RHIC Data and Analysis Preservation Round Table

06/05/2025

Introduction

Content

1. Estimated hardware resources required for 5-10 years from now and DAP implementation

Data volume

These are projections, including the 2025 run

[PB]	PHENIX	sPHENIX	STAR
RAW	20	160-300	130
Analysis Objects	5	50-100 (one processing)	*45*
Others	10	50-100 (prev. processing)	?

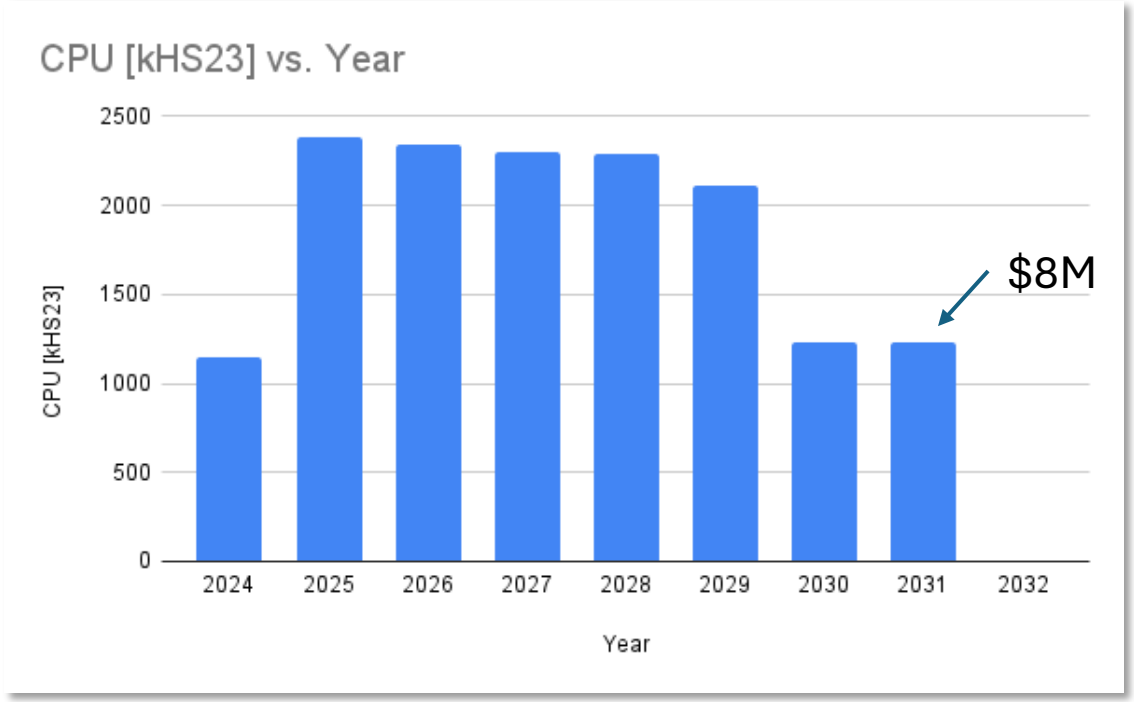
Consistency with the current Tape system accounting needs to be checked.
For example, DST volume (2025 run is not included)

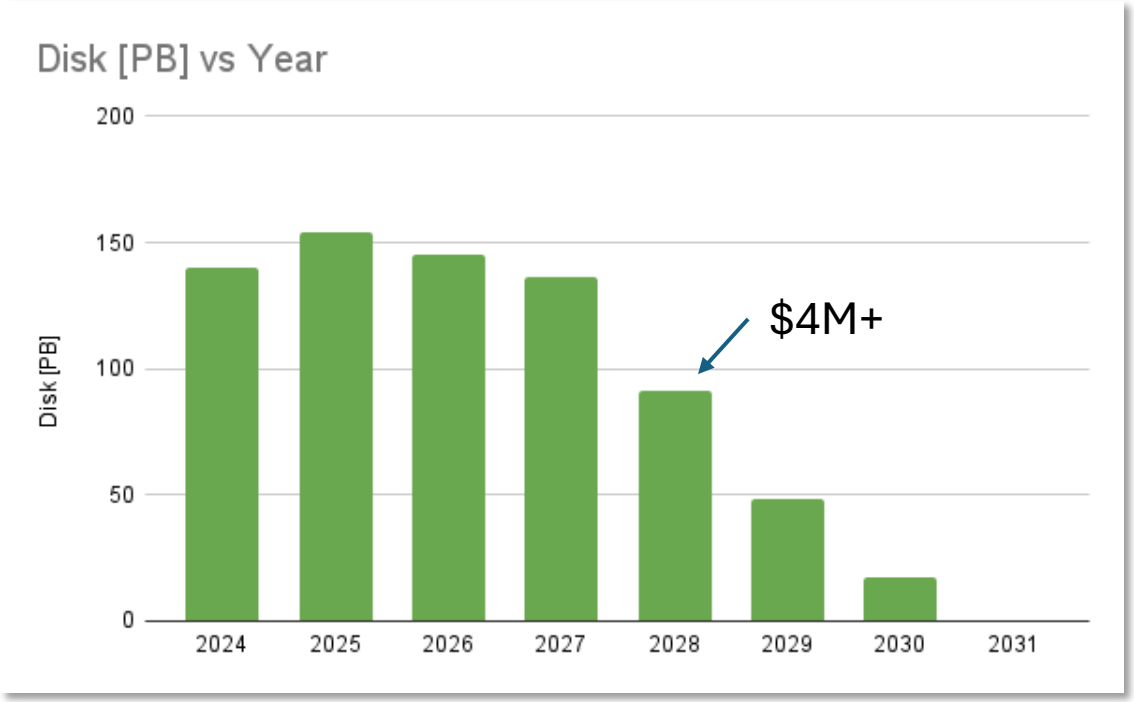
[PB]	Phenix	sPhenix	Star	Sum
DST	9		*50*	59
Raw	17	95	68	181
Raw 2nd copy	6		4	9
Archive	1		3	3
Sum	32	95	124	252

Resources projections

- **Main constraint**: Hardware resources currently deployed will not be renewed past their lifetime
- Hardware lifetime:
  - CPU: 7 years
  - Main Disk Storage / Tape Buffer: 5 years
  - Tape: Repack and change of technology every 10 years
- **Reality**: Hardware available for DAP will be limited

Comparison with expected resources





Critical Gap for funding for disk beyond 2029 and CPU beyond 2031

Shigeki will present the details, including Tape systems

The 2030 Challenge

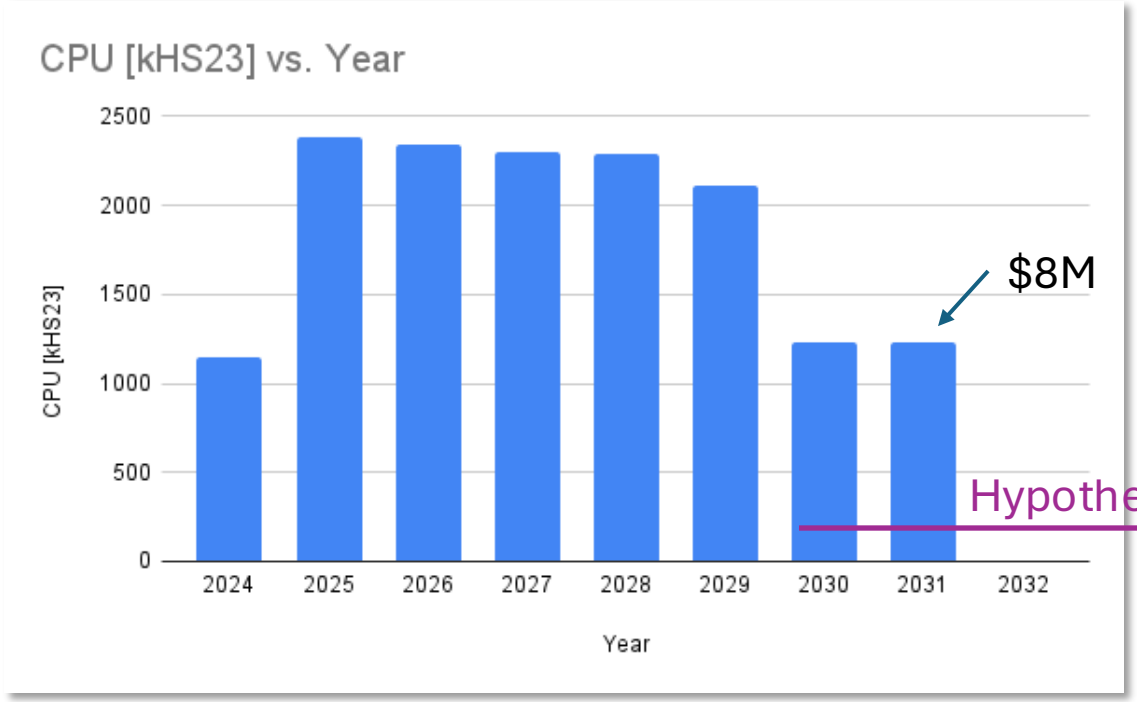
- Targets past 2030 – **THIS IS AN EXERCISE**
  - 30 PB of disk storage
  - Today's STAR usage of CPU for analysis (20% of total)
  - Parameters can/will be adjusted
- Limited budget:
  - Not all Analysis Objects fit on disk. Requirement 100-150 PB.
  - Addressing the gap: Rely on tape as main storage.
  - Analysis is tape dependant

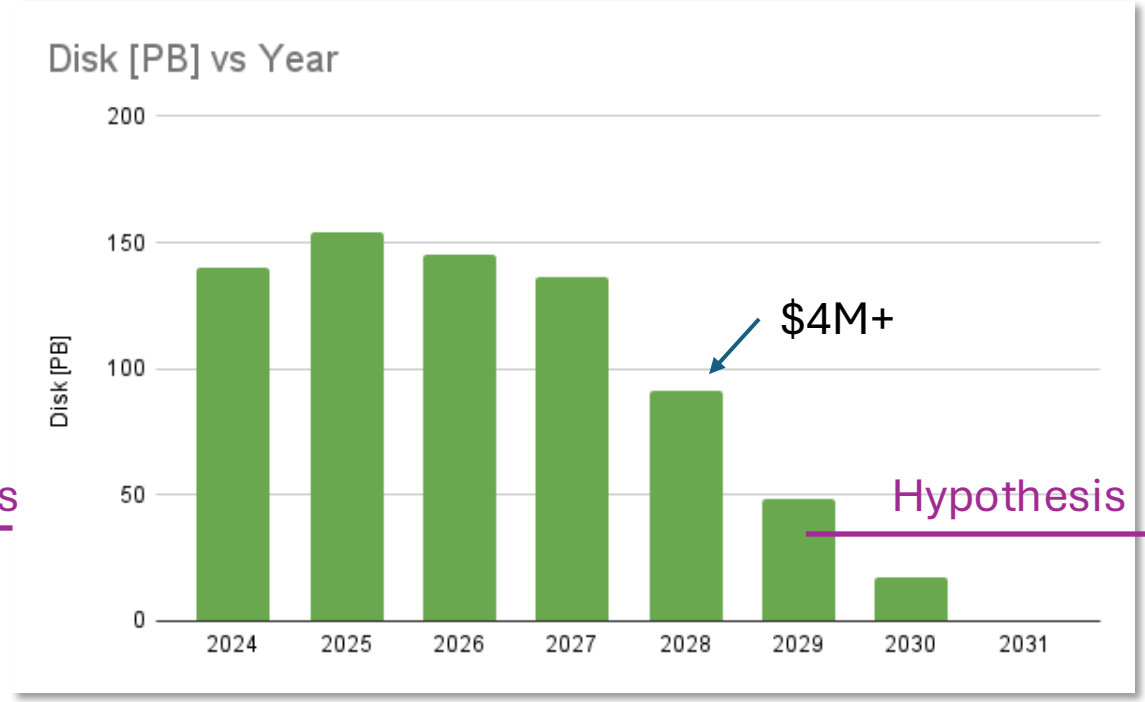
Solutions?

Making it work with limited resources

- Efficient data carousel service
- Review of data format/content
- Advanced compression
- Prioritized data placement

Comparison with expected resources





Hardware requirements for CPU & Disk to support DAP start
CPU: 2032
Disk: 2030

Questions to the experiments

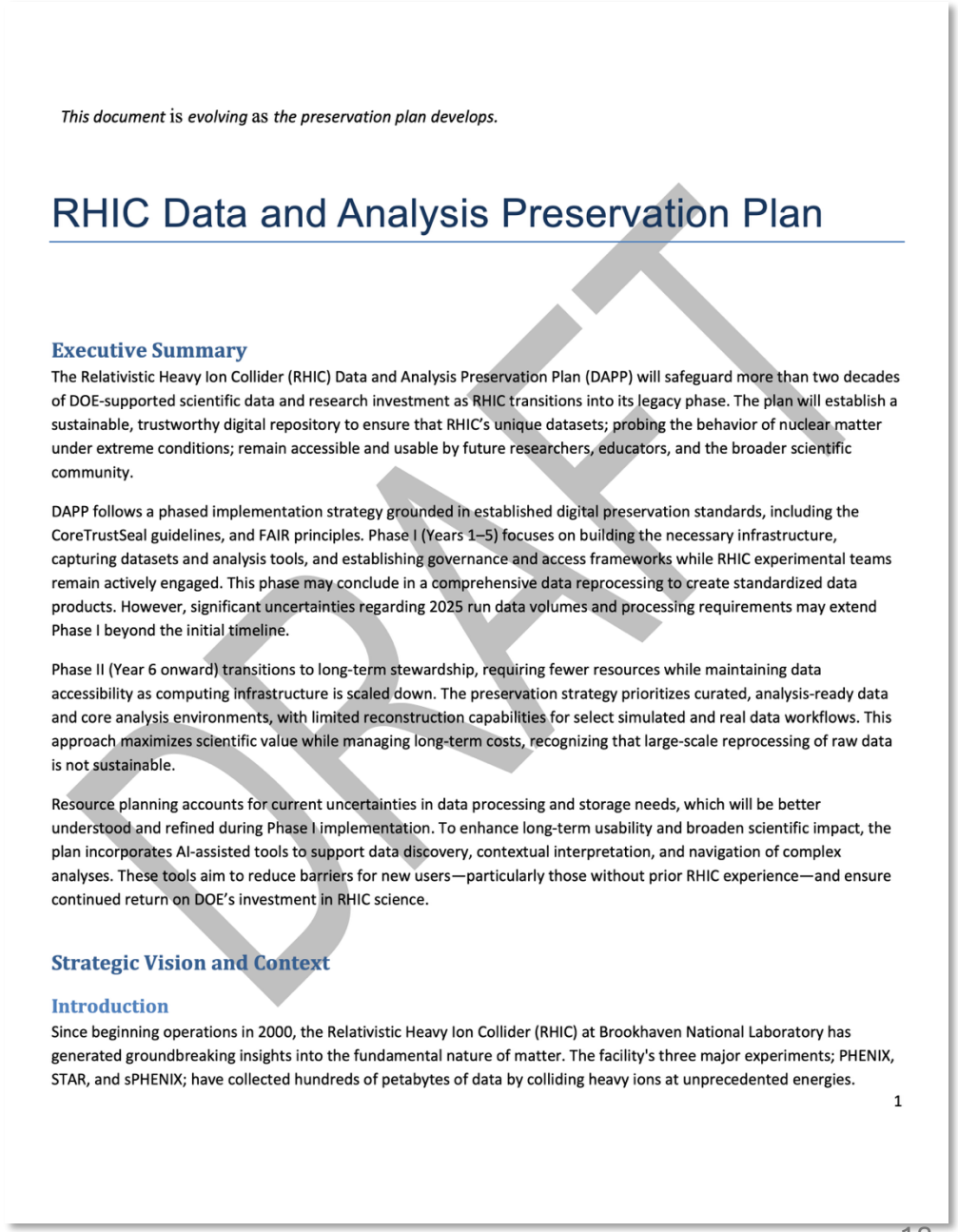
- Do you see **2030** as a realistic timeframe for **entering the data preservation phase**, where analysis becomes the primary activity? If not, what are the main obstacles?
- How do you **anticipate CPU and disk storage** needs will evolve over the **next five years** compared to current levels? Rough estimates and timelines are perfectly fine.
- What do you expect the **primary computing activities** will be around **2030**?
- If your computing allocations remain at current levels, when do you believe the transition to preservation mode would realistically take place?

Why These Questions Matter

- These questions help to
  - Define a realistic timeline for entering Preservation phase
  - Infrastructure planning
  - Identify obstacles
  - Guide resource allocation (tape vs disk)
- Expected Outcomes
  - Estimate transition timeline
  - Shape transition to preservation
  - Guidance for the Data and analysis Preservation plan

Draft of the DAP Plan (the DAPP)

- **Overview:**
  - A **first draft** of the DAPP has been prepared.
  - It captures ideas and priorities discussed during the RHIC DAPP roundtable series.
  - The document was initially expected to be released by this summer.
- **Next Steps:**
  - The draft will be shared shortly.
  - Everyone is encouraged to review and provide feedback.
  - Dedicated time will be allocated in upcoming meetings for discussion and refinement.



This document is evolving as the preservation plan develops.

RHIC Data and Analysis Preservation Plan

DRAFT

Executive Summary

The Relativistic Heavy Ion Collider (RHIC) Data and Analysis Preservation Plan (DAPP) will safeguard more than two decades of DOE-supported scientific data and research investment as RHIC transitions into its legacy phase. The plan will establish a sustainable, trustworthy digital repository to ensure that RHIC's unique datasets; probing the behavior of nuclear matter under extreme conditions; remain accessible and usable by future researchers, educators, and the broader scientific community.

DAPP follows a phased implementation strategy grounded in established digital preservation standards, including the CoreTrustSeal guidelines, and FAIR principles. Phase I (Years 1–5) focuses on building the necessary infrastructure, capturing datasets and analysis tools, and establishing governance and access frameworks while RHIC experimental teams remain actively engaged. This phase may conclude in a comprehensive data reprocessing to create standardized data products. However, significant uncertainties regarding 2025 run data volumes and processing requirements may extend Phase I beyond the initial timeline.

Phase II (Year 6 onward) transitions to long-term stewardship, requiring fewer resources while maintaining data accessibility as computing infrastructure is scaled down. The preservation strategy prioritizes curated, analysis-ready data and core analysis environments, with limited reconstruction capabilities for select simulated and real data workflows. This approach maximizes scientific value while managing long-term costs, recognizing that large-scale reprocessing of raw data is not sustainable.

Resource planning accounts for current uncertainties in data processing and storage needs, which will be better understood and refined during Phase I implementation. To enhance long-term usability and broaden scientific impact, the plan incorporates AI-assisted tools to support data discovery, contextual interpretation, and navigation of complex analyses. These tools aim to reduce barriers for new users—particularly those without prior RHIC experience—and ensure continued return on DOE's investment in RHIC science.

Strategic Vision and Context

Introduction

Since beginning operations in 2000, the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory has generated groundbreaking insights into the fundamental nature of matter. The facility's three major experiments; PHENIX, STAR, and sPHENIX; have collected hundreds of petabytes of data by colliding heavy ions at unprecedented energies.

1

Today

1. Hardware projections - Shigeki

- Next meeting: **Thursday, 06/12 - Tentatively**