



## Remote Institute Tasks

Frank Filthaut

11 February 2002

- Monte Carlo production
- Algorithm development
- Alignment, calibration
- Data analysis
- Data reconstruction



## Starting remarks

- **Charge:** try to see which tasks remote institutes “can do best”
  - ◆ Best for collaboration or for institute? (Try to address both)
  
- I’m assuming that the idea is that it “should not matter where data are stored”
  - ◆ Only alternative would be to replicate data in individual institutes
    - Seems very impractical to achieve this in general given exclusive streaming strategy
      - ⚙ Related: tape costs make it undesirable to replicate substantial datasets
    - Thumbnail format probably the exception
    - Book-keeping issues need to be considered (versioning)
  - ◆ I will not discuss the case of institutes unable to meet necessary technical requirements
    - Technicalities reduced to being able to log on to machines at centres that do meet those requirements



## Starting remarks

### □ Try to address two sorts of issues:

#### ◆ Technical

- (hardware) resources: CPU, storage, network
- (software) know-how (SAM, DØ code releases)

#### ◆ Sociological

- What does the collaboration as a whole need?
- How do the institutes operate?
  - ⚙ Are graduate students, post-docs, staff normally at the lab or in the home institute?
  - ⚙ Is it possible to come to FNAL regularly?
- To what extent does communication from the remote institutes work?
  - ⚙ News, e-mail, mailing lists adequate?
  - ⚙ (conference) phone calls, video ⇒ availability, **quality**
  - ⚙ Time differences

Apologies if all of this is too obvious



## Monte Carlo production

Technical requirements:

### □ CPUs

- ◆ Farm, either dedicated or desktop

### □ Sufficient storage space

- ◆ Tape robot (unless idea is to ship all output elsewhere)

### □ Network: reasonable bandwidth to storage medium (local?)

- ◆ If the collaboration as a whole is to profit from this, need to be able to transfer to outside world

### □ Software

- ◆ Stand-alone code releases in form of mini-tar files: easy
  - Use well-tested code versions
  - Should be (and is) centrally managed

### □ Man power

- ◆ 1-2 administrators (+ technical assistance)

Few non-technical issues



# Algorithm development

## Technical requirements:

### □ CPUs

- ◆ Few desktop nodes suffice

### □ Storage space

- ◆ Moderate (~4 GB / full release + data files for test purposes)
  - Large amount of data may be required for specific cases

### □ Software

- ◆ Need code releases

- Relatively isolated piece of code: use of “production” or “test” releases not very relevant

- ✿ Production release: higher chance that other code you need is actually working

- ✿ ... but not necessarily the most up-to-date

- ✿ When inserting code into CVS: have to deal with possible clashes in similar way

- ✿ Potential problems with (DØ-external) code availability under upd

- Otherwise: will have to resort to “test” releases

### □ Man power

- ◆ Some administration (code installation)

# Algorithm development



## Sociological aspects:

### □ For isolated efforts:

- ◆ Less urgent need for frequent consultation with others
- ◆ Can be done by single institute or group of institutes working closely together

### □ For efforts requiring feed-back using data or otherwise non-isolated:

- ◆ Communication issues become more important
  - Bound to be easier if inconveniences like time differences are not an issue
  - Scrutiny in choice of topic helpful



# Alignment, calibration

Technical requirements:

## □ CPUs

### ◆ (Small) farm

- Assume a reasonable turnaround time is desired: in general, likely to require reprocessing (from raw / reco level) significant amount of data

## □ Storage space

### ◆ Same “significant amount” of data – but what data?

- In general, specific streams of limited use

- ☼ “Only” convenient for electrons, muons, photons?

- Pick selected events from general stream

- ☼ Exceedingly tedious for reasonable event samples

- Hand-pick selected runs

- ☼ Most resource-friendly – but tedious and too limiting?

## □ Alternatively: SAM transfers

- ◆ requires high bandwidth network

## □ Software

- ◆ Need (most) recent code releases ⇒ “test” releases



## Alignment, calibration

Sociological aspects:

- This is guaranteed to need a lot of feedback concerning:
  - ◆ Detector status
    - Has to come from FNAL!
    - Database should eventually solve this
      - ✦ On what timescale?
  - ◆ Peculiarities of (recent) code releases
    - Level of documentation is not such that understanding is trivial



# Data analysis

## Technical requirements:

### □ CPUs

- ◆ Highly dependent on analysis (stream / trigger)
  - But in general at tuple / thumbnail level  $\Rightarrow$  manageable?

### □ Storage space

- ◆ Again analysis dependent
  - May involve significant amount of data except in Thumbnail format
    - ✿ Various ideas for copying and “locally” storing (significant fraction of) data
  - Similar issue for MC

### □ SAM transfers again the alternative

- ◆ requires high bandwidth network (analysis at tuple level even more I/O limited)

### □ Software

- ◆ Tuple level: ROOT sufficient
  - Correction & algorithm packages running on tuples?
- ◆ Thumbnail: need (recent?) DØ code release to read data



# Data analysis

## Sociological aspects:

### □ Requires

- ◆ Information on available / appropriate calibrations (e.g. jet energy scale)
- ◆ Information on appropriate data samples (good runs)
- ◆ Feedback from within (and outside) physics group
  - The obvious... it helps to choose topics for which there exists nearby expertise
    - ⚙ Good example: French SUSY efforts

### □ It **must be** possible to do an analysis “at home”

- ◆ Not all graduate students can stay at FNAL indefinitely
- ◆ For teaching staff even more difficult to leave home institute
- ◆ Financial constraints

### □ The required facilities are currently clearly not in place

# Data reconstruction



Technical requirements:

## □ CPUs

- ◆ Significant need (CPU resource scarcity at FNAL primary reason for doing this at remote institutes?)

## □ Storage space and network requirements

- ◆ Not clear whether requirements are stringent
  - Does FNAL stay primary storage, or is distributed storage a possibility?
  - Output best stored at a location with good network access ⇒ centres

## □ Software

- ◆ Need DØ code release to (re-)reconstruct data
  - “certified” but yet recent version
  - Should be centrally managed
- ◆ Will need database access
- ◆ Reliability (robustness against loss of data) must be addressed

There need not be many non-technical aspects!!



## Conclusion

- Data analysis: crucial but non-trivial
  - ◆ Emphasis largely on non-technical issues
- Data reconstruction: desirable?
  - ◆ Technical issues determine feasibility
- Alignment & calibration: possible but difficult
  - ◆ Technical hurdles; availability of pertinent information even more of an issue than for data analysis
- Algorithm development: doable
  - ◆ But be careful with choice of topic
- Monte Carlo production: easy
  - ◆ Well contained, collaboration can accommodate wide range of configurations
  
- It's a burden on people present at FNAL to ensure information flow so that collaboration as a whole can function efficiently
  - ◆ I sincerely hope that they will take up this challenge