Data Polymorphism: Enabling Novel Data Representations For Next-Generation Scientific Data Management And Visualization

---

Overview:

The overarching goal of this project is to design and develop *an efficient and scalable framework to generate effective data representations* for exascale data management and visualization, to address the unprecedented amount and velocity of scientific data produced by simulations, instruments, and observations. Combining the recent advancements in data compression and error control theories, our key idea is to provision data in the most appropriate representation for fast and accurate visualization and analytics, allowing for polymorphic representations of the same data under distinct scenarios. This is motivated by the fact that the same data will be used for multiple scientific analyses with different requirements, and these analyses could be costly if degrees of freedom are not reduced during data compression. By representing data in progressive and multiresolution formats, we aim to fundamentally address the issues of single error tolerance and no reduction in degrees of freedom in state-of-the-art scientific data compressors. Research outcomes will be disseminated as software products and integrated into state-of-the-art data management and visualization tools after thorough validation with both edge devices and leadership computing systems. Success of this project will significantly reduce the time to insights for a wide range of applications running on diverse systems, thus advancing scientific discoveries across multiple disciplines.

Intellectual Merit:

The proposed research will leverage progressive and multiresolution representations to enable polymorphism on scientific data with three thrusts, to address the growing gap between *the generation speed and amount of scientific data* and *the bandwidth for data transmission and processing speed for data analytics*. First, a unified framework will be developed to generate progressive and multiresolution representations for data in both structured grid and unstructured grids, such that scientific data can be provisioned in any designated precision. Second, intelligent algorithms will be designed to determine the best-fit data representation under diverse user and system requirements. This includes rigorous analysis to preserve the outcome of downstream data analytics in addition to the precision of raw data. Third, we will ensure high performance, scalability, and portability in our software products, to allow for practical use in large-scale computing systems with heterogeneous architectures. This includes tailored optimizations for distributed processing and heterogenous systems with advanced accelerators. The proposed methods will be integrated into the leading scientific data management and visualization software including ADIOS-2 and VTK/VTK-m with usability and transparency in mind.

Broader Impacts:

The project will enhance research and engineering in advanced cyberinfrastructure by delivering publicly available software to the community. Results of this project will be submitted for publication in relevant peer-reviewed conferences and journals, and the related research topics will be integrated into workshops. Furthermore, this proposed work will contribute to the education, training, and workforce development of future cyberinfrastructure users and developers. In particular, an integrated educational program will be designed to foster interest in advanced cyberinfrastructure within both computer science programs and computational science programs, to engage undergraduates and graduates in STEM with interdisciplinary collaborations. Additional training will be provided in the form of summer workshops with a focus on underrepresented students through collaboration with the PI's current cyberTraining program. The PI will also host seminars and data reduction competitions in both the university and the local high schools for broad participation.