

# Towards the HL-LHC-scale: I/O

Brian Bockelman

Section I: What's going to happen  
in the future?

# Analysis Estimates at the HL-LHC

- Compared to the Data Challenge exercise, a struggle with analysis at the HL-LHC has been generating agreed-upon minimal analysis models.
  - For the DC, we have a target of ~10Tbps moved worldwide.
    - Can it change tomorrow? Yes!
    - Can we find alternative constructs or counter-examples? Yes!
    - Is the error bar less than 30%? Probably not...
  - We've been unable to determine a consensus figure, much to the community's detriment.
- What's the "10Tbps" metric for analysis?

# Models are powerful!

- Having a simplified mental model is essential to communicating what we do!
  - Remember the MONARCH model of tiered sites? It was never implemented as written but allowed the LHC to communicate the power of distributed computing. United the community behind a philosophy and vision.
- It is understood that plans change!
  - Yes, ML could completely upend analysis tomorrow.
  - Yes, the solution set for “how is analysis done?” is multi-dimensional.
- Having no model is equivalent to having no plan.
  - So, let’s get out the napkin and calculate!

# Strawman Proposal #1 (using CMS terminology)

- Physics analysis is based completely on NanoAOD.
- Primary + simulated datasets add up to 20% of the official CMS NanoAOD 2028 dataset.
- User would like to derive quantities (ntuples) from the selected data, resulting in a user-managed dataset 20% of the size of the input.
  - This derivation is done once a week.
  - Derivation is CPU-bound at **N** HZ; heavily inference based and GPU-bound at **M** Hz.
  - Goal: derivation is done “over lunch”. Time-to-first-event is “instant”
- Once the user’s ntuple are completed, they want to explore and create histograms and refine.
  - Goal: Time to ‘histograms with 1% of events’ is “instant”. Time to ‘all histograms created’ is “cup of coffee”.

# Strawman Proposal #2

## (using ATLAS terminology)

- Same as before *but* the analysis requires reachback to xAOD for 5% of events in the analysis.
  - The reachback occurs when the user ntuples are derived.
  - The events in PHYSLITE are joined with the columns read from the xAOD.
  - Important alternate: need 5% of all xAOD events.
- Goals:
  - Keep the user ntuple creation in the same order magnitude of time of strawman #1.

# I/O for Analysis Facilities

- I'm not the right person to add the “physics coloring” to these strawman to convince others scaling up is reasonable.
  - However, having a single – preferably ~45 – use cases / 2-page whitepapers written up is immensely valuable for communicating between the physics groups and computing facilities.
- For example, to do strawman #1 at an AF, we need:
  - ~350Gbps sustained from storage (XCache?) to support the user ntuple creation.
  - ~800Gbps sustained from storage to support the histogram creation.

# Can we simplify to a spreadsheet?

<b>CMS Assumption</b>	<b>Value</b>	<b>Notes</b>		<b>Derived CMS Quantities</b>	<b>Value</b>
Events / Year (B)	150	100B sim / 50B data		NanoAOD size / version / yr (TB)	750
NanoAOD event size (KB)	5				
MiniAOD size (KB)	200				
"Cup of Coffee" (min)	5				
"Lunch time" (min)	60				
<b><u>Strawman Analysis #1</u></b>					
<b>Analysis Assumption</b>	<b>Value</b>	<b>Notes</b>		<b>Derive Analysis Quantities</b>	<b>Value</b>
Percent NanoAOD events read	20%			Events read / user ntuple gen (B)	30
Size user ntuple event (KB)	1			User ntuple size (TB)	30
				Event rate / ntuple gen (MHz)	8
				Data rate / ntuple gen (Gbps)	333
				Event rate / ntuple read (MHz)	100
				Data rate / ntuple read (Gbps)	800



Section II: What are reasonable targets?

# Don't forget, computers are fast...

- Unlike CPU and memory, networking and NVMe costs have dropped shockingly fast.
- A few reference points:
  - 15TB NVMe => \$2,500
  - 32 x 100GbE switch => \$12k (\$15k to upgrade to 32 x 400GbE).
  - Upgrade WN to 100GbE NIC => \$300
  - WN with (relatively slow) 48C/96HT => \$5k
  - 4xL40 GPU host => \$40k

# Can we design out an AF on a napkin?

- With 4 x 15TB NVMe, XCache should saturate a 100GbE connection.
  - A single switch would be non-blocking if the CPU sink was co-located.
- It would take 13 servers to host a single copy of NanoAOD.
  - More than sufficient bandwidth to support the defined operations.
- All within a \$250k budget.
  - For capacity beyond this, scale out to the local T2.

Purchase	Cost	Count	Total Cost	Notes
WN with 60TB NVMe	\$15,000	13	\$195,000	Sufficient for one NanoAOD copy
Switch/networking	\$15,000	1	\$15,000	
GPU host with 4xL40	\$40,000	1	\$40,000	
		Totals	\$250,000	

# Back to reality

Would this napkin AF work as-is? **Probably not...**

- While the ntuple event creation rate / CPU core is reasonable (13KHz), the event rate / GPU is not (2.1MHz).
- We only went through strawman #1; insufficient disk space for strawman #2.
- User ntuples would need to go on distributed storage, not accounted for here.
- We assume the user can utilize 100% of the system. In reality, there's collision for "generate ntuples over lunch".
  - Even ignoring bursts, barely enough CPU cores to support 100 analysts generating their ntuples once a week.
- Then again ... this is the 2028 workload using 2023 prices!

# So what's the point?

1. I/O-wise, HL-LHC analyses are not particularly demanding for today's hardware.
  - They are demanding for the type of hardware at our production facilities.
  - The costs are low enough to consider 25-50% scale demos today.
2. Given a few input parameters, facilities can iterate on cost models and build multi-year purchase plans.
  - Hardware bought today will be in use at the HL-LHC.
  - We need to start cost models now to roll them into the budget for HL-LHC.
3. Even simplified use cases – like strawman #1 – can produce sufficient inputs for facility design.
  - Only leading-order estimates are needed. The highest risk appears to be around the costs for GPUs within the model!

# Section III: Data Preservation and Open Access

# The Nelson Memo

About a year ago, the “Nelson Memo” came out instructing funding agencies to generate new open data policies:


- “... make publications and their supporting data ... accessible without an embargo on their free and public release”



EXECUTIVE OFFICE OF THE PRESIDENT  
OFFICE OF SCIENCE AND TECHNOLOGY POLICY  
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson   
Deputy Assistant to the President and Deputy Director for Science and Society  
Performing the Duties of Director  
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31<sup>st</sup>, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

# The Nelson Memo

- Taken at a literal level, it means LHC raw data should be published at the same time as the corresponding papers.
- As with any rules, there are exceptions and carve-outs:
  - I.e., maximize “appropriate sharing” where *appropriate* “preserves the balance between the relative value of long-term preservation and access and the associated cost and administrative burden”
- I’ve heard this described as “open data where you reasonably can”.
  - “I might publish more from this dataset” is no longer an acceptable excuse.
  - “Agreements with international collaborations” likely will be.
  - How will US funding agencies view “but we fund M&O costs through data access rights”? Probably OK, yet to be seen...

**Don’t focus on the legalese, focus on the opportunity**



# A new world of Open Access

- Funding agencies have had a year to digest the Nelson memo.
  - NSF released their [policy in June](#); will come into effect for proposals submitted by January 2025.
- Next, I expect agencies to have opportunities for leaders to show a vision on how to execute this! Many non-trivial questions...
  - Data volume is an obvious challenge! In such a regime, is a separate data archive (as is done today) feasible?
    - If now, how to give managed access to our production facilities?
  - Open Access has an equity component! Releasing an exabyte of data on a 1Gbps link isn't really "open", is it? How do groups from non-R1 universities use open data?
    - What centers need to exist? Is the IRIS-HEP Open Data Facility a first version of what will need to exist for HL-LHC?

# Questions?

This project is supported by the National Science Foundation under Cooperative Agreements OAC-1836650 and PHY-2323298. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.