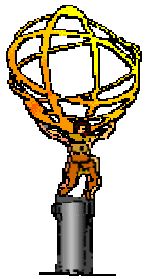
ATLAS Computing Model & Service Challenges

Roger Jones

12th October 2004

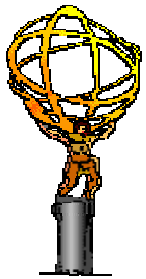
CERN

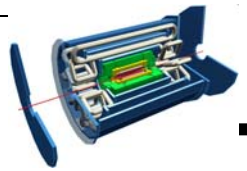


Computing Model

- **A new revision started late spring 03**
  - Workshop in June 04, input from CDF, DØ and ATLAS
- **Very little input so far from some key communities:**
  - detector groups on calibration and alignment computing needs
    - **Calibration input on 2 October 2004!**
  - physics groups on data access patterns
    - **This is a major concern, as some have unrealistic ideas!**
- **Still large uncertainties on the final event sizes**
  - Huge potential impact on access and on costs!
  - With the advertised assumptions, we are at the limit of available disk
  - RAW data cannot be bigger because of TDAQ bandwidth

The ATLAS System





~Pb/sec

Event Builder

10 GB/sec

PC (2004) = ~1 kSpecInt2k

Event Filter
~7.5MSI2k

450 MB/sec

- Some data for calibration and monitoring to institutes
- Calibrations flow back

T0 ~5MSI2k

- ~5 Pb/year
- No simulation

~ 75MB/s/T1 for ATLAS

Tier 0

Tier 1

US Regional Centre

Dutch Regional Centre

French Regional Centre

UK Regional Centre (RAL)

- ~2MSI2k/T1
- ~2 Pb/year/T1

10 Tier 1s assumed

≥622Mb/s links

Tier 2

Northern Tier ~200kSI2k

Tier2 Centre ~200kSI2k

Centre ~200kSI2k

Centre ~200kSI2k

- ~200 Tb/year/T2

≥622Mb/s links

Lancaster ~0.25TIPS

pool

anchester

Sheffield

Physics data cache

100 - 1000 MB/s links

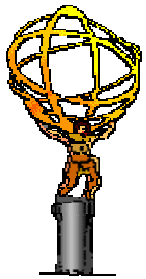
Desktop

Workstations

Each Tier 2 has ~15-20 physicists working on one or more channels

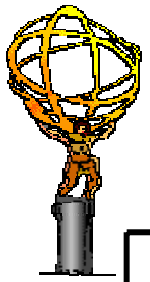
Each Tier 2 should have the full AOD, TAG & relevant Physics Group summary data

Tier 2 do bulk of simulation



Computing Resources

- **Assumption:**
  - 50 days running in 2007 and 2008 (over the year break)
  - Tier-0 has raw+calibration data+first-pass ESD
  - CERN 'tier-1' is analysis-only (big tier-2)
- **Notes:**
  - One-off purchase of disk buffer is needed in first year
    - Allows coherent data sets available while reprocessing
  - Efficiencies folded-in to the numbers
  - Model trys to capture analysis activity
    - The analysis test in phase 3 of DC2 will be important to answer some of the questions
    - Especially hard to quantify year-1 semi-private reprocessing need, but estimates included

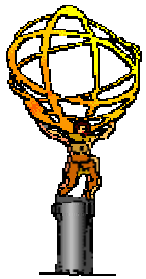


The input Numbers

	Rate(Hz)	sec/year	Events/y	Size(MB)	Total(TB)
Raw Data (inc express etc)	200	1.00E+07	2.00E+09	1.6	3200
ESD (inc express etc)	200	1.00E+07	2.00E+09	0.5	1000
General ESD	180	1.00E+07	1.80E+09	0.5	900
General AOD	180	1.00E+07	1.80E+09	0.1	180
General TAG	180	1.00E+07	1.80E+09	0.001	2
Calibration (ID, LAr, MDT)					44 (8 long-term)
MC Raw			2.00E+08	2	400
ESD Sim			2.00E+08	0.5	50
AOD Sim			2.00E+08	0.1	10
TAG Sim			2.00E+08	0.001	0
Tuple				0.01	

Nominal year 10^7 s

Accelerator efficiency 50%



Year 1 T0 requirements

Table Y1.1 **CERN T0 : Storage requirement**

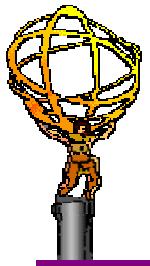
	Disk (TB)	Tape (TB)	
Raw	0	3040	→
General ESD (prev..)	0	1000	→
Total	0	4040	

ESD is 24% of Tape
ESD 0.5MB

Table Y1.2 **CERN T0 : Computing requirement**

	Reconstr.	Reprocess.	Calibr.	Cent.Analysis	User Analysis	Total (kSI2k)
CPU (KSI2k)	3529	0	529	0	0	4058

Note that the calibration load is evolving
Aim here is for steady-state requirements (then vire resources for start-up)

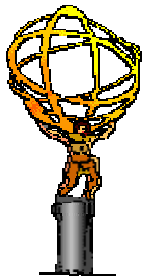


CERN Super-T2 (First year)

Table Y1.3 Storage requirement

	Disk (TB)	Auto.Tape (TB)
Raw	46	0
General ESD (curr.)	26	0
General ESD (prev.)	0	18
AOD (curr.)	257	0
AOD (prev.)	0	4
TAG (curr.)	3	0
TAG (prev.)	0	2
ESD Sim (curr.)	143	0
ESD Sim (prev.)	0	100
AOD Sim (curr.)	29	0
AOD Sim (prev.)	0	20
Tag Sim (curr.)	0.2	0
Tag Sim (prev.)	0	0.2
Calibration	57	
User Data (100 users)	173	0
Total	733	144

- **Small-sample chaotic reprocessing 170kSI2k**
- **Calibration 530kSI2k**
- **User analysis ~810kSI2k**
- **This site does not share in the global simulation load**
  ESD is 23% of Disk
  ESD is 82% of Tape
- **The start-up balance would be very different, but we should try to respect the envelope**



Year 1 T1 Requirements

Table Y1.5	T1 : Storage requirement	
	Disk (TB)	Auto.Tape (TB)
Raw	46	320
General ESD (curr.)	257	90
General ESD (prev..)	129	90
AOD	283	36
TAG	3	0
Calib	57	0
RAW Sim	0	40
ESD Sim (curr.)	29	10
ESD Sim (prev.)	14	10
AOD Sim	31	4
Tag Sim	0	0
User Data (20 groups)	69	0
Total	918	600

Typical Tier-1
Year 1 resources

This includes a '1year, 1 pass' buffer

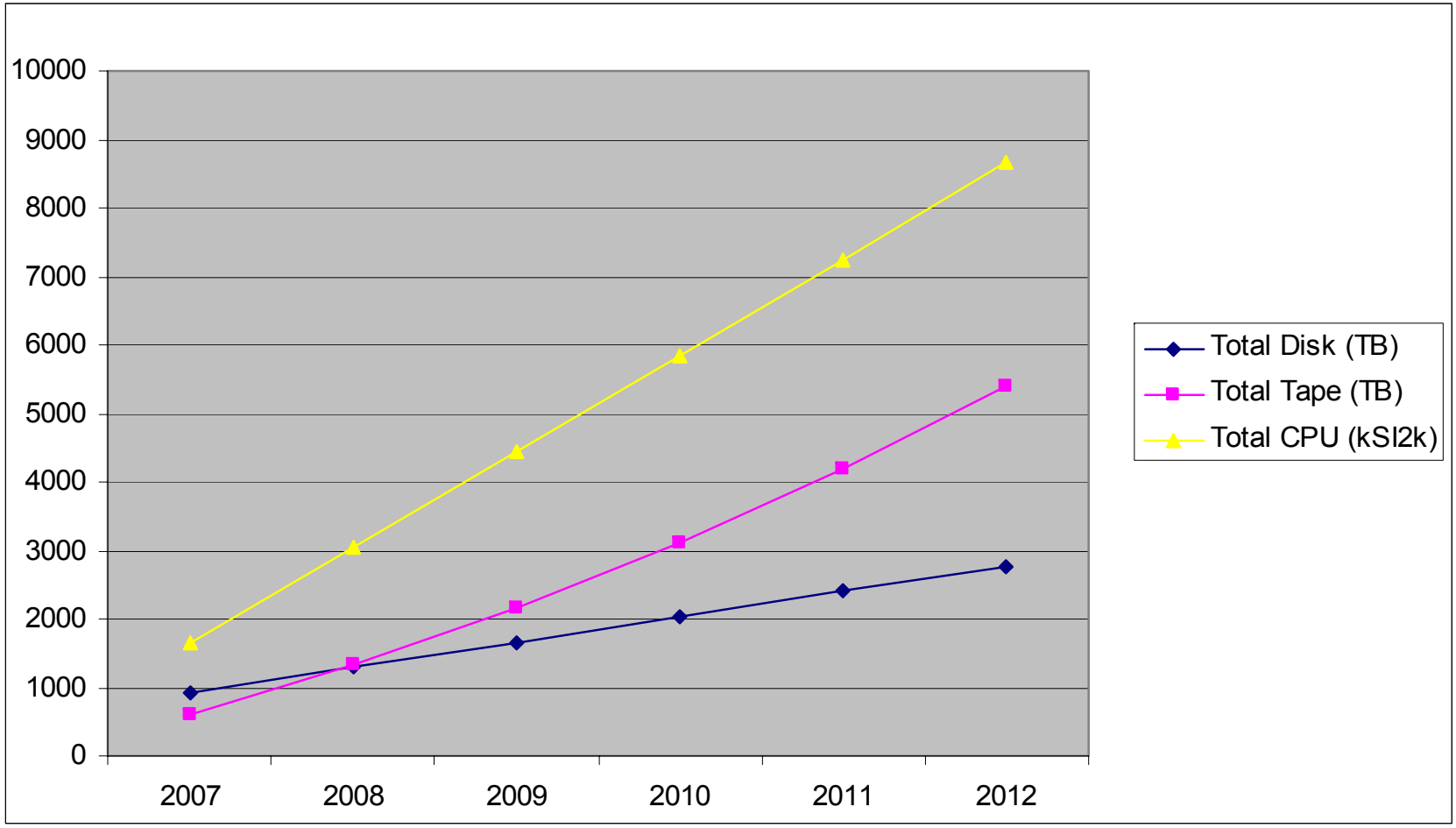
ESD is 47% of Disk
ESD is 33% of Tape

Current pledges are ~55% of this requirement
Making event sizes bigger makes things worse!

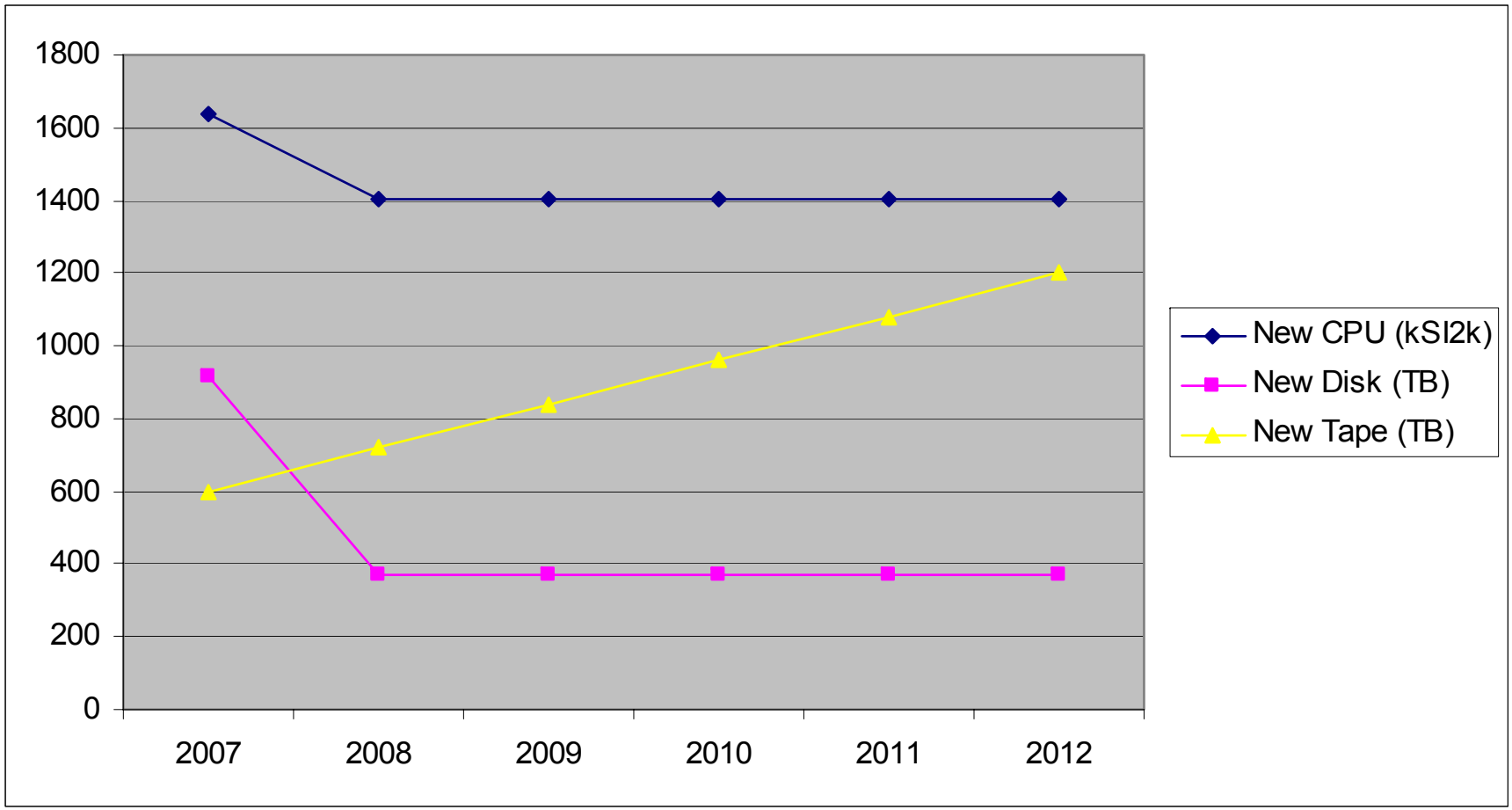
Estimate about 1660kSi2k for each of 10 T1s

Central analysis (by groups, not users) ~1200kSI2k

Single T1 Evolution (totals)



Single T1/year (per year)



Tier-2 (First year)

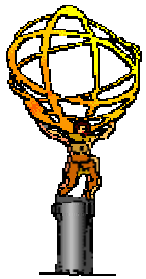
Table Y1.7 Typical Storage requirement

	Disk (TB)
Raw	1
General ESD (curr.)	13
AOD	64
TAG	3
RAW Sim	0
ESD Sim (curr.)	3
AOD Sim	7
Tag Sim	0
User Group	17
User Data	26
Total	134

- User activity includes some reconstruction (algorithm development etc)
- Also includes user simulation
- T2s also share the event simulation load, but not the output data storage

Table Y1.8 Typical Computing requirement

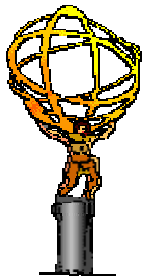
	Reconstruction.	Reprocessing	Simulation	User Analysis	Total (kSI2k)
CPU (KSI2k)	0	0	22	121	143



Overall Year-1 Resources

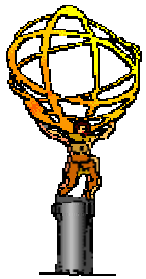
	CERN	All T1	All T2	Total
Tape (Pb)	4.3 Pb	6.0 Pb	0.0 Pb	10.3 Pb
Disk (Pb)	0.7 Pb	9.2 Pb	5.4 Pb	15.3 Pb
CPU (MSI2k)	5.6 MSI2k	16.6 MSI2k	5.7 MSI2k	27.9 MSI2k

If T2 supports private analysis, add about 1 TB and 1 kSI2k/user



Issues

- **Lively debate on streaming and exclusive streams**
  - e.g. ESD/AOD meeting last Friday
- **Model always had streaming of AOD etc**
  - No problem with post T0 streaming of RAW/ESD or small strims
  - We need typical selections to determine the most efficient streams
- **Level of access to RAW?**
  - Depends on functionality of ESD
  - Discussion of small fraction of DRD – augmented RAW data
- **Much bigger concerns about non-exclusive streaming**
  - How do you handle the overlaps when you spin over 2 streams?
  - Real use cases needed to calculate the most efficient access
- **On the input side of the T0, assume following:**
  - Primary stream – every physics event
    - **Publications should be based on this, uniform processing**
  - Calibration stream – calibration + copied selected physics triggers
    - **Need to reduce latency of processing primary stream**
  - Express stream – copied high-pT events for 'excitement' and (with calibration stream) for detector optimisation
    - **Must be a small percentage of total**

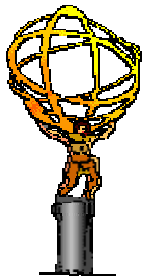


Networking – T0-T1

- EF⇔T0 maximum 300MB/s (450MB/s with headroom)
- If EF away from pit, require 7GB/s for SFI inputs (10x10Gbps with headroom)
- Offline networking off-site now being calculated with David Foster
- Recent exercise with (almost) current numbers
- Full bandwidth estimated as requirement*1.5(headroom)*2(capacity)

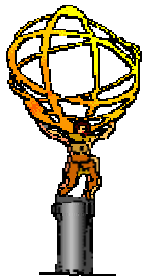
➔ Propose dedicated networking test beyond DC2

	RAL (typical T1?)	T0 Total
ATLAS (MB/s)	72	874
T1 Total ATLAS Gbps	1.7	14.4
T1 Gbps full	3.5	43
Assumed Gbps	10	70



Additional Bandwidth

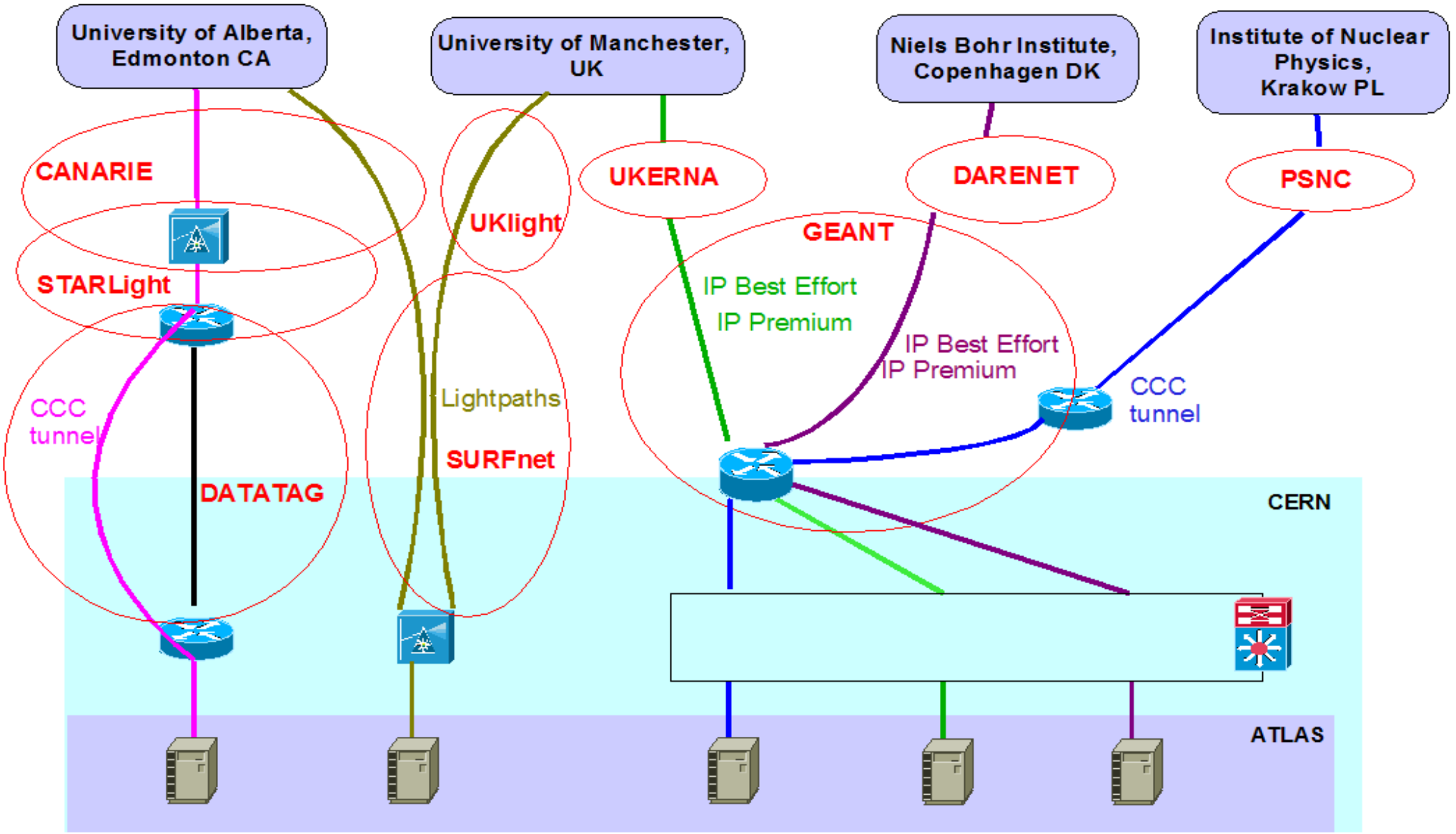
- **There will also be traffic direct from the online to the outside world**
  - **Monitoring – low volume overall, but may be in bursts**
  - **Calibration – generally low volume, but some – MDT for example – may be large for short periods (~Gbps)**
  - **A possibility (for dedicated periods) is offline event filtering**
    - **Full rate would be ~10x10Gbps**
    - **More likely a reduced stream, several Gbps for short periods**
    - **Big issues here, but should not be excluded a priori**



Organisation/Resources

- **The T0-T1 transfer tests are much delayed**
  - Now in Nov-Dec 04, so parallel with Service Challenge?
  - Organisation of tools is also slow
  - Oxanna Smirnova co-ordinating
- **Need for higher-volume tests to**
  - Stress the network
  - Stress the data management tool (Don Quixote)
- **Willing participants/partners:**
- **Lancaster (with Richard Hughes-Jones, Manchester):**
  - Through ESLEA, interest in dedicated light-paths
  - 2-year post to be released shortly, available early in new year
- **CERN**
  - Brian Martin and Catalin Meirosu
  - Continuing online tests
  - Need to be integrated with general effort

ATLAS remote farms – network connectivity



Catalin.Meirosu@cern.ch, Edoardo.Martelli@cern.ch - last update: 20040930

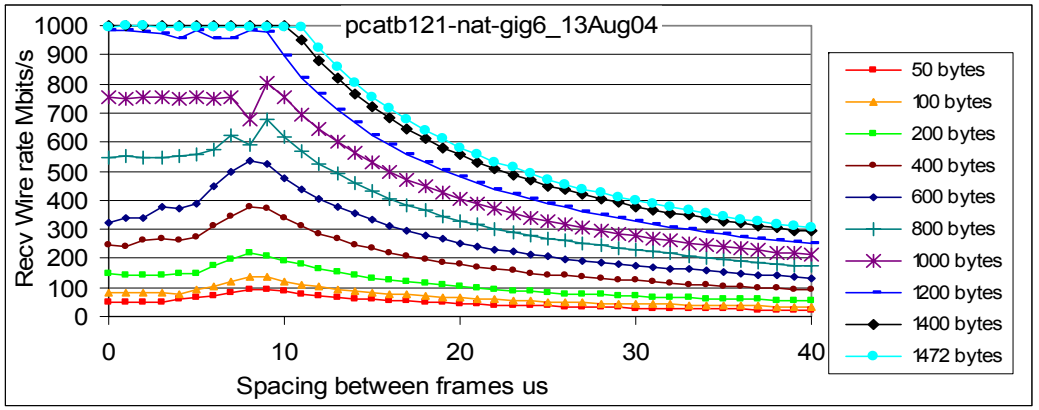
Bandwidth measurements

- **The networking is layered, from the physical transmission media (layer 1) to the application (layer 7)**
- **Tests at layer 2,3**
  - **relevant when the remote and the local sites are logically in the same LAN**
  - **example: throughput between CERN – INP Krakow, August 2004:~1000 Mbit/s**
- **Layer 4 tests: TCP, UDP**
  - **Relevant for general-purpose, Internet-style connectivity**
  - **Performed tests between Geneva and Copenhagen, Edmonton, Krakow, Manchester**
  - **Test equipment: server PCs, running patched Linux kernels and open source software for network measurements**

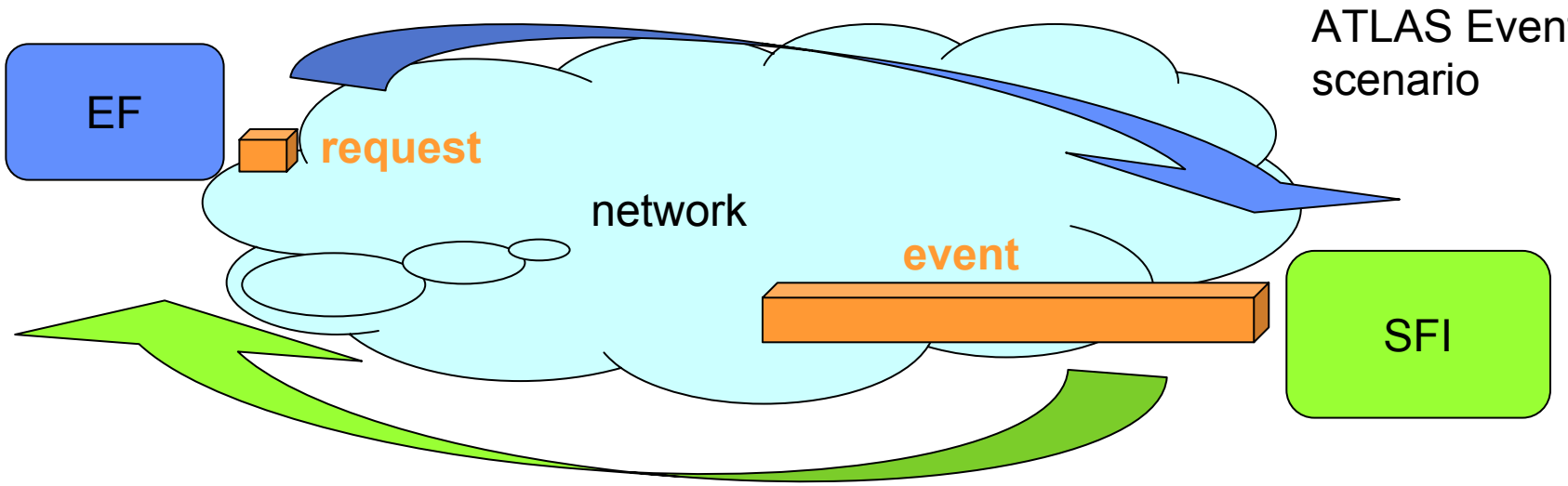
Example: Geneva – Manchester

•The network can sustain 1Gbps of UDP traffic, but the average server has problems with smaller packets

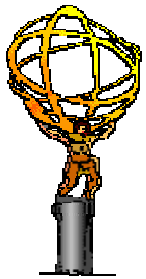
• Degradation for packets smaller than ~1000bytes, caused by the PC receiving the traffic



Real application in an ATLAS context



- **Simple request-response program**
  - **Emulation of the request-response communication between the SFI and EFD in the Event Filter**
  - **Runs over TCP/IP**
  - **The client sends a small request message**
  - **The server answers with an up to 2 MB message**
- **Results … to be understood**

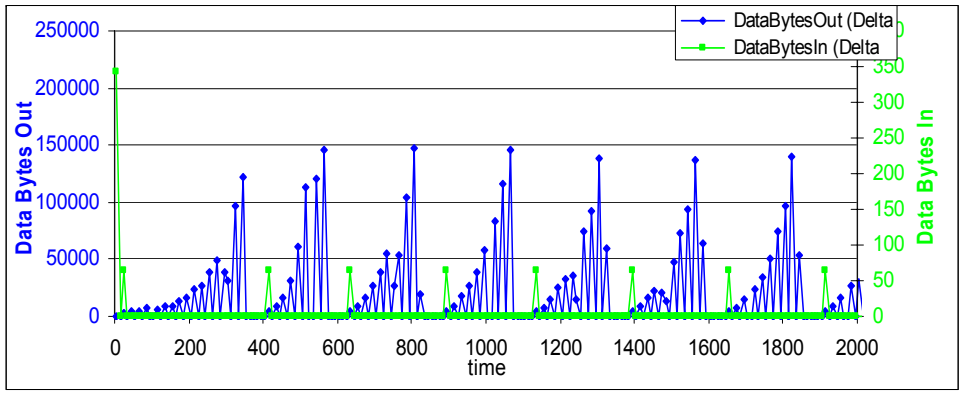


Request – Response results, CERN – Uni. Manchester connection

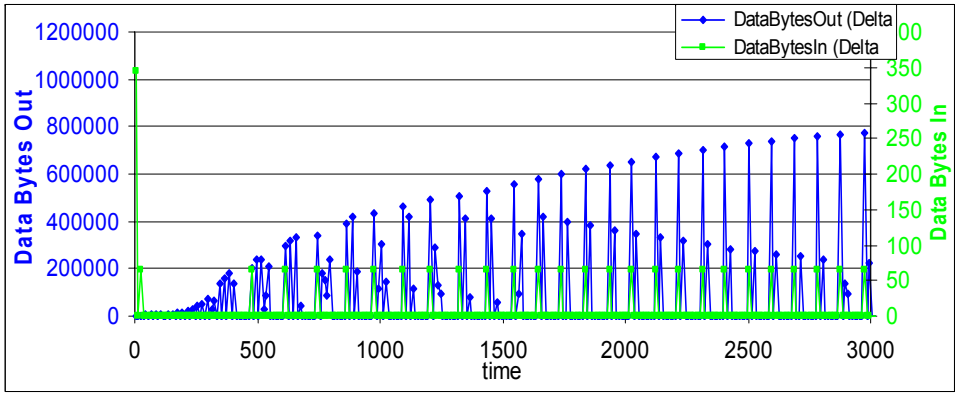
- Good response if the TCP stack is properly tuned, poor if not
  - Need to understand TCP implementation issues, not only the generic protocol
- 800Mbit/s achievable with tuned stack, 120 Mbit/s without – the same end nodes were used in both cases !

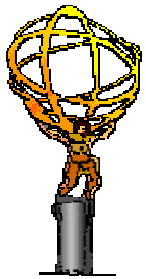
- 64 byte Request green
  1 MByte Response blue

Out-of-the-box TCP settings



Tuned TCP stack





Conclusions and Timetable

- **Computing Model documents required by end of year**
  - We will not now have completed the DC2 Tests by then, especially the analysis tests
  - We can have serious input from detector calibrators and physics groups (sample sizes, access patterns)
  - We also need to know (urgently) about off-site networking from online (calibration, monitoring, …)
  - Event access times
- **Computing Model review in January 2005 (P McBride)**
  - We need to have serious inputs at this point
- **Documents to April RRBs**
- **MoU Signatures in Summer 2005**
- **Computing & LCG TDR June 2005**