Research and Reviews: Journal of Global Research in Computer Science

Extended Abstract
Vol. 9, Iss. 4
2018

Data Mining 2016: Alignment and convergence of knowledge discovery and HPC- Thomas Sterling , Indiana University

Thomas Sterling
Indiana University, USA

Data analytics in its many forms has rapidly expanded to interact scientific, industrial, and societal application domains. But as more problem spaces yield to this expanding genre of computing, the demand for capabilities is expanding. Simultaneously, high performance computing (HPC) systems and methods is experiencing significant change in form and performance with the asymptotic convergence with nano-scale semiconductor feature size and therefore the end of Moore's law even with exascale performance anticipated in the early years of the next decade. Historically these two processing domains are largely independent but now a growing consensus is driving them together, aligning their respective modalities and catalyzing a synergistic convergence. A major premise folks Presidential Executive Order resulting in the National Strategic Computing Initiative stipulates that the merger of massive data and numeric intensive computing be a constituent of national exascale charter. This presentation will describe the many shift in system architecture and operational methodologies which will be required to simultaneously answer the challenges of the top of Moore's law and the graph processing approaches, potentially dynamic which will augment the more conventional matrix-vector oriented computation. It will discuss the likely importance of dynamic adaptive resource management and task scheduling essential to dramatic improvements in scalability and efficiency for exascale computing and the way these changes will be applied to knowledge discovery.

To answer today's increasingly complex and data-intensive science questions in experimental, observational and computational sciences, we are developing methods in three interrelated R&D areas: (i) We are creating new scalable data analysis methods capable of running on large-scale computational platforms to respond to increasingly complex lines of scientific inquiry. (ii) Our new computational design patterns for key analysis methods will help scientific researchers take full advantage of rapidly evolving trends in computational technology, such as increasing cores per processor, deeper memory and storage hierarchies, and more complex computational platforms. The key objectives are high performance and portability across DOE computational platforms. (iii) By combining analysis and processing methods into data pipelines for use in large-scale HPC platforms—either standalone or integral to a larger scientific workflow—we are maximizing the opportunities for analyzing scientific data using a diverse collection of software tools and computational resources.

Despite tremendous progress made in biological imaging that has yielded tomograms with ever-higher resolutions, the segmentation of cell tomograms into organelles and proteins remains a challenging task. The difficulty is most extreme in the case of cryo-electron tomography (cryo-ET), where the samples exhibit inherently low contrast due to the limited electron dose that can be applied during imaging before radiation damage occurs. The tomograms have a low signal-to-noise ratio (SNR), as well as missing-wedge artifacts caused by the limited sample tilt range that is accessible during imaging. While SNR can be improved by applying contrast enhancement and edge detection methods, these algorithms can also generate false connectivity and additional artifacts that degrade the results produced by automatic segmentation programs. If the challenges can be overcome, automatic segmentation approaches are of great interest. However, the achievement of this vision is precluded today by the complexity of the specimen and the SNR limitations described above. State of the art machine learning results are not generally suitable for deep mining, in fact, the situation in cryo-ET is quite the opposite: the highest quality segmentations are produced by hand, representing effort levels ranging from days to months. Segmentation tools could be vastly improved if they were constructed to take into account prior knowledge, minimizing the sensitivity to noise and false connection. To the best of our knowledge, there are no methods using specific contextual information about biological structures as restraints for segmentation. Nor are there approaches that incorporate active learning with feedback from the user, which would provide guidance as to the correctness of the segmentation. We are developing new machine learning techniques to facilitate the segmentation, extraction, visualization, and annotation of biological substructures within 3D tomograms obtained from a variety of imaging modalities.

Biography

Thomas Sterling is a Professor of Intelligent Systems Engineering at the Indiana University School of Informatics and Computing. He serves as the Chief Scientist and Associate Director of the Center for Research in Extreme Scale Technologies (CREST). After receiving his PhD from MIT in 1984 as a Hertz Fellow, he has been engaged in research fields associated with Parallel Computing System Structures and Semantics. He is the co-author of 6 books and holds 6 patents. He was the Recipient of the 2013 Vanguard Award.

Email: tron@iu.edu