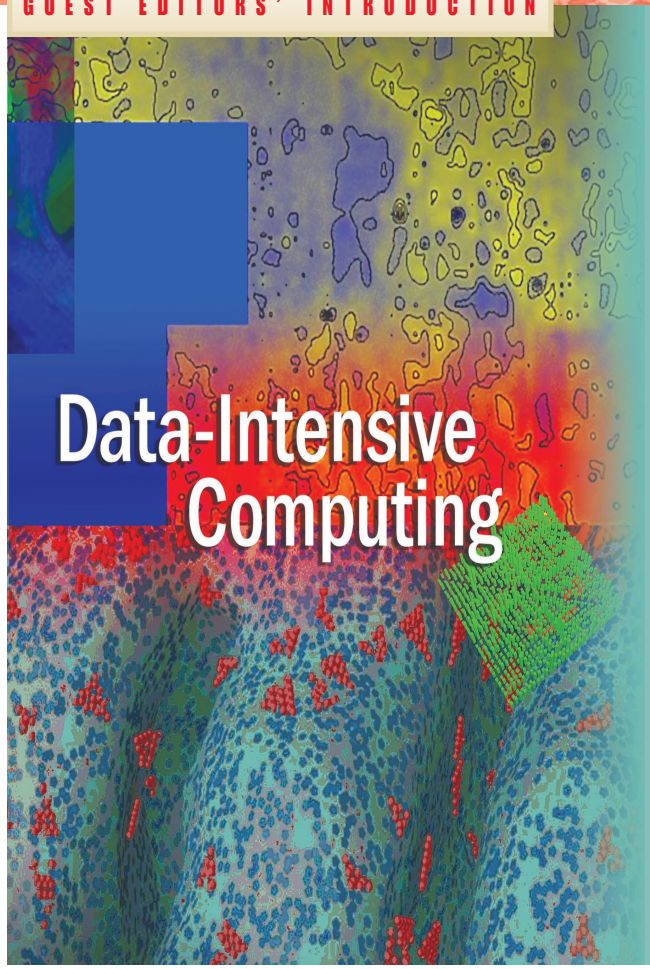

Data-Intensive Computing in the 21st Century

Ian Gorton, Pacific Northwest National Laboratory
Paul Greenfield, CSIRO
Alex Szalay, John Hopkins University
Roy Williams, Caltech

The deluge of data that future applications must process—in domains ranging from science to business informatics—creates a compelling argument for substantial increased R&D targeted at discovering scalable hardware and software solutions for data-intensive problems.

In 1998, William Johnston delivered a paper at the 7th IEEE Symposium on High-Performance Distributed Computing[1] that described the evolution of data-intensive computing over the previous decade. While state of the art at the time, the achievements described in that paper seem modest in comparison to the scale of the problems researchers now routinely tackle in present-day data-intensive computing applications.

More recently, others including Tony Hey and Anne Trefethen,[2] Gordon Bell and colleagues,[3] and Harvey Newman and colleagues[4] have described the magnitude of the data-intensive problems that the e-science community faces today and in the near future. Their descriptions of the data deluge that future applications must process, in domains ranging from science to business informatics, create a compelling argument for R&D to be targeted at discovering scalable hardware and software solutions for data-intensive problems. While petabyte datasets and gigabit data streams are today's frontiers for data-intensive applications, no doubt 10 years from now we'll fondly reminisce about problems of this scale and be worrying about the difficulties that the looming exascale applications are posing.

Fundamentally, data-intensive applications face two major challenges:

- managing and processing exponentially growing data volumes, often arriving in time-sensitive streams from arrays of sensors and instruments, or as the outputs from simulations; and
- significantly reducing data analysis cycles so that researchers can make timely decisions.

There is undoubtedly an overlap between data- and compute-intensive problems. Figure 1 shows a simple diagram that can be used to classify the application space between these problems.

Purely data-intensive applications process multi-terabyte to petabyte sized datasets. This data commonly comes in several different formats and is often distributed across multiple locations. Processing these datasets

Published by the IEEE Computer Society
0018-9162/08/$25.00 © 2008 IEEE

typically takes place in multiple-step analytical pipelines that include transformation and fusion stages. Processing requirements typically scale near-linearly with data size and are often amenable to straightforward parallelization. Key research issues involve data management, filtering and fusion techniques, and efficient querying and distribution.

Data/compute-intensive problems combine the need to process very large datasets with increased computational complexity. Processing requirements typically scale superlinearly with data size and require complex searches and fusion to produce key insights from the data. Application requirements may also place time bounds on producing useful results. Key research issues include new algorithms, signature generation, and specialized processing platforms such as hardware accelerators.

We view data-intensive computing research as encompassing the problems in the upper two quadrants in Figure 1. The following are some applications that exhibit these characteristics.

Astronomy. The Large Synoptic Survey Telescope (LSST; www.lsst.org) will generate several petabytes of new image and catalog data every year. The Square Kilometer Array (SKA; www.skatelescope.org) will generate about 200 Gbytes of raw data per second that will require petaflops (or possibly exaflops) of processing to produce detailed radio maps of the sky. Processing this volume of data and making it available in a useful form to the scientific community poses highly challenging problems.

Cybersecurity. Anticipating, detecting, and responding to cyberattacks requires intrusion-detection systems to process network packets at gigabit speeds. Ideally, such systems should provide actionable results in seconds to minutes rather than hours so that operators can defend against attacks as they occur.

Social computing. Sites such as the Internet Archive (www.archive.org/) and MySpace (www.myspace.com) store vast amounts of content that must be managed, searched, and delivered to users over the Internet in a matter of seconds. The infrastructure and algorithms required for websites of this scale are challenging, ongoing research problems.

DATA-INTENSIVE COMPUTING CHALLENGES

The breakthrough technologies needed to address many of the critical problems in data-intensive computing will come from collaborative efforts involving several disciplines, including computer science, engineering, and mathematics. The following list shows some of the advances that will be needed to solve the problems faced by data-intensive computing applications:

- new algorithms that can scale to search and process massive datasets;

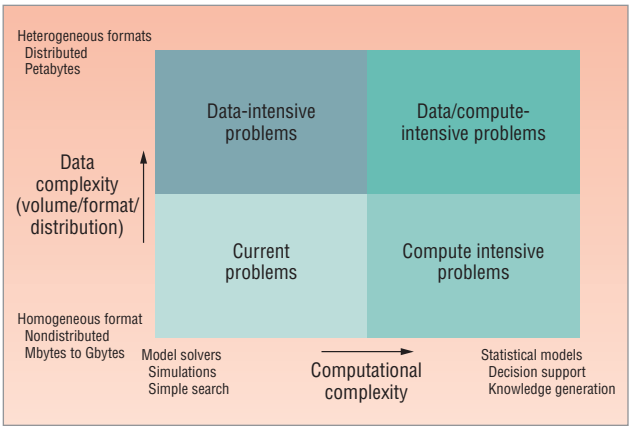

Figure 1. Data-intensive computing. Data-intensive computing research encompasses the problems in the upper two quadrants.

- new metadata management technologies that can scale to handle complex, heterogeneous, and distributed data sources;
- advances in high-performance computing platforms to provide uniform high-speed memory access to multi-terabyte data structures;
- specialized hybrid interconnect architectures to process and filter multigigabyte data streams coming from high-speed networks and scientific instruments and simulations;
- high-performance, high-reliability, petascale distributed file systems;
- new approaches to software mobility, so that algorithms can execute on nodes where the data resides when it is too expensive to move the raw data to another processing site;
- flexible and high-performance software integration technologies that facilitate the plug-and-play integration of software components running on diverse computing platforms to quickly form analytical pipelines; and
- data signature generation techniques for data reduction and rapid processing.

IN THIS ISSUE

This special issue on data-intensive computing presents five articles that address some of these challenges.

In "Quantitative Retrieval of Geophysical Parameters Using Satellite Data," Yong Xue and colleagues discuss the remote sensing information service grid node, a tool for processing satellite imagery to deal with climate change.

In "Accelerating Real-Time String Searching with Multicore Processors," Oreste Villa, Daniele Paolo Scarpazza, and Fabrizio Petrini present an optimization strategy for a popular algorithm that performs exact string matching against large dictionaries and offer solutions to alleviate memory congestion.

"Analysis and Semantic Querying in Large Biomedical Image Datasets" by Joel Saltz and colleagues describes a set of techniques for using semantic and spatial information to analyze, process, and query large image datasets.

"Hardware Technologies for High-Performance Data-Intensive Computing" by Maya Gokhale and colleagues offers an investigation into hardware platforms suitable for data-intensive systems.

In "ProDA: An End-to-End Wavelet-Based OLAP System for Massive Datasets," Cyrus Shahabi, Mehrdad Jahangiri, and Farnoush Banaei-Kashani describe a system that employs wavelets to support exact, approximate, and progressive OLAP queries on large multidimensional datasets, while keeping update costs relatively low.

We hope you will enjoy reading these articles and that this issue will become a catalyst in drawing together the multidisciplinary research teams needed to address our data-intensive future.

References

1. W. Johnston, "High-Speed, Wide Area, Data-Intensive Computing: A Ten-Year Retrospective," *Proc. 7th IEEE Symp. High-Performance Distributed Computing*, IEEE Press, 1998, pp. 280-291.
2. T. Hey and A. Trefethen, "The Data Deluge: An e-Science Perspective;" www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf.
3. G. Bell, J. Gray, and A. Szalay, "Petascale Computational Systems," *Computer*, Jan. 2006, pp. 110-112.
4. H.B. Newman, M.H. Ellisman, and J.A. Orcutt, "Data-Intensive E-Science Frontier Research," *Comm. ACM*, Nov. 2003, pp. 68-77.

Ian Gorton *is the chief architect for Pacific Northwest National Laboratory's Data-Intensive Computing Initiative. His research interests include software architectures and middleware technologies. He received a PhD in computer science from Sheffield Hallam University. Gorton is a member of the IEEE Computer Society. Contact him at ian.gorton@pnl.gov.*

Paul Greenfield *is a research scientist in Australia's Commonwealth Scientific and Industrial Research Organisation. His research interests are distributed applications, the analysis of genetic sequence data, and computer system performance. He received an MSc in computer science from the University of Sydney. Greenfield is a member of the IEEE and the ACM. Contact him at paul.greenfield@csiro.au.*

Alex Szalay *is the Alumni Centennial Professor in the Department of Physics and Astronomy at the Johns Hopkins University. His research interests include large spatial databases, pattern recognition and classification problems, theoretical astrophysics, and galaxy evolution. He received a PhD in //**what discipline**??// from //**degree-granting institution**//. He is a member of //**relevant professional organizations**//. Contact him at szalay@jhu.edu.*

Roy Williams *is a senior scientist at the Center for Advanced Computing at Caltech. His research interests are astronomical transients, virtual observatory, and big data. He received PhD in physics from Caltech. He is a member of the IEEE and the AAS. Contact him at cacr.caltech.edu.*