NVIDIA IndeX Accelerated Computing for Visualizing Cholla's Galactic Winds

Evan Schneider
Princeton University
es26@astro.princeton.edu

Brant Robertson
University of California Santa Cruz
brant@ucsc.edu

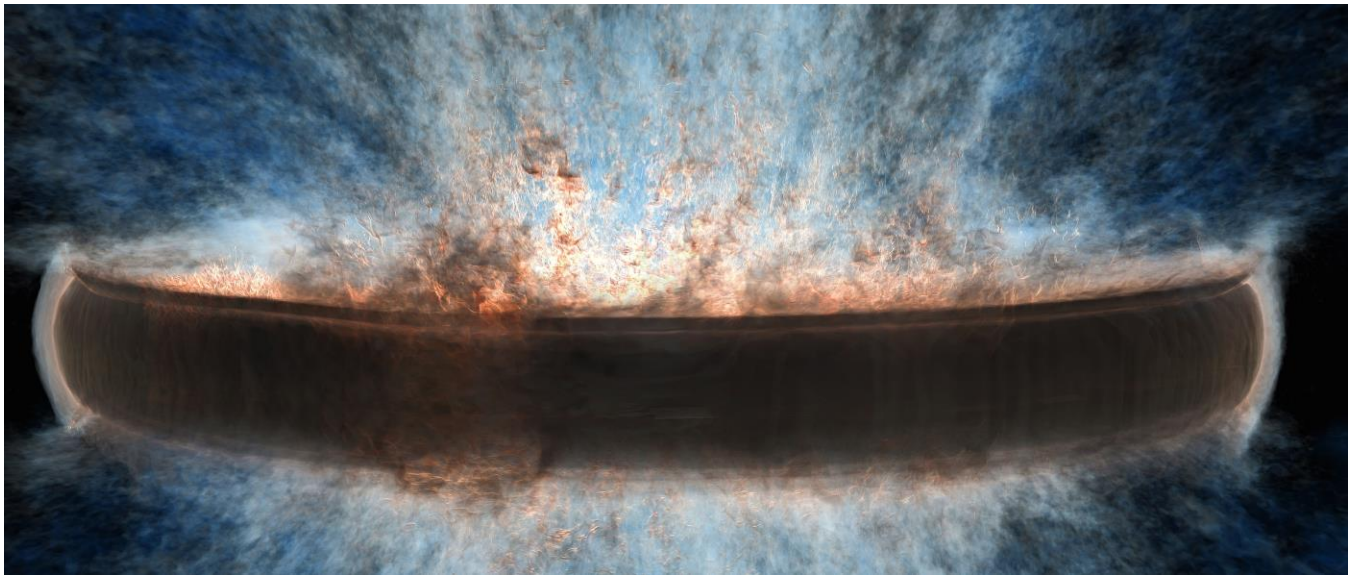
Alexander Kuhn
NVIDIA
akuhn@nvidia.com

Christopher Lux
NVIDIA
clux@nvidia.com

Marc Nienhaus
NVIDIA
mnienhaus@nvidia.com

Abstract

Galactic winds – outflows of gas driven out of galaxies by the combined effects of thousands of supernovae – are a crucial feature of galaxy evolution. By removing gas from galaxies, they regulate future star formation, and distribute the dust and heavy elements formed in stars throughout the Universe. Despite their importance, a complete theoretical picture of these winds has been elusive. Simulating the complicated interaction between the hot, high pressure gas created by supernovae and the cooler, high density gas in the galaxy disk requires massive computational resources and highly sophisticated software. In addition, galactic wind simulations generate terabytes of output posing additional challenges regarding the effective analysis of the simulated physical processes. In order to address those challenges, we present NVIDIA IndeX as a scalable framework to visualize the simulation output. The framework features a streaming-based architecture to interactively explore simulation results in distributed multi-GPU environments. We demonstrate how to customize specialized sampling programs for volume and surface rendering to cover specific analysis questions of galactic wind simulations. This provides an extensive level of control over the visualization while efficiently using available resources to achieve high levels of performance and visual accuracy.



Galactic Winds. When a galaxy experiences a period of rapid star formation, the combined effect of thousands of supernovae exploding within the disk drives gas to flow out of the galaxy and into its surroundings, as seen in the simulation rendered here. These "galactic winds" are a crucial feature of galaxy evolution, Simulating galactic winds requires simulations with extremely high resolution, as can only be achieved with scalable codes like Cholla and massive computational resources like Titan.

1. Introduction

Galactic outflows are the result of supernovae exploding within a galaxy. In galaxies that are experiencing a particularly high rate of star formation, the combined effects of thousands of these bombs going off in rapid succession creates regions of high pressure, low density gas that carve holes in the galaxy disk and explode out of it, powering a galactic wind. This high velocity gas continues to interact with the cooler, denser gas in a galaxy's disk, driving it out several thousand light years or beyond. Our modern theoretical picture of galaxy evolution indicates that most, if not all, galaxies have gone through one or more periods of hosting powerful galactic winds, making a complete understanding of this process critical for our understanding of how galaxies evolve over

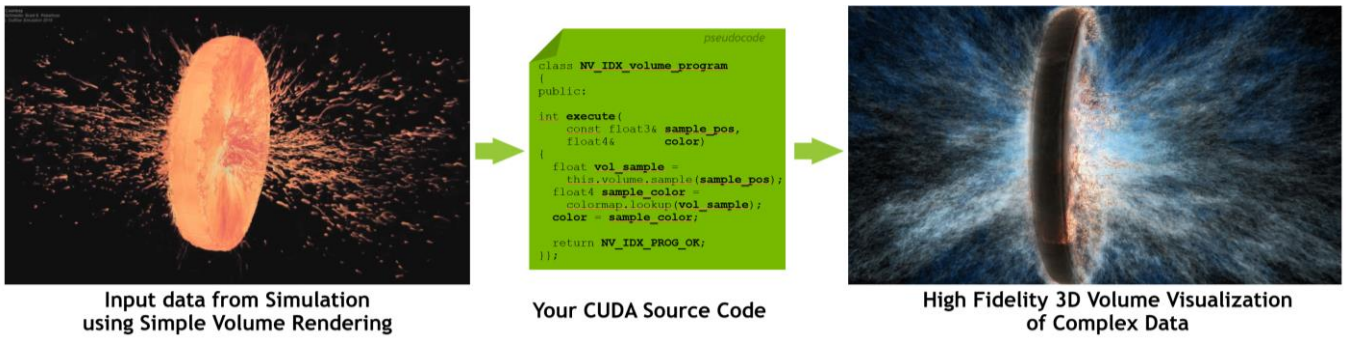


Figure 1. NVIDIA IndeX Accelerated Computing (XAC) Technology. The XAC technology provides the infrastructure to program, compile and execute CUDA code at runtime enabling scientist and laymen alike to carve out complex structure and create high-fidelity data visualization from raw input data.

cosmic time. Massive supercomputing resources combined with modern astrophysical simulation software and scalable analysis packages are making such an understanding possible for the first time.

In order to accurately simulate a galactic wind, we require simulation volumes that are both large enough to capture galaxy-scale features, like the disk, and have high enough resolution to capture the complicated hydrodynamical interactions that govern the evolution of the winds themselves. These simulations typically have billions of cells and produce terabytes of data, making their analysis a research problem comparable in scale to running the simulation! Until this point, we have been restricted to relatively simple analyses that rely on aggregate statistics of the data produced, for example, calculating the average density and velocity of gas in the wind as a function of its distance from the disk. With sophisticated analysis tools like IndeX, we are able to delve into more complicated features of these winds, examining individual structures like the small, dense clouds seen embedded within the wind in the cover image. This analysis tool is therefore opening new territory in our ability to recover information from galactic wind simulations, vastly increasing the value of these expensive numerical experiments.

2. Related Work

Volume rendering has a long history stretching back to the original slice-based approach of Drebin [1]. NVIDIA IndeX is spiritually similar to the ray casting method of Levoy [2], with a block-based decomposition approach such as that first proposed by Hadwiger et al. [4]. For distributed parallel volume rendering on the GPU, most modern systems follow the ray-guided approach of Fogal et al. [10]. We refer readers to the full survey by Beyer et al. [3]. More recent distributed rendering research has pursued efficient compositing approaches built on top of CPU volume rendering [5, 6]. On the GPU, research has explored novel interconnects such as GPUDirectRDMA [7]. NVIDIA IndeX [8, 9] is a scalable, distributed volume visualization product that builds on similar concepts, utilizing the latest hardware technologies such as NVLink/NVSwitch and the NVIDIA software technologies/stack including CUDA and DiCE.

3. Cholla

Cholla: Computational Hydrodynamics On || Architectures.

Cholla is a GPU-based hydrodynamics code that was designed to be massively-parallel and extremely efficient, and has been run on some of the largest supercomputers in the world. Originally envisioned as a solution to the relatively large fractional computational expense that hydrodynamics comprises in astrophysical simulations, it has since evolved into a software that can compute many complicated physical processes, from gravity to radiative cooling – all while taking advantage of the architectural advantages of the GPU. When run on modern GPU systems such as ORNL's Summit [12], Cholla can perform hydrodynamics calculations at a rate of over 50 million cell updates per second per GPU, allowing us to run simulations at unprecedented resolution. These simulations take advantage of Cholla's excellent weak scaling in order to create datasets consisting of billions of cells, with individual time snapshots approximately a terabyte in size. Many such snapshots are produced over the course of a single simulations, making analysis of spatial properties a challenge, and temporal analysis even more complex.

4. Visualizing Ever-Increasing Dataset Sizes

Powerful and scalable simulation systems, such as Cholla, are able to take advantage of the immense computational resources provided by todays supercomputers, such as the ORNL Titan [11], the machine on which these simulations were run. The simulations run on large supercomputers today already produce tens to hundreds of terabytes of n-dimensional, time-dependent volumetric datasets. With every advancement of the simulation systems or the availability of even larger compute resources, such as the ORNL Summit supercomputer [12], the resulting datasets will quickly grow to the petabyte scale.

Interactive visual examination and exploration of simulation results is key to monitoring the simulation output and eventually studying and gaining new insight into the simulation matter. The sheer size of the simulation data we are presented with dramatically limits the scope of the available visualization software systems, that are able to directly work on the unaltered data without potentially lossy data compression.

The NVIDIA IndeX rendering system[1] utilizes GPU clusters for the scalable, real-time visualization of large-scale datasets, consisting of volumetric data as well as surface data primitives. NVIDIA IndeX overcomes the limitations of single computer systems, e.g., limited GPU memory and compute capacity, by utilizing the distributed resources of multi-GPU environments, starting from single multi-GPU nodes or multi-node GPU clusters and up to full supercomputers. Through the effective allocation of the appropriate GPU cluster resources, larger dataset sizes are efficiently processed on GPUs while ensuring interactive framerates. Additionally, the efficient use of multi-GPU cluster systems enables the analysis of data at its original resolution; no additional pre-processing steps for data reduction or abstraction (e.g., level-of-detail) are required. This ensures high visual fidelity of the rendering and reduces the risk of losing visual features or introducing artifacts due to the additional processing. Furthermore, the effective parallelization of the visualization process allows for enhanced visualization techniques (c.f., Figure 3 and Figure 4) at interactive frame rates. The definition of such enhanced techniques is crucially enabled through the NVIDIA IndeX Accelerated Computing Technology (XAC) (c.f., Section 4.2.). It provides a simple-to-use mechanism to define and modify visualization strategies interactively.

4.1. Distributed, Interactive Rendering

The main approach to the distributed data handling in NVIDIA IndeX is data subdivision. The source datasets are split into spatially disjoint subsets. These subsets are assigned to individual nodes and eventually GPUs in the allocated visualization (sub-) cluster. This assignment is steered by varying factors such as memory footprint as well as expected processing overhead for the rendering of particular data subsets. The NVIDIA IndeX system aims to evenly distribute the data and rendering workload amongst its processing nodes for an efficient parallel rendering approach.

Data import for NVIDIA IndeX is handled either through user-defined distributed import callbacks or it is handled in-Situ/in-Trans during the simulation process. The import callbacks allow users to input their custom or proprietary data formats into the visualization process. For each data subset an import request is generated and directed to the runtime application to fill in the required data. These requests are processed in parallel and allow users to read data from storages specific to their application setup. With simulation data results in the range of hundreds of terabytes even temporarily storing them might be a limiting factor due to constraints of the storage as well as bandwidth to the storage system. For such use-cases NVIDIA IndeX supports in-Situ and in-Trans workflows. In these cases, the simulation data is either visualized directly on the simulation nodes that process a particular data subset or the results are transferred to an allocated visualization (sub-)cluster for the visualization without intermediate storage of the data.

Each GPU in a NVIDIA IndeX visualization cluster performs the rendering based on a ray-casting approach directly on the GPUs based on custom NVIDIA CUDA kernels. Regular-grid volume data, as generated by the Cholla simulation, is processed using a front-to-back volume ray-casting approach. Each volume subset is rendered individually by a GPU and results in a small frame-buffer fragment. These fragments are transferred and assembled to the final rendering result by a parallel compositing process. This process also distributes the compositing workload (data transfer and image processing) amongst different nodes in the visualization cluster in order to provide the final rendered image with as little added latency as possible.

While the core NVIDIA IndeX rendering system returns the rendering result as a raw image, NVIDIA IndeX is shipped with a supporting application layer that provides a client-server application with integrated video streaming functionality. On top of this functionality a browser-based HTML5 viewer-client application is provided. This application allows easy access to a running, distributed visualization process from a range of devices. The viewer application provides user interface elements to interactively set rendering parameters, such as camera, transfer function, volume filter quality and time step parameters. It further allows the user to interactively modify or define the NVIDIA IndeX Accelerated Computing Technology programs applied to the visualization – directly from a browser window without immediate access to the running application process, rendering machine or cluster.

4.2. NVIDIA IndeX Accelerated Computing Technology for Scientific Data Visualization

One of the defining features of the NVIDA IndeX rendering system is the ability to directly control the appearance of the visualization through an extensible programming interface called NVIDIA IndeX Accelerated Computing Technology (XAC).

XAC programs are similar to shading programs in the traditional rasterization pipeline, such as vertex or fragment programs. In our ray-casting based pipeline small XAC shading programs are applied to data samples. We principally differentiate two sample-program types: volume sample and surface sample programs (c.f., Figure 2). Surface sample programs are executed during the rendering approach for intersections of viewing rays with geometric surface primitives in the scene, such as triangle meshes or probing planes. Volume sample programs are executed for each sample taken during the volume ray-casting sampling traversal.

[1] NVIDIA IndeX is available as an SDK or Paraview Plugin. It is free for the scientific community and for non-commercial use.

Figure 2. Programming the inner loop of the visualization pipeline. The XAC technology enables scientists to inject CUDA code into the inner loop of the ray-caster/tracer that is then evaluated at each volume sample and at each geometry intersection. The diagram on the left illustrates (top) the volume samples and geometry intersections along a ray cast through a scene and (bottom) that the evaluated XAC shader code determines the color at these positions. The snapshot of NVIDIA IndeX viewer on the right shows a Galactic Winds visualization, the XAC shader code and the color map editor that are both used to create the visualization.

The program types share a lot of common features offered by the NVIDIA IndeX XAC interface. It is possible for XAC programs to access or sample other data primitives present in the scene. This opens a wide range of visualization possibilities. Volumetric datasets can be explored or more finely investigated by applying special analysis programs to probing geometries. Multiple volume datasets can be combined into a multi-valued visualization emphasizing different features of the datasets in a single display. Furthermore, the XAC interface offers a set of utility functions to access and apply materials, lights and colormaps to the data visualization or to generate, for instance, volume gradients for particular data samples.

The XAC programs are defined as small CUDA C++ programs and applied to dataset elements the general scene description. These programs are compiled and incorporated into the inner rendering loop of the rendering kernels instantly at runtime. This facilitates easy, quick and interactive iterations on the visualization strategies. With the scalable architecture of NVIDIA IndeX complex XAC visualization kernels can always be handled at interactive frame rates. The second part of the accompanying video demonstrates a live visualization session modifying several XAC programs interactively.

5. Results and Scientific Value

The ***Cholla Galactic OutfLow Simulations*** (CGOLS) suite utilized an ORNL INCITE allocation on the Titan supercomputer to simulate galactic winds on scales of ~10 kiloparsecs with a resolution of ~5 parsecs - over an order of magnitude higher resolution than had ever been used to simulate a single galaxy. The final CGOLS suite consists of five simulations that range from a simple wind model with a single spherical injection region 300 parsecs in radius at the center of the galaxy, to more complex models that allow the winds to cool radiatively, or break the spherical symmetry of the injection region. Later simulations increased the physical realism of the supernova feedback by allowing the mass and energy injection rates within each star cluster to vary with time according to stellar population synthesis models. The video of this Supercomputing submission and the images in Figure 3 show a fly-through of the density field at a single point in time for a CGOLS simulation where clusters were distributed throughout the disk.

The ability to render these massive datasets relatively easily and with such high fidelity is leading to breakthroughs in our understanding of the physical properties at work within galactic winds. In particular, we now have the ability to capture and visualize the way that small, dense clouds of disk gas are lofted into the wind and then torn apart and shredded into long filaments, as seen in the cover image. The large surface areas of these filaments allow the disk gas to mix efficiently with the hot, low density gas produced within the supernova clusters, which enhances momentum transfer from the hot gas to the cold, solving a long-standing question about how cool gas is accelerated in galactic winds. In addition, the ability to render datasets using combinations of more than one attribute – for example, gas internal energy as well as density – is helping us uncover the relationships between gas in different phases in the outflow. As we continue to analyze these valuable datasets, we will continue to improve our understanding of the physics at work in galactic winds, and thus, our understanding of the process of galaxy evolution.

5.2. Dataset Sizes and Systems

The datasets presented in the associated images and videos are generated by the Cholla Galactic OutFlow Simulations suite on the Titan supercomputer. All images and videos are generated using the NVIDIA IndeX rendering system:

- The live presentation shown during the Supercomputing 2018 keynote [9] showed 7TB of time-varying volumetric data. The visualization was run on a GPU cluster consisting of 28 DGX-1 machines with each containing 8 Quadro V100 GPUs using 32GB VRAM. The demonstration with a framerate of at least 20 frames per second. The visualization cluster was situated in Santa Clara and the visualization results were video-streamed in real-time to the conference show-floor in Dallas.
- The technical overview part of the accompanying video shows a 3.2TB time series of the galactic wind simulation volumetric dataset consisting of 400 time-steps. The live visualization session was captured running on 18 DGX-1 machines running in real-time above 20 frames per second.
- The images shown in Figures 3 and 4 as well as the first part of the accompanying video show a single snapshot of a simulated time-series. This volume dataset is 8GB in size and was visualized interactively on a regular workstation running a single Quadro V100 GPU.

Acknowledgements

The simulations were carried out on the ORNL Titan supercomputer. BER acknowledges support from NASA grant 80NSSC18K0563, contract NNG16PJ25C.

References

[1] Drebin, Robert A. and Carpenter, Loren and Hanrahan, Pat, Volume Rendering, ACM Siggraph Computer Graphics 22(4), 1988.
[2] Levoy, Marc, Efficient Ray Tracing of Volume Data, ACM Transactions on Graphics (TOG), 9(3), 1990.
[3] Beyer, Johanna and Hadwiger, Markus and Pfister, Hanspeter, State-of-the-Art in GPU-Based Large-Scale Volume Visualization, Computer Graphics Forum, 34(8), 2015.
[4] Hadwiger, Markus and Sigg, Christian and Scharsach, Henning and Bühler, Khatja and Gross, Markus, Real-time Ray-casting and Advanced Shading of Discrete Isosurfaces, Computer graphics forum, 24(3), 2005.
[5] Biedert, Tim and Werner, Kilian and Hentschel, Bernd and Garth, Christoph, *A Task-Based Parallel Rendering Component for Large-scale Visualization Applications*, Eurographics Symposium on Parallel Graphics and Visualization, Alexandru Telea and Janine Bennett (Eds.). The Eurographics Association, 2017.
[6] Wu, Qi and Usher, Will and Petruzza, Steve and Kumar, Sidharth and Wang, Feng and Wald, Ingo and Pascucci, Valerio and Hansen, Charles D , VisIt-OSPRay: Toward an Exascale Volume Visualization System, EGPGV, 2018.
[7] Grosset, AV Pascal and Prasad, Manasa and Christensen, Cameron and Knoll, Aaron and Hansen, Charles, TOD-tree: Task-Overlapped Direct Send Tree Image Compositing for Hybrid MPI Parallelism and GPUs, IEEE transactions on visualization and computer graphics}, 23(6), 2016.
[8] NVIDIA Index. https://developer.nvidia.com/index, 2019.
[9] NVIDIA IndeX in NVIDIA Supercomputing 2018 Keynote. https://www.youtube.com/watch?v=PQbhxfRH2H4#t=1h02m00s, 2018.
[10] Fogal, Thomas and Schiewe, Alexander and Krüger, Jens, An Analysis of Scalable GPU-Based Ray-Guided Volume Rendering, IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV)}, 2013.
[11] Oak Ridge National Laboratory (ORNL) Titan Supercomputer. URL: https://www.olcf.ornl.gov/titan/, 2012.
[12] Oak Ridge National Laboratory (ORNL) Summit Supercomputer. URL: https://www.olcf.ornl.gov/summit/, 2018.

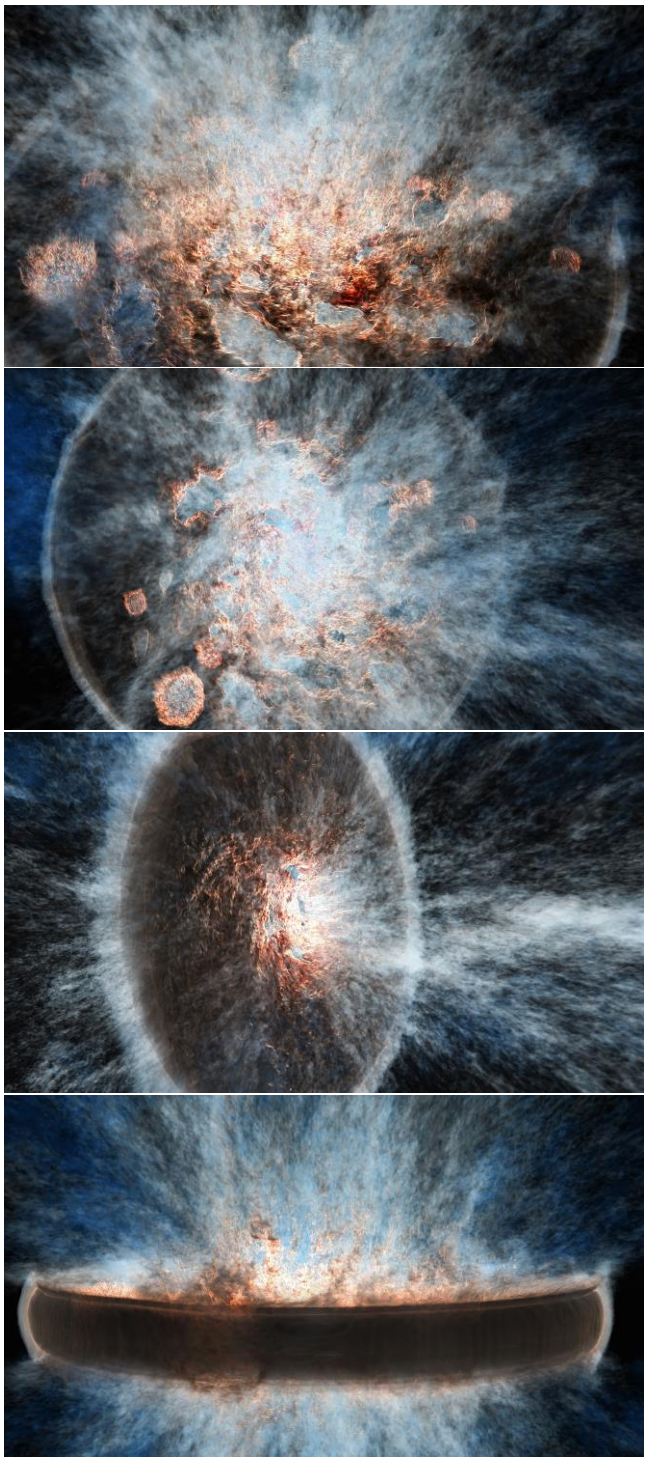


Figure 3. Series of Stills showing the Galactic Winds. The above renderings are still of a fly-through of the density field at a single point in time for a Cholla Galactic Outflow Simulation where clusters were distributed throughout the disk. Such visualizations allow us to inspect the structure of the wind, and thus elucidate relationships between gas in different phases. For example, this rendering highlights the filamentary structure of high density gas in the outflow, which increases its surface area and aids in momentum transfer from one phase to another.

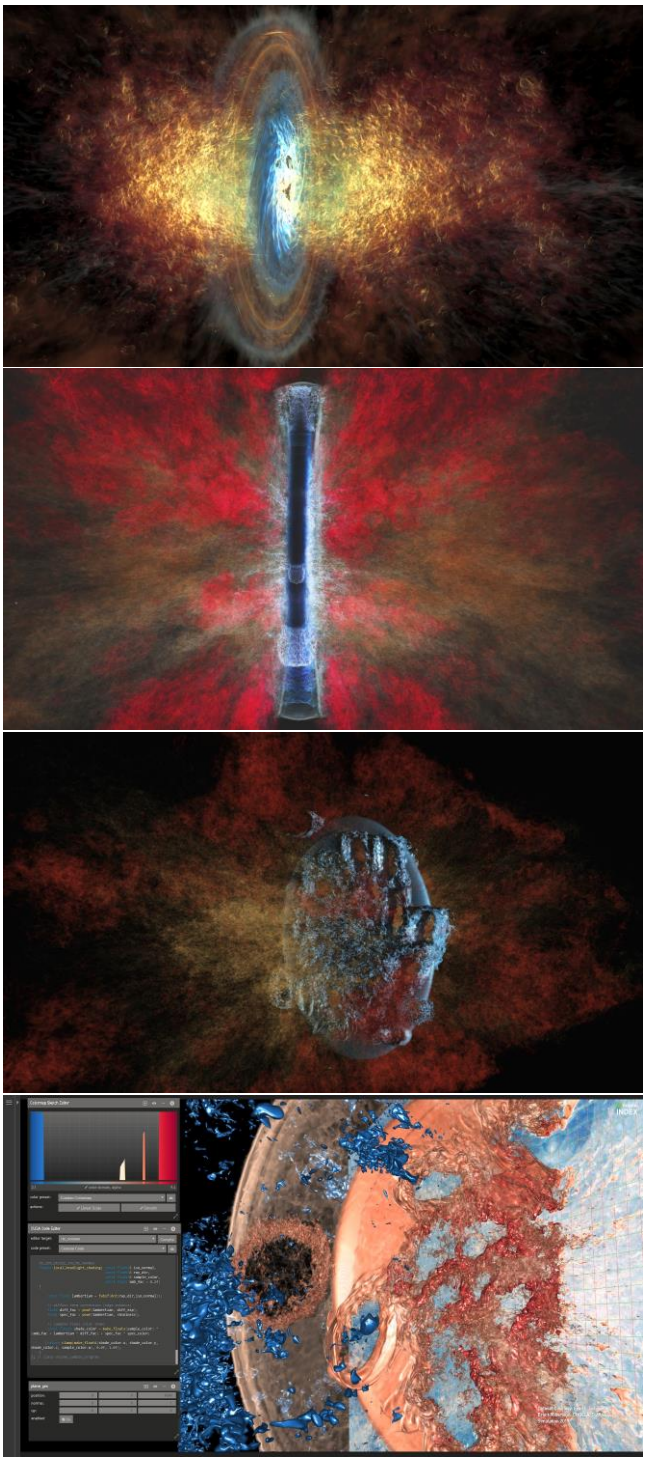


Figure 4. Variations of Visualizations. Minor changes to the original XAC shader code and the transfer functions result in visualizations that can highlight different aspects of the science and help focus on specific features that were not visible before. The image at the bottom is a snapshot of the NVIDIA IndeX viewer. The viewer helps scientist to explore data interactively using the XAC code editor and the transfer function editor and by positioning additional geometry into the scene, e.g., giving spatial reference.