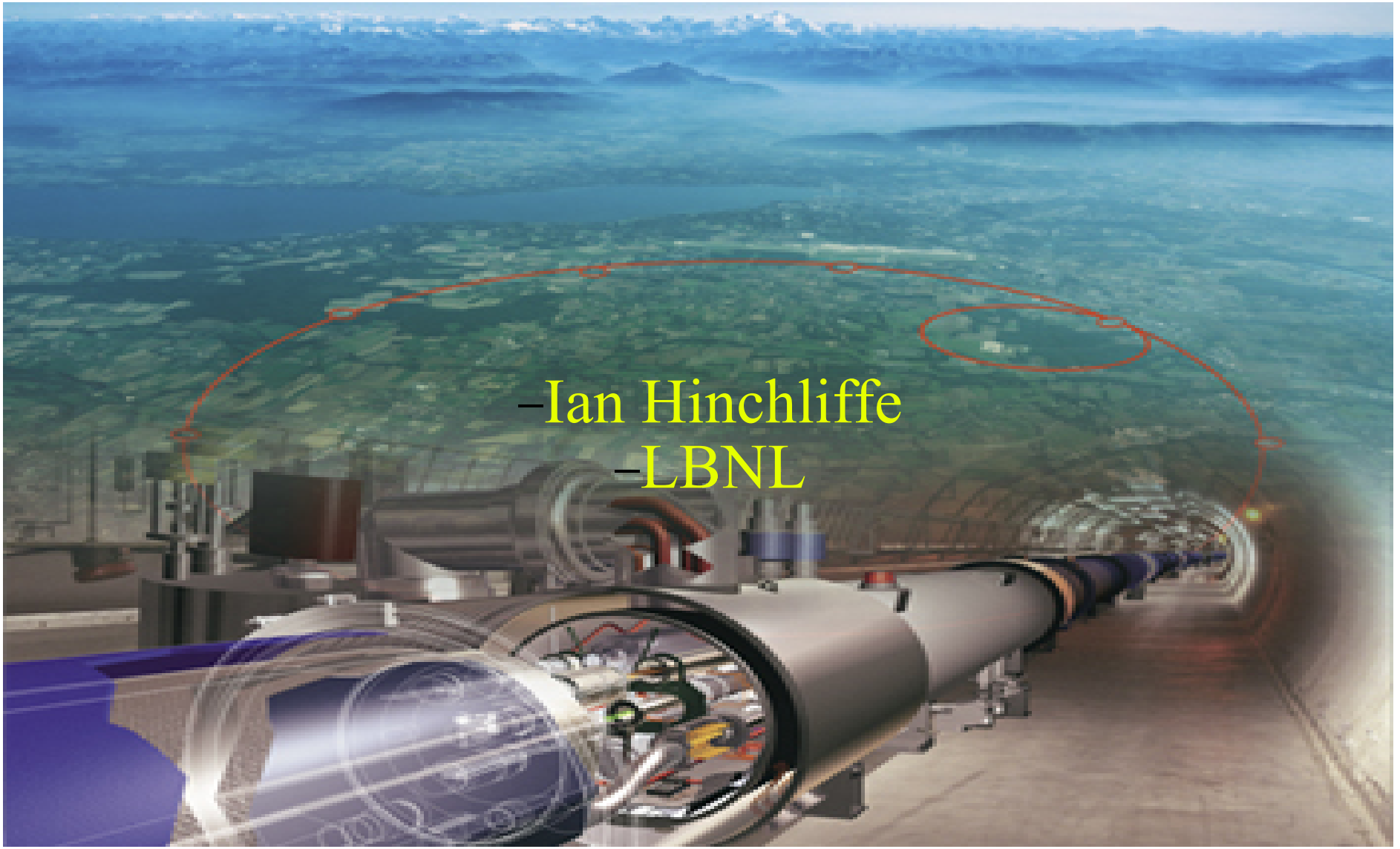
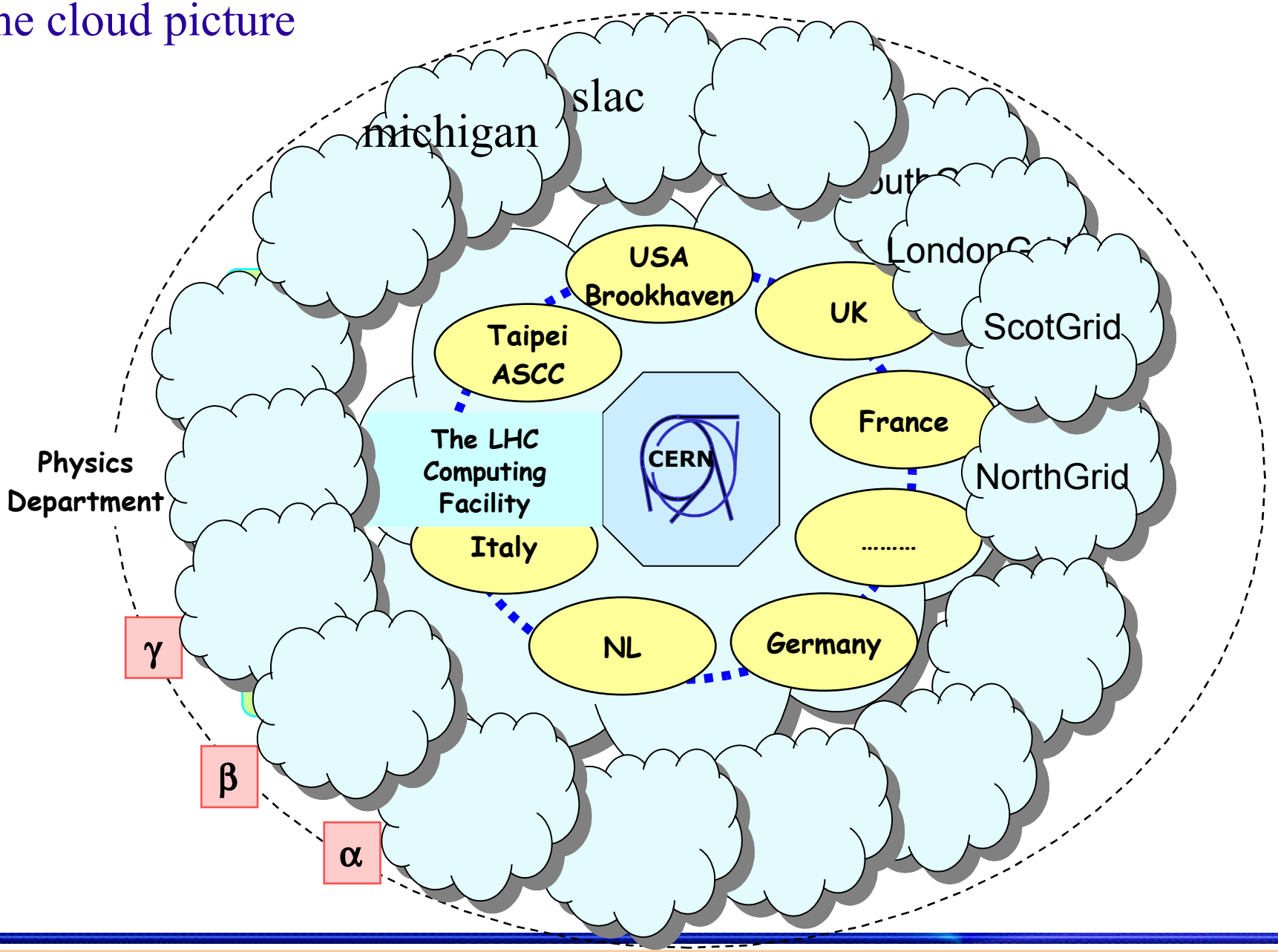
# Aspects of ATLAS computing model



# Outline

- Components
- Data flow
- Resources and requirements
- Data access patters
- Early data and evolution
- Apologies for the quality of this talk: we are in the middle of FDR2 data prep crisis: I have filched most of the slides from other atlas talks (mainly Roger Jones)

# The cloud picture



# Data Flow

- EF farm  T0
  - 320 MB/s continuous
- T0 Raw data  Mass Storage at CERN
- T0 Raw data  Tier 1 centers
- T0 ESD, AOD, TAG  Tier 1 centers
  - 2 copies of ESD distributed worldwide

Tier 0 view

- T1  T2
  - Some RAW/ESD, All AOD, All TAG
  - Some group derived datasets
- T2  T1
  - Simulated RAW, ESD, AOD, TAG
- T0  T2 Calibration processing?

Tier 2 view

# CERN

- Tier-0:
  - Prompt first pass processing on express/calibration & physics streams with old calibrations - calibration, monitoring
  - Calibrations tasks on prompt data
  - 24-48 hours later, process full physics data streams with reasonable calibrations
    → Implies large data movement from T0 →T1s
- CERN Analysis Facility
  - Access to ESD and RAW/calibration data on demand
  - Essential for early calibration
  - Detector optimization/algorithmic development
- Tier 3 for CERN users: access limited to CERN group
  -

# Data streams

- Data coming from SFO will be "streamed" by trigger
  - Streams are inclusive
    - Events enter more than one stream
    - More flexibility for subsequent reprocessing
    - More robust
    - But must be careful of disk wastage
    - Stream content will vary with luminosity, experience
      - Very early data may have less streams
      - Exercised in FDR (later):
      - Stream content will vary with luminosity, trigger configuration and detector performance
      - Of order 6 streams, Muon, E/gamma, Jet, Express, Min bias, Bphys
- Comments on express stream in FDR talk

# Outside CERN

- Tier-1:
  - Reprocess 1-2 months after arrival with better calibrations
  - Reprocess all resident RAW at year end with improved calibration and software
    → Implies large data movement from T1↔T1 and T1 → T2
- ~30 Tier 2 Centers distributed worldwide Monte Carlo Simulation, producing ESD, AOD, ESD, AOD  Tier 1 centers
  - On demand user physics analysis of shared datasets
  - Limited access to ESD and RAW data sets
  - Simulation (some at Tier 1s in early years)
    → Implies ESD, AOD, ESD, AOD  Tier 1 centers
- Tier 3 Centers distributed worldwide
  - Physics analysis
  - Data private and local - summary datasets

# Event sizes and processing rates

These are from TDR and were used as basis for resources:

Current situation is worse

**Table 2-2** The assumed event data sizes for various formats, the corresponding processing times and related operational parameters.

| Item | Unit | Value |
|---|---|---|
| Raw Data Size | MB | 1.6 |
| ESD Size | MB | 0.5 |
| AOD Size | kB | 100 |
| TAG Size | kB | 1 |
| Simulated Data Size | MB | 2.0 |
| Simulated ESD Size | MB | 0.5 |
| Time for Reconstruction (1 ev) | kSI2k-sec | 15 |
| Time for Simulation (1 ev) | kSI2k-sec | 100 |
| Time for Analysis (1 ev) | kSI2k-sec | 0.5 |
| Event rate after EF | Hz | 200 |
| Operation time | seconds/day | 50000 |
| Operation time | days/year | 200 |
| Operation time (2007) | days/year | 50 |
| Event statistics | events/day | $10^7$ |
| Event statistics (from 2008 onwards) | events/year | $2 \cdot 10^9$ |

# Actual performance today: real data

- Trigger runs at 200 Hz out of EventFilter
- Event sizes from FDR1, 13.0.40, (no truth, MC is bigger): note 10**31
  - ESD are approx 700KB(averaged over streams_
  - AOD are 160Kb
  - DPD are unknown at this stage, but some plans are worrying
  - ttbar events are about 30% larger
  - RAW 2.6 GB (known problem in RDO to BS converter)
  - AOD are reducing: some jet collections were dropped last week
  - New numbers from FDR2 next week
- Processing time exceeds budget by factor of few, code improving, may have to lower trigger rate depending on LHC machine performance.
- Memory usage is critiical.
- Heavy Ions are a special case: worry about it later

_

# Actual performance today: simulation

- Model assumed simulated production of 20% of real data volume using full G4.
- Current G4 and tunings is approximately 8 times slower than assumed
  - Defaults in release 14 are twice as slow as defaults in release 12
  - Strong arguments related to calorimeter calibration
  - Can probably cope this year due to late start: 10 TeV simulation is about to run (meeting today to discuss this)
  - Resources must come from somewhere: less user CPU at Tier 2, less simulation???
  - Role of fast simulation still unclear (will depend on experience with data)
    - Atlfast II is much faster (comparable to reco time): memory usage is an issue here
    - Parameterized G4 is about 2 times faster
  - Simulated events are twice budget if only RAW (HITS) are retained: but we currently keep RDO and sum is 5 times budget
- My comment: there is a serious long term issue here

# Resources

| | CPU (MSi2k) | | Disk (PB) | | Tape (PB) | |
|---|---|---|---|---|---|---|
| | 2008 | 2010 | 2008 | 2010 | 2008 | 2010 |
| Tier-0 | 3.7 | 6.1 | 0.15 | 0.5 | 2.4 | 11.4 |
| CERN Analysis Facility | 2.1 | 4.6 | 1.0 | 2.8 | 0.4 | 1.0 |
| Sum of Tier-1s | 18.1 | 50 | 10 | 40 | 7.7 | 28.7 |
| Sum of Tier-2s | 17.5 | 51.5 | 7.7 | 22.1 | | |
| Total | 41.4 | 112.2 | 18.9 | 65.4 | 10.5 | 41.1 |

- Some CPU may be memory limited: mainly an issue for reco
- Disk space likely to be critical: larger event sizes for everything!
  - 

# Tier 1 cloud

- Tier 1 cloud (10 sites of very different size) contains:
  - 10% of RAW on disk, the rest on tape
  - 2 full copies of current ESD on disk
  - A full AOD/TAG at each Tier 1
  - A full set of group DPD
- Access is scheduled, through *ANALYSIS and PHYSICS* groups, and for production
- Users do not run jobs on T1: for production or group based activities
- RAW data reprocessing will occur at T1 (one pass per year: more at the beginning)
- Note that BNL is atypical
  - Complete ESD copy available to all collaborators (not just US)
  - Also functions as a "giant T3" got US:atlas
  - T3 part is available to all users
  - T3=total resource – pledged resource

# Tier 1 cloud: usage

- Group analysis will produce
  - Deep copies of subsets, group DPD
  - Dataset definitions
  - TAG based selections
- Characterised by access to full ESD and sometimes RAW
  - This is resource intensive
  - Must be a scheduled activity
  - Can back-navigate from AOD to ESD as at same site only.
  - Can harvest small samples of ESD (and some RAW) to be sent to Tier 2s
  - Must be agreed by physics and detector groups
- *T*rain model (scheduled access)
  - Efficiency and scheduling gives gains in access. Some form of co-ordination is needed
  - The 'big train' model has been discussed, but requires an infrastructure was not obviously emerging - but EventView may allow a train in the architecture
  - A model for human organization:
    - Group production co-ordinator (may also be simulation co-ordinator)
    - DPD production role per group (OTSMOU task)
    - Can co-ordinate effort in a flexible way
    - Allows management of 'production' space for Group DPD data
  - *Same requirement for group production role to validate and sign-off on group production tasks*

# Tier 2 cloud: usage

- ~30 Tier 2 sites of very, very different size contain:
- Some of ESD and RAW
  - In 2008: 30% of RAW and 150% of ESD in Tier 2 cloud
  - In 2009 and after: 10% of RAW and 30% of ESD in Tier 2 cloud
  - This will largely be 'pre-placed' in early running
  - recall of small samples through the group production at T1
- Additional access to ESD and RAW in CAF
  - 1/18 RAW and 10% ESD: may only be available for calibration work
- 10 copies of full AOD on disk
- A full set of official group DPD *(production area)*
- Lots of small group DPD *(in production area)*
- User data *(in 'SCR$MONTH') (more on this later)*
- Access is 'on demand'

# Tier 2 cloud: usage

- Restricted Tier 2s and CAF
  - *Note: CAF is 'on demand', group analysis at T1 is scheduled*
  - Can specialise some Tier 2s for some groups
  - All Tier 2s are for ATLAS-wide usage
- Most ATLAS Tier 2 data should be 'placed' with lifetime ~ months
  - *Lifetime matches ~4 group DPDs a year*
  - Tier 2 bandwidth is vastly lower, job efficiency higher
  - *Group DPD in 'production' area and 'pinned' to disk*
- Role and group based quotas are essential (but are not emerging quickly!)
  - CPU fair-shares are quite easily done
  - We can at least easily split 'production' from user space
  - Quotas to be determined per group not per user
  - User files can be garbage collected - effectively ~SCR$MONTH unless 'adopted' by a physics/detector group
  - 'Adoption' implies the group 'production' role moves (or reallocates) the files into the group's production quota and 'pins' it
    - The details of this migration need to be fleshed-out
    - The details of the garbage collection also need to be fleshed-out
    - Has to be by ATLAS, as deleted files need to be removed from the catalogues

# Early data comments

- Storage: has to be in place before usage
  - Possible to use in the short term
    - More ESD - so long as you clear the extra events for new data
    - Bigger AOD - so long as you reduce it later
  - Hard to remove AOD features from users
  - Note: augmented AOD for well defined subsets or tasks is not a problem
    - This is a use case for group DPD!
- CPU:
  - At the Tier 1, the CPU is going to be busy much of the time
    - The full reprocessing will obviously wait for calibration and algorithmic development, so the capacity is available until then
    - *The group analysis/big trains have a large resource allocation: the balance between DPD production and reprocessing is adjustable*
  - We anticipate some samples being reprocessed often
    - Output short-lived
    - 'Proper' processing for physics results
    - Must beware inconsistent processing, especially if inclusive streams

# User issues

- The Tier 1s and Tier 2s are collective - if the data is on disk, you (@ a T2) or your group (@ a T1) can run on it
- For any substantial data access, jobs go to the data
  - Users initially thought data goes to the job! Cannot be sustained
  - Better for network better for job efficiency
- Data for Tier 3s should be pulled from Tier 2s using ATLAS tools
  - Tier 3s need to ensure adequate networking
  - We need to monitor (and potentially control) traffic

# DPD's on T3?

- Naive assumption
  - Small ESD samples, some AOD, mainly tuples
- Space issues
  - If 1DPD event is 10kB, 10M ev/TB; 5% of year is 5TB
    - I expect about 1-2TB/user at T3 in 2008
    - Some users already use ~1TB with tuples etc on top of this.
  - Can a typical Tier3 handle this, may be similar to Tier 2 commitment and cost
    - Large data movement (v. large extra load on DQ2)
      - Why not use the Tier 2?
      - Why not just extract what you want?
    - Note: large data movement to Tier 3s will be throttled (outline policy presented without objection to the CB)
- Users may be overoptimistic based on perfectly streamed simulated data or small FDR volume

# Analysis

Analysis model broken into two components

- @ Tier 1: Scheduled central production of augmented AOD, tuples & TAG collections from ESD

→ Derived files moved to other T1s and to T2s

- @ Tier 2: On-demand user analysis of augmented AOD streams, tuples, new selections etc and individual user simulation and CPU-bound tasks matching the official MC production

→ Modest job traffic between T2s

→ Tier 2 files are not private, but may be for small sub-groups in physics/detector groups

→ Limited individual space, copy to Tier3s

# Group based Analysis

- Group analysis will produce
  - Deep copies of subsets
  - Dataset definitions
  - TAG selections
- Characterised by access to full ESD and sometimes RAW
  - This is resource intensive
  - Must be a scheduled activity
  - Can back-navigate from AOD to ESD at same site
  - Can harvest small samples of ESD (and some RAW) to be sent to Tier 2s
  - Must be agreed by physics and detector groups
- Big Trains etc
  - Efficiency and scheduling gains access. Some form of co-ordination is needed
  - If analyses are blocked into a 'big train';
    - Each wagon (group) has a wagon master )production manager
    - Must ensure will not derail the train
  - Train must run often enough (every ~2 weeks)
  - Trains can also harvest ESD and RAW samples for Tier 2s (but we should try to anticipate and place these subsets)

# Reality check

- We cannot keep all RAW data on disk, and
- We cannot sustain random access to RAW on tape
  - Modification for early running:
    - We have some flexibility to increase RAW and ESD on disk temporarily in all Tiers
    - The fraction also decreases with year of data taking
- The disk RAW data is be pre-selected as far as possible
- ~50% RAW and ESD at Tier 2s must also be preselected
  - Any additional needed later, requests above ~20Gbytes/day need to be requested, not grabbed
    - ESD can be delivered in a few hours
    - RAW on tape may take ~week, but can be prioritised
      - All Raw from tape must be requested

# User issues: storage

- User space is provided at T2 and T3: latter is "private" resource
- How is user space at T2 managed
  - Individual quotas mapped to grid cert: No tools
  - Giant aged "scratch space": part of original model
  - New proposal
    - Assign users to particular T2
    - Space then managed "locally"
  - New proposal

# User space on the Grid (1)

- Many discussions took place recently on this matter. Two extremes:
  - "Traditional" computing model:
    - All ATLAS VO members can write to all user space (ATLASUSERDISK token) wherever this token is defined
      - Usually Tier-2s
    - All this space is defined as "scratch" with a retention time of a week to a month
      - This retention time gives users time to decide if they wish to keep the data longer in their private site, promote them to group level and move them to a group area, or delete them
    - Problem: how to guarantee a minimum retention time? The SE may be flooded by other users that force all existing data out
    - Also, how to synchronize the SE with the catalogue(s)? How would the garbage collector work in practice? Would it be a central or local operation?
  - Recent proposal:
    - Allocate all ATLASUSERDISK space to (national or local) groups
      - Pro: no garbage collection, but if the SE gets full, people are readily findable
      - Contra: it breaks down the concept of the Grid and of equal access for all ATLAS members
      - Contra: it forces all analysis jobs to write their output to the SE where the user is authorized to write (lots of unnecessary data movement in real time)

Dario Barberis: User Data

# User space on the Grid (2)

- My proposal:
  - Define the minimal need for user-managed (home) space on the Grid and allocate it at Tier-2s
    - Assuming everyone has at least one "friendly" Tier-2
      - Some ATLAS-wide agreement may be needed to make sure that this is actually true
    - There is additional user space at Tier-3s, which are not under central control
  - Set up all the rest of disk space at Tier-2s as a scratch area with garbage collection
    - Set up a garbage collector that cleans (for example):
      - All non-catalogued files older than a few days
      - All catalogued files older than a month (cleaning the catalogue too, of course)
        - Possibly sending a warning a week in advance
  - Set up distributed analysis tools so that they follow one of these work models:
    1) Jobs write the output to the local SE; either an automatic data transfer is triggered at the end of the job, or it is then the user's responsibility to recover the data or let it fade away
    2) Jobs write the output to the default SE of the user; if that fails for whichever reason, write the output to the local SE or have a list of failover SEs
  - Almost all tools to implement this proposal exist
    - Some may need some further development and tuning to make sure that we are not going to lose data

# Comments about 2008 (and 2009?)

- Data volume unknown
- Machine running unknown
  - Can use gaps to reprocess at T0: not possible in steady state
- More access to RAW and ESD needed.
- Utility of DPD unclear

# Summary and concerns

- Model is well developed but will need to adapt to data
- Most critical issue is disk space
  - Event sizes too big.
  - Do we need both HITS and RDO for simulation?
- Simulation may be limited by CPU
- Memory is an issue for reconstruction.