RFC: Data Flow From The MICE Experiment

J.J. Nebrensky, (Brunel University, Uxbridge UB8 3PH, UK),

MICE-NOTE-COMP-252

This document sketches out the flow of data from the MICE experiment, as I currently understand it. This includes not only illustrating the structure of the data flow, but also setting out a consistent vocabulary with which to describe it. Many aspects of this data flow (fig. 1) are either misunderstood by me, currently undecided, not yet implemented, or simply have never been considered before; so feedback is both welcomed and *essential*.

Background information about job submission and file storage on the Grid can be found in previous MICE Notes ([1], [2]) and the references therein. In particular the first two sections of Note 247 [2] are meant to provide a gentle introduction to Grid data storage from the MICE perspective, and timid MICE may wish to read those first.

1 Online Reconstruction

Raw data from the DAQ readout is stored in the online buffer (a Tier 0 for MICE), and used by the online farm in the MICE LCR for online reconstruction (fig. 1). The results (fitted tracks, etc.) are stored as ROOT files also in the online buffer and are available to the shifters to monitor the experiment. The online reconstruction output is not expected to be kept permanently but will simply be discarded after an appropriate interval. The online buffer is currently expected to be able to hold about 5 days worth of RAW data produced at a rate of just under 30 MB/s (see Appendix 1.)

2 Storage to Tape

The raw data recorded by MICE is to be archived on tape within the CASTOR system at RAL (currently in the Atlas centre, hopefully moving to R89 this summer). An optical link – shown in rose in fig. 1 – with 1 Gbps bandwidth is to be installed between the MLCR and CASTOR to provide a pathway for this data. The RAL CASTOR is thus a "Tier 1" resource within the MICE workflow much as it is a "Tier 1" for LHC data.

The CASTOR tape subsystem prefers files to be about 1 GB in size, with a minimum size for normal use of 200 MB.

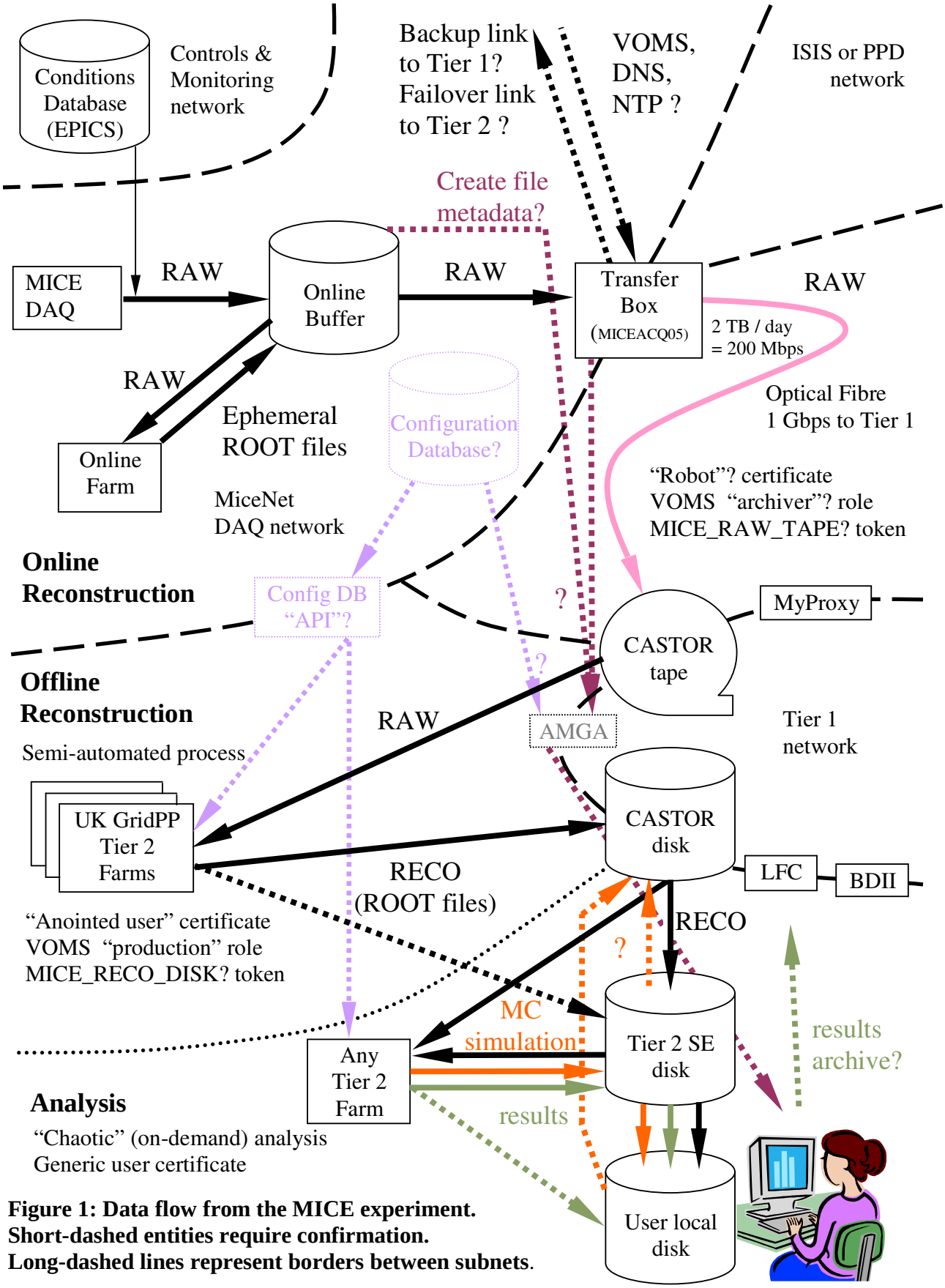

Figure 1: Data flow from the MICE experiment. Short-dashed entities require confirmation. Long-dashed lines represent borders between subnets.

The Tier 1 team have indicated that they would prefer us to use "Grid" authentication and protocols if possible; this implies using the SRM protocol to trigger the transfers[1]. Using Grid protocols from the beginning should also help us ensure that all results are properly and consistently registered in the LFC, metadata catalogue, etc. SRM would thus be hidden behind higher-level *lcg-** commands [2].

The proposed implementation is a software agent (known here as the "data mover") installed on a machine referred to here as the "transfer box" and located in the MLCR rack room. The transfer box will be dual-homed with network interfaces on both the DAQ network and the optical uplink. The data mover will read in RAW data from the online buffer, gzip[2] it for storage?, and then upload the files to CASTOR and register them in LFC, e.g. using the *lcg-cr* command. On CASTOR they may be held on an interim disk pool before actually being written to tape, so the data mover will check periodically to make sure the data has been correctly transferred and is actually on tape, after which the data mover can delete it from the online buffer[3]. The data mover should store at least the LFNs and SURLs of the uploaded data locally, so that the data can still be found on tape in case of a disastrous failure of the LFC.

Grid data transfers require authentication by X.509 certificate proxies. As the data mover will be an autonomous system potentially running 24/7, the correct route is to use a "robot" certificate, rather than one belonging to a specific user. (Except that no-one in the UK has used robot certificates yet, so we will probably start off with a couple of humans also registered, just in case[4].) Those entities authorised to carry out these transfers will be added to the "archiver" role in VOMS (anybody got a better name?) which will have the privilege of writing into the MICE_RAW_TAPE space token[5] on CASTOR and thus be the only way[6] to write (and thus (1) delete from and (2) use up MICE quota) on tape.

The RAW data is currently the *only* data flow we are expecting to archive on tape in CASTOR (?).

In order to function properly the transfer box will need access to other Tier 1 services (LFC, BDII, etc.) as well as auxiliary services that may not necessarily be provided by the Tier 1 (i.e. routed via the fibre): DNS, NTP[7], access to OS & middleware repositories for updates, and access to the VOMS server (Manchester, UK). Regarding disaster planning, it would also be useful to have an alternative route from the transfer box to the Tier 1 (in case the optical fibre is damaged) and a possible route from the transfer box to

[1] Using `rfio` authenticated against RAL CSF accounts is the fallback non-Grid route.
[2] gzipping is not expected to significantly compress the data, but it would package the RAW data into files that incorporate checksum information making it easier to detect data corruption.
[3] e.g. call srmls and check that the file locality includes NEARLINE.
[4] There is also the wrinkle that robot certificates used in portals require the private key to be stored on a secure hardware token ("at least FIPS 140-1/2 level 2,"), though as the data mover isn't a portal there may be other ways to address the relevant security concerns.
[5] A "space token" is a named allocation of storage, with a defined size, access control, curation policy etc.
[6] Actually CASTOR won't support VOMS until end 2009, so in the interim the DNs will also have to be added to CASTOR by hand. But we shouldn't need more than three entries at most, preventing confusion.
[7] For some reason the cryptographic authentication used on the Grid requires client and server clocks to be synchronised to better than one second. NTP is thus overkill, but common and convenient.

the Grid in general so that data can be copied to other sites if the Tier 1 itself has problems (need to work out how to handle loss of the LFC though). (These connections should not be assumed to be able to provide the full bandwidth of the optical link.) There will also be a need to SSH into the transfer box from within RAL during development and testing. The visitors' network can provide access to the auxiliary services, but the firewall has been found to block Grid data transfers. There is still discussion ongoing regards the connectivity that will be provided by the DAQ network; if it cannot satisfy the transfer box requirements then it would be necessary to negotiate adding another network interface connecting to the ISIS or PPD networks.

It would be useful to be able to backup Controls and Monitoring data (the "EPICS Channel Archiver") to tape. Connecting the transfer box directly would compromise the Controls and Monitoring Network, so if there is any such requirement a suitable mechanism for data transfer (RS-232, USB stick) needs to be decided before a suitable data mover can be written. In the meantime, it is believed that the data volume is sufficiently low (< 10 GB per month) that e.g. an external USB hard disk would suffice. These backups would be done by hand monthly, rather than as a continuous data flow. As they will involve files significantly larger than those preferred by CASTOR and will need to be done by a number of people, it may be wise to create a separate "archivist" VOMS role and corresponding space token[1].

3 Offline Reconstruction

Offline reconstruction will extract information such as particle tracks from the RAW data into ROOT files (referred to as "RECO"). This will be a semi-automated process, i.e. one triggered by the Production Manager when the need arises. The computation will be carried out via the Grid; as there will be insufficient resources available on the compute farm at the Tier 1, GridPP [3] have instead undertaken to provide equivalent computing power spread across Tier 2 sites within the UK.

It is not yet clear where the RECO data will be stored. One possibility is to write it directly to CASTOR disk (rather than tape) so that it can readily be copied elsewhere (e.g. to the US), but this implies an enormous number of inbound transfers to the Tier 1 during the reconstruction process. A better approach will be to simply save it to the farm's declared default storage element ("SE") for MICE, which will usually be local to the Tier 2 – this will spread the RECO data across a number of UK sites, but it can still be located transparently through the LFC. The CASTOR disk pools will still be needed as a fallback in case the local SEs have problems.

Storage for RECO should be protected by the MICE_RECO_DISK space token to only allow writing by the production manager. This will provide a dedicated space quota, and prevent accidental deletion by end users. Replication of the entire dataset to Tier 2's abroad could also be done under the aegis of the Production Manager.

[1] These could be shared with a number of other possible "occasional" use cases.

The individual anointed as Production Manager will be granted the existing "production" role in VOMS; this will allow them access both to any computing resources set specifically aside for offline reconstruction and to write to storage protected by the MICE_RECO_DISK space token. The process will thus use a grid proxy generated from their personal certificate.

4 Analysis

The final analysis process does not yet seem to have been defined; in particular there hasn't been any mention of any centralised, automated activity analogous to that underpinning the offline reconstruction.

Currently, the tools in place already allow analysis activity by independent individual users: the LFC allows users to either replicate files to the SE at any Tier 2 site supporting MICE [2] and analyse them through the Grid, or else to download them to local disk and work with them directly. The LCG tools also allow users to upload to the Grid and share their results via the LFC. These will write into the "generic" MICE storage (rather than that covered by a space token) and may require users to manage their usage of space on their "home" SE. Such on-demand activity ("chaotic analysis") will be authorised by vanilla proxies derived from certificates held by any member of the MICE VO.

The Grid resources available for user analysis will include at least some of those used for offline reconstruction (though with different privileges/restrictions), as well as other sites around the world – hence the use of a dotted line in fig. 1 to separate analysis and offline reconstruction processes.

As stated above, it is usually more efficient for output data from each Grid job to be stored at the site where it was run, rather than directly writing to a central location. As Grid jobs normally have a shell wrapper around the actual executable being used, it was proposed in [2] that individual users identify SEs close to or at their home institute that they trust to act as a fallback, should those at remote sites fail. The name of the chosen SE can be accessed in (preferably standardised) wrapper scripts as the ${MICE_HOME_SEID} variable, set e.g. using the "Environment" JDL attribute.

Will any of the results (e.g. those used in a particular publication) need to be archived somewhere central, or are individuals responsible for preserving their own data?

The current model also allows users to read the RAW data and do their own reconstruction. Are there any circumstances where read access to the RAW data would need to be restricted to the Production Manager? How often are users likely to want access to the RAW data (we don't want the CASTOR tape drives tied up repeatedly reading data – it should be mirrored on disk if it is to be regularly accessed)?

Some possible analysis data flows are included in fig. 1 in sage green.

5 Simulation

There will probably be some Monte Carlo simulation done at some point. The present mode of working is shown in fig. 1 in orange: data from simulations run on the Grid is stored on SEs at Tier 2 sites and later retrieved by the user. Will this data need to be stored in a central location such as the RAL Tier 1, given that other users can still access it from the Tier 2 SEs? Will it need to be archived *on tape*? Will there be someone like a "Simulation Production Manager" to oversee this, or will it all be ad hoc by end users?

6 Other

Fig. 1 also outlines two other flows of data – the configuration database (lavender) and file metadata (plum) – with which the primary experimental data must interface.

Both offline reconstruction and the analysis jobs will need to query the configuration DB. The mechanism for achieving this is still undecided, but probably a visible server holding a replica or snapshot of an appropriate view will be better than direct connection to the master server.

Analysis jobs should hopefully only require access to those RECO files specifically containing events of interest. It will therefore be necessary to provide a "metadata catalogue" that allows the user to identify a list of files relevant to a particular analysis; as yet neither the technology nor the required criteria have been identified [4].

Conclusions

Lots of stuff still needs to be settled, e.g.

- Is everyone happy with the terms *RAW* and *RECO*?
- What network services and connectivity will be needed?
- Any better names for *transfer box*, *data mover*, the *archiver* and *archivist* roles, etc.?
- Which data needs to be preserved on tape?
- Does it need replicating to tape at other Tier 1s?
- Are there any other data transfer or storage use cases that will require additional VOMS roles or space tokens?
- All data will be readable by anyone.

This note has identified several flavours of data within MICE: RAW, RECO, analysis outputs, and simulations. As can be seen from the complexity of figure 1, ensuring that they are all correctly preserved and made available will not be trivial. Although Grid tools provide us with some ready-made building blocks, it is still necessary to put them together in the right way to ensure the whole structure meets our requirements.

It is thus imperative that we agree and understand the basic attributes of the four data flavours listed above:

- volume (the total amount of data, the rate at which it will be produced, and the size of the individual files in which it will be stored)
- lifetime (ephemeral or longer lasting? will it need archiving to tape?)
- access control (who will create the data? who is allowed to see it? can it be modified or deleted, and if so who has those privileges?)

As it says above, please comment!

Appendix 1: Data Rate[5]

For each particle trigger (pt) without zero suppression we have:

TOF TDC: Maximum 108 hits, 4 Bytes per hit → Max 432 Bytes/pt

TOF fADC after firmware upgrade: 60 samples per channel → 13 kBytes/pt

KL fADC after fADC firmware upgrade: 60 samples per channel → 6 kBytes/pt

CKOV fADC: 300 samples per channel, 1 Byte per sample → 2.4 kBytes/pt

Tracker: 5536 Bytes per tracker/pt → 10.8 kBytes/pt

TOTAL: ~33 kBytes/pt

There will be about 500 particle triggers per spill, and one spill per second, implying a data rate of about 16.5 MB/s.

Electron Muon Ranger (coming up after spring 2010):

TDC: about 3000 channels, 2 Bytes/ch → 6 kBytes /pt

fADC: about 50 channels, 300 samples/ch, 1 Byte/sample → 15 kBytes /pt

TOTAL for EMR: 21 kBytes /pt → 10.5 MB/s

All these figures are without zero suppression - they are real upper limits.

Note that the fADC firmware upgrade will happen before data taking starts, so the larger pre-upgrade data rates are not relevant to this Note.

Acknowledgements

Thanks to everyone that takes the trouble to reply.

References

1. D. Forrest: "*The Grid & MICE*" MICE Note 246 (2009)

2. J.J. Nebrensky: "*Draft Grid Storage Namespace Guidelines*" MICE Note 247 (2009)

3. The GridPP collaboration: "GridPP: Development of the UK Computing Grid for Particle Physics" *Journal of Physics G: Nuclear and Particle Physics* **32**, pp. N1-N20 (2006) or see http://www.gridpp.ac.uk/

4. H. Nebrensky: "*Grid Update*" MICE Collaboration Meeting (CM23), January 2009

5. Jean-Sebastien Graulich: Private Communication, April 2009