

CIOReview

The Navigator for Enterprise Solutions

MARCH 2-2015

CIOREVIEW.COM

Company of the Month



Gary Gauba, Chairman & CEO, Cognilytics

Entrepreneur of the Month



Richard E. Malinowski
Founder & President,
REMTCS

General Dynamics Mission Systems Delivering Real-time Intelligence through HPC

Nadia Short,
VP & GM,
Cyber Systems,
General Dynamics
Mission Systems



#202, Fremont, CA-94538
44790, S Grimmer Blvd
CIO REVIEW

HPC and Big Data: A View from the Corner of Science and Industry

By David A. Bader, Chair of the School of Computational Science and Engineering & Executive Director of High Performance Computing, Georgia Institute of Technology



David A. Bader

With the last decade's rapid advances in computational power and the resulting explosion of available data, two major areas of computational application—business and scientific research—are now converging. And this digitally fluid world of data science and application is providing remarkable value for industry.

High performance computing, once an area reserved for technical or scientific application, has for some time now affected the realm of business and enterprise computing. What does this mean for business? It means improved knowledge of customers, more efficient operations, and accelerated response to changes or to problems, among other advantages.

Many businesses have seen the light and are investing in personnel with expertise in analytics and data-specific algorithms like Hadoop. Smaller companies also now have the option of buying big data solutions for business intelligence and analytics from third-party companies.

With a dual career in both industry and scientific research, I enjoy a broad and fascinating view of this ever-changing HPC terrain. Over in the research sciences, we are transforming methods for studying the very complex. This shift affects how scientists approach, for example, studies of genome-wide associations, the interplay of genes, and climate modeling.

Science has traditionally operated by deciphering time-consuming cause-and-effect relationships. Meaningful

associations are now emerging from heaps of various data types that can be merged and analyzed. Trends in large data sets can be searched, scrutinized and visualized—all techniques that dramatically speed up research time and allow us to tackle a huge range of problems at scales from tiny to enormous in terms of physical size, duration and complexity.

HPC capabilities also permit powerful simulations that speed product development. From detergents to jet engines, products can now be developed without intermediate prototypes or several iterations of laboratory tests. All of these advances for science are also proportional advances for industry.

Our world today is one of growing torrential streams of information that can provide valuable information to inform decisions related to business intelligence, market analysis and social trends using data from social media and user behaviors. At Georgia Tech, we are forging new ways to combine big data with HPC research to craft solutions to business and social problems.

Take cybersecurity. Our work in cyberanalytics is an excellent example of applying these techniques at the intersection of HPC, big data, research, and real-world problems. Indeed, the word “cybersecurity” often evokes an image of outsiders trying to hack into our computer systems, particularly those of large corporations. But many incidents that go unreported in the media are insider events, which some reports say comprise up to a third of threats and can be more costly or damaging than those orchestrated by outsiders. We often wonder how these trusted individuals went about their malicious work unnoticed, but the key to detecting and preventing the breach was likely there, buried in data that could not—until recently—be interpreted and quickly acted upon.

Every time we use a key card to open a door, send an email or invoke any number of computer actions, we leave a digital trace. Security officers analyze this information to determine our patterns and identify potential threats. But the massive scale datasets are often unstructured and challenging to inspect.

At Georgia Tech, we conduct analytic research on graphs. Graphs are networks with up to trillions of connections, and they help us discover patterns and relationships hidden deep within massive amounts of data. These graphs are comprised of interconnected vertices and edges that change over time. In the realm of cybersecurity, the vertices may represent computers, and the edges represent their interactions. By designing fast theoretic algorithms on large-scale graphs, we can produce insights in near real time.

Emerging graph technology at Georgia Tech has the potential to quickly sift massive amounts of data and correctly prioritize the

most likely short list of results for a fast response. The aim is to develop the best and most efficient way to prevent future malicious activities where we work and live.

The media notably cover large security breaches, but for each such report, going unnoticed by the public are numerous other data breaches in small companies, which are often less resilient and have a harder time recovering. Many of these companies fold permanently after an attack.

The new way of thinking calls for businesses both large and small to deeply embed into their cultures the cybersecurity priorities, investments in expertise, and an increasing awareness of emerging technologies like those in use at Georgia Tech.

Also, our work in graph analytics can lead to a suite of far-ranging and useful applications, such as:

- Unraveling the driving forces in graphs that change as they grow, such as social networks. We are expanding what we know about scientific computing graphs and moving into large social and information networks, whose informatics are much more challenging.
- Analyzing massive streaming complex networks in real-time, with applications in public health, transportation, evacuation, security, drug design, water supplies and full-scale socioeconomic systems.
- Visualizing massive graphs. Growing amounts of data require new methods for meaningful interpretation.

In addition to advancing data analysis, we are also actively researching computer architectural requirements to maximize the performance of graph analyses across a variety of problem types. How can we integrate across different algorithms, programming models and architectures to address new challenges? The research includes exploring how best to combine cloud computing with in-memory parallel computing. This work lays a foundation to take on some of the most difficult problems in the world today, from computational biology and genomics, massive-scale data analytics with a focus on parallel algorithms, to combinatorial optimization, and vast social networks.

Georgia Tech is playing its role to build a new computational future, one that includes creative new architectures and software approaches, improved energy efficiency, altered computing paradigms to ingest more information, and reshaped scientific and business workflows. In exchange, we will handle large scale, complex problems more quickly and accurately. We will have more targeted approaches to managing everything from malicious activities to public health and environmental concerns.

The view from this corner is indeed compelling. Our basic assumptions change almost daily. What will tomorrow bring? **CR**