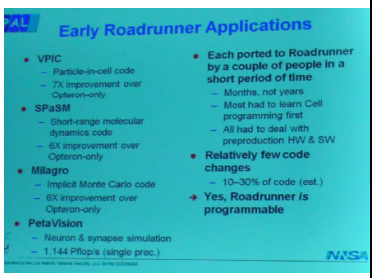

Why This Architecture?
Attempt to optimize application performance given multiple constraints
Cost
Flexibility
Power & cooling
Floor space
Delivery schedule
Roadrunner's hybrid architecture deemed the best solution
Hybrid may be the new trend in HPC
1970s: HPC is scalar; LANL is an early adopter of vector (Cray 1 #1)
1980s: HPC is vector; LANL is an early adopter of data-parallel (TMC CM-1)
1990s: HPC is data-parallel; LANL is an early adopter of distributed-memory (TMC CM-5)
2000s: HPC is distributed-memory; LANL is an early adopter of hybrid (Roadrunner)

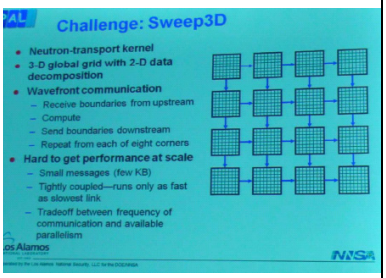

Memory Subsystem Performance
Better
29.28
0.89
5.41
SPE
PPE
Opteron
Core type
Bandwidth (GB/s)
PPE core provides 78% more peak flop/s than Opteron core
– but –
PPE observes only 16% of Opteron's memory bandwidth
Our experience
PPE not fast enough for significant computation
PPE speed on small kernels is about 1/3 Opteron speed
On Roadrunner, PPEs are best used for shuttling data between SPEs and Opterons

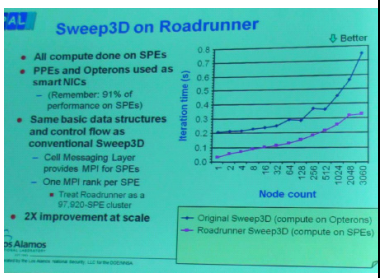

Communication Subsystem Performance
Key contributor to performance of many parallel applications
Complexities of measuring on Roadrunner
Multiple networks: Element Interconnect Bus, FlexIO, PCI Express, HyperTransport, InfiniBand
Different low-level protocols: put/get vs. send/receive
Our approach
Normalize all protocols to send/receive
MPI send/receive for Opteron-Opteron communication
DaCS send/receive for PPE-Opteron communication
Cell Messaging Layer for SPE-SPE communication
Measure ping-pong performance (half round-trip time) across each interconnect type

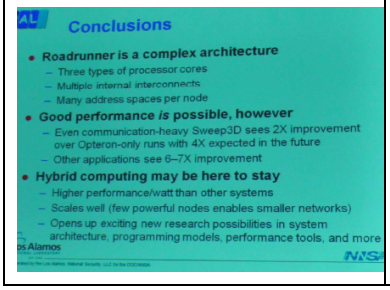

Small-Message Communication Time
SPE
EIB
SPE
0.3µs
SPE
EIB
FlexIO
0.8µs
EIB
SPE
PPE
PCIe
HT2100
3.2µs
HT
Opt
Opt
HT
HT2100
NIC
2.1µs
IB
NIC
HT2100
HT
Opt

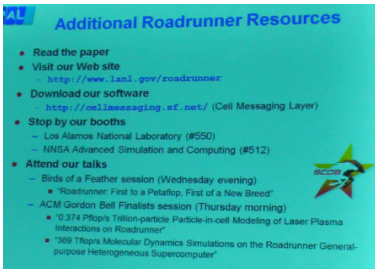

Large-Message Communication Time
Better
PPE–Opteron (same node)
Opteron–Opteron (different nodes)
SPE–SPE (same blade, different Cells)
SPE–SPE (same Cell)
Time (µs)
Message Size (B)
PPE–Opteron is bandwidth bottleneck
Same link technology as Opteron–Opteron (PCI Express x8)
As SW matures, we expect performance to improve to current MPI+IB performance


Early Roadrunner Applications
VPIC
Particle-in-cell code
7X improvement over Opteron-only
SPaSM
Short-range molecular dynamics code
6X improvement over Opteron-only
Milagro
Implicit Monte Carlo code
6X improvement over Opteron-only
PetaVision
Neuron & synapse simulation
1.144 Pflop/s (single prec.)
Each ported to Roadrunner by a couple of people in a short period of time
Months, not years
Most had to learn Cell programming first
All had to deal with preproduction HW & SW
Relatively few code changes
10–30% of code (est.)
→ Yes, Roadrunner is programmable


Challenge: Sweep3D
Neutron-transport kernel
3-D global grid with 2-D data decomposition
Wavefront communication
Receive boundaries from upstream
Compute
Send boundaries downstream
Repeat from each of eight corners
Hard to get performance at scale
Small messages (few KB)
Tightly coupled—runs only as fast as slowest link
Tradeoff between frequency of communication and available parallelism


Sweep3D on Roadrunner
Better
All compute done on SPEs
PPEs and Opterons used as smart NICs
(Remember: 91% of performance on SPEs)
Same basic data structures and control flow as conventional Sweep3D
Cell Messaging Layer provides MPI for SPEs
One MPI rank per SPE
Treat Roadrunner as a 97,920-SPE cluster
2X improvement at scale
Iteration time (s)
Node count
Original Sweep3D (compute on Opterons)
Roadrunner Sweep3D (compute on SPEs)


Conclusions
Roadrunner is a complex architecture
Three types of processor cores
Multiple internal interconnects
Many address spaces per node
Good performance is possible, however
Even communication-heavy Sweep3D sees 2X improvement over Opteron-only runs with 4X expected in the future
Other applications see 6–7X improvement
Hybrid computing may be here to stay
Higher performance/watt than other systems
Scales well (few powerful nodes enables smaller networks)
Opens up exciting new research possibilities in system architecture, programming models, performance tools, and more


Additional Roadrunner Resources
Read the paper
Visit our Web site
http://www.lanl.gov/roadrunner
Download our software
http://cellmessaging.sf.net/ (Cell Messaging Layer)
Stop by our booths
Los Alamos National Laboratory (#550)
NNSA Advanced Simulation and Computing (#512)
Attend our talks
Birds of a Feather session (Wednesday evening)
"Roadrunner: First to a Petaflop, First of a New Breed"
ACM Gordon Bell Finalists session (Thursday morning)
"0.374 Pflop/s Trillion-particle Particle-in-cell Modeling of Laser Plasma Interactions on Roadrunner"
"369 Tflop/s Molecular Dynamics Simulations on the Roadrunner General-purpose Heterogeneous Supercomputer"

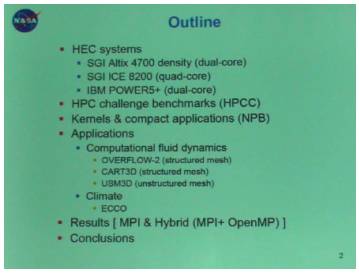
② Papers「Scientific Application-based Performance of SGI Altix 4700, IBM POWER 5+, and SGI Altix ICE 8200 Supercomputers.」by Subhash Saini et. al. (NASA Ames Research Center)

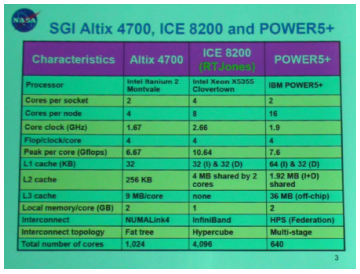
NASA に導入された SGI ICE のアプリケーション性能について、POWER5+や Altix(Itanium 2)と比較結果の発表である。NASA の SGI ICE 8200EX は、今回の TOP500 の 3 位に入賞した、約 0.5 ペタクラスの計算機である。本検討は、このシステムの一部を利用したものと考えられる。同じ数の Core を用いた計算では、POWER5+が最も性能がよく、SGI ICE は、半分程度の性能が出るというものであった。NASA の SGI ICE は、LINPACK 性能は良いものの、1Core あたりのメモリが 1GB であるなど、アプリケーション実行のためには使い勝手が良いとはいえないようである。

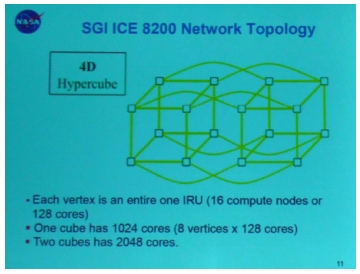
論文の PDF は、http://scyourway.nacse.org/conference/view/pap452。

発表のポイントは次の通りである。

- NASA が所有する 3 種のスパコン（SGI Altix 4700、SGI ICE 8200、Power5+）について、アプリケーションベースで評価した。
- SGI ICE は、低価格な Xeon を用い、Core 数を増やして LINPACK 性能を稼いでいる機種である。NASA に導入された SGI ICE の 1Core あたりのメモリは 1GB である。Core のネットワークトポロジーは 4 次元ハイパーキューブであり、128core×16 = 2048core が 1 つの unit となっている。
- Core の単体性能としては Xeon よりも Power5+や Itanium2(Montvale)がよいという当然の結論である。特に、メモリバンド幅の差が顕著である。
- NASA が利用している 4 つのアプリケーションについて性能評価した結果からは、アプリケーション性能は Power5+が一番である。SGI ICE はメモリバンド幅が小さいために、アプリケーション性能が出ず、Power5+や Altix の半分程度の性能である。なお、512Core までの評価である。
- また、MPI と OpenMP のハイブリッドコードは、Altix や ICE では性能が出せないようである。これらでは、ハイブリッドでは性能が出せず、MPI だけのコードにした方が良いようである。





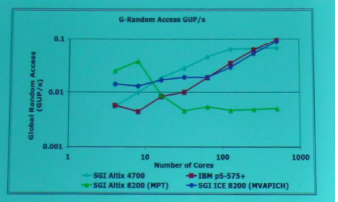


HPC Challenge Benchmarks

- Basically consists of 7 benchmarks
  - HPL: floating-point execution rate for solving a linear system of equations
  - DGEMM: floating-point execution rate of double precision real matrix-matrix multiplication
  - STREAM: sustainable memory bandwidth
  - PTRANS: transfer rate for large data arrays from memory (total network communications capacity)
  - Random Access: rate of random memory integer updates (GUPS)
  - FFTE: floating-point execution rate of double-precision complex 1D discrete FFT
  - Latency/Bandwidth: ping-pong, random & natural ring

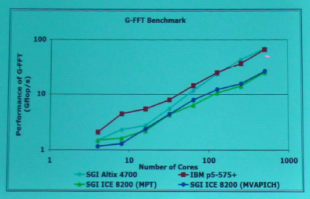
12

G - Random Access Benchmark



17

G - FFTE Benchmark



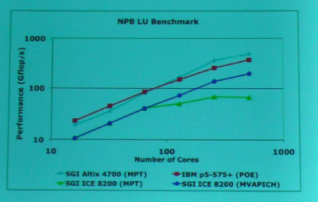
21

NAS Parallel Benchmarks (NPB)

- Kernel benchmarks
  - MG: multi-grid on a sequence of meshes, long- & short-distance communication, memory intensive
  - FT: discrete 3D FFTs, all-to-all communication
  - IS: integer sort, random memory access
  - CG: conjugate gradient, irregular memory access and communication
  - EP: embarrassingly parallel
- Application benchmarks
  - BT: block tri-diagonal solver
  - SP: scalar penta-diagonal solver
  - LU: lower-upper Gauss Seidel

22

NPB Class C LU Benchmark



27

OVERFLOW-2

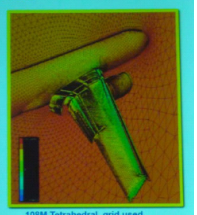
Space shuttle and boosters simulated in Overflow-2

- Code Description:
  - Computational Fluid Dynamics code for the compressible Navier-Stokes equations (NASA Langley)
  - Finite differences in space, implicit time stepping.
  - Handles geometric complexity via overlapping grids
  - Fortran, some C
  - Has been used for multiple projects including Space Shuttle Launch Vehicle, subsonic transport aircraft
- Code Characteristics
  - MPI SPMD model
  - Natural parallelism with multiple overset grids
  - Further grid splitting for load balancing
  - Memory bound
  - Not communication intensive

29

USM3D

- USM3D is an tetrahedral based cell-centered Navier-Stokes flow solver using an unstructured meshes
- It is part of the NASA Tetrahedral Unstructured Software System (TetrUSS) suite
- Routinely used to predict aerodynamic parameters like lift, drag, and detailed airflow about candidate aircraft and aerospace vehicle designs
- **The model is memory bound and latency bound for large number of cores**



108M Tetrahedral grid used in Boeing 777 Simulation

31

CART3D

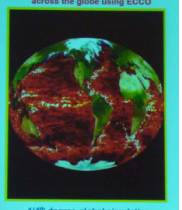
- Inviscid analysis package, Cartesian structured meshes
- Surface modeling, mesh generation, data extraction
- Space-Filling-Curve based partitioner and mesh coarsener
- Each sub-domain has own local grid hierarchy
- **CPU-intensive, memory bound for large multigrid levels**

30

ECCO

Magnitudes of velocity across the globe using ECCO

- Estimating the Circulation and Climate of the Ocean (ECCO) is a global ocean circulation model solving the hydrostatically approximated time dependent Reynolds averaged Navier-Stokes equations in 3D
- Utilizes a finite volume discretization of the equations of motion that is 2nd order accurate in time and space
- **Memory bound.**
- **Scaling is ultimately latency bound for large number of cores**
- **Significant amount of I/O**



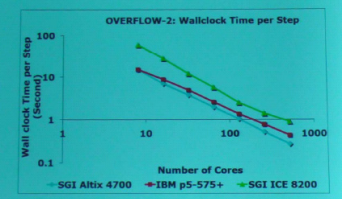
1/4th degree global simulation

32

Applications Characteristics

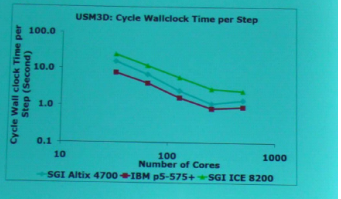
Application	Discipline	Programming Language	Programming Paradigm	Compute Intensity	Memory Intensity	Comm. Intensity	I/O
Cart3D	CFD unstructured Cartesian meshes (24 M)	C & C++	MPI	high	med	low	low
ECCO	Climate (50 M)	Fortran	MPI	high	med	high	high
Overflow-2	CFD Structured meshes (36 M)	Fortran	MPI & Hybrid MPI + OpenMP	high	high	low	low
USM3D	CFD unstructured Tetrahedral meshes (10 M)	Fortran	MPI	high	high	med	low

Total Wallclock Time for OVERFLOW-2



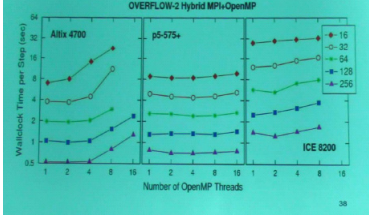
34

Wall Clock Time for USM3D



36

Overflow-2: Hybrid (MPI + OpenMP)



38

Conclusions

Applications performance is best on POWER5+

- Memory subsystem performance is crucial. ICE has lowest memory bandwidth per core and POWER5+ has the highest.
- Network topology is important. Fat tree of Altix 4700 is better than hypercube of ICE 8200.
- Software is important. Performance of MVAPICH is about 10% better than MPT on ICE 8200 and in some cases by 30%
- OpenMP does not perform well. Performance of hybrid codes (MPI+OpenMP) is sometimes same as MPI but usually inferior to pure MPI. On ICE it was lowest in spite of lowest latency within a node. It performs best on the POWER5+.
- Memory per core should be at least 2 GB. ICE has 1 GB/core
- For consistently good performance, a balance between the performance of core, memory and interconnect is needed

Price has not been considered in this study, which is an extremely important consideration

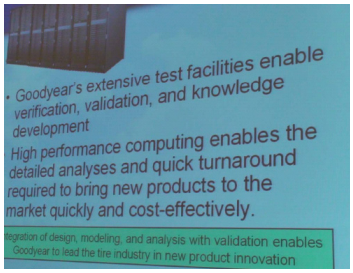
39

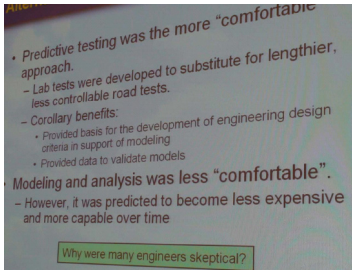
③ Masterworks「Enabling Analysis-based Tire Development with High Performance Computing」by Loren Miller (Goodyear Tire & Rubber Company)

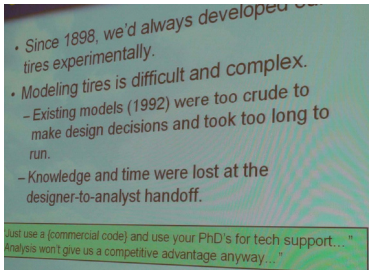
Goodyear が、1993 年より継続しているエネルギー省の CRADA プロジェクトの下で、有限要素法(FEM)によるタイヤのトレッドパターンの大規模シミュレーション手法を、Sandia 研究所との共同研究として開発した経緯に関する発表である。企業単独では解決できなかった課題を、Sandia のコードを用いることで解決し、実際の製品開発に大きく役立っていることはすばらしいと感じた。

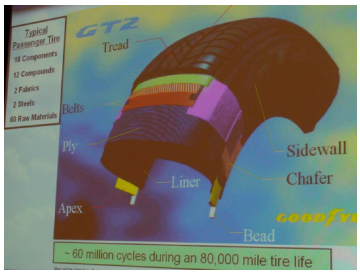
発表のポイントは次の通りである。

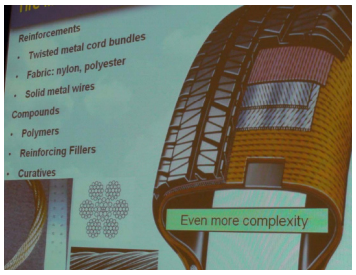
- タイヤの開発において、燃費性能と安全性の向上が重要な課題である。
- Goodyear は、HPC を、詳細に解析する手段であるとともに、低コストで短期間に市場に製品を送り出す手段として捕らえている。
- タイヤのモデリングは、困難で複雑である。タイヤは、60 種類の材料から多数のコンポーネットから構成されている。特に、ゴム材料は、高分子や補強のためのフィラーなどからなり、Mullin's effect、Payne effect など複雑な性質を示し、扱うのが難しい。
- 1994 年に Goodyear で計算できる FEM モデルの自由度の最大値は 90,000 自由度であった。タイヤトレッドのモデルに必要な最小のサイズは 250,000 自由度と見積もられ、手が出る状況ではなかった。そこで、1993 年より、Sandia 研究所との共同研究で、エネルギー省の CRADA(Cooperative Research and Development Agreement)プロジェクトを実施した。Sandia 研究所が作成したソルバーを用いることで、計算コストを大幅に減らすことに成功し、大規模な FEM 計算が可能になった。
- タイヤトレッドの FEM のシミュレーションにより、タイヤ製品の開発期間を 3 年から 1 年に短縮することに成功した。
- 現在では、128 ノードの Linux クラスターを利用しており、RunFlat タイヤの開発で、非常に活躍している。

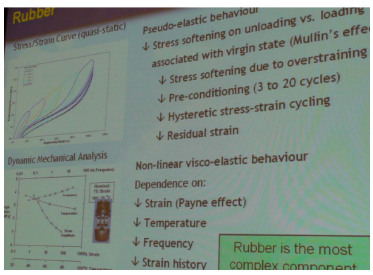


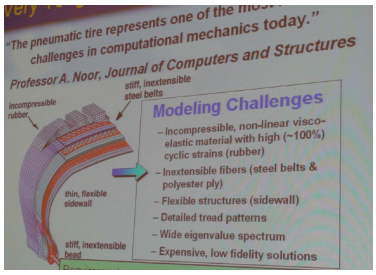

Predictive testing was the more "comfortable" approach.
– Lab tests were developed to substitute for lengthier, less controllable road tests.
– Corollary benefits:
• Provided basis for the development of engineering design criteria in support of modeling
• Provided data to validate models
Modeling and analysis was less "comfortable".
– However, it was predicted to become less expensive and more capable over time
Why were many engineers skeptical?

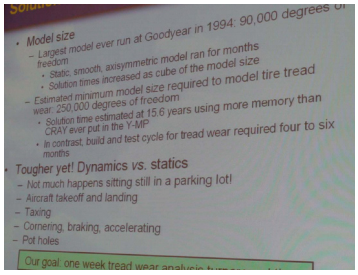

Since 1898, we'd always developed
tires experimentally.
Modeling tires is difficult and complex.
– Existing models (1992) were too crude to make design decisions and took too long to run.
– Knowledge and time were lost at the designer-to-analyst handoff.
Just use a (commercial code) and use your PhD's for tech support..."
Analysis won't give us a competitive advantage anyway..."

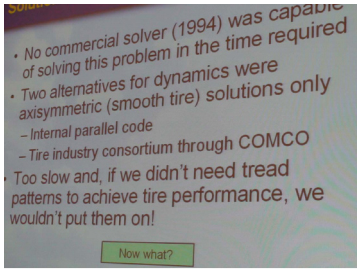

GT2
Typical Passenger Tire
18 Components
12 Compounds
2 Fabrics
2 Steels
60 Raw Materials
Tread
Belts
Ply
Liner
Apex
Sidewall
Chafer
Bead
~ 60 million cycles during an 80,000 mile tire life

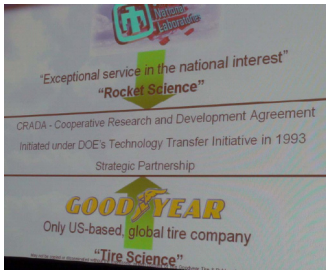

Reinforcements
Twisted metal cord bundles
Fabric: nylon, polyester
Solid metal wires
Compounds
Polymers
Reinforcing Fillers
Curatives
Even more complexity

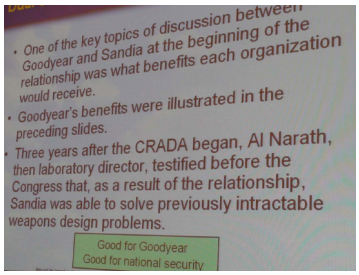

Rubber
Stress/Strain Curve (quasi-static)
Dynamic Mechanical Analysis
Pseudo-elastic behaviour
↓ Stress softening on unloading vs. loadin
associated with virgin state (Mullin's effe
↓ Stress softening due to overstraining
↓ Pre-conditioning (3 to 20 cycles)
↓ Hysteretic stress-strain cycling
↓ Residual strain
Non-linear visco-elastic behaviour
Dependence on:
↓ Strain (Payne effect)
↓ Temperature
↓ Frequency
↓ Strain history
Rubber is the most complex component

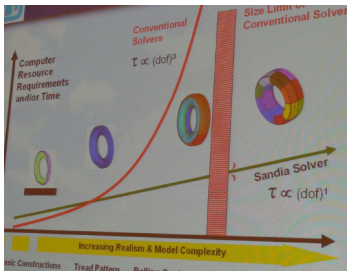

"The pneumatic tire represents one of the most challenges in computational mechanics today."
Professor A. Noor, Journal of Computers and Structures
incompressible rubber
stiff, inextensible steel belts
thin, flexible sidewall
stiff, inextensible bead
Modeling Challenges
– Incompressible, non-linear visco-elastic material with high (~100%) cyclic strains (rubber)
– Inextensible fibers (steel belts & polyester ply)
– Flexible structures (sidewall)
– Detailed tread patterns
– Wide eigenvalue spectrum
– Expensive, low fidelity solutions

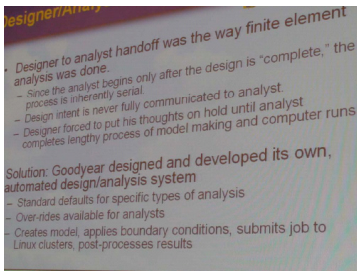

Model size
– Largest model ever run at Goodyear in 1994: 90,000 degrees of freedom
• Static, smooth, axisymmetric model ran for months
• Solution times increased as cube of the model size
– Estimated minimum model size required to model tire tread wear: 250,000 degrees of freedom
• Solution time estimated at 15.6 years using more memory than CRAY ever put in the Y-MP
• In contrast, build and test cycle for tread wear required four to six months
Tougher yet! Dynamics vs. statics
– Not much happens sitting still in a parking lot!
– Aircraft takeoff and landing
– Taxing
– Cornering, braking, accelerating
– Pot holes

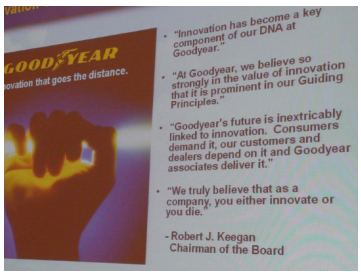

No commercial solver (1994) was capable of solving this problem in the time required
Two alternatives for dynamics were axisymmetric (smooth tire) solutions only
– Internal parallel code
– Tire industry consortium through COMCO
Too slow and, if we didn't need tread patterns to achieve tire performance, we wouldn't put them on!
Now what?


"Exceptional service in the national interest"
"Rocket Science"
CRADA - Cooperative Research and Development Agreement
Initiated under DOE's Technology Transfer Initiative in 1993
Strategic Partnership
GOODYEAR
Only US-based, global tire company
"Tire Science"


One of the key topics of discussion between Goodyear and Sandia at the beginning of the relationship was what benefits each organization would receive.
Goodyear's benefits were illustrated in the preceding slides.
Three years after the CRADA began, Al Narath, then laboratory director, testified before the Congress that, as a result of the relationship, Sandia was able to solve previously intractable weapons design problems.
Good for Goodyear
Good for national security


Conventional Solvers
τ ∝ (dof)³
Size Limit of Conventional Solver
Computer Resource Requirements and/or Time
Sandia Solver
τ ∝ (dof)¹
Increasing Realism & Model Complexity


Designer to analyst handoff was the way finite element analysis was done.
– Since the analyst begins only after the design is "complete," the process is inherently serial.
– Design intent is never fully communicated to analyst.
– Designer forced to put his thoughts on hold until analyst completes lengthy process of model making and computer runs
Solution: Goodyear designed and developed its own, automated design/analysis system
– Standard defaults for specific types of analysis
– Over-rides available for analysts
– Creates model, applies boundary conditions, submits job to Linux clusters, post-processes results


GOODYEAR
novation that goes the distance.
"Innovation has become a key component of our DNA at Goodyear."
"At Goodyear, we believe so strongly in the value of innovation that it is prominent in our Guiding Principles."
"Goodyear's future is inextricably linked to innovation. Consumers demand it, our customers and dealers depend on it and Goodyear associates deliver it."
"We truly believe that as a company, you either innovate or you die."
- Robert J. Keegan
Chairman of the Board

④ Posters「Performance of Electronic Structure Calculations on Blue Gene and Cray XT4 Computers」by Andrey Asadchev et. al. (Iowa State Univ. & Ames Lab.)

GAMESS の中にある 2 次の摂動で電子相関を簡便に見積もる方法である MP2 法のチューニングについての発表である。このチューニングでは、BLAS2 と BLAS3 を駆使して行列変換を行うことがポイントである。20 アミノ酸からなる TRP タンパクについて、425 の active 軌道と 1107 の仮想軌道を用い計算した結果である。変換行列に 1.25TB のメモリが必要である。結果的には、128Core の Blue Gene で、64Core の Cray XT4 と同じ程度の計算時間で計算可能になったとのことであった。まだ改良の余地があり論文未発表とのことである。発表者に、Cray と Blue Gene のどちらの方がアプリケーション性能を出せるか質問したところ、当然 Cray であるという回答であった。

⑤ Posters「Caching on a Chip Multi Vector Processor」by Akihiro Musa et. al. (東北大学 & NEC)

現状のベクトル CPU はマルチコア化されていない上、キャッシュが無いので、メモリバンド幅が勝負である。SX-10 あたりでは、マルチコア化は必須である。SX-6 相当の性能のベクトル CPU 4 つとメインメモリの間に Shared Vector Cache をつけてマルチコア化した試作品を作成し、東北大の既存のマルチスレッド最適済プログラム（ベクトル向き）を用い、メモリバンド幅の重要性を検証した。スケーラビリティは、バンド幅に依存していることが明確になった。Shared Vector Cache に関しては、Non-blocking cache と 32 sub-cache の機能だけでは不十分で、現状の SX-9 のベクトルキャッシュには実装されていない Miss Status Handling Register がないと、キャッシュヒット率が上がらないとのことであった。

⑥ Posters「Storage Performance Analysis in ES and ES2」by Ken'ichi Itakura et. al. (JAMSTEC)

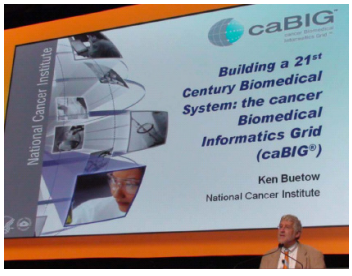
次期地球シミュレータ(ES2)のディスクシステムにおいて、Stage IN と Stage OUT の制御について検討した研究である。地球シミュレータでは、計算時の I/O 性能を高めるために、バッチジョブの実行前にローカルディスクにファイル転送（Stage IN）し、バッチ終了後に出力ファイルを転送(Stage OUT)する機能を用いている。640 ノードの ES では、160 ノードを 1Gbps で 1 台の交換機に接続し、その交換機 4 台を 2Gbps でディスクシステムに接続している。160 ノードからなる ES2 では、16 ノードを 4Gbps で 1 台の交換機に接続し、その交換機 10 台を 4Gbps でディスクシステムに接続している。この研究では、ある 1 日の実際に地球シミュレータで流れたジョブのデータサイズと計算時間を用いて、Stage IN と Stage OUT にかかる時間について検討している。ES の運用データを分析すると、Stage OUT の時間については、ばらつきがあり、極端に遅いものがあった。原因は、Stage IN が優先的に帯域を使うためである。計算時間が半分で、データの量が同じと仮定する場合、Stage IN の時間は 37418sec から 9076sec に、Stage OUT の時間は 782sec から 94sec に削減することが分かったとのことである。データ量の増加が想定されること、Stage IN の優先度設定についての最適値の検討などが今後の課題であり、3 か月分のデータを用いてさらに検討を深めるとのことであった。

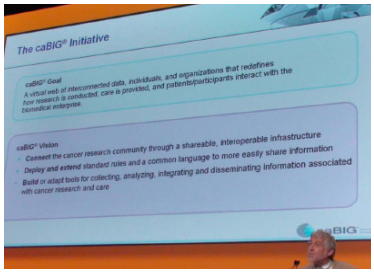
⑦ Invited Speakers「Developing an interoperable IT Framework to Enable Personalized Medicine」by Ken Bertow (National Cancer Institute)

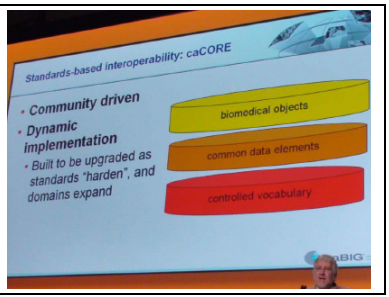
癌患者に関する医療情報や解析ツールをがんセンターの間で共有する caBIG という grid システムについての紹介が行われた。実際にどのような成果が上がっているかの統計については紹介が無く、どこまで実用に至っているのかの情報を得ることはできなかった。

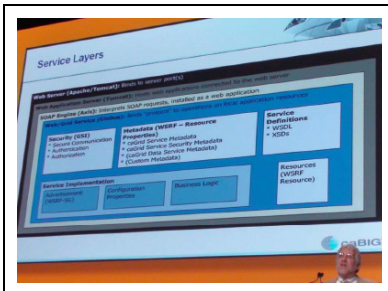
発表のポイントは次の通りである。

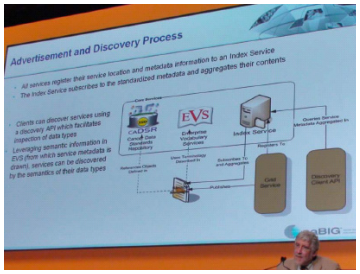
- 21 世紀の創薬としては、遺伝情報などに基づくテーラーメード創薬を目指す考えである。
- バイオメディカルの分野においては、情報が孤立していることが問題であった。臨床／画像診断／病理／分子バイオロジーの 4 分野の連携を促すために、caBIG という情報交換の枠組みとして、各分野の情報の API を整理して Grid ベースの Web サービスを構築した。
- caBIG は、56 の NCI のがんセンターと 16 の地域のがんセンターをつないでいる。National Health Information Network との接続や、英国、インド、中国、ラテンアメリカとの連携も予定している。
- caGrid、NHLBI Grid、NCTI ONIX の 3 つのグリッド情報は、相互に情報交換を行い、Grid のグリッドを実現している。
- ユーザーに対しては、single web のポータルサイトを構築している。Gene 情報、Genome 情報、臨床情報および、解析ツールを提供している。
- テーラーメード創薬のために、BIG Health Consortium を設立した。政府、企業、個人を結んでいる。

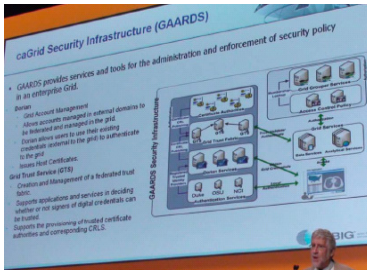


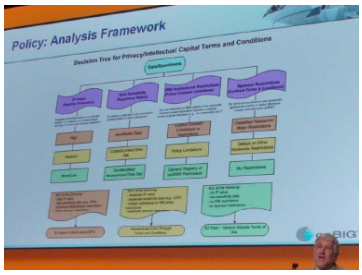


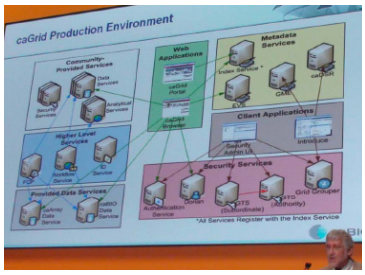


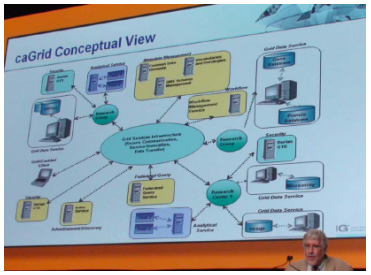

Service Layers

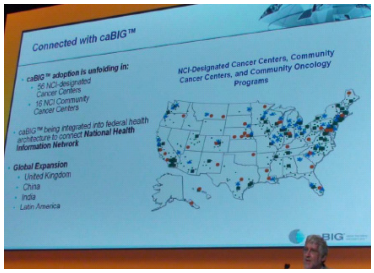

Advertisement and Discovery Process

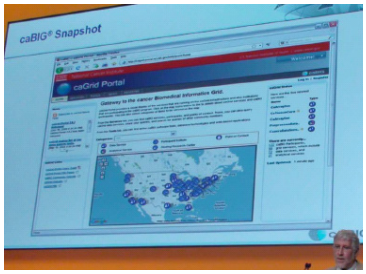

caGrid Security Infrastructure (GAARDS)

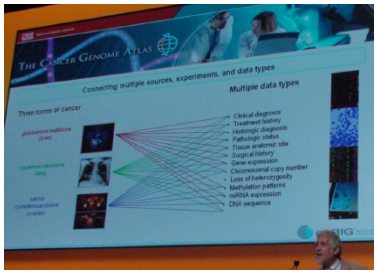

Policy: Analysis Framework

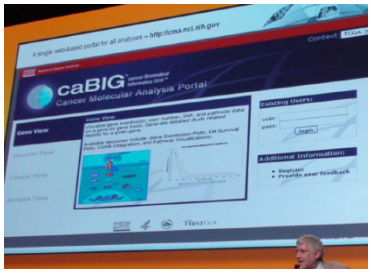

caGrid Production Environment


caGrid Conceptual View

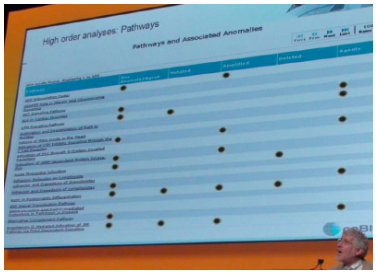

Connected with caBIG™

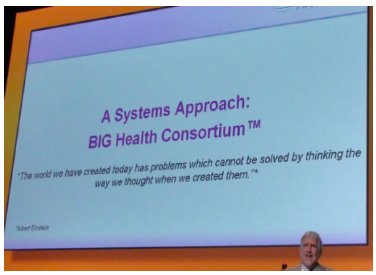

caBIG® Snapshot

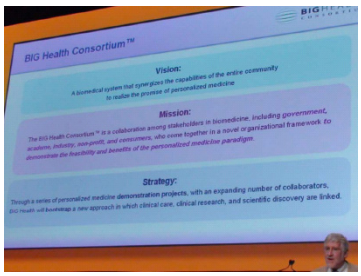

The Cancer Genome Atlas

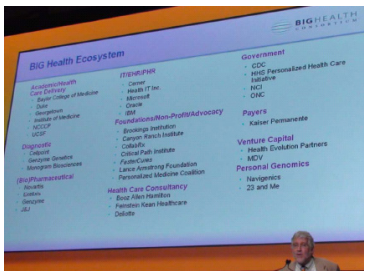

caBIG
Cancer Molecular Analysis Portal


3D Structure Viewer


High order analyses: Pathways


A Systems Approach:
BIG Health Consortium™


BIG Health Consortium™

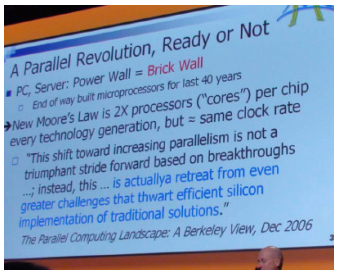

BIG Health Ecosystem

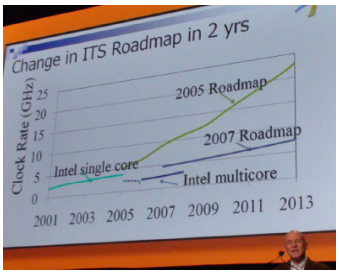
⑧ Invited Speakers「Parallel Computing Landscape: A View from Berkley」
by Dave Patterson (U.C. Barkeley)

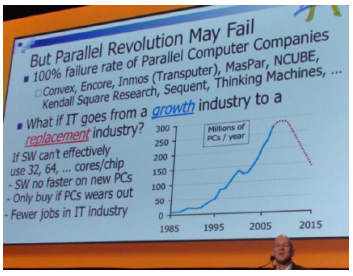
UC バークレーに設立された新たな並列ソフトウェアの研究センターの紹介であった。CPU の単体性能の成長は止まり、CPU 数の増加で性能を上げていることから、FPGA を用いて 90MHz の RICS チップ 1008Core の並列環境を構築し、並列ソフトウェアの研究を行っているとのことである。

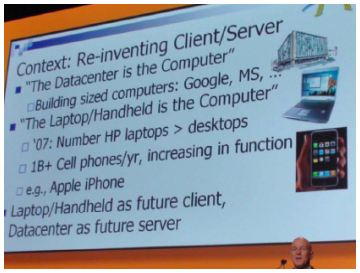
発表のポイントは次の通りである。

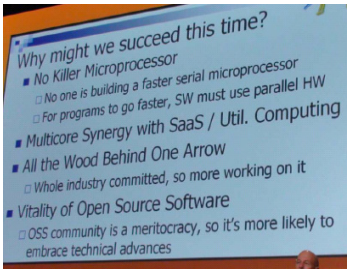
- CPU や半導体のロードマップは、2005 年に比べ、2007 年では大幅に下方修正されている。キラーCPU の登場は期待することはできない。プログラムを早くするには、並列ハードウェアの利用が必須である。今後、IT 産業は成長産業から、置き換え産業になると予想される。
- コンピュータというものは、Google などのデータセンターがサーバとして、Laptop や Handheld が端末として利用されるものとなってきている。
- 100 以上の core に対して、生産的かつ経済的、ポータブルなソフトウェアの開発を行うために、新しいセンターを設立した。
- アプリケーションとしては、多数の CPU を利用し高性能な計算が要求される音楽、イメージ検索、冠状動脈障害、パラレルブラウザなどを扱うこととした。必要な計算要素に分解し、最適化を検討した。
- パラレルソフトウェアの開発では、効率の追求と生産性の追求を分離した。効率を追求するのはプログラマーの 10%であり、残りの 90%は生産性を追及しているからである。
- FPGA 上に 12 個の 32-bit RICS Core (90MHz)を構築し、84 個の FPGA を接続した 1008 core の"RAMP Blue"という試験用計算機を構築した。
- 今後、電源やパフォーマンスのボトルネック解析を行い、最適化を図っていく考えである。

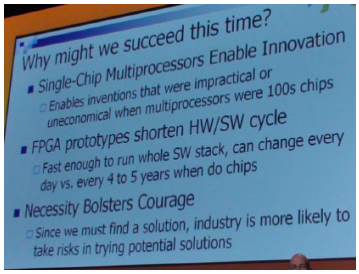


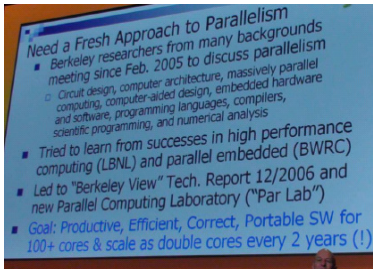


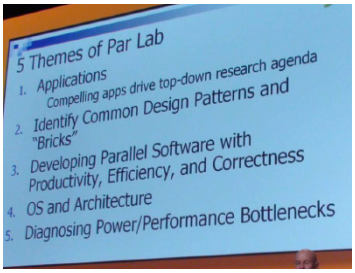


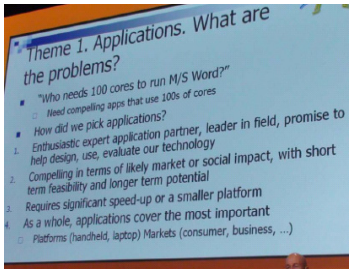

Context: Re-inventing Client/Server
"The Datacenter is the Computer"
Building sized computers: Google, MS, ...
"The Laptop/Handheld is the Computer"
'07: Number HP laptops > desktops
1B+ Cell phones/yr, increasing in function
e.g., Apple iPhone
Laptop/Handheld as future client,
Datacenter as future server

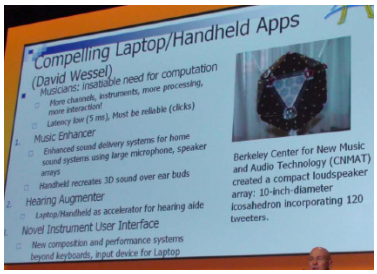

Why might we succeed this time?
No Killer Microprocessor
No one is building a faster serial microprocessor
For programs to go faster, SW must use parallel HW
Multicore Synergy with SaaS / Util. Computing
All the Wood Behind One Arrow
Whole industry committed, so more working on it
Vitality of Open Source Software
OSS community is a meritocracy, so it's more likely to embrace technical advances

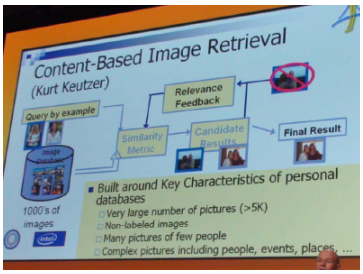

Why might we succeed this time?
Single-Chip Multiprocessors Enable Innovation
Enables inventions that were impractical or uneconomical when multiprocessors were 100s chips
FPGA prototypes shorten HW/SW cycle
Fast enough to run whole SW stack, can change every day vs. every 4 to 5 years when do chips
Necessity Bolsters Courage
Since we must find a solution, industry is more likely to take risks in trying potential solutions

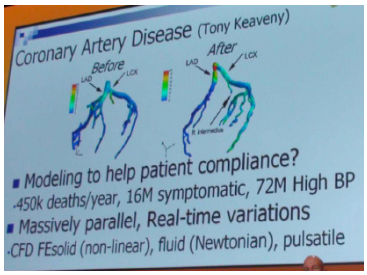

Need a Fresh Approach to Parallelism
Berkeley researchers from many backgrounds meeting since Feb. 2005 to discuss parallelism
Circuit design, computer architecture, massively parallel computing, computer-aided design, embedded hardware and software, programming languages, compilers, scientific programming, and numerical analysis
Tried to learn from successes in high performance computing (LBNL) and parallel embedded (BWRC)
Led to "Berkeley View" Tech. Report 12/2006 and new Parallel Computing Laboratory ("Par Lab")
Goal: Productive, Efficient, Correct, Portable SW for 100+ cores & scale as double cores every 2 years (!)

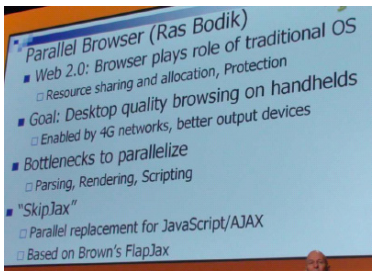

5 Themes of Par Lab
1. Applications
Compelling apps drive top-down research agenda
2. Identify Common Design Patterns and "Bricks"
3. Developing Parallel Software with Productivity, Efficiency, and Correctness
4. OS and Architecture
5. Diagnosing Power/Performance Bottlenecks

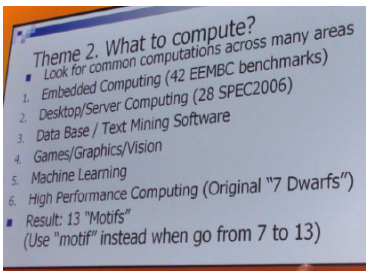

Theme 1. Applications. What are the problems?
"Who needs 100 cores to run M/S Word?"
Need compelling apps that use 100s of cores
How did we pick applications?
1. Enthusiastic expert application partner, leader in field, promise to help design, use, evaluate our technology
2. Compelling in terms of likely market or social impact, with short term feasibility and longer term potential
3. Requires significant speed-up or a smaller platform
4. As a whole, applications cover the most important
Platforms (handheld, laptop) Markets (consumer, business, ...)

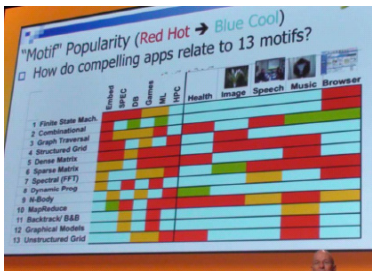

Compelling Laptop/Handheld Apps
(David Wessel)
Musicians: Insatiable need for computation
More channels, instruments, more processing, more interaction!
Latency low (5 ms), Must be reliable (clicks)
1. Music Enhancer
Enhanced sound delivery systems for home sound systems using large microphone, speaker arrays
Handheld recreates 3D sound over ear buds
2. Hearing Augmenter
Laptop/Handheld as accelerator for hearing aide
3. Novel Instrument User Interface
New composition and performance systems beyond keyboards, input device for Laptop
Berkeley Center for New Music and Audio Technology (CNMAT) created a compact loudspeaker array: 10-inch-diameter icosahedron incorporating 120 tweeters.

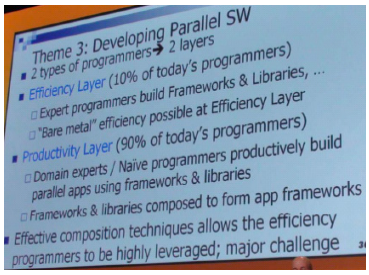

Content-Based Image Retrieval
(Kurt Keutzer)
Query by example
Relevance Feedback
Similarity Metric
Candidate Results
Final Result
1000's of images
Built around Key Characteristics of personal databases
Very large number of pictures (>5K)
Non-labeled images
Many pictures of few people
Complex pictures including people, events, places, ...

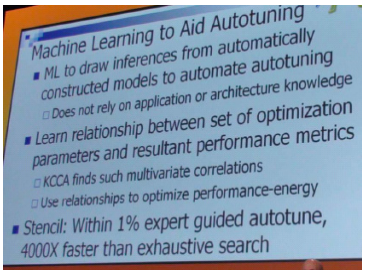

Coronary Artery Disease (Tony Keaveny)
Before
After
Modeling to help patient compliance?
450k deaths/year, 16M symptomatic, 72M High BP
Massively parallel, Real-time variations
CFD FEsolid (non-linear), fluid (Newtonian), pulsatile

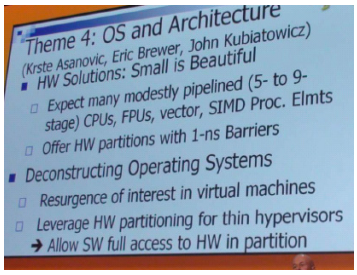

Parallel Browser (Ras Bodik)
Web 2.0: Browser plays role of traditional OS
Resource sharing and allocation, Protection
Goal: Desktop quality browsing on handhelds
Enabled by 4G networks, better output devices
Bottlenecks to parallelize
Parsing, Rendering, Scripting
"SkipJax"
Parallel replacement for JavaScript/AJAX
Based on Brown's FlapJax

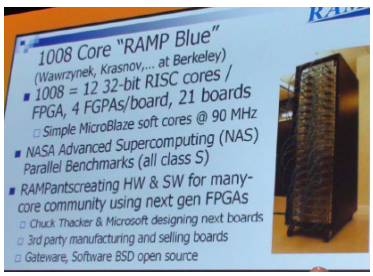

Theme 2. What to compute?
Look for common computations across many areas
1. Embedded Computing (42 EEMBC benchmarks)
2. Desktop/Server Computing (28 SPEC2006)
3. Data Base / Text Mining Software
4. Games/Graphics/Vision
5. Machine Learning
6. High Performance Computing (Original "7 Dwarfs")
Result: 13 "Motifs"
(Use "motif" instead when go from 7 to 13)

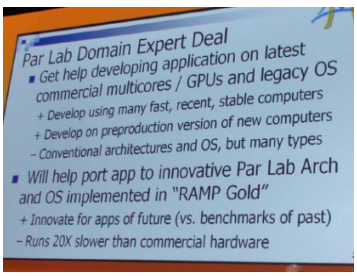

"Motif" Popularity (Red Hot → Blue Cool)
How do compelling apps relate to 13 motifs?
Health
Image
Speech
Music
Browser
1 Finite State Mach.
2 Combinational
3 Graph Traversal
4 Structured Grid
5 Dense Matrix
6 Sparse Matrix
7 Spectral (FFT)
8 Dynamic Prog
9 N-Body
10 MapReduce
11 Backtrack/ B&B
12 Graphical Models
13 Unstructured Grid

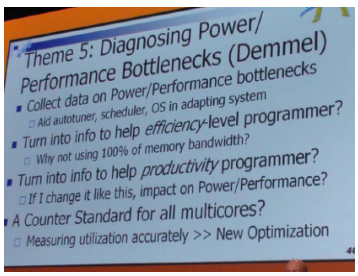

Theme 3: Developing Parallel SW
2 types of programmers → 2 layers
Efficiency Layer (10% of today's programmers)
Expert programmers build Frameworks & Libraries, ...
"Bare metal" efficiency possible at Efficiency Layer
Productivity Layer (90% of today's programmers)
Domain experts / Naïve programmers productively build parallel apps using frameworks & libraries
Frameworks & libraries composed to form app frameworks
Effective composition techniques allows the efficiency programmers to be highly leveraged; major challenge


Machine Learning to Aid Autotuning
ML to draw inferences from automatically constructed models to automate autotuning
Does not rely on application or architecture knowledge
Learn relationship between set of optimization parameters and resultant performance metrics
KCCA finds such multivariate correlations
Use relationships to optimize performance-energy
Stencil: Within 1% expert guided autotune, 4000X faster than exhaustive search


Theme 4: OS and Architecture
(Krste Asanovic, Eric Brewer, John Kubiatowicz)
HW Solutions: Small is Beautiful
Expect many modestly pipelined (5- to 9-stage) CPUs, FPUs, vector, SIMD Proc. Elmts
Offer HW partitions with 1-ns Barriers
Deconstructing Operating Systems
Resurgence of interest in virtual machines
Leverage HW partitioning for thin hypervisors
→ Allow SW full access to HW in partition

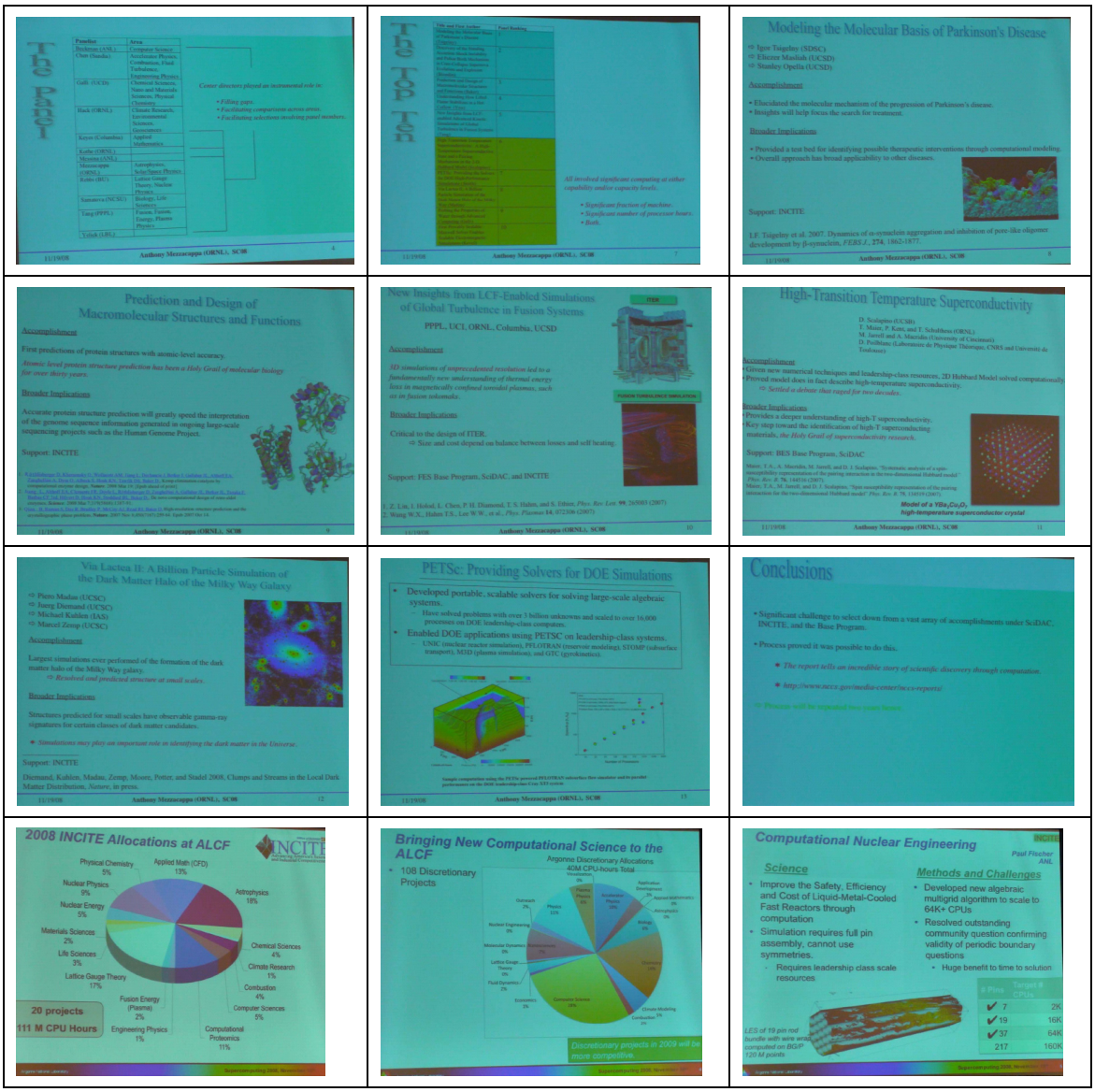

1008 Core "RAMP Blue"
(Wawrzynek, Krasnov,... at Berkeley)
1008 = 12 32-bit RISC cores / FPGA, 4 FGPAs/board, 21 boards
Simple MicroBlaze soft cores @ 90 MHz
NASA Advanced Supercomputing (NAS) Parallel Benchmarks (all class S)
RAMPantscreating HW & SW for many-core community using next gen FPGAs
Chuck Thacker & Microsoft designing next boards
3rd party manufacturing and selling boards
Gateware, Software BSD open source


Par Lab Domain Expert Deal
Get help developing application on latest commercial multicores / GPUs and legacy OS
+ Develop using many fast, recent, stable computers
+ Develop on preproduction version of new computers
– Conventional architectures and OS, but many types
Will help port app to innovative Par Lab Arch and OS implemented in "RAMP Gold"
+ Innovate for apps of future (vs. benchmarks of past)
– Runs 20X slower than commercial hardware


Theme 5: Diagnosing Power/ Performance Bottlenecks (Demmel)
Collect data on Power/Performance bottlenecks
Aid autotuner, scheduler, OS in adapting system
Turn info into to help efficiency-level programmer?
Why not using 100% of memory bandwidth?
Turn into info to help productivity programmer?
If I change it like this, impact on Power/Performance?
A Counter Standard for all multicores?
Measuring utilization accurately >> New Optimization

⑨ BOF「Petascale Computing Experiences on Blue Gene/P」

実質的には、DOE の INCITE(Innovative and Novel Computational Impact on Theory and Experiment)プログラムなどの下での Blue Gene を用いた計算事例の紹介のみであった。会場から具体的なチューニングの状況について興味がある旨の声が出ていたが、技術的な部分には対応していない発表であった。オークリッジ研究所においては、INCITE プログラムの下で、たんぱく質の MD 計算、ITER の計算、ダークマターに関する粒子系の計算、大規模代数計算などが行われたとのことであった。アルゴンヌ研究所では、CFD や原子力技術をはじめ、20 のプロジェクトに計算資源をアサインしているとのことであった。



⑩ BOF「User Experiences with Large SGI Altix ICE」

SGI ICE は Xeon ベースのクラスターシステムである。TOP500 の 25 位以内としては、NASA(3 位)、ニューメキシコ計算センター(12 位)、フランス GENCI(14 位)、Total 石油調査(17 位)が導入しているスパコンである。

1 Core あたりのメモリは、2～4GB がトレンドである。2GB で導入したところは、少し苦労しているような感じであった。量子化学計算のように、メモリの大きさがものを言う場合には向いていないという感じであった。安物は安物で苦労しなくてはいけないという感覚は共通のようである。

フランス GENCI のシステムは､32 ラック中 24 ラックを計算用に利用している。設置面積は 63m² で、消費電力は 640kW である。1/4ITER トーラスのシミュレーションでは、1360 億点を 4096Core 上で計算するとのことである。次期 IPCC に利用する計算では、864Core を用い、28km メッシュの 1442×1021×46 点について 300 年間分の計算を行うとのことである。現在、グランドチャレンジとして、燃焼シミュレーションを 8192Core で進めており、12000Core まで引き上げるとのことである。

SGI ICE の利点は、gpgpu や clearspeed を導入するハイブリッド構成が可能なことや、効率、安定性がよいことである。また、ユーザーの満足度は高い。なお、IPoIB のマイナーなバグがあったりするが、改良されるであろう。

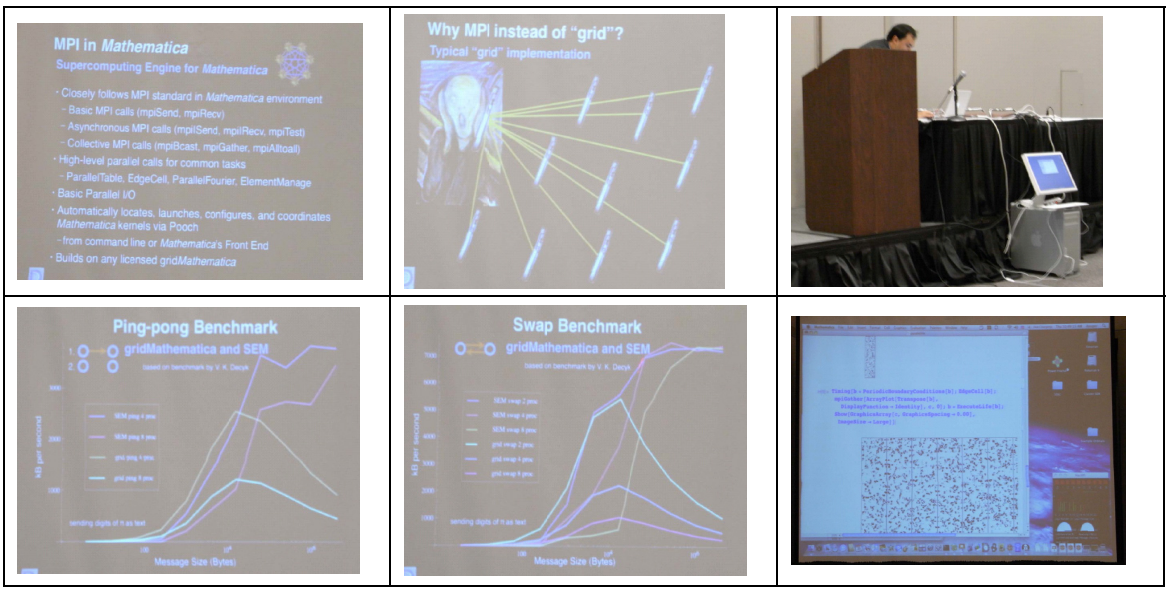
実用の計算では、チープシステムにおいてはファイルの I/O 周りが厳しくなりがちで、そのあたりは仕方ないと考えている様子であった。

⑪ Exhibiter Events「Supercomputing Engine for Mathematica」by D. E. Dauger (Dauger Research Inc.)

Matheamtica 上で MPI を利用できるための環境構築に関する発表であった。MPI プログラミングを Mathematica でそのまま行うことができるライブラリとみなすことができる。Mathematica 上で大規模計算を通常しないとは考えられるが、初学者にとって役立つこともあることから今後の発展が期待される。

発表のポイントは次の通りである。

- 既存製品の gridMathematica では、データが大きくなると使い物にならなくなるとのことである。grid タイプの計算では、データを 1 つの core に集約することとなるため、マスターノードで処理できなくなるためである。
- その対応として､MPI による Domain Decomposition の計算を実現すべく、MPI の Send、Recv、Gather などの機能を備えた環境を整備した。
- gridMathematica のライセンス上で動作するように構築されている。
- 具体的には、Pooch という Dauger Research の MPI に関連するソフトを用いて複数の Mac でクラスターを構成した環境において、Mathematica の FrontEnd から MPI と同じような機能をするコマンドを発行することで MPI 並列計算を行うことができる。
- セルオートマトンやプラズマダイナミクスのシミュレーションについて、領域分割した計算が可能である。
- gridMathematica と比較し、実際に大規模な計算が可能である。



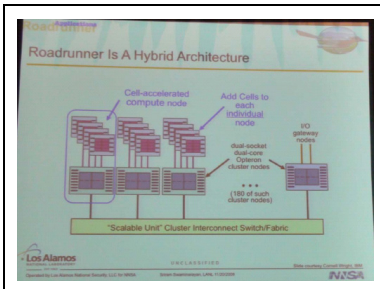
⑫ ACM Gordon Bell Finalist「369 Tflop/s Molecular Dynamics Simulations on the Roadrunner General-purpose Heterogeneous Supercomputer」by Sriram Swaminarayan et. al. (Los Alamos)

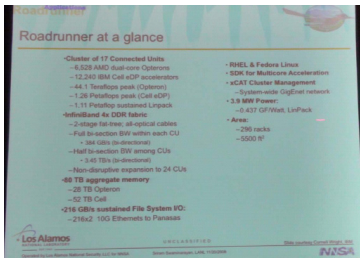
Roadrunner 上で MD コードを最適化した研究の発表であった。Cell の MD プログラムのチューニングをしっかりやっている印象を受けた。また、個人的には、Cell-SPE と Opteron の間の通信が多いような気がしたが、実際にやってみなと判断がつかないと思われる。

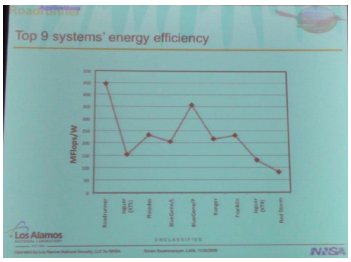
本発表の論文の PDF は、http://scyourway.nacse.org/conference/view/gb124 を通じ IEEE のアーカイブや ACM のデジタル図書館から取得できる。

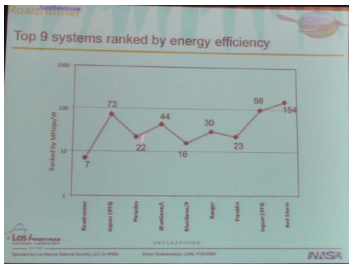
発表のポイントは次の通りである。

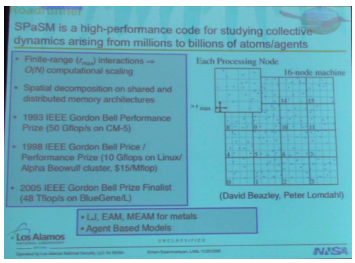
- Roadrunner は、17 のユニットに、6528 個の Dual-core Opteron と 12240 個の Cell を搭載した計算機である。Opteron のピークは 44.1 テラ flops で、Cell が 1.26 ペタ flops である。ユニット内では 384GB/s の InifiniBand で接続され、ユニット間は 3.45TB/s の InfiniBand で、24 ユニットまで接続できる。ファイルシステムとは、216GB/s の I/O が可能である。
- Roadrunner は Mflops/W の値が圧倒的に低い、エコなシステムである。
- SPaSM(Scalable Parallel Short-range Molecular dynamics)は、1990 年代にロスアラモスで開発された領域分割型の並列 MD コードである。2005 年に Blue Gene/L 上で動作するように移植されている。SPaSM は、金属の計算で利用されるとともに、感染症のエージェント計算にも用いられている。
- 原子間の力の計算では、領域分割を用いるのが、(当然) 最も効率的である。
- スタンドアロンの Cell での MD は、Opteron の 3.5 倍の速度であった。
- Cell は、ベクトル向けのデータ構造である必要があり、データ構造の見直しを行った。SPaSM の通信は、1timestep でプロセスあたり 50000 メッセージとなることがある。今回は、6 までに通信を減らした。このようにしても、通信とデータの入れ替えに 20%の時間を使っている。
- Opteron は通信を担当し、Cell の SPU で他の処理を行った。
- 結果としては、Opteron の 6～9 倍まで高速化することができた。
- 平衡系の問題やエージェントシミュレーションなどでは、単精度にすることが可能である。その場合さらに、多くの flops 値を稼ぐことができる。8 ユニットの場合、倍精度では 195 テラ flops であるが、単精度では 471 テラ flops となる。

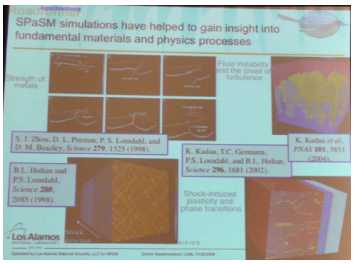

Roadrunner Is A Hybrid Architecture
Cell-accelerated compute node
Add Cells to each individual node
I/O gateway nodes
dual-socket dual-core Opteron cluster nodes
(180 of such cluster nodes)
"Scalable Unit" Cluster Interconnect Switch/Fabric
Los Alamos
UNCLASSIFIED
NNSA

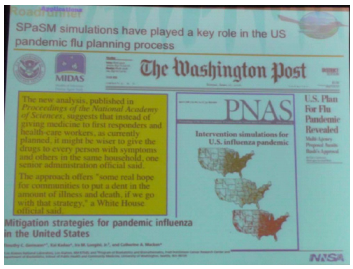

Roadrunner at a glance
Cluster of 17 Connected Units
~6,528 AMD dual-core Opterons
~12,240 IBM Cell eDP accelerators
~44.1 Teraflops peak (Opteron)
~1.26 Petaflops peak (Cell eDP)
~1.11 Petaflop sustained Linpack
InfiniBand 4x DDR fabric
2-stage fat-tree; all-optical cables
Full bi-section BW within each CU
~384 GB/s (bi-directional)
Half bi-section BW among CUs
~3.45 TB/s (bi-directional)
Non-disruptive expansion to 24 CUs
80 TB aggregate memory
~28 TB Opteron
~52 TB Cell
216 GB/s sustained File System I/O:
216x2 10G Ethernets to Panasas
RHEL & Fedora Linux
SDK for Multicore Acceleration
xCAT Cluster Management
System-wide GigEnet network
3.9 MW Power:
~0.437 GF/Watt, LinPack
Area:
~296 racks
~5500 ft²
Los Alamos
UNCLASSIFIED
NNSA

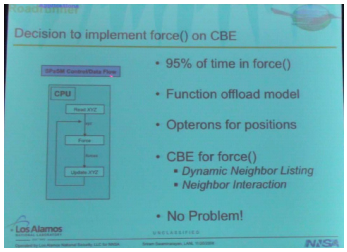

Top 9 systems' energy efficiency
Mflops/W
Los Alamos
UNCLASSIFIED
NNSA

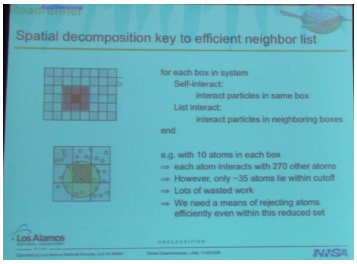

Top 9 systems ranked by energy efficiency
Ranked by Mflops/W
73
22
44
16
30
23
98
154
7
Los Alamos
UNCLASSIFIED
NNSA

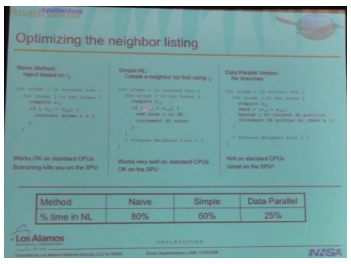

SPaSM is a high-performance code for studying collective dynamics arising from millions to billions of atoms/agents
Finite-range (r_max) interactions ⇒ O(N) computational scaling
Spatial decomposition on shared and distributed memory architectures
1993 IEEE Gordon Bell Performance Prize (50 Gflop/s on CM-5)
1998 IEEE Gordon Bell Price / Performance Prize (10 Gflops on Linux/ Alpha Beowulf cluster, $15/Mflop)
2005 IEEE Gordon Bell Prize Finalist (48 Tflop/s on BlueGene/L)
Each Processing Node
16-node machine
(David Beazley, Peter Lomdahl)
LJ, EAM, MEAM for metals
Agent Based Models
Los Alamos
UNCLASSIFIED
NNSA

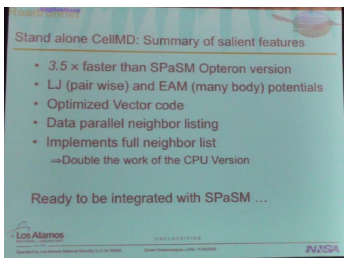

SPaSM simulations have helped to gain insight into fundamental materials and physics processes
Strength of metals
Fluid instability and the onset of turbulence
S. J. Zhou, D. L. Preston, P. S. Lomdahl, and D. M. Beazley, Science 279, 1525 (1998).
K. Kadau, T.C. Germann, P.S. Lomdahl, and B.L. Holian, Science 296, 1681 (2002).
K. Kadau et al., PNAS 101, 5851 (2004).
B.L. Holian and P.S. Lomdahl, Science 280, 2085 (1998).
Shock-induced plasticity and phase transitions
Los Alamos
UNCLASSIFIED
NNSA

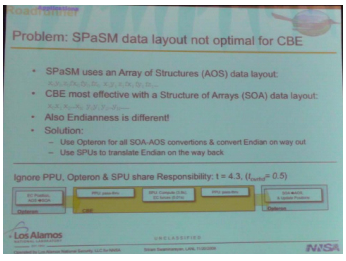

SPaSM simulations have played a key role in the US pandemic flu planning process
MIDAS
The Washington Post
PNAS
U.S. Plan For Flu Pandemic Revealed
Intervention simulations for U.S. influenza pandemic
Mitigation strategies for pandemic influenza in the United States
Los Alamos
UNCLASSIFIED
NNSA

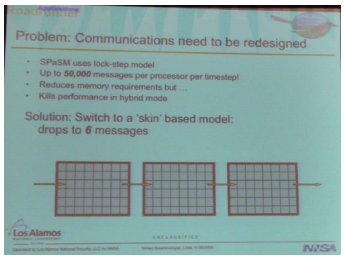

Decision to implement force() on CBE
CPU
95% of time in force()
Function offload model
Opterons for positions
CBE for force()
Dynamic Neighbor Listing
Neighbor Interaction
No Problem!
Los Alamos
UNCLASSIFIED
NNSA

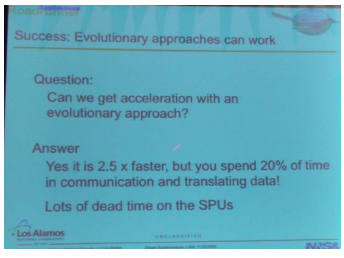

Spatial decomposition key to efficient neighbor list
for each box in system
Self-interact:
interact particles in same box
List interact:
interact particles in neighboring boxes
end
e.g. with 10 atoms in each box
⇒ each atom interacts with 270 other atoms
⇒ However, only ~35 atoms lie within cutoff
⇒ Lots of wasted work
⇒ We need a means of rejecting atoms efficiently even within this reduced set
Los Alamos
UNCLASSIFIED
NNSA

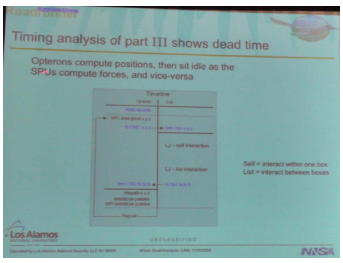

Optimizing the neighbor listing
Method
% time in NL
Naive
80%
Simple
60%
Data Parallel
25%
Los Alamos
UNCLASSIFIED
NNSA

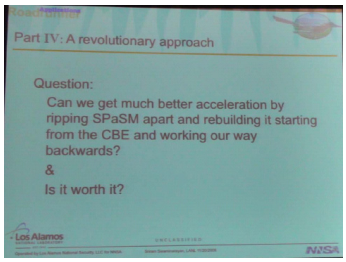

Stand alone CellMD: Summary of salient features
3.5 × faster than SPaSM Opteron version
LJ (pair wise) and EAM (many body) potentials
Optimized Vector code
Data parallel neighbor listing
Implements full neighbor list
⇒Double the work of the CPU Version
Ready to be integrated with SPaSM …
Los Alamos
UNCLASSIFIED
NNSA

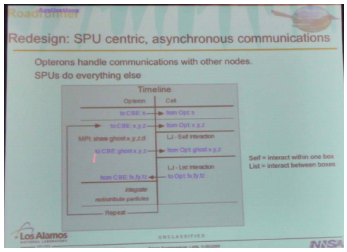

Problem: SPaSM data layout not optimal for CBE
SPaSM uses an Array of Structures (AOS) data layout:
CBE most effective with a Structure of Arrays (SOA) data layout:
Also Endianness is different!
Solution:
Use Opteron for all SOA-AOS convertions & convert Endian on way out
Use SPUs to translate Endian on the way back
Ignore PPU, Opteron & SPU share Responsibility: t = 4.3, (t_cell/t_o = 0.5)
Los Alamos
UNCLASSIFIED
NNSA

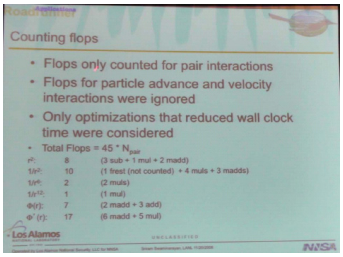

Problem: Communications need to be redesigned
SPaSM uses lock-step model
Up to 50,000 messages per processor per timestep!
Reduces memory requirements but …
Kills performance in hybrid mode
Solution: Switch to a 'skin' based model: drops to 6 messages
Los Alamos
UNCLASSIFIED
NNSA

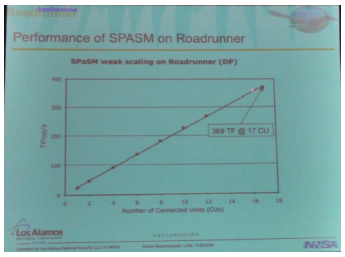

Success: Evolutionary approaches can work
Question:
Can we get acceleration with an evolutionary approach?
Answer
Yes it is 2.5 x faster, but you spend 20% of time in communication and translating data!
Lots of dead time on the SPUs
Los Alamos
UNCLASSIFIED
NNSA

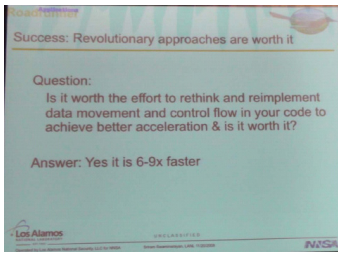

Timing analysis of part III shows dead time
Opterons compute positions, then sit idle as the SPUs compute forces, and vice-versa
Timeline
Self = interact within one box
List = interact between boxes
Los Alamos
UNCLASSIFIED
NNSA

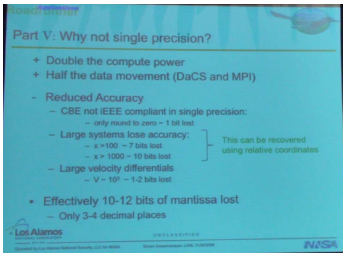

Part IV: A revolutionary approach
Question:
Can we get much better acceleration by ripping SPaSM apart and rebuilding it starting from the CBE and working our way backwards?
&
Is it worth it?
Los Alamos
UNCLASSIFIED
NNSA

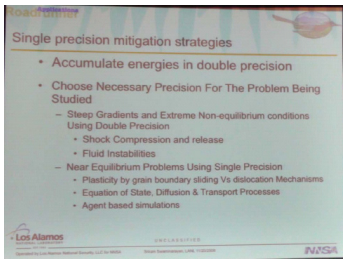

Redesign: SPU centric, asynchronous communications
Opterons handle communications with other nodes.
SPUs do everything else
Timeline
Opteron
Cell
Self = interact within one box
List = interact between boxes
Los Alamos
UNCLASSIFIED
NNSA

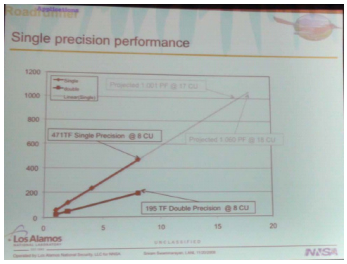

Counting flops
Flops only counted for pair interactions
Flops for particle advance and velocity interactions were ignored
Only optimizations that reduced wall clock time were considered
Total Flops = 45 * N_pair
(3 sub + 1 mul + 2 madd)
(1 frest (not counted) + 4 muls + 3 madds)
(2 muls)
(1 mul)
(2 madd + 3 add)
(6 madd + 5 mul)
Los Alamos
UNCLASSIFIED
NNSA

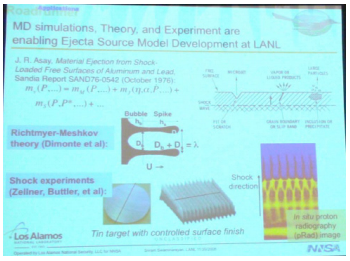

Performance of SPASM on Roadrunner
SPaSM weak scaling on Roadrunner (DP)
369 TF @ 17 CU
Number of Connected Units (CUs)
Los Alamos
UNCLASSIFIED
NNSA

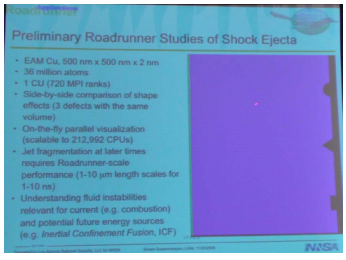

Success: Revolutionary approaches are worth it
Question:
Is it worth the effort to rethink and reimplement data movement and control flow in your code to achieve better acceleration & is it worth it?
Answer: Yes it is 6-9x faster
Los Alamos
UNCLASSIFIED
NNSA

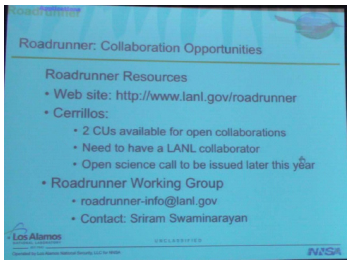

Part V: Why not single precision?
Double the compute power
Half the data movement (DaCS and MPI)
Reduced Accuracy
CBE not IEEE compliant in single precision:
only round to zero ~ 1 bit lost
Large systems lose accuracy:
x >100 ~ 7 bits lost
x > 1000 ~ 10 bits lost
This can be recovered using relative coordinates
Large velocity differentials
Effectively 10-12 bits of mantissa lost
Only 3-4 decimal places
Los Alamos
UNCLASSIFIED
NNSA


Single precision mitigation strategies
Accumulate energies in double precision
Choose Necessary Precision For The Problem Being Studied
Steep Gradients and Extreme Non-equilibrium conditions Using Double Precision
Shock Compression and release
Fluid Instabilities
Near Equilibrium Problems Using Single Precision
Plasticity by grain boundary sliding Vs dislocation Mechanisms
Equation of State, Diffusion & Transport Processes
Agent based simulations
Los Alamos
UNCLASSIFIED
NNSA


Single precision performance
471TF Single Precision @ 8 CU
195 TF Double Precision @ 8 CU
Los Alamos
UNCLASSIFIED
NNSA


MD simulations, Theory, and Experiment are enabling Ejecta Source Model Development at LANL
Richtmyer-Meshkov theory (Dimonte et al):
Shock experiments (Zellner, Buttler, et al):
Shock direction
Tin target with controlled surface finish
in situ proton radiography (pRad) image
Los Alamos
UNCLASSIFIED
NNSA


Preliminary Roadrunner Studies of Shock Ejecta
EAM Cu, 500 nm x 500 nm x 2 nm
36 million atoms
1 CU (720 MPI ranks)
Side-by-side comparison of shape effects (3 defects with the same volume)
On-the-fly parallel visualization (scalable to 212,992 CPUs)
Jet fragmentation at later times requires Roadrunner-scale performance (1-10 μm length scales for 1-10 ns)
Understanding fluid instabilities relevant for current (e.g. combustion) and potential future energy sources (e.g. Inertial Confinement Fusion, ICF)
Los Alamos
UNCLASSIFIED
NNSA


Roadrunner: Collaboration Opportunities
Roadrunner Resources
Web site: http://www.lanl.gov/roadrunner
Cerrillos:
2 CUs available for open collaborations
Need to have a LANL collaborator
Open science call to be issued later this year
Roadrunner Working Group
roadrunner-info@lanl.gov
Contact: Sriram Swaminarayan
Los Alamos
UNCLASSIFIED
NNSA

⑬ BOF「The HPC Advisory Council Initiative」

このBOFは、実質的には、60を超えるベンダー企業、大学とのHPCに関するコンソーシアムのミーティングであった。HPC技術、人的ネットワーク、教育、ベストプラクティスなどを共有している。このBOFでは、下記のことについて、フリーディスカッションがなされた。特筆すべき意見は見られなかった。直感的には、お付き合い程度で、機能していないコンソーシアムのように感じられた。

- HPC利用の上での最大の弱点は？
- ベンダーとエンドユーザーの関係のレベルは？
- HPCリソースをちゃんと使っているのか？
- 貴社は、本コンソーシアムの人的にネットワークにどういう経験について情報提供できるか？
- よりよい生産性のためにHPCを使いたがるユーザーに思いとどまらせた事例はあるか？
- 生産性か効率か？違いはあるのか？
- 本コンソーシアムでの活動経験は？

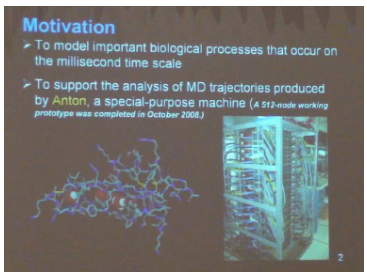
⑭ Papers 「A Scalable Parallel Framework for Analyzing Terascale Molecular Dynamics Trajectories」by Tiankai Tu et. al. (D. E. Shaw)

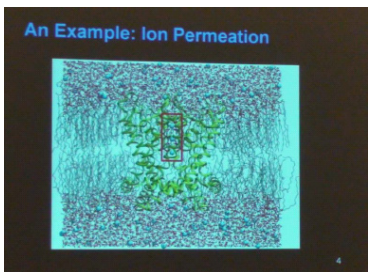
MD 専用計算機 Anton の稼動後に予想される膨大なデータについて、データ解析を並列に行う手法の検討である。2 万原子の中から 2 つのイオンの位置座標の時系列データを取り出す手法を提案している。提案の手法は分散入力がポイントであり、性能はファイルシステムに依存するところが大きいと考えられる。出力はファイル出力ではなく、可視化アプリのメモリバッファへの出力を実装している。手早く計算結果を見るためには、有用なツールである。

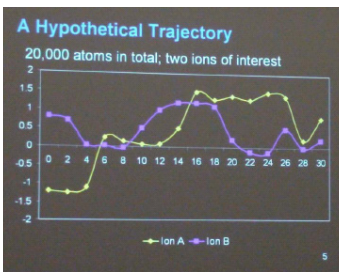
論文の PDF は、http://scyourway.nacse.org/conference/view/pap318。

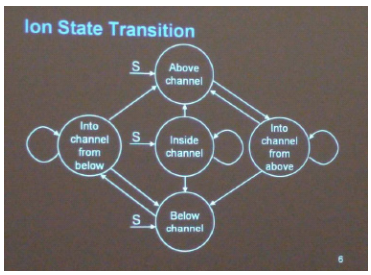
発表のポイントは次の通りである。

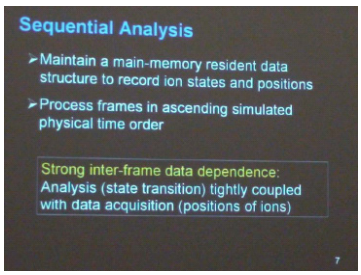
- Anton は、512 ノードのプロトタイプが 2008 年 10 月に稼動している。
- 膜タンパクでのイオンの動き（軌道）に注目していることから、約 2 万の原子の時系列データから特定のイオンの時系列データを取り出す必要がある。
- 10 ピコ秒ごとにスナップショットの出力を想定し、1 ミリ秒の長さでは、1 億毎のスナップショットで 60TB のデータサイズとなる。ローカルディスクのバンド幅を 100MB/s とすると、読み込みに 1 週間以上かかる。
- HiMACH は、Google type の Map Reduce による処理を念頭においた、MD 軌道を解析する Python プログラムである。フレームごとのデータの取得を行い、その後、通信により、フレーム間のデータ解析を行う手法である。
- 512Core を用いると、Single Core に比べ 100 倍以上速くなる。
- 2 万原子の中から 2 つのイオンの位置座標の時系列データを取り出す場合、1TB のデータ解析が 15 分。解析結果は、VMD の内部バッファへ投入する。

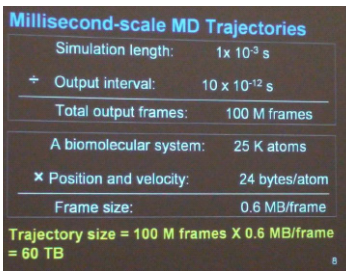












Problems with Sequential Analysis

Millisecond-scale trajectory size :	60 TB
Local disk read bandwidth:	100 MB / s
Time to fetch data to memory:	> 1 week
Analysis time:	Varied
Time to perform data analysis :	**Weeks**

Sequential analysis lack the computational, memory, and I/O capabilities!

9

A New Data Analysis Model

- Specify which frames to be accessed
- Decouple data acquisition from data analysis

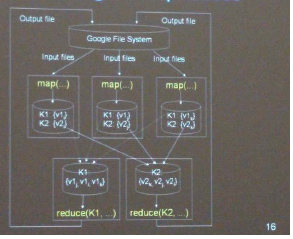
Trajectory definition

Stage1: Per-frame data acquisition

Stage 2: Cross-frame data analysis

11

Inspiration: Google's MapReduce



16

Trajectory Analysis Cast Into MapReduce

- Per-frame data acquisition (stage 1): map()
- Cross-frame data analysis (stage 2): reduce()
- Key-value pairs: connecting stage1 and stage2
  - Keys: categorical identifiers or names
  - Values: including timestamps
  - Examples: $(ion\_id_j, (t_k, x_{jk}, y_{jk}, z_{jk}))$

17

HiMach

- A MapReduce-style API that allows users to write Python programs to analyze MD trajectory
- A parallel runtime that executes HiMach user programs in parallel on a Linux cluster automatically
- Performance results on a Linux cluster:
  - 2 orders of magnitude faster on 512 cores than on a single core
  - 1 TB data analyzed in 15 minutes

19

Chained Reduce Operations

- Support multiple rounds of analysis
- Exploit a Python language feature (the `yield` statement)

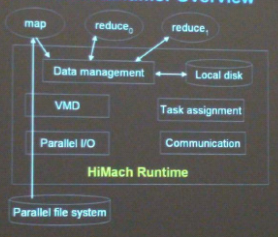
Trajectory definition

Per-frame data acquisition (map)

Cross-frame data analysis (reduce)

22

The HiMach Runtime: Overview



24

Task Assignment

- Assumption: equal-weighted tasks
- Partitioning of the workload:
  - A block data decomposition algorithm
  - No inter-processor communication when assigning map tasks
  - One collective operation (`MPI_Allgather`) when assigning reduce tasks

25

Parallel I/O

- Assumption:
  - Frames organized as megabyte-sized files
  - Files stored in NFS or in a parallel file system
- How to read the frames:
  - No communication needed to coordinate I/O
  - Reads issued by HiMach
  - Frame data passed to user-defined `map()`
- How to write output results:
  - Writes issued by user programs directly

26

Interaction with VMD

- VMD: Visual Molecular Dynamics (UIUC)
- Run an instance of VMD on each core
- Load molecule structure files
- Push frame into VMD's internal buffer
- Support frame caching and prefetching

27

Experiment Setup

- Cluster configuration:
  - A commodity Linux cluster with 128 nodes
  - Two Intel Xeon 2.66 GHz quad-core processors per node
  - Two 250 GB SATA disks organized in RAID 1 per node
  - CentOS 4.6 with a Linux kernel of 2.6.22
  - Nodes connected via a Gigabit Ethernet connection
- Storage configuration:
  - PVFS 2.7.0 (a parallel file system developed at Argonne National Laboratory) on 64 nodes
  - Both the I/O server and the metadata servers on all the 64 nodes

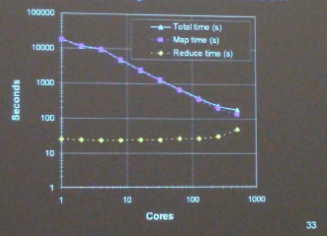
31

Experiment Setup

- Driving applications:
  - Ion permeation: Time series analysis (154 GB)
  - Sliding-window RMSD: Measurement of structural "proximity" between two frames (85 GB)

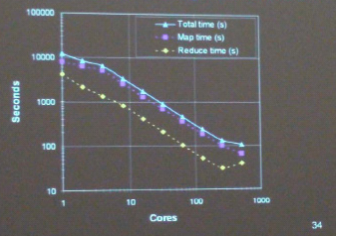
32

Strong Scalability: Ion Permeation



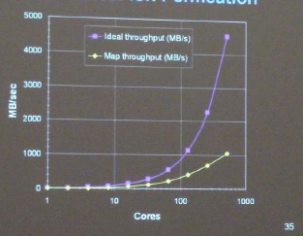
33

Strong Scalability: Sliding-Window RMSD



34

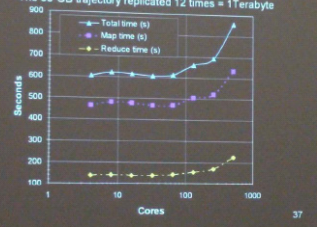
I/O Performance: Ion Permeation



35

Weak Scalability: Sliding-Window RMSD

The 85-GB trajectory replicated 12 times = 1Terabyte



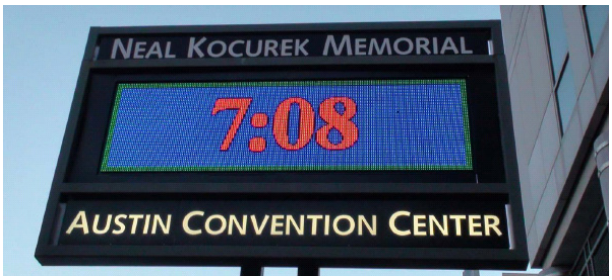
37

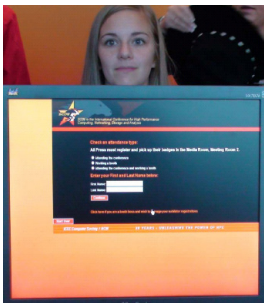
Summary

- HiMach = MapReduce + MPI + Python + VMD
  - + A new MD trajectory parallel analysis model
  - + A new interface for defining trajectories
  - + A novel method of running VMD in parallel
  - + An extension to MapReduce to support chained reduce operations
- Potentially applicable to data analysis for other types of simulations

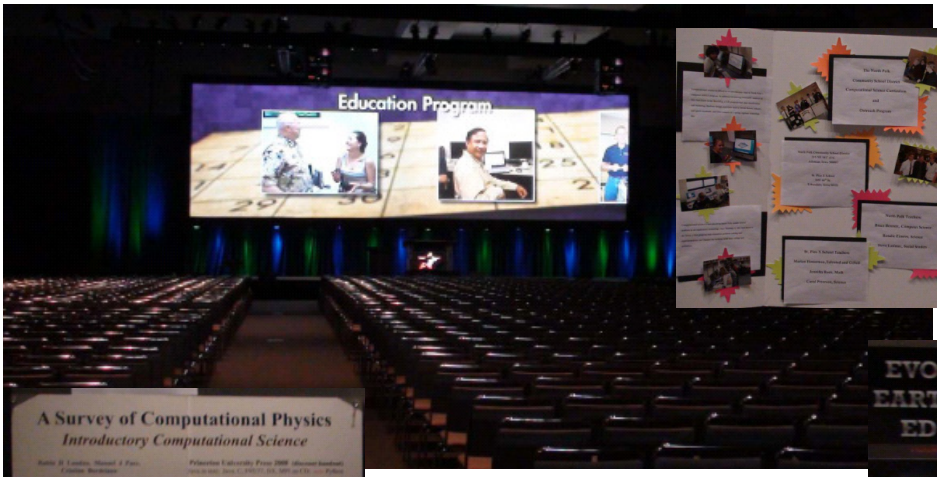
38

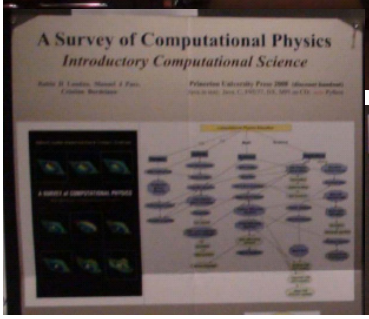
● SC08 の会場へ

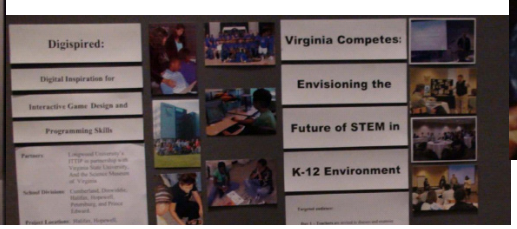




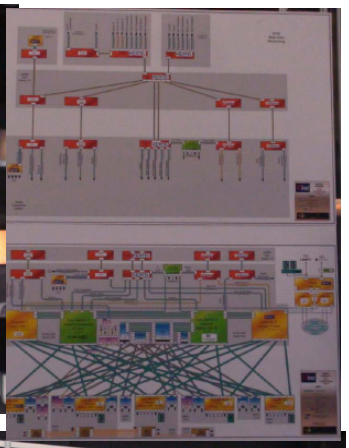
● Education Program

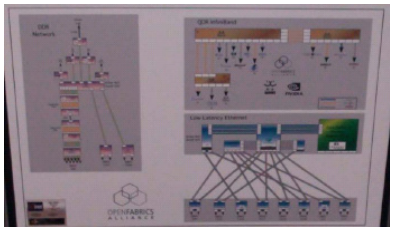


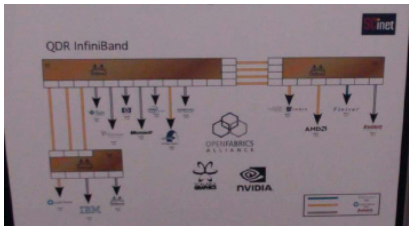


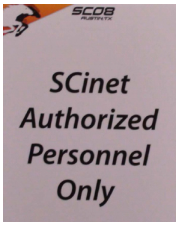


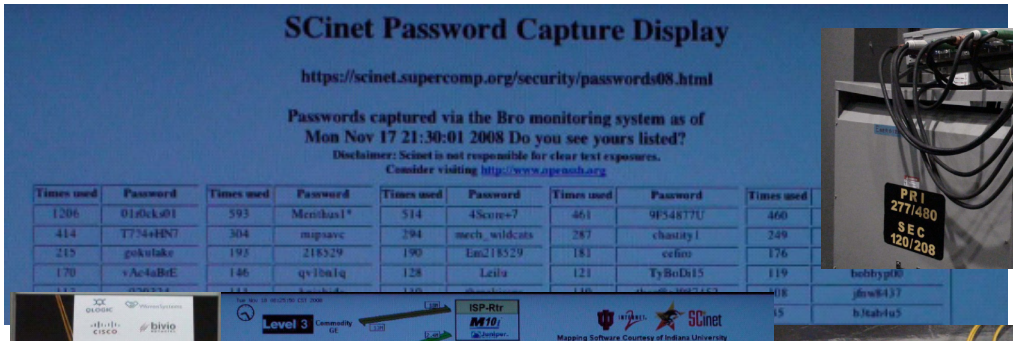
● SC を支えるネットワーク

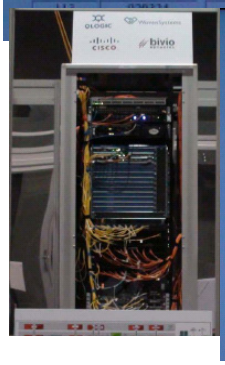


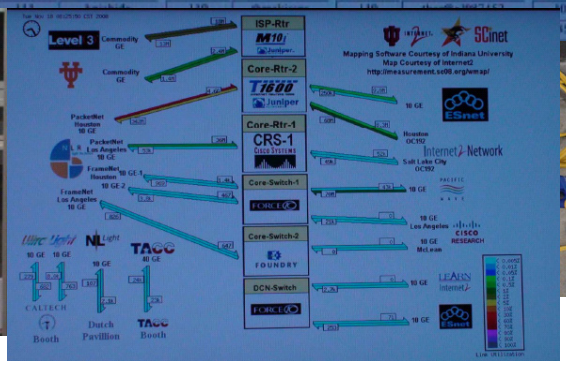






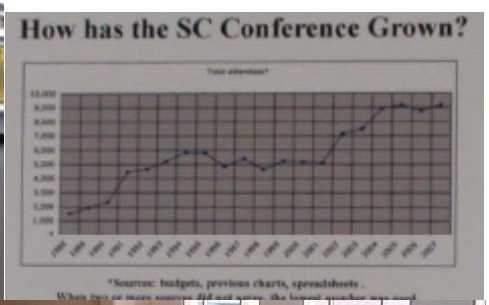


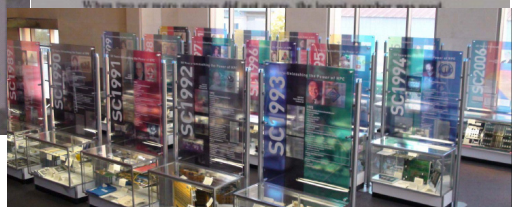




● SC20 年の歴史

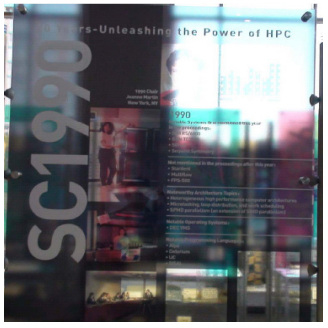

WELCOME
TO THE SC08
20TH ANNIVERSARY
EXHIBITS

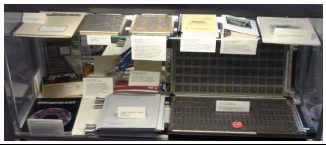

How has the SC Conference Grown?


SC1991
SC1992
SC1993
SC1994
SC2006


SC1988


SC1989

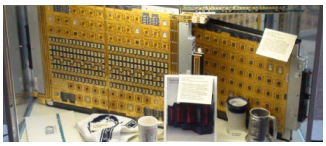

SC1990

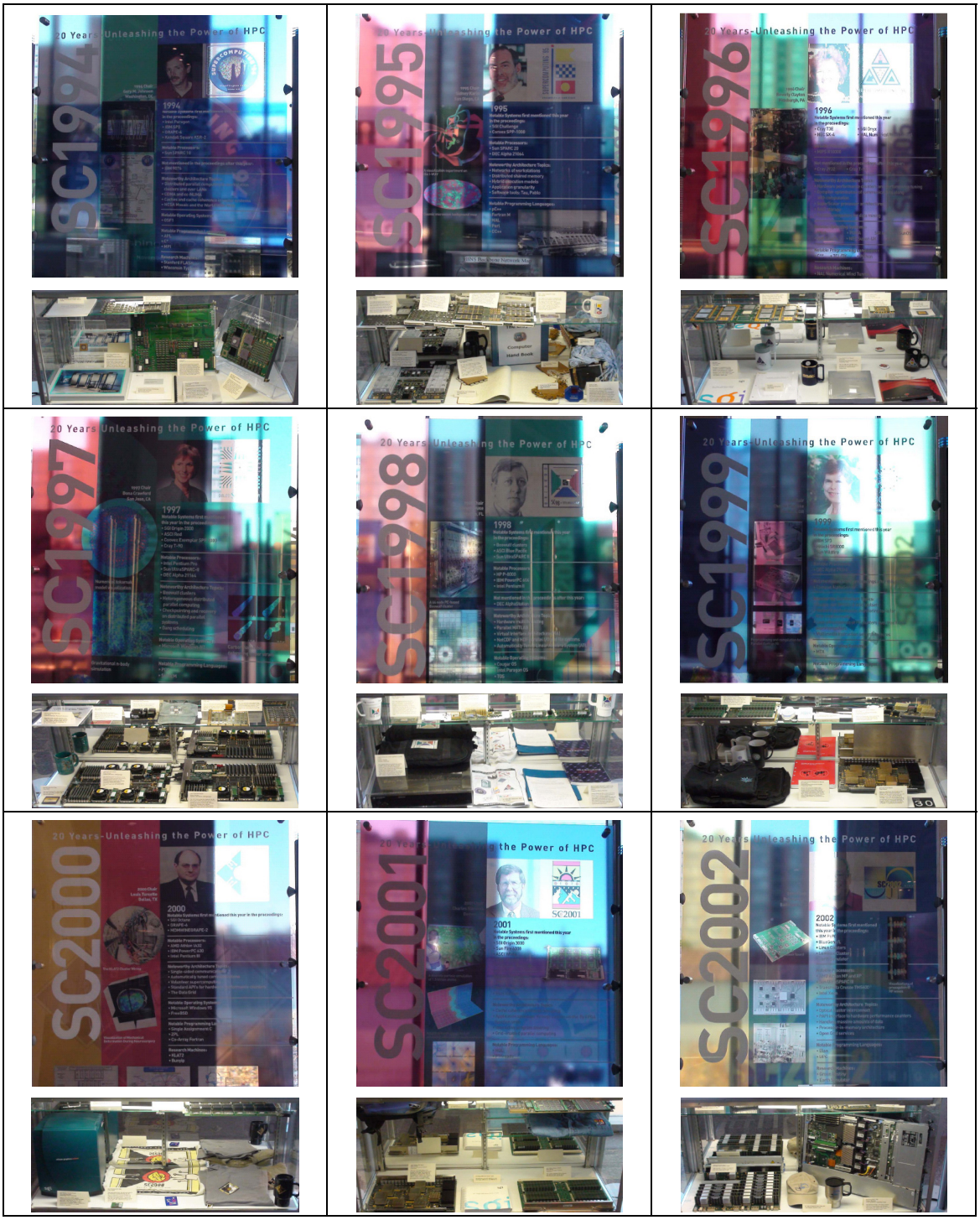


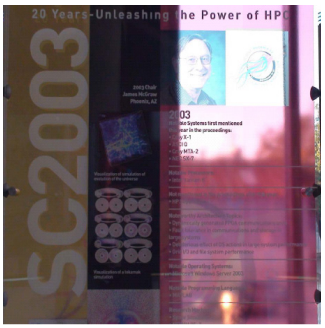

SC1991

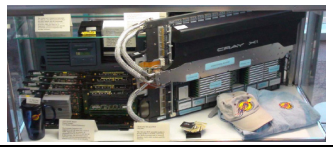

SC1992

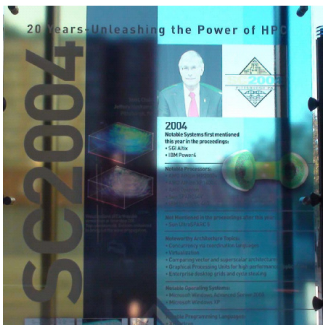

SC1993

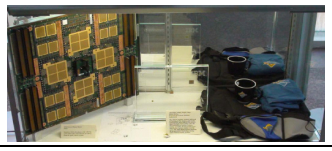


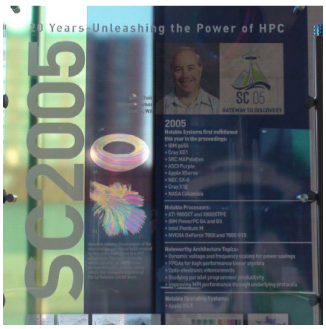

20 Years-Unleashing the Power of HPC
SC1994
SC1995
SC1996
SC1997
SC1998
SC1999
SC2000
SC2001
SC2002

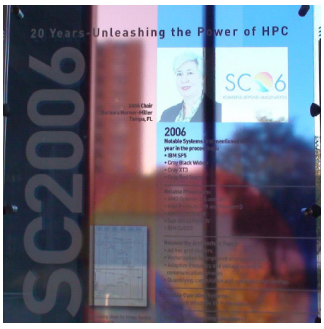

20 Years-Unleashing the Power of HPC
SC2003

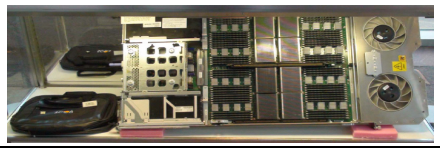


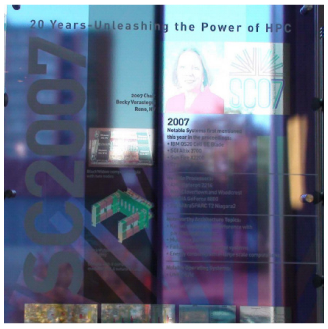

20 Years-Unleashing the Power of HPC
SC2004
2004




20 Years-Unleashing the Power of HPC
SC2005
2005


20 Years-Unleashing the Power of HPC
SC2006
2006



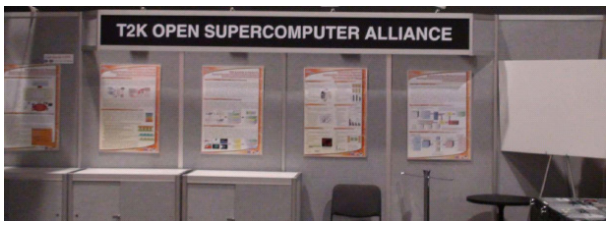

20 Years-Unleashing the Power of HPC
SC2007
2007

● 展示会場　（日本コレクション）


GRAPE Projects


IST/CMC - Osaka University

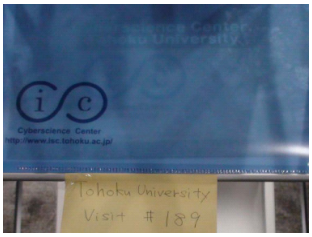

DOSHISHA UNIVERSITY

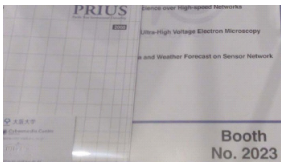

T2K OPEN SUPERCOMPUTER ALLIANCE

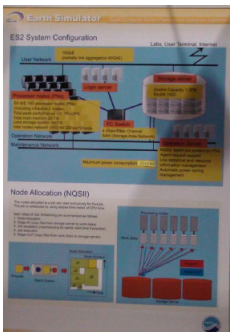

RIST
Research Organization for Information Science & Technology
High-End Computing Large-Scale Simulation


Hokkaido University

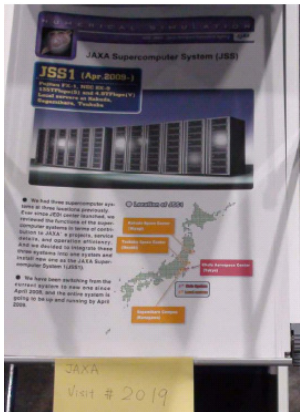

E-Science Infrastructure
AIST

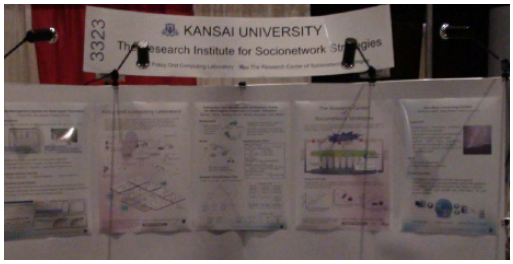

Tohoku University
Visit # 189


PRIUS
Booth
No. 2023

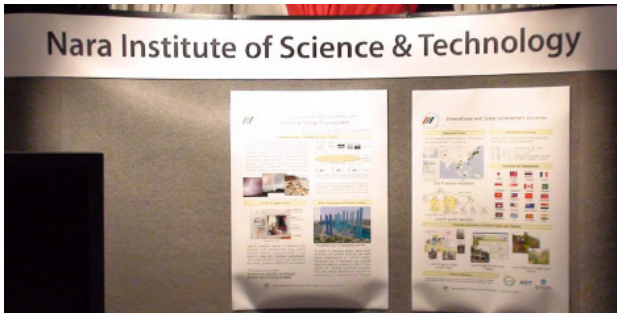

Earth Simulator
ES2 System Configuration
Node Allocation (NQSII)


地球シミュレータ
Earth Simulator
地球シミュレータセンター
The Earth Simulator Center
JAMSTEC
Visit # 2628


JAXA Supercomputer System (JSS)
JSS1
JAXA
Visit # 2019


3323
KANSAI UNIVERSITY


統計数理研究所
The Institute of Statistical Mathematics
http://www.ism.ac.jp/
Promoting Research on Theories and Methodologies for Statistical Inference and Reasoning Based on Data
「データに基づく合理的推論の仕組みの研究を推進します。

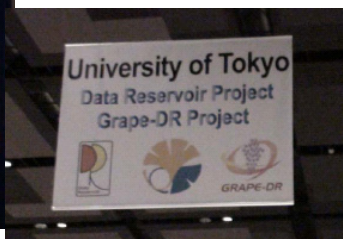

Nara Institute of Science & Technology

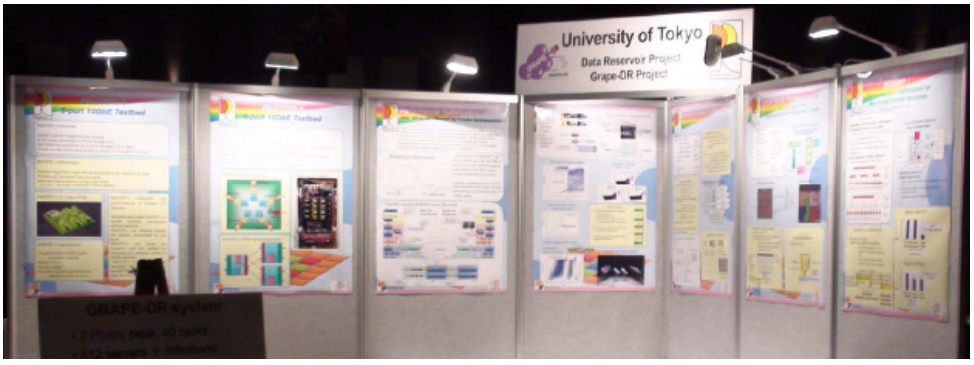

NiCT
National Institute of Information and Communications Technology
SPARC
NiCT


KYUSHU UNIVERSITY


TOKYO TECH


JAIST
Japan Advanced Institute of Science and Technology


University of Tokyo
Data Reservoir Project
Grape-DR Project
GRAPE-DR


University of Tokyo
Data Reservoir Project
Grape-DR Project

● ペタ flops 級の計算機を持つ大御所のブース


NASA
NASA @ SC08
Inspiration, Innovation, Discovery


Raytheon


NSF
National Science Foundation
National Center for Atmospheric Research
University Corporation for Atmospheric Research
CISL


Argonne
Argonne
Argonne
TeraGrid

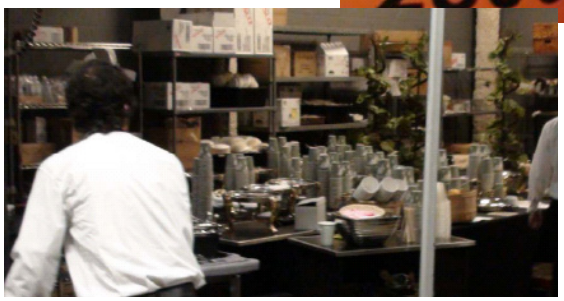

HPC
HPC
Disaster Recovery
Comprehensive Security

● 代表的なベンダーのブース

● おまけ


More
Exhibits


MORE
EXHIBIT
2800-330




Security
Personnel
Only
Beyond this Point


Thank You
for Attending