Storage history & Ceph Plans at US Tier-1

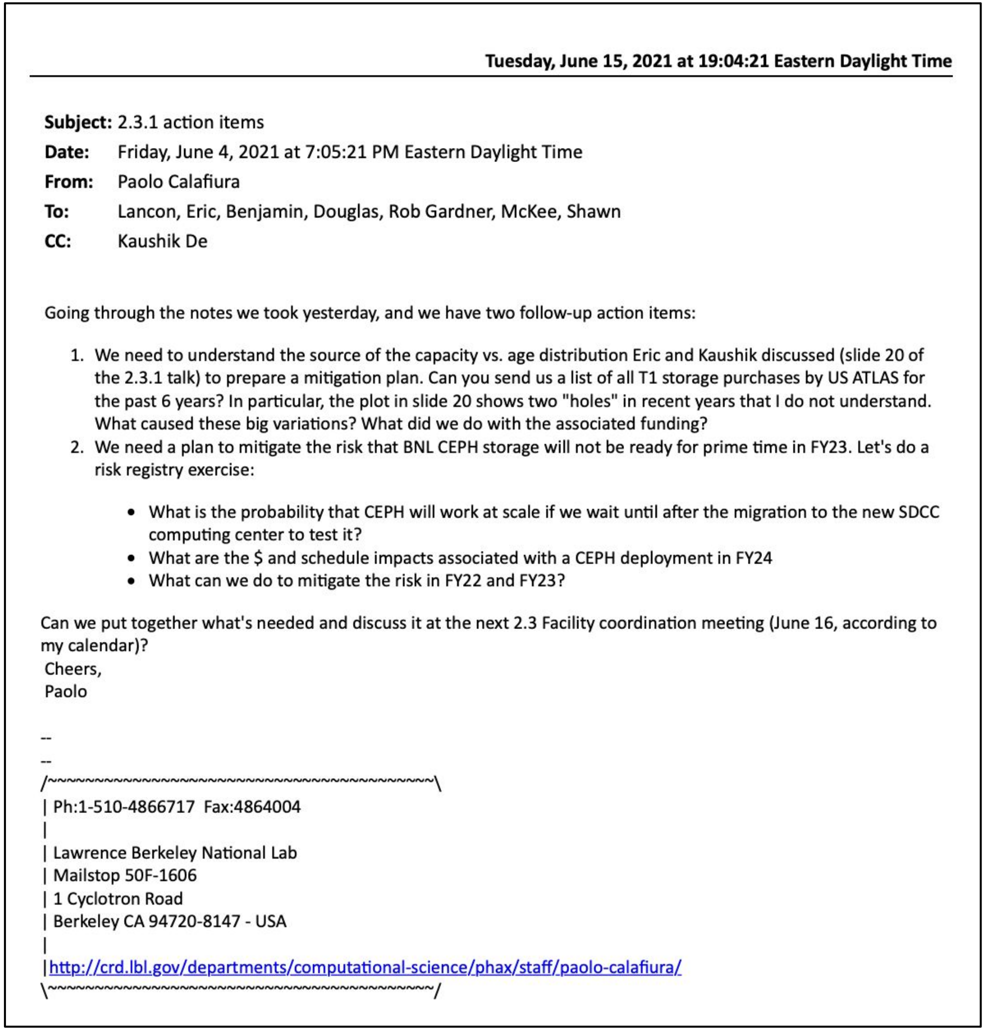
June 16, 2021

Pre-scrubbing WBS 2.3(.1) follow up

Brookhaven National Laboratory

Paolo's pre-scrubbing follow-up questions

1. Disk storage purchase history
2. Ceph plan



Tuesday, June 15, 2021 at 19:04:21 Eastern Daylight Time

Subject: 2.3.1 action items
Date: Friday, June 4, 2021 at 7:05:21 PM Eastern Daylight Time
From: Paolo Calafiura
To: Lancon, Eric, Benjamin, Douglas, Rob Gardner, McKee, Shawn
CC: Kaushik De

Going through the notes we took yesterday, and we have two follow-up action items:

1. We need to understand the source of the capacity vs. age distribution Eric and Kaushik discussed (slide 20 of the 2.3.1 talk) to prepare a mitigation plan. Can you send us a list of all T1 storage purchases by US ATLAS for the past 6 years? In particular, the plot in slide 20 shows two "holes" in recent years that I do not understand. What caused these big variations? What did we do with the associated funding?
2. We need a plan to mitigate the risk that BNL CEPH storage will not be ready for prime time in FY23. Let's do a risk registry exercise:
   - What is the probability that CEPH will work at scale if we wait until after the migration to the new SDCC computing center to test it?
   - What are the $ and schedule impacts associated with a CEPH deployment in FY24
   - What can we do to mitigate the risk in FY22 and FY23?

Can we put together what's needed and discuss it at the next 2.3 Facility coordination meeting (June 16, according to my calendar)?
Cheers,
Paolo

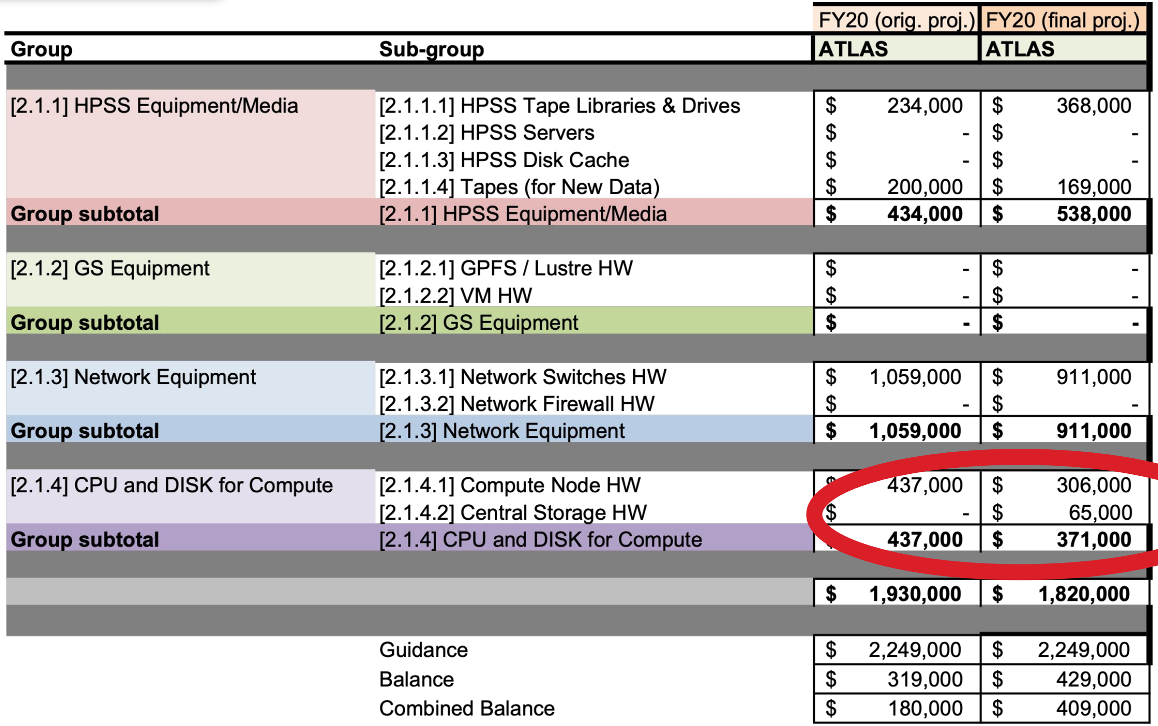
--
--
/~~~~~~\
| Ph:1-510-4866717 Fax:4864004
|
| Lawrence Berkeley National Lab
| Mailstop 50F-1606
| 1 Cyclotron Road
| Berkeley CA 94720-8147 - USA
|
|http://crd.lbl.gov/departments/computational-science/phax/staff/paolo-calafiura/
\~~~~~~/

Storage history

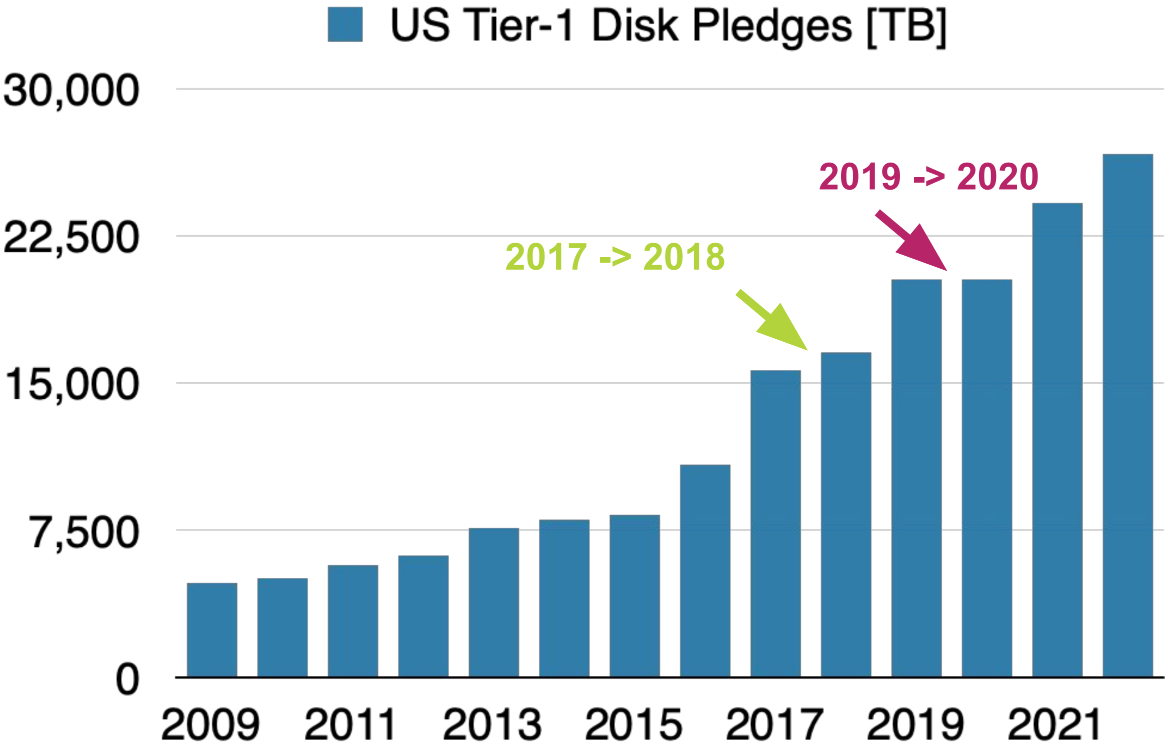
Can you send us a list of all T1 storage purchases by US ATLAS for the past 6 years? In particular, the plot in slide 20 shows two "holes" in recent years that I do not understand. What caused these big variations? What did we do with the associated funding?

- The list of Tier-1 storage purchases since 2015 is provided on the slide #4
  - Spending, equipment purchased year by year
- The 2 “holes” (2018 & 2019) were documented and presented in 2018 and 2020 scrubbing (SLAC & Seattle) and in subsequent scrubbings
  - 2018: funding came late and purchase was combined with 2019 one
  - 2020: no increase of pledged resources was needed in 2020 (graph next slide)
- The big variations are caused by:
  - The increase of disk resources needed by ATLAS which are not increasing linearly over time (graph next slide)
  - Little increase from 2017 to 2018 (see next slide), followed by no increase from 2019 to 2020 (see next slide)
- The funding received was used to increase the disk capacity when needed (to follow ATLAS requirements) or to refresh the infrastructure (2018 & 2020) -- see slide #5

FY20 accounting Equipment



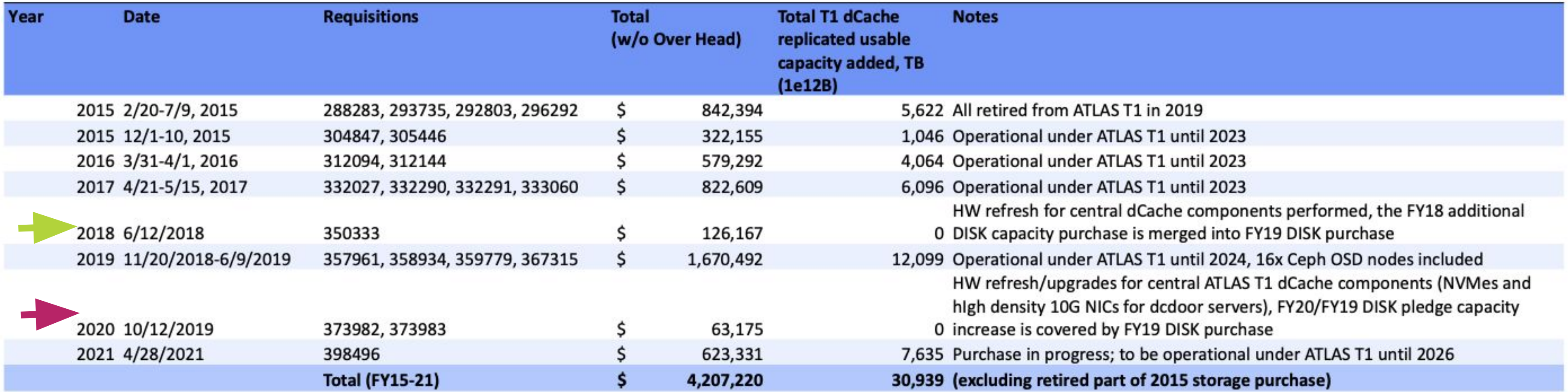
		FY20 (orig. proj.)	FY20 (final proj.)
Group	**Sub-group**	**ATLAS**	**ATLAS**
[2.1.1] HPSS Equipment/Media	[2.1.1.1] HPSS Tape Libraries & Drives	$ 234,000	$ 368,000
	[2.1.1.2] HPSS Servers	$ -	$ -
	[2.1.1.3] HPSS Disk Cache	$ -	$ -
	[2.1.1.4] Tapes (for New Data)	$ 200,000	$ 169,000
Group subtotal	[2.1.1] HPSS Equipment/Media	**$ 434,000**	**$ 538,000**
[2.1.2] GS Equipment	[2.1.2.1] GPFS / Lustre HW	$ -	$ -
	[2.1.2.2] VM HW	$ -	$ -
Group subtotal	[2.1.2] GS Equipment	**$ -**	**$ -**
[2.1.3] Network Equipment	[2.1.3.1] Network Switches HW	$ 1,059,000	$ 911,000
	[2.1.3.2] Network Firewall HW	$ -	$ -
Group subtotal	[2.1.3] Network Equipment	**$ 1,059,000**	**$ 911,000**
[2.1.4] CPU and DISK for Compute	[2.1.4.1] Compute Node HW	$ 437,000	$ 306,000
	[2.1.4.2] Central Storage HW	$ -	$ 65,000
Group subtotal	[2.1.4] CPU and DISK for Compute	**437,000**	**$ 371,000**
		$ 1,930,000	**$ 1,820,000**
	Guidance	$ 2,249,000	$ 2,249,000
	Balance	$ 319,000	$ 429,000
	Combined Balance	$ 180,000	$ 409,000



These are CRSG recommendations not ATLAS requests

http://wlcg-cric.cern.ch/core/vopledgereq/list/

List of Tier-1 storage purchases



Year	Date	Requisitions	Total (w/o Over Head)	Total T1 dCache replicated usable capacity added, TB (1e12B)	Notes
2015	2/20-7/9, 2015	288283, 293735, 292803, 296292	$ 842,394	5,622	All retired from ATLAS T1 in 2019
2015	12/1-10, 2015	304847, 305446	$ 322,155	1,046	Operational under ATLAS T1 until 2023
2016	3/31-4/1, 2016	312094, 312144	$ 579,292	4,064	Operational under ATLAS T1 until 2023
2017	4/21-5/15, 2017	332027, 332290, 332291, 333060	$ 822,609	6,096	Operational under ATLAS T1 until 2023
2018	6/12/2018	350333	$ 126,167	0	HW refresh for central dCache components performed, the FY18 additional DISK capacity purchase is merged into FY19 DISK purchase
2019	11/20/2018-6/9/2019	357961, 358934, 359779, 367315	$ 1,670,492	12,099	Operational under ATLAS T1 until 2024, 16x Ceph OSD nodes included
2020	10/12/2019	373982, 373983	$ 63,175	0	HW refresh/upgrades for central ATLAS T1 dCache components (NVMes and hIgh density 10G NICs for dcdoor servers), FY20/FY19 DISK pledge capacity increase is covered by FY19 DISK purchase
2021	4/28/2021	398496	$ 623,331	7,635	Purchase in progress; to be operational under ATLAS T1 until 2026
		Total (FY15-21)	**$ 4,207,220**	**30,939**	**(excluding retired part of 2015 storage purchase)**

$ values are w/o OH (OverHead factor varies from year to year: 1.1309x in FY19, 1.1344x in FY20, 1.1401x in FY21)

Ceph plans

Original pre-Covid Ceph plan

- Repurpose old (2012-2015 purchases) retired dCache storage into a Ceph cluster to be used for hosting secondary copy in dCache starting FY21
  - Avoid buying additional storage to be installed in old data center
  - Get operational experience with Ceph & Erasure Code (EC) at about ~10 PB JBOD scale (no WLCG site operating Ceph EC/dCache/JBOD) --- e.g. **operating a very low cost large scale storage solution**
- Covid made this plan inapplicable
  - Regular (1+1 replicated) storage had to be purchased in FY21
  - Experience with Ceph EC at scale still needs to be developed

Current storage plan

- Starting FY23, additional new storage hardware will only be deployed with Ceph under dCache, with Erase Code (EC)
  - 1.25 replication factor for main storage (replacement of 1+1 replicated dCache)
  - Presented at last scrubbing
- FY23 disk storage budget associated to that plan considering anticipated ATLAS disk requests: $1.1M
  - Alternative option: deployment of regular 1+1 dCache instead: $0.6M additional

Paolo’s 3 questions about Ceph plans

What is the probability that CEPH will work at scale if we wait until after the migration to the new SDCC computing center to test it?

- We plan to test Ceph at scale over the next months, in order to be ready to make an inferred decision, by summer 2022, before ordering FY23 equipment
- Plan:
  - Includes deployment of Ceph as a solution for sPHENIX experiment at BNL
  - 6 months of operation at scale test within ATLAS environment

What are the $ and schedule impacts associated with a CEPH deployment in FY24

- In case tests are unsuccessful, Ceph deployment as main storage solution will be postponed to FY24
- Regular dCache (1+1 replicated) storage will be deployed instead in FY23 for an additional estimated cost of ~$0.6M

What can we do to mitigate the risk in FY22 and FY23?

- Testing will be performed during FY22 to come to a decision before FY23
    - Ceph testing for sPHENIX experiment
    - Ceph+dCache stress tests outside of ATLAS
    - 6 months tests within ATLAS environment beside the Tier-1
    - Plan includes 3 months contingency
- Experience sharing with
    - Other large sites operating Ceph (RAL, CERN, …)
    - dCache team for optimized Ceph-dCache interface solution