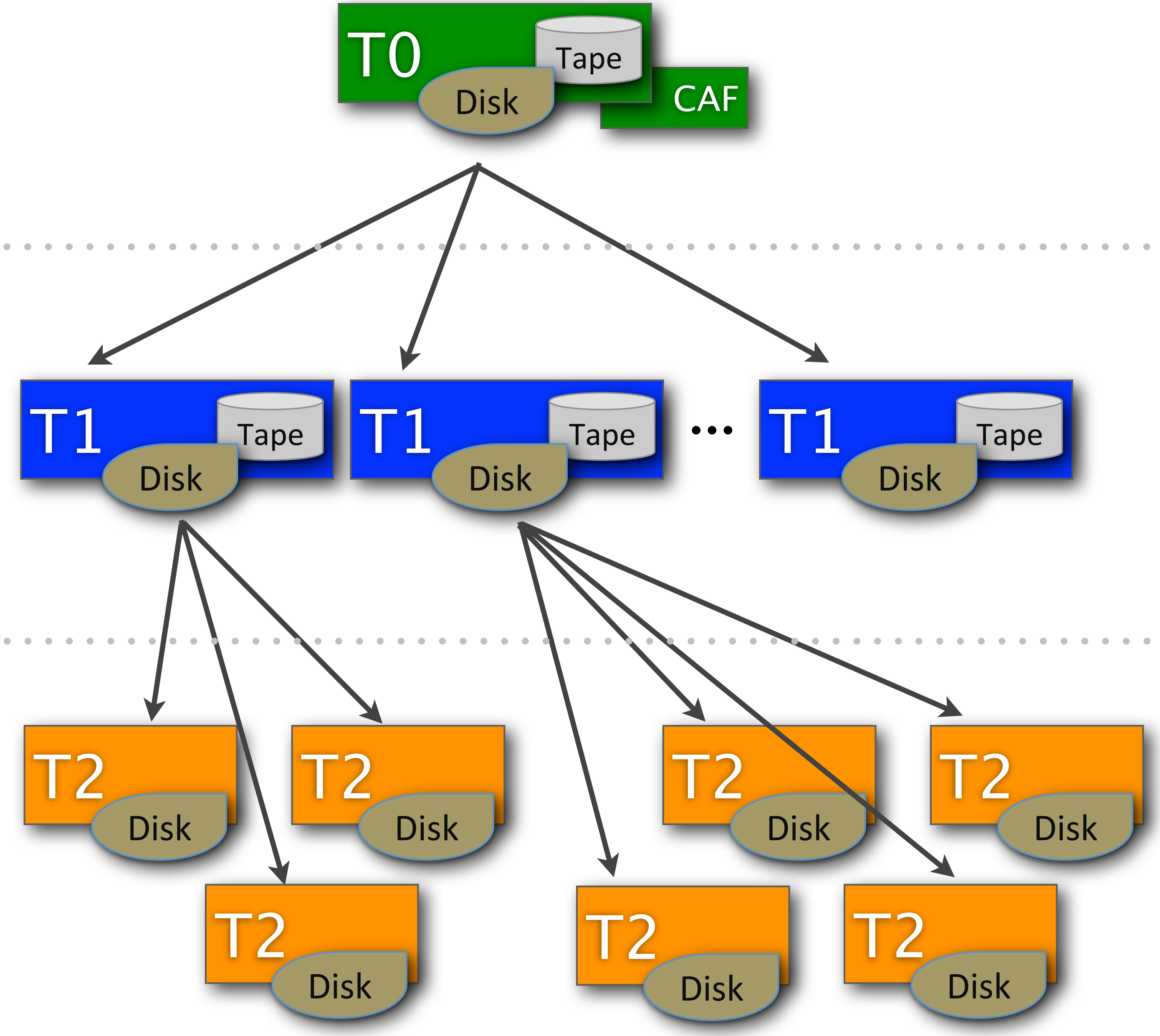
CMS Computing resource planning:

some food for thoughts in SuperB

Daniele Bonacorsi

[deputy CMS Computing coordinator – University of Bologna, Italy]

CMS Tiers and tasks



- Prompt processing
- Archival Storage
- Data export

- Organized processing
- Storage
- Data serving
- Production

- Unpredictable analysis
- Production

Resource planning

A realistic planning for computing resources (2-3 years ahead):

- precise - to the best knowledge and/or possible extrapolation
- granular - at a reasonable level
- easy to be kept up to date - with realtime feedback from computing operations

It should take into account:

- The machine plans
  - major impact on the overall resource planning and management
- The volume and type of data
  - not only from LHC but also the derived data (reprocessing, skimming, ...) and their relative importance
- The number and peculiarities of Tiers
  - 1 Tier-0 center, 7 Tier-1 sites, >40 Tier-2 sites, a growing number of Tier-3 sites
  - technical differences, that lead to different strengths and weaknesses
- The interaction with and input from other CMS projects
  - Mainly: Offline, Trigger, Physics, ...
- Any migrations to new tools and solutions (internal to CMS-Computing)
  - Avoid destructive interference: e.g. adoption of new solutions, once planned, must be folded in

Set this up is just the start: it needs to be maintained and regularly updated

- Any form/tool you prefer. CMS opted for just a unique resources spreadsheet.

Input parameters [1/2]

Live secs: **5.2 M secs**

- ~200 days of running at ~30% live time, spread over months

Expected average trigger rate in CMS: **~300, 400 Hz** in 2011-2012

- the system has demonstrated ability to record substantially higher trigger rates
- rate limited primarily by computing processing and analysis resources

Initial overlap factor between primary datasets: **~1.25**

- simulation using early versions of trigger menu for various luminosity scenarios

Tier-0 keep-up factor: **~0.75**

- fraction of the incoming trigger rate the T0 can process in real-time
  - if <1, T0 is still processing data in the time between fills
- important in T0 resource needs calculations
  - e.g. HI in 2010 showed capability of refilling for collisions in 3-4 hrs
  - if too low, it has the potential for not allowing the T0 to keep up with incoming data

More “facility” parameters ...

- number of Tier-{1,2} sites, Tier-{0,1,2} (+CAF) installed vs. pledges
  - for processing capacity, {archival,disk} storage
- efficiency for organized and analysis processing
- fraction of MC events compared to data
- {Data,MC} {RAW,RECO,AOD} fraction on disk T1
- T2 Space per User, # users /T2, Production space needed per T2, Passes through data at T2 /month, Disk Fill factor
- More...

Input parameters [2/2]

CAVEAT: Focus on pp here.
Same work is done for HI also

Additionally, relevant "CMS" parameters computed by expected pp PU scenarios

- events reco size and time are more correlated to PU conditions than to year

Sizes

- RAW evt size (data) was estimated by 2010 experience
- Simulation remains the same
- RECO and AOD sizes grow with the increase in the nb of interactions per crossing
- RECO size scrubbed by Offline, but still high w.r.t Computing Model
  - actions: migration to AOD, plus aggressive clean-up campaigns of older reconstruction versions

Times

- Reconstruction time scales roughly linearly with the nb of PU evts
  - number of tracks in the event as a significant driver of the reco speed

Parameter (pp)		Expected PU scenarios			
		0	4	8	16
RAW evt size (data)	[MB]	0.24	0.32	0.39	0.72
RAW evt size (MC)	[MB]	1.5	1.5	1.5	1.5
RECO evt size (data)	[MB]	0.26	0.39	0.53	0.81
RECO evt size (MC)	[MB]	0.36	0.49	0.63	0.91
AOD evt size (data)	[MB]	0.13	0.17	0.21	0.30
AOD evt size (MC)	[MB]	0.18	0.22	0.26	0.35
Repacker time	[HS06s]	3	6	7	8
RECO time (data)	[HS06s]	16	28	43	92
Gen-Sim time (MC)	[HS06s]	500	500	500	500
Redigi-Rereco time (MC)	[HS06s]	37	65	93	164

From these, you should be able to compute:

CAVEAT: the actual lists are more detailed and include more items

- total {pp,HI} evts /month and /yr
- {data,express} breakdown
- total MC evts /month and /yr

- T0: {pp,HI}{RECO,express,repacker,validation} CPU required
- T0: {pp,HI} VOboxes budget
- T0: CPU usability reduction factors
- T0: Analysis/Simulation resources
- T0: % of CPU pledge used
- T0: {RAW,RECO,AlcaReco} data volume on tape in {pp,HI}
- T0: predictions for tape available/used
- T0: Castor pools capacity (all buffers)

- CAF: {express, prompt-reco,MC,RelVal} data volume
- CAF: predictions for CPU available/used
- CAF: predictions for CAF {disk,tape}
- CAF T2 (in all details)

- T1: CPU needed for data reco for {current,previous} yr
- T1: CPU needed for MC redigi/rereco for {current,previous} yr
- T1: CPU needed for skims
- T1: CPU needed for new MC production rounds
- T1: % of CPU pledge used
- T1: {data,MC} RAW data volume and duplication factor
- T1: prompt-reco data volume
- T1: {data,MC} rereco data volume for {current,previous} yr
- T1: RECO data volume /month and delete factor
- T1: skims data volume /month and delete factor
- T1: {data,MC} {RAW,RECO, AOD} volume on tapes
- T1: {data,MC} AOD delete factors and turn factor
- T1: skims data volume on {disk,tapes}
- T1: predictions for tape available/used
- T1: {data,MC} {RAW,RECO, AOD} volume on disk
- T1: predictions for disk available/used

- T2: {data,MC} {RECO,AOD} on disk
- T2: {Production,User} Space on T2
- T2: total T2 disk available/used
- T2: {analysis,MC} processing needed
- T2: predictions for % of {T1,T2} needed for {analysis,MC}

This (and more) is what your computing infrastructure/sites need to know.

Tier-0 requests (example from last CRSG)

Monthly breakdowns available

CMS Tier-0, *300 Hz* CPU [kHS06]	Year		
	2011	2012	2013
Express	5	8	0
Prompt-RECO	44	53	0
Repack	3	3	0
AlCa workflow	1	1	0
RelVal/Validation	6	6	0
VOBoxes	9	11	0
Analysis	0	0	60
MC production	0	0	20
Total	**68**	**82**	**80**

CPU: requests do not grow in 2013

- CERN CPUs available in 2013 for ana/sim
- large integrated data sample, need to alleviate resource shortage at T2s

NOTE: CAF resources are in separate tables, not folded in here.

CMS Tier-0, *300 Hz* Disk [TB]	Year		
	2011	2012	2013
Streamer pool	500	500	0
Input Buffer	50	50	0
Export Buffer	248	248	0
Production space	200	200	0
Total	**998**	**998**	

Disk: breakdown into different buffers

- mostly, workflow-based

CMS Tier-0, *300 Hz* Tape [TB]	Year		
	2011	2012	2013
RAW (pp)	4317	5793	0
RECO (pp)	8633	10330	0
AlCaRECO (pp)	415	595	0
Total	**13365**	**16718**	

Tape: scales with nb of evts collected /yr

Tier-1 requests (example from last CRSG)

Monthly breakdowns available

CMS Tier-1, *300 Hz*	Year		
CPU [kHS06]	2011	2012	2013
Processing	130	160	160

CPU: requests driven by reco times, total volume of data, time allocated to complete a processing pass

CMS Tier-1, *300 Hz*	Year		
Tape [TB]	2011	2012	2013
RAW (data)	2452	4039	4039
RECO (data)	7037	8991	9243
AOD (data)	2224	3740	5001
RAW (MC)	10616	15544	17758
RECO (MC)	7433	14489	18107
AOD (MC)	3866	6837	8309
Skims	1811	2397	2473
Total	**35438**	**56036**	**64930**

CMS Tier-1, *300 Hz*	Year		
Disk [TB]	2011	2012	2013
RAW (data)	2200	2100	2100
RECO (data)	2551	2926	2926
AOD (data)	4089	7595	7108
RAW (MC)	1081	1585	2089
RECO (MC)	887	1297	2130
AOD (MC)	1992	3139	4888
Skims	1700	2300	2500
T1 temp disk	1600	2200	2700
Total	**16100**	**23141**	**26441**

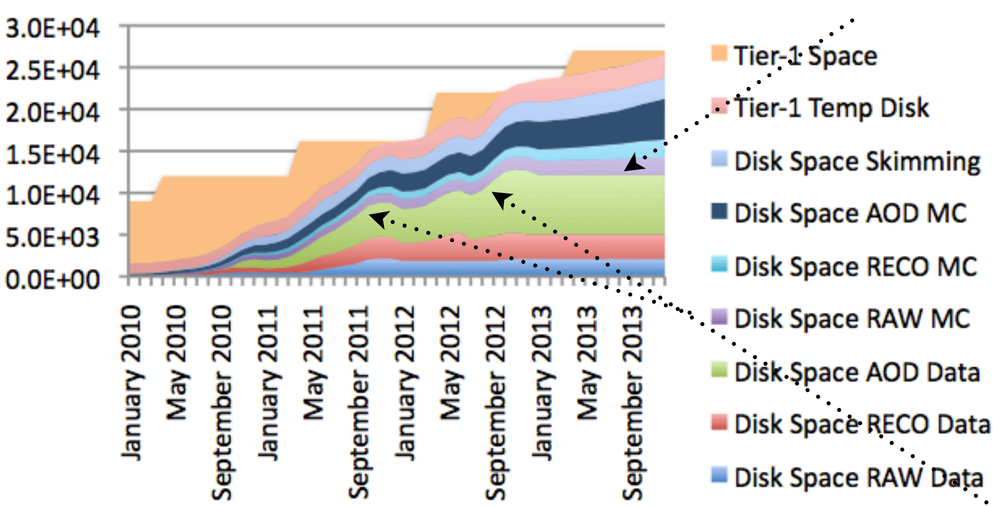
Disk: 1 copy of current RECO + current year's RAW + 10% of preceding RECO + 10% of all simulations

- ✦ No more need for full AOD replica sets at all T1s
  - reduced AOD size + full-mesh transfer model

Tape: stage-in back from tape whatever is not on disk

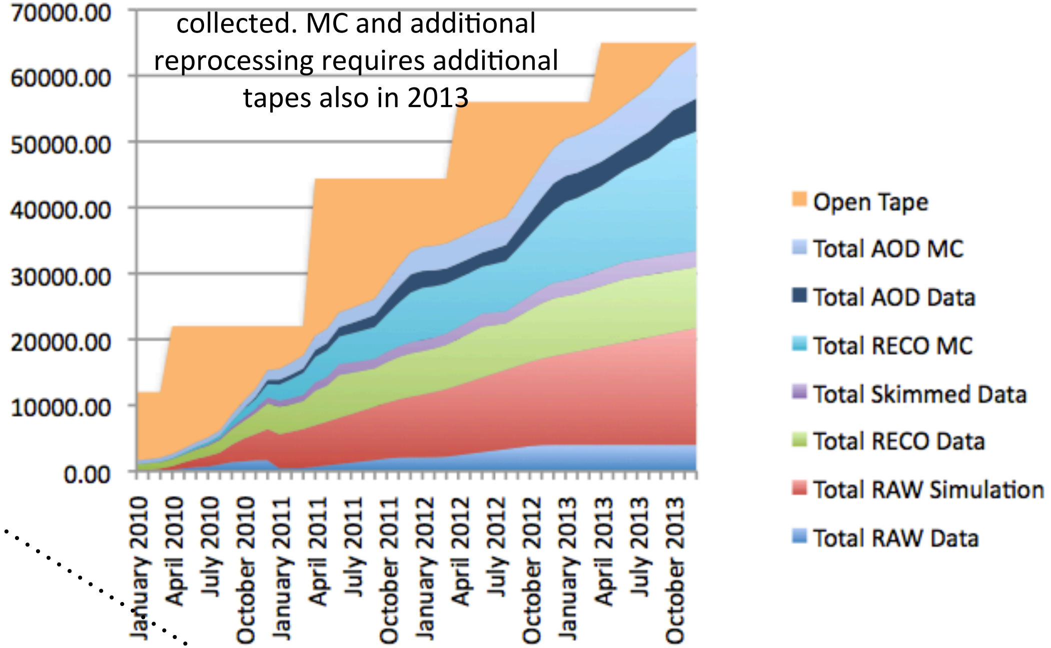
T1 resources evolution

Tier-1 Disk Storage



Disk increase flattens in 2013 in the absence of new data

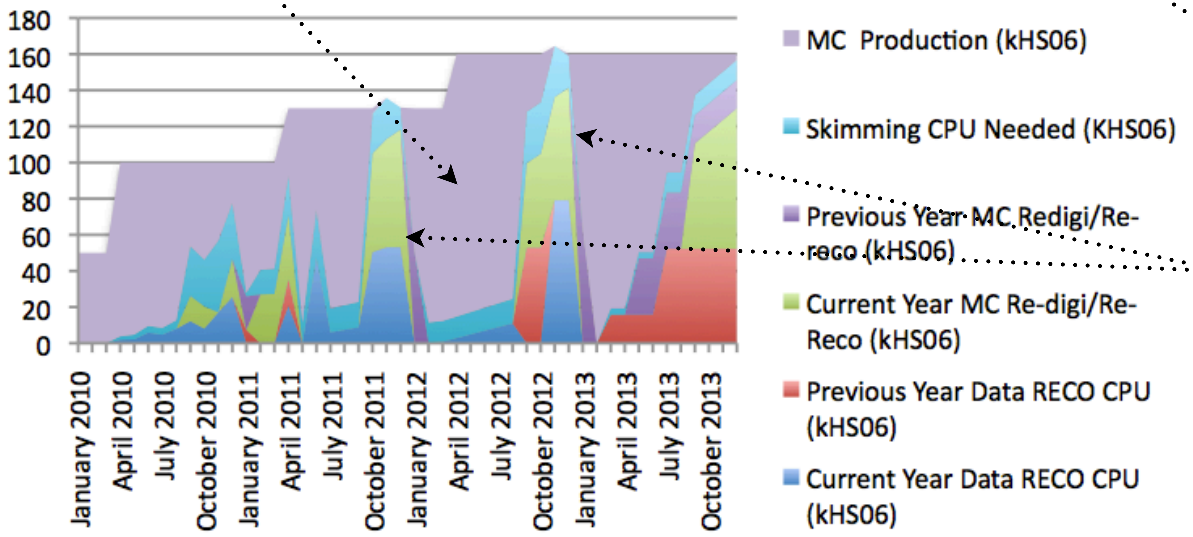
Tier-1 Tape Usage



Linear rise when new data is collected. MC and additional reprocessing requires additional tapes also in 2013

Profit of no-scheduled-processing periods to fill with MC production

Tier-1 Processing Resources



The changes in slopes corresponds to period of running interspersed with TS and deletion campaigns of old versions

Processing needs within a fixed time window produced peaks

(before Confs, and lined up with other LHC experiments).

Useful: you can proactively prepare for multi-VO processing periods (e.g. at T1s)

Tier-2 requests (example from last CRSG)

Monthly breakdowns available

CMS Tier-2, *300 Hz* **CPU** [kHS06]	Year		
	2011	2012	2013
Analysis	195	280	280
Production	120	120	120
Total	**315**	**400**	**400**

CMS Tier-2, *300 Hz* **Disk** [TB]	Year		
	2011	2012	2013
RECO (data)	2415	1000	1000
AOD (data)	1683	8500	7747
RECO (MC)	9270	3060	5000
AOD (MC)	3431	9862	10001
User Space on T2s	2400	3600	3600
Production Space on T2s	1000	1000	1000
Total	**20198**	**27022**	**28348**

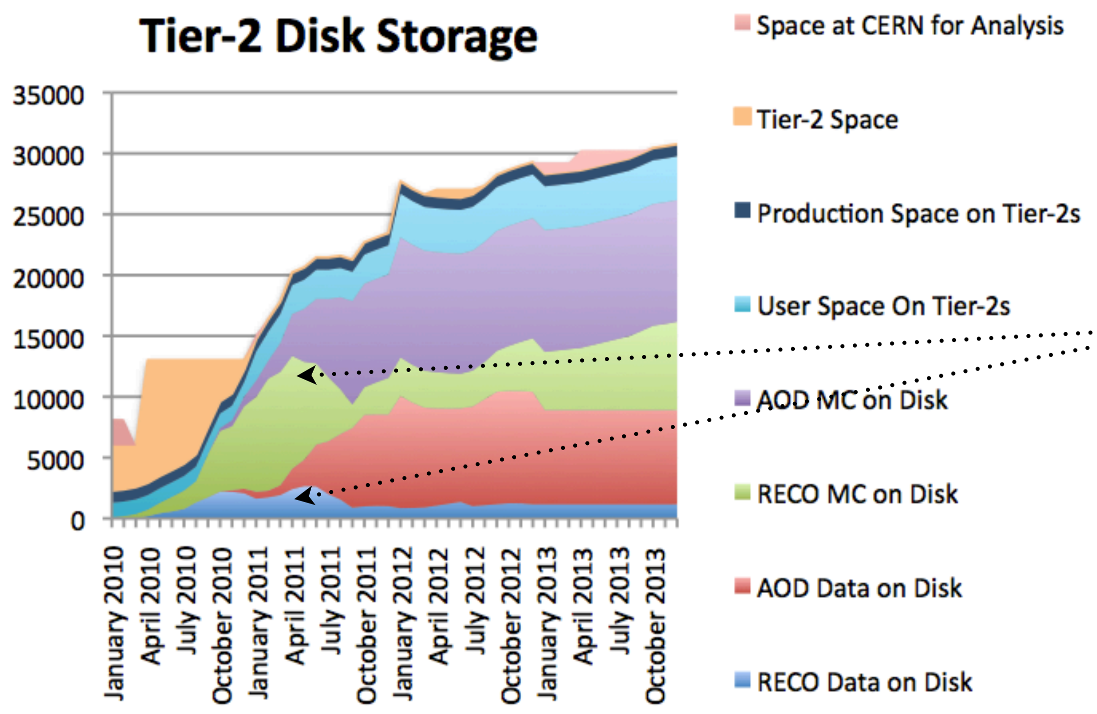
The total amount of resources for analysis scale strongly with the transition from RECO to AOD

- ✦ Smooth so far (see next slide)
- ✦ Assumption in the planning:
  - within 6-8 months from the start of 2011, 50% of the analysis activity would have been performed using AOD
  - This will increase eventually to 90% at the end of 2011

CPU: we moved part of production to T1s to free slots for distributed analysis at T2s

Disk: we assume to also use CERN/CAF resources for analysis in 2013

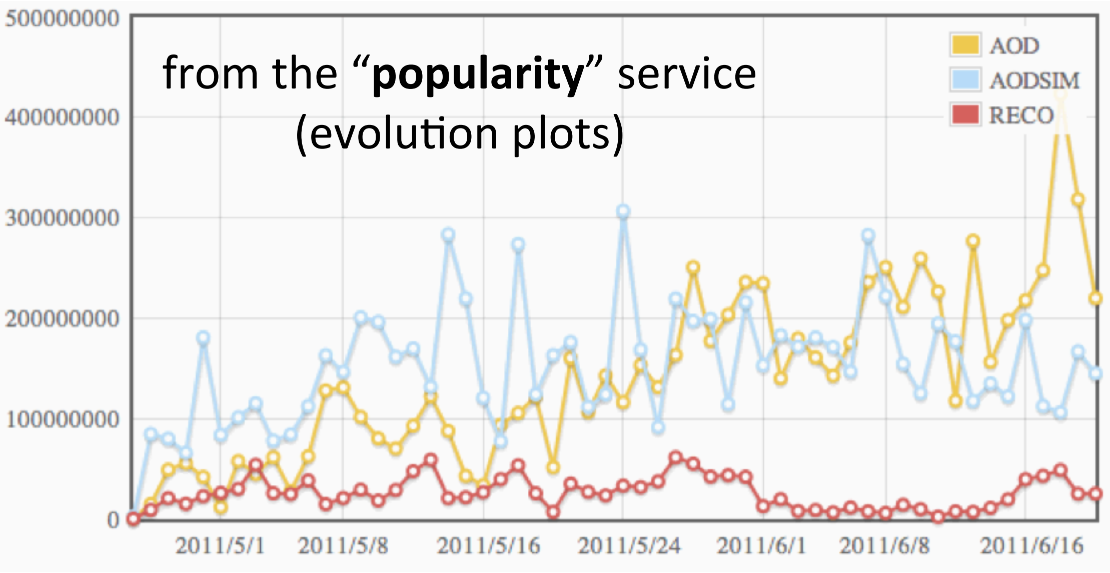
T2 resources evolution



The proportion of RECO increases at the beginning when most of the analysis is on this data format,and decreases as CMS transitions to AOD.

So far, smooth transition to AODs

- it needs to be closely monitored, though.



from the “**popularity**” service (evolution plots)

400 Hz

In 2012, we could have 5E33 cm^{-2} s^{-1} and 16 pp/crossing

- ✦ Bandwidth increase of ~100 Hz would significantly improve discovery potentials
  - e.g. Higgs to WW: Physics claims a 10% increase in dilepton efficiency by bandwidth increase of 75 Hz

The resource request increase varies between 10% and 30% higher than needed to support 300Hz

Supporting 400 Hz during 2011 would require some additional operation model changes

- ✦ the time for reprocessing would need to increase
  - freezing SW and calibrations earlier in the year to be ready for Confs
- ✦ allow high priority analyses (that benefit from the higher trigger rate) to have access to the limited processing resources

What matters in this context:

- ✦ the computing required to support 400Hz, as well as any scenario different from the 'reference' one, is relatively easy to extract
  - just vary some parameters in the resource planning spreadsheet

CMS Tiers, *400 Hz*	% increase over 300 Hz	
	2012	2013
T0 CPU	22%	22%
T0 disk	0%	0%
T0 tape (+HI)	10%	10%
CAF CPU	18%	18%
CAF disk	11%	11%
T1 CPU	25%	25%
T1 disk	23%	30%
T1 tape	7%	11%
T2 CPU	12%	12%
T2 disk	30%	20%

Outlook

We “used” CMS as an example in a data-taking context. Any experiment needs a realistic planning on computing resources

- as soon as possible. Needed for the infrastructure/sites to get prepared.

CMS uses a flexible tool with monthly breakdown on most categories

- it maintains the fundamentals of the CMS Computing Model (and its evolutions) and combines with our best understanding from the operational experience we have with collision data
- some work to set it up, some work to update and maintain it
- also open to the C-RSG: they used it and were able to recompute all CMS figures

Useful only if tuned real-time with Computing operations

1. start with clear assumptions and produce reasonable predictions
2. updated plans and/or actual resource utilization folded in month by month
   - e.g. hard data on utilization available and discussed in weekly Operations internal meetings
3. assumptions smoothly fade out, predictive power grows

New experiments may need something similar

- whatever works may be just fine. But don’t fail to prepare, or be prepared to fail.