Leveraging High Performance Computing for Managing Large and Evolving Data Collections

Ritu Arora (rauta@tacc.utexas.edu)

Maria Esteva (maria@tacc.utexas.edu)

Jessica Trelogan (j.trelogan@austin.utexas.edu)

TACC

THE UNIVERSITY OF TEXAS AT AUSTIN

TACC

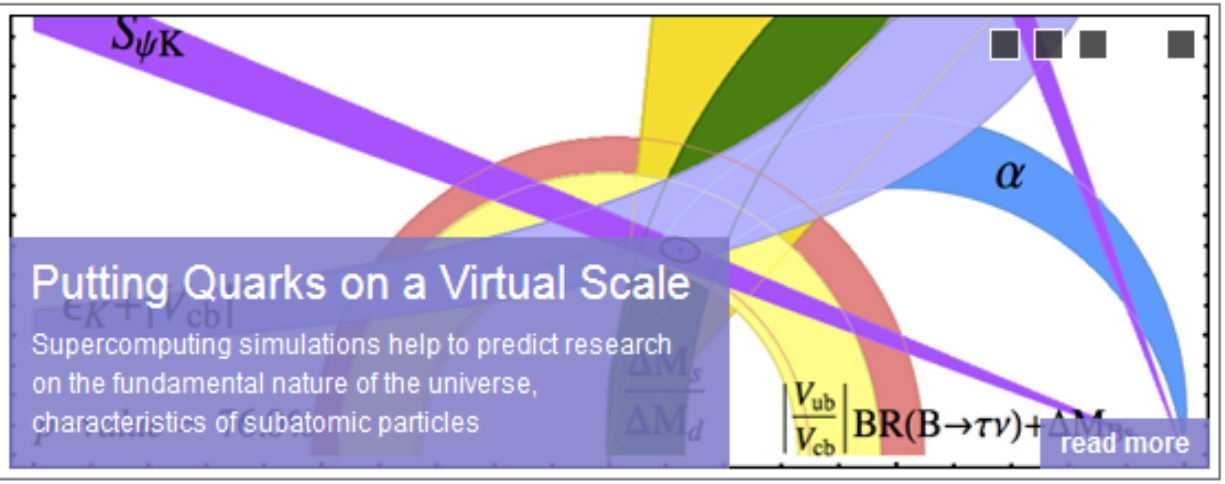
TEXAS ADVANCED COMPUTING CENTER

Powering Discoveries That Change The World

THE UNIVERSITY OF TEXAS AT AUSTIN

About | Systems & Software | Research & Development | User Services | Partnerships | Education & Outreach | News

TACC designs and provides the most powerful computing technologies in the world to enable discoveries that change the world.



Putting Quarks on a Virtual Scale

Supercomputing simulations help to predict research on the fundamental nature of the universe, characteristics of subatomic particles

read more

NEWS

ISC'14 S. African Student Cluster Team Visits Austin

@TACC February Newsletter

The latest news, events, and outreach activities delivered directly to you every month - sign up today

Dan C. Stanzione Jr. Named Acting Director of Texas

HOW CAN TACC HELP?

Researchers:

We help you determine which computing, analytics, storage, visualization & portal technologies help solve your most challenging problems.

LEARN MORE

Technology Collaborators:

We help you develop technologies that make advanced computing systems more powerful, reliable, and useful by more people

SEARCH

Search TACC

TACC USER PORTAL

A user management system for TACC accounts, projects, and allocations

TACC MOBILE IOS APP

TACC

THE UNIVERSITY OF TEXAS AT AUSTIN

THE INSTITUTE OF CLASSICAL ARCHAEOLOGY

The Institute of Classical Archaeology (ICA) carries out multi-disciplinary archaeological research, conservation, and cultural resource management projects in the territory of ancient Greek colonies in southern Italy and on the Black Sea coast of Ukraine. ICA was established as a research unit in the College of Liberal Arts at the University of Texas, Austin, in 1974. Over the past 30 years, with major support from public and private sources, ICA has developed long-term projects to explore the agricultural hinterlands of ancient **Metapontum** and **Croton** in Italy and the territory and urban area of ancient **Chersonesos** in Crimea, Ukraine. Principal collaborators include the Archaeological Superintendencies of Calabria and Basilicata and the National Preserve of Tauric Chersonesos at Sevastopol. ICA's publications and research have brought it international recognition as a leader in the study of rural populations in the Greek and Roman world.

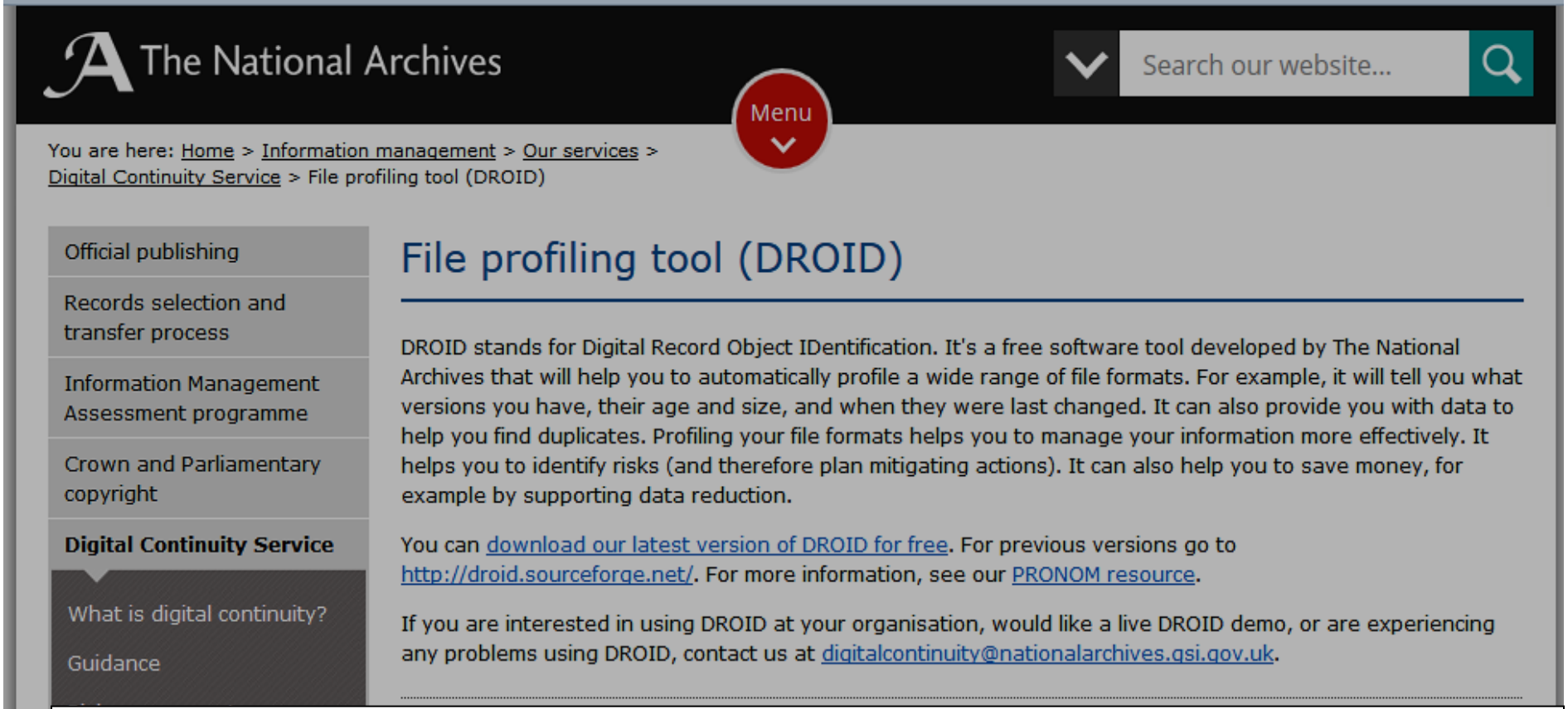
Chersonesos on UNESCO World Heritage List

Croton • Metapontum • Chersonesos

russkaya versia

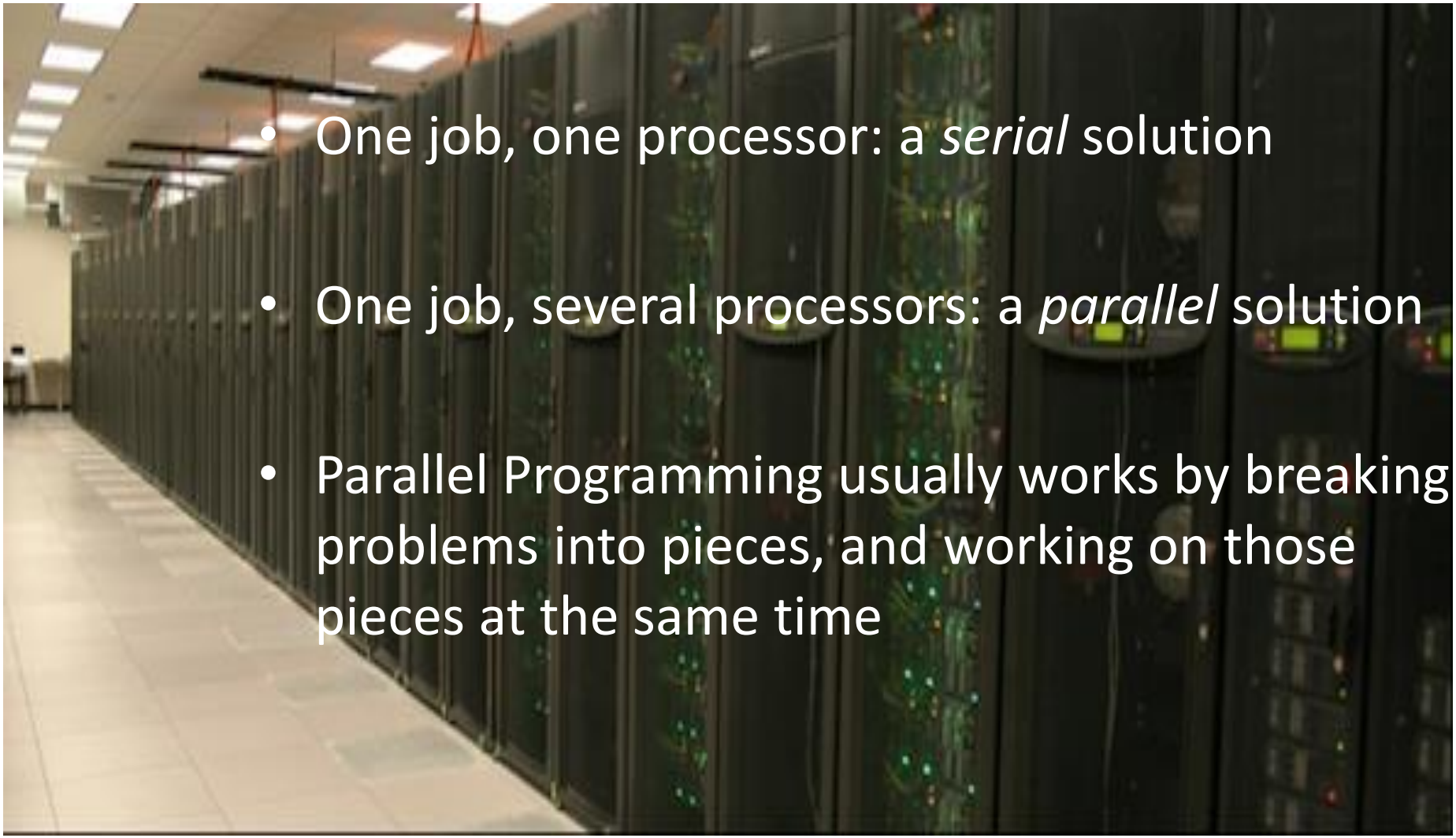
Visual Representation

- The visual representation of the collection at a given point is shown in the picture
- This representation is developed using the metadata extracted using DROID and other scripts
- The collection is currently 4.3 TB and has approximately 1 million files in deeply nested directories



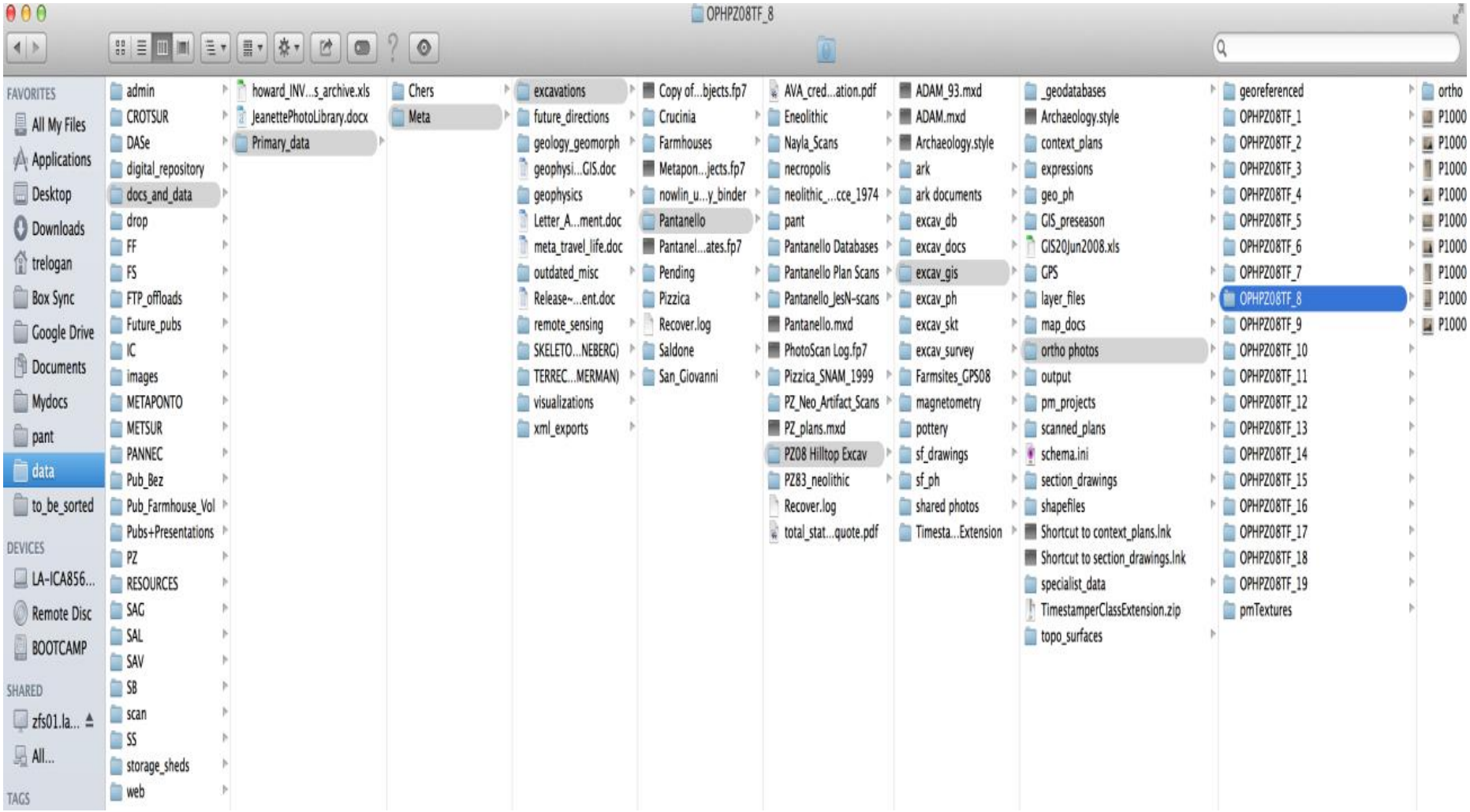
- Running one DROID instance to profile the 4.3 TB archeological data collection with mixed file formats takes
  - about two days to profile the collection on a slower processor
  - **6 to 8 hours** on a platform equipped with 2.7 GHz processors

Serial Versus Parallel

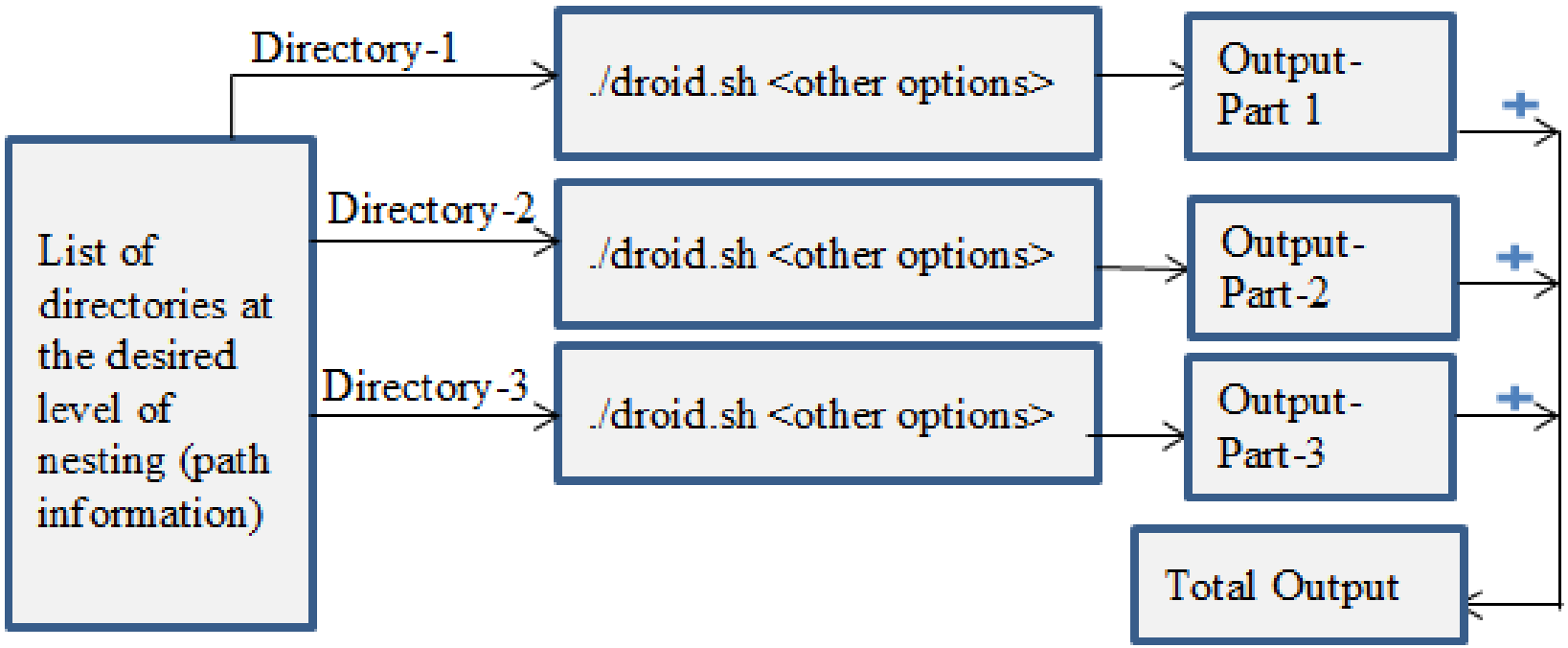


- One job, one processor: a *serial* solution
- One job, several processors: a *parallel* solution
- Parallel Programming usually works by breaking problems into pieces, and working on those pieces at the same time

Deeply Nested Archaeological Collection

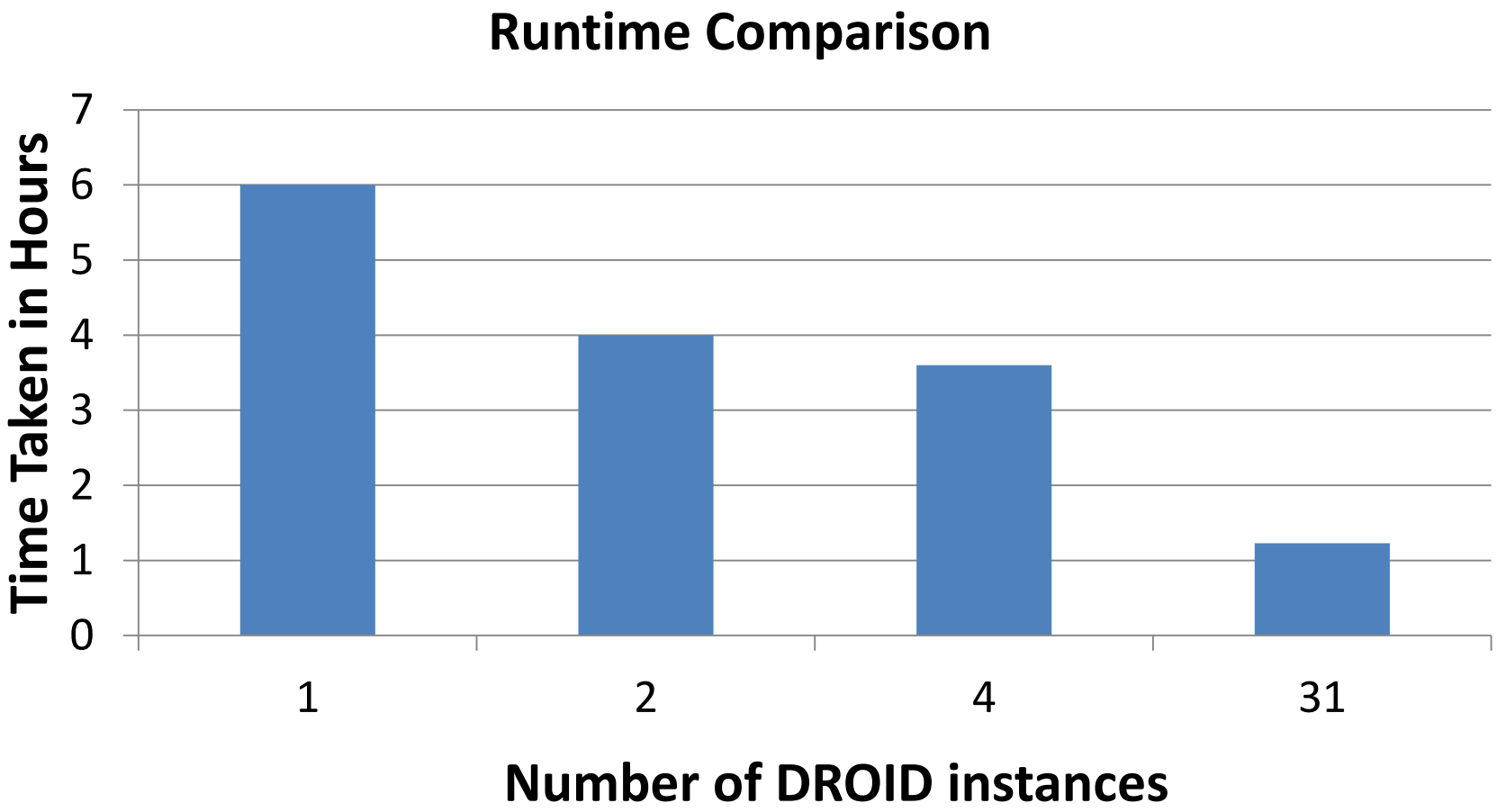


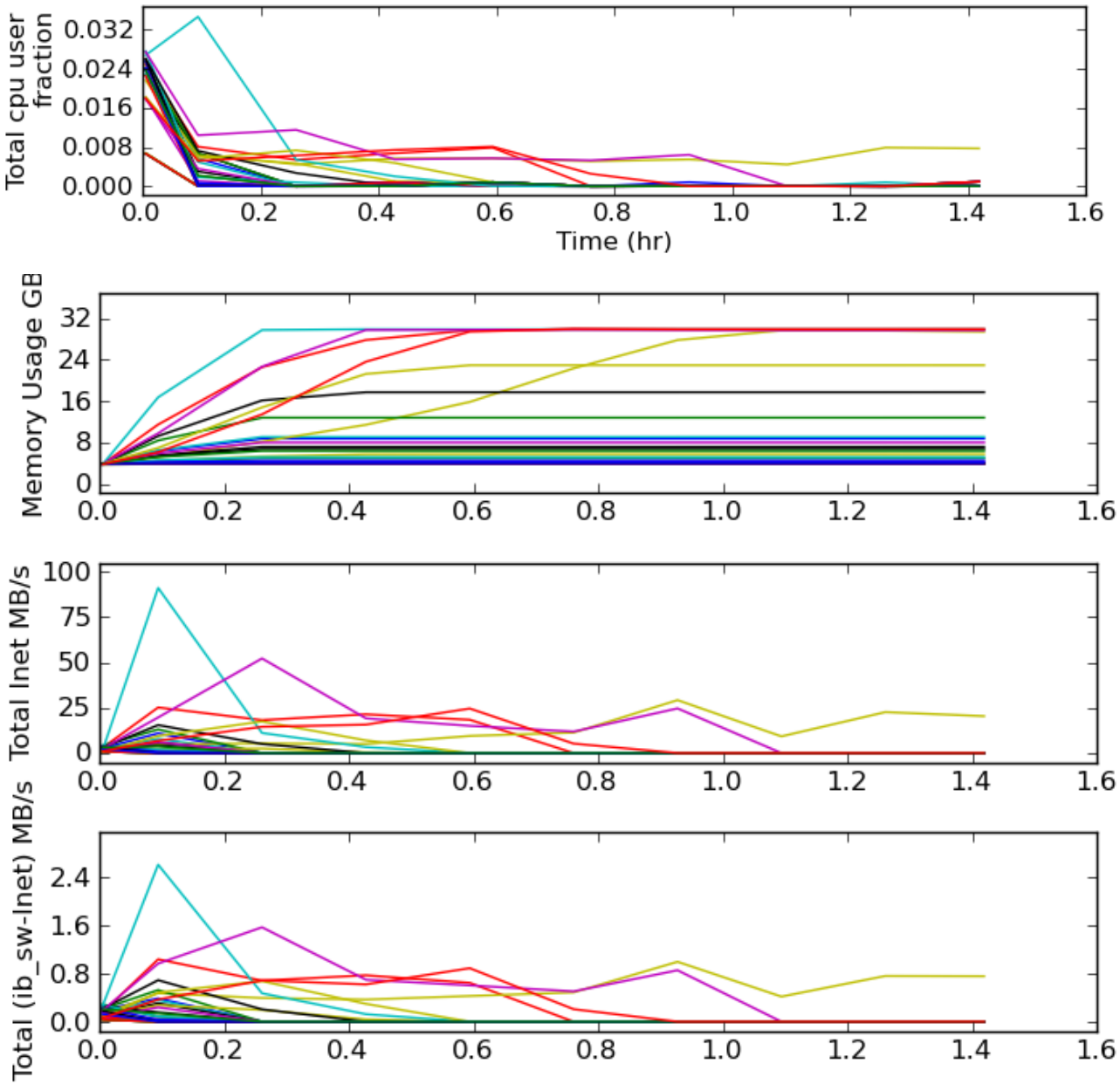
Running DROID in Parallel



1. **We ran 31 parallel instances of DROID after coarsely dividing the workload (directories) for file-profiling into 31 directories**
2. **The entire work-flow (work-load distribution, submitting multiple DROID jobs on the supercomputer, combining the results of multiple DROID runs, etc.) is mostly automated, few manual steps though**

Time-taken when Different Number of DROID Instances are Used Simultaneously

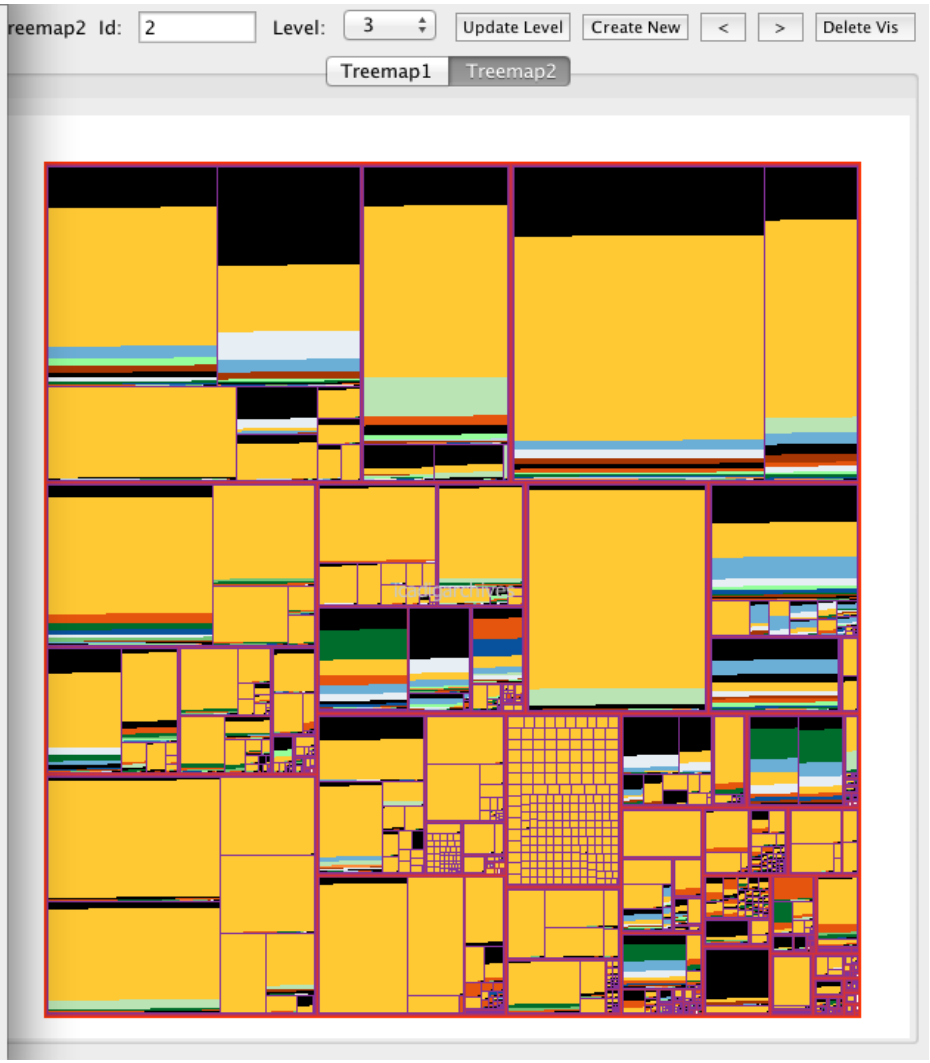




Running 31 Parallel Instances of DROID

Graphs show CPU usage, memory usage, and network bandwidth usage during parallel runs of DROID on Y-axis and time consumed on X-axis

Collection Before and After Cleaning



Lessons Learnt

- Network bottlenecks in transferring large data from remote storage locations to computing resources
  - These can be mitigated by access to high-speed networks, or by storing data in close proximity to the computing resources

- Need for software tools that are scalable/flexible enough to be used in high performance computing environments which are mostly Linux-based
  - Challenges in using DROID on our Stampede supercomputer

- Training of data curators is critical
  - Basic Linux training, running the scripts in the workflow, troubleshooting

Conclusion

- For efficiently managing large and evolving data collections, there is a need for fast data processing (reduced time-results)
  - Think data-intensive computing
  - Think open-science supercomputing resources and services

- We have demonstrated that data curation workflows can be adapted on supercomputing platforms to reduce the overall time-to-results

- Hence, on-demand data curation of large data collections is possible through the usage of scalable software tools and hardware infrastructure

Thanks!

Questions or Comments?