Argonne NATIONAL LABORATORY

I/O Infrastructure Support and Development

David Malon
malon@anl.gov

ATLAS Software Technical Interchange Meeting
9 November 2015

U.S. DEPARTMENT OF ENERGY

Introduction

- Lots of ongoing I/O-related activities, in support of Run 2 data-taking, in performance tuning and related work, in distributed event service development, …
- Peter and Jack will provide details about many of these things
- I will keep this introduction brief (and it has not been so very long since the last Software and Computing Workshop), providing just a bit of background, and mentioning one or two of the many things that do not warrant standalone presentations

Just a few words on I/O performance

- Ongoing I/O performance work encompasses a range of development and tuning activities, the latter based upon data (monitoring, explicit performance measurements, …)
  - E.g., development work on how to reduce the overhead of reading an event, how to support efficient event selection/rejection, …
  - And tuning of split levels, flush settings, and so on, to balance a number of considerations (I/O speed, memory and storage footprints, …) for a range of use cases
- Peter will cover some of these items in greater detail
- Rather than risk saying things that Peter will tell you about (and better than I), I will use this opportunity to advertise the ATLAS distributed I/O performance working group
  - Intended to be a cross-domain forum, with core software, analysis framework and tools, distributed computing, and site deployment experts all represented

I/O performance working group

- Sample of recent issues
  - Monitoring of analysis access patterns (not just which data types, but which data)
  - Robustness of error report propagation through our many software layers
  - What is really happening in I/O under the covers (numbers and sizes of reads as seen by the storage layer, …)
  - Expected and unexpected costs of class-versus-branch access modes
  - Xrootd issues arising in AthenaMP
  - Affects of current I/O and cache settings on performance in HPC clusters
- "External" experts join us as needed
  - Andy from the xrootd team, Philippe from the ROOT team, multiple CERN IT people
- Will not say more about these here, but if these topics interest you, please join
  - The group has value as a cross-domain discussion forum ("do the core experts understand why we are seeing this behavior in our analysis?"), but it is more valuable if effort can be brought to it
    - Otherwise items end up on the "to do someday as time permits" list
  - One authorship-qualifying project was completed based upon work for this group; more would not be so hard to define
- Fortnightly meetings on Tuesdays at 17:00 CERN time

Metadata and metadata infrastructure

- Emerging as an area that requires substantially more attention, and soon
  - where soon==now
- Already facing metadata issues in current data-taking
  - Jack will talk about some of these
- And while some of these are type-specific and/or specific to details of our current implementations, some point to issues we will face more generally
- A future frameworks issue, certainly, but metadata propagation models are already being challenged in AthenaMP and in distributed event service work
  - And in propagation from Athena-based processing to non-Athena-based analysis

Metadata and metadata infrastructure

- Current Athena metadata propagation infrastructure is incident-driven
  - For good reasons, historically—file boundaries are and should be asynchronous to Gaudi/Athena state transitions
  - And to the extent that we can treat file boundaries, not as fundamental (certainly not to physics processing), but as artifacts of storage, we should
- The use of incidents is the most frequently cited example of the need to rethink this infrastructure, but it is not the only consideration
- We should use the opportunity to rethink what is in in-file metadata as a matter of convenience versus necessity, how we augment content and how we find it, what the alternatives might be, which components need to see this content when and why, and how various kinds of bookkeeping may be done robustly in an environment in which it is possible that no single component in a single job may see "all" input (or output) events
- And of course the strategy must be consistent with strategies we are formulating to deal with conditions and other time-varying data
  - Partly but not only because IOV data constitute a big part of in-file metadata today
- Expect that metadata infrastructure will be a focus of more than one technical discussion in more than one context this week, e.g., both
  - Addressing current lumi block accounting problems
  - Planning infrastructure evolution for future frameworks and emerging processing models

Another advertisement: metadata validation

- (Restricting attention to validation of *in-file metadata* for today, but not forever)
- In-file metadata validation is often not part of technical or physics validation, or of checks done by developers as part of the tag approval process
  - And it shows
- We need to change this
- And metadata validation issues arise in additional contexts that do not or should not require event data validation
  - For example, in merging and splitting (e.g., AthenaMP) stages in ATLAS workflows
- Currently in the process of defining more precisely what is meant by in-file metadata validation, and formulating specific tasks (authorship-qualifying and/or for OTP credit)
- See DM's presentation at https://indico.cern.ch/event/455519/ for a handful of task descriptions as starting points

I/O in support of framework prototyping

- July framework workshop (Peter and Marcin present) was the occasion of the last substantive update
- Principal July objective: allow noninterfering reading and writing in prototype ✓
  - Separate POOL/APR persistency service instances for input and output
  - Separate converter instances for input and output
  - And so on
- This suffices to allow the non-I/O components of multithreaded prototyping and development to proceed, but it is not the end of the story
- Since then, Marcin has undertaken an inventory of I/O components, particularly those inherited from POOL, to understand where our thread safety challenges might lie
  - See subsequent slides
- Identified too a few issues that could affect robustness even of the current prototype's implementation

Near- and medium-term development plans in support of AthenaMT

1. Address the areas in which separated input and output services could in principle step on one another's toes in the current implementation
2. Ensure that I/O components are able in principle to support multiple output persistency services
3. In parallel:
   - Revisit "direct" ROOT conversion services (an area of development that has been dormant for a while) and how they might be used and useful in the future framework
   - Address streamlining and evolution of POOL/APR components for a future framework, with redesign as needed, beyond the minimum needed to support the prototype effort

- 1. and 2. should be achievable this calendar year (!?) in the absence of too many preemptive priorities
- 3. is work for the first half of 2016