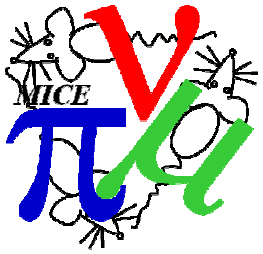
Notes from Data Flow Workshop

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



MICE-MIN-COMP-255

This document summarises the discussions at the MICE Data Flow Workshop held at Brunel University on 30th June 2009.

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. The proposed data flow is described in MICE Note 252 [3].

1 Metadata

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].

For many applications – such as analysis – you will want to identify the list of files containing the data that matches some parameters. This is done by a "metadata catalogue". For MICE this doesn't yet exist.

A metadata catalogue can in principle return either the GUID or an LFN – it shouldn't matter which as long as it's properly integrated with the other Grid services.

We have several databases (configuration DB, EPICS, e-Logbook) where we should be able to find all sorts of information about a run/timestamp, but how do we know which runs to be interested in, for our analysis? We need an "index" to the MICE data, and for this we need to define the set of "index terms" that will be used to search for relevant datasets.

Thus, if I wanted to analyse some data, I might search for all events with a particular:

- Run, date/time
- Step
- Beam – e^-, π, p, μ (back or forward)
- Nominal 4-d / transverse normalised emittance
- Diffuser setting
- Nominal momentum – beam momentum, MICE cooled? momentum
- Configuration:
    - Magnet currents (nominal)
    - Physical geometry (The MC should hold a pointer to a geometry in the config DB, not all the info itself)
- Absorber material, which units are filled
- RF phase and voltage (for each cavity)
- MC Truth, MCsimulated configuration

Anything else?

I'm not sure what the conclusion was – the list was generally agreed but there was opposition to creating yet another copy of existing information in yet another database, but no concrete proposals for an alternative. Note that you can't get the info from the data stream (as you don't yet have the file), it's not in the configuration database, and the EPICS Archiver can't be queried from off-site. A read-only replica of selected data seems to be most of the way to a metadata catalog from a different direction.

2 Management

A "role" is a combination of duties and privileges, with a specific aim. These are distinct from those of the person fulfilling that role. The Operations Manager ("MOM") is an example of a continuous role, enacted by different people over time. Some roles may be so specialised that only a particular person can do them; others can have many people in them at the same time.

=> Don't equate roles with FTEs!

For the data flow, the privileges (such as the ability to write to tape) associated with roles are enforced by VOMS. They may also require space tokens to be set up (on a site-by-site basis).

Roles identified prior to the workshop:

- Online reconstruction manager (needs to liaise with DAQ and archiver over online buffer access patterns)

- Archiver ("mvr") (storage of RAW data to tape; eventually to be a robot)

- Offline reconstruction manager ("???") (submits jobs, collates results, partly automatable)

- Simulation Production Manager ("production") (submits jobs?, collates results?, ensures good data is preserved, partly automatable) I'm not sure if this is to be a big centralised system that actually runs the simulations, or just someone who co-ordinates individual users running their own.

- Analysis manager?

- Data Manager ("gdm") (moving data around Tier 2s, LFC consistency, point of contact)
  The offline reconstruction, simulation production etc. managers will be responsible for actually transferring their data to a sensible location. The Data Manager is more of a human, supervisory role – negotiating quotas at sites, monitoring usage, raising LFC issues with its sysadmins and acting as a public point of contact, e.g. if a site needs to take storage off-line for maintenance will any data need replicating elsewhere for the duration?

- Archivist ("archivist") (storage of miscellaneous data to tape – an occasional, human role)

- VO manager ("VO-Admin")(assigns people to roles in VOMS)

- Operations Manager (MOM)

- Shifter – possible need for EPICS access? Otherwise they're just an ordinary user at a specific location.

Unfortunately I missed the discussion about this at the workshop (no electricians to be seen for years, then **four** turn up at once!) – I gather the above were accepted. We also need to add to the list:

- Software manager ("lcgadmin") (ensures working G4MICE available at grid nodes)

Next steps

- Confirm names to be used for roles (probably no spaces) and get them created in VOMS

- Decide which humans should fulfil each role

3 Data Volumes

Several flavours of data have been identified [3] 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?)
- service level (availability, resilience against short-term outages, bandwidth)

These attributes will then be represented by the SRM space tokens [5] associated with the storage that Grid sites make available to MICE. See also [6] regarding some issues with tape spaces and robots.

3.1 For RAW data:

volume: the total amount of data: 27 TB, rate at which it will be produced: 30 MB/s, and the size of the individual files in which it will be stored: 1-2 GB

lifetime: permanent. will it need archiving to tape: yes. Replication: yes, there should always be an off-site copy but only on disk

access control: who will create the data: archiver. Who is allowed to see it: all. Can it be modified or deleted: no

"service level": desired availability: write 24/7 if ISIS up, else none. Allowable outage: 48 hrs

(The 27 TB is derived from the CM24 500 million events figure. That implies that all MICE steps add up to less than a fortnight's data taking – is that right?)

3.2 For RECO data:

volume: the total amount of data: ???, the rate at which it will be produced: ???MB/s, and the size of the individual files in which it will be stored: ??? GB

lifetime: most longer lasting, publications data sets permanent. will it need archiving to tape: only the publications. Replication: multiple disk

access control: who will create the data: offline reco manager. Who is allowed to see it: all. Can it be modified – no – or deleted: yes and if so who has those privileges: offline reco manager

“service level”: desired availability: write ??? if ISIS up. Allowable outage: ??? hrs

(I’ve seen a claim of 6 TB for RECO data somewhere)

(Could be bigger than RAW – Overall about same as includes sim reco)

(Not important to be on Tier 1)

3.3 For analysis output:

volume: the total amount of data: tiny, the rate at which it will be produced: tinyMB/s, and the size of the individual files in which it will be stored: ??? GB

lifetime: most ephemeral, publication permanent. will it need archiving to tape: publication will. Replication no

access control: who will create the data: archivist, user. Who is allowed to see it: all. Can it be modified or deleted: no(pubs) yes(rest) and if so who has those privileges: creator

“service level”: desired availability: write ??? if ISIS up. Allowable outage: ??? hrs Uses LFC+VOMS – need sufficient notice of maintenance

Don’t need to talk to DB during analysis process.

Vast majority will be ephemeral created by users on local disk or Tier 2.

3.4 For simulation-RAW:

volume: the total amount of data: 10xRAW, the rate at which it will be produced: 30 MB/s, and the size of the individual files in which it will be stored: 1-2 GB

lifetime: keep some, ditch crap. will it need archiving to tape: no. Replication: Tier2 disks

access control: who will create the data: SPM, beamline grp. Who is allowed to see it: all. Can it be modified – no – or deleted: yes – delete crap and if so who has those privileges SPM

“service level”: desired availability: write ??? if ISIS up Allowable outage: ??? hrs Will all be at Tier 2. Uses LFC+VOMS – need sufficient notice of maintenance

(maybe extra x2 JC)

(Includes g4beamline stuff – group role for beamline group)

3.5 Other data to be kept (“archivist”):

volume: the total amount of data: 1 TB, the rate at which it will be produced: tinyMB/s, and the size of the individual files in which it will be stored: various

lifetime: permanent. will it need archiving to tape: yes. Replication no

access control: who will create the data: archivist. Who is allowed to see it: all. Can it be modified or deleted: no

"service level" – irrelevant as chaotic anyway

(I know about the Tracker QA data, KEK testbeam, Field maps)

Next steps

- Confirm space token names for various data flavours
- Meet with Tier 1 and/or GridPP and get space tokens/quotas created on relevant sites

4 File Catalogue Namespace

Also, we need to agree on a consistent namespace for the file catalogue

Proposal ([2], [4]):

We get given `/grid/mice/` by the server

Five upper-level directories:

`Construction/`
: historical data from detector development and QA

`Calibration/`
: needed during analysis (large datasets, c.f. DB)

`TestBeam/`
: test beam data

`MICE/Step`*`n`*`/Date`
: DAQ output and corresponding MC simulation

`MICE/Beam/`
: g4beamline beam simulations

If the archiver creates a new directory, how does it ensure that only it and MCmanager can create new file in it?

Reco goes near to RAW.

`/grid/mice/users/name`
: For people to use as scratch space for their own purposes, e.g. analysis

Encourage people to do this through LFC – helps avoid "dark data"

LFC allows Unix-style access permissions.

Next steps

- Create top level and have LFC admin write-protect it.
- Start uploading QA data…

5 Data Integrity

I've previously raised the question of whether we should compress data files before uploading in order that we can check their integrity down the line with e.g. with `gunzip -t …`

However, for recent SE releases a checksum is calculated automatically when a file is uploaded, using the lightweight Adler32 algorithm. This could be checked when the file is transferred between SEs, or the value retrieved to check local copies. The workshop felt that this removed any need for compression, though for critical files (e.g. RAW) we should also do the checksum ourselves locally to ensure that the initial upload is uncorrupted.

6 Data Archiver/Mover

T2K have agreed to allow us to modify their archiver ("QOS") for MICE:

http://www-pnp.physics.ox.ac.uk/~west/t2k/discussions/daq_archive_docs/Archiver_User_Manual.html

The work is to be done by a summer student at RAL.

7 Questions

How does one run become the next – what triggers it, who confirms, how is it propagated?

How does "data" come out of the DAQ and get turned into files?

How do we know a run is complete => that a file is closed?

Does the GB file size for CASTOR match the online reco sample rate? => Could the data mover trigger the online reco?

That ol' online buffer round-robin thang

Replication of data to other Tier 1s

Should EPICS monitor the data mover?

8 Actions

Work out desired CASTOR resources, interface (SRM?) and QoS. Meet with Tier 1 and iterate.

Draw up list of VOMS roles and get them created.

Draw up list of space tokens by tier and role.

Create LFC namespace, set permissions, upload existing data

Get archiver robot certificate

Identify needed Tier 2 resources

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[7]

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.

[We expect ~500x10^6 good muons through MICE]

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. J.J. Nebrensky: "*RFC: Data Flow From The MICE Experiment*" MICE Note 252 (2009)

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

5. A. Domenici and F. Donno: "Static and Dynamic Data Models for the Storage Resource Manager v2.2" *Journal of Grid Computing* **7** (1) pp. 115-133 (2009)
DOI: 10.1007/s10723-008-9110-3

6. With regard to RAL Castor tape storage see e.g.
http://www.gridpp.rl.ac.uk/blog/2009/06/10/step09-tape-drive-performance/ and
http://www.gridpp.rl.ac.uk/blog/2009/06/12/step09-tape-migration-stream-policies/

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

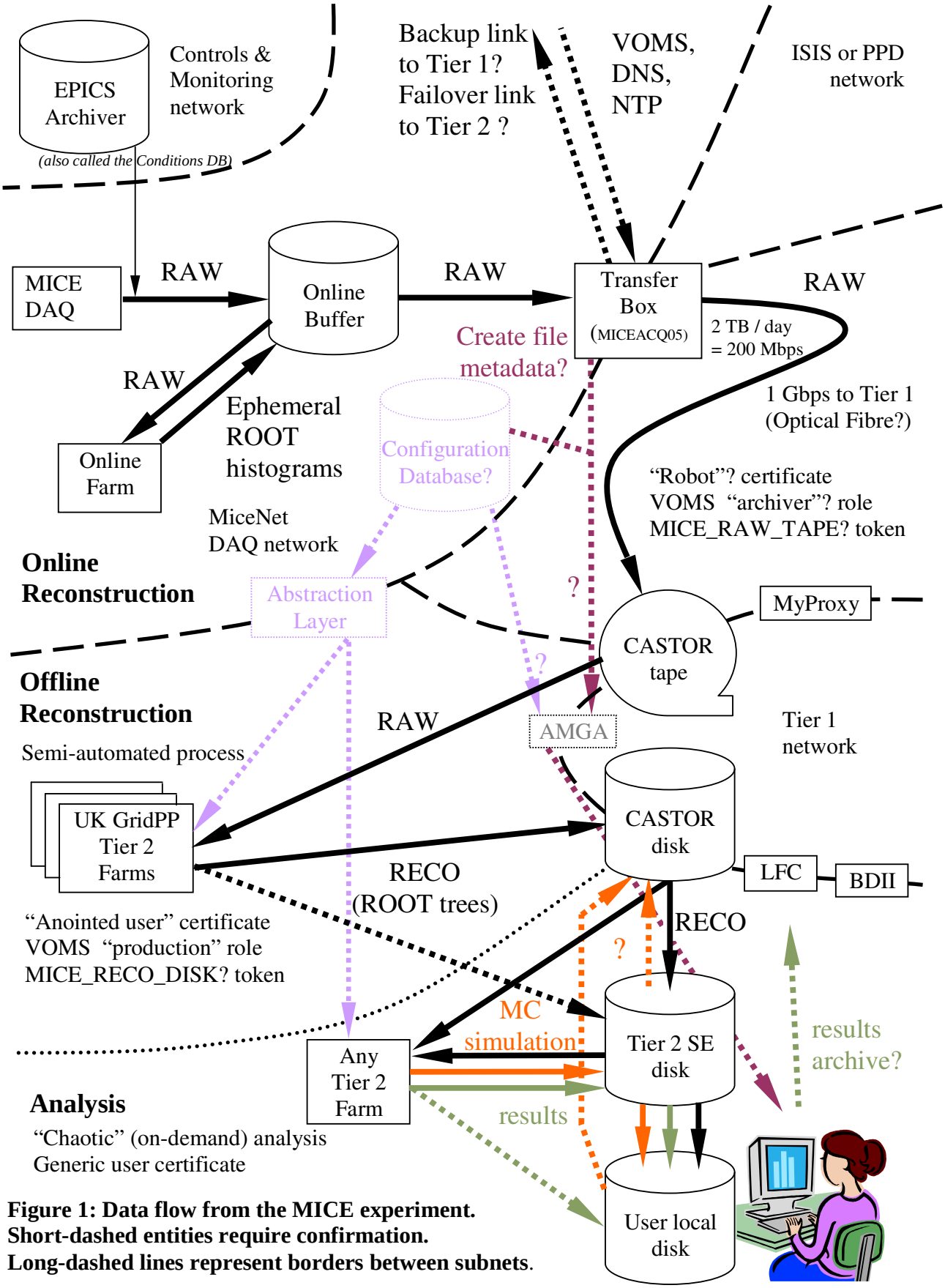


Figure 1: Data flow from the MICE experiment. Short-dashed entities require confirmation. Long-dashed lines represent borders between subnets.