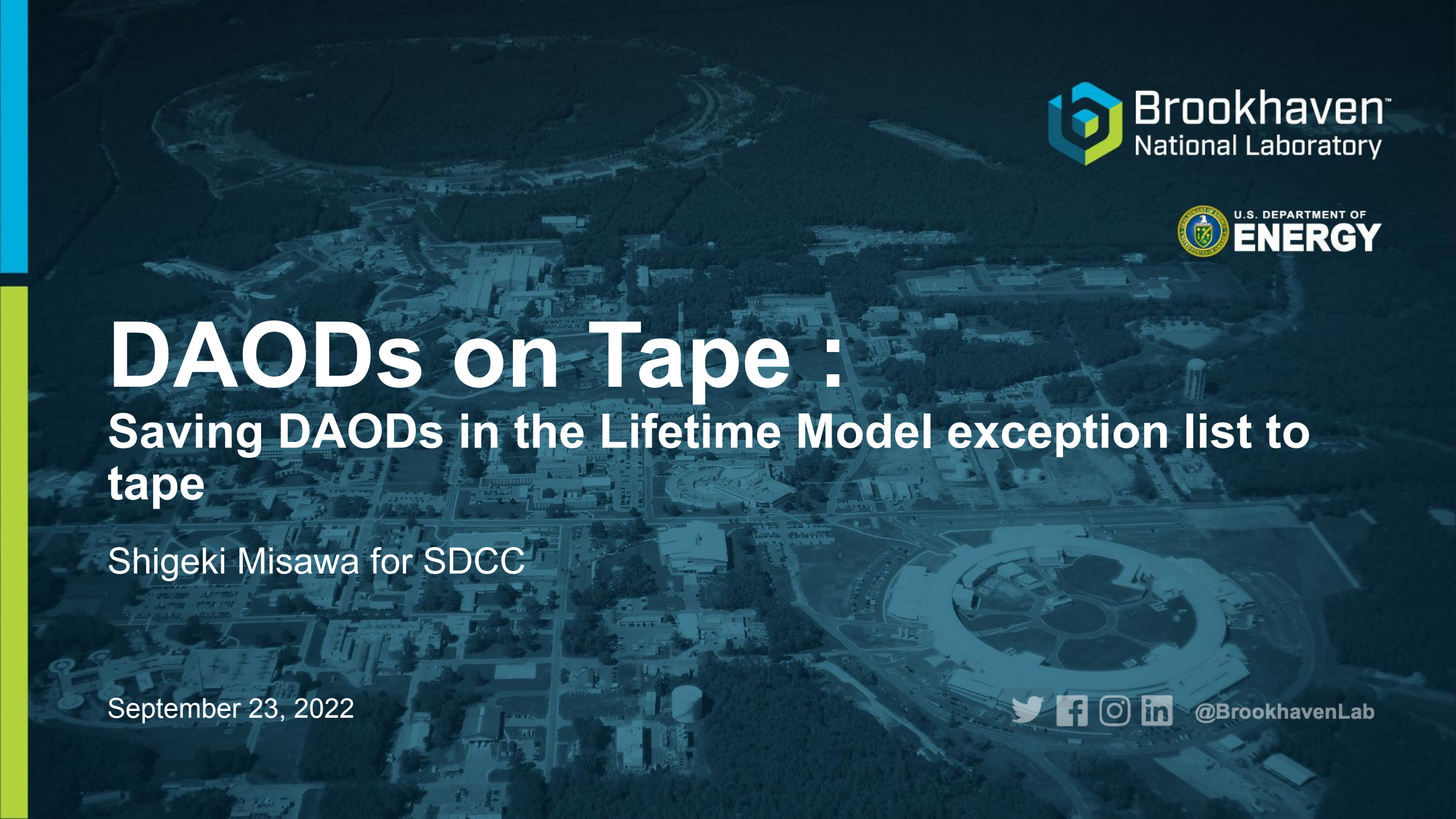


Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

DAODs on Tape :

Saving DAODs in the Lifetime Model exception list to tape

Shigeki Misawa for SDCC

September 23, 2022

@BrookhavenLab

Proposed New Uses of Tape

- Proposal to move DAOD datasets in Lifetime Model exception list to tape [1,2]
- Two new workflows have been proposed to handle requests for these DAODs on tape [2,3]
  - DAODs on Demand - Allow users to access DAOD directly from tape
  - DOADs from AOD on Demand - Satisfy requests for DAOD on demand by recreating them from their parent AODs on tape

[1] https://docs.google.com/presentation/d/1F0hrxzH3DuSaqygf36Y2IlxYddE2KnM_1P8gXDJJfUU/edit#slide=id.gcb8421af42_0_88
[2] Data Carousel R&D ideas (cern.ch)
[3] https://indico.cern.ch/event/1189287/contributions/4998966/attachments/2491909/4279511/DAOD_nAOD_popularity_Aug2022.pdf

Tape System Limitations

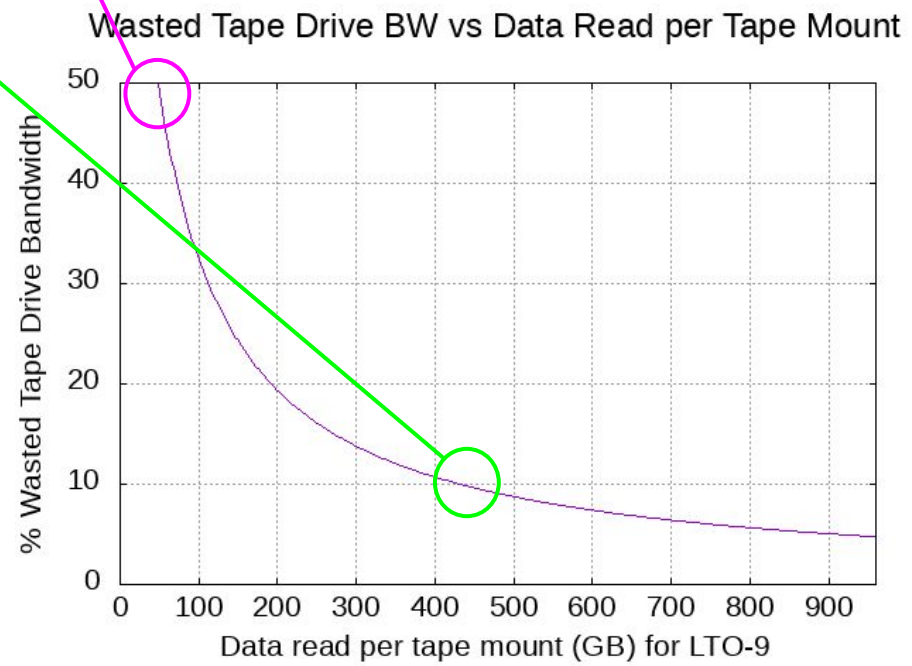
- By default there is one QoS (Quality of Service) level
  - You get it when you get it
  - DAOD read requests are competing with other ATLAS read requests
    - LTO-9 - end to end tape read takes 13.1 hours
  - Read optimizations favor maximizing data throughput over reducing time to first/last byte in a dataset
    - But some fair share mechanisms are in place to prevent complete starvation
  - Additional QoS levels requires development effort
- “Smart Writing” (Tailoring dataset layout on tape to optimize reading)
  - Items brought up in meetings are only enablers of smart writing
    - e.g., Minimizing dataset transfer windows, dataset manifests, using file families
  - Achieving smart writing in reality requires additional work “under the covers”.
    - All enablers may not be usable by all sites
    - Uncontrollable factors may make it impractical or impossible to achieve complete smart writing
      - e.g., Striping over tapes to achieve bandwidth requirements may conflict with dataset colocation goals to reduce the cost of tape mounts
  - Pareto principle (80/20 rule) applies when trying to achieve complete smart writing

Resource Limitations

- @ < 50GB read per tape, mount costs start to dominate
  - For data blobs < 50 GB, metric is # data blobs retrieved per minute (not MB/sec)
  - Effective use of tape drives requires reading at least 480GB of data from a tape
- There are limits to increasing # tape drives
  - $$ (cost)
  - ~128 tape drives max per tape library
    - Also all drives may not be usable as the library must accommodate multiple generations of tape drives
- Tape Mounts are also limited
  - 700 tape mounts per hour per tape library

$$F_{Waste} = \frac{1}{1 + T_{Read}/T_{Move}}$$

F_{Waste} = Fraction of wasted BW
T_{Move} = Mount/Seek time
T_{Read} = Time Reading



Brookhaven National Laboratory

Dataset Ingest - Unanswered Questions

- What ingest bandwidth needs to be supported (burst and average) ?
  - Need 28 tape drive-days per petabyte @400MB/sec per tape drive running 24x7
  - Alternatively, how quickly do DAODs need to be removed from disk ?
    - This impacts network bandwidth, disk cache configuration and required # of tape drives
- Dataset transfer window size distributions requires enumeration to determine impact on disk capacity and ability to colocate datasets
- Number of dataset being transferred concurrently of interest for the same reasons
- Ability to tune dataset transfers to an advantageous configuration
- Are there correlated datasets
  - Datasets that will be read back together
- DAOD dataset size distribution
  - Verify that average DAOD dataset size is representative of most datasets

DAOD Access (DAOD on Demand)

- Read access requirements needed
  - What staging bandwidth (BW off tape) is require for a dataset AND what export bandwidth (BW from disk to client) is required ?
    - Latter will be higher if multiple clients read the same dataset
  - Are there upper bounds on access latency that cannot be exceeded ?
    - Time to the first and last byte in the dataset, from the time that the initial request was issued
  - How long does the dataset need to be available ?
    - This affects the size of the disk cache
  - How large is the request time window for a dataset
    - i.e., time between receiving first stage request for a dataset file to the last stage request for dataset file.
  - Aggregate DAOD/AOD access requirements
    - What number (peak and average) of jobs generating tape request are expected at any given time
      - i.e., what aggregate capability needs to be deployed

Backup Slides

Candidate DAOD dataset information

- “Short lived data to tape” slide [1] states the datasets of interest satisfy the following conditions
  a. Flagged by DDM Operations for deletion in quarterly run of the Lifetime Model Campaign, **AND**
  b. Flagged by users for a “lifetime extension”
  c. Total volume of datasets moved to tape is ~4PB per campaign
     - (55 to 85 LTO cartridges, depending on generation)
  d. Less than 5% of datasets moved to tape will be accessed in the 12 months after storage. Slide #5 in [2] seems to suggest that higher percentage will be accessed with 12 months

[1] https://docs.google.com/presentation/d/1F0hrxzH3DuSaqygf36Y2IlxYddE2KnM_1P8gXDJJfUU/edit#slide=id.gcb8421af42_0_88
[2] https://indico.cern.ch/event/1189287/contributions/4998966/attachments/2491909/4279511/DAOD_nAOD_popularity_Aug2022.pdf

Dataset Lifecycle[1]

- All datasets from a given campaign will be stored at one Tier 1
- Tier 1 will host datasets from at most one campaign on tape
  - Destination of campaign data will rotate among the Tier 1s
  - Campaign data will be deleted after one year
  - For any given Tier 1, all previous campaign data will be complete deleted, before receiving the next round of campaign data.
- Dedicated spacetoken will created for the datasets

[1] https://docs.google.com/presentation/d/1F0hrxzH3DuSaqygf36Y2IlxYddE2KnM_1P8gXDJJfUU/edit#slide=id.gcb8421af42_0_88
[2] https://indico.cern.ch/event/1189287/contributions/4998966/attachments/2491909/4279511/DAOD_nAOD_popularity_Aug2022.pdf