

NERSC/LBNL 視察報告

鴨志田良和 宮寄洋
東京大学情報基盤センター

東京大学情報基盤センター スーパーコンピューティング部門の鴨志田と宮寄は、2010 年 11 月 10 日から 12 日までの 3 日間、米国カリフォルニア州の Lawrence Berkeley National Laboratory と、この研究所に設置された NERSC (National Energy Research Scientific Computing Center) を視察した。本稿はその報告である。

NERSC とは

NERSC は、米国のエネルギー省 (DOE) の主要な科学技術計算施設であり、世界有数の大規模な計算機を保有して計算科学における様々な発見をけん引している施設の一つである。カリフォルニア州の Lawrence Berkeley National Laboratory の Oakland Scientific Facility (OSF) に設置されており、サンフランシスコのベイエリアを走る鉄道、BART の “19th Street” 駅のすぐそばにある。

3000 以上の利用者は全て DOE の研究プロジェクトのもとで研究を行っており、その内容は気象モデリング、材料科学、初期宇宙のシミュレーション、高エネルギー物理、たんぱく質の構造解析などさまざまである。

NERSC は、その前身である National Magnetic Fusion Energy Computer Center が 1990 年に名称を変更したもので、当時は、National Energy Research Supercomputer Center (NMFEC) という名称であった。現在の Scientific Computing Center の名称は、1996 年に採用されたもので、単にスパコンのサービスを提供するだけでなく、科学技術計算の生産性をより高めていくという考え方を示したものである。

NERSC 訪問概要

我々は、NERSC における運用や設備管理、ユーザーサポートなどの専門家と以下に示すトピックごとに 30 分から 1 時間程度のミーティングを持ち、意見交換を行った。

- Overview of NERSC
- Presenting HPC Information on the Web
- NERSC/Cray Center of Excellence for Programming Models and Performance Modeling
- Programming Libraries in the Advanced Computational Software (ACTS) Collection
- NERSC 6 Benchmarking
- HPC Consulting
- NERSC 6 Procurement
- Cloud Computing and HPC
- Optimizing for Energy Efficiency

- Lustre Monitoring Tool (LMT)
- NERSC Global Filesystem and Storage
- NERSC Operations

また、東京大学情報基盤センターと、スーパーコンピューティング部門で行われている研究の概要についての説明、計算機室や電源設備の見学、Berkeley 研究所のメインキャンパスの研究室訪問も行った。

以降、上記のミーティングの中で、NERSC に設置されている計算機、システム運用、ユーザーサポート、そしてシステムの調達について、他のミーティングでの話題も交えつつ、その内容を述べる。

設備

NERSC 内の最新のスーパーコンピュータは Hopper と呼ばれる、Cray XE6 システムである (図 1)。Hopper は、1990 年に NMFEC が NERSC に名称を変更してから 6 代目のスパコンであるため、NERSC-6 とも呼ばれている。Hopper は、全 6,392 ノード、各ノードは 2 つの AMD の 12 コア CPU (‘Magny-Cours’) 2.1GHz を搭載し、全体で 153,408 コアを有する計算機である。2010 年 11 月時点でのランキングで Top 500 ランキングの 5 位に入った Hopper は、訪問時は運用に入る前の様々なテストを行っている最中であった。電源事情は日本に限らず米国でも厳しいようで、Hopper の導入にあたっては、受電電力の増強も行ったそうである (図 2)。他にも、1 世代前の Franklin (Cray XT4)、中規模クラスタの Carver、クラウド計算のテストベッドである Magellan (どちらも IBM iDataPlex) などの計算機が計算機室内に設置されていた。

各計算機は NERSC Global Filesystem (NGF) と呼ばれるファイルシステムに接続している。NGF は IBM General Parallel File System (GPFS) で構築されており、多くの計算機システムのホームディレクトリや、プロジェクト用のディレクトリを提供している。この NGF により、利用者は計算機システム間でファイルを共有することができる。また、データのアーカイブ・バックアップ用に、High Performance Storage System (HPSS) が提供されている。HPSS は 140TB のディスクキャッシュを持つテープストレージシステムである。

サンフランシスコは日本と同様に地震の多い場所であるため、地震対策も行われている。各計算機ラックは、ISO-Base という免震設備の上に設置されている。ISO-Base は、2 枚のプレートとその間に挟まれた鉄のベアリングでできており (図 3)、地震の際は滑らかに横運動をして振動に耐えることができる。

計算機室の一角には、運用スタッフが常駐するコントロールルームがある。ガラス張りの壁で仕切られているため、コントロールルームからは計算機室が一望できる。各計算機ラックの上・中・下段に設置された温度センサーの測定結果はこのコントロールルームからグラフィカルに見ることができるようになっていて、効果を実際に確かめながら冷却設備の設定変更を行うのだと、ネットワーク・サーバーグループのリーダーである Brent Draney 氏が説明してくれた。

運用

システムの運用については、運用グループリーダーの Steve Lowe 氏が語ってくれた。NERSC では、週 7 日 24 時間の運用を行うために、9 人の NERSC のスタッフが運用にあたっている。夜間 (23 時から 7 時)、休日も 3 人体制で対応を行っている。Nagios 等で死活監視を行い、ダウン時の Web、E-mail 対応、ベンダーへの連絡を行う。ログインできない、パスワードを忘れた等の比較的簡単な質問も一日あたり 2~3 件あるが、システムグループやユーザーサービスグループのアプリケーションの専門家などと、複数名で問題対応にあたる必要がある場合にはチャットも活用しているという。NERSC が Oakland に移ったのは 1996 年であるが、電力や設置面積などの都合から、現在 Berkeley のメインキャンパスに新たな建物を建設し、計算機室を設ける計画があるそうである。この場合、運用スタッフは 2 か所に分散せざるを得ないため、限



図 1: Hopper の説明を行う Carter 氏



図 2: 電源設備の説明を行う Draney 氏

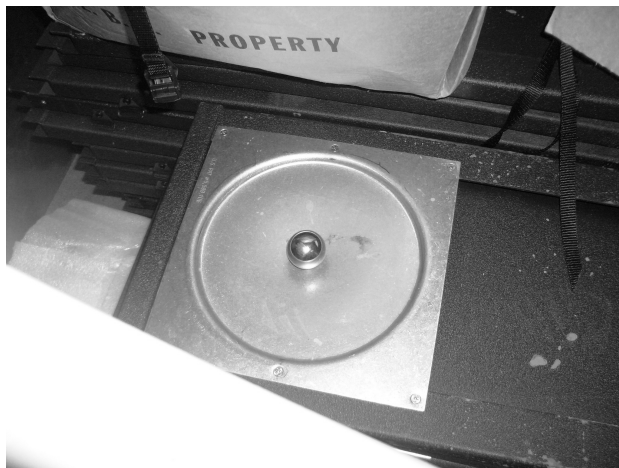


図 3: ISO-Base のプレートとベアリング

られた人数のスタッフで運用を続けるためには、拠点あたりの人数を減らすなどの効率化を図る必要があるだろうということだった。本学でも、次に導入予定のスパコンは本郷キャンパスではなく、柏キャンパスに設置される予定であるため、同様の悩みを抱えている。

NERSC では、本学のシステムで行っているような定期メンテナンスは実施せず、OS や他のソフトウェアのアップグレードやセキュリティに関わる変更を行う必要がある場合などに適宜停止のスケジュールを行う。通常は停止の1週間前にアナウンスを行うが、Global File System など、影響範囲が大きい変更の場合は、1か月前にアナウンスを行う。ハードウェアの部分的な故障は2週間に1回くらいで、システム全体がダウンすることはほとんどないという。主要な並列計算機である Franklin と Carver は、年間 95%以上稼働している。

ポータルとユーザーサポート

NERSC の計算資源の割り当てを受けるためには、研究プロジェクトが DOE から資金を供給されているものである必要がある。そのうえで、プロジェクトごとに計算機利用の計画を提出し、審査を受けたのちに計算資源が割り当てられる。使用した CPU 時間とノード数に基づいて従量制で仮想的に課金され、プロジェクトに割り当てられた上限に達するまで計算資源を利用することが可能である。実際の利用料はかからない。NERSC の利用者は約 3,000 名で、研究プロジェクトの数は約 400 である。

利用者には専用の Web ポータルが提供され、投入されたジョブの状態 (実行中、開始待ちなど) や、システム全体の様子 (定期的に実行されているベンチマークの結果など) を閲覧することができる。また、終了したジョブの統計情報を閲覧することもできる。Web ポータルは Integrated Performance Monitoring (IPM) ツールと統合されている。IPM は並列ジョブのプロファイリングツールで、IPM ライブラリをリンクして実行したジョブの場合は、MPI 呼び出しの種類ごとの統計など、より詳細な情報をポータル上でもグラフィカルに閲覧することができる。

NERSC では利用者からの問い合わせ対応にも力を入れている。各プロジェクトで使用しているコードの種類は約 500 存在するため、質問の内容も多岐にわたる。これを、化学、気象、計算物理、物質、コンパイラ、数学、I/O、システム評価などに専門性を持つ 9 人のコンサルタントが担当する。電話、メール、Web 上のシステムなどから問い合わせを行うことが可能であり、ユーザーサービスグループのリーダー、Katie Antypas 氏によると、一日あたり 12 件程度の問い合わせがあるそうである。月曜から金曜、午前と午後で主担当を分担し、問い合わせの一時対応を行う。必要に応じて担当を変更するなどして、基本的には 4 時間以内に回答することとしている。全体の 80%は 3 日以内に解決することである。RightNow というシステムを使用して、問い合わせチケットの発行や担当のアサインを行っている。問い合わせの内容によっては、利用者のコードを見て対応を行うこともある。

調達

NERSC の Hopper システムの調達について、Hopper の調達プロジェクトのリーダーであり、現在はパークレイ研究所の Computing Sciences Deputy を務める Jonathan Carter 氏が説明してくれた。Carter 氏は今回の NERSC 訪問のホスト役を快諾してくれた人物である。

Hopper は、2008 年 9 月に RFP(提案要求) が公開されたのち、2009 年 7 月に Cray との契約が締結された、その後、システムは 2 段階に分けて導入された。Phase 1 で導入されたのは、664 ノードの Cray XT5 システムである。総 CPU コア数は Phase 2 の 1/30 程度であるが、2010 年の 2 月にすでに運用を開始している。Phase 2 の Cray XE6 システムは現在 Cray、NERSC のスタッフと一部のユーザによってテスト中であり、本格運用開始は 2011 年の夏以降となる予定である。

仕様の策定に当たっては、電源、ハードウェア、基本的なソフトウェアなどの最小限の要求のほか、性能については、STREAM、PSNAP、IOR等の基本的なベンチマークのほかに、Sustained System Performance (SSP) という指標で評価を行うこととしている。SSPは、スパコンのピーク性能がアプリケーションの実際の性能と次第に離れていくという近年の観察に基づいて、なるべく多くのアプリケーション範囲をカバーするように選定された7つのベンチマーク (CAM/GAMESS/GTC/MAESTRO/MILC/PARATEC/IMPACT-T) の実行結果 (TFLOPS) の平均値である。導入ベンダーは、計算機の設置後、もし性能要件を満たさない場合はベンチマークアプリケーションのチューニング等を実施することが求められる。

各ベンチマークは運用開始後も定期的に実行・計測され、システムの異常発見や、ハード増設等のシステム変更の影響の確認のためなどに利用される。

おわりに

NERSC の Director を務める Katherine Yelick 氏は、神戸の京速 (けいそく) コンピュータ「京 (けい)」や、最先端研究基盤事業で整備される大規模な共有ファイルシステムの動向にとっても興味があるようで、ハードウェア構成などを熱心に聞いていた。

今回の訪問で、NERSC のメンバーと行ったディスカッションは今後の情報基盤センターのスパコンの運用にも大変役立つものとなった。今後も互いに情報交換を行い、共同研究等に発展させることが重要であると考えられる。最後に、多忙な時期に我々を迎えてくれた、Yelick 氏、Carter 氏を始めとする NERSC のメンバーの皆さんに心からの感謝をしたい。