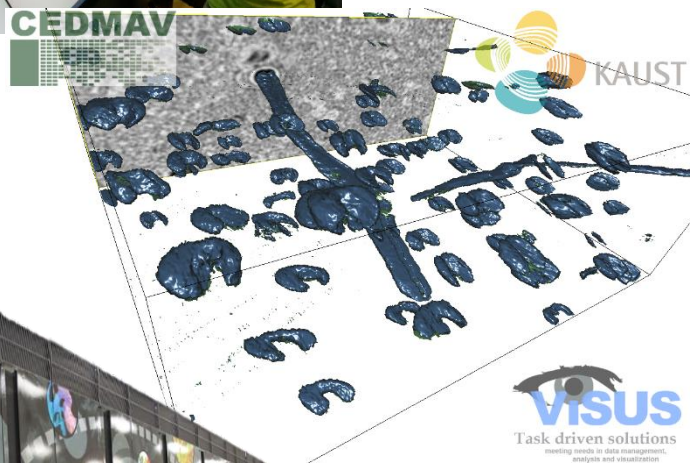
Big Scientific Data Made Simple

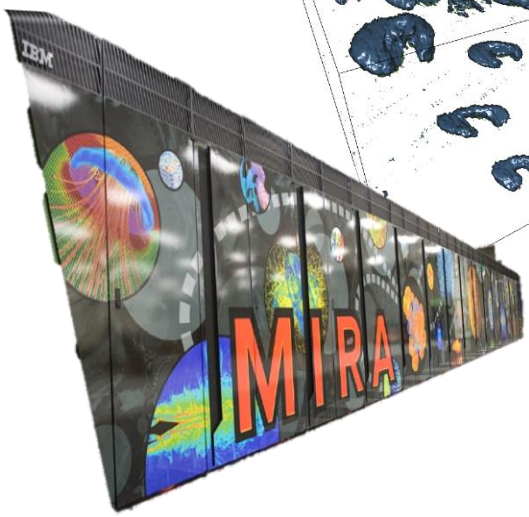
A Hands-on Tutorial in Data Generation, Processing, and Delivery for
High Performance Computing and High Resolution Imaging

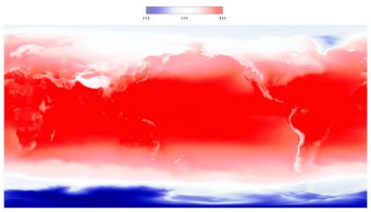
Dates: June 28-29 From 8:30am to 12noon
Location: KAUST Library Computer Classroom
Registration: http://tiny.cc/KAUST_BDM15_registration

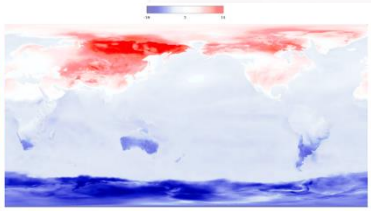
Abstract. Massive amounts of scientific data are an increasing challenge for scientists and engineers. Difficulties once limited to developing efficient output systems now involve the entire data lifespan, from creation and distribution through analysis and visualization. In this tutorial we will provide hands-on training in the use of a tool chain that improves each stage of data management while seamlessly integrating into existing infrastructures. The tool chain begins with an I/O library that scales on HPC platforms with over 700K cores (e.g. on MIRA at ANL), directly producing a multi-resolution format which enables a fast, flexible post-processing. This is complemented by a lightweight web-based server that enables data streaming for remote access as well as on-demand data conversion. The environment is completed with tools for data analysis and advanced visualization that can significantly shorten the time from simulation or experiment to scientific insight.

Instructors: V. Pascucci, S. Petruzza, R. Khurram, B. Hadri, S. Kumar (remotely from Salt LakeCity), and P.-T. Bremer (remotely from Lawrence Livermore National Laboratory).





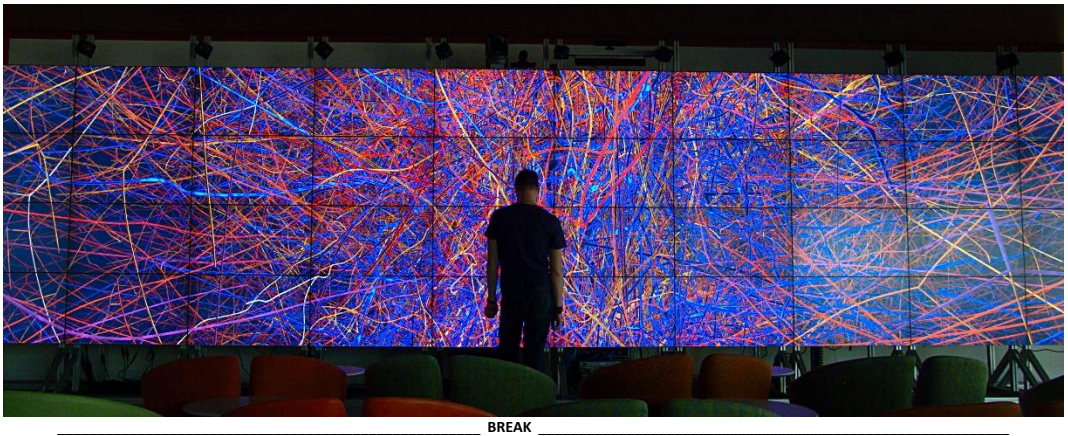




Detailed outline of the tutorial: Day 1

Morning of day 1 (June 28)

Topics	Minutes
Introduction and outline for the day: what users should expect and what are the prerequisites for taking full advantage of the tutorial.	**10**
Demonstrations of the full data management pipeline. - Local visualization of large dataset that are larger than the main memory of the current computer: (a) combustion simulation (b) brain microscopy. - Visualize data remotely: (a) server on Shaheen, (b) server in Utah, (c) server at Lawrence Livermore National Laboratory. - Visualize microscopy data while generated. - Run basic checkpoint/restart S3D proxy code: (1) select which variables to save and time frequency, (2) restart at any given point, (3) ability to visualize in streaming mode to monitor simulation. - Monitor a simulation from iPhone/iPad.	**20**
PIDX I/O: The library provides fully scalable file I/O solution, while preserving a simple and high-performance read interface. Have been scaled on multiple platforms and more than to 700K cores. Description of the characteristics of the library and the output file format. General presentation of PIDX structure, why it has advantages. Details of the different pipelines that can be set with PIDX infrastructure.	**45**
PIDX checkpoint and restart: Walking the participants of the tutorial though the process of downloading and compiling the library. API that needs to be used to link the library to a simulation. Practical considerations to use PIDX efficiently for checkpointing and restarts.	**15**
BREAK	
Demo on Display Wall	
BREAK	
PIDX advanced topics (AMR, ROI, reduced resolution, compression): Demonstrate advanced features of PIDX such as handling AMR simulations, and providing support for region-of-interest and reduced resolution, along with usage of state of the art block based compression techniques.	**40**
Server Installation and update: detailed description of the server installation process, prerequisites (apache). Testing that the data server is working properly. Connecting to a client and making sure that proper caching allows to maximize effective use of the server.	**25**
On demand data conversion: adding to the server the capability to access a number of datasets that are not already stored in streaming format.	**25**

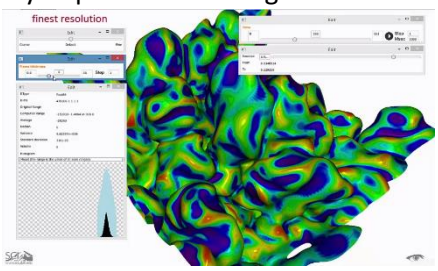


Detailed outline of the tutorial: Day 2

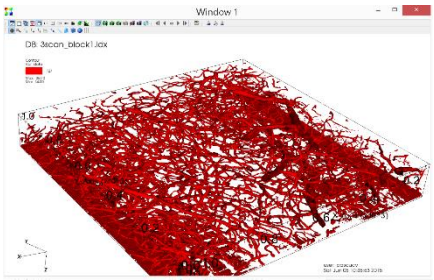
Topics **Minutes**

Morning of day 2 (June 29)

In-situ data analysis and reduction and post-processing exploration: In-situ analytics compute on an HPC platform (Titan) scaling over 200K core. Massive data reduction and interactive feature study in post-processing still with the ability of parameter change. **40**



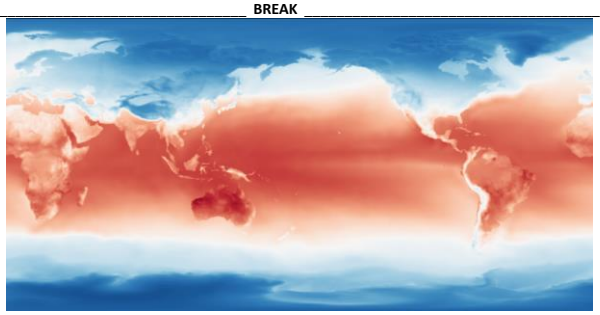
Real-time data analysis in streaming mode: Use of a complex analysis pipeline for understanding distribution of heat release over time for a KAUST combustion simulation. Interactive parameter study to validate stability of the results with respect to variation of the inputs. Generalization to other science cases. **20** **10**



Visit (or Paraview): Demonstrate and practice the visualization of data generated from PIDX using visit (or paraview) visualization framework. All normal processing applies. Show how to simply take advantage of the multiresolution nature of the data format to manage large datasets on commodity hardware.

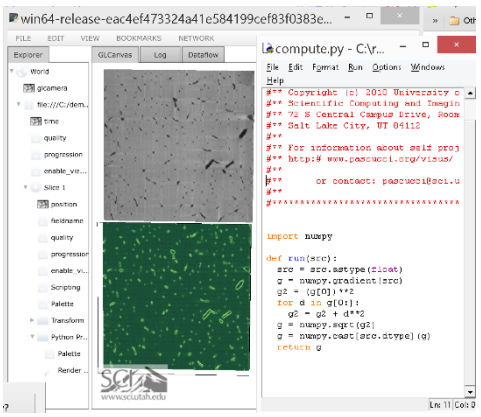
Ensemble analysis climate: Usage of the framework to analyzing collections of simulations. Focus on the case study of groups of runs of climate simulations as well as the comparison across groups of runs yielded by different code that may have different characteristics such as domain resolution. **20**

BREAK



BREAK

Scripting capabilities: Enabling advanced data processing and analysis capabilities vis scripting. Light scripting fully embedded in the streaming dataflow. Full python access for general connection to external libraries. **35**



Manual data conversion: Demo fast parallel post-process conversion of data to the format originally dumped by PIDX. **20**

Data from experiments: Collecting data from experiments with direct storage or quick conversion in IDX format. Enabling immediate data exploration. **20**

Conclusions and Q/A **15**