Computing
for the CERN prototype test: 2015 Q3,Q4 and 2016 Q1 Milestones

Maxim Potekhin
BNL

CERN single-phase meeting
June 12, 2015

Philosophy:

- Computing systems developed for, and utilized to support the CERN prototype are not meant to be a computing prototype for the full DUNE experiment due to timing, effort profile and difference in scope.
- At the same time, it's an important stepping stone and a way to put in place software engineering practices which will benefit DUNE in years to come.
- An opportunity to evaluate existing solutions for use in DUNE.
- We expect minimal amount of manpower to be available during the workup to the test, so components and systems may have to be developed and deployed sequentially and in advance. For this reason, **data handling will be our focus in the next 9 months** since it's the foundation of the distributed architecture in any experiment.
- One motivation that will drive our system design is the ability to have "express streams" for data processing, in order to identify and correct potential problems quickly since the schduled beam time will be limited and the measurement program is rich, hence no second chances.

Sharing?

- Meaningful discussion with WA105 of possible sharing infrastructure components such as data handling is yet to happen: everyone was too busy, we are still working up to the CDR review and things in general have not settled (e.g. there is no official S&C organization in DUNE at the moment).

Work Areas

Data handling:

- Capturing and replicating raw data from the detector is priority and we chose to decouple this part of data handling from more general offline data management to reduce risk and provide more flexibility in the development of data processing system. Prime copy of the data will be at CERN, with at least one full replica at FNAL and potentially two more (could be partial) – at BNL and LBL/NERSC.
- Do we envisage additional European sites (outside of CERN)? Need opinions.
- Production runs will take place at a handful of well managed facilities with FNAL being the hub, while subsequent analyses will be distributed to a larger number of smaller sites. For that stage, we will consider utilizing XRootD-based storage federation based on experience such as AAA in CMS and FAX in ATLAS.
- Need to have well-designed Metadata.

Computational power:

- The default option for computing power is utilizing FNAL facilities but given the scale (a few thousand cores considered optimal) need to explore solutions elsewhere (including tools that exist at FNAL to federate different classes of resources e.g. OSG, cloud etc).
- Software validation and distribution will need to remain focus areas.

Priority Items for the next few months

- Establish a reliable and empowered contact(s) at CERN, in order to coordinate resource allocation, interface to CERN infrastructure etc
- As a follow up to the previous item, we need to understand who and how provisions various hardware and software components (cf. database servers as an example)
- Continue to narrow down storage requirements (currently "O(PB)" in the proposal) in order to inform our FNAL (and other) collaborators so proper planning of resources can start in a timely fashion - lead time for securing dedicated PB-scale storage is not negligible!
- Continue to evaluate existing data handling systems with a view to reuse those, e.g.
  a) LHC experiments - Rucio, Phedex etc
  b) FNAL: SAM/IFDH
  c) Daya Bay. IceCube: Spade (see backup slides)
- The choice need to be made on a ~1 month scale so we can go forward. In addition to reliability and performance, the criteria include the level of infrastructure and manpower required to deploy and operate the system, and the scope and volume of additional software development that may be needed for adoption.
- Have to select a Metadata system (SAM is the likely candidate).
- Prepare plans and pilots for processing.
- Software infrastructure

A possible timeline and effort profile

- NB: To be revisited after the DUNE S&C Organization is formed – what's included is the absolute minimum that must be accomplished. This is real effort and it needs real support.
- 2015 Q3
  - Coordination with CERN and US sites 0.2FTE
  - Evaluation of data handling systems, tech downselect 0.2FTE
  - Specification and prototyping of the "sandbox" for the pilot deployment 0.2FTE
  - Software provisioning - plans and prototyping (for reco) 0.1FTE
  - Storage federation (xrootd) R&D and reuse 0.1FTE
- 2015 Q4
  - Design of Metadata v1.0 0.2FTE
  - Selection and configuration of the Metadata Sytem 0.2FTE
  - Initial pilot deployment (CERN+endpoints elsewhere) 0.1FTE
  - Integration testing (including Metadata) 0.2FTE
  - Storage federation prototyping 0.1FTE
- 2016 Q1
  - Full pilot deployment of the data replication system 0.1FTE
  - "Data Challenge" - mock data distribution at scale - CERN to the US 0.2FTE
  - Plans and prototypes for production system 0.3FTE
  - Storage federation 0.2FTE

Summary

- These are initial thoughts for the time horizon of about 9 months.
- Issues such as calibrations (from the computing perspective) are left for later when they are better understood.
- By the end of this time, we won't yet have an end-to-end system for data processing (nor do we aim to) - the focus will be on the data handling and replication from CERN to other sites.
- The goal is to create conditions for further development of the complete reconstruction/production chain – to be designed later in 2016.

Backup Slides

Data Transport

Evaluating the LHC experience:

- ATLAS - currently deployed data movement machinery is experiment-specific and various couplings exist that will make reuse difficult (metadata design reflects specific workflow patterns in ATLAS and is split between at least two different databases).
- CMS - being investigated, awaiting feedback from CMS contacts, initial impressions are roughly similar to the above.
- Both would be impossible to use without a massive rewrite if we were to utilize a different metadata system (e.g. SAM).
- Substantial “external” expertise needed to deploy and operate either.

What else?

- There is a proposal to utilize Spade (IceCube, Daya Bay) - current and recent “in-house” expertise available, deployment is described as light-weight and essentially quick. Plug-ins can be utilized to interface an external metadata system. Well suited to the domain (i.e. is used to transport raw data from DAQ to remote storage). Choice of transport layer. Interaction with DAQ is well understood. Monitoring is available.
- SAM+IFDH: a choice of language binding, flexibility with protocols, "SAM batteries" included.

Dayabay SPADE Topology

SPADE developed in IceCube, used in Daya Bay & ALS

Data movement orchestration

Data catalog and warehouse with Python APIs

Transfer protocol agnostic: scp, gridftp, bbcp, Globus Online, (RDMA)