Mathematics for Analysis of Petascale Data (MAPD)

Workshop Sponsored by the
Office of Advanced Scientific Computing Research
U.S. Department of Energy

June 3–5, 2008
Hilton Rockville Hotel
Rockville, Maryland

Both now and increasingly for the foreseeable future, scientists must address the challenges posed by petascale data sets. These sets may be produced by high-resolution simulations on massively parallel computers in complex applications, such as climate modeling and fusion calculations. They may also result from experiments and observational studies, such as those in cosmology and high-energy physics. Extracting scientific knowledge from these massive data sets has become both increasingly difficult and increasingly necessary as computer systems have grown larger and experimental devices more sophisticated. Mathematical techniques from several fields, including but not restricted to statistics, machine learning, image analysis, and pattern recognition, have long been used to analyze scientific data. However, many existing methods fail to provide adequate robustness, scalability, and combinatorial tractability when applied to petascale data sets.

The goal of this workshop is to engage mathematical scientists and applications researchers to define a research agenda for developing the next-generation mathematical techniques needed to meet the challenges posed by petascale data sets. Specific objectives are to:

- understand the needs of various scientific domains,
- delineate appropriate mathematical approaches and techniques,
- determine the current state-of-the-art in these approaches and techniques, and
- identify the gaps that must be addressed to enable the effective analysis of large, complex data sets in the next five to ten years.

A principal outcome of the workshop will be a report submitted to the Applied Mathematics Program in DOE's Office of Advanced Scientific Computing Research that will include recommendations for developing a research agenda.

Dr. Homer Walker
Program Manager for Applied Mathematics
Office of Advanced Scientific Computing Research
Office of Science
U.S. Department of Energy
www.sc.doe.gov/ascr/Research/AppliedMath.html
walker@ascr.doe.gov
(301) 903-1465

Tuesday, June 3:

7:00–8:00 Registration and Continental Breakfast

8:00–8:20 Welcome. "State of DOE Mathematics Program, Vision for Mathematics for Petascale Data", Homer Walker, DOE

8:20–8:30 Agenda and process description. Philip Kegelmeyer, Sandia National Laboratories.

8:30–9:30 Application focus: Astrophysics.

- Application domain challenges and issues. Speaker: Kirk Borne, George Mason University. Title: ""Data Science Challenges from Distributed Petascale Astronomical Sky Surveys".

  I will address the computational and data science challenges that will be imposed by the new petascale astronomical sky surveys that are coming online in the next decade. These challenges include the development of efficient mathematical models and parallel algorithms for several tasks: data mining and petascale statistical analyses across hundreds of dimensions, multi-dimension N-point correlation analyses, dimensionality reduction, condensed representations, and outlier/anomaly/interestingness detection and characterization. This includes the development of new scalable algorithms for statistical tests that typically scale as a power of the number of objects — our problem is that the number of objects may soon be 10^10 or larger. In addition, since the knowledge discovery potential of these large databases increases factorially as more datasets are combined, we seek robust approximate statistical learning and sampling algorithms for massive distributed data mining, in which a limited number of dimensions and training samples are communicated to the algorithm. For example, efficient tree structures, hierarchical aggregations, vector space factorizations, multipole expansions, or fractal representations of the information content of petascale databases will enable and enhance mining and visualization of the data for new scientific knowledge discovery. The above challenges are imposed both by the anticipated observational data collections and by new petascale computational model outputs. The enormous potential for grand-scale numerical simulations enabled by new petascale computing will generate enormous quantities of theoretical data. In many cases, multiple generations of models will be simulated to test against the observational data. This data-model matching leads to model validation and new scientific understanding, but it also leads to enormous combinatorial challenges, as the number of input parameters and model output parameters continue to grow as our models are refined. These petascale simulations thus impose enormous computational challenges in the areas of optimization, maximum likelihood estimation, and computational steering. Finally, as the numerical simulations grow in size and complexity, we seek faster methods of numerical integration and differentiation over gridded data structures.

- Mathematics commentary. Speaker: Alex Gray, Georgia Tech. Title: "Computational Mathematics for Large-Scale Data Analysis"

  Driven by the challenges in astronomical data analysis laid out in the pre-

vious talk, this talk will consider a selected survey of the computational mathematics, both current and future, which can be brought to bear. First, what are the data analysis (machine learning and statistics) methods we want to perform on large datasets in general, and how much do they cost? Then, what is a short-list of the basic mathematical operations underlying these methods, or bottletneck subroutines for which we can focus on developing fast algorithms? I'll discuss aggregations, N-body-like problems, graphical model inference, linear algebra, and optimization; for each I'll discuss the main mathematical challenge and highlight at least one current interesting solution. I'll close with some general solution avenues for the future, including the deep interplay between statistics and computation, visual analytics, and algorithm/software synthesis.

9:30–10 Break

10–11 Application focus: Networks

- Application domain challenges and issues. Speaker: George Karypis, University of Minnesota.
- Mathematics commentary. Speaker: *also* George Karypis, University of Minnesota.
- Title of the combined talk: "Drug and Probe Discovery and its Mathematical Challenges"

11–12 Application focus: Nano/Chemistry

- Application domain challenges and issues. Speaker: Thomas Schulthess, Oak Ridge National Laboratory. Title: "Understanding emergent properties in nanoscale systems: anticipating the role of petascale computing and data analysis."

  This talk will begin with a brief introduction to emergent properties, and a review of the traditional roles of experiment and theory. This will lead into a discussion of the *new* role theory and simulations can play given petascale computing. This role will be illustrated by the Hubbard model studies of the pairing mechanism, and further examined in the light of new experiments (neutrons, X-ray, EM & EELS), and via a consideration of the interface between simulations and experiment.

- Mathematics commentary. Speaker: Devinder Sivia, ISIS/RAL. Title: "Data analysis in condensed matter science".

12–1 Lunch

1–2:30 First Application Breakout Sessions. Objective: go into detail, listing the mathematics challenges and needs of each domain. Capture in a short list to be distributed to the mathematics breakouts.

1. Astrophysics. Moderator: Peter Nugent, Lawrence Berkeley National Laboratory. Scribe: Kevin Bowyer, Notre Dame.
2. Networks. Moderator: Jim McCalley, Iowa State University. Scribe: George Karypis, University of Minnesota.
3. Fusion Physics. Moderator: Linda Sugiyami, MIT. Scribe: Juan Meza, Lawrence Berkeley National Laboratory.
4. Combustion. Moderator: Mitch Smooke, Yale. Scribe: Habib Najm, Sandia National Laboratories.

5. Cybersecurity. Moderator: Deb Frincke, Pacific Northwest National Laboratory . Scribe: Mine Altunay, Fermi National Accelerator Laboratory.

2:30–3 Break

3–4:30 First Mathematics Breakout Sessions. Objective: Extract, list, and prioritize the math challenges gleaned from the problem statement presentations, 6 parallel sessions.

1. Statistics, Part 1. Moderator: Jon Kettenring, Drew University. Scribe: George Ostrouchov, Oak Ridge National Laboratory.

   In this session, we will consider three broad categories of statistical issues as they relate to petascale data: analysis of streaming data, analysis of non-streaming very large data sets, and the design of computer experiments. For streaming data, questions include: What estimators work sequentially and require no storage of the prior data stream? How do we detect change points within a data stream? For non-streaming data sets, potential focuses include problem-aware error analysis and sampling strategies for estimation. The design of computer experiments considers how to construct sets of code runs that provide data appropriate to the analysis goals. If for example, I have large numbers of input and output parameters, what set of runs do I make to apportion the variance of the outputs to variations in the inputs?

2. Data Reduction, Part 1: Dimensionality Reduction. Moderator: Lee Jameson, National Science Foundation. Scribe: Cathy Jiao, Oak Ridge National Laboratory.

   High-dimensional spaces are ubiquitous in massive data sets either directly or due to the analysis being performed for structural understanding of variability in the underlying phenomena. Examples of high dimensionality include organism variability in gene expression across thousands of genes, time variability across thousands of spatial locations in climate (or fusion, or combustion), and combinatorial complexity of dependence structure among simulated or measured quantities. Methods that exploit the blessings of high-dimensionality, such as concentration of measure and approach to continuum are needed to extract low-dimensional structures that are key to understanding application phenomena. Also likely to be useful are hierarchical data representations, which permit one to move up or down in the granularity of the data, as appropriate to the problem at hand The complexity of these methods often conflicts with petascale data and fast, approximate, adaptive, or entirely new methods are needed. This session will define the principal research directions in dimension reduction that are needed across applications.

3. Uncertainty Quantification, Part 1. Moderator: Habib Najm, Sandia National Labs. Scribe: Roger Ghanem, University of Southern California.

   Assessing the scope and accuracy of computational models is accomplished by incorporating measurements and expert judgment with the models. Mathematical issues at the petascale include constructing efficient uncertainty representations, developing appropriate algorithms for carrying out the transfer of uncertainty from inputs to the relevant model outputs, and (perhaps unique to large/petascale data), developing the mathematical apparatus to

support the navigation of the uncertainty representation associated with the models and petascale data.

In addition, uncertainty quantification encompasses the general goal of utilizing physical data (from experiments and observations) and complex simulations to make key scientific inferences and predictions. While traditional "verification and validation" focuses on the simulation model (does it solve the mathematical equations correctly? Does it adequately simulate the physical system?), this session specifically asks: How can we make use of both physical data and simulations to better understand physical processes and to gain predictive understanding?

4. Optimization Part 1. Moderator: Sven Leyffer, Argonne National Laboratory. Scribe: Duncan Temple Lang, University of California at Davis.

   Optimization techniques are an important component in data analysis. Partitioning, clustering, dimension reduction, classification algorithms, and statistical learning in general form an integral part of data analysis, all of which rely on optimization techniques. Recently there has been a tremendous growth in areas such as convex programming, combinatorial optimization, interior point methods, and mixed integer nonlinear programming methods to name just a few. Advances in these fields have allowed a broader application of optimization to numerous areas in data analysis. As the volume of the data explode in size there is a pressing need for novel data structures and efficient new scalable algorithms, especially on high performance parallel computing platforms.

5. Graph/Network Analysis Part 1. Moderator: Paul Whitney, Pacific Northwest National Laboratory. Scribe: Mac Hyman, Los Alamos National Laboratory.

   This session is organized around the need to develop a new mathematical foundation for network science that is grounded in empirical data and will enable design of efficient, stable and secure networks. We will explore the development of tools, abstractions, and methodologies that allow reasoning about large-scale networks, as well as techniques for modeling networks characterized by noisy and incomplete data. A central challenge at the petascale is capturing, understanding and ultimately exploiting high-variability phenomena observed in networks, including the heavy-tail characteristics of Internet traffic as a special case. We will examine the different root causes for observed high-variability phenomena with the aim of developing new approaches to explanatory modeling that capture features critical to control of highly-engineered systems such as the Internet. In particular we will explore the analysis of massive multi-scale data collected from large-scale operational networks and methods for distributed control that take account of network stochasticity.

6. Machine Learning, Part 1 Moderator: Larry Hall, University of South Florida. Scribe: Kirk Borne, George Mason University.

   Unsupervised machine learning is very useful in finding relevant patterns, in aiding understanding, in data that is not yet well understood. Supervised machine learning is similarly useful in focusing on the pertinent subsets of a

larger data set, and in other automated post-processing. However, as data increases in size, it falls increasingly subject to a host of ills that complicate the use of machine learning. In addition to sheer scale (which complicates the methods with worse than linear costs, such as SVMs, NNs, and clustering), large data is typically increasingly "skew", with the subsets of interest becoming a vanishingly small fraction of the whole. Large data is also likely large because it is higher dimensional, or temporal, or highly correlated (as in sensitivity studies), all properties which pose unique challenges. Finally, large data is very likely uncurated data, collected with less oversight and control. So the data itself is often noisier. Worse, the associated truth data, if there is any, is often inadequate in size (motivating active and semi-supervised training methods) or riddled with error itself. This session will examine current and future machine learning methods with an eye to addressing all of these difficulties.

4:30–5 Regroup and process check. Anything to tweak for tomorrow? Any emergent topics for one of the uncommitted breakouts? Philip Kegelmeyer, Sandia National Laboratories.

Wednesday, June 4:

7:00-7:45 Continental Breakfast

7:45–8:00 Agenda and process review. Philip Kegelmeyer, Sandia National Laboratories.

8–9 Application focus: Biology.

- Application domain challenges and issues. Speaker: Dan Rokhsar, University of California at Berkeley. Title: "Genomes, Genetics, and Diversity"

  New sequencing technologies are emerging driven by human biomedical applications. This talk will introduce the characteristics of these new datasets, and outline the computational hurdles in both converting the raw data into useful genetic information, and in answering questions related to population structure, dynamics, and evolution.

- Mathematics commentary. Speaker: Vipin Kumar, University of Minnesota. Title: "Application of Association Patterns Mining Techniques to Genomic Data"

  A wide variety and large amounts of genomic data is available now in the form of data on genetic variations, e.g., Single Nucleotide Polymorphisms (SNPs), and data on protein interactions, e.g., protein networks and complexes. Valuable biological knowledge can be derived from this data. For example, interaction data can be used to obtain information about protein function and the mechanisms involved in the accomplishment of these functions. As another example, data on genetic variations such as SNPs has created the possibility of discovering important connections between genetic factors and disease phenotypes, which, in turn, allows for the possibility of personalized medicine. But these data sets pose a number of challenges such as high dimensionality, noise, missing data, etc. This talk will describe how some of these challenges can be addressed by association analysis techniques developed in the field of data mining by providing examples using SNP and protein interaction data.

9–10 Application focus: Earth System Modeling

- Application domain challenges and issues. Speaker: Bill Collins, University of California Berkeley and Lawrence Berkeley National Laboratory. Title: "Extreme Climate Change: Scaling Laws, and Scale Invariance".
- Mathematics commentary. Speaker: Micheal Wehner, Lawrence Berkeley National Laboratory. Title: "Challenges in the analysis of petascale climate model output datasets".

  Climate model output is unique among simulation datasets in that it is widely and freely shared amongst the international climate science community. As part of the Fourth Assessment Report of the International Panel on Climate Change (IPCC AR4), over twenty global climate modeling groups ran ensembles of numerical experiments whose output data was collected and disseminated by the Program for Climate Model Diagnosis and Intercomparison at Lawrence Livermore National Laboratory. It is estimated that several hundred refereed scientific papers written over the last five years utilize portions of this database.

  Global climate models continue to increase in resolution, both horizontally and vertically. The resulting steady increase in output dataset size poses significant challenges to both the distribution of the data as well as to the analysis itself. A near term challenge presents itself for the planned Fifth Assessment Report (AR5) of the IPCC resulting from a quadrupling of resolution from the AR4 to the AR5 models. Datasets are expected to be close to two orders of magnitude larger. A much larger challenge is posed by the possibility of exascale computing which could enable global cloud system resolving models. These models, running at the kilometer scale, would produce datasets at least four orders of magnitude larger than the AR4 models.

  The talk will survey a few selected current methods of climate model analysis and present hard numbers on the sizes of future datasets.

10–10:30 Break

10:30–12 Second Application Breakout Sessions. Objective: go into detail, listing the mathematics challenges and needs of each domain. Capture in a short list to be distributed to the mathematics breakouts.

1. Biology. Moderator: Chris Oehmen, Pacific Northwest National Laboratory. Scribe: Ghaleb Abdulla, Lawrence Livermore National Laboratory.
2. Nano/Chemistry. Moderator: John Larese, University of Tennessee. Scribe: Jim Davenport, Brookhaven National Lab.
3. Accelerator Physics. Moderator: Dave Morrison, Brookhaven National Laboratory. Scribe: Alex Gray, Georgia Tech.
4. Visualization. Moderator: Valerio Pascucci, Lawrence Livermore National Laboratory. Scribe: Terence Critchlow, Pacific Northwest National Laboratory.
5. Earth System Modeling. Moderator: Deb Agarwal, Lawrence Berkeley National Laboratory. Scribe: Duncan Temple Lang, University of California at Davis.

12–1 Lunch

1–2:30 Second Mathematics Breakout Sessions. Objective: Extract, list, and prioritize the math challenges gleaned from the problem statement presentations, 4 parallel ses-

sions.

1. Streaming data analysis. Moderator: Terence Critchlow, Pacific Northwest National Laboratory. Scribe: Ghaleb Abdulla, Lawrence Livermore National Laboratory.

   Data streams are an abstraction for a large, possibly infinite, sequentially accessed collection of information. Three of the defining characteristics of data streams are that the data is far too large to fit in memory, even on a supercomputer, each data element is read only once, in the order in which it was created, and the data needs to be processed in real time as it is accessed. This abstraction holds for a number of interesting scientific applications such as NMR machines, HEP experiments, and telescopic imagery. It can also be considered to hold for petascale simulation data, where practical considerations limit the number of times the data set can be read. For purposes of this workshop, the challenges lie in developing new algorithms and approaches that will allow complex data analysis, such as categorization, clustering, pattern recognition and trend detection, to be performed on streaming data sets.

2. Statistics, Part 2. Moderator: Alyson Wilson, Los Alamos National Laboratory. Scribe: Steve Sain, National Center for Atmospheric Research.

   In this session, we will consider three broad categories of statistical issues as they relate to petascale data: analysis of streaming data, analysis of non-streaming very large data sets, and the design of computer experiments. For streaming data, questions include: What estimators work sequentially and require no storage of the prior data stream? How do we detect change points within a data stream? For non-streaming data sets, potential focuses include problem-aware error analysis and sampling strategies for estimation. The design of computer experiments considers how to construct sets of code runs that provide data appropriate to the analysis goals. If for example, I have large numbers of input and output parameters, what set of runs do I make to apportion the variance of the outputs to variations in the inputs?

3. Data Reduction, Part 2: Feature Extraction and Tracking. Moderator: George Ostrouchov, Oak Ridge National Laboratory. Scribe: Roger Ghanem, University of Southern California.

   Feature extraction is an important first step in many scientific analyses. The term "feature" can mean either a low level characteristic of the objects in the data, or some object itself. Mathematical challenges in feature extraction include enhancing the quality of the data, for example, by noise removal, robust techniques for identifying and extracting objects of interest in the data, and techniques for extracting low-level features from the data in a manner that is invariant to scale, translation, and rotation. A specific challenge in the context of this workshop is the extraction of ill-defined features spread across distributed petabyte-sized data files. Another related challenge is the selection of key low level features for use in building reduced dimension models.

2:30–3 Break

3–4:30 Last Mathematics Breakout Sessions. Objective: Extract, list, and prioritize the math

challenges gleaned from the problem statement presentations, 6 parallel sessions.

1. Uncertainty Quantification, Part 2. Moderator: David Scott, Rice University. Scribe: Steve Sain, National Center for Atmospheric Research.

   Assessing the scope and accuracy of computational models is accomplished by incorporating measurements and expert judgment with the models. Mathematical issues at the peta-scale include constructing efficient uncertainty representations, developing appropriate algorithms for carrying out the transfer of uncertainty from inputs to the relevant model outputs, and (perhaps unique to large/petascale data), developing the mathematical apparatus to support the navigation of the uncertainty representation associated with the models and petascale data.

   In addition, uncertainty quantification encompasses the general goal of utilizing physical data (from experiments and observations) and complex simulations to make key scientific inferences and predictions. While traditional "verification and validation" focuses on the simulation model (does it solve the mathematical equations correctly? Does it adequately simulate the physical system?), this session specifically asks: How can we make use of both physical data and simulations to better understand physical processes and to gain predictive understanding?

2. Optimization Part 2. Moderator: Juan Meza, Lawrence Berkeley Laboratory. Scribe: Juan Meza, Lawrence Berkeley Laboratory.

   Optimization techniques are an important component in data analysis. Partitioning, clustering, dimension reduction, classification algorithms, and statistical learning in general form an integral part of data analysis, all of which rely on optimization techniques. Recently there has been a tremendous growth in areas such as convex programming, combinatorial optimization, interior point methods, and mixed integer nonlinear programming methods to name just a few. Advances in these fields have allowed a broader application of optimization to numerous areas in data analysis. As the volume of the data explode in size there is a pressing need for novel data structures and efficient new scalable algorithms, especially on high performance parallel computing platforms.

3. Graph/network Analysis, Part 2. Moderator: Robert Calderbank, Princeton University. Scribe: Kevin Bowyer, Notre Dame.

   This session is organized around the need to develop a new mathematical foundation for network science that is grounded in empirical data and will enable design of efficient, stable and secure networks. We will explore the development of tools, abstractions, and methodologies that allow reasoning about large-scale networks, as well as techniques for modeling networks characterized by noisy and incomplete data. A central challenge at the petascale is capturing, understanding and ultimately exploiting high-variability phenomena observed in networks, including the heavy-tail characteristics of Internet traffic as a special case. We will examine the different root causes for observed high-variability phenomena with the aim of developing new approaches to explanatory modeling that capture features critical to control of

highly-engineered systems such as the Internet. In particular we will explore the analysis of massive multi-scale data collected from large-scale operational networks and methods for distributed control that take account of network stochasticity.

4. Machine Learning, Part 2. Moderator: Chandrika Kamath, Lawrence Livermore National Laboratory. Scribe: Terence Critchlow, Pacific Northwest National Laboratory.

   Unsupervised machine learning is very useful in finding relevant patterns, in aiding understanding, in data that is not yet well understood. Supervised machine learning is similarly useful in focusing on the pertinent subsets of a larger data set, and in other automated post-processing. However, as data increases in size, it falls increasingly subject to a host of ills that complicate the use of machine learning. In addition to sheer scale (which complicates the methods with worse than linear costs, such as SVMs, NNs, and clustering), large data is typically increasingly "skew", with the subsets of interest becoming a vanishingly small fraction of the whole. Large data is also likely large because it is higher dimensional, or temporal, or highly correlated (as in sensitivity studies), all properties which pose unique challenges. Finally, large data is very likely uncurated data, collected with less oversight and control. So the data itself is often noisier. Worse, the associated truth data, if there is any, is often inadequate in size (motivating active and semi-supervised training methods) or riddled with error itself. This session will examine current and future machine learning methods with an eye to addressing all of these difficulties.

4:45–5 Regroup and process check. Anything to tweak for tomorrow? Any issues to be addressed tonight to ease tomorrow's reports? Philip Kegelmeyer, Sandia National Laboratories.

6:00 Conference Dinner, in the Regency Room of the Hilton Rockville.

7:00 Keynote presentation, Prof. Emery Brown of MIT and Harvard Medical School, "Dynamic Signal Processing Analysis of Brain Function."

Neural systems encode representations of biological signals in the firing patterns of their spike trains. Spike trains are point process time-series and their codes are both dynamic and stochastic. Although the signal is often continuous, its representation in the nervous systems is as a high-dimensional point process time-series. Because neural spike trains are point processes, standard signal processing techniques for continuous data are of limited utility in the analysis of neural systems. Accurate processing of neural signals requires the development of quantitative techniques to characterize correctly the point process nature of neural spiking activity. We discuss our research on the use of the state-space paradigm for point process observation to characterize neural systems and discuss three applications characterizing: the response threshold of neurons in primary auditory cortex; the dynamics of subthalamic nuclei neurons in patients suffering from Parkinson's disease; and how ensemble spiking activity of rat hippocampal neurons maintain a dynamic representation of the animal's position in its environment. Brain dynamics can also be measured indirectly at the scalp through the brain's electrical and magnetic fields. As a final example,

we present a high-dimensional EM algorithm for a state-space model and use it to compute simultaneous state (dipole sources) and parameter estimates for magnetoencephalography (MEG) inverse problems.

Thursday, June 5:

7:00-8:00 Continental Breakfast

8:00-10:00: Reports and discussion from the seven mathematics discipline areas (plus "Other", as appropriate.) Roughly 15 minutes each.

10:00–10:30 Break

10:30–12 General discussion, compilation of major themes and conclusions, drafting of report outline.

Noon: Adjourn.

1–4: Writing session for organizing committee.