Profiling Analysis at Startup

Jim Cochran
Iowa State

Charge from Rob & Michael

Best guess as to the loads analysis will place on the facilities (starting from physics working group plans, job and task definitions) :

- likely dataset distributions to the Tier 2 (content and type)
- users submission and data retrieval activity

Outline:

- Review run1 load estimate (moving target) and concerns
- Focus on expectations from 1st 1-2 months
- Recommendation on dataset "distributions" (my personal suggestion)

To set the scale

From USATLAS db (Dec. 2008):

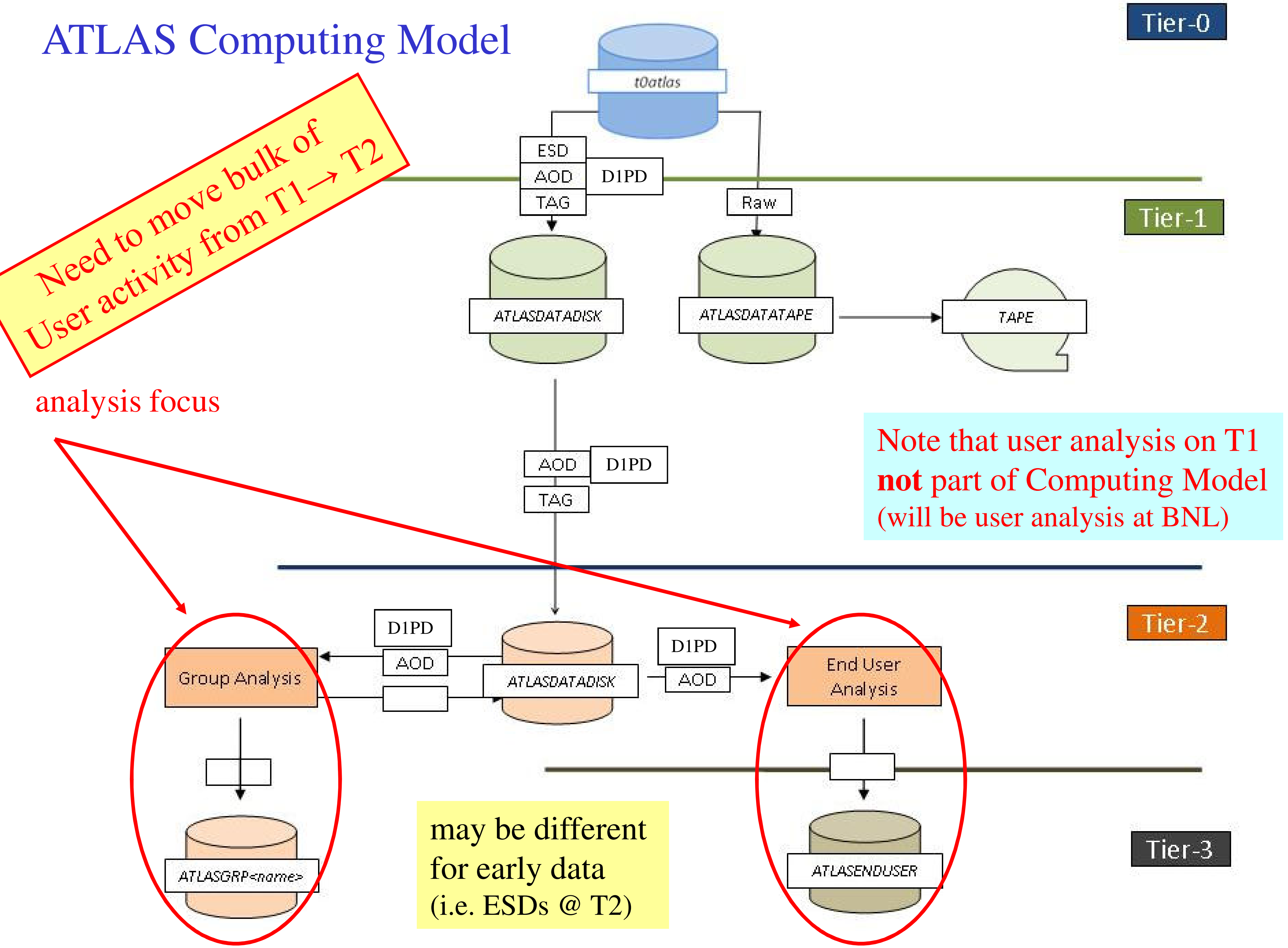
Total people in US ATLAS = 664

Total students = 140
Total postdocs = 90
Total research scientists = 105
Total faculty = 140

~375 potential analyzers

During STEP09, supported ~200 non-HC users (see Kaushik's talk)
- but activity profile likely was not what we will have with real data

ATLAS Computing Model



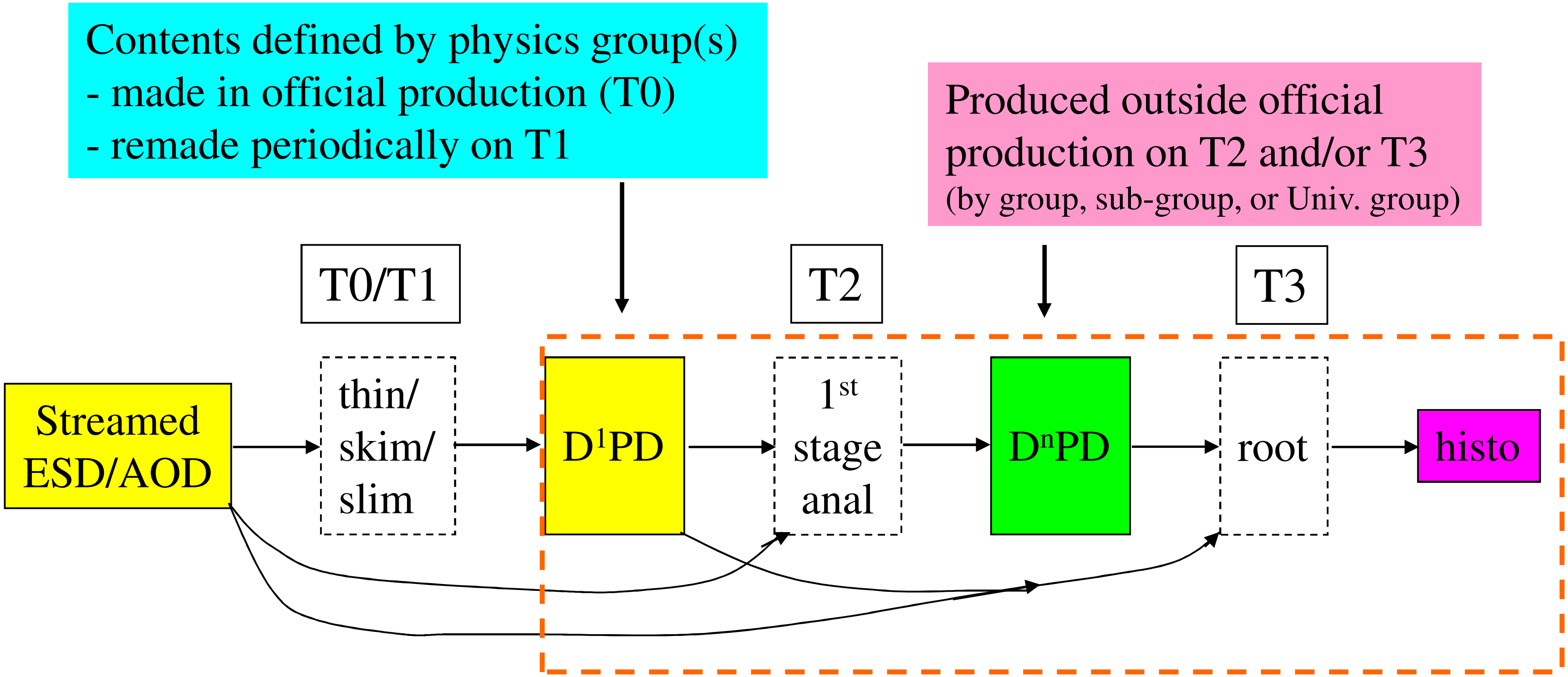
Need to move bulk of User activity from T1 → T2

analysis focus

Note that user analysis on T1 **not** part of Computing Model (will be user analysis at BNL)

may be different for early data (i.e. ESDs @ T2)

ATLAS Analysis Model – analyzer view



Contents defined by physics group(s)
- made in official production (T0)
- remade periodically on T1

Produced outside official production on T2 and/or T3
(by group, sub-group, or Univ. group)

T0/T1

T2

T3

Streamed ESD/AOD

thin/ skim/ slim

D^{1}PD

1st stage anal

D^{n}PD

root

histo

ESD/AOD, D^{1}PD, D^{2}PD - POOL based

D^{3}PD - flat ntuple

Expected analysis patterns for early data

Assume bulk of group/user activity will happen on T2s/T3s
(define user accessible area of T1 as a T3 [BAF/WAF])

Assume final stage of analysis (plots) happens on T3s (T2s are not interactive)
[except for WAF]

<u>Two primary modes</u>:

(1) Physics group/user runs jobs on T2s to make tailored dataset (usually D^3PD)
(potential inputs: ESD,AOD,D^1PD)
resultant dataset is then transferred to user's T3 for further analysis

(2) group/user copies input files to specified T3 (potential inputs: ESD,AOD,D^1PD)
On T3 group/user either generates reduced dataset for further analysis or performs final analysis on input data set

Choice depends strongly on capabilities of T3, size of input data sets, etc.

Also, expect some users to run D^3PD analysis jobs directly on T2 analysis queues

Analysis Requirements Study Group: Initial Estimate

Charge: estimate resources needed for the analysis & performance studies planned for 2009 & 2010

- considerable overlap with some activities of T3 Task Force
  (worked together, many of the same people)

Motivation: Management needs input from US physics community in order to make decisions/recommendations regarding current and future facilities

Received responses from 39/43 institutions

guessed for missing 4

Basic idea:

(1) predict (based on institutional polling) US based analyses (2009-2010)

(2) classify as: performance, physics-early, physics-late (sort by input stream)

(3) make assumptions about repetition rate (expect to vary with time)

(4) compute needed storage and cpu-s (using benchmarks)

early = months 1-4
late = months 5-11

Additional inputs & assumptions

# of data events

months 1-4: 2×10^6 s × 200 Hz = 4×10^8 events

months 5-11: 4×10^6 s × 200 Hz = 8×10^8 events

based on current CERN 2009/2010 plan: 1.2×10^9 evts with 1/3 before April 1 and 2/3 after April 1

streaming fractions*

Performance (ESD/pDPD)

egamma	muon	track	W/Z(e)	W/Z(m)	W/Z(T)/mET	gamjet	minbias
0.36	0.17	0.46	0.36	0.17	0.46	0.36	0.10

GB pointed out that
I missed the jet pDPD
(will fix in next draft)

Physics: 2009 (AOD/D^{1}PD)

egamma	muon	jet/mET
0.36	0.17	0.46

Physics: 2010 (AOD/D^{1}PD)

egamma	muon	jet/mET
0.36	0.17	0.46

* pDPD Streaming fractions were found on the pDPD TWiki in Jan 2009; they are no longer posted – regard as very preliminary

more inputs/assumptions

Institutional response summary:

Performance (ESD → D^{3}PD)

egamma	muon	track	W/Z(e)	W/Z(m)	W/Z(T)/mET	gamjet	minbias
26	28	18	16	19	15	9	15

Physics: 2009 (AOD → D^{3}PD)

egamma	muon	jet/mET
31	37	4

Physics: 2010 (AOD → D^{3}PD)

egamma	muon	jet/mET
19	18	3

Assume analyses begun in 2009 continue in 2010

In anticipation of the need for institutional cooperation (common DPDMaker, common D3PD):

minimal cooperation

Performance								Phys (2009)			Phys (2010)		
egam	mu	trk	W/Z(e)	W/Z(m)	W/Z(T)/mET	gamjet	minbias	egam	mu	jet/mET	egam	mu	jet/mET
15	15	8	8	10	8	5	8	16	18	2	10	8	1

maximal cooperation

Performance								Phys (2009)			Phys (2010)		
egam	mu	trk	W/Z(e)	W/Z(m)	W/Z(T)/mET	gamjet	minbias	egam	mu	jet/mET	egam	mu	jet/mET
7	7	4	4	5	4	3	4	8	9	1	5	4	1

supermax cooperation

Performance								Phys (2009)			Phys (2010)		
egam	mu	trk	W/Z(e)	W/Z(m)	W/Z(T)/mET	gamjet	minbias	egam	mu	jet/mET	egam	mu	jet/mET
2	2	1	0	0	1	1	1	1	1	1	1	1	1

Tier2 CPU Estimation: Results

compare needed cpu-s with available cpu-s:

kSI2k-s needed/10^{10}

	m1-4	m5-11
all analyses independent	8.1	17
minimal cooperation	4.2	8.5
maximal cooperation	2.1	4.3
supermax cooperation	0.4	1.0

kSI2k-s available/10^{10}

US Tier2s	4	13

Note that having every analysis make its own D^3PDs is **not** our model!

We have always known that we will need to cooperate

Available Tier2 cpu should be sufficient for 2009-2010 analyses

Storage plans (as I understand them)

Decided

Included in LCG pledge: T1: All AOD, 20% ESD, 25% RAW
each T2: 20% AOD (and/or 20% D^{1}PD ?)

2 copies of AODs/D^{1}PDs (data+MC) are distributed over US T2s

1 copy of ESD (data only) distributed over US T2s (expect only for 2009-2010)
(may be able to use perfDPDs in some cases)

D^{1}PDs initially produced from AODs as part of T0 production, replicated to T1s, T2s
D^{1}PDs will be remade from AODs as necessary on the T1

Not Decided

Final content of D^{1}PDs

Streaming strategy for D^{1}PDs (3 options under consideration - very active area of discussion)

Too early to make decisions about D^{2}PDs

Tier2 Storage Estimation: results

Recall from slide 20,
T2 beyond pledge storage must accommodate 1 set of AODs/D1PDs, 1 set of ESDs, & needs of users/groups (what we are trying to calculate)

subtracting AOD/D^{1}PD and ESD storage from beyond pledge storage, we find

Available for individual users:
m1-4: 0 TB
m5-11: 0 TB
17 TB if we assume only 20% ESD

no level of cooperation is sufficient here

We have insufficient analysis storage until Tier2 disk deficiency is resolved

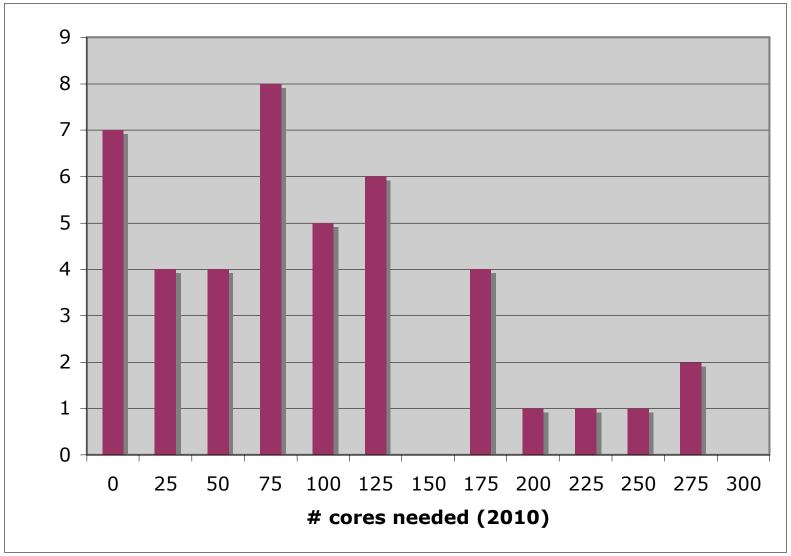
Tier3 CPU & Storage Estimation

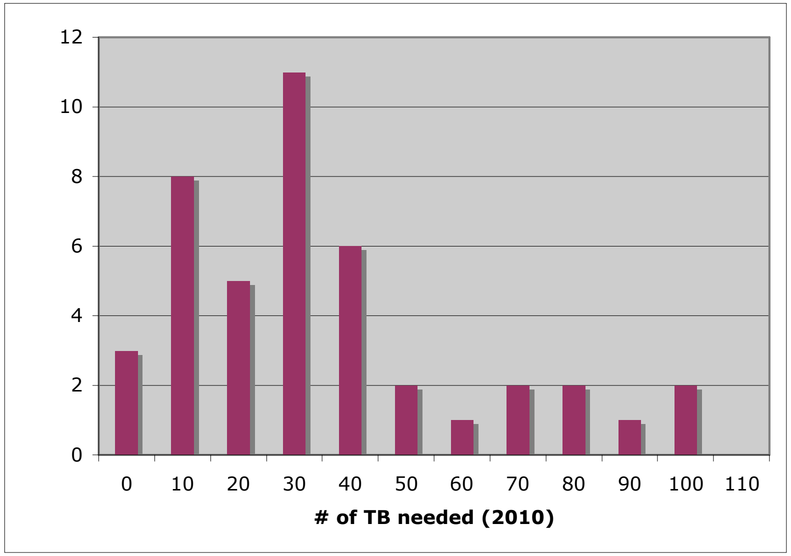
T3 CPU and disk calculations essentially same as T2 calculations

Sum # kSI2k-s needed for all analyses at institution (m1-4, m5-11)
Compare with # kSI2k-s available in 1 hour at institution's T3

Sum # TB needed for all analyses at institution, compare with # TB available

Each institution was sent estimate & comparison (week of 3/23): asked to correct/confirm – many responses so far (still incorporating)





most T3s have insufficient resources for planned activities - hope remedy from ARA/NSF

Scale of “missing” T3 resources

Assuming no ARA/NSF relief

Summing “missing” T3 resources (for both institutions with & without existing T3 hardware):

	# cores	TB
m1-4	447	161
m5-11	1018	366

Assuming 2 kSI2k/core

This sets the scale for a potential “single” US analysis facility (T3af)

Or, if we sort by geographic region:

	# cores (m1-4)	# cores (m5-11)	TB (m1-4)	TB (m5-11)
Western	111	290	40	105
Midwestern	36	113	13	41
Eastern	299	614	108	221

Already 2 de-facto T3af’s: Brookhaven Analysis Facility (BAF) (interactive T1)
Western Analysis Facility (WAF) at SLAC (interactive T2)

Exact sizes of these “facilites” difficult to pin down

T2→T3 data transfer

Tuning 0: No Tuning
Tuning 1: see Sergei’s talk

Need: sustained rate of hundred of Mbps

Single dq2_get command from ANL ASC

T2 Site	Tuning 0	Tuning 1
AGLT2_GROUPDISK	-	62 Mbps
BNL-OSG_GROUPDISK	52 Mbps	272 Mbps
SLACXRD_GROUPDISK	27 Mbps	347 Mbps
SWT2_CPG_GROUPDISK	36 Mbps	176 Mbps
NET2_GROUPDISK	83 Mbps	313 Mbps
MWT2_UC_MCDISK	379 Mbps	423 Mbps

Single dq2_get command from Duke

T2 Site	Tuning 0	Tuning 1
AGLT2_GROUPDISK	-	150 Mbps
BNL-OSG_GROUPDISK	38 Mbps	42 Mbps
SLACXRD_GROUPDISK		98 Mbps
SWT2_CPG_GROUPDISK	28 Mbps	? Mbps
NET2_GROUPDISK	38 Mbps	120 Mbps
MWT2_UC_MCDISK		173 Mbps

Readiness Summary

US Resources

Tier2 CPU – ok

Tier2 Disk – analysis will be negatively impacted if deficiency is not resolved

Tier3s –

most have insufficient resources for planned activities - hope for ARA/NSF

incorporating T3s into the US T1/T2 system (& testing them) is **urgent** priority

support for T3s expected to be a major issue (see tomorrow's talks)

Readiness Testing

Expect ~200 US-based analyses to start in 1st few months of running

By now T2 analysis queues are in continuous use but larger scale testing is needed

Increasingly expansive robotic & user tests are being planned

Not well tested: large scale data transfers from the T2s to T3s – **this is urgent**

Expectations: 1st 1-2 months

months 1-2: 1×10^6 s × 200 Hz = 2×10^8 events

We already know T2 cpu is sufficient

Storage needed for Beyond LCG Pledge (not yet counting user D3PD needs):

	Size/event	1 copy (m1-2)
ESD	700 kB	140 TB
perfDPD	NA	
AOD	170 kB	34 TB
D^{1}PD	30 kB	6 TB
Sim AOD	210 kB	126 TB

For now assuming no factor due to inclusive streaming

Assuming # of MC events = 3x # of data events

→ will not duplicate this [will count only against pledge]

This gives a total disk storage requirement (before user output needs) of

m1-4: 90 TB

Maybe ok if we're not using all pledged resources

Recommendation on dataset "distributions"

(my personal suggestion)

Recall

Included in LCG pledge: T1: All AOD, 20% ESD, 25% RAW
each T2: 20% AOD (and/or 20% D[1]PD ?)

2 copies of AODs/D[1]PDs (data+MC) are distributed over US T2s

1 copy of ESD (data only) distributed over US T2s (expect only for 2009-2010)
(may be able to use perfDPDs in some cases)

For 1st 2 months:

AOD/D[1]PD storage should not be a problem
(assume distribution is handled automatically ?)

For ESDs and pDPDs, should have **all** streams available on T2s

who decides these ?
when ?

Should we consider associating specific streams to specific T2s ?

Final Comment

Need to perform testing beyond the “20% of T2” level

In lead up to big conference, we will likely be asked to suspend production
And allocate all (or almost all) of T2s to analysis
- we need to test that we can support this (more intensive HC ?)

Backup Slides

Storage needed for Beyond LCG Pledge (not yet counting user D3PD needs)

		1 copy	1 copy
	Size/event	(m1-4)	(m5-11)
ESD	700 kB	280 TB	840 TB
perfDPD	NA		
AOD	170 kB	68 TB	204 TB
D^{1}PD	30 kB	12 TB	36 TB
Sim AOD	210 kB	252 TB	756 TB

For now assuming no factor due to inclusive streaming

Assuming # of MC events = 3x # of data events

→ will not duplicate this [will count only against pledge]

This gives a total disk storage requirement (before user output needs) of

m1-4: 180 TB
m5-11: 1080 TB
408 TB if we assume only 20% ESD

Available Tier2 resources & what’s left over for user D^3PD output

values from M. Ernst JOG Apr09 contribution

	2009 pledge	2009 installed	2009 users	2009 Q3 installed	2009 Q3 users
cpu (kSI2k)	6210	10,026	3816	11,220	5010
disk (TB)	2115	1964	-151	2540	425

US currently behind on 2009 disk pledge

assume m1-4 ~ 2009
m5-11 ~ 2009 Q3

recall ESD, AOD, D^1PD expectations
(previous slide)

m1-4: 180 TB
m5-11: 1080 TB
408 TB if we assume only 20% ESD

→

Available for individual users (D3PD output):
m1-4: 0 TB
m5-11: 0 TB
17 TB if we assume only 20% ESD

Expect actual allocation to be somewhat dynamic, monitored closely by RAC