ARGONNE
ATPESC 10
EXTREME-SCALE COMPUTING

Data Intensive Computing and I/O

ATPESC 2022 **August 5, 2022**

Rob Latham, **Phil Carns**, Shane Snyder,
Scot Breitenfeld, Mike Brim, and Greg Nawrocki

ECP Argonne NATIONAL LABORATORY

Welcome to Track 3 of ATPESC 2022:

Data Intensive Computing and I/O

We want to help you answer the following questions today:

1. What are the key things that I need to know about HPC data storage?
2. What data management tools are available, and how do I use them?
3. How can my application access data more efficiently?

Today’s topics at a high level

- Morning:
  - Introductory concepts
  - System walkthroughs
  - Instrumentation
  - Data movement
- Afternoon
  - I/O libraries
    - MPI-IO
    - PnetCDF
    - HDF5
  - Understanding and tuning performance
  - Discussion

Building up more detail as the day goes on

Meet your lecturers (Argonne staff)

Phil Carns is a computer scientist at ANL focused on measurement, modeling, and development of data services. He has made key contributions to influential storage research projects including Mochi, Darshan, CODES, and PVFS.

Rob Latham is a principal software development specialist at ANL who strives to make applications use I/O more efficiently. He has played a prominent role in the ROMIO MPI-IO implementation, the PVFS file system, and the PnetCDF high level library.

Shane Snyder is a software engineer at Argonne National Laboratory. His research interests include the design of high-performance distributed storage systems and the characterization and analysis of I/O workloads on production HPC systems.

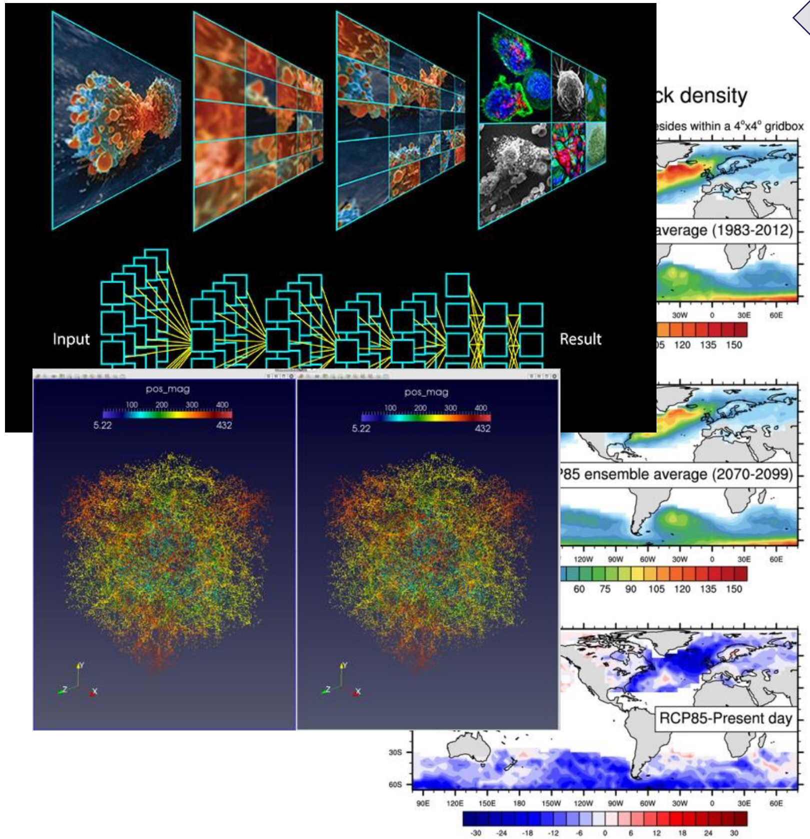
Meet your lecturers (Expert guests)

Scot Breitenfeld specializes in HPC application use of HDF5 at The HDF Group. He has implemented, troubleshot, and tuned HDF5 for a broad spectrum of HPC applications and third-party HDF5 based libraries for a variety of platforms.

Greg Nawrocki is the Director of Customer Engagement of the Globus Department at the University of Chicago. He has also worked in high energy physics, the television and consumer products industry, and as the co-founder of a data analytics company.

Mike Brim is an R&D staff member in the National Center for Computational Sciences at Oak Ridge National Laboratory. Michael has over 20 years of research experience in scalable tools, middleware, and systems software for HPC systems.

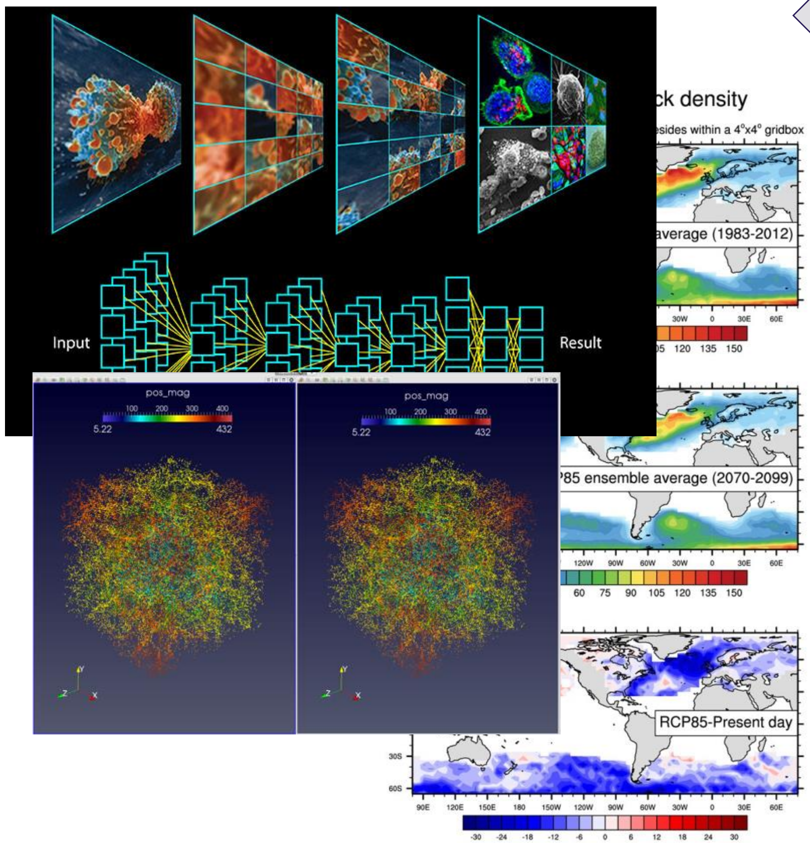
Why we do what we do:
bridging the gap between science and storage systems



There are many different high performance storage technologies available. How can we use these technologies to meet the needs of scientists?

We need techniques, algorithms, and software to bridge the “last mile” between storage systems and scientific applications.

Why we do what we do:
bridging the gap between science and storage systems



Examples of how we do this:

- Building/optimizing data services
- Operating data centers
- Understanding how storage is used
- Predicting how storage will be used
- **Putting new data storage technology into the hands of scientists**

Hands on exercises: https://github.com/radix-io/hands-on

Logistics for ATPESC-IO

ATPESC attendees have a dedicated reservation on Theta (ALCF) today for experiments and exercises from noon to 9pm, but you are welcome to compile and run jobs on any of the ATPESC systems.

- Agenda:
  - https://extremecomputingtraining.anl.gov/agenda-2022/#Track-3
- Discussion and questions:
  - Please ask questions as we go!
  - At least one of us will be monitoring the #track-3-io slack channel at all times.
    - We can provide one-on-one help and relay questions to lecturers if needed.
- Hands-on exercises and machine reservations:
  - See https://github.com/radix-io/hands-on
  - We don’t have much time blocked specifically for hands-on exercises.
  - Please work on exercises at your own pace.
  - Continue to reach out to us through the remainder of the ATPESC program if you have questions.

Thanks!

Any questions about logistics before we roll up our sleeves and get to work?