第 3 回先進スーパーコンピューティング環境研究会（ASE 研究会，Advanced Supercomputing Environment）

中 島 研 吾

東京大学情報基盤センター

2009 年 1 月 20 日（火）15 時から 18 時まで，東京大学情報基盤センター大会議室において，第 3 回先進スーパーコンピューティング環境研究会（ASE 研究会，Advanced Supercomputing Environment）が開催された。今回は，「ペタスケールアプリケーション開発支援環境に関する国際ワークショップ」と題して，科学技術振興機構 戦略的創造研究推進事業（CREST）との共催にて実施された。国内の大学，研究機関，企業から合計 20 名の出席者があり，活発な議論が行われた。

プログラムは表 1 に示すように，招待講演 1 件と情報基盤センタースーパーコンピューティング研究部門に所属する教員による 3 件の講演から構成されている。

表 1　第 3 回 ASE 研究会　プログラム

講演者	講演タイトル
Jonathan Carter （ NERSC , Lawrence Berkeley National Laboratory, USA）（招待講演）	Collaborative Benchmarking at NERSC and the T2K Open SuperComputer (Tokyo)
石川　裕（東大・情報理工／情報基盤センター）	An Overview of Seamless and Highly-Productive Parallel Programming Environment Project
片桐孝洋（東大・情報基盤センター）	Towards Petascale Eigensolver and Its Auto-tuning Methodology
中島研吾（東大・情報基盤センター）	Parallel Multistage Preconditioners by Hierarchical Interface Decomposition on T2K Open Supercomputer (Tokyo) with Hybrid Parallel Programming Models

招待講演者の Jonathan T. Carter 博士[1]はローレンスバークレイ国立研究所（Lawrence Berkeley National Laboratory）の National Energy Research Scientific Computing (NERSC) Center で User Service Group Leader を勤めており, スーパーコンピュータの導入, 運用の責任者の一人である。スパコンの導入，運用を円滑に実施するための評価手法，ベンチマーキングの研究を行っており，今回は NERSC における実例も交えてシステムの評価手法について紹介があった。今回は Carter 博士のご好意により，後掲のように当日の発表資料を掲載しているので，詳細についてはそちらをご覧いただきたいが，非常に興味深かったのは：

・ 様々なユーザーアプリケーションの特徴を網羅し，実際のアプリケーションに基づいて作成された 7 種類のベンチマークテストによって評価を実施している。

[1] http://www.nersc.gov/~jcarter/

・ いわゆる「検収」を，運用前のテストだけでなく，実運用環境でも実施している。また，実行性能，入出力性能を常時モニターしており，例えばコンパイラやライブラリのバージョンアップ等による様々な影響についても考慮している。

という点であり，今後の調達，センター運用においても参考になる点が多かった。

Carter 博士の今回の来日の目的は，これらのベンチマークを Hitachi HA8000 クラスタシステム（T2K オープンスパコン（東大））に適用し，評価することである。現在 NERSC で稼動中の Cray XT4 システムは T2K オープンスパコンと同様，AMD Quad-core Opteron（Barcelona）2.3GHz をしており，両者の比較という点からも興味深い。結果については今後，本「スーパーコンピューティングニュース」で紹介して行きたい。

引き続いて実施された，センター3 教員による講演は，現在，筑波大学，京都大学と共同で実施されている「シームレス高生産・高性能プログラミング環境」（代表：石川裕教授（東大））に関連したものである。本プロジェクトは文部科学省 次世代 IT 基盤構築のための研究開発「e-サイエンス実現のためのシステム統合・連携ソフトウェアの研究開発（テーマ①：高生産・高性能計算機環境実現のためのシステムソフトウェアの研究開発）」として実施されているものであり，本学では「高効率・高可搬性ライブラリ」を担当している。第 3 回 ASE 研究会では，プロジェクトの概要（石川），自動チューニング機構の開発（片桐），マルチコア環境での最適化（中島）について紹介した。

次回（第 4 回）ASE 研究会は 2009 年 3 月 27 日（金）14 時から 17 時を予定している。基調講演 1 件（Julien Langou 博士（University of Colorado, Denver）），招待講演 2 件（伊藤祥司博士（理化学研究所），村上弘博士（首都大学東京））が開催の予定である。

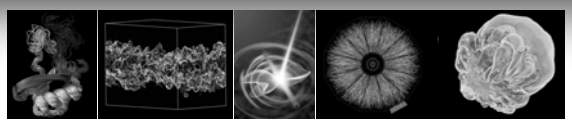
写真：Jonathan T. Carter 博士による講演の一コマ

Collaborative Benchmarking at NERSC and the T2K Open SuperComputer (Tokyo)

Jonathan T. Carter

NERSC, Lawrence Berkeley National Laboratory

以下に第 3 回先進スーパーコンピューティング環境研究会（ASE 研究会，Advanced Supercomputing Environment）における Jonathan T. Carter 博士（NERSC, Lawrence Berkeley National Laboratory）の講演資料を掲載する。



Collaborative Benchmarking at NERSC and the T2K Open SuperComputer (Tokyo)

Jonathan Carter
NERSC Division, Berkeley Lab
JTCarter@lbl.gov

Information Technology Center, University of Tokyo
January 20 2009

NERSC Mission

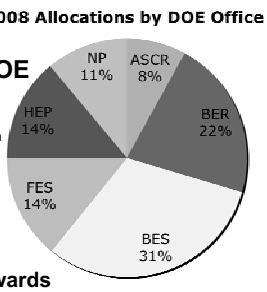
The mission of the National Energy Research Scientific Computing Center (NERSC) is to *accelerate the pace of scientific discovery* by providing high performance computing, information, data, and communications services for *all DOE Office of Science (SC) research.*

Science Over the Years

NERSC is enabling new science in all disciplines, with over *1,500 refereed publications* in 2007

NERSC is the Production Facility for DOE SC

- **NERSC serves all areas**
  - **~3000 users, ~400 projects**
- **Allocations managed by DOE**
  - **10% INCITE awards:**
    - **Large allocations, extra service**
    - **Used throughout SC; not just DOE mission**
  - **70% Production (ERCAP) awards:**
    - **From 10K hour (startup) to 5M hour**
    - **Only available at NERSC**
  - **10% each NERSC and DOE/SC reserve**
- **Award mixture offers**
  - **High impact through large awards**
  - **Broad impact across science domains**



NERSC 2008 Configuration

Large-Scale Computing System

Franklin (NERSC-5): Cray XT4
- 9,660 nodes; 38,640 cores
- ~37 Tflop/s sustained SSP (355 Tflops/s peak)

Clusters

Bassi (NCSb)
- IBM Power5 (888 cores)

Jacquard (NCSa)
- LNXI Opteron (712 cores)

PDSF (HEP/NP)
- Linux cluster (~1K cores)

NERSC Global Filesystem (NGF)

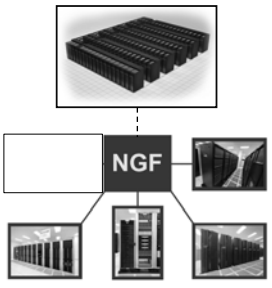
230 TB; 5.5 GB/s

HPSS Archival Storage
- 44 PB capacity
- 10 Sun robots
- 130 TB disk cache

Analytics / Visualization
- Davinci (SGI Altix)

NERSC Global Filesystem (NGF)

- **After thorough evaluation and testing phase in production**
- **Based on IBM GPFS**
- **Seamless data access from all of NERSC's computational and analysis resources**
- **Single unified namespace makes it easier for users to manage their data across multiple system**
- **First production global filesystem spanning four platforms, three architectures, and four different vendors**



2005-2010 Plan

2005: NERSC Five-Year Plan

- **Three trends identified:**
  - **widening gap between sustained application performance and peak**
  - **emergence of large, multidisciplinary computational science teams**
  - **flood of scientific data from both simulations and experiments**
- **Requirements/Trend Analysis led to the NERSC Five-Year Plan**
- **Approved by DOE and included:**
  - **NERSC-5 (Franklin, Cray XT4)**
  - **NERSC-6: 3-4x NERSC-5 sustained performance; initial delivery in 2009**

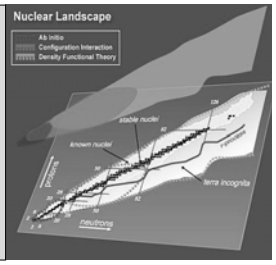
Science Benefits from Increased Computational Capability

- **2007 Franklin experience**
  - **Accepted in October 2007**
  - **DOE allocated the machine starting in January 2008**
  - **Before acceptance, full NERSC workload was running**
  - **Before and after acceptance, all runs on Franklin were "free" (not charged against user's allocation)**
- **Resulting experience:**
  - **Franklin was 80%-95% utilized within a week of acceptance**
  - **Users consumed 5x more compute time than they were allocated in 2007 (14x for largest users)**
  - **Users took the opportunity to produce science results; experiment with new algorithms, and scale to new levels**

Nuclear Physics

- **Calculation: High accuracy *ab initio* calculations on O^{16} using no-core shell model and no-core full configuration interaction model**
- **PI: James Vary, Iowa State**

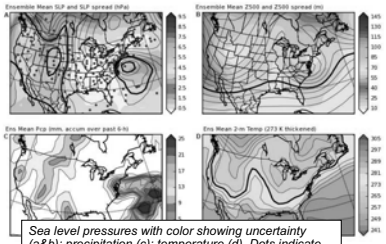
- **Science Results:**
  - **Most accurate calculations to date on this size nuclei**
  - **Can be used to parametrize new density functionals for nuclear structure simulations**
- **Scaling Results:**
  - **4M hours used; 200K allocated**
  - **12K cores; vs 2-4K before Franklin uncharged time**
  - **Diagonalize matrices of dimension up to 1 billion**



Validating Climate Models

- **INCITE Award for "20th Century Reanalysis" using an Ensemble Kalman filter to fill in missing climate data since 1892**
- **PI: G. Compo, U. Boulder**

- **Science Results:**
  - **Reproduced 1922 Knickerbocker storm**
  - **Data can be used to validate climate and weather models**
- **Scaling Results:**
  - **3.1M CPU Hours in allocation**
  - **Scales to 2.4K cores**
  - **Switched to higher resolution algorithm with Franklin access**

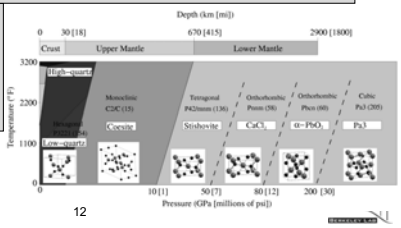


Sea level pressures with color showing uncertainty (a&b); precipitation (c); temperature (d). Dots indicate measurements locations (a).

Modeling Dynamically and Spatially Complex Materials for Geoscience

- **Calculation: Simulation of seismic waves through silicates, which make up 80% of the Earth's mantle**
- **PI: John Wilkins, Ohio State University**

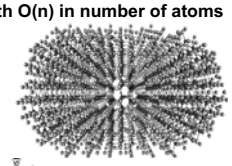
- **Science Result**
  - **Seismic analysis shows jumps in wave velocity due to structural changes in silicates under pressure**
- **Scaling Result**
  - **First use for elastic constants**
  - **8K core vs. 128 on allocated time**

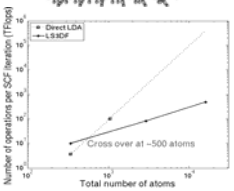


Nanoscience Calculations and Scalable Algorithms

- **Calculation: Linear Scaling 3D Fragment (LS3DF). Density Functional Theory (DFT) calculation numerically equivalent to more common algorithm, but scales with O(n) in number of atoms rather than $O(n^3)$**
- **PI: L.W. Wang, LBNL**

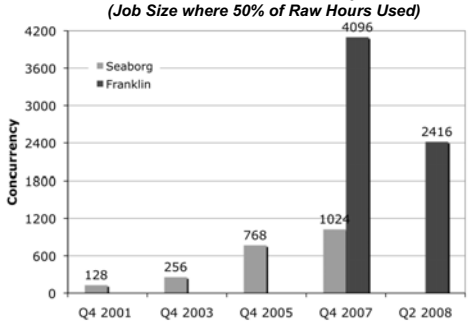
- **Science Results**
  - Calculated dipole moment on 2633 atom CdSe quantum rod, $Cd_{961}Se_{724}H_{948}$.
- **Scaling Results**
  - Ran on 2560 cores
  - Took 30 hours vs many months for $O(n^3)$ algorithm
  - Good parallel efficiency (80% on 1024 relative to 64 procs)





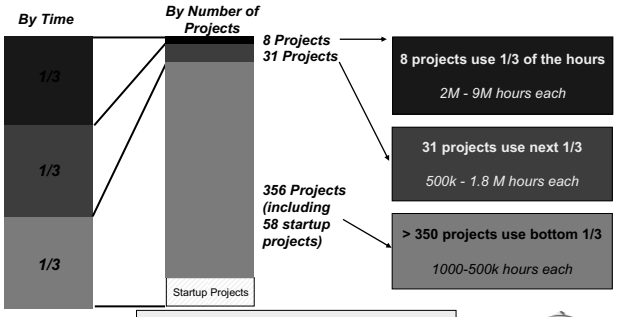
Concurrency Level is Constrained by Availability

Median Job Concurrency
(Job Size where 50% of Raw Hours Used)



NERSC Allocation Breakdown

Total NERSC allocation time divided into thirds



Middle projects have insufficient resources for large production runs

NERSC Response

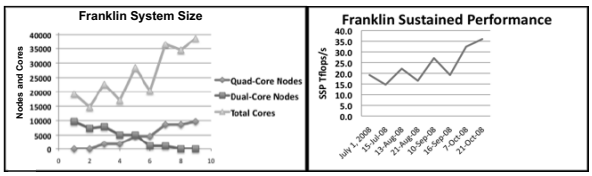
- **Further progress in these key science missions (and others) requires increased computational capability.**
- **NERSC Strategy: Increase user scientific productivity via timely introduction of the best new technologies designed to benefit the broadest subset of the NERSC workload**

=> Upgrade Franklin
=> Commence NERSC-6.

Franklin Quad Core Upgrade

- **In-place, no-interruption upgrade taking place between July and October, 2008.**
- **All 9,672 nodes change from 2.6-GHz dual core to 2.3-GHz quad core.**
- **QC nodes have 8 GB memory, same average GB/core as on DC Franklin.**
- **Memory from 667 MHz to 800 MHz.**

Maintaining Service While Improving Service



Phase	Start Date	Number of Dual Core Racks	Number of Quad Core Racks	Sustained Performance (SSP Tflops/s)	SSP Tflop/s-Days
Before	July 1, 2008	102	0	19.2	
1	15-Jul-08	78	0	14.7	425.8
2a	13-Aug-08	84	18	22.2	177.3
2b	21-Aug-08	54	18	16.5	330.4
3a	10-Sep-08	54	48	27.1	162.6
3b	16-Sep-08	12	48	19.2	403.2
4a	7-Oct-08	0	92	32.5	454.6
4b	21-Oct-08	0	102	36.0	

Measuring Success

"For better or for worse, benchmarks shape a field."
David Patterson, UCB CS267 2004

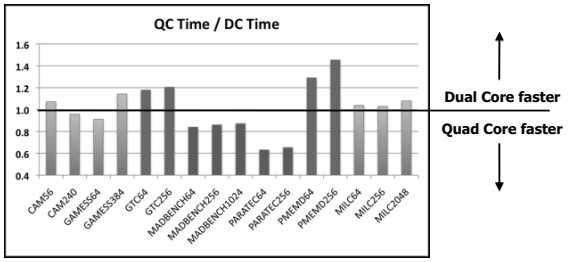
"Benchmarks are only useful insofar as they model the intended computational workload."
Ingrid Bucher & Joanne Martin, LANL, 1982

NERSC-5 Application Benchmarks

Benchmark	*Science Area*	*Algorithm Space*	*Base Case Concurrency*	*Problem Description*	*Lang*	*Libraries*
CAM	Climate (BER)	Navier Stokes CFD	56, 240 Strong scaling	D Grid, (~.5 deg resolution); 240 timesteps	F90	netCDF
GAMESS	Quantum Chem (BES)	Dense linear algebra	64, 384 (Same as Ti-06)	DFT gradient, MP2 gradient	F77	DDI, BLAS
GTC	Fusion (FES)	PIC, finite difference	64, 256 Weak scaling	10 particles per cell	F90	
PMEMD	Life Science (BER)	Particle Mesh Ewald	64, 256 Strong scaling		F90	
MadBench	Astrophysics (HEP & NP)	Power Spectrum Estimation	64,256, 1024 Weak scaling	Vary Npix; 730 MB per task, 200 GB disk	F90	Scalapack, LAPACK
MILC	Lattice Gauge Physics (NP)	Conjugate gradient, sparse matrix; FFT	64, 256, 2048 Weak scaling	16^4 Local Grid, ~4,000 iters	C, assem.	
PARATEC	Material Science (BES)	DFT; FFT, BLAS3	64, 256 Weak scaling	250-686 Atoms, 1372 bands, 10 iters	F90	Scalapack, FFTW

QC / DC Comparison

NERSC-5 Benchmarks

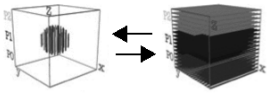


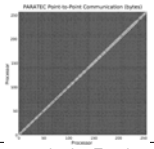
Compare time for *n* cores on DC socket to time for *n* cores on QC socket.

Data courtesy of Helen He, NERSC USG

PARATEC: Parallel Total Energy Code

- **Captures the performance of ~70% of NERSC material science computation.**
- **Planewave DFT; calculation in both Fourier and real space; custom 3-D FFT to transform between.**
- **Uses MPI / SCALAPACK / FFTW / BLAS3**
- **All-to-all data transpositions dominate communications.**





Communication Topology for PARATEC from IPM.

PARATEC: Performance

Medium Problem (64 cores)

	Dual Core	Quad Core	Ratio
FFTs[1]	425	537	1.3
Projectors[1]	4,600	7,800	1.7
Matrix-Matrix[1]	4,750	8,200	1.7
Overall[2]	2,900 (56%)	4,600 (50%)	1.6

- **[1] Rates in MFLOPS/core from PARATEC output.**
- **[2] Rates in MFLOPS/core from NERSC-5 reference count.**
- **Projector/Matrix-Matrix rates dominated by BLAS3 routines.**

=> SciLIB takes advantage of wider SSE in Barcelona-64.

NERSC Sustained Performance

- **7 application benchmarks**
- **Two machines (DC & QC)**

- **How do we summarize performance?**
- **How do we express computing capability over time?**

Sustained System Performance (SSP)

- **Aggregate, un-weighted measure of sustained computational capability relevant to NERSC's workload.**
- **Geometric Mean of the processing rates of seven applications multiplied by *N*, # of cores in the system.**
  - **Largest test cases used.**
- **Uses floating-point operation count predetermined on a reference system by NERSC.**

$$\text{SSP in TFLOPS} = \frac{N * \sqrt[7]{\prod_i P_i}}{1000}$$

NERSC Composite SSP Metric

The time for the largest concurrency run of each full application benchmark is used to calculate the SSP.

NERSC-5 SSP

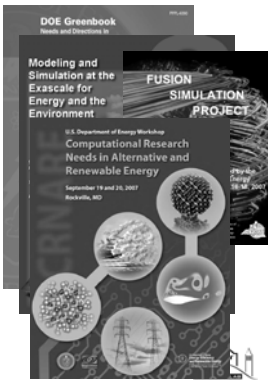
CAM	GAMESS	GTC	PMEMD	MADBench	MILC	PARATEC
240	384	256	256	1024	2048	256
Climate Modeling	Quantum Chemistry	Fusion	MolDyn	Astro	Lattice QCD	MatSci DFT

For Franklin DualCore, (N = 19,344) 19.3 Tflop/s out of 101 Tflop/s peak
QuadCore, (N = 38,640) 37.6 Tflop/s out of 355 Tflop/s peak

NERSC Next-Generation System

NERSC-6 Project Overview

- **Acquire the next major NERSC computing system**
  - **Goal: 70-100 Sustained TF/s on representative applications (NERSC-6 SSP)**
  - **Fully-functional machine accepted in FY10 and available for DOE allocation**
  - **RFP release September 4, 2008.**
  - **Approach designed to select the best machine for science with greatest flexibility for both NERSC and vendors.**



NERSC-6 Benchmarks

- **New codes/methods address evolution of the workload, emerging programming models, algorithms**
  - **New SSP applications: MAESTRO and IMPACT-T**
  - **UPC, AMR, implicit and sparse methods**
  - **Comprehensive workload study:**
  - http://www.nersc.gov/projects/procurements/NERSC6/NERSC6Workload.pdf
- **Largest concurrency increases from 2,048 to 8,196**
  - **Increased focus on strong scaling**
- **Two ways for vendors to run benchmarks…**

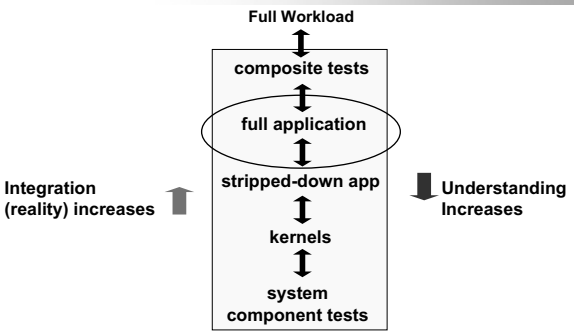
Base Case for Application Runs

- **LCD for comparison among proposed systems.**
- **Limits the scope of optimization.**
  - **Modifications only to enable porting and correct execution.**
- **Limits allowable concurrency to prescribed values.**
- **MPI-only (even if OpenMP directives present).**
- **Fully packed nodes.**
- **Libraries okay (if generally supported).**
- **Hardware multithreading okay, too.**
  - **Expand MPI concurrency to occupy hardware threads.**

Optimized Case for Application Runs

- Allow the Offeror to highlight features of the proposed system.
- Applies to seven SSP apps only, all test problems.
- Examples:
  - Unpack the nodes;
  - Higher (or lower) concurrency than reference base case;
  - Hybrid OpenMP / MPI;
  - Source code changes for data alignment / layout;
  - Any / all of above.
- Caveat: SSP based on total number of processors blocked from other use.

Use a Hierarchy of Benchmarks

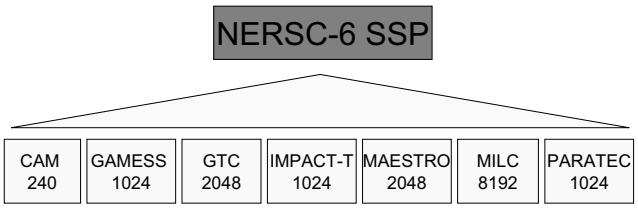


NERSC-6 Application Benchmarks

Benchmark	*Science Area*	*Algorithm Space*	*Base Case Concurrency*	*Problem Description*	*Lang*	*Libraries*
CAM	Climate (BER)	Navier Stokes CFD	56, 240 Strong scaling	D Grid, (~.5 deg resolution); 240 timesteps	F90	netCDF
GAMESS	Quantum Chem (BES)	Dense linear algebra	256, 1024 (Same as Ti-09)	DFT gradient, MP2 gradient	F77	DDI, BLAS
GTC	Fusion (FES)	PIC, finite difference	512, 2048 Weak scaling	100 particles per cell	F90	
IMPACT-T	Accelerator Physics (HEP)	PIC, FFT component	256,1024 Strong scaling	50 particles per cell	F90	FFTW
MAESTRO	Astrophysics (HEP)	Low Mach Hydro; block structured-grid multiphysics	512, 2048 Weak scaling	16 32^3 boxes per proc; 10 timesteps	F90	Boxlib
MILC	Lattice Gauge Physics (NP)	Conjugate gradient, sparse matrix; FFT	256, 1024, 8192 Weak scaling	8x8x8x9 Local Grid, ~70,000 iters	C, assem.	
PARATEC	Material Science (BES)	DFT; FFT, BLAS3	256, 1024 Strong scaling	686 Atoms, 1372 bands, 20 iters	F90	Scalapack, FFTW

NERSC-6 Composite SSP Metric

The largest concurrency run of each full application benchmark is used to calculate the composite SSP metric



For each benchmark measure
- *FLOP counts on a reference system*
- *Wall clock run time on various systems*

Open Benchmark Suite

- Collect data from other HPC centers using application benchmarks
- Starting to run on T2K (Tokyo) cluster this week
  - Compare multi-socket nodes to single socket
  - Compare Myrinet to Cray HSN
  - Compare Hitachi compilers to PGI, Pathscale

Acknowledgements

- Kengo Nakajima, ITC, Tokyo University
- Kathy Yelick, NERSC Director
- John Shalf, Harvey Wasserman, NERSC SDSA
- Nick Cardo, NERSC Franklin Project Lead
- Helen He, Katie Antypas, NERSC USG
- Joe Glenski, et al., Cray

This work was supported by Office of Advanced Scientific and Computational Research, Office of Science, US Department of Energy under Contract No. DE-AC02-05CH11231