Data Intensive Computing and I/O

ATPESC 2020

Rob Latham, **Phil Carns**, Quincey Koziol,
Kathryn Mohror, Sarp Oral, and Shane Snyder

July 31, 2020

Thank you for joining us for Track 3 of ATPESC 2020!

Data Intensive Computing and I/O

We hope that today's lectures will help you answer the following questions:

How do HPC storage systems work?

What tools are available to assist with data management?

How can I access data more efficiently?

Topics

- Morning:
  - Introductory concepts and tools
  - MPI-IO and PnetCDF
- Afternoon
  - HDF5
  - Architectures
  - Tuning
  - Discussion

Building up more detail as the day goes on

ATPESC attendees have a dedicated reservation on Ascent (OLCF) and Theta (ALCF) today for experiments and exercises. See the link at the top of each slide for details.

Meet your lecturers

Phil Carns is a principal software development specialist at ANL who works on measurement, modeling, and development of data services. He has made key contributions to a variety of storage research projects, including Mochi, Darshan, CODES, and PVFS.

Quincey Koziol is a principal data architect at LBNL where he drives scientific data architecture discussions and participates in NERSC system design activities. He was the principal architect for the HDF5 project and a founding member of the HDF Group.

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.

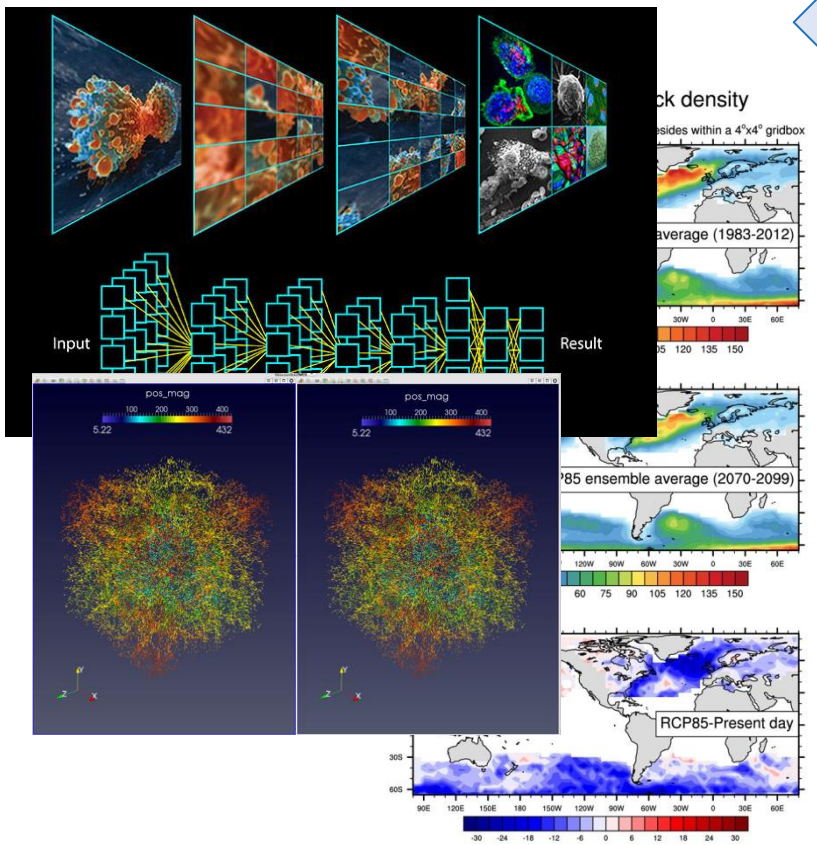
Kathryn Mohror is a computer scientist at LLNL who focuses on research for improving I/O performance of applications. She currently leads the UnifyFS and Scalable Checkpoint/Restart (SCR) projects.

Meet your lecturers (continued)

Sarp Oral is the Group Leader for the Technology Integration Group and a Senior Research Scientist at the National Center of Computational Sciences (NCCS) Division of Oak Ridge National Laboratory. His research interests are parallel I/O, benchmarking, high-performance computing and networking, fault-tolerance.

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

Your lecturers' day job: bridging the gap between applications and storage systems

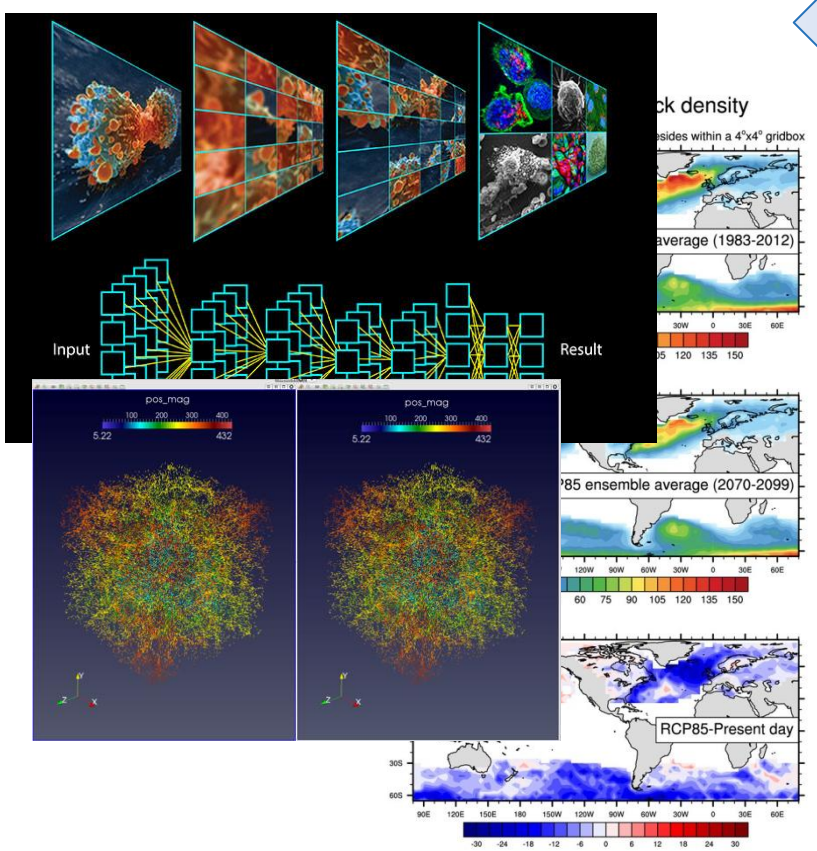


Techniques, algorithms, and software to bridge the "last mile" between scientific applications and storage systems.

Argonne NATIONAL LABORATORY

ECP EXASCALE COMPUTING PROJECT

Your lecturers' day job: bridging the gap between applications and storage systems



This means:

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

Argonne NATIONAL LABORATORY

ECP EXASCALE COMPUTING PROJECT

Logistics for ATPESC-IO

- Agenda:
  - https://extremecomputingtraining.anl.gov/agenda-2020/#Track-3
- Discussion and questions:
  - Please ask questions as we go!
  - At least one of us will be monitoring the #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://xgitlab.cels.anl.gov/ATPESC-IO/hands-on
  - Unfortunately we don't have a lot of time blocked for hands-on exercises.
  - Please work on exercises of interest at your own pace.
  - Continue to reach out to us through the remainder of the ATPESC program if you have questions.

Hands on exercises: https://xgitlab.cels.anl.gov/ATPESC-IO/hands-on

Thanks!

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