

Chapter 1

Executive Summary

The ASCAC Subcommittee on Synergistic Challenges in Data-Intensive Science and Exascale Computing has reviewed current practice and future plans in multiple science domains in the context of the Big Data and the Exascale Computing challenges that they will face in the future. The review drew from public presentations, workshop reports and expert testimony. Data-intensive research activities are increasing in all domains of science, and exascale computing is a key enabler of these activities. We briefly summarize below the key findings and recommendations from this report from the perspective of identifying investments that are most likely to positively impact both data-intensive science goals and exascale computing goals.

Finding 1: There are opportunities for investments that can benefit both data-intensive science and exascale computing. There are natural synergies among the challenges facing data-intensive science and exascale computing. Data-intensive science relies on the collection, analysis and management of massive volumes of data, whether they are obtained from scientific simulations or from experimental facilities. In the former case, investments in exascale computing will be essential for the simulations needed by data-intensive science. Further, in both cases (simulation or experimental), investments in exascale systems or, more generally, in “extreme-scale” systems¹ will be necessary to analyze the massive data involved in DOE’s science missions.

For example, high-throughput filtering and analysis capabilities are essential when processing large volumes of data generated by science instruments. While the computational capability needed within a single data analysis tier of an experimental facility may not be at the exascale, extreme scale processors built for exascale systems will be well matched for use in different tiers of data analysis, since these processors will be focused on optimizing the energy impact of data movement. An industry ecosystem for building exascale computers will include the creation of higher-volume extreme-scale system components which will be beneficial for data analysis solutions at all scales. For example, innovative memory hierarchies in data-intensive architectures (as in Figure 2.2) will be very useful for the analytics components of exascale systems.

The Exascale Computing Initiative has also identified the need for innovations in applications and algorithms so as to address fundamental challenges in extreme-scale systems related to concurrency, data movement, and resilience. Innovative solutions to these challenges will jointly benefit analysis and computational techniques for both data-intensive science and exascale computing. Finally, advances in networking (as projected for future generations of ESNet technology) will also

¹As in past reports, we use “exascale systems” to refer to systems with an exascale capability and “extreme-scale systems” to refer to all classes of systems built using exascale technologies which include chips with hundreds of cores and different scales of interconnects and memory systems.

benefit both data-intensive science and exascale computing.

Finding 2: Integration of data analytics with exascale simulations represents a new kind of workflow that will impact both data-intensive science and exascale computing.

In the past, the computational science workflow was represented by large-scale simulations followed by off-line data analyses and visualizations. Today's ability to understand and explore gigabyte and some petabyte spatial-temporal high-dimensional data in this workflow is the result of decades of research investment in data analysis and visualization. However, exascale data being produced by experiments and simulations are rapidly outstripping our current ability to explore and understand them. Exascale simulations require that some analyses and visualizations be performed while data is still resident in memory, so-called *in-situ* analysis and visualization, thus necessitating a new kind of workflow for scientists. In addition, we need new algorithms for scientific data analysis and visualization along with new data archiving techniques that allow for both in-situ and post processing of petabytes and exabytes of simulation and experimental data. This new kind of workflow will impact data-intensive science due to its tighter coupling of data and simulation, while also offering new opportunities for data analysis to steer computation.

In addition, in-situ analysis will impact the workloads that high-end computers have traditionally been designed for. Even for traditional floating-point-intensive applications, the addition of analytics will change the workload to include (for example) larger numbers of integer operations and branch operations than before. Design and development of scalable algorithms and software for mining big data sets, as well as an ability to perform approximate analysis within certain time constraints will be necessary for effective in-situ analysis. In the past, different assumptions were made for designing high-end computing systems vs. analysis and visualization systems. Tighter integration of simulation and analytics in the science workflow will impact co-design of these systems for future workloads, and will require development of new classes of proxy applications to capture the combined characteristics of simulations and analytics.

Finding 3: There is an urgent need to simplify the workflow for data-intensive science.

Analysis and visualization of increasingly larger-scale data sets will require integration of the best computational algorithms with the best interactive techniques and interfaces. The workflow for data-intensive science is complicated by the need to simultaneously manage large volumes of data as well as large amounts of computation to analyze the data, and this complexity is increasing at an inexorable rate. These complications can greatly reduce the productivity of the domain scientist, if the workflow is not simplified and made more flexible. For example, the workflow should be able to transparently support decisions such as when to move data to computation or computation to data. The recent proposal for a Virtual Data Facility (VDF) will go a long way in simplifying the workflow for data-intensive science because of its integrated focus on data-intensive science across the DOE ASCR facilities.

Finding 4: There is a need to increase the pool of computer and computational scientists trained in both exascale and data-intensive computing.

Earlier workflow models allowed for a separation of concerns between computation and analytics that is no longer possible as computation and data analysis become more tightly intertwined. Further, the separation of concerns allowed for science to progress with personnel that may be experts in computation or in analysis, but not both. This approach is not sustainable in data-intensive science where the workflow for computation and analysis will have to be co-designed. There is a need for investments to increase the number of computer and computational scientists trained in both exascale and

data-intensive computing to advance the goals of data-intensive science.

Recommendation 1: The DOE Office of Science should give higher priority to investments that can benefit both data-intensive science and exascale computing so as to leverage their synergies. The findings in this study have identified multiple technologies and capabilities that can benefit both data-intensive science and exascale computing. Investments in such dual-purpose technologies will provide the necessary leverage to advance science on both data and computational fronts. For science domains that need exascale simulations, commensurate investments in exascale computing capabilities and data infrastructure are necessary for advancement. In other domains, extreme-scale components of exascale systems will be well matched for use in different tiers of data analysis, since these processors will be focused on optimizing the energy impact of data movement. Further, innovations in applications and algorithms to address fundamental challenges in concurrency, data movement, and resilience will jointly benefit data analysis and computational techniques for both data-intensive science and exascale computing. Finally, advances in networking (as projected for future generations of ESNet technology) will also benefit both data-intensive science and exascale computing.

Recommendation 2: DOE ASCR should give higher priority to investments that simplify the science workflow and improve the productivity of scientists involved in exascale and data-intensive computing. We must pay greater attention to simplifying human-compute-interface design and human-in-the-loop workflows for data-intensive science. To that end, we encourage the recent proposal for a Virtual Data Facility (VDF) because it will provide a simpler and more usable portal for data services than current systems. A significant emphasis must be placed on developing a collection of scalable data analytics and data mining algorithms and software components that can be used as building blocks for sophisticated analytics pipelines and flows. We also recommend the creation of new classes of proxy applications to capture the combined characteristics of simulation and analytics, so as to help ensure that computational science and computer science research in ASCR are better targeted to the needs of data-intensive science.

Recommendation 3: DOE ASCR should adjust investments in programs such as fellowships, career awards, and funding grants, to increase the pool of computer and computational scientists trained in both exascale and data-intensive computing. There is a significant gap between the number of current computational and computer scientists trained in both exascale and data-intensive computing and the future needs for this combined expertise in support of DOE's science missions. Investments in ASCR such as fellowships, career awards, and funding grants should look to increase the pool of computer and computational scientists trained in both exascale and data-intensive computing.