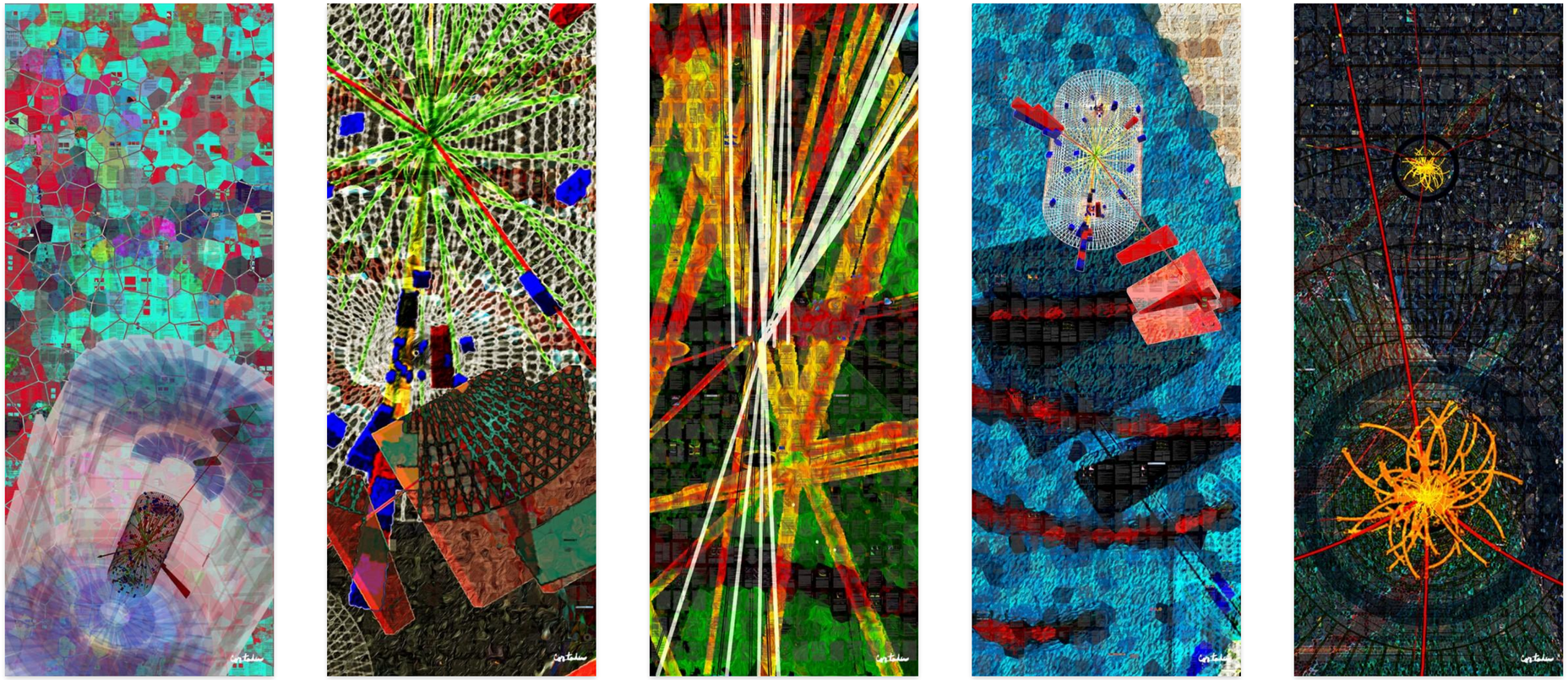
U.S. CMS Operations Program

FNAL FCRSG CMS

D. Mason for U.S. CMS Software and Computing Operations Program



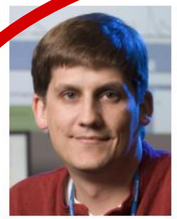
What we plan to talk about

- USCMS Ops program
  - This talk will be **completely facility focused**
- Project Planning and Risks
  - Brief 2021 Summary/Milestones achieved
- 2022 Resources
  - Forecasted resource needs into HL-LHC era
    - Brief drill down on tape, including WLCG Network/Tape Challenges
    - Run 3 and HL-LHC estimates, schedule change effect
    - HEPCloud and EAF

U.S. CMS S&C Operations Program Organization

The U.S. CMS S&C Operations Program Execution Team (PET)


Software & Computing
Oliver Gutsche/Fermilab
Tulika Bose/U.Wisconsin, deputy
Margret Votava/Fermilab, deputy
Robert Tuck/Princeton, project support


Facilities
David Mason
(Fermilab)
Andrew Melo
(Vanderbilt)

WBS 22.1

Operations
Dirk Hufnagel
(Fermilab)
Mia Liu
(Purdue)


WBS 22.3


Software
Saba Sehrish
(Fermilab)
Kevin Lannon
(Notre Dame)

WBS 22.2

HL-LHC R&D
Lindsey Gray
(Fermilab)
Michael Hildreth
(Notre Dame)


WBS 22.4

USCMS Computing Facilities

- The USCMS Fermilab Facility is a piece of the FNAL Institutional Cluster, and consists of the Tier 1 and LPC
- USCMS Provides computing and storage resources to CMS via FNAL Tier 1 and University based Tier 2 and 3 computing facilities.
  - **FNAL Tier 1 provides high availability computing and custodial data storage, as well as the analysis facility supporting the LPC.**
  - USCMS Tier 2's provide reliable compute and storage to CMS for both user analysis and central production
- The Operations Program budgets for USCMS Tier 1 and 2 hardware and support.
- In the following we summarize our process and estimates for estimating hardware needs and resulting budgets into the HL-LHC era (currently through 2030) for the Tier 1 ~~and 2's:~~

Project planning & risks

- Milestones and performance goals are tracked using ProjectManager.com

S&C ID	TASK NAME	DURATION	PLANNED ...	PLANNED ...	PERCE...	KPI ...
SC-M-WBS1-22-3	Decision Point: Should Tier-2s use Erasure-Coded Storage	126 days	1/6/2022	6/30/2022	100%	
	Tier 1 Pledge Deployment	**567 days**	**1/28/2021**	**3/31/2023**	**60%**	
SC-M-WBS1-23-3	**Deploy 2023 WLCG T1 Pledge**	**262 days**	**3/31/2022**	**3/31/2023**		
	Report WLCG Pledge Deployed	262 days	3/31/2022	3/31/2023		
SC-M-WBS1-21-4	**Deploy 2022 WLCG T1 Pledge**	**306 days**	**1/28/2021**	**3/31/2022**	**100%**	
	Update 5+ year T1 resource plan/forecast	65 days	6/2/2021	8/31/2021	100%	
	Extract Run 3 resource plan from LPC heads	88.96 days	1/28/2021	6/1/2021	100%	
	Ops program internal review	0 days	9/1/2021	9/1/2021	100%	
	PB of Retired HW from T1 into LPC EOS	242 days	1/28/2021	12/31/2021	100%	
	Report WLCG Pledge Deployed	1 day	3/31/2022	3/31/2022	100%	
	Tier 1 Purchases	**633 days**	**1/28/2021**	**7/3/2023**	**13%**	
	Purchase 2021 CPU	**226 days**	**2/19/2021**	**12/31/2021**	**100%**	
	Funds moved to EQ	1 day	2/19/2021	2/19/2021	100%	
	RFP submitted	85 days	2/22/2021	6/18/2021	100%	
	Vendor Responses Received	19 days	6/21/2021	7/15/2021	100%	
	PO Awarded	14 days	7/16/2021	8/4/2021	100%	
	Hardware Delivered	9 days	11/19/2021	12/1/2021	100%	
	HW Deployed	22 days	12/2/2021	12/31/2021	100%	
	Purchase 2021 Disk	**163.04 days**	**2/19/2021**	**10/6/2021**	**100%**	
	Req submitted	1 day	2/19/2021	2/19/2021	100%	
	Disk Arrays Deployed	64 days	2/19/2021	5/19/2021	100%	
	Servers Deployed	132.04 days	4/5/2021	10/6/2021	100%	
	Purchase 2022 CPU	**567 days**	**1/28/2021**	**3/31/2023**		
	Deploy	567 days	1/28/2021	3/31/2023		
	Purchase 2022 Disk	**619 days**	**1/28/2021**	**6/13/2023**	**50%**	
	RFP process started	567 days	4/12/2021	6/13/2023	100%	
	Hardware Delivered	567 days	1/28/2021	3/31/2023		
	Spectra Additional 20 Tape Drive Purchase 2022	371 days	1/31/2022	7/3/2023		
	LPC GPU Replacement 2022	566.96 days	1/28/2021	3/31/2023		
	Move T1 to LTO9 Media (Begin Writing)	1 day	1/2/2023	1/2/2023		
	T1 M8 to LTO9 Migration	522 days	6/5/2023	6/3/2025		
SC-P-WBS1-1	**Performance Goal: Meet CMS Site Readiness Metrics at T1**	**328 days**	**10/1/2021**	**1/3/2023**		**99.5**
	Performance Goal: Meet CMS Site Readiness Metrics at T1 (...	**64 days**	**1/3/2022**	**3/31/2022**		**99.5**
	CY22Q1	64 days	1/3/2022	3/31/2022		99.5
	Performance Goal: Meet CMS Site Readiness Metrics at T1 (...	**328 days**	**10/1/2021**	**1/3/2023**		
	CY22Q2	66 days	4/1/2022	7/1/2022		
	CY22Q3	66 days	7/1/2022	9/30/2022		
	CY22Q4	328 days	10/1/2021	1/3/2023		
SC-P-WBS1-2	**Performance Goal: LPC Availability**	**261 days**	**1/3/2022**	**1/2/2023**	**30%**	**99.2**
	LPC Interactives dual stack	22 days	3/2/2022	3/31/2022	100%	
	CY22Q1	63.96 days	1/3/2022	3/31/2022	100%	99.2
	CY22Q2	65 days	4/1/2022	6/30/2022		
	CY22Q3	66 days	7/1/2022	9/30/2022		
	CY22Q4	66 days	10/3/2022	1/2/2023		
	Performance Goal: Provide Necessary Tier 1 Services 2022	**990 days**	**3/18/2019**	**12/30/2022**	**30%**	
	Batch System Dual Stack	19 days	2/1/2022	2/25/2022	100%	
	Tier 1 IPv4/6 dual stack	**1 day**	**3/18/2019**	**3/18/2019**	**100%**	
	Storage	1 day	3/18/2019	3/18/2019	100%	
	2022 WLCG March Tape Challenge	5 days	3/7/2022	3/11/2022	100%	
	Dcache upgrade to 7x series	**44 days**	**3/2/2022**	**5/2/2022**	**100%**	
	Disk	44 days	3/2/2022	5/2/2022	100%	
	Tape	44 days	3/2/2022	5/2/2022	100%	
	Placeholder Random Thing to Make this Last The Year	260 days	1/3/2022	12/30/2022		
SC-M-WBS1-22-10	**Migrate T1 to OSG 3.6**	**135 days**	**2/24/2022**	**8/31/2022**	**31%**	
	Migrate XRootD to OSG 3.6	135 days	2/24/2022	8/31/2022		
	Migrate HTCondor-CE to OSG3.6	129.96 days	2/24/2022	8/24/2022	100%	
	Batch Farm to OSG 3.6	1 day	3/2/2022	3/2/2022	100%	
	Storage to OSG 3.6	131 days	3/2/2022	8/31/2022		

- Risks tracked in FNAL IPPM – major risks expected in 2022:
  - Cost increases and delivery delays
    - CPU purchase from 2021 had >6 month delivery delay
    - Have seen ~10% cost increases on first 2022 purchases (network and tape drives)
  - Russia-Ukraine conflict
    - Likely that JINR Tier 1 will not be available from June
    - 20% of CMS Tier 1 capacity (FNAL is 40%)
    - Have so far accepted 500 TB of JINR data, included with 2022 tape pledge
    - CMS completing validation of tape deletion campaign
  - Deviations from expected Run 3 and especially heavy ion data-taking rates

2021-22 Milestones & Activities

2021 Milestones Achieved

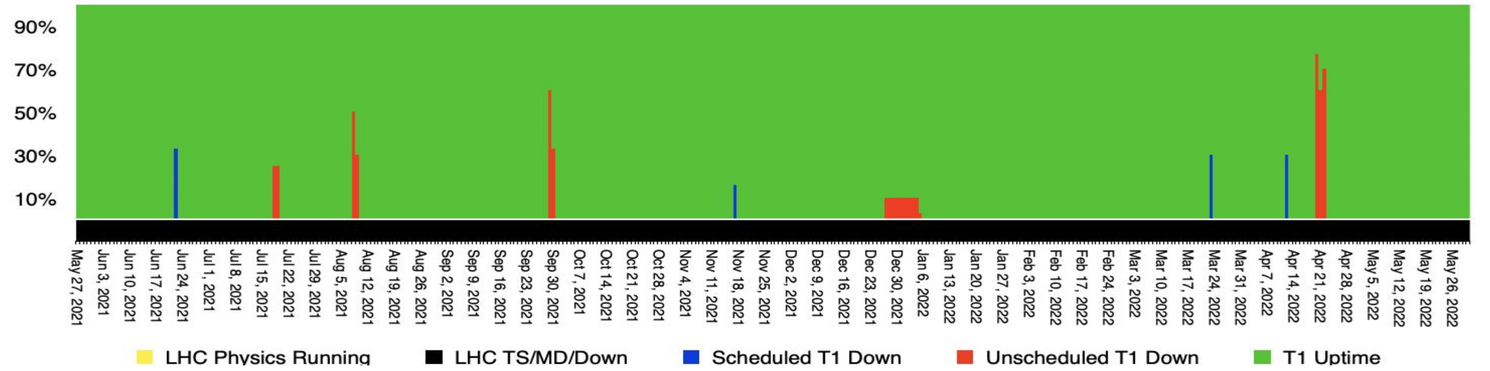
- Deployed 2022 WLCG Pledge
  - Though needed to hold onto old hardware longer than we would have due to significant delivery delay with 2021 CPU purchase. PO issued in August, needed by April 1, deployed by end of April
- Completed migration out of Oracle library and then:
- Decommissioned Oracle Library at T1 (see later slides)
- All CMS batch and storage resources at FNAL dual stack, transfers moved to HTTPS

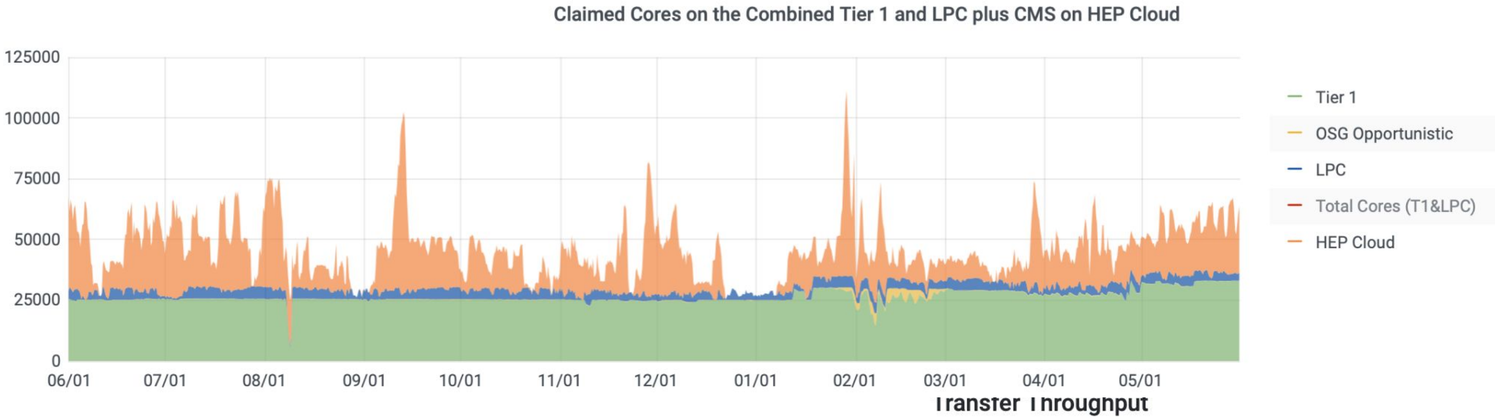
Milestones for 2022

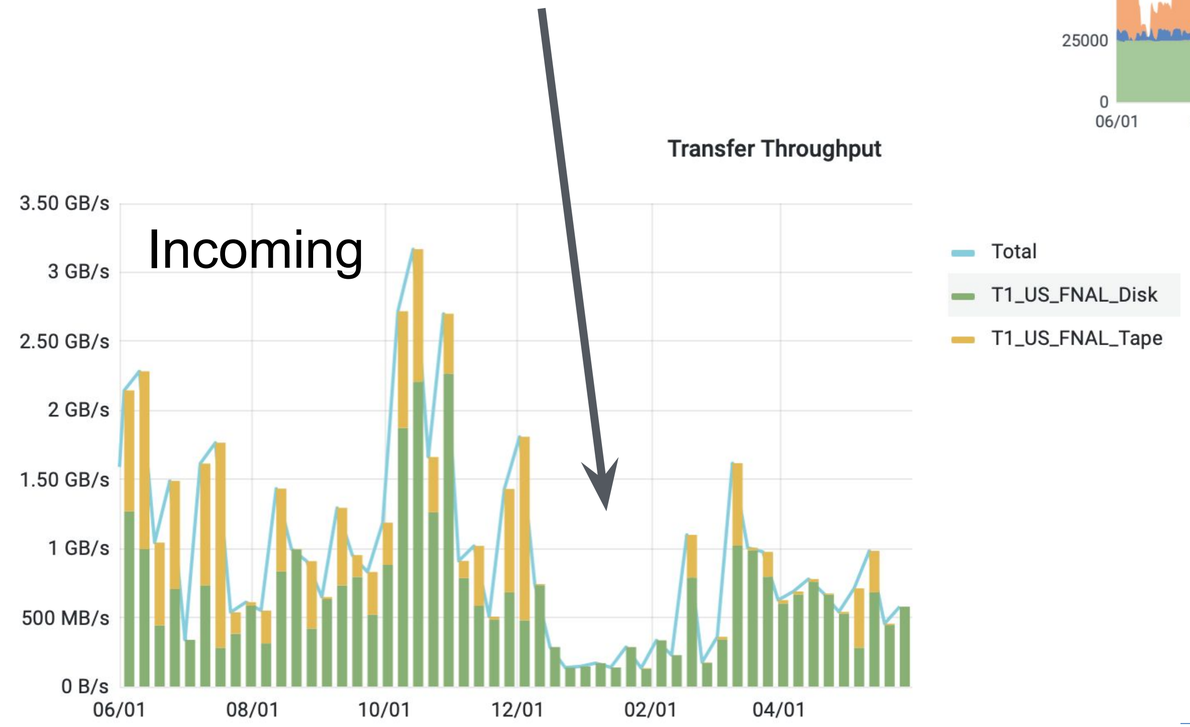
- Purchases for 2023 WLCG Pledge
  - Disk req started, moderate CPU and tape (more later)
  - Budget for update/replacement of interactive GPU, hardware for EAF and NVME for storage R&D
- Run 3 and HI datataking
- Testing and planning for transition to CTA in ~2024
- Investigating erasure coding for LPC EOS, and if LTO9 cost effective M8→LTO9 start

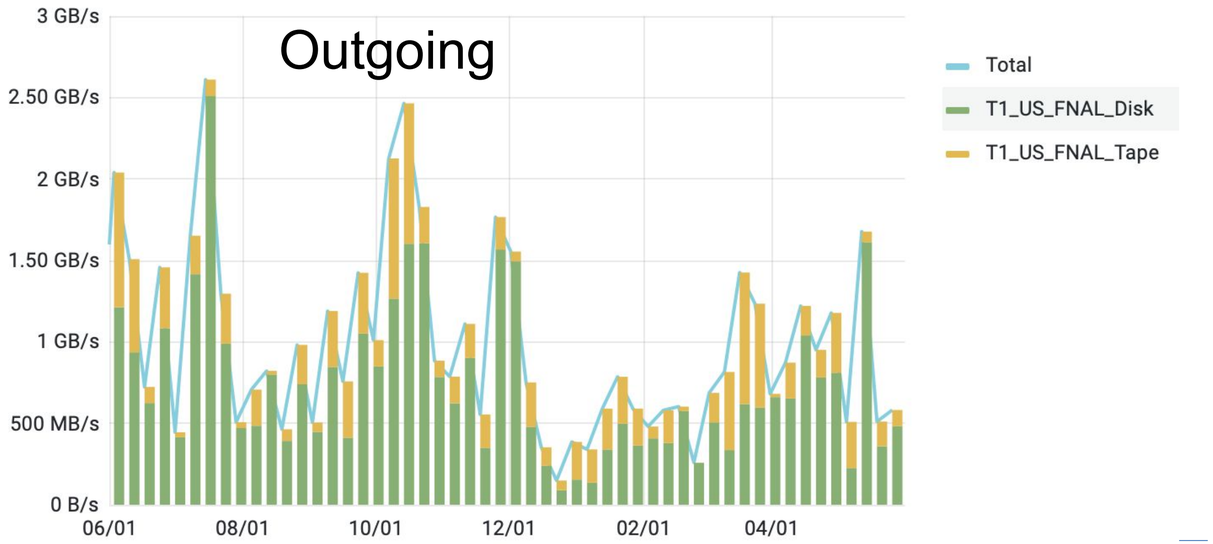
Performance

- Required to meet >98% readiness
  - **98.5% over the past year**
- CMS exhausted 2021 tape pledge by December 2021









Units

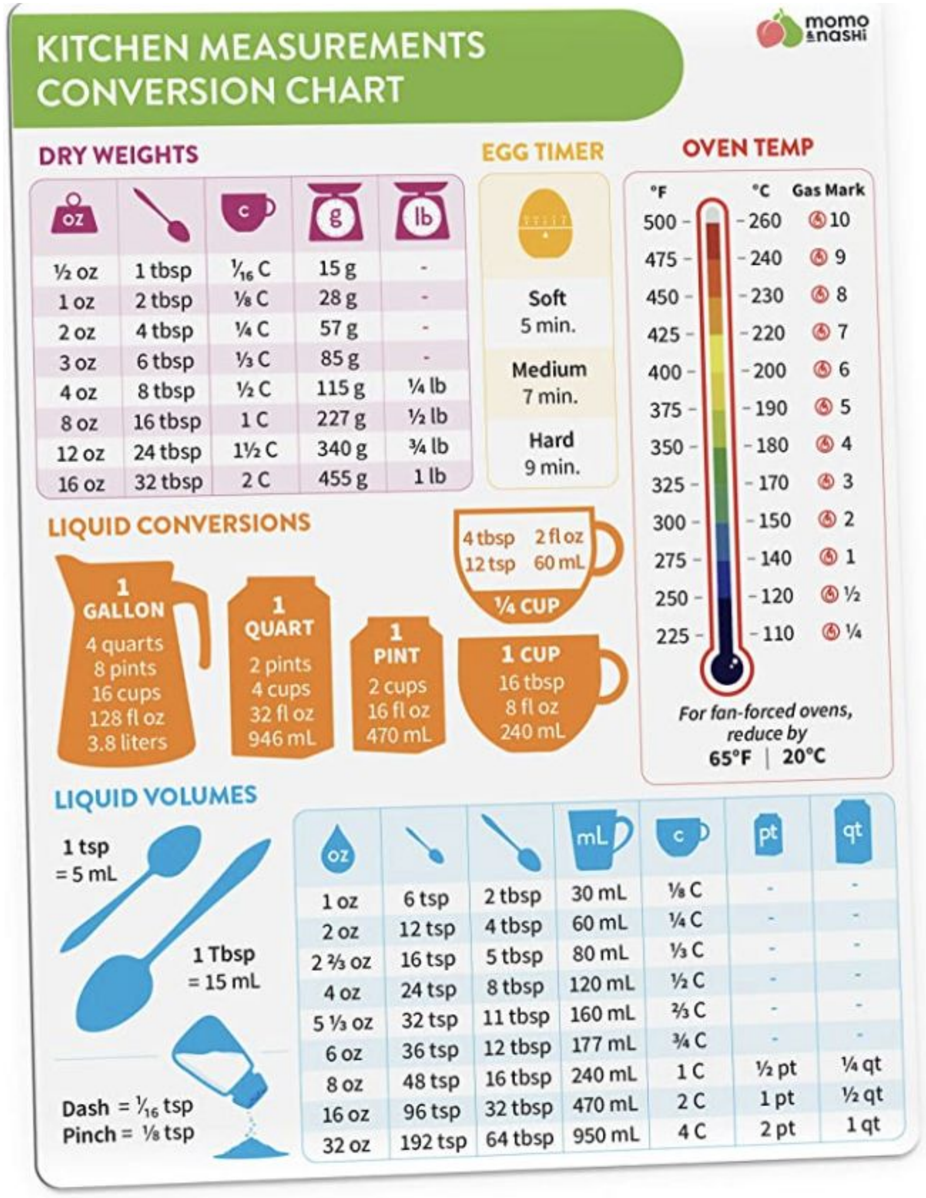
For the last decade and a half, WLCG has specifed CPU in kHS06*-years.

With the mix of CPU's we currently have in the T1:

Core-hr/year = (kHS06)/12.1 * 365 * 24

I.e. the 2022 pledge of 292kHS06 is about 24k cores, or 211Mcore-hr. (at high availability)

*WLCG moving to HEPScore this year



USCMS Computing Facilities

- Schedule changes – Run 3 now has an additional year in 2025, HL-LHC now starts 2029!
  - NOTE what we will see here is from our review this winter, based on the old schedule.
- In the short term, CMS requests were approved last spring.
- 2023 is a preliminary request, with expected minor adjustments this coming spring.
- For the rest of Run 3, we follow past experience and assume a continued growth rate of 10% per year.
  - We have reasonable confidence in this, given the expected similarity of Run 3 with Run 2.

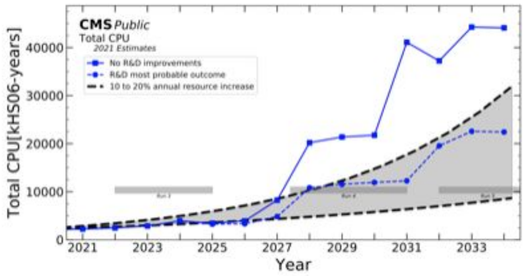
CMS		2021 Approved Request - Spring '20	2022 Approved Request - Spring '21	2023 Preliminary Request - Fall '21	Increase wrt 2022
CPU [kHS06]	Tier-0	500	540	720	33%
	Tier-1	670	730	800	10%
	Tier-2	1,070	1,200	1,350	13%
	Total	**2,240**	**2,470**	**2,870**	**16%**
Disk [PB]	Tier-0	30	35	45	29%
	Tier-1	77	84	98	18%
	Tier-2	92	98	117	19%
	Total	**199**	**216**	**260**	**20%**
Tape [PB]	Tier-0	120	155	228	47%
	Tier-1	230	260	316	22%
	Total	**350**	**415**	**544**	**31%**

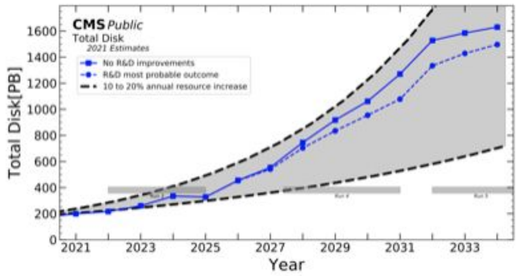
FNAL is 40% of T1!

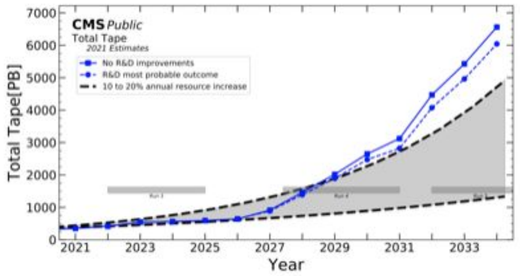
- Following Run 3, and into the first run of HL-LHC, we base our estimates on the CMS HL-LHC resource projections provided for the HLCC review in November.
  - As we all know "It is very difficult to predict – especially the future".
  - And yet we will proceed – the difficulty, and therefore uncertainties, increase with future distance.

Resource requirement modeling – CPU

CPU time, disk and tape time projected requirements estimated to be required annually for CMS processing and analysis needs.

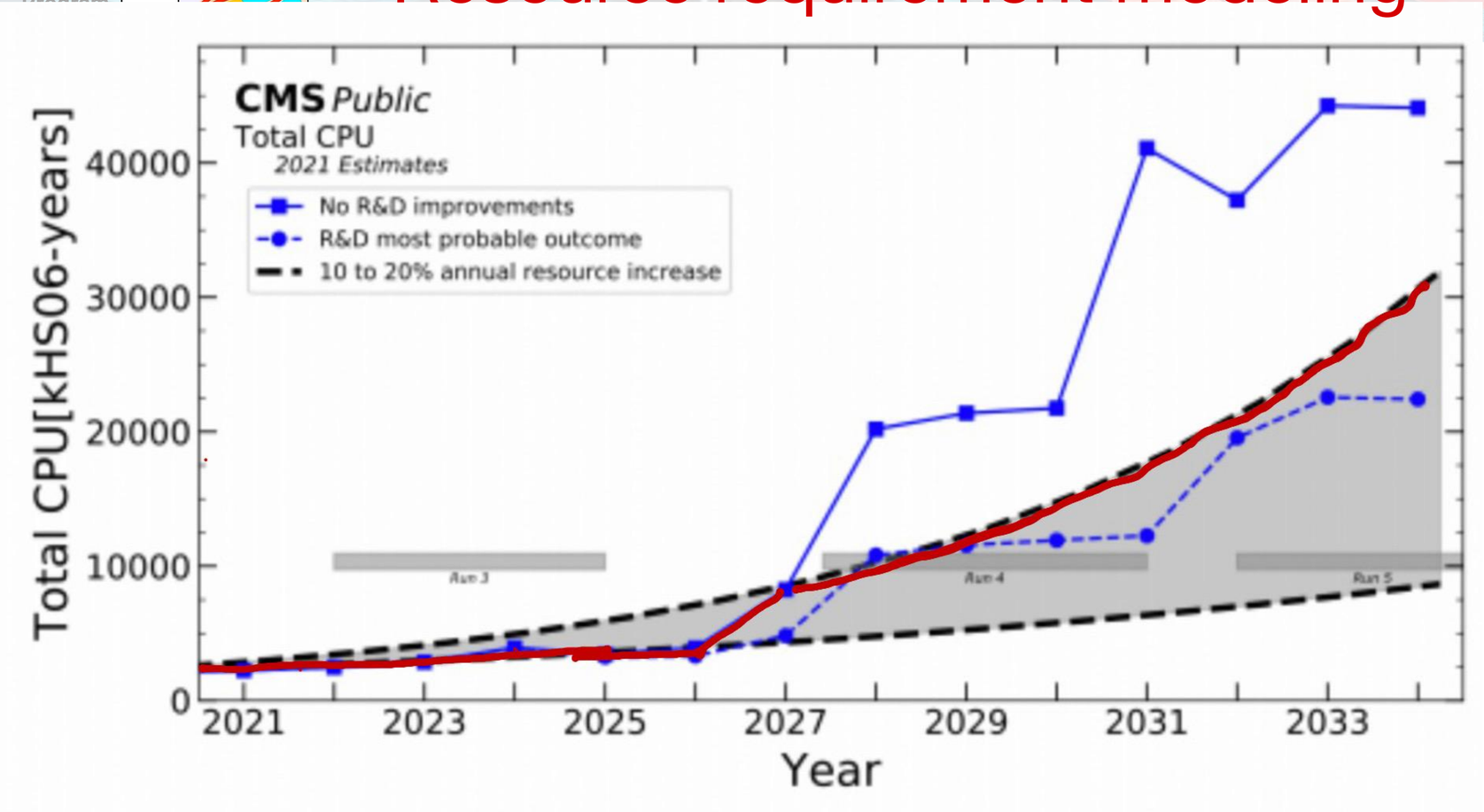


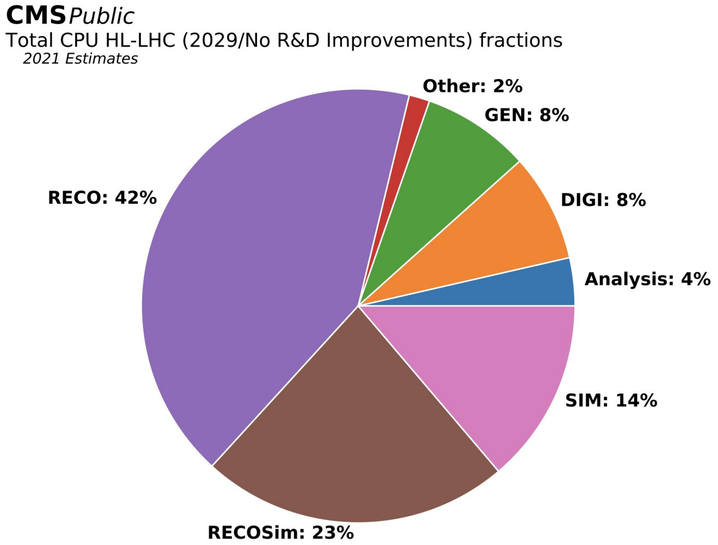




- Above are CMS model forecast results into HL-LHC for CPU, Disk and Tape.
- The blue lines are model forecasts, with dotted being most likely, after R&D goals realized
- The grey band follows 5 and 10% hardware/$ improvement
- Differences within the two sets give an indication of their uncertainties
  - For blue lines, its to what extent CMS is able to achieve expected improvements in software and capability
  - For grey its to what extent hardware cost/resource improves over time.
- With these forecasts and uncertainties in mind, in what follows we generally “bet” on CMS succeeding in making expected improvement, and aim to simplify.

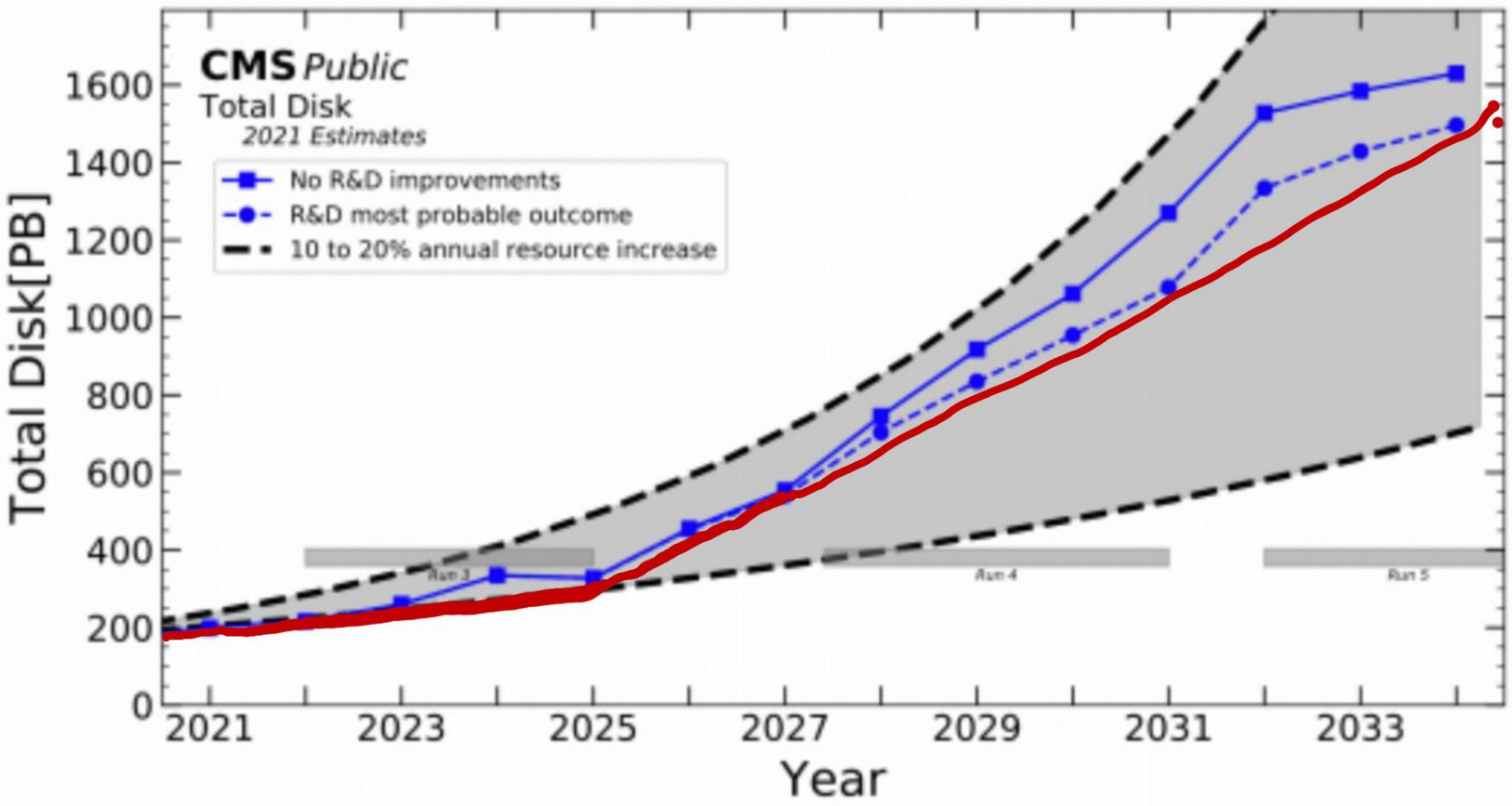
Resource requirement modeling – CPU

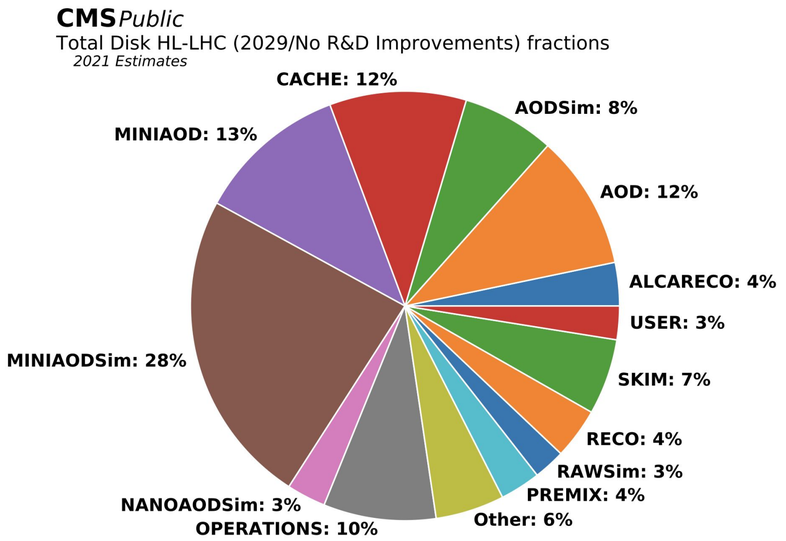




- Red sketch illustrates our budgeting assumptions based on the LHCC forecasts
- For CPU we stay with the past 10% increase assumption through 2026, then catch up to 20% curve in 2027, stay with 20% following.

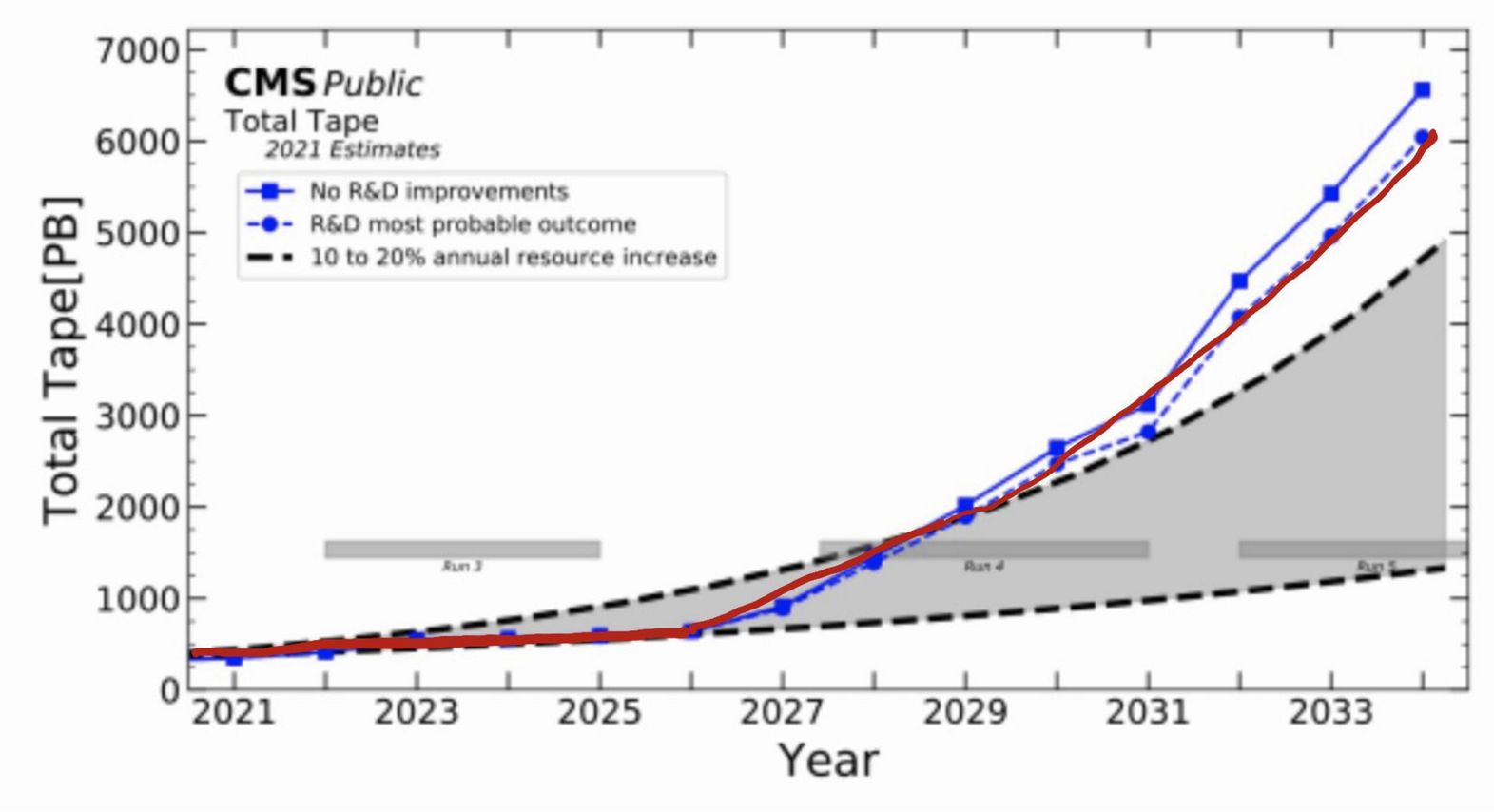
Resource requirement modeling – Disk

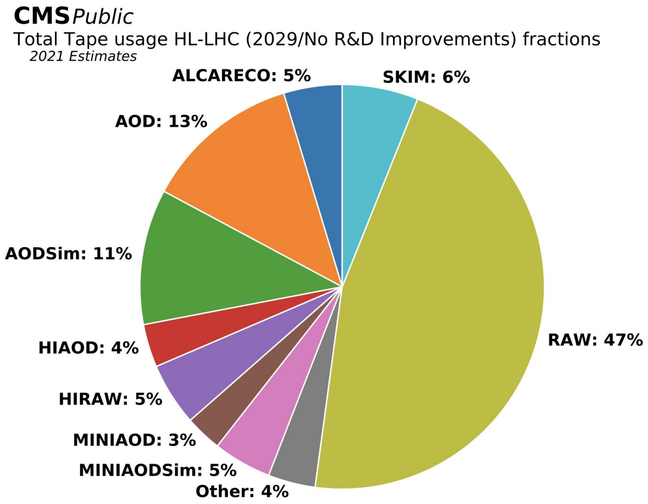




- For disk we stay with 10% increases until 2025, in 2026 and 2027 catch up to 15% increase curve, stay at 15% increase/year following.

Resource requirement modeling – Tape



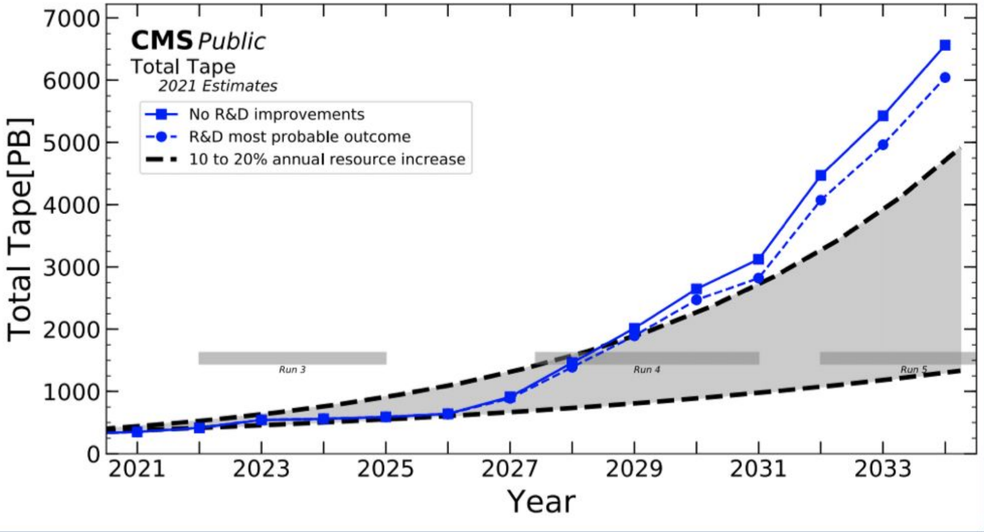


- For tape we stay with the ~10% /year increase through 2026, then catch up to 30% curve by 2029, stay at 30%/year following.

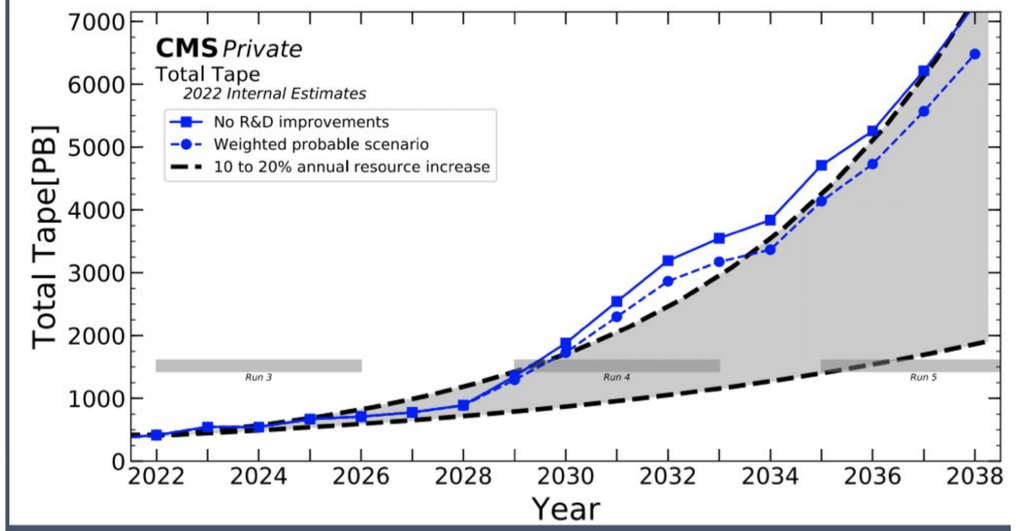
To give a sense of schedule change effect

Updated results – Total Tape (Internal)

LHCC review (last public result)



After changes



Preliminary modeling of the effect of the schedule changes – tape most extreme, in general we're pushing the "most probable" into the gray band

Tape Interlude

Types of Tape

There are many different kinds of tape, but adhesive tape usually consists of a narrow strip of backing material coated with adhesive on one side.

Related Guides

How to Prep a House for Painting

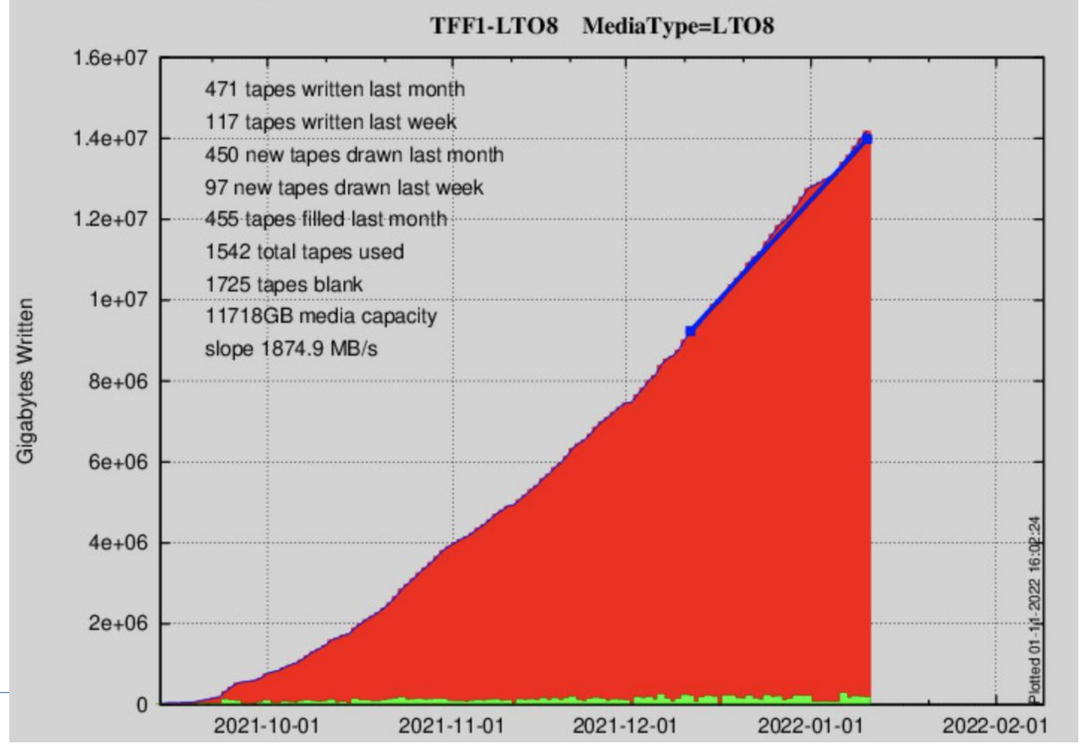
How to Paint Stripes on a Wall

How to Paint a Room

How to Paint Trim

Spectra Logic Library progress

- **In 2021 a new tape library was purchased to replace retiring Oracle libraries and provide capacity needed for Run 3.**
- Delivered end of May, commissioned end of August 2021
- Involved 2 months of integration work to enable in ENSTORE (first Spectra library)
- Growing pains and some initial scaling issues resolved by October
- **Migration from Oracle into this library now ~~nearing completion!~~ COMPLETED**
- **With the new library in place, moved equivalent of entire 2022 pledge increase in 3 months!**
  - **Demonstrates readiness for Run 3 capacity**
- **Many thanks to SCD-SDS for significant improvements to migration system enabling this success!**

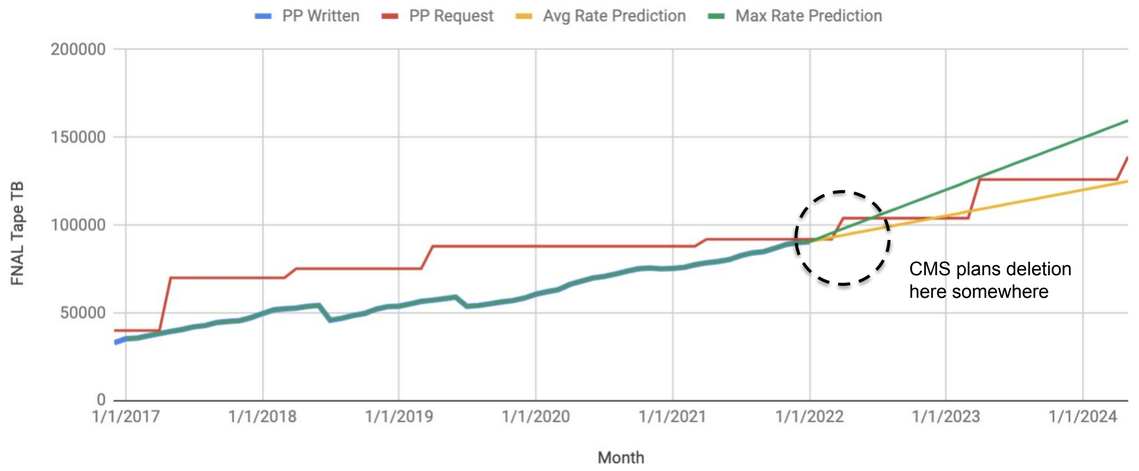


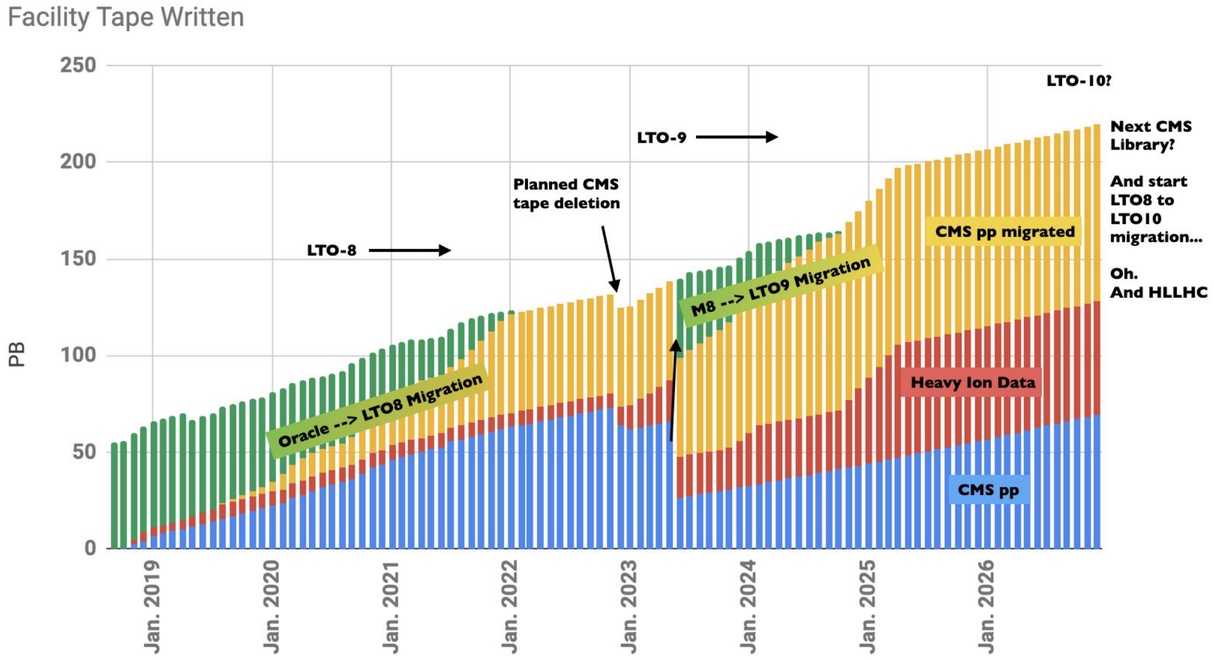
Ashes to Ashes, Dumpster to Dumpster

Successfully decommissioned this spring!

Tape Utilization & Predictions

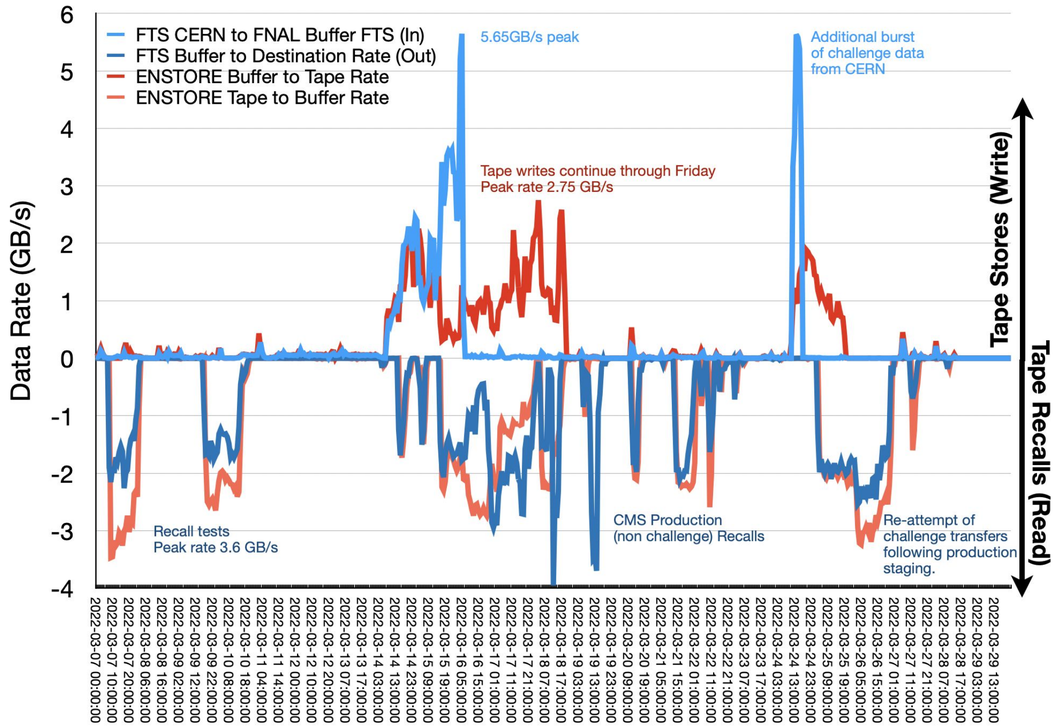
- Top right: Recent tape usage compared to CMS pledge request levels for FNAL.
  - Red is the CMS pledge request for FNAL
  - Yellow is extrapolation of 5 year average write rate.
  - Green is extrapolation of maximum achieved monthly average rate by CMS.
- Until 2022 CMS underutilized FNAL tape, asked for lower increases in 2020, 21 to "catch up"
- Bottom plot shows tape plan up to 2027
  - Anticipate adding LTO9 drives to new Spectra Logic library in 2022
    - Initially writes LTO8 media, in 2023 move to LTO9 media, start migrating M8 from IBM to Spectra LTO9
  - ~~Will need to migrate M8 -->LTO9 from 2023 to regain space in IBM library~~ Could happen sooner depending on LTO9 $/TB





Tape Challenges and Challenges for Tape

- A series of data challenges has been laid out to prepare for HL-LHC.
  - October and then this March the goals have been focussed around Run 3 readiness and rates.
  - FNAL met goals for this, but CMS struggled to keep the fire fed to really push things.
- Run 3 will be next “challenge”, though expected to be similar to Run 2.
- Next significant challenge for us will be Heavy Ion run at the end of this year.
  - Expect about 10 PB being produced at CERN in a few weeks, in Nov-Dec 2022.
  - Another ~5 PB to be produced during the year 2023.
  - Expect we can intake about 2-3 GB/s while continuing to support PP at normal rates. =few months transfer
  - More concrete planning discussions with CMS HI will be happening soon.



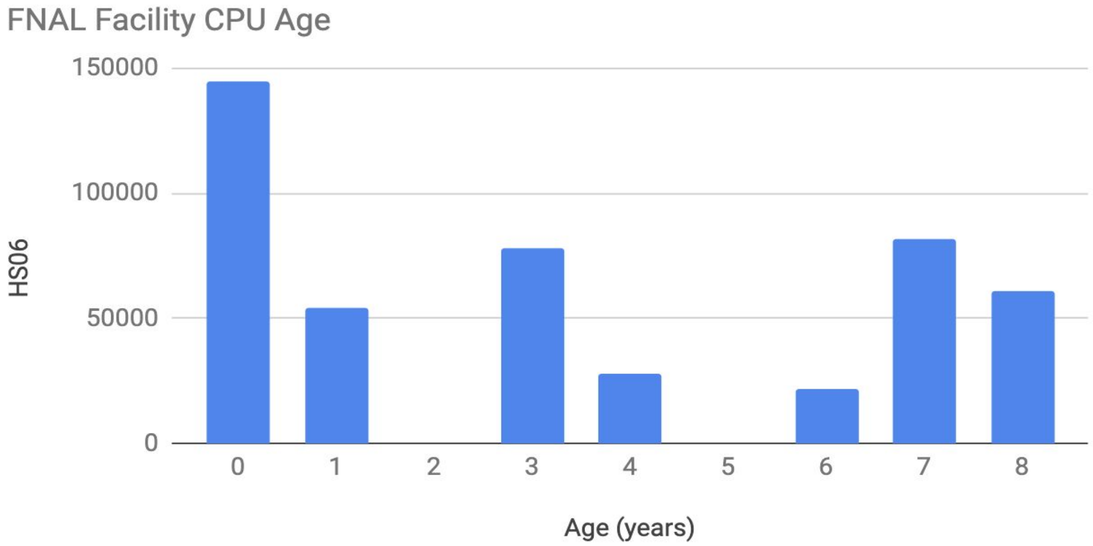
U.S. CMS Facilities Estimated Resource Requests

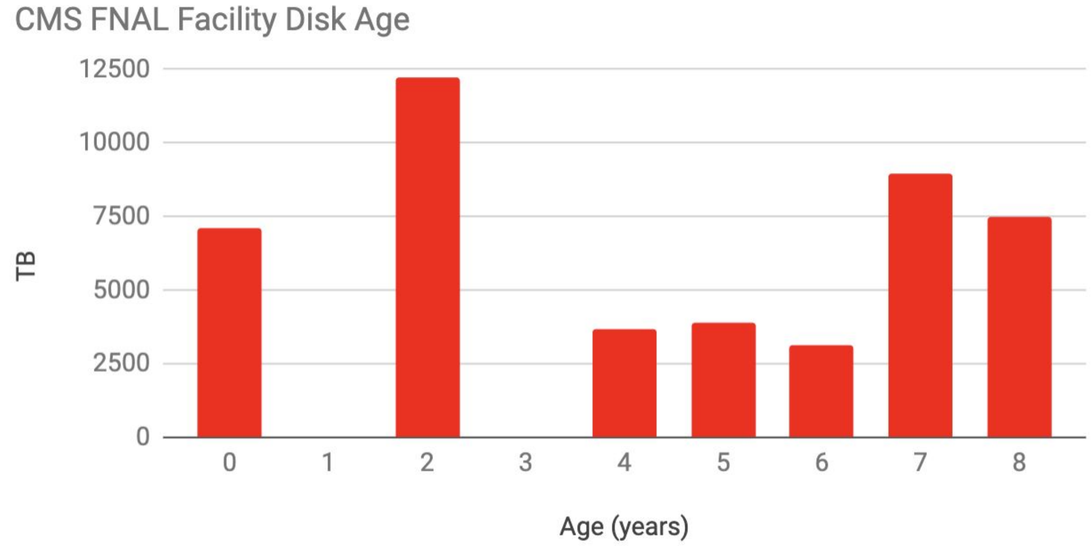
For FNAL	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
CPU kHS06	268	292	**320**	352	387	667	800	960	1152	1383
	% change	+9%	**+10%**	+10%	+10%	+72%	20%	20%	20%	20%
Disk PB	30.7	33.2	**39.2**	43.1	47.4	59.3	71.2	81.9	94.2	108.4
	% change	+8%	**+18%**	+10%	+10%	+25%	+20%	+15%	+15%	+15%
Tape PB	92	104	**126**	139	153	168	219	284	369	481
	% change	+13%	**+22%**	+10%	+10%	+10%	+30%	+30%	+30%	+30%
For Tier-2	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
CPU kHS06	268	300	**338**	371	408	678	813	976	1,171	1,406
	% change	+12%	**+13%**	+10%	+10%	+66%	+20%	+20%	+20%	+20%
Disk PB	23.0	24.5	**29.3**	32.2	35.3	44.3	53.2	61.2	70.3	80.9
	% change	+7%	**+19%**	+10%	+10%	+25%	+20%	+15%	+15%	+15%

- 2024 and 2025 assume 10% resource increases until requests are finalized
  - Prior to Run 4 ramp up from 10%/year increases to 15-30%
- Assume conservative 5% resource/$ improvements year to year in following budget estimates.

Tier 1 Hardware ages and retirements

- Hardware lifetime at the Tier 1 is maintained 3 years past 5 year warranty
  - Based on experience with hardware reliability
  - CPU safer to hold on to longer. Stateless, if fails you only lose that moment's jobs.
  - Relying on aging disk is more risky, a failure can result in data loss.
    - Our modular RAID architecture is very resilient, older hardware moved to replicated EOS in the LPC.
    - Storage refresh in 2019 allowed us to retire >8 year old hardware from Tier 1.
- Large quantities of resources, purchased in Run 1 are due to retire this year and next.





M&S Total Estimates to 2030

Preliminary Not for Distribution

	Run 3					HLLHC			
Year	2022	2023	2024	2025	2026	2027	2028	2029	2030
CPU	$628,616	$670,934	$417,771	$1,264,966	$797,338	$1,054,019	$859,428	$1,120,657	$624,356
Disk	$1,074,938	$880,311	$685,529	$981,652	$1,200,948	$849,865	$1,464,104	$875,410	$345,513
Tape	$656,329	$669,871	$606,074	$691,137	$1,222,781	$956,961	$903,797	$983,543	$1,462,024
Network	$318,986	$318,986	$318,986	$425,315	$425,315	$425,316	$425,316	$425,316	$425,316
Total	**$2,713,677**	**$2,540,103**	**$2,028,360**	**$3,363,070**	**$3,646,382**	**$3,286,161**	**$3,652,645**	**$3,404,927**	**$2,857,212**

- 2022 budget benefits from CPU buy ahead in 2021
  - Largest purchase will be disk to cover increase and replace expected retiring hardware
- Storage costs are approximately ⅔ of the overall budget through Run 3
- HL-LHC estimates simplified from model forecasts used in HLCC review this past fall

- Dominant purchase this year will be 14 PB of disk (ongoing)
- CPU as well, also tape media purchase this summer for PP and especially HI
- Will participate in ongoing GPU purchase

EAF and HEPCloud

- Marched toward, then needed to reverse course in deploying HTCondor infrastructure in OKD due to dual stack requirements.
- EAF making very good progress
  - LPC users though have been moving from “first beta testers” to more earnest testing of EAF.
  - In this year we will need to work out the Venn diagram of analysis/gpu/jupyter production commissioning.
  - As well as synergies with Tier 2 based analysis facilities
- HEPCloud managed to exhaust primary allocation and several additions in 2021!
  - As well as several XSEDE NSF allocations
  - During past year:
    - 136M core-hours from HEPCloud
    - 77M opportunistic from OSG
    - **Combined is just a hair under 217M provided by T1!**
  - Anvil and Perlmutter now onboarded!

Summary

- **FNAL Facility is READY FOR RUN 3**
  - **Continues to perform well and set the bar for other CMS sites (thank you!)**
  - **Much learned in WLCG challenges – passed what we needed to pass!**
- 2022 begins Run 3 (Physics data taking in mid July!)
- CMS Schedule has been updated, additional Run 3 year in 2025, HL-LHC now begins in 2029.
- HI run at end of this year will be a challenge – HI run at end of Run 3 will be double that challenge
- Risks to watch are supply chain and Russia/Ukraine situation
- Purchasing process for 2023 started, largest is disk purchase
- Working out commissioning plan for EAF and CTA will be important activities this year