Tutorial: OpenViSUS Streaming-based Large-Data Visualization

Lecturer: Prof. Valerio Pascucci (University of Utah)

Abstract

In recent times, we have seen data creation consistently outpace the infrastructure for its storage and utilization. With the growing size of scientific simulations, increasing resolution of sensors and advent of big data, this chasm between production and utilization is wider than ever. The ability to generate large amounts of data leads to major bottlenecks when it comes to its movement, analysis and visualization. We present an end-to-end data management framework (available at http://ViSUS.org) to improve the three key stages of the data life-cycle: generation, movement and analysis. We begin by presenting PIDX, a highly scalable and tunable parallel I/O library that writes data in a multiresolution format.

Following which, we will present a suite of customized visualization and analysis tools that take advantage of the explicit nature of the data format and enable interactive analytics. A containerized web server will allow users to import/convert, share and stream their data. Attendees will be able to access the web server to explore large-scale datasets from their browser and perform interactive analysis and visualization using the framework through a Jupyter notebook interface. Through the tutorial, we will teach the audience the core design principles of this framework and also provide hands on training on how to use all the components.

Overview and goals of the tutorial

The tutorial gives an overview of the challenges faced with large-scale data management and presents a new set of techniques to tackle them. In particular, we will present the following three technologies: (i) PIDX, an open-source parallel I/O framework that enables HPC applications to directly write data in an analytics-appropriate multiresolution data format. The data format allows fast access to spatial subsets at varying resolution allowing scientists to decide the scale at which they want to perform their analysis task, avoiding the need to read or even to store data at finer scales. We will go over the API of the I/O library and teach how to integrate existing applications with the I/O system. (ii) Data streaming server: attendees will learn how to make their multiresolution data available for streaming using a containerized data portal and streaming server. (iii) Web-based and standalone viewers designed to enable analysis and visualization of data at varying resolution scale using javascript and python interfaces.

Target audience

Our target audience are researchers with high resolution (and throughput) imaging devices and developers of large-scale scientific simulations that routinely have to deal with large amounts of data. More specifically, we are aiming for application teams that are currently constrained by either: (i) file I/O; (ii) post-processing resources; or (iii) the inability to adequately process or move large data sets.

Tentative Program

Topic	**Time**
Introduction: big data from HPC and high throughput microscopes, challenges and objectives of this tutorial • Interactive visualization of large scale data from different domains • Quick demo of the entire framework: simulation producing data, real-time streaming visualization, simple scripting for interactive analysis **Hands-on 1:** Connect to the web viewer to explore live data interactively	40 min
Questions time	10 min
Data generation: simulations on HPC systems • PIDX, design, scalability • Code example of “simulation code” doing checkpoint and restart with PIDX Data access (local): • Access and visualize data with Paraview or VisIt visualization framework • Access and visualize data with standalone viewer	30 min
Questions time	10 min

Break (30 min.)

Data access (streaming): • Introduction of the streaming server • Demonstrate setting up of server and data portal using Docker • Publish (and share) the generated data • Showcase the framework for large image data, demo of the data portal for conversion of stack of image data to the multiresolution data format • Real time conversion monitoring using web interface **Hands-on 2:** Access the data portal to configure, import and visualize new data	30 min
Questions time	10 min
Data analysis • Python deployment (how to get OpenVisus with pip and conda) • Use Jupyter notebook for analysis and visualization of datasets • Interactive/progressive scripting in python **Hands-on 3:** Connect to the Jupyter server and interact with remote data using python **Hands-on 4 (advanced):** users can use their own python environment to query remote data and perform analysis using arbitrary resolutions of the data	40 min
Conclusions and resources (github, wiki, etc.)	5 min
Questions	10 min