Virtualized Big Data: Reproducing Simulation Output on Demand

Salvatore Di Girolamo
Department of Computer Science
ETH Zurich
digirols@inf.ethz.ch

Torsten Hoefler (advisor)
Department of Computer Science
ETH Zurich
htor@inf.ethz.ch

ACM Reference format:
Salvatore Di Girolamo and Torsten Hoefler (advisor). 2017. Virtualized Big Data: Reproducing Simulation Output on Demand. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 2 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 EXTENDED ABSTRACT

Many scientific simulations produce vast amounts of data that is stored in parallel filesystems or large-scale databases to be analyzed later. For example, the climate community requires petabytes of storage to save simulation outputs. The European Centre for Medium-Range Weather Forecasts (ECMWF) alone had an archive of 100 PiB in 2015, which experienced an annual growth rate of 45% [6]. This means, that by 2020, the ECMWF archive will be near a Zettabyte. Similarly, groups of astrophysicists unite to experiment with large-scale simulations using trillions of particles to better understand the universe [7]. A single such simulation creates more than 20 PiB output and virtual observatories often store thousands of simulation results for which the accuracy (in terms of timesteps) is limited by the available storage [3, 8]. Astrophysics and climate are just two communities that are facing large-scale simulation data today but as we proceed into the age of simulation [10], such big data problems will soon affect the majority of sciences.

The data produced by such hero-runs is extremely valuable and is analyzed by thousands of researchers over the course of decades. Scientists are used to a data-backed analysis workflow where the data exists in files or databases. Specifically, this workflow addresses two requirements: (1) data can conveniently be analyzed with any traversal pattern (e.g., time-reverse or random access) and (2) the exact same data can later (often years) be re-analyzed to reproduce the results. Even *in-situ* analysis techniques [4] cannot always satisfy the trajectories that analysis tools require and it is often not possible for thousands of researchers to re-run the whole simulation. Furthermore, reproducibility is often mandated by regulatory bodies especially in climate science. Thus, the data-backed analysis has been established as the de-facto standard for data analysis.

Yet, storing the massive amounts of ever-growing simulation output data is a large part of the big data challenge in scientific computing. This is primarily due to the high storage costs and the fact that compute capabilities grow faster than storage capacities and bandwidths. Furthermore, traditional storage schemes are sub-optimal for long-term archiving [2] and we need to find alternatives that allow us to store data redundantly at reasonable cost.

Fortunately, simulation output data is different from data that is collected by sensors in the field because the scientific application that initially produced the data can be used to re-create it at any point. Since computation can be significantly cheaper than storage, it is worth exploring the computation-storage tradeoff for large-scale scientific data. This is especially attractive because datacenters have a fixed storage capacity and scientists are simply not able to store the output of some simulations. Yet, a system that only stores a subset of the simulation data and recreates the remaining data on demand could enable simulations and analyses that are impossible today. This would not only enable new scientific breakthroughs but the system's cost would shrink with the computation costs.

In this work, we introduce SDaVi, a Simulation Data Virtualizer, a system that virtualizes offline-data access by automatically re-running simulation for computing data that is not on disk.

The proposed workflow consists in having scientists setting up the initial simulation in order to produce the restart files. The amount of restart files to produce is a simulation parameter, hence the required storage can be tuned. Later, different analysis tools access the virtualization layer through standard data-access interfaces such as netCDF [9] or HDF-5 [5]. Wrappers to such interfaces are provided in order to enable data virtualization. The system can work in a fully-transparent way, accepting requests from off-the-shelf analysis tool, that are not aware of the virtualized environment. Otherwise, additional info on the access pattern can be provided by virtualization-aware tools. Also, analysis tools can be grouped in synchronized sets to minimize the re-simulation computing time.

If the requested data is not available in the SDaVi-managed caching hierarchy, a new simulation is launched in order to re-create the data. We remark that simulations can be restarted on different devices than the original simulation, e.g., smaller GPU systems, because the simulated time intervals are less demanding. SDaVi keeps track of the analysis tools' access patterns and apply prefetching strategies in order to minimize the re-simulation time and increase the aggregate simulation bandwidth (i.e., the analysis tool consumes the data faster than it is produced).

This approach requires that the simulations can be re-started and produce bit-reproducible outputs. We remark that most simulation tools already support checkpoint/restart to recover from system failures. Bit-reproducibility is required for good scientific practice and can be added relatively easily using standard techniques [1].

Overall, SDaVi offers a viable path towards exa-scale scientific simulations, by exploiting the growing computing power and relaxing the storage capacity requirements.

Conference'17, Washington, DC, USA
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

REFERENCES

[1] Arteaga, A., O.Fuhrer, and Hoefler, T. Designing Bit-Reproducible Portable High-Performance Applications. In *Proceedings of the 28th IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (Apr. 2014), IEEE Computer Society.

[2] Baker, M., Keeton, K., and Martin, S. Why traditional storage systems don't help us save stuff forever. In *Proceedings of the First Conference on Hot Topics in System Dependability* (Berkeley, CA, USA, 2005), HotDep'05, USENIX Association, pp. 7–7.

[3] Bernyk, M., Croton, D. J., Tonini, C., Hodkinson, L., Hassan, A. H., Garel, T., Duffy, A. R., Mutch, S. J., Poole, G. B., and Hegarty, S. The theoretical astrophysical observatory: Cloud-based mock galaxy catalogs. *The Astrophysical Journal Supplement Series 223*, 1 (2016), 9.

[4] Bethel, E. W., Childs, H., and Hansen, C. *High Performance Visualization: Enabling Extreme-Scale Scientific Insight.* CRC Press, 2012.

[5] Folk, M., Cheng, A., and Yates, K. Hdf5: A file format and i/o library for high performance computing applications. In *Proceedings of Supercomputing* (1999), vol. 99, pp. 5–33.

[6] Grawinkel, M., Nagel, L., Mäsker, M., Padua, F., Brinkmann, A., and Sorth, L. Analysis of the ecmwf storage landscape. In *Proceedings of the 13th USENIX Conference on File and Storage Technologies* (Berkeley, CA, USA, 2015), FAST'15, USENIX Association, pp. 15–27.

[7] Potter, D., Stadel, J., and Teyssier, R. Pkdgrav3: Beyond trillion particle cosmological simulations for the next era of galaxy surveys. *arXiv preprint arXiv:1609.08621* (2016), 1.

[8] Ragagnin, A., Dolag, K., Biffi, V., Bel, M. C., Hammer, N. J., Krukau, A., and Margarita Petkova, D. S. An online theoretical virtual observatory for hydrodynamical, cosmological simulations. *arXiv preprint arXiv:1612.06380* (2016).

[9] Rew, R. K., and Davis, G. P. The unidata netcdf: Software for scientific data access. In *Sixth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology* (1990), pp. 33–40.

[10] Winsberg, E. *Science in the age of computer simulation.* University of Chicago Press, 2010.