Progress Towards Petascale Applications in Biology: Status in 2006

Craig A. Stewart[1], Matthias Mueller[2], Malinda Lingwall[3]

[1]Office of the Vice President for Information Technology, Indiana University, Bloomington, IN; [2]Center for Information Services and High Performance Computing, Technische Universitaet Dresden; [3]University Information Technology Services, Indiana University, Bloomington, IN
stewart@iu.edu, matthias.mueller@tu-dresden.de, mlingwal@indiana.edu

Abstract. Petascale computing is currently a common topic of discussion in the high performance computing community. Biological applications, particularly protein folding, are often given as examples of the need for petascale computing. There are at present biological applications that scale to execution rates of approximately 55 teraflops on a special-purpose supercomputer and 2.2 teraflops on a general-purpose supercomputer. In comparison, Qbox, a molecular dynamics code used to model metals, has an achieved performance of 207.3 teraflops. It may be useful to increase the extent to which operation rates and total calculations are reported in discussion of biological applications, and use total operations (integer and floating point combined) rather than (or in addition to) floating point operations as the unit of measure. Increased reporting of such metrics will enable better tracking of progress as the research community strives for the insights that will be enabled by petascale computing.

Keywords: Computational biology, grand challenge problem, high performance computing, life sciences, peak theoretical capacity, petabytes, petaflops, petascale computing.

1 Introduction

The worldwide high performance computing (HPC) community is at present highly focused on petascale computing – a common topic of discussion in press releases, grant solicitations, conferences, and technical papers. Biology in general and protein structure in particular are often important themes in discussion of petascale computer applications. The government of Japan and the Institute of Physical and Chemical Research (RIKEN) announced in 2003 plans to create a high performance computing system with 1 petaflops peak theoretical capability to model protein folding [1]. In the United States, the National Science Foundation (NSF) and the US Department of Energy (DOE) have each announced programs designed to develop and implement petaflops supercomputers, in both cases with biology among the driving applications. The DOE has announced plans to install a supercomputer with 1 petaflops peak theoretical capability in 2008 [2], while the NSF's target is 1 petaflops sustained

performance achieved by 2010-2011 [3]. Most recently, the RIKEN Institute announced that their Protein Explorer system has been clocked at a peak theoretical capability of 1 petaflops [4]. The era of petascale computing in biology is here – at least by one measure.

The purpose of this paper is to assess the current state of progress toward petascale computing in biology. Petascale is used here to indicate applications that use petaflops of computing power, petabytes of data, or both. We present data combed from the literature on execution rates of applications in biology and other sciences, as well as information on the size of publicly available data sets. Based on examination of the currently available data, we make recommendations about ways in which performance of applications and size of databases could be reported so that the research community could better track progress in capabilities of biological applications.

2 Methods and Materials

There are several ways to measure computational speed: peak theoretical capability (the maximum number of operations that could possibly be completed by a computer given the number of instructions per clock cycle and number of clock cycles per second); peak achieved performance on benchmark applications (especially the Linpack benchmark program, which is used in rankings for the Top500 List of the fastest supercomputers in the world [5]); and peak achieved performance on a "real" applications that solve some current scientific problem.

To assess progress in scale of applications in biology and other disciplines, we combed the literature and the World Wide Web for examples of particularly large computations in biology and, for purposes of comparison, other scientific disciplines. Because there is little consistency in how the performance of large biological applications is reported, we also solicited information directly from leading supercomputing centers. The progress of application performance can be understood only in the context of the progress in the capabilities of hardware systems. For comparisons of hardware capabilities we compiled information on the peak theoretical capability of general and special-purpose supercomputers. Key sources of information included papers about Gordon Bell prizes from the ACM/IEEE SCxy supercomputing conferences [6-9] and the Top500 List [5]. To assess progress toward petascale data used in biology, we examined the current sizes of major public biological data sources.

3 Results

Figure 1 demonstrates the well-understood progress of the peak theoretical capability of the top-ranked system on the Top500 List. In terms of systems that run the Linpack benchmark, statistical extrapolation from all previous Top500 Lists suggests that the top system on that list will reach a peak theoretical capability of 1 petaflops in November 2009 and achieved Linpack performance of 1 petaflops in June 2012.

Figure 1 also shows peak theoretical capability of several special-purpose systems of note. The MD-GRAPE and GRAPE systems are not included on the Top500 list since they perform molecular dynamics and astrophysical N-body calculations, respectively, and cannot run the Linpack benchmark suite. Figure 1 also shows current aggregate TFLOPS for the combined BOINC project [10], and two subcomponents of that system – SETI@Home [11], the largest BOINC project overall, and ROSETTA@Home [12], the largest biological application within the BOINC system for which aggregate performance data are available. Table 1 details the systems shown in Figure 1.

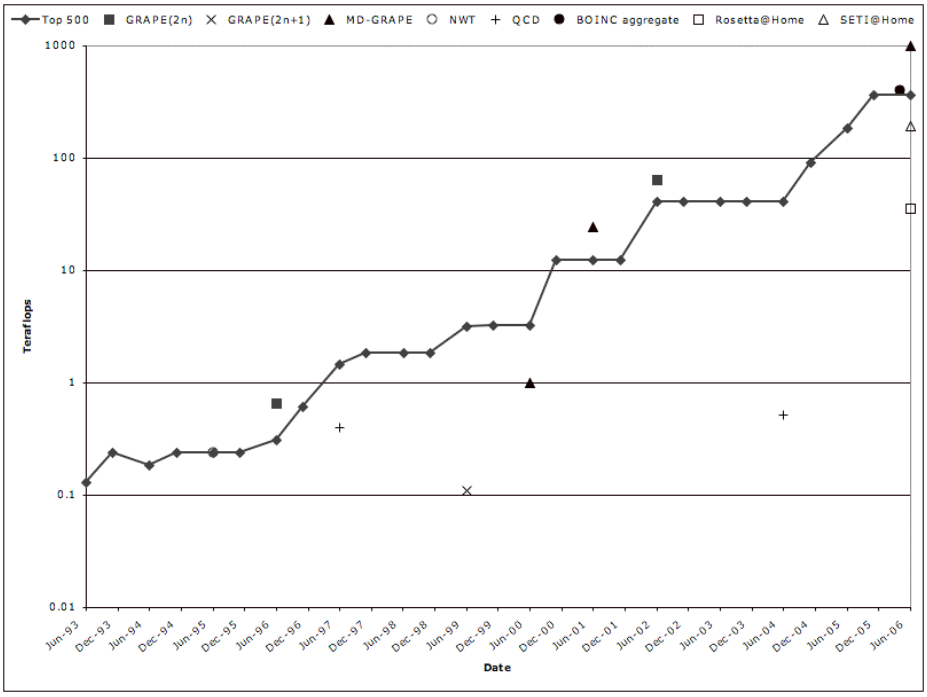

Fig. 1. Peak theoretical capacity of high performance computing systems over time. Shown are the peak theoretical capacity of the #1 ranked system on the Top500 List since its inception, along with the peak theoretical capability of selected special-purpose computing systems. Special-purpose systems represented include the Numerical Wind Tunnel, GRAPE family, MD-GRAPEs, specialized QCD systems, and distributed BOINC applications [4], [5], [8], [10-19].

Table 1. Data about systems in Figure 1.

System	Classification	Peak theoretical capacity	Year	Reference
MDGRAPE-3	MD-GRAPE	1 PF	2006	4
BOINC combined statistics	BOINC aggregate	400.85 TF	2006	10
SETI@Home	SETI@Home	191.233 TF	2006	11
GRAPE-6	GRAPE(2n)	63.4 TF	2002	13
Rosetta@Home	Rosetta@Home	35.654 TF	2006	12
MDGRAPE-2	MD-GRAPE	24.6 TF	2001	14
MDGRAPE-2	MD-GRAPE	1 TF	2000	15
GRAPE-4	GRAPE(2n)	0.66 TF	1996	8
QCDOC	QCD	0.512 TF	2004	16
QCDSP	QCD	0.4 TF	1997	17
Numerical Wind Tunnel	NWT	0.2 TF	1995	18
GRAPE-5	GRAPE(2n+1)	0.11 TF	1999	19

Figure 2 shows progress in sustained performance on several applications since the inception of the Top500 List. Included are the top achieved Linpack performance from the Top500 List and the top performance achieved on several heroic applications. Table 2 details the applications shown in Figure 2.

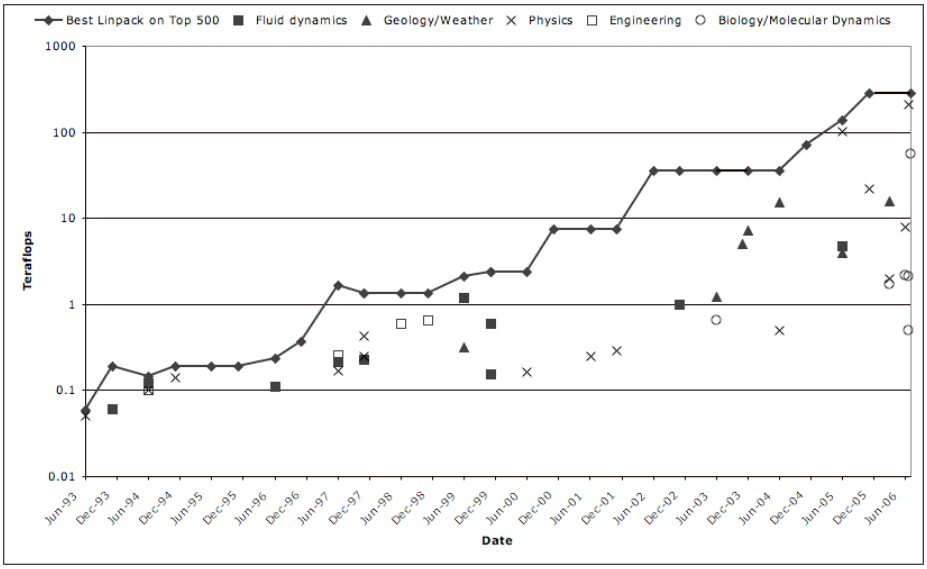

Fig. 2. Achieved floating point computation rates for applications in several disciplines. Included are the Linpack performance data of the #1 system on the Top500 List since its inception, and other applications that have reported high floating point operation rates. [5-9], [20-44]

Table 2. Data about applications in Figure 2.

Application	Discipline	Peak achieved rate	Year	Reference
Qbox	Physics	207.3 TF	2006	20
Solidification simulations	Physics	103 TF	2005	21
Peptide simulation	Biology/Molecular dynamics	55 TF	2006	22
Qbox	Physics	22.02 TF	2005	23
Corona simulation	Geology/Weather	15.6 TF	2006	24
Earth Simulator	Geology/Weather	15.2 TF	2004	25
LSMS	Physics	8 TF	2006	26
Weather forecast (NWS)	Geology/Weather	7.3 TF	2003	27
Earth Simulator	Geology/Weather	5 TF	2003	28
Lattice Boltzmann model	Fluid dynamics	4.7 TF	2005	29
Weather forecast (NOAA)	Geology/Weather	4 TF	2005	30
Blue Matter	Biology/Molecular dynamics	2.2 TF	2006	31
NAMD	Biology/Molecular dynamics	2.08 TF	2006	32
VASP	Physics	2 TF	2006	33
CPMD	Biology/Molecular dynamics	1.7 TF	2006	33
Wave propagation solver	Geology/Weather	1.21 TF	2003	34
Turbulence simulation	Fluid dynamics	1.18 TF	1999	35
DOWSER	Fluid dynamics	1 TF	2002	36
First principles calculation	Engineering	0.657 TF	1998	37
NAMD	Biology/Molecular dynamics	0.65 TF	2003	38
Parallel Eigensolver	Engineering	0.605 TF	1998	39
Turbulence simulation	Fluid dynamics	0.6 TF	1999	35
NAMD	Biology/Molecular dynamics	0.5 TF	2006	32
Finite element analyses	Physics	0.5 TF	2004	40
Tree-code method	Physics	0.43 TF	1997	9
Hairpin vortices simulation	Geology/Weather	0.319 TF	1999	41
Cactus	Physics	0.292 TF	2001	42
MP-QUEST	Engineering	0.256 TF	1997	9
Cactus	Physics	0.249 TF	2001	42
Quark modeling	Physics	0.246 TF	1997	9
Pronto	Fluid dynamics	0.225 TF	1997	9
MPSalsa	Fluid dynamics	0.212 TF	1997	9
Tree-code method	Physics	0.17 TF	1997	9
Bunyip	Physics	0.163 TF	2000	43
Unstructured mesh CFD	Fluid dynamics	0.156 TF	1999	44
Sound wave computation	Physics	0.143 TF	1994	7
Numerical Wind Tunnel	Fluid dynamics	0.12 TF	1994	7
Numerical Wind Tunnel	Fluid dynamics	0.111 TF	1996	8
Composite modeling	Engineering	0.1 TF	1994	7
Radar scattering	Physics	0.1 TF	1994	7
Boltzmann equation	Fluid dynamics	0.06 TF	1993	6
Crack modeling on CM-5	Physics	0.05 TF	1993	6

We collected information about the size of data sets used in several fields of research in order to study progress in data-centric life sciences research as compared to other disciplines. Table 3 shows the sizes of several important data sets. In many

cases these databases tend to report their size in terms of numbers of records (or in the case of sequence databases number of sequences). Indiana University maintains a repository of copies of many of these data sets, and we determined the size in petabytes of these data sets from those copies.

Table 3. Current size of some exemplars of databases used in the life sciences as of summer 2006, compared with key exemplars from other disciplines. The size of datasets marked with an * were determined from copies of data downloaded to Indiana University from the original resources.

Database name	Discipline	Current estimated size
BaBar	High-energy physics	2 PB [45]
National Virtual Observatory	Astronomy	> 0.5 PB [46]
NCBI*	Biology	~ 0.005 PB
Regenstrief Medical Records System	Medicine	0.004 PB [47]
Protein Data Bank	Biology	0.0007 PB [48]
EarthScope	Geology	0.0004 PB [49]
PubChem*	Chemistry	~ 0.0001 PB
Swiss-Prot*	Biology	0.00000087 PB

4 Discussion

There are notable accomplishments in terms of peak performance of biological applications. The top performance in terms of floating point execution rate that we have been able to find for a biological application is 55 teraflops on a special-purpose MDGRAPE-3 system with a peak theoretical capability of 415 teraflops (an efficiency of 13.25%) [22]. This application simulated the formation of amyloid fibrils including 14 million atoms. The top performance in terms of floating point execution rates on a general-purpose supercomputer is approximately 2.2 teraflops with Blue Matter software on 80% of an 11.5 teraflops Blue Gene/L supercomputer (an efficiency of approximately 24%) [31], using the 92,000 atom ApoA1 benchmark. (The Blue Matter software is discussed in this volume in the paper by Fitch et al, "Progress in Scaling Biomolecular Simulations to Petaflop Scale Platforms.") Another application of note in terms of instruction rate is NAMD, which can operate at 2.08 teraflops in a 2.7 million atom simulation on a system with a peak theoretical capacity of 9.83 teraflops (an efficiency of approximately 21%) [32]. Based on the data we have been able to obtain, these seem to be the top biologically-oriented applications in terms of rates of floating point executions. There is a fairly strong contrast between the achieved rate of floating point operations on biological codes, the peak theoretical performance of systems available today, and the peak achieved performance on other scientific applications.

The progress of the peak theoretical capability of HPC systems, and of Linpack performance on these systems, is progressing steadily toward petascale computing. Special-purpose systems based on GRAPE and MD-GRAPE boards have on several occasions managed faster peak theoretical capability than the top system on the Top500 List. This trend is in evidence at present, as the MDGRAPE-3 is the basis for

the RIKEN Institute's Protein Explorer, the first system with a reported peak theoretical capability of 1 petaflops. The fastest supercomputer in the world according to the June 2006 Top500 List (among those capable of running the Linpack Benchmark) is the 367 teraflops IBM BlueGene/L system at Lawrence Livermore National Laboratory, larger than but otherwise similar to the system used for the Blue Matter software calculations mentioned above. Plans announced by the US Department of Energy and National Science Foundation will thus result in implementation of systems of 1 petaflops peak theoretical capability (2008) and 1 petaflops achieved performance (2010-2011) more quickly than would be predicted on the basis of extrapolation from the existing Top500 list data.

In terms of performance of applications other than Linpack, the highest rate of floating point executions reported to date are from simulation of crack formation in 1,000 Molybdenum atoms with the Qbox application [20], [23]. Qbox on the 367 teraflops LLNL BlueGene/L system has achieved a peak execution rate of 207.3 teraflops – 56.5% of peak theoretical capability (as compared to 73.8% of peak achieved on the Linpack benchmark). Another notable physics application is LSMS [26], which ran on Pittsburgh Supercomputer Center's Cray XT3 at just over 8 teraflops – 82% of peak theoretical capability (as compared to 80.2% of peak achieved on the Linpack benchmark.) This LSMS run performed an ab initio quantum calculation of an iron nanoparticle of more than 4,400 atoms.

There seem to be fewer data available at present regarding high rates of floating point executions for heroic biological applications – and fewer than seem available for other disciplines. This is at least in part because HPC applications in biology have been in existence for less time (and are still less prevalent) than disciplines such as material science, physics, and computational fluid dynamics. In addition, performance results for biological codes are most often reported in ways that are directly meaningful to the time to solution of the particular problem at hand. Wall clock times, and decreases thereof, to solve a particular problem are perhaps the most common metric overall; total CPU hours used is also a common metric. In the case of protein folding, wall clock time per time step (or simulated time steps per unit wall clock time) is often used. In the case of genome sequence comparisons, number of sequences compared per unit time is a common metric. In the case of phylogenetic inference, the number of evolutionary trees analyzed per wall clock hour is commonly used. Researchers in the life sciences often do not collect and report the performance of their applications in terms of floating point operations. For example, two of the authors of this paper participated in an HPC Challenge project at SC2003, in which many collaborators created a global computational grid to run fastDNAml, a program for inferring evolutionary relationships [50]. We reported our results in terms of rate of analysis of trees, total number of processors used, etc. but did not instrument the code to measure actual floating point executions. Had we tried to do so, we would not have managed to get the application running during the time period of the HPC Challenge at SC2003. Similarly, high throughput applications such as Folding@home [51] and fightAIDS@home [52] involve thousands of computers working simultaneously on particular parts of a large-scale biological problem, but the rate at which work is done is not reported in terms of floating point calculation rates.

Floating point operation rates are mentioned specifically in major grant solicitations, and are thus of some practical import to the high performance computing

community [53]. However, rates of floating point operations have two limitations as a measure of biological applications. One is that improving time to solution may involve decreasing execution rates. For example, the floating point rates for NAMD today are roughly 30% lower than in the code version in 2002 [38] because the underlying algorithms are more efficient [32].

A second limitation, perhaps more specific to biological operations, is the relative importance of integer operations in biological applications. The performance of the DOTTER program [54] was carefully analyzed in terms of total operations because of the predominance of integer mathematics in that application [55]. Understanding the application performance was possible only by including integer operations in the analysis. BLAST and other important bioinformatics applications also use integer operations extensively. Roughly two thirds of the mathematical operations in NAMD are integer operations [32]. To the extent that execution rates provide a means to compare the behavior of diverse biological applications, total operation rates (integer and floating point) would likely be a better basis for comparison than floating point operation rates alone. This poses the question of how to factor in the importance of operand length. Double precision reals are the basis for the standard Linpack benchmark, and there seems little reason at present to deviate from that approach in general (although there may be interesting exceptions [56]). As regards integer operations, when reporting rates it is probably best to specify the integer length – but it may make sense in the context of biological applications to count operations without regard to operand length. To do otherwise and somehow correct for length would likely penalize clever coding schemes that take advantage, for example, of the four letter alphabet of nucleotides (A,C,G,T).

In addition to measuring rates of operation execution, it will likely be useful to measure the total amount of computation that contributed to a particular analysis or simulation. For example, some of the largest biological computations performed to date in terms of total computer operations involve NAMD simulations of an entire ribosome in 2005 [57] and the tobacco mosaic satellite virus [58]. The former seems to be the largest simulation of a biological structure (in terms of CPU hours) ever published; the latter is the first ever molecular simulation of an entire life form. A useful measure of total calculation effort comparable across applications and systems might be simply total operations, or a measure analogous to the kilowatt-hour – that is, the PetaOPS-hour. Given the diversity of biologically oriented applications, it simply may not be possible to capture the performance of applications with a single metric. However, reporting total operation rates (integer and floating point) and total operation counts or PetaOPS-hours, in addition to other measures, will enable better comparisons among biological applications. Such comparisons are only a means, and the ends desired are biological insights rather than high operation rates. Still, tracking the progress of operation rates as a means will enable us to better determine if the oft-discussed ends (new insights and knowledge) are in evidence as the capabilities of our means progress.

The sizes of public biological data sets are growing rapidly, but life sciences data sets are still well away from the petabytes range and well smaller than the size of data sets found in other disciplines. Data sets in the range of 2 petabytes are available now in high energy physics research with 20 petabytes planned by 2008 [45]. The Terashake earthquake simulation run at the San Diego Supercomputer Center

generated a data set of 45 TB [59]. In contrast, the largest publicly available biological data set is at present approximately 5 TB. Graphs of the amount of data contained in NCBI's Genbank data set show dramatic rates of growth [60], and that dramatic rate of growth creates an impression that may obscure the size of the actual data set: in 2006, the actual aggregate size of the data set is still well under a terabyte. Likewise, a recent demonstration at Indiana University included an analysis of some of the chemical properties of all of the compounds in PubChem [61] in less than 10 minutes – a significant accomplishment from the standpoint of obtaining information from a comparatively large data set (more than 19 million records), yet the input data amounted to less than 100GB. Other very large and notable data-centric initiatives in the life sciences include BIRN [62], eDiaMoND [63], and NEON [64]. Aggregated sets of data in clinical practice and held by pharmaceutical companies may be much larger. For example, the Regenstrief Institute [47] holds an aggregate of 4 TB of clinical data. While reporting of biological database size in number of records, or number of sequences, or number of compounds is common, more routine reporting of database size in terms of actual disk storage space would be useful in comparability across disciplines in discussing the size of data sets.

Sterling et al [65] produced the first careful analysis of the opportunities and challenges in achieving petascale computing. In their 1994 workshop, they identified several candidates for petaflops applications, including protein folding, modeling of circulation in the human body, and data-intensive applications using petabytes or exabytes of data. Stevens [66], CIBIO [67], and Atkins [68] provide more recent analyses of opportunities for petascale biological applications. Stevens [66] outlined eight categories of potential petascale applications; of these, five categories were related to molecular structure, function, and dynamics; other categories included sequence analysis, whole genome metabolic modeling, and population modeling. A recent NSF-sponsored workshop on petascale applications in biology reinforced many of these ideas, and added novel ideas such as ecological simulations linked to climage models and real-time patient profiling [69].

Based on data currently available, molecular dynamics codes clearly scale to the highest operation rates achieved on monolithic supercomputers and are likely candidates to be the first applications to achieve petaops calculation rates. One model of circulatory function in the human body – ATREE – creates large-scale models of biological function by employing computational physics codes (including turbulence) to solve biological problems. These codes have been implemented on the NSF-funded TeraGrid [70], linking simulation of many components of the human arterial system. By linking many HPC systems ATREE is a likely candidate to achieve extremely high mathematical operation rates in a grid environment. In terms of data-intensive applications, several examples given by Stevens [66] involve coarse-grained (and often very complex) parallel analyses of large data sets; such data-parallel applications are also good candidates for achieving very high operation rates. All in all, the current state of affairs is consistent with many of the predictions made by Sterling et al. more than a decade ago.

5 Conclusion

There are many ways to count what are petascale applications in biology; by one measure at least the era of petascale biology begins in 2006 with the successful operation of the Protein Explorer at a peak theoretical capability of 1 petaflops. Many obstacles remain between the state of the art in 2006 and biological applications that achieve petaops calculation rates and process petabytes of data. In tracking the progress toward petascale biological applications it will be helpful to report application characteristics in ways that will enable better comparisons across applications. For applications, routine reporting of calculation rates in terms of total petaoperations per second, and total computing power in petaoperations or PetaOPS-hours for particular simulations, would be helpful. For data-intensive applications, more routine reporting of data set size in tera- or petabytes would be helpful. Petascale applications are only a means to an end; the ends are new insights about the function of biological systems and better human health. Still, tracking progress of the means will enable some insight as to whether the ends anticipated are being achieved.

Acknowledgements. This research was supported in part by the Indiana Genomics Initiative and the Indiana Metabolomics and Cytomics Initiative. The Indiana Genomics Initiative of Indiana University and the Indiana Metabolomics and Cytomics Initiative of Indiana University are supported in part by Lilly Endowment, Inc. The authors also wish to thank IBM, Inc. for support via Shared University Research Grants and partnerships via IU's relationship as an IBM Life Sciences Institute of Innovation. Indiana University also thanks the TeraGrid partners; IU's participation in the TeraGrid is funded by National Science Foundation grant numbers 0338618, 0504075, and 0451237. The early development of this paper was supported by a Fulbright Senior Scholars award from the Council for International Exchange of Scholars (CIES) and the United States Department of State to Dr. Craig A. Stewart; Matthias Mueller and the Technische Universität Dresden were hosts. Many reviewers contributed to the improvement of the ideas expressed in this paper and are gratefully appreciated; Thom Dunning, Robert Germain, Chris Mueller, Jim Phillips, Richard Repasky, Ralph Roskies, and Allan Snavely are thanked particularly for their insights.

References

1. Taiji M., Narumi T., Ohno Y., Futatsugi N., Suenaga A., Takada N., Konagaya A. "Protein Explorer: A Petaflops Special-Purpose Computer System for Molecular Dynamics Simulations," *sc*, p. 15, ACM/IEEE SC 2003 Conference (SC'03), 2003. http://csdl.computer.org/dl/proceedings/sc/2003/2113/00/21130015.pdf
2. "Vendor Spotlight: Cray to Deliver Petaflop Supercomputer to ORNL in 2008." HPCwire. 16 June 2006. Accessed 30 August 2006. http://www.hpcwire.com/hpc/694425.html
3. NSF Cyberinfrastructure Council. "NSF's Cyberinfrastructure Vision for 21st Century Discovery." 20 July 2006. Accessed 31 August 2006. http://www.nsf.gov/od/oci/ci-v7.pdf

4. Taiji M., Yamashita Y., Nakanishi T. “Completion of a one-petaflops computer system for simulation of molecular dynamics.” 2006. Press release. Accessed 17 August 2006. http://www.riken.go.jp/engn/r-world/info/release/press/2006/060619/index.html
5. Top500 Supercomputer Sites. Accessed 1 September 2006. http://top500.org/lists
6. Karp A.H., Heller D., Simon H. “1993 Gordon Bell Prize Winners.” *IEEE Computer*, January 1994, pp. 69-75. http://csdl.computer.org/dl/mags/co/1994/01/r1069.pdf
7. Karp A.H., Heath M., Heller D., Simon H. “1994 Gordon Bell Prize Winners.” *IEEE Computer,* January 1995, pp. 68-74. http://csdl.computer.org/dl/mags/co/1995/01/r1068.htm
8. Karp A.H., Geist A., Bailey D. “1996 Gordon Bell Prize Winners.” *IEEE Computer*, January 1997, pp. 80-85. http://csdl.computer.org/dl/mags/co/1997/01/r1080.htm
9. Karp A.H., Lusk E., Bailey D.H. “1997 Gordon Bell Prize Winners.” *IEEE Computer,* January 1998, pp. 86-92. http://csdl.computer.org/dl/mags/co/1998/01/r1086.htm
10. BOINC combined statistics (BOINCstats). Accessed 20 September 2006. http://www.boincstats.com/stats/project_graph.php?pr=bo
11. SETI@Home (BOINCstats). Accessed 20 September 2006. http://www.boincstats.com/stats/project_graph.php?pr=sah
12. Rosetta@Home (BOINCstats). Accessed 20 September 2006. http://www.boincstats.com/stats/project_graph.php?pr=rosetta
13. Makino J., Kokubo E., Fukushige T., Daisaka H. “A 29.5 Tflops simulation of planetesimals in Uranus-Neptune region on GRAPE-6.” *sc*, p. 34, ACM/IEEE SC 2002 Conference (SC'02), 2002. http://csdl.computer.org/dl/proceedings/sc/2002/1524/00/15240034.pdf
14. Narumi T., Kawai A., Koishi T. "An 8.61 Tflop/s Molecular Dynamics Simulation for NaCl with a Special-Purpose Computer: MDM." *sc*, p. 11, ACM/IEEE SC 2001 Conference (SC'01), 2001. http://csdl.computer.org/dl/proceedings/sc/2001/1990/00/19900011.pdf
15. Narumi T., Susukita R., Koishi T., Yasuoka K., Furusawa H., Kawai A., Ebisuzaki T. "1.34 Tflops Molecular Dynamics Simulation for NaCl with a Special-Purpose Computer: MDM." *sc*, p. 54, ACM/IEEE SC 2000 Conference (SC'00), 2000. http://csdl.computer.org/dl/proceedings/sc/2000/9802/00/98020054.pdf
16. Boyle P.A., Chen D., Christ N.H., Clark M., Cohen S., Dong Z., Gara A., Joo B., Jung C., Levkova L., Liao X., Liu G., Mawhinney R.D., Ohta S., Petrov K., Wettig T., Yamaguchi A., Cristian C. "QCDOC: A 10 Teraflops Computer for Tightly-Coupled Calculations." *sc*, p. 40, ACM/IEEE SC 2004 Conference (SC'04), 2004. http://csdl.computer.org/dl/proceedings/sc/2004/2153/00/21530040.pdf
17. Chen D., Chen P., Christ N.H., Edwards R.G., Fleming G., Gara A., Hansen S., Jung C., Kahler A., Kasow S., Kennedy A.D., Kilcup G., Luo Y.B., Malureanu C., Mawhinney R.D., Parsons J., Sexton J., Sui C., Vranas P. "QCDSP: A Teraflop Scale Massively Parallel Supercomputer." *sc*, p. 52, ACM/IEEE SC 1997 Conference (SC'97), 1997. http://csdl.computer.org/dl/proceedings/sc/1997/1982/00/19820052.pdf
18. Yoshida M., Nakamura A., Fukuda M., Nakamura T., Hioki S. "Quantum Chromodynamics Simulation on NWT." *sc*, p. 65, ACM/IEEE SC 1995 Conference (SC'95), 1995. http://csdl.computer.org/dl/proceedings/sc/1995/2568/00/25680065.pdf
19. Kawai A., Fukushige T., Makino J. "$7.0/Mflops Astrophysical N-Body Simulation with Treecode on GRAPE-5." *sc*, p. 67, ACM/IEEE SC 1999 Conference (SC'99), 1999. http://csdl.computer.org/dl/proceedings/sc/1999/1966/00/19660067.pdf
20. Johnston D., Smith J., Acocella K. “NNSA announces new mark for world’s fastest supercomputer.” Press release. 22 June 2006. Accessed 12 September 2006. http://www.llnl.gov/pao/news/news_releases/2006/NR-06-06-07.html

21. Streitz F.H., Glosli J.N., Patel M.V., Chan B., Yates R.K., deSupinski B.R., Sexton J., Gunnels J.A. “100+ TFlop Solidification Simulations on BlueGene/L.” November 2005. SC 2005. 14 August 2006. http://sc05.supercomputing.org/schedule/pdf/pap307.pdf
22. Narumi T., Ohno Y., Okimoto N., Koishi T., Suenaga A., Futatsugi N., Yanai R., Himeno R., Fujikara S., Taiji M., Ikei M. “A 55 TFLOPS Simulation of Amyloid-forming Peptides from Yeast Prion Sup35 with the Special-purpose Computer System MDGRAPE-3.” To appear in *sc*, ACM/IEEE SC 2006 Conference (SC'06), 2006.
23. Gygi F., Yates R.K., Lorenz J., Draeger E.W., Franchetti F., Ueberhuber C.W., de Supinski B.R., Kral S., Gunnels J.A., Sexton J.C.. "Large-Scale First-Principles Molecular Dynamics simulations on the BlueGene/L Platform using the Qbox code," *sc*, p. 24, ACM/IEEE SC 2005 Conference (SC'05), 2005.
24. “SDSC Helps Scientists Accurately Simulate Sun’s Corona.” 3 August 2006. Accessed 11 October 2006. http://www.sdsc.edu/Press/2006/08/080306_corona.html
25. Kageyama A., Kameyama M., Fujihara S., Yoshida M., Hyodo M., Tsuda Y. "A 15.2 TFlops Simulation of Geodynamo on the Earth Simulator," *sc*, p. 35, ACM/IEEE SC 2004 Conference (SC'04), 2004. http://csdl.computer.org/dl/proceedings/sc/2004/2153/00/21530035.pdf
26. “Science, the XT3 and TeraGrid: An Interview with PSC Scientific Directors Michael Levine and Ralph Roskies.” Pittsburgh Supercomputing Center. June 2006. Accessed 13 September 2006. http://www.psc.edu/publicinfo/news/2006/2006-06-09-xt3.php
27. Handwerk, B. “Faster Supercomputers Aiding Weather Forecasts.” National Geographic News. 29 August 2005. Accessed 11 October 2006. http://news.nationalgeographic.com/news/2005/08/0829_050829_supercomputer.html
28. Komatitsch D., Tsuboi S., Ji C., Tromp J., "A 14.6 billion degrees of freedom, 5 teraflops, 2.5 terabyte earthquake simulation on the Earth Simulator," *sc*, p. 4, ACM/IEEE SC 2003 Conference (SC'03), 2003. http://csdl.computer.org/dl/proceedings/sc/2003/2113/00/21130004.pdf
29. Lammers P., Wellein G., Zeiser T., Hager G. “Have the Vectors the Continuing Ability to Parry the Attack of the Killer Micros.” In: Resch M., Bönisch T., Benkert K., Furui T., Seo Y., Bez W. (eds) High Performance Computing on Vector Systems, Volume 1, 25-37. Springer:2006.
30. Curns, T. “WEATHER FORECASTING ON A 'REMOTE' SUPERCOMPUTER?” HPCwire. 13 August 2004. Accessed 11 October 2006. http://www.hpcwire.com/hpcwire/hpcwireWWW/04/0813/108178.html
31. Germain, Robert. "Re: Two requests." Personal correspondence. 29 August 2006.
32. Phillips, Jim. "Re: Fwd: NAMD performance in FLOPS?" Personal correspondence. 29 August 2006.
33. Tiyyagura S.R. et al. “TERAFLOPS Sustained Performance with Real World Applications.” Accepted for publication in: “Performance Characterization of the World’s Most Powerful Supercomputers,” special issue of the International Journal of High Performance Computing Applications (IJHPCA). Guest-edited by L. Oliker and R. Biswas. To appear 2007.
34. Akcelik V., Bielak J., Biros G., Epanomeritakis I., Fernandez A., Ghattas O., Kim E.J., Lopez J., O’Hallaron D., Tu G. Urbanic J. "High Resolution Forward And Inverse Earthquake Modeling on Terascale Computers," *sc*, p. 52, ACM/IEEE SC 2003 Conference (SC'03), 2003. http://csdl.computer.org/dl/proceedings/sc/2003/2113/00/21130052.pdf
35. Mirin A.A., Cohen R.H., Curtis B.C., Dannevik W.P., Dimits A.M., Duchauneau M.A., Eliason D.E., Schikore D.R., Anderson S.E., Porter D.H., Woodward P.R., Shieh L.J., White S.W. "Very High Resolution Simulation of Compressible Turbulence on the IBM-SP System." *sc*, p. 70, ACM/IEEE SC 1999 Conference (SC'99), 1999. http://csdl.computer.org/dl/proceedings/sc/1999/1966/00/19660070.pdf

36. Tajkhorshid, E., Nollert, P., Jensen, M. O., Miercke, L. J., O'Connell, J., Stroud, R. M., and Schulten, K. "Control of the Selectivity of the Aquaporin Water Channel Family by Global Orientational Tuning." (2002) Science 296, 525-530
37. “Gordon Bell Prize Winners.” June 2000. SC2000. 14 August 2006. http://www.sc2000.org/bell/pastawrd.htm
38. Tajkhorshid E., Aksimentiev A., Balabin I., Gao M., Isralewitz B., Phillips J.C., Zhu F., Schulten K. “Large scale simulation of protein mechanics and function.” In: Frederic M. Richards, David S. Eisenberg, and John Kuriyan, editors, *Advances in Protein Chemistry*, volume 66, pp. 195-247. Elsevier Academic Press, New York, 2003.
39. Sears M.P., Stanley K., Henry G. "Application of a High Performance Parallel Eigensolver to Electronic Structure Calculations." *sc*, p. 54, ACM/IEEE SC 1998 Conference (SC'98), 1998. http://csdl.computer.org/comp/proceedings/sc/1998/8707/00/87070054.pdf
40. Adams M.F., Bayraktar H.H., Keaveny T.M., Papadopoulos, P. “Ultrascalable Implicit Finite Element Analyses In Solid Mechanics With Over A Half a Billion Degrees of Freedom.” *sc*, p. 34, ACM/IEEE SC 2004 Conference (SC'04), 2004. http://csdl.computer.org/dl/proceedings/sc/2004/2153/00/21530034.pdf
41. Tufo H.M., Fischer P.F., Papka M.E., Blom K. "Numerical Simulation and Immersive Visualization of Hairpin Vortices." *sc*, p. 62, ACM/IEEE SC 1999 Conference (SC'99), 1999. http://csdl.computer.org/dl/proceedings/sc/1999/1966/00/19660062.pdf
42. Allen G., Dramlitsch T., Foster I., Karonis N.T., Ripeanu M., Seidel E., Toonen B. “Supporting Efficient Execution in Heterogeneous Distributed Computing Environments with Cactus and Globus.” *sc*, p. 52, ACM/IEEE SC 2001 Conference (SC'01), 2001. http://csdl.computer.org/dl/proceedings/sc/2001/1990/00/19900052.pdf
43. “ANU DCS Technical Services Group.” 2000. Australian National University. Accessed 14 August 2006. http://tsg.anu.edu.au/Projects/Beowulf/
44. Anderson W.K., Gropp W.D., Kaushik D.K., Keyes D.E., Smith B.F., "Achieving High Sustained Performance in an Unstructured Mesh CFD Application," *sc*, p. 69, ACM/IEEE SC 1999 Conference (SC'99), 1999. http://csdl.computer.org/dl/proceedings/sc/1999/1966/00/19660069.pdf
45. Teige, Scott. “Re: Question about HEP databases.” Personal correspondence. 31 August 2006.
46. Hanisch, Robert. “Re: [nvo-feedback] Amount of data currently accessible through NVO?” Personal correspondence. 31 August 2006.
47. Miller, Theda. “Re: [Fwd: Size of RMRS database(s)?]” Personal correspondence. 31 August 2006.
48. Research Collaboratory for Structural Bioinformatics. “Protein Data Bank Annual Report for July 2004 – June 2005.” Accessed 15 September 2006. http://www.rcsb.org/pdbstatic/general_information/news_publications/annual_reports/annual_report_year_2005.pdf
49. “EarthScope Distribution Statistics from the IRIS DMC.” Accessed 15 September 2006. http://www.iris.edu/earthscope/stats/
50. Stewart C.A., Hart D., Aumüller M., Keller R., Müller M., Li H., Repasky R., Sheppard R., Berry D.K., Hess M., Wössner U., Colbourne J. “A Global Grid for Analysis of Arthropod Evolution.” *grid*, pp. 328-337, Fifth IEEE/ACM International Workshop on Grid Computing (GRID'04), 2004. http://csdl.computer.org/dl/proceedings/grid/2004/2256/00/22560328.pdf
51. “Folding@Home Stats.” Folding@home distributed computing. 2006. Accessed 14 August 2006. http://folding.stanford.edu/stats.html
52. fightAIDS@home. Accessed 31 August 2006. http://fightaidsathome.scripps.edu/
53. NSF Solicitation 06-573. “Leadership-Class System Acquisition - Creating a Petascale Computing Environment for Science and Engineering.” 5 June 2006. Accessed 31 August 2006. http://nsf.gov/funding/pgm_summ.jsp?pims_id=13649

54. Dotter: A dot-matrix program with interactive greyscale rendering for genomic DNA and Protein sequence analysis. Accessed 1 September 2006. http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html
55. Mueller C., Dalkilic M., Lumsdaine A. "High-Performance Direct Pairwise Comparison of Large Genomic Sequences," *ipdps*, p. 199a, 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) - Workshop 7, 2005. http://csdl.computer.org/dl/proceedings/ipdps/2005/2312/08/23120199a.pdf
56. Feldman, M. "Less is More: Exploiting Single Precision Math in HPC." HPCwire. 16 June 2006. Accessed 21 September 2006. http://www.hpcwire.com/hpc/692906.html
57. Sanbonmatsu K.Y., Joseph S., Tung C.S. "Simulating movement of tRNA into the ribosome during decoding." Proc Natl Acad Sci U S A. 2005 Oct 25.
58. Freddolino, P.L. et al. "Molecular dynamics simulations of the complete satellite tobacco mosaic virus." Structure 14, 437-449 (2006).
59. Tooby P. "TeraShake: Simulating the BIG ONE on the San Andreas Fault." *EnVision* Volume 20, Number 1, 2004. pp 4-7. http://www.npaci.edu/envision/v20.1/Envision-2004.pdf
60. Genbank Statistics. Accessed 1 September 2006. http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html
61. PubChem. Accessed 1 September 2006. http://pubchem.ncbi.nlm.nih.gov/
62. Biomedical Informatics Research Network (BIRN). Accessed 1 September 2006. http://www.nbirn.net/
63. eDiaMoND grid computing project. Accessed 1 September 2006. http://www.ediamond.ox.ac.uk/index.html
64. "NEON: National Ecological Observatory Network." Accessed 15 September 2006. http://www.neoninc.org/
65. Sterling T., Messina P., and Smith P.H. *Enabling Technologies for Petaflops Computing*. Cambridge, Massachusetts: MIT Press, 1995.
66. Stevens R. "Trends in Cyberinfrastructure for Bioinformatics and Computational Biology." *CTWatch QUARTERLY* Volume 2, Number 3, August 2006. http://www.ctwatch.org/quarterly/articles/2006/08/trends-in-cyberinfrastructure-for-bioinformatics-and-computational-biology/
67. "Building a Cyberinfrastructure for the Biological Sciences (CIBIO): A BIO Advisory Committee Workshop." July 2003. Accessed 31 August 2006. http://research.calit2.net/cibio/archived/CIBIO_Overview_Report.pdf
68. Atkins D.E., Droegemeier K.K., Feldman S.I., Garcia-Molina H., Klein M.L., Messerschmitt D.G., Messina P., Ostriker J.P., Wright M.H. "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure." January 2003. Accessed 1 September 2006. http://www.nsf.gov/od/oci/reports/atkins.pdf
69. Petascale Computing in the Biological Sciences. NSF-funded workshop held August 29-30, 2006. Accessed 11 October 2006. http://www.sdsc.edu/PMaC/BioScience_Workshop/biosciences.html
70. Dong S., Insley J., Karonis N.T., Papka M., Binns J. and Karniadakis G.E. 'Simulating and visualizing the human arterial system on the TeraGrid." Future Generation Computer Systems, The International Journal of Grid Computing: Theory, Methods and Applications, to appear, 2006.