Argonne NATIONAL LABORATORY

… for a brighter future

U.S. Department of Energy

UChicago ▶ Argonne LLC

Office of Science U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

Analysis Performance and I/O Optimization

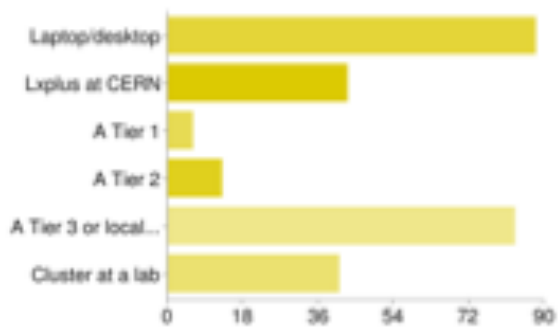
Jack Cranshaw, Argonne National Lab

October 11, 2011

Data for Analysis (September 2011 Survey Results)

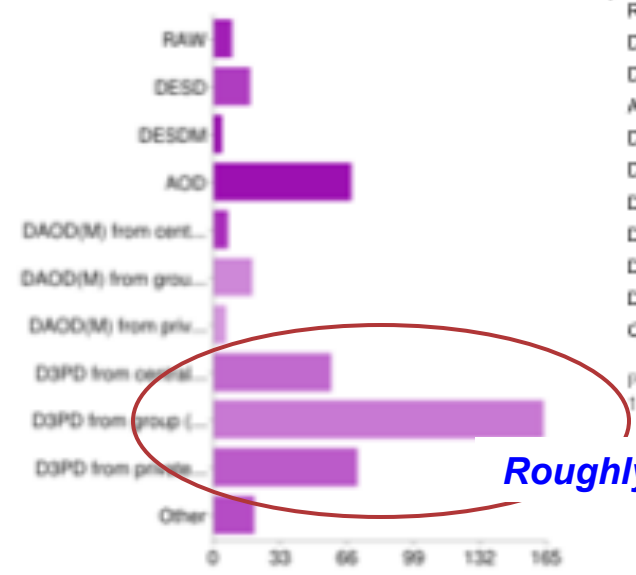
- September PAT Workshop Agenda

4. What is your primary computing resources for data analysis (i.e. where you login)?



Laptop/desktop	88	32%
Lxplus at CERN	43	16%
A Tier 1	6	2%
A Tier 2	13	5%
A Tier 3 or local cluster	83	30%
Cluster at a lab	41	15%

13. What data formats do you use for your physics analysis (multiple answers possible)? For each one, please state where you get them from



RAW	9	4%
DESD	18	7%
DESDM	4	2%
AOD	68	28%
DAOD(M) from central production	7	3%
DAOD(M) from group (e.g., SM, Top, Higgs,..) production	19	8%
DAOD(M) from private production	6	2%
D3PD from central production	58	24%
D3PD from group (e.g., SM, Top, Higgs,..) production	163	67%
D3PD from private production	71	29%
Other	20	8%

People may select more than one checkbox, so percentages may add up to more than 100%.

Roughly 2/3 of analysis is done on centrally produced ntuples.

POOL Optimizations

- Much of the non-D3PD analysis is done on POOL data products (DESD, AOD, DAOD, …)
- Optimization of this area is already underway (Argonne, Orsay, …).
- For Example:
  - By looking at the details of how data is retrieved and used, the splitting levels of the stored data were optimized to give significant performance increases.
  - The amount of resources used by each process was also reduced.
- Close collaboration with ROOT team for feedback and requests.

Results:		
	Read all events	Read 1% of events
AOD 1 split, 30MB TTree	55 ms/ev.	270 ms/ev.
AOD 2 un-split, flush 10 evt.	35 ms/ev.	60 ms/ev.
Difference	**~30 % faster read**	**4 – 5 times faster read**
Memory	**Reduced by 50 – 100 MB**	

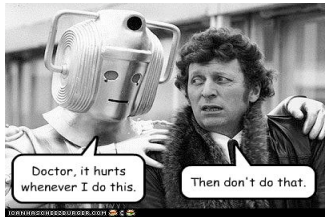
Organizational Meeting

- Analysis Performance Priorities Meeting held at Argonne in mid September.
  - Rik Yoshida, Doug Benjamin, Nils Krumnack, Jack Cranshaw, Peter van Gemmeren, David Malon
- Broad Mandate: “Optimize the time from data taken to physics results presented.”
- Specific Mandate: Address broad mandate by looking at improvements in I/O.
- Team Building
  - Build on expertise and previous efforts for POOL.
  - Leverage the resources available at Argonne (Tier3, local Tier2).
  - Leverage connections to other parts of the project.
    - *David Malon: Event Store*
    - *Rik Yoshida: Analysis Support*
    - *Doug Benjamin: Tier 3 Technical Coordination*
    - *Peter van Gemmeren: ROOT Integration*
  - *Developers*: Jack Cranshaw (coordinator), Nils Krumnack, Doug Benjamin, Peter van Gemmeren.

Current Work Plan: Status

- Establish connections to stakeholders
  - Physics Analysis Tools (workshop, week of 9/26)
  - ROOT team (met with Philippe, Fons, and Rene on 10/3).
- Survey of current usage patterns
  - Start collecting examples
    - *Shuwei's Top TSelector example*
    - *…*
  - Plan to make these into standard tests once instrumented.
- Instrument Code
  - Look at ROOT provided tools like *TTreePerfStats*.
  - Working on understanding what more is needed.
- Look at lessons from previous optimization
  - *TTreeCache* is currently not used much for analysis.
    - *TTreeCache makes reading more efficient by learning which branches are used most frequently and caching them.*

Current Work Plan: Next Steps



- Work with PAT to make sure documentation on how to write performant ntuple analyses is available and develop tools.
  - The D3PDReader project gives us a place to both implement improvements and install instrumentation.
- Using examples and relevant variables start compiling performance measurements.
- Start using TTreeCache in analysis examples.
  - Understand issues when used with PROOF.
  - Training of cache for random access mode of analysis may be different from that of *read all* mode seen in reconstruction. The learning mechanism may need to be customized.
  - Compare TTreeCache performance on LAN and WAN.
- Look at new ROOT features being prepared for release.
  - Pre-fetching in a background thread.
  - Disk resident TTreeCache.
- Provide Philippe/Rene with examples of D3PD and codes for comment.
- Begin looking at improving efficiency of disk/memory usage.
- Use locally available resources as sandbox, then expand into larger scale as the project matures.

Analysis Model Evolution 2012

- The LHC has passed into the phase of luminosity ramp up in its current configuration.
  - Initial threshold discovery opportunities have mostly finished.
  - Pileup in the detectors is already large and increasing, so event data is getting more complex.
- Data volumes are growing and are limited only by the acceptable trigger rate/Tier0 storage.
- Expected developments
  - Analysis will still be most efficient/fun when done ‘interactively’ on local resources.
    - *These users will continue in this mode as long as they can, but will have to become more and more selective as data accumulates. (see DPD train discussion, PAT Wkshp).*
  - As more advanced analysis techniques are developed, analyses done on D(3+)PD will naturally get more complex.

Summary

- An effort has been started to look at optimizing post-athena data access.
- This effort naturally focuses on D3PD access as most of the analysis user community uses this format.
- This effort builds on previous successful efforts with POOL data.
- We are working with the PAT team to develop tools.
- Other efforts in this area are being monitored and we will work with them when appropriate.
  - For example Sebastien's work with benchmarking in mana and SFrame.
- This effort is currently using ANL resources as a test bed, but eventually these developments will be deployed more broadly.
  - Any feedback from facilities on concerns or interesting measurements is welcome.