CMS T1 / T2 Estimates

→ CMS perspective:

- Part of a wider process of resource estimation
- Top-down Computing Model -> real per-site estimates
- More detail exists than is presented in the Megatable

→ Original process:

- CMS had a significant resource shortfall (esp. T1)
- To respect pledges -> ad hoc descoping of CM

→ After new LHC planning

- New top-down planning roughly matches overall pledged resource
- Allow resource requirements at T1 centres to float a little
- Establish self-consistent balance of resources

→ Outputs

- Transfer capacity estimates between centres
- New guidance on balance of resources on T1/T2

Inputs: CMS Model

→ Data rates, event sizes

- Trigger rate: ~300Hz (450MB/s)
- Sim to real ratio is 1:1 (though not all full simulation)
- RAW (sim) 1.5 (2.0) MB/evt; RECO (sim) 250 (400) kB/evt
- All AOD is 50kB/evt

→ Data placement

- RAW/RECO: one copy across all T1, disk1tape1
- Sim RAW/RECO: one copy across all T1, on tape with 10% disk cache
  - How is this expressed in diskXtapeY formalism?
  - Is this formalism in fact appropriate for resource questions…?
- AOD: one copy at each T1, disk1tape1

Inputs: LHC, Centres

→ 2008 LHC assumptions

- 92 days of 'running' (does not include long MD periods)
- 50% efficiency during 'running'
- Practical implication: the T0 is 100% busy for this time
- Input smoothing at T0 required; assume queue < few days
- T0 output rate is flat during 'running' (straight from T0 capacity)
- More precise input welcomed + would be useful
  - Not expected to have strong effects upon most of the estimates

→ Efficiencies, overheads, etc

- Assume 70% T1/T2 disk fill factor (the 30% included in expt reqt)
- Assume 100% tape fill factor (i.e. any overhead owned by centre)
- T1 CPU efficiency back to 75 / 85% (chaotic/shed)

Centre Roles: T1

→ T1 storage reqts:

- Curation of assigned fraction of RAW
    - Assigned raw data fractions 1st fundamental input to T1/T2 process
- Storage of corresponding RECO / MC from associated T2 centres
    - Association of T1/T2 2nd fundamental input to T1/T2 process
- Hosting of entire AOD

→ T1 processing reqts:

- Re-reconstruction: RAW -> RECO -> AOD
- Skimming; group and end-user bulk analysis of all data tiers
- Calibration, alignment, detectors studies, etc

→ T1 connections

- T0 -> T1: Prompt RAW/RECO from T0 (to tape)
- T1 <->T1: Replication of new AOD version / hot data
- T1 -> T2; T2 -> T1 (see below)

Centre Roles: T2

- T2 storage reqts:
  - Caching of T1 data for analysis; no custodial function
  - Working space for analysis groups, MC production
- T2 processing reqts:
  - Analysis / MC production only
  - Assume ratio of analysis:MC constant across T2
- T1 -> T2 dataflow:
  - AOD: comes from any T1 in principle, often from associated T1
    - For centres without ‘local’ T1, can usefully share the load
  - RECO: must come from defined T1 with that sample
  - Implies full T1 -> T2 many-to-many interconnection
    - Natural consequence of storage-efficient computing model
- T2 -> T1 dataflow:
  - MC data always goes to associated T1

T1/T2 Associations

Centre	Streams	Associated T2
FZK	5	German T2, Poland, Switzerland
IN2P3	6	French T2, China, Belgium
PIC	2	Spain T2, Portugal
CNAF	7	INFN T2, Hungary
ASGC	5	Taipei, India, Pakistan
RAL	5	UK T2, Estonia, Finland
FNAL	20	US T2, Brazil
CERN		Russia, Ukraine

→ NB: These are working assumptions in some cases

→ Stream “allocation” ~ available storage at centre

Centre Roles: CERN CAF / T1

- CAF functionality
  - Provides short-latency analysis centre for critical tasks
  - e.g. detector studies, DQM, express analysis, etc
  - All data available in principle
- T1 functionality
  - CERN will act as associated T1 for RDMS / Ukraine T2
  - Note: not a full T1 load, since no T1 processing, no RECO serving
  - There is the possibility to carry out more general T1 functions
    - e.g. second source of some RECO in case of overload
  - Reserve this T1 functionality to ensure flexibility
    - Same spirit as the CAF concept
- CERN non-T0 connections
  - Specific CERN -> T2 connection to associated centres
  - Generic CERN -> T2 connection for service of unique MC data, etc
  - T1 <-> CERN connection for new-AOD exchange

Transfer Rates

→ Calculating data flows

- T0->T1: data rates, running period
  - Rate is constant during running, zero otherwise
- T1<->T1; total AOD size, replication period (currently 14 days)
  - High rate, short duty cycle (so OPN capacity can be shared)
  - Short repl. period driven by disk reqd for multiple AOD copies
- T1->T2: T2 capacity; refresh period at T2 (currently 30 days)
  - This gives the average rate only - *not a realistic use pattern*
- T2->T1: total MC per centre per year

→ Peak versus average (T1 -> T2)

- Worst-case peak for T1 is sum of T2 transfer capacities
  - Weighted by data fraction at T1
- Realistically, aim for: average_rate < T1_capacity < peak_rate
- Difference between peak / avg is uniformly a factor 3-4
- Better information on T2 connection speeds will be needed

Outputs: Rates

	FZK	IN2P3	PIC	CNAF	ASGC	RAL	FNAL	CERN
OPN in	91	86	94	87	90	88	105	95
OPN out	48	78	30	74	55	67	215	263
T2 in avg	7	15	6	11	9	13	30	9
T2 out avg	63	96	53	94	76	84	248	47
T2 out peak	314	336	203	236	213	329	953	208

→ Units are MB/s

→ These are *raw rates*: no catchup (x2?), no overhead (x2?)

- Potentially some large factor to be added
- A common understanding is needed

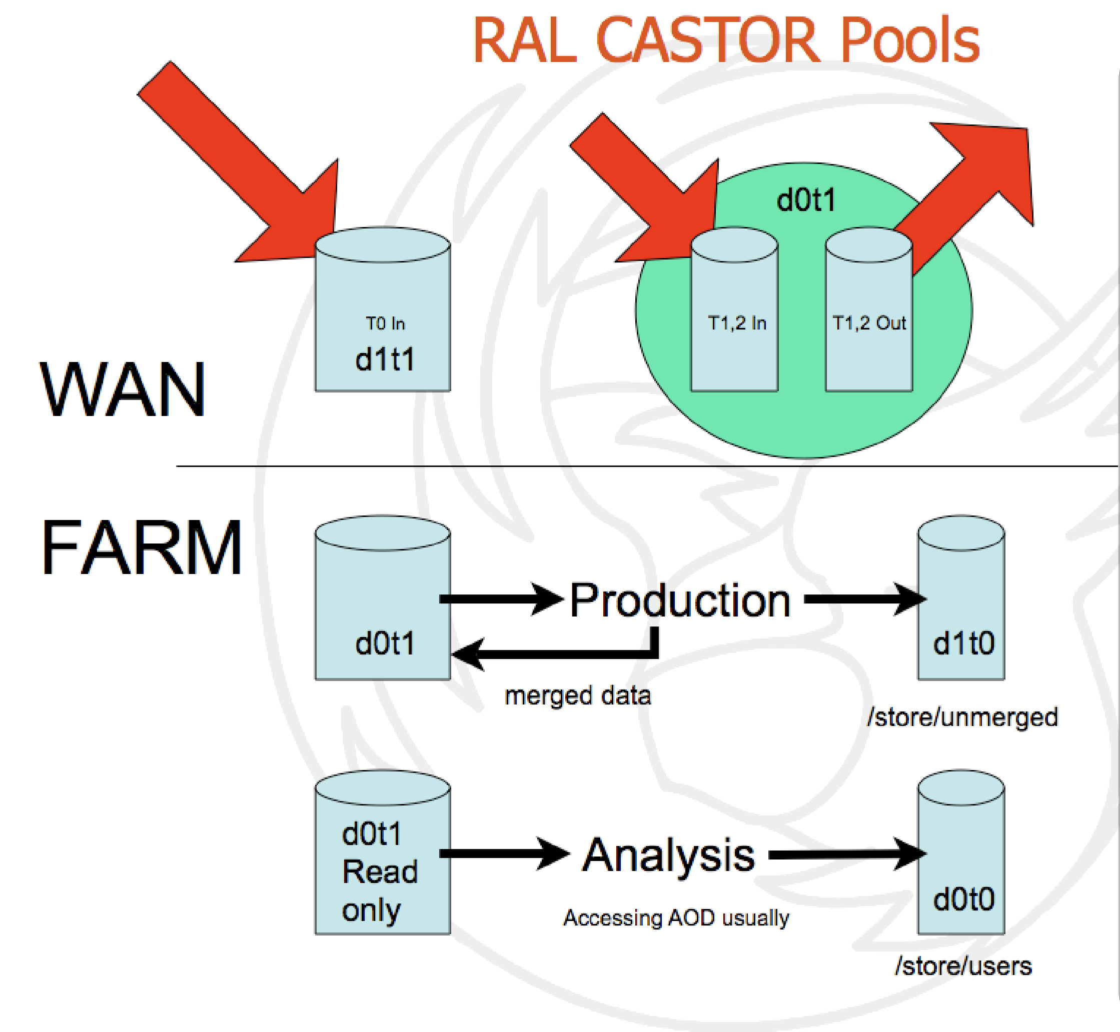
→ FNAL T2-out-avg is around 50% US, 50% external

Outputs: Capacities

	P.CPU	P.Disk	P.Tape	CPU	Disk	Tape	Tr Buf	"Resource
	kSI2k	TB	TB	kSI2k	TB	TB	TB	
FZK	1200	650	900	999	647	1025	21	97%
IN2P3	1490	780	1180	1616	758	1554	28	106%
PIC	760	350	835	620	415	581	18	96%
CNAF	1925	875	735	1541	822	1546	27	99%
ASGC	1530	675	585	1139	657	1134	23	98%
RAL	1330	620	1280	1379	673	1320	25	106%
FNAL	4256	1986	4700	4456	1916	4458	60	99%

- → "Resource" from a simple estimate of relative unit costs
  - CPU : Disk : Tape at 0.5 : 1.5 : 0.3 (a la B. Panzer)
- → Clearly some fine-tuning left to do
  - But is a step towards a reasonably balanced model
- → Total is consistent with top-down input to CRRB, by construction
- → Storage classes are still under study
  - Megatable totals are firm, but diskXtapeY categories are not
  - This may be site-dependent (also, details of cache)

e.g. RAL Storage Planning



Data Types:

AOD - summary data. On disk always. Periodically in transfer to T1,2 at high rate.

RECO - Produced at T0, reprocessed at T1's 3* per year. Should be on disk.

SIMRECO - Produced at T2's, should be on tape at least.

RAW - Data from detector, transferred from T0. On disk.

SIMRAW - Produced at T2's, should be on tape at least.

RECO, SIMRECO, RAW and SIMRAW are custodial data and must be stored on tape.

Comments / Next Steps?

- → T1 / T2 process:
  - Has been productive and useful; exposed many issues
- → What other information is useful for sites?
  - Internal dataflow estimates for centres (-> cache sizes, etc)
  - Assumptions on storage classes, etc.
  - Similar model estimates for 2007 / 2009+
  - Documentation of assumed CPU capacities at centres
- → What does CMS need?
  - Feedback from sites (not overloaded with this so far)
  - Understanding of site ramp-up plans, resource balance, network capacity
  - Input on realistic LHC schedule, running conditions, etc
  - Feedback from providers on network requirements
- → Goal: detailed self-consistent model for 2007/8
  - Based upon real / guaranteed centre, network capacities…
  - Gives at least an outline for ramp-up at sites, global experiment
  - Much work left to do…

Backup: Rate Details

		FZK	IN2P3	PIC	CNAF	ASGC	RAL	FNAL	CERN
AOD exch 2008									
	Size	6.00048	7.200576	2.400192	8.400672	6.00048	6.00048	24.00192	0
	Sim size	3.685386784	8.557931637	3.652552651	6.574790494	5.072080546	7.449602483	19.31099211	5.701463296
	Rate out	48.04	78.17	30.02	74.28	54.92	66.72	214.85	28.28
	Rate in	91.21	86.19	94.21	86.83	90.06	88.09	63.41	94.50
FEVT transfer 2008									
	Size	210.0168	252.02016	84.00672	294.02352	210.0168	210.0168	840.0672	0
	Rate	26.25	31.50	10.50	36.75	26.25	26.25	105.00	0.00
	(Out rate)								262.50
RECO transfer 2008									
	Size	30.0024	36.00288	12.00096	42.00336	30.0024	30.0024	120.0096	
	Rate	24.80	29.76	9.92	34.73	24.80	24.80	99.21	41.79
Rate 2008 (MB/s)									
	OPN in	91.21	86.19	94.21	86.83	90.06	88.09	105.00	94.50
	OPN out	48.04	78.17	30.02	74.28	54.92	66.72	214.85	262.50
	T2 in avg	7.00	15.00	6.00	11.00	9.00	13.00	30.00	9.00
	T2 out avg	63.00	96.00	53.00	94.00	76.00	84.00	248.00	47.00
	T2 out peak	314.00	336.00	203.00	236.00	213.00	329.00	953.00	208.00

Backup: Capacity Details

		RAW disk	RECO disk	Sim disk	Sim tape	Ana store	Transfer Buf	Old RECO	AOD disk	RECO repl.	Sim frac	Tot tape	Tot disk	No_sim tape	No_sim disk
		TB	TB	TB	TB	TB	TB	TB	TB			TB	TB	TB	TB
T1 storage 2008															
	FZK	180.01	30.00	17.69	246.92	50	0.00	6.00	144.01		6.14%	1025.06	647.13	736.55	435.30
	IN2P3	216.02	36.00	41.08	573.38	60	0.00	7.20	144.01		14.26%	1553.83	757.99	883.85	489.52
	PIC	72.01	12.00	17.53	244.72	20	0.00	2.40	144.01		6.09%	580.57	414.56	294.62	272.66
	CNAF	252.02	42.00	31.56	440.51	70	0.00	8.40	144.01		10.96%	1545.89	821.84	1031.16	543.73
	ASGC	180.01	30.00	24.35	339.83	50	0.00	6.00	144.01		8.45%	1133.62	656.64	736.55	435.30
	RAL	180.01	30.00	35.76	499.12	50	0.00	6.00	144.01		12.42%	1319.75	672.94	736.55	435.30
	FNAL	720.06	120.01	92.69	1293.84	200	0.00	24.00	144.01		32.18%	4457.99	1916.01	2946.18	1248.52
	CERN	0.00	0.00	27.37	382.00	0	0.00	0.00	144.01		9.50%	446.35	273.71	0.00	164.23
	Total	1800.14	300.02	288.02	4020.32	500.00	0.00	60.00	1152.09		1.00	12063.06	6160.82	7365.45	4024.55

		Mevts	Sim Mevts	Tot Mevts	Tot s Mevts	re-RECO	sim-reRECO	Selection	Calib	Tot with effs	Nxt_yr disk
						kSI2k	kSI2k				
T1 CPU 2008											
	FZK	120.01	73.71	140.23	86.13	285.41	175.29	327.48	15.00	998.64	150.01
	IN2P3	144.01	171.16	168.27	199.99	342.49	407.06	532.79	18.00	1616.21	156.01
	PIC	48.00	73.05	56.09	85.36	114.16	173.73	204.64	6.00	619.56	132.01
	CNAF	168.01	131.50	196.32	153.65	399.58	312.73	506.32	21.00	1541.09	162.01
	ASGC	120.01	101.44	140.23	118.53	285.41	241.25	374.36	15.00	1138.75	150.01
	RAL	120.01	148.99	140.23	174.09	285.41	354.34	454.74	15.00	1378.97	150.01
	FNAL	480.04	386.22	560.91	451.28	1141.64	918.52	1464.40	60.00	4456.26	240.02
	CERN	0.00	114.03	0.00	133.24	0.00	271.19	192.77	0.00	576.07	120.01