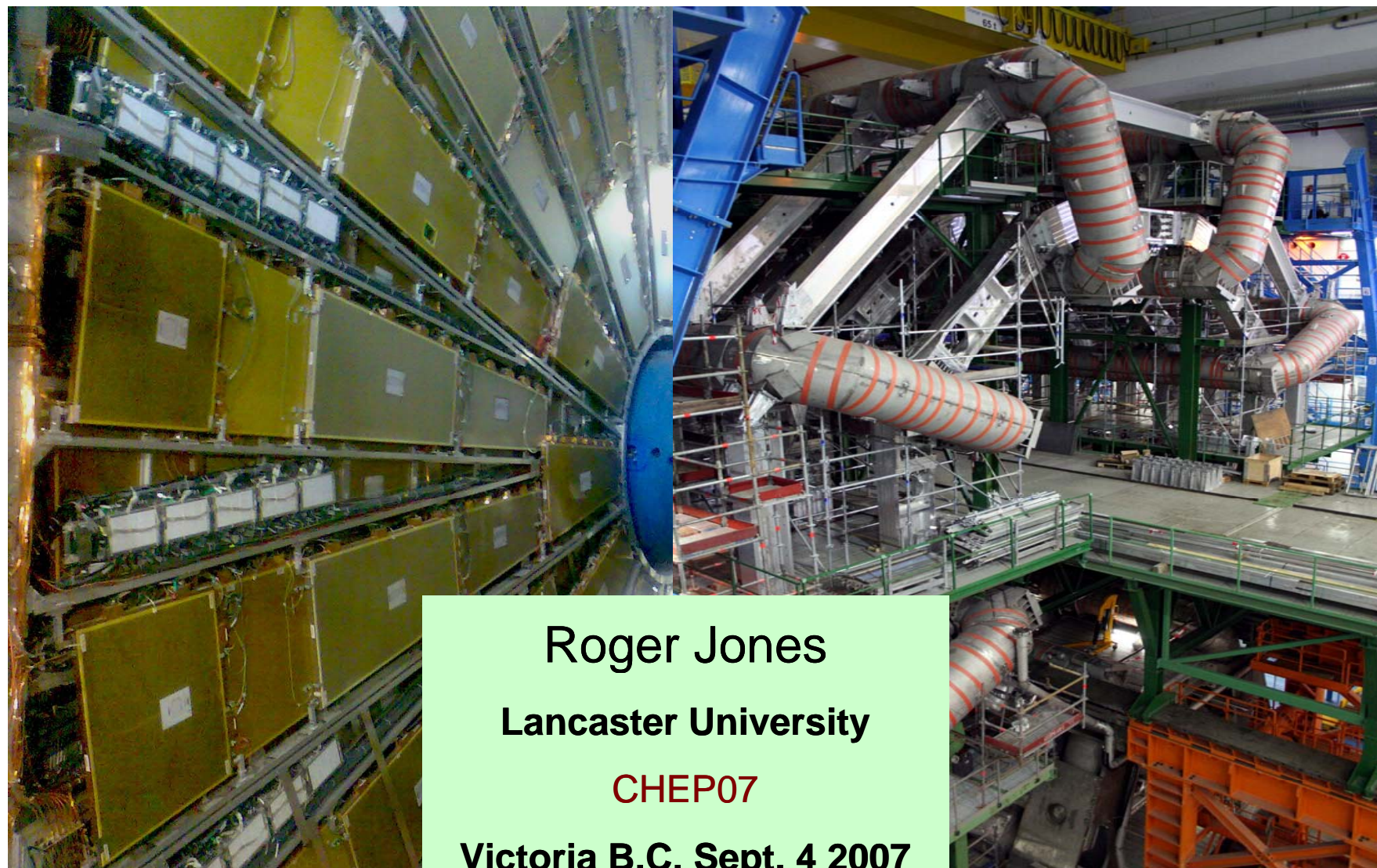


The ATLAS Computing Model



Roger Jones
Lancaster University
CHEP07
Victoria B.C. Sept. 4 2007



Overview



- **Brief summary ATLAS Facilities and their roles**
- **Commissioning**
 - Cosmics running: M3-M6
 - Dummy Data
 - T0/T1
 - Full Dress Rehearsals
- **Data Distribution**
 - CPU, Disk, Mass Storage
- **Data Access**
 - Streaming
 - TAGs
- **Operational Issues and Hot Topics**



Transition to Service versus Development



No more 'clever' developments for the next 18 months!

- **Focus now must be physics analysis performance**
 - Which means integration, deployment, testing, documentation, EDM
- **Pragmatic addition of limited vital additional functionality**
- **Stimulation will come from the physics!**

But also many 'big unsolved problems' for later:

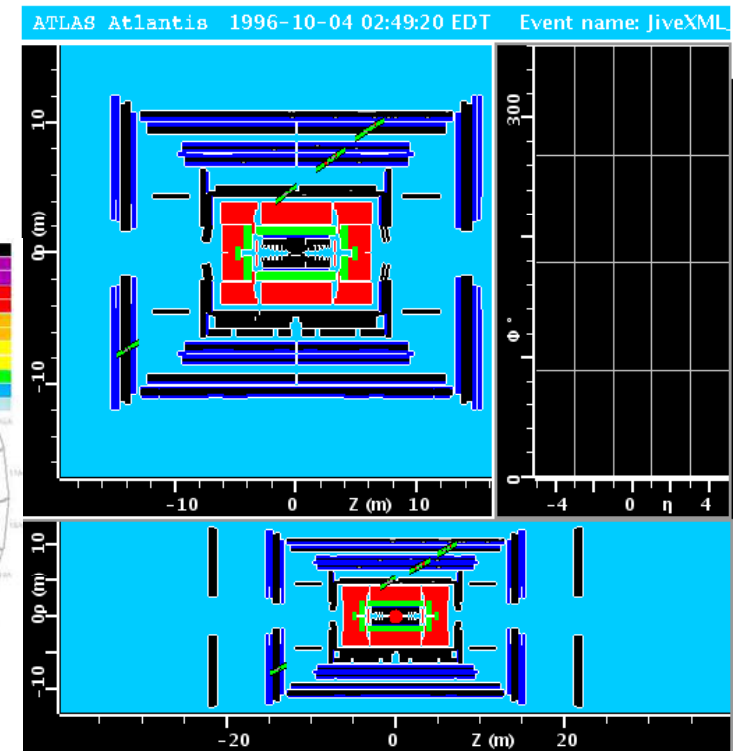
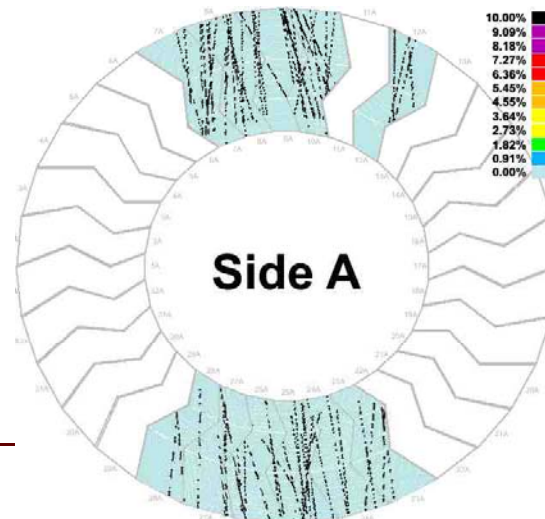
- How can we store data more efficiently?
- How can we compute more efficiently?
- How can we reduce memory profiles?
- How far can we move data reduction earlier in the chain?
- How should we use virtualisation?
- What tasks can really be made interactive, and which are desirable?
- How do we use really high-speed communications?



Computing Resources



- **No major change in high-level Computing Model, gradual evolution**
 - Computing Model Review, Jan 2005
 - Computing Technical Design reports July-October 2005
 - Revised planning Summer 07
- **A Big Change - Planning now beginning to be informed by real data!**
 - Detector commissioning has a real payload
 - Updates in the accelerator schedule
 - Updating with each revision:
 - **Event sizes**
 - Large improvements
 - **Reconstruction times**
 - **Simulation times**
 - Constant tension!
 - **This is still evolving!**
 - **Memory an issue**





ATLAS Requirements start 2008, 2010



	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

•Note the high ratio of disk to cpu in the Tier 2s

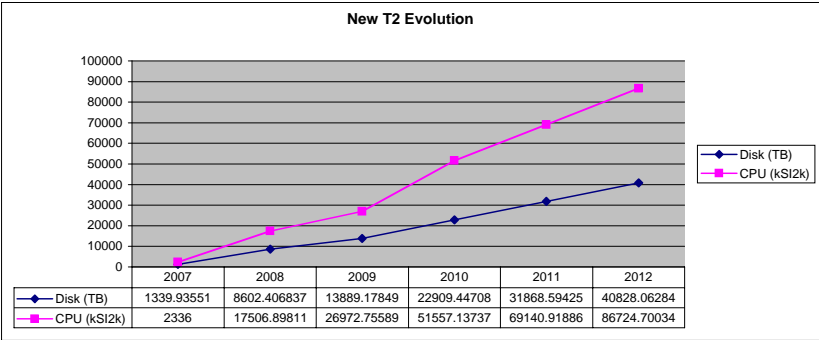
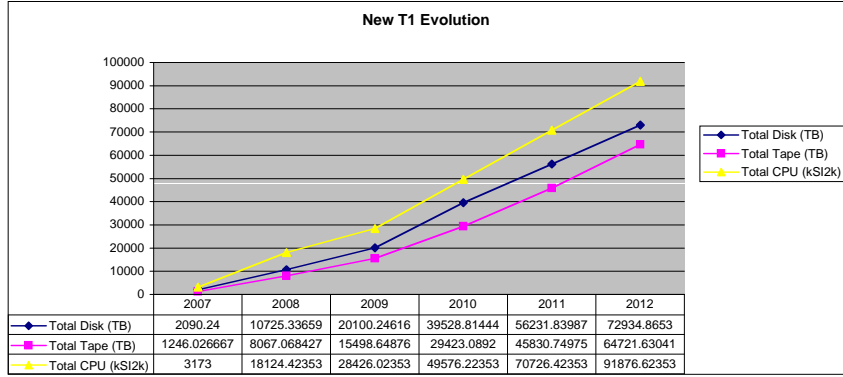
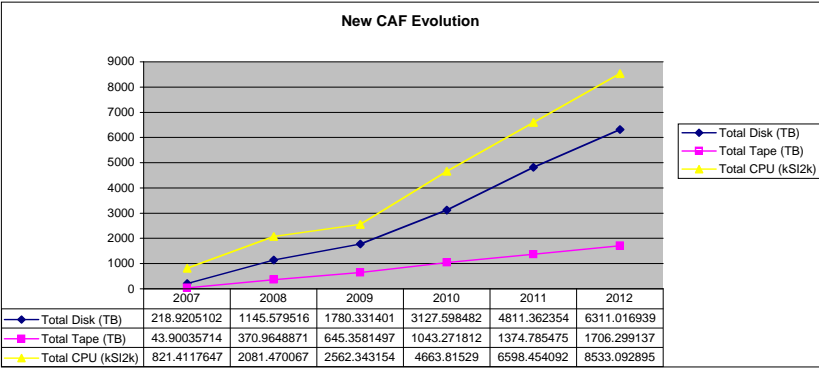
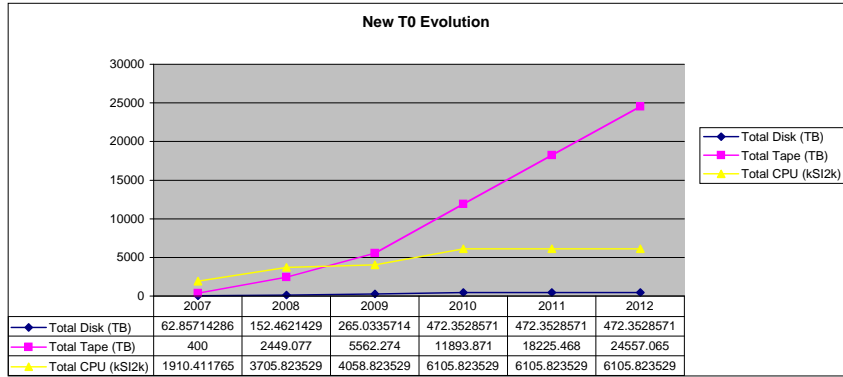
- Not yet realised
- May require adjustments



Resource Plan



- Recall the early data is for calibration and commissioning
- This is needed either from collisions or cosmics etc
- Change of schedule makes little change to the resource profile



• The issue of evolution post 2012 is now under active consideration



Roles for Sites



The roles in our model are remarkably stable

- **CERN Tier-0:**
 - Prompt first pass processing on express/calibration & physics streams with old calibrations - calibration, monitoring
 - Calibration tasks on prompt data
 - 24-48 hours later, process full physics data streams with reasonable calibrations
- **CERN Analysis Facility**
 - Access to ESD and RAW/calibration data on demand
 - Essential for early calibration / Detector optimization / algorithmic development
- **10 Tier-1s worldwide (2/3 of collaboration):**
 - Reprocess 1-2 months after arrival with better calibrations
 - Reprocess all resident RAW at year end with improved calibration and software
- **30+ Tier 2 Facilities distributed worldwide**
 - On demand user physics analysis of shared datasets
 - Limited access to ESD and RAW data sets
 - Simulation (some at Tier 1s in early years)
- **Tier 3 Centers distributed worldwide**
 - On demand physics analysis
 - Data private and local - summary datasets



Advice for Tier 3s



- **A Tier 3 is not a defined set-up, nor does it have any formal commitment**
- **Working definition: 'Tier 3' facilities are for local use, not for all of ATLAS**
 - Need a Grid User Interface
 - There are some ATLAS common requirements
 - *It is also desirable to have a CE and SE to allow occasional use by ATLAS for additional production capacity at local discretion*
- **ATLAS Tier 3 task force is starting to give recommendations and to describe possible solutions**
 - Aim is to help sites
 - This will be advisory, not prescriptive!
- **They can come in different forms, many ideas**
 - Dedicated cpu and disk racks
 - Fraction of fabric for Tier 2s
 - Desktop clusters
 - Enumerate the possibilities, match to cases



Roles for People



- **Need for tens of groups**
 - It is clear we will have 20+ physics/detector groups
 - We are also seeing justified regional / national groups
 - We envisage in a typical group:
 - Production role
 - Installer
 - User
- **We really needs groups and roles to be known and understood by the middleware**
 - Different roles, different quotas / fairshares / access to SRMs
 - We can live without individual quotas etc for now
 - But we need to be able to manage at the group level
 - At present, we are only able to split storage between production and user by 'force majeure'!
 - There is a danger here of divergent VO-specific solutions



Data Movement Policy



- **We still believe that an efficient system requires**
 - **The pre-placement of data**
 - Early on, try to place all of the Express Stream ESD at Tier 2s
 - Express stream determines most RAW data placed on disk
 - **Multiple instances**
 - **Jobs going to the data**
 - The brokers and tools like GANGA steer jobs to close CE
- **The Tier 1s should serve (most) close Tier 2 data needs**
 - Full AOD set and group Derived Physics Datasets (DPD) in each cloud
 - Pre-placed small RAW and ESD samples in Tier 2s
- **This requires that**
 - **Tools place the data quickly and efficiently**
 - But if that fails, 'wildcat' data movements by individuals makes the situation worse for everyone
 - **Policies that do the same**
 - **If only complete datasets are to be moved**
 - ☺ The datasets must be closed in a reasonably short time
 - ☹ ~~The copy to be replicated must itself be copied quickly/completely~~



User Data Movement Policy



- **Users need to access the files they produce**
 - This means they need (ATLAS) data tools on Tier 3s
- **There is a risk: some users may attempt to move large data volumes (to a Tier 2 or Tier 3)**
 - SE overload
 - Network congestion
- **ATLAS policy in outline:**
 - O(10GB/day/user) who cares?
 - O(50GB/day/user) rate throttled
 - O(10TB/day/user) user throttled!
 - Planned large movements possible if negotiated
- **But how can we enforce this?!**
 - The first line of defence is user education, but it is not enough





Interaction Between Computing Models



- **So far, the computing models have been tested largely independently**
 - They are (quite rightly) designed to optimize the operation of the individual VO
 - Many sites and many networks are shared
- **How will the VO Computing Models interact?**
 - E.g. LHCb are the only VO with on-demand analysis at Tier 1s
 - Different access patterns etc
 - Probably a manageable interaction, but we need experience
 - E.g. on-demand data access in CMS from Tier 2 to any Tier 1
 - Very different network usage c.f. ATLAS
- **The proposed joint exercises will be very important to iron-out issues in advance of real daay**



Access Optimization - Streaming



- **All discussions are about optimisation of data access**
- **ATLAS now plans streaming of RAW, ESD, AOD**
- **Streams derived from trigger information**
 - Does not change with processing version
 - Inclusive versus exclusive being evaluated based on recent tests
- **Current tests have ~5 streams, plus an overlap stream for exclusive streaming**
 - Evaluating bookkeeping load, multi-stream analyses etc
- **Further refinement of event selection using TAGs**



Event data Model



- **Lots of progress on all formats**
 - Trade size versus flexibility
 - Much more will be possible from AOD that originally planned
 - AOD size still under control
- **Big discussion on Derived Physics Data**
 - This can be made by physics/detector groups or individuals
 - Much work on possible common formats - good for groups
 - Some use cases are more demanding (large samples, partial data required)
 - Studies of
 - Skimming (event selections)
 - Thinning (selecting containers or objects from a container)
 - Slimming (selection of properties of an object)
 - Also studies of alternate formats
 - Hope this is only needed in a few cases



Optimised Access - TAGs



- **RAW, ESD and AOD will be streamed to optimise access**
- **The selection and direct access to individual events is via a TAG database**
 - TAG is a keyed list of variables/event
 - Overhead of file opens is acceptable in many scenarios
 - Works very well with pre-streamed data
- **Two roles**
 - Direct access to event in file via pointer
 - Data collection definition function
- **Two formats, file and database**
 - **Now believe large queries require full database**
 - Multi-TB relational database; at least one, but number to be determined from tests
 - Restricted it to Tier0 and a few other sites
 - Does not support complex 'physics' queries
 - **File-based TAG allows direct access to events in files (pointers)**
 - Ordinary Tier2s hold file-based primary TAG corresponding to locally-held datasets
 - Supports 'physics' queries
- **See performance and scalability test talk from Helen McGlone**



Commissioning Plans



- **We are now getting real data, at realistic rates to validate the model**
 - **M3 (mid-July)**
 - Cosmics produced about 100TB in 2 weeks
 - Stressed offline by running at 4 times the nominal rate (32LAr sample test)!
 - **M4 now underway - August 23 - early September**
 - **Expect about**
 - Total data volume: RAW = 66 TB , ESD + AOD = 6 TB
 - 20TB RAW data and 6TB ESD at RAL
 - 2TB ESD at 5 Tier 2 sites
 - Data distribution as for real data
 - Currently writing at 200MB/sec, half nominal
 - RAW and ESD now appearing at RAL
 - **M5 will be similar - October 16-23**
 - **M6 will run from end December until real data**
 - Incremental goals, reprocessing between runs
 - Will run close to nominal rate
 - Maybe ~420TB by start of run, plus Monte Carlo
 - **T1 should treat this as valuable data, but may only live for about a year**
- **Full rate T0 processing OK**
- **Data exported to 5 / 10 T1's and stored OK, and did more !**
- **For 2 / 5 T1's exports to at least 2 T2's not yet all done**
- **Quasi-rt analysis in at least 1 T2 OK, and did more !**
- **Reprocessing in Sept. in at least 1 T1 not yet started**

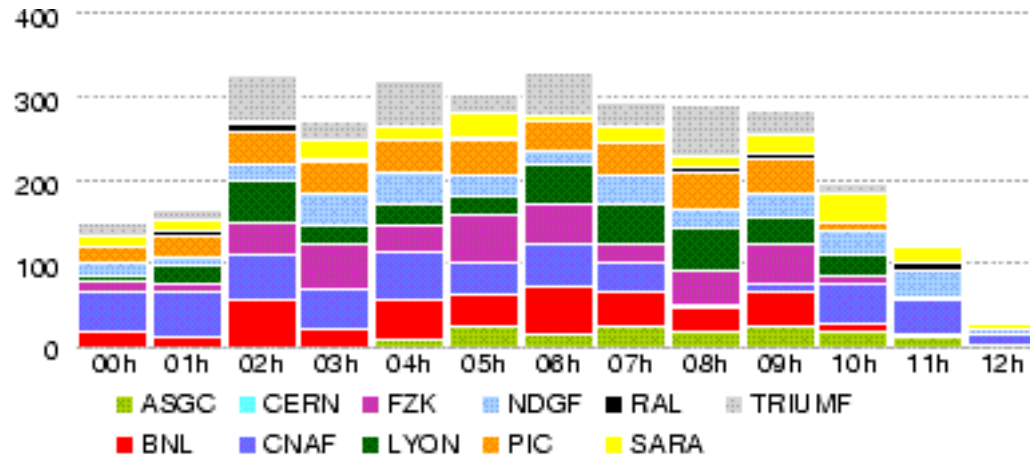


• Full chain from DAQ to DA in last week of Aug07

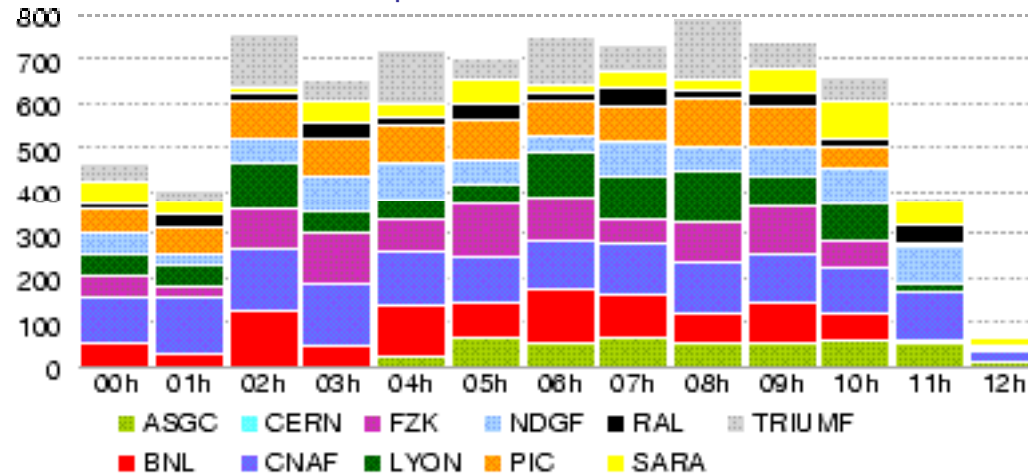


M4 Data Movement

• Throughput (MB/s)



• Completed File Transfers





FDR production goals



- **Simulated events injected in the t/daq**
 - Realistic physics mix in bytestream format incl. luminosity blocks
 - Real data file & dataset sizes, trigger tables, data streaming
- **T0/T1 data quality, express line, calibration running**
 - Use of conditions database
- **T0 reconstruction: ESD, AOD, TAG, DPD**
 - ⇒ Exports to T1&2's
- **Remote analysis**
 - **@ the T1's**
 - Reprocessing from RAW → ESD, AOD, DPD, TAG
 - Remake AOD from ESD
 - Group based analysis → DPD
 - **@ the T2&T3's**
 - Root based analysis
 - Trigger aware analysis with Cond. and Trigger db
 - No MC truth, user analysis
 - MC/Reco production in parallel



FDR Schedule



Round 1

1. **Data streaming tests** **DONE**
2. **Sept/Oct 07 Data preparation** **STARTS SOON**
3. **End Oct07: Tier 0 operations tests**
4. **Nov07-Feb08. Reprocess at Tier1, make group DPD's**

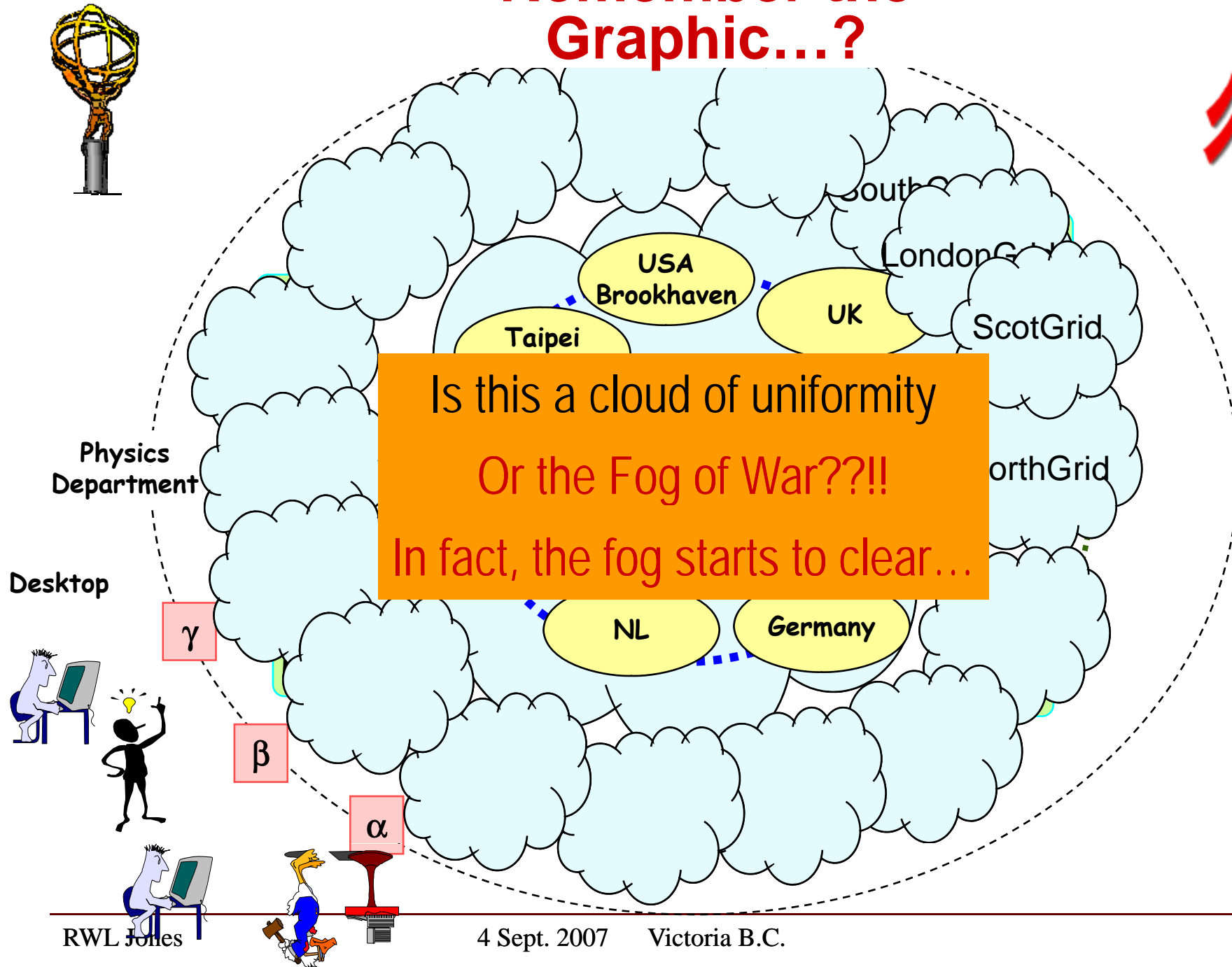
Round 2

ASSUMING NEW G4

1. **Dec07-Jan08 New data production for final round**
2. **Feb08 Data prep for final round using**
3. **Mar08. Reco final round ASSUMING SRMv2.2**
4. **Apr08. DPD production at T1's**
5. **Apr08 More simulated data prod in preparation for first data.**
6. **May08 final FDR**

- **First pass production should be validated by year-end**
 - **Reprocessing will be validated months later**
 - **Analysis roles will still be evolving**
- ¡Expect the unexpected!

Remember the Graphic...?





Summary



- **The computing model has (so far) stood up well**
- **The localization of data to clouds future proofs!**
- **Scheduled production is largely solved**
- **On-demand analysis, data management & serving many users are the mountains we are climbing now**
- **Users are important to getting everything working**
 - **'No Pain, No Gain!'**

