Table of Contents

How to update your HS06 Conversion Factor:

- Sites can update their APELNormalFactor by submitting a pull request here: https://github.com/opensciencegrid/topology/.
- Documentation for how to do that can be found here: https://opensciencegrid.org/docs/common/registration/#modifying-existing-resources.

DRAFT Deployment Guidance for 2018

We don't know what the 2018 hardware budget will be yet, but we in the University Facilities area have been thinking about priorities for this year's hardware deployment, and we would like to get feedback from each of the sites.

The storage and processing pledges don't go up much in 2019, which is a shutdown year for the LHC, relative to 2018. We use the current calendar year's hardware budget to provision for the following year's pledge, which activates on April 1st each year. Given that most hardware will not be deployed until late in the calendar year, close to the start of the two-year shutdown, we do not see a need to massively increase resources at the sites at this time. We are also running more than 50% of the total CMS production activity in the U.S., which is problematic for the operations program when talking to funding agencies. Instead, we can focus on reinforcing site infrastructure to better support current operations and prepare for future needs.

Our consensus is that since we have enough capacity to cover the 2019 estimated processing pledge times three (250,000 HS06 in total for the seven U.S. Tier-2 sites, while the current total deployment is 773,650 HS06), that we prioritize 2018 purchases as follows:

- Maintain (but not increase) current processing capacity, with respect to any planned worker node retirements.
- Deploy the storage pledge plus a buffer of +1 PB to enable U.S. physics analysis. The April 1, 2018 storage pledge to CMS is 2,500 TB and this increases to 2,800 TB for April 1, 2019, and so should be deployed by the end of 2018. Therefore, the storage deployment goal for April 1, 2018 is 3,500 TB and for the end of calendar 2018 is 3,800 TB.
- Take the opportunity to make any infrastructure upgrades rather than increase capacity, i.e. networking, cooling, etc.
- Spend any remaining funds on storage or processing, as you see fit or according to any pricing advantages you are facing. We recommend storage purchases over processing, which is already very well-provisioned.

Please send us your comments or criticisms about this guidance. Sites which lease rather than own equipment may have quite different opinions or constraints. We'd like to hear them too.

2018 CMS Tier-2 resource request, per site:

	2018	2019
CPU HS06	32,143	35,714
Storage TB	2,500	2,786

Purchasing Power Estimates Estimates for 2018:

- $10/HS06 processing
- $55/TB storage

Nominal site in January 2018:

- 110,500 HS06
- 3,300 TB

or a ratio of 33.5 HS06/TB, somewhat more in favor of processing than the previous year. Assuming 10% of storage and 5% of processing is retired each year, the steady-state replacement cost using the above

purchasing power amounts is 330TB * $55/TB + 5,525 HS06 * $10/HS06 = $73.400.

U.S. CMS Tier-2 Facilities Deployment Status

This table is used to track deployment status at the U.S. CMS Tier-2 sites. Sites should update the table on the page as they deploy new equipment. This page reflects the "true" deployments paid for from U.S. CMS Tier-2 funds, as distinct from what we report to CMS as what we've pledged.

This page had been deprecated. Please update the new google doc here.

Site	CPU (HS06)	Batch Slots	Physical Cores	Space for hosting (TB)	WAN (Gb/s)	Last update
T2_US_Caltech	100408	9096	4548	4120	100	06/27/18
T2_US_Florida	143942	8398	8398	4076	100	05/15/19
T2_US_MIT	110917	10616	10616	4000	100	11/09/19
T2_US_Nebraska	112395	10304	5152	5000	100	08/08/19
T2_US_Purdue	127213	8444	5952	4600	100	11/26/19
T2_US_UCSD	122247	10757	5456	3328	80	6/24/19
T2_US_Wisconsin	128151	12888	6444	4100	100	02/04/19
Total HEP Sites	845273	70503	46566	29224		
T2_US_Vanderbilt	unk	4396	2198	3200	100	03/21/17

Notes:

- Florida Slots are reduced due to Florida HPC 5 year hardware retirement policy. We are negotiating with HPC to extend the retired cores, but very unlikely to get the extention.
- Purdue compute nodes adhere to a strict 5 year hardware retirement policy. Some Purdue storage nodes run outside their warranty period, 170 TB of the above 3900 TB is out of warranty.
- Caltech Total Available storage is 4120 TB. 7281 TB RAW disk space for HDFS (3640 TB usable - replication 2). In addition to HDFS, Caltech maintains 330 TB xrootd cache (no replication) and a 300 TB Ceph development instance. (CEPH Instance Replication is changing and for acounting purposes factor of 2 is taken. We test different errasure codings and this makes always different available space. This space is not used for Ops). Overall usable 3640 + 330 + 150 = 4120TB. Overall RAW 7281 + 330 + 300 = 7911 TB RAW
- UCSD has a flexible replication scheme, plus a 288 TB xrootd cache which is in production.

Storage breakdown study - May 2019

All numbers are in units of TB = 10^12 bytes. Total capacity for hosting data in the main storage system should add up to the sum total usage in /store plus free space available for writing data. Total capacity includes the capacity for hosting data in the main storage system plus caches and any other space.

Site	Raw Disk	Capacity	User	Group	Unmerged	/store	Buffer	Free	Caches	Other	TOTAL	Hos
T2_US_Caltech	7911	3090	27	92	40	2402	550	588	330	0	3720	2
T2_US_Florida	5923	3939	199	2	28	2750	207	1189	0	0	3939	2
T2_US_MIT	8000	3800	640	0	208	3641	200	159	0	0	3800	3
T2_US_Nebraska	7070	3400	146	108	123	2934	336	270	0	74	3400	3
T2_US_Purdue	8806	3721	194	230	94	3244	413	406	0	41	3721	3
T2_US_UCSD	5500	2000	168	348	30	1572	444	432	1000	60	3060	2
T2_US_Wisconsin	8200	3590	555	13	83	2650	510	740	0	50	3640	2
Totals	51410	23540	1929	793	606	19193	2660	3784	1330	225	25280	20

- "Raw disk" = all un-replicated disk, purchased with Tier-2 funds and still in use

- "Capacity" = usable space in the main storage element, after subtracting any operationally necessary buffers
- "User" = usage in /store/user namespace
- "Group" = usage in /store/group namespace
- "Unmerged" = usage in /store/unmerged and /store/temp namespaces. Not broken out in some cases but the /store number is correct.
- "/store" = total PhEDEx space plus /store/user, /store/group, and /store/unmerged, i.e. all of /store namespace.
- "Buffer" = amount of free replicated disk space needed for good performance of the storage system
- "Free" = free space actually available for writing new replicated data in the main storage system
- "Caches" = total space in xrootd caches (generally not replicated)
- "Other" = see explanation below, typically user space not under /store.
- "TOTAL" = Storage capacity in the main storage system plus any caches
- "Hosted" = total in /store and any caches
- FTE+* = count of any un-costed labor which supports the Tier-2 program but not paid for by the operations budget

- Site Notes:
    - Caltech has a 15% buffer for hadoop. There is an empty 300TB ceph storage system included in the total but not in free space in the main storage system. They had 0.2 FTE from campus computing in 2018 for the admin transition
    - Florida uses a RAID system so that the ratio of Raw to usable space is 720/504. A buffer of approximately 5% free space is needed for good performance. Approximately 0.1 FTE non-costed labor. Site has also ordered an additional 500TB usable for 2019, partly with leftover 2018 funds. Did not ask for /store/unmerged but the /store number is correct. Some numbers were reported to us in TiB not TB and are adjusted in the table.
    - MIT: According to PhEDEx, there are 3,520 TB of data hosted of which 627 TB is in the heavy-ions group which should not be counted for the HEP program, leaving 2,893 TB. /store/user folder is 1700 TB of usable storage. HI users take 1060 TB the rest (640TB) is in HEP. Heavy Ion purchased 2544 TB of usable storage, 5088 TB raw. the only US Tier-2 center of LHCb will be included into our setup in the next months. The purchase (~$150k) is being prepared and should go out this week. This will add opportunistic resources. Normally they do not want to go over 95% filling the storage. Right now at 81% full.
    - Nebraska: [not confirmed] For things like ancillary system administration, network maintenance and operation, assistance with opportunistic resources at HCC, and other primarily personnel contributions not paid for by CMS, 0.5 FTE is a reasonable guess. Operational buffer is 9.5% of the main storage system. Replication factor of 2.19 is slightly higher than 2x since they use treble replication in some spaces like unmerged. "Other" includes 55TB for LIGO, 20TB for Brian, 6TB for dteam, and 4TB for cvmfs, plus about 123TB in unmerged replicated 3x, so
    - Purdue: Has 60TB of raw disk not deployed that is under warranty. Performance buffer is usually 10-15%. We estimate ~0.23 FTE of support from different research computing personnel whose salaries are not covered by the operations program. Most notable is the 0.1 FTE contribution of Laura Theademan, a research computing program manager who provides CMS project management support. Other contributions include HPC engineers (0.05 FTE) and Data Center Management (0.05 FTE) teams supporting the Community Cluster program. We also receive support from our networking department (0.02 FTE) and research computing user support staff (0.01 FTE). Purdue CMS equipment purchases are exempt from facilities and administrative costs.
    - UCSD has 60TB of raw disk for transient user space, listed under "other". There is a 10% free disk buffer for performance. Includes 240 TB of raw disk not fully deployed in the main storage system but purchased with 2018 funds.
    - Wisconsin has 50 TB of replicated disk for transient user space, listed under "other". The un-costed labour that supports the site amounts to 0.05 - 0.075 FTE.

Opportunistic Computing Resources

Information from Sites

Many Tier-2 sites allow (often) seamless usage for CMS of opportunistic computing resources through the existing Tier-2 infrastructure. List here please the opportunistic computing resources (CPU and/or storage) available at each site over and above the values in the table above.

Site	Batch Slots	Space for hosting (TB)	Last update
T2_US_Caltech	0*		01/12/2017
T2_US_Florida	~4,000		02/19/2015
T2_US_MIT	1500	900	02/19/2015
T2_US_Nebraska	~4,000		05/11/2016
T2_US_Purdue	29,660**		02/1/2018
T2_US_UCSD	6000***		12/03/2015
T2_US_Vanderbilt	800		05/25/2016
T2_US_Wisconsin	~1,500		07/14/2016
Total	>23,900	900	

(*) Disabled until 2017 HEP cluster upgrade.
(* * *) Comet at SDSC with allocation

** Available via PBS standby queue with 4 hour walltime

Discovery of Opportunistic Batch Slots with the Global Pool

Between March 11-27 2015, the machine name and number of CPUs of any worker node where a glidein was run in the glideinWMS Global Pool was recorded and analyzed to find the maximum number of uniquely identifiable batch slots at any given site. The results for the U.S. Tier-2 sites are given in the table below, with an overall global summary. Generally, opportunistic machines at Purdue and Florida were easily accessible.

Site	Batch Slots
T2_US_Caltech	5,685
T2_US_Florida	8.782
T2_US_MIT	5,660
T2_US_Nebraska	9,688
T2_US_Purdue	22,458
T2_US_UCSD	4,643
T2_US_Vanderbilt	2,598
T2_US_Wisconsin	7,252
All Sites	205,080
All Tier-1 Sites	43,229
All Tier-2 Sites	141,720

The measurement was repeated between September 23 and November 2, 2015. The rossmann cluster was excluded from the counting at Purdue since it was retired during the month of October.

Generally about 50% of the opportunistic resources are available to CMS on any given day.

Other resources:

- Comet at UCSD?

Site	Total Batch Slots	Less Purchased	Opportunistic	Reported Opp. Feb. 2015
T2_US_Caltech	6164	-5780	384	200
T2_US_Florida	10194	-4126	6068	4000
T2_US_MIT	7256	-5200	2056	0
T2_US_Nebraska	9557	-5840	3717	3000
T2_US_Purdue	16217	-6436	9781	9200
T2_US_UCSD	5329	-5256	73	0
T2_US_Vanderbilt	5274	-4396	878	0
T2_US_Wisconsin	10573	-7860	2713	1500
Total	70564	-44894	25670	17900

Cluster Management

Poll on configuration and cluster management tools, May 2015:

Site	Tools
T2_BR_SPRACE	
T2_BR_UERJ	Kickstart and Ansible(Red Hat's resources)
T2_US_Caltech	Foreman 1.16 and Puppet 5
T2_US_Florida	Florida HiperGator SIS system (Image)
T2_US_MIT	
T2_US_Nebraska	Cobbler 2.6 and Puppet 4.8 (opensource)
T2_US_Purdue	Foreman 1.2 and Puppet 2.7
T2_US_UCSD	Foreman (Latest) and Puppet 3.7.5 (4.x)
T2_US_Vanderbilt	CFEngine 3.5 (cluster + storage) Saltstack 2016.11.3 (storage)
T2_US_Wisconsin	Puppet 3.7.3, ganeti (VM cluster)

Responsible: JamesLetts

JamesLetts - 2016-11-07

This topic: CMSPublic > USCMSTier2Deployment
Topic revision: r153 - 2020-04-01 - JamesLetts