The Sequential Access Model for Run II Data Management and Delivery

Wolfgang Kowald, *Lee Lueking*, Kurt Ruthmansdorpher,
Heidi Schellman, Vicky White.

URL: www-d0.fnal.gov/~lueking/cdr2dma/sequential.html.

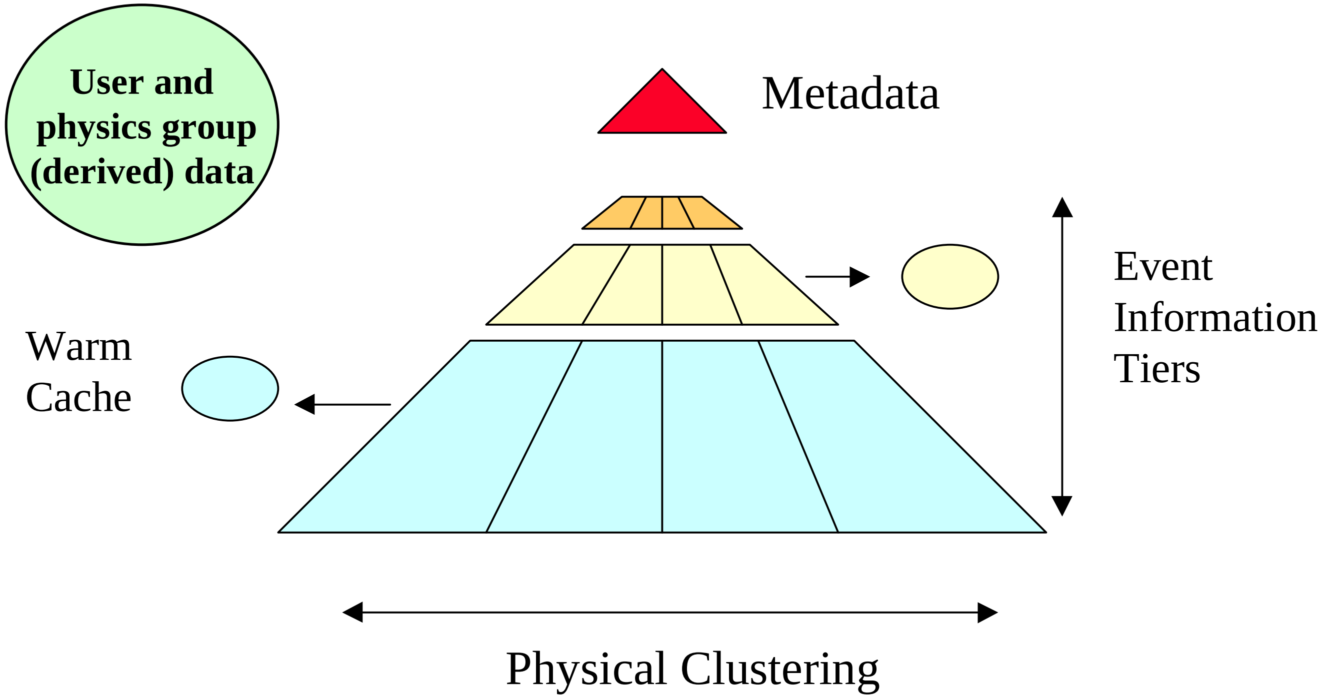
Run II Von Ruden Offline Review

October 13, 1997

What is The Sequential Access Model?

- Sequential events: Data is stored in files as sequential events.
- Data Tiers: Each event is stored in each of several data tiers.
  - » The Event Data Unit (EDU) is the unit of data stored in each tier.
  - » Physical event size: EDU5=5kB/event, EDU50=50kB/event, et cetera.
- Physical streaming (clustering): data categories based on Trigger or reconstructed info. (Following example; n categories)
  - » Inclusive: Each stream is complete; shares events with other streams: n streams.
  - » Exclusive:Each stream is either a single category, or overlap: (2^n-1) streams.
  - » Sub-event: Pieces of events are physically clustered together.
- Metadata: Event Management Database (EDB); Information about the data - event-level, file-level, run-level ...

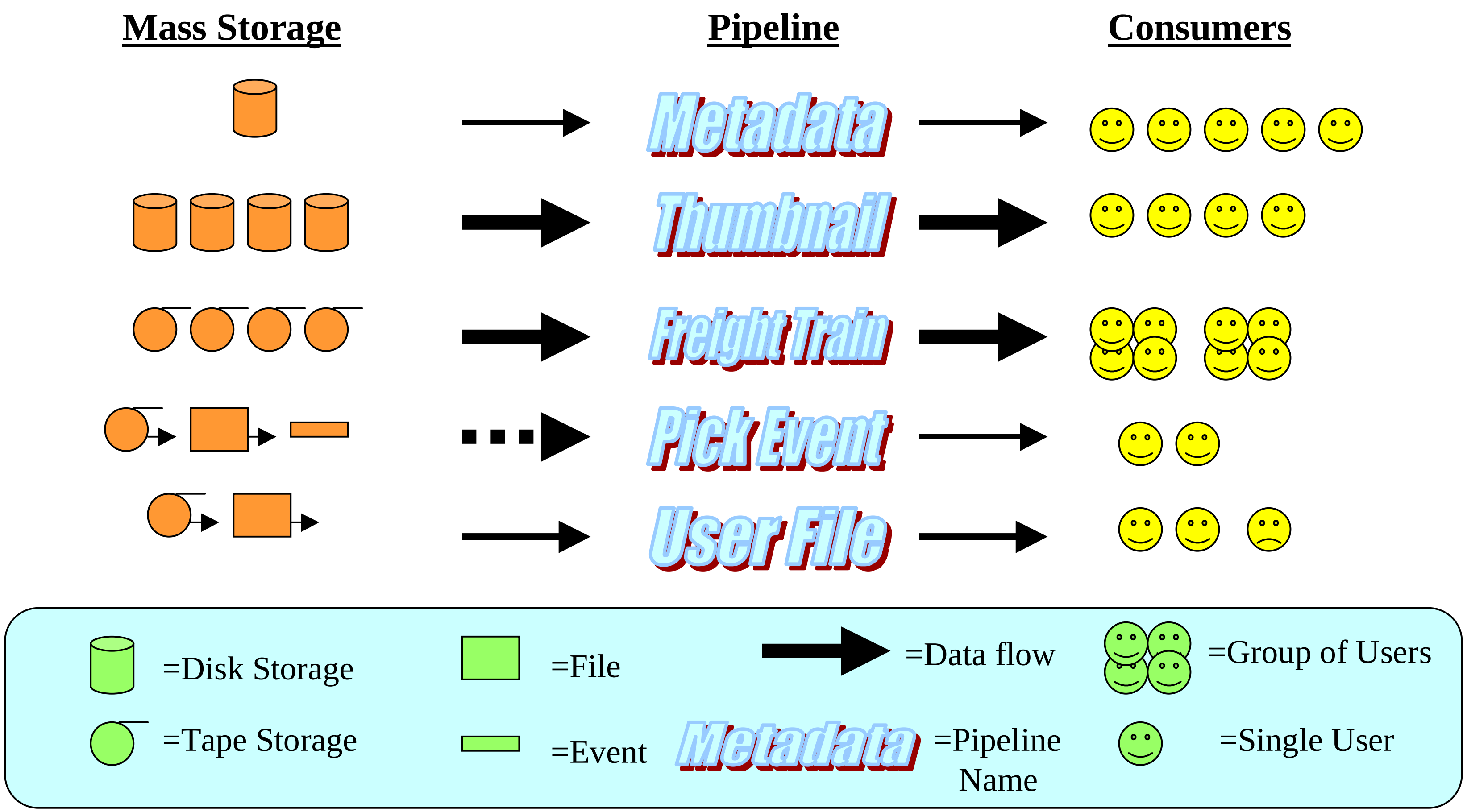
Data Organization



How Do I Access Data?

- Pipelines: Data access channels tailored for particular processing and analysis patterns.
- Pipeline segments: Tapes, drives + ATL + HSM + network, group-shared and user-private analysis disk.
- Defined pipelines (currently):
  - » Metadata: Access to event trigger & other EDB info.
  - » Thumbnail: Disk resident sketch of each event.
  - » Freight Train: Large data stream event server.
  - » Event Picking: Random event selection from any data tier.
  - » Small Data-set: One or a few files from any data tier.

Data Access



Mass Storage
Pipeline
Consumers
Metadata
Thumbnail
Freight Train
Pick Event
User File
=Disk Storage
=Tape Storage
=File
=Event
=Data flow
Metadata
=Pipeline Name
=Group of Users
=Single User

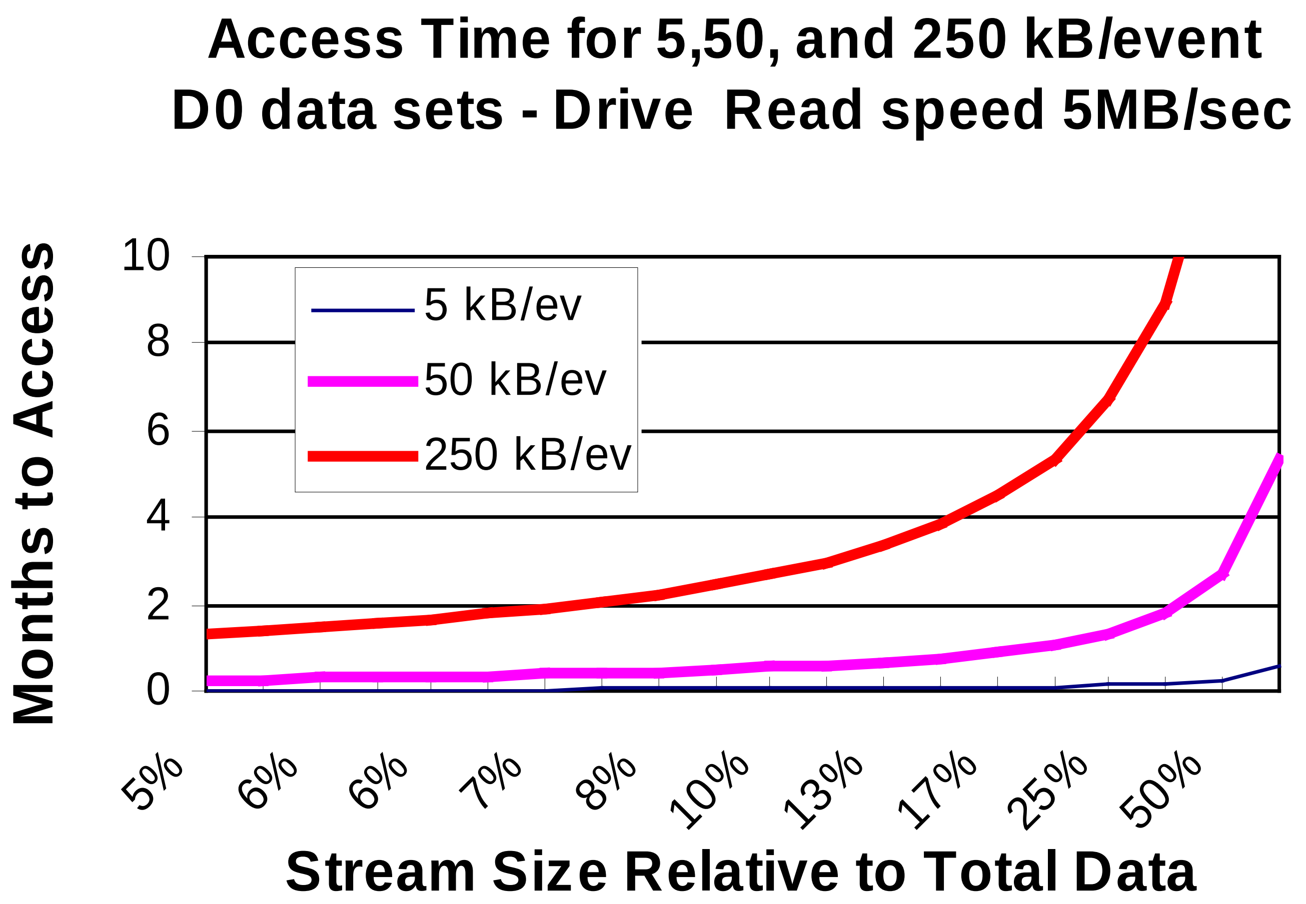
CDF & D0 Specifications

- Data tiers

	CDF	D0	Comment
Events	1.8e9	1.2e9	
Raw size	250kB	250kB	
Big EDU	130 kB	300 kB *	*< 20% of events
Med. EDU	63 kB	50-100 kB	
Small EDU	10 kB	5-10 kB	Stored on disk
Metadata	10-50 B	40-50 B	Stored on disk

- CDF
  - » 20 streams with 30% overlap. Based on L3.
  - » 10% express stream
  - » Phases: ramp-up, stable analysis.
  - » May consider sub-event partitioning.
- D0
  - » 20-30 exclusive streams preferred. Based on L3 & Reco information.
  - » 10% Warm caches of RAW and DST data.
  - » Possible on-demand reconstruction.

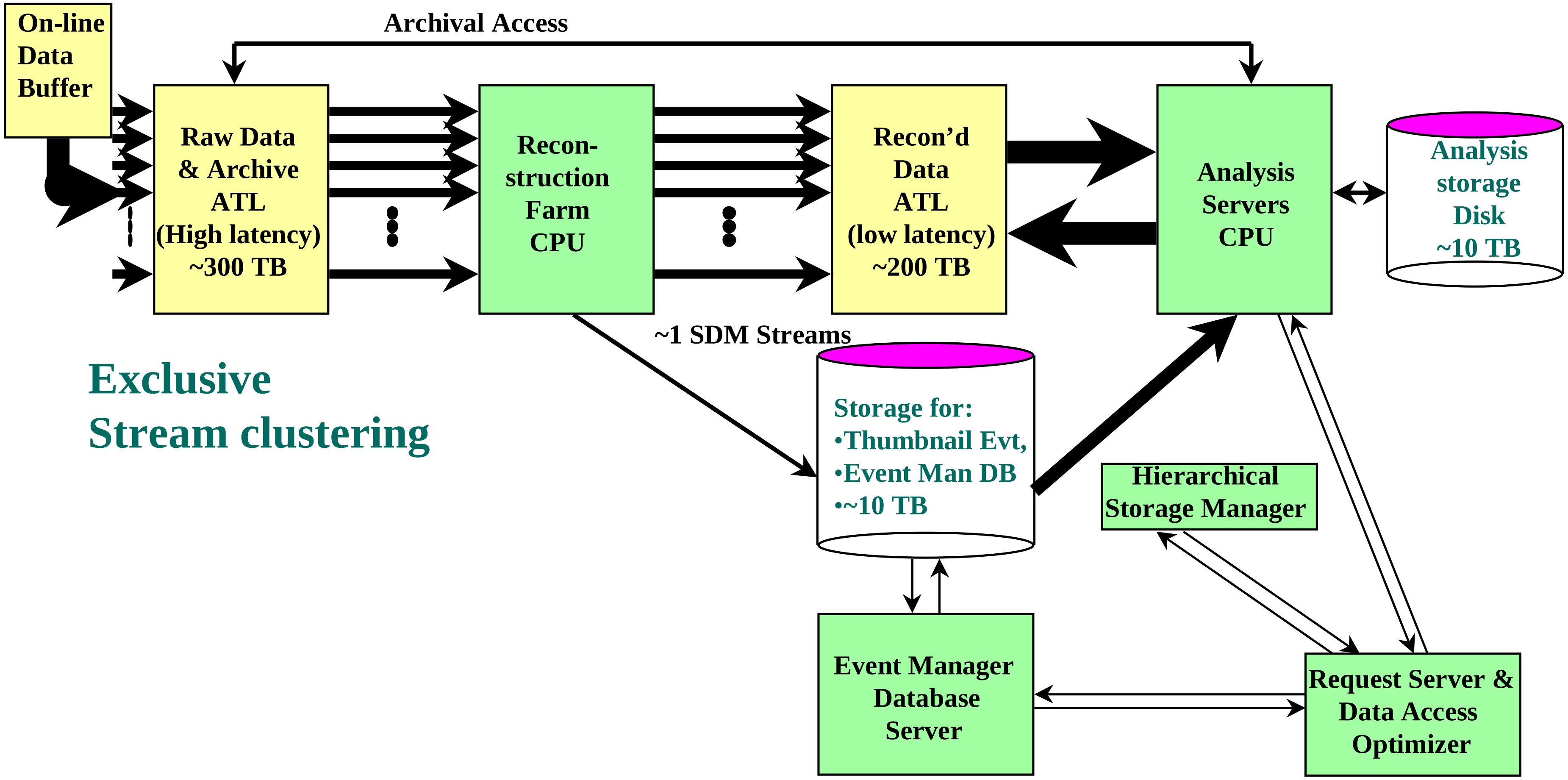
Will it Scale to Run II?



Components of the Design

- Physical Streaming: Possible stream categories.
- Event Management Database (EDB): Well organized metadata with event-level granularity.
- Request Server & Data Access Optimizer: All requests are submitted and processed through a central clearinghouse.
- Thumbnail event
- User and administrative interfaces.
- More about pipelines
- Other tightly linked components and concerns
  - HSM system to manage data on tape
  - Tightly coupled data and analysis CPU
  - Integration with Reconstruction Farm Processing
  - Storage and Access to user or group data sets derived from primaries

Data Management & Access Model

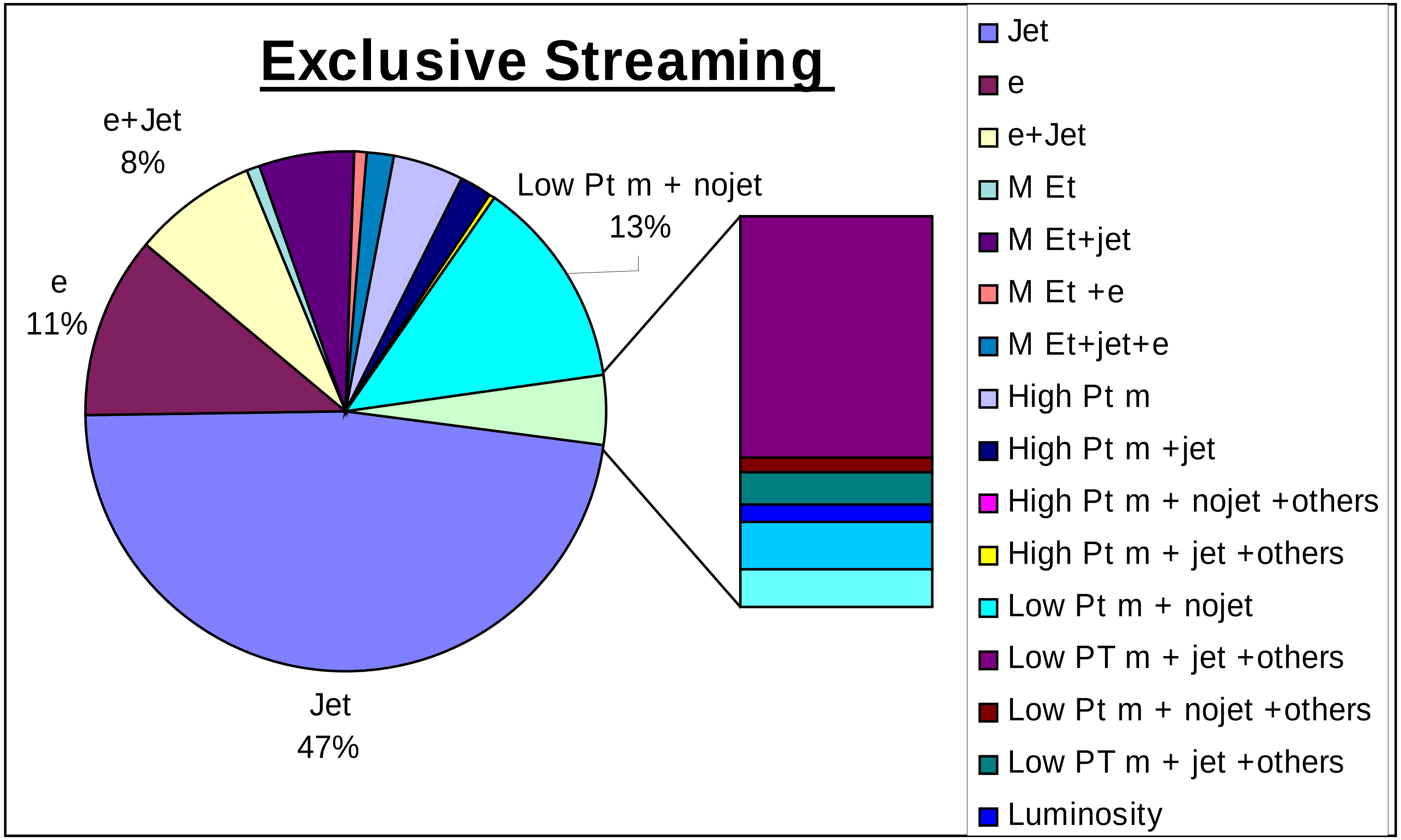


Exclusive Stream clustering

SDM = Drive-Month = 5MB/sec x 26E5sec = 13 TB
SDD = Drive-Day = 5MB/sec x 86E3sec = 0.43 TB

Stream Size Distribution

- Study based on 100k events from Run I D0 data sample



Streaming Concerns

- Data Integrity (Exclusive mode):
  - » Problem: Loss of data in overlap streams is serious.
  - » Solution: Multiple copies of small overlap streams will be required.
- Luminosity Tracking (Exclusive mode):
  - » Problem: Difficult to track integrated luminosities.
  - » Solution: Track Luminosity for each trigger in Virtual Partitions (VP) with fixed event ranges. Map physical streamed partitions onto the VPs.
- Quality Cuts (Exclusive mode):
  - » Problem: Post-reconstructed streams defined with quality cuts may be complicated to use.
  - » Solution: Use quality tags for overlap streams.
- Sorting Streams to Serial Media:
  - » Problem: Difficult to get each stream sorted into its own set of tapes. Solution: Compromise by mixing small streams on tapes. Exploring various hardware options.

Contents of the EDB

- EDB will require ~50 bytes/event; total size 50-100 GB
- Anticipate use of commercial database (considering Oracle).

Event Level Info (1.2×10^9 entries):
Run[8], Event[8],Trigger bits[32],
Filter bits[32], Cluster Category[2],
Virtual Stream[16]

File Event Catalog (5×10^6 tables):
Event,Record, byte offset for each event

File-level info (5×10^6 entries):
Filename[40],run[8],Cluster Category[2],
Event range[16], Merge links[?]

Processing info (1×10^3 entries):
processing features, e.g. reco version,
filter steps, etc.

Run-level info (5×10^4 entries):
Links to on-line info, conditions db,
luminosity, additional trigger info, etc.

Logical Containers (1×10^3 tables):
Lists defined by physics-groups or
users as interesting data sets.

Request Server & Access Optimizer

- All user requests go through a central Request Server.
  - » The request is evaluated.
  - » User is sent an estimate time to execute.
  - » In conjunction with AO, request is processed.
- The Access Optimizer pools and sorts requests to efficiently use hardware.

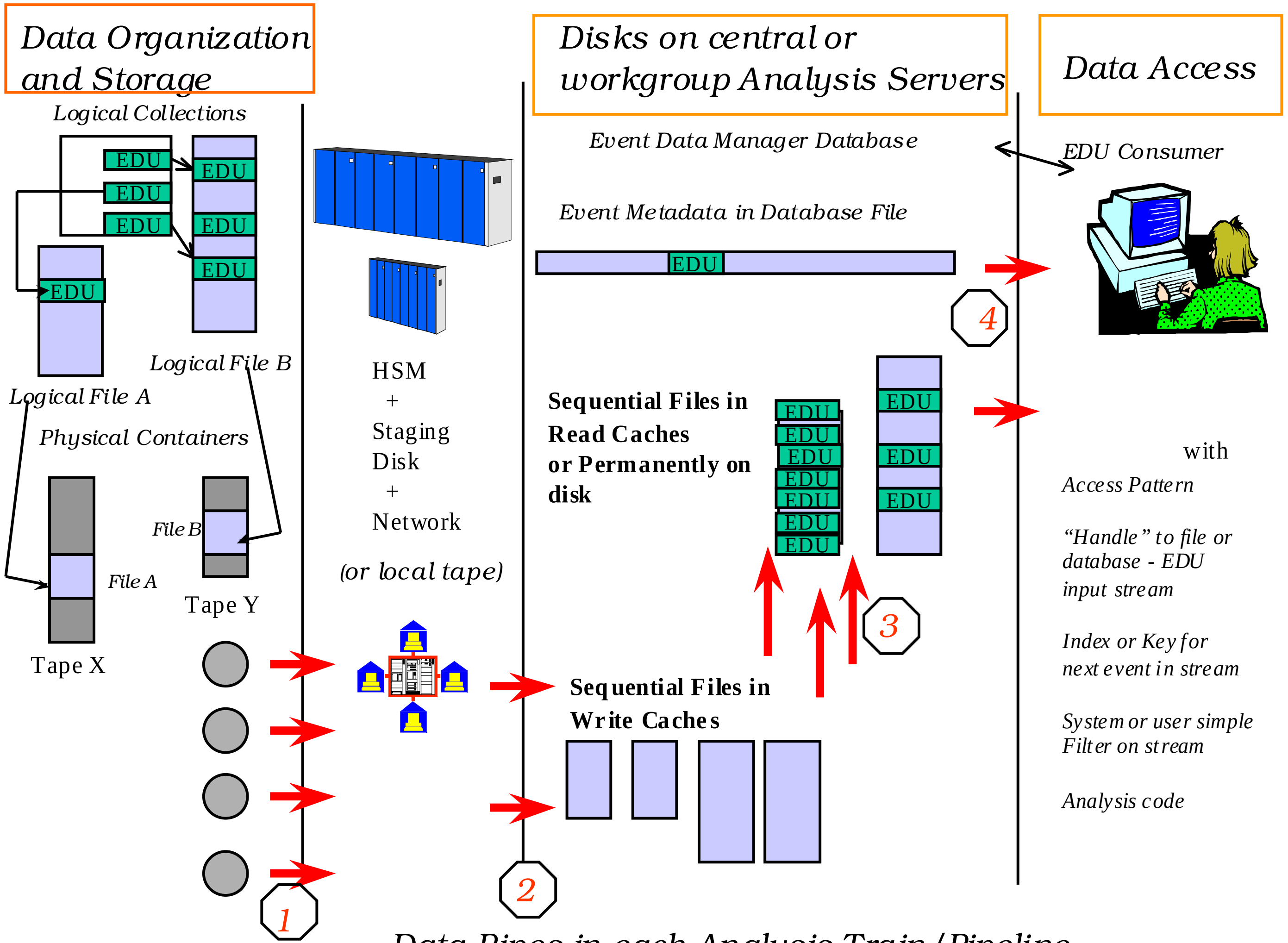
Thumbnail Data

- Highly evolutionary data set
- Tightly integrated to EDB
- Efficient and flexible access required
- Possible Format Options
  - » OODBMS
  - » Conventional N-tuple
  - » DSPACK, TRYBOS
  - » Other e.g. ROOT

User and Administrative Interfaces

- **User interface**
  - » GUI for creating datasets based on EDB info. Can also use pre-defined sets.
  - » Batch interface for just-in-time data staging.
- **Administrative interface**
  - » Information to understand access patterns
  - » charts, graphs, tables showing hardware usage.

Pipelines



Data Pipes in each Analysis Train/Pipeline

Advantages of SAM

- Easy to understand, implement & debug
- Implementation can be staged
- Applicable to all stages of data taking
- Little potential for corruption
- Consistency is easily verified
- Data and Management are de-coupled
- Well matched to the technical expertise available at the lab

Summary and Plan

- Mission: To provide an “integrated” system to deliver sequential event data to users which is efficient, reliable, easily managed/used and scales to anticipated demand.
- Currently: Developing detailed model which includes usage patterns, hardware bandwidths, database design and functionality.
  - » Testing with spreadsheet evaluation.
  - » Work closely with RIP to understand & test ATL and HSM needs.
- Future (4-6 months): Create a real Event Management db and populate it. Simulate and prototype all major sub-systems.