Contribution ID: **190** Type: **Invited Talk**

SAVIME - Simulation Analysis and Visualization in-Memory

Tuesday, 3 December 2019 11:20 (20 minutes)

The increasing computational power of HPC systems fosters the development of complex numerical simulations of phenomena in different domains, such as medicine [1], Oil & Gas [2] and many other fields [3,4,5]. In such applications, a huge amount of data in the form of multidimensional arrays is produced and need to be analyzed and visualized enabling researchers to gain insights about the phenomena being studied.

Scientists also generate huge multidimensional arrays through environmental observations, measurements of physical conditions and other types of sensors. For instance, satellite data for Earth's weather, oceans, atmosphere and land [6] are kept in the form of multidimensional arrays in scientific file formats. Data collected by sensors in physics experiments, such as the ones conducted in the photon studies by SLAC National Accelerator Laboratory [7], are also represented and processed in the form of multidimensional arrays.

Machine learning is another context in which multidimensional arrays are present. They are the basic input format for the heavily optimized linear algebra algorithms implemented in deep learning frameworks, such as: TensorFlow, Keras and Torch. Deep Learning algorithms were able to achieve superhuman performance for image recognition problems in the past few years [8], and they are among the most promising alternative for tackling difficult problems in Natural Language Processing, Image and Video Recognition, Medical Image Analysis, Recommendation Systems and many others. Thus, managing these large arrays in the context of deep learning is a very important task.

The traditional approach for managing data in multidimensional arrays in scientific experiments is to store them using file formats, such as netCDF and HDF5. The use of file formats, and not a database management systems (DBMS), in storing scientific data has been the traditional choice due to the fact that DBMSs are considered inadequate for scientific data management. Even specialized scientific data management systems, such as SciDB [10], are not very well accepted for a myriad of reasons listed in

- the impedance mismatch problem [12,13], that makes the process of ingesting data into a DBMS very slow.
- the inability to directly access data from visualization tools like Paraview Catalyst [14] and indexing facilities like FastQuery [13].
- the Inability to directly access data from custom code, which is necessary for domain specific optimized data analysis.

However, by completely dismissing DBMSs, some nice features also become unavailable. Including the access for out-the-box parallel declarative data processing with the usage of query languages and query

optimization, and management of dense and sparse matrices. In this talk, we will present SAVIME, a Database Management System for Simulation Analysis and Visualization in-Memory. SAVIME implements a multi-dimensional array data model and a functional query language. The system is extensible to support data analytics requirements of numerical simulation applications.

References

[1] Pablo J. Blanco, M. R. Pivello, S. A. Urquiza, and Raúl A. Feijóo. 2009. On the potentialities of 3D-1D coupled models in hemodynamics simulations.Journal of Biomechanics 42, 7 (2009), 919–930.
[2] Cong Tan et al. 2017. CFD Analysis of Gas Diffusion and Ventilation Protection in Municipal Pipe Tunnel. In Proceedings of the 2017 International Conference on Data Mining, Communications and Information Technology (DMCIT '17). ACM, New York, NY, USA, Article 28, 6 pages. Https: //doi.org/10.1145/3089871.3089904
[3] Igor Adamovich et al. 2015. Kinetic mechanism of molecular energy transfer and chemical reactions in low-temperature air-fuel plasmas. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 373 (08 2015). https://doi.org/10.1098/rsta.2014.0336
[4] Mollona, Edoardo, Computer simulation in social sciences, Journal of Management & Governance, May, 2008, N(2) V(12).
[5] Mitsuo Yokokawa, Ken'ichi Itakura, Atsuya Uno, Takashi Ishihara, and Yukio Kaneda. 2002. 16.4Tflopss Direct Numerical Simulation of Turbulence by a Fourier Spectral Method on the Earth Simulator. In Proceedings of the 2002 ACM/IEEE Conference on Supercomputing (SC '02). IEEE Computer Society Press, Los Alamitos, CA, USA, 1–17. http://dl.acm.org/citation.cfm?id=762761.762808 .
[6] R. Ullman and M. Denning. 2012. HDF5 for NPP sensor and environmental data records. In 2012 IEEE International Geoscience and Remote Sensing Symposium. 1100–110
[7] Jack Becla, Daniel Wang, and Kian-Tat lim. 2011. Using SciDB to Support Photon Science Data Analysis. (01 2011).
[8] Dan Ciresan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. 2012. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. In Advances in Neural Information Processing Systems 25 , F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2843–2851.
[9] Z Zhao. 2014. Automatic library tracking database at NERSC. (2014). https://www.nersc.gov/assets/altdatNERSC.pdf, J. ACM, Vol. 37, No. 4, Article 111. Publication date: August 2019.
[10] Paradigm4. 2019. SciDB. http://www.paradigm4.com/ [Online; accessed 01-sep-2019].
[11] Haoyuan Xing, Sofoklis Floratos, Spyros Blanas, Suren Byna, Prabhat, Kesheng Wu, and Paul Brown. 2017. ArrayBridge: Interweaving declarative array processing with high-performance computing. arXiv e-prints , Article arXiv:1702.08327 (Feb 2017).
[12] Spyros Blanas, Kesheng Wu, Surendra Byna, Bin Dong, and Arie Shoshani. 2014. Parallel Data Analysis Directly on Scientific File Formats. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD '14) . ACM, New York, NY, USA, 385–396.
[13] Luke Gosink, John Shalf, Kurt Stockinger, Kesheng Wu, and Wes Bethel. 2006. HDF5-FastQuery: Accelerating Complex Queries on HDF Datasets Using Fast Bitmap Indices. In Proceedings of the 18thInternational Conference on Scientific and Statistical Database Management (SSDBM '06). IEEE Computer Society, Washington, DC, USA, 149–158
[14] Utkarsh Ayachit, Andrew Bauer, Berk Geveci, Patrick O'Leary, Kenneth Moreland, Nathan Fabian,and Jeffrey Mauldin. 2015. ParaView

Catalyst: Enabling In Situ Data Analysis and Visualization. In Proceedings of the First Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization (ISAV2015) . ACM, New York, NY, USA, 25–29

Supported Student

Primary author: Dr PORTO, Fabio (LNCC)

Co-author: LUSTOSA, Hermano (National Laboratory of Scientific Computing (LNCC))

Presenter: Dr PORTO, Fabio (LNCC)

Session Classification: HPC Applications

Track Classification: Earth Systems Modelling