Object Stores for CMS data

Nick Smith, Bo Jayatilaka, David Mason, Oliver Gutsche,
Alison Peisker, Robert Illingworth, Chris Jones (FNAL)

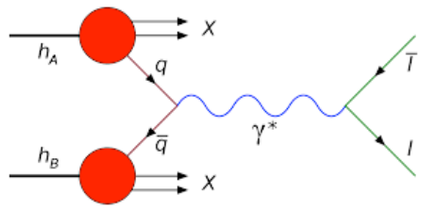
ACAT 2022

28 Oct 2022

Rothko, Number 19 (1949)

Primary dataset

Abstract, “what kind of events.”
e.g. hard scatter process for simulation, trigger filter for data

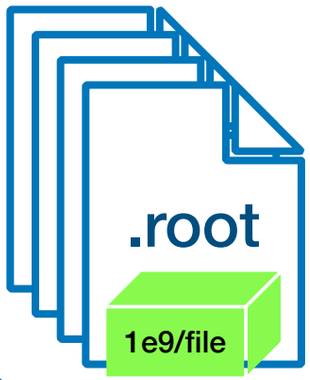


Data tiers

AOD

1e5/event

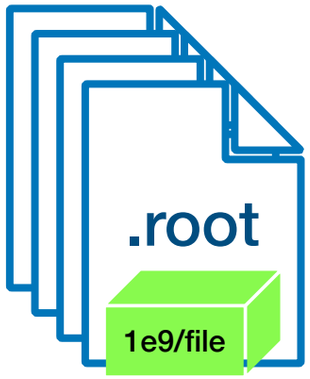
Data columns pertaining to low-level reconstruction



MiniAOD

1e4/event

Calibrated physics objects
Particle-flow candidates

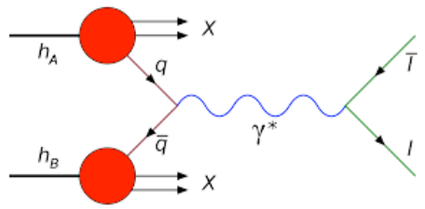


. . .

Data volume
order of magnitude
[bytes]

Primary dataset

Abstract, “what kind of events.”
e.g. hard scatter process for simulation, trigger filter for data

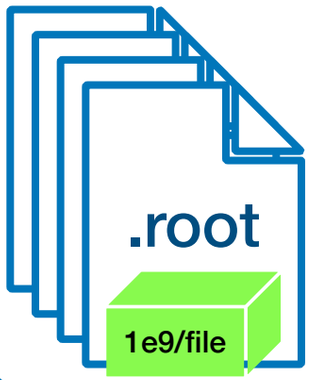


Data tiers

AOD

1e5/event

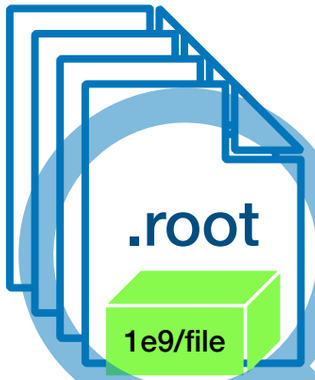
Data columns pertaining to low-level reconstruction



MiniAOD

1e4/event

Calibrated physics objects
Particle-flow candidates

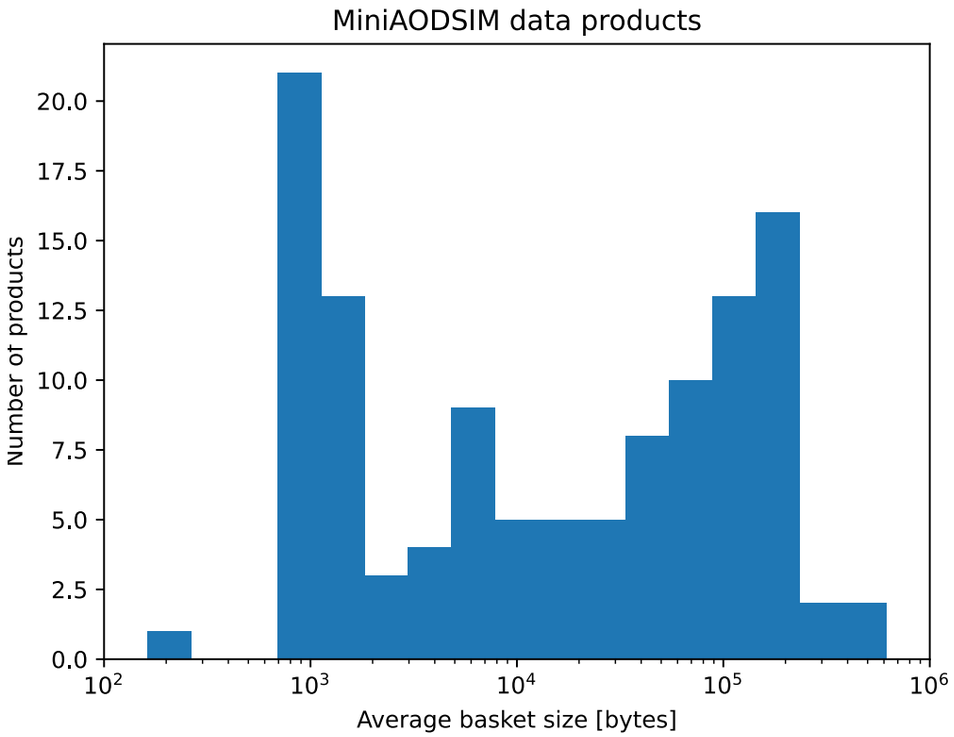


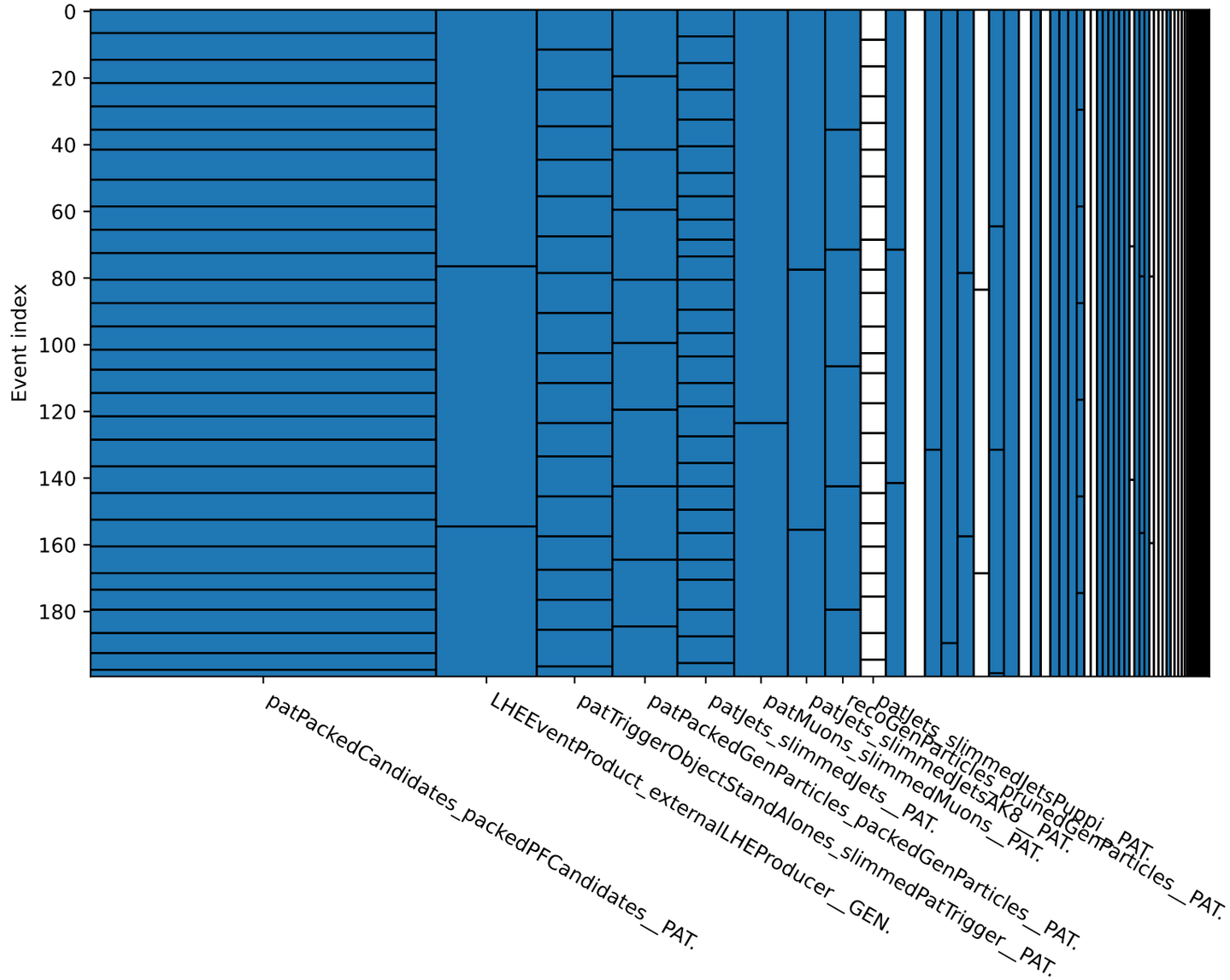
. . .

Data volume
order of magnitude
[bytes]

File format

- Event Data Model (TTree)
- Branch: metadata about C++ data type, basket positions
- Basket: serialized C++ objects stored contiguously*

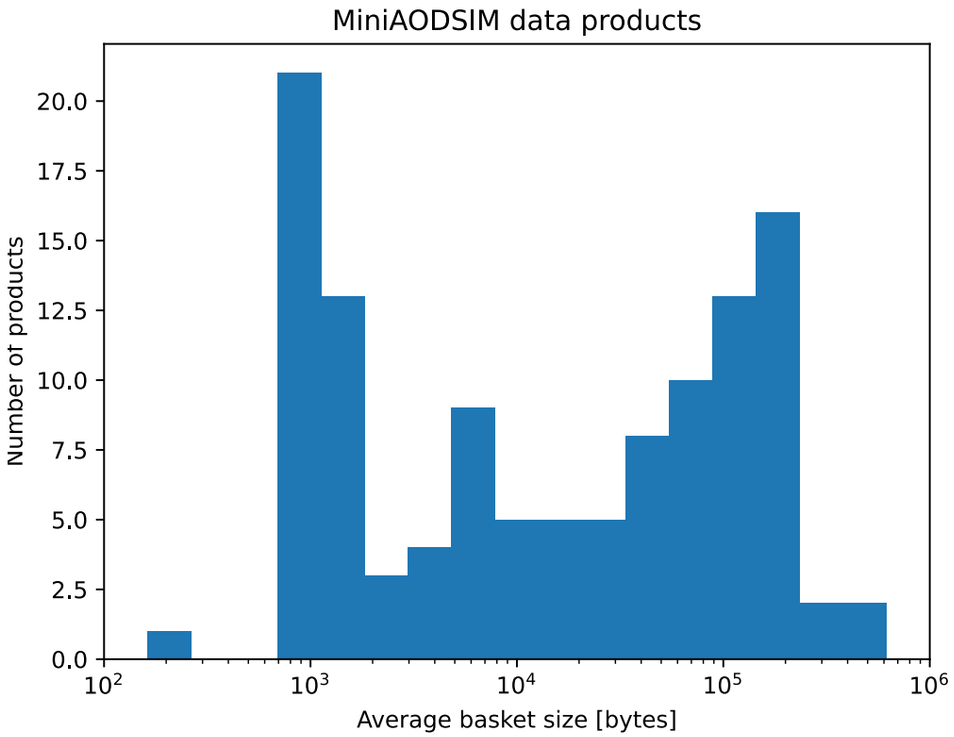


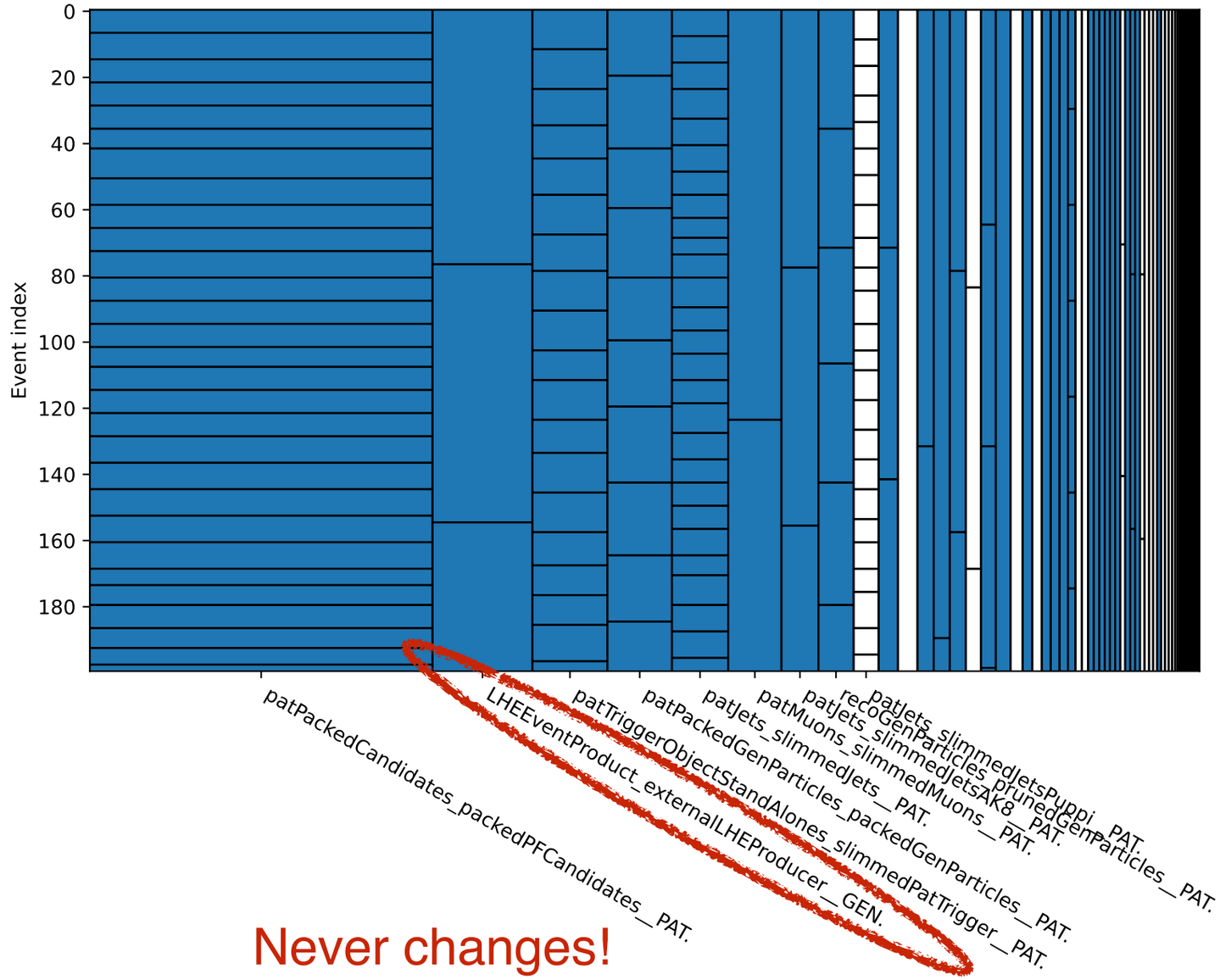


Fermilab

File format

- Event Data Model (TTree)
- Branch: metadata about C++ data type, basket positions
- Basket: serialized C++ objects stored contiguously*



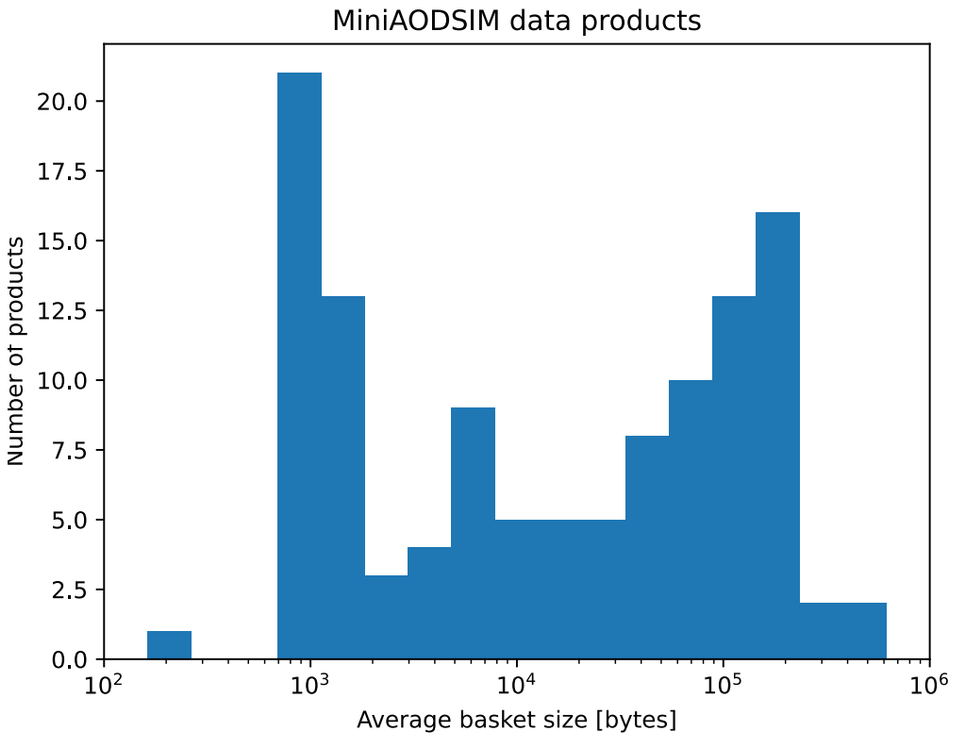


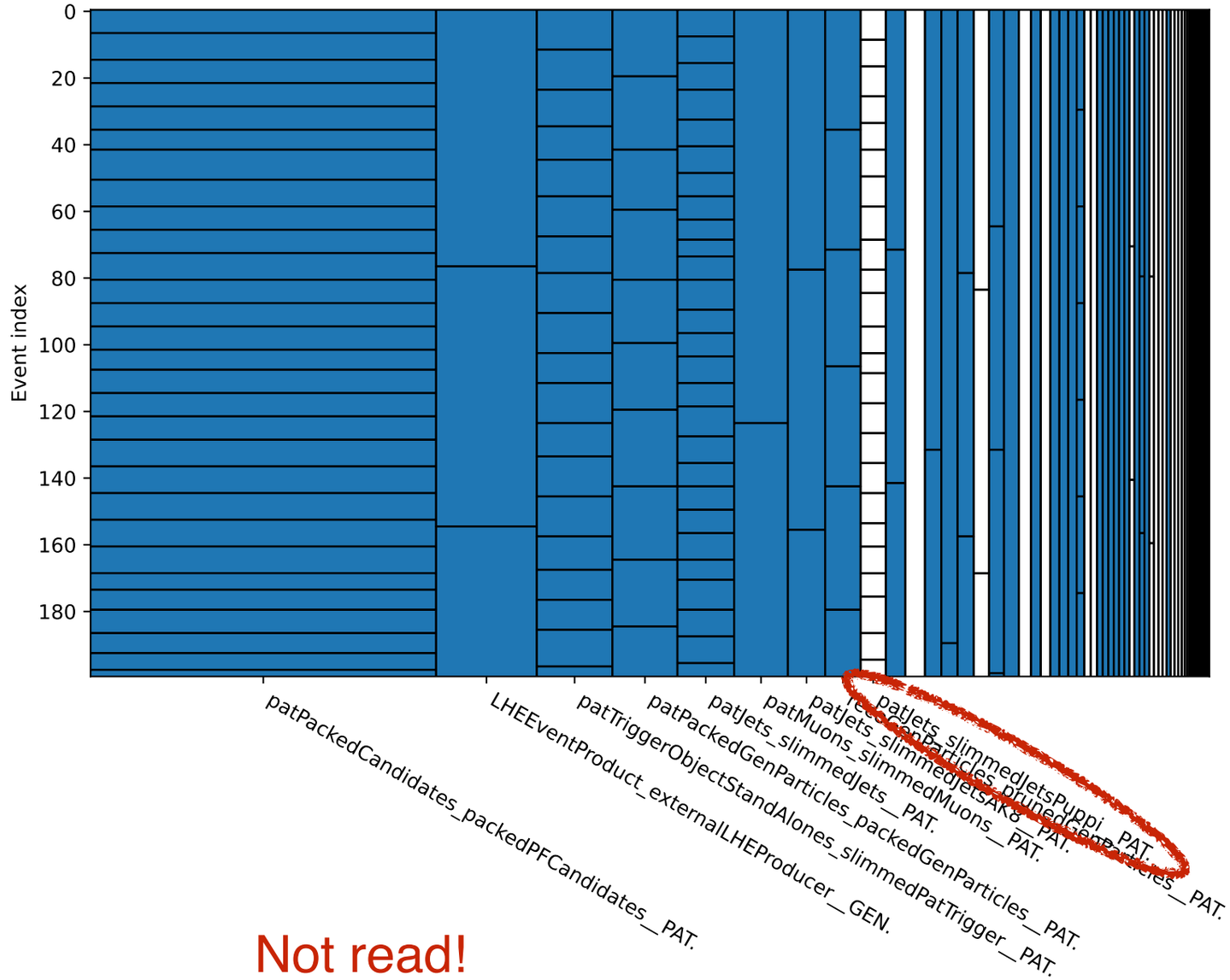
Never changes!

Fermilab

File format

- Event Data Model (TTree)
- Branch: metadata about C++ data type, basket positions
- Basket: serialized C++ objects stored contiguously*

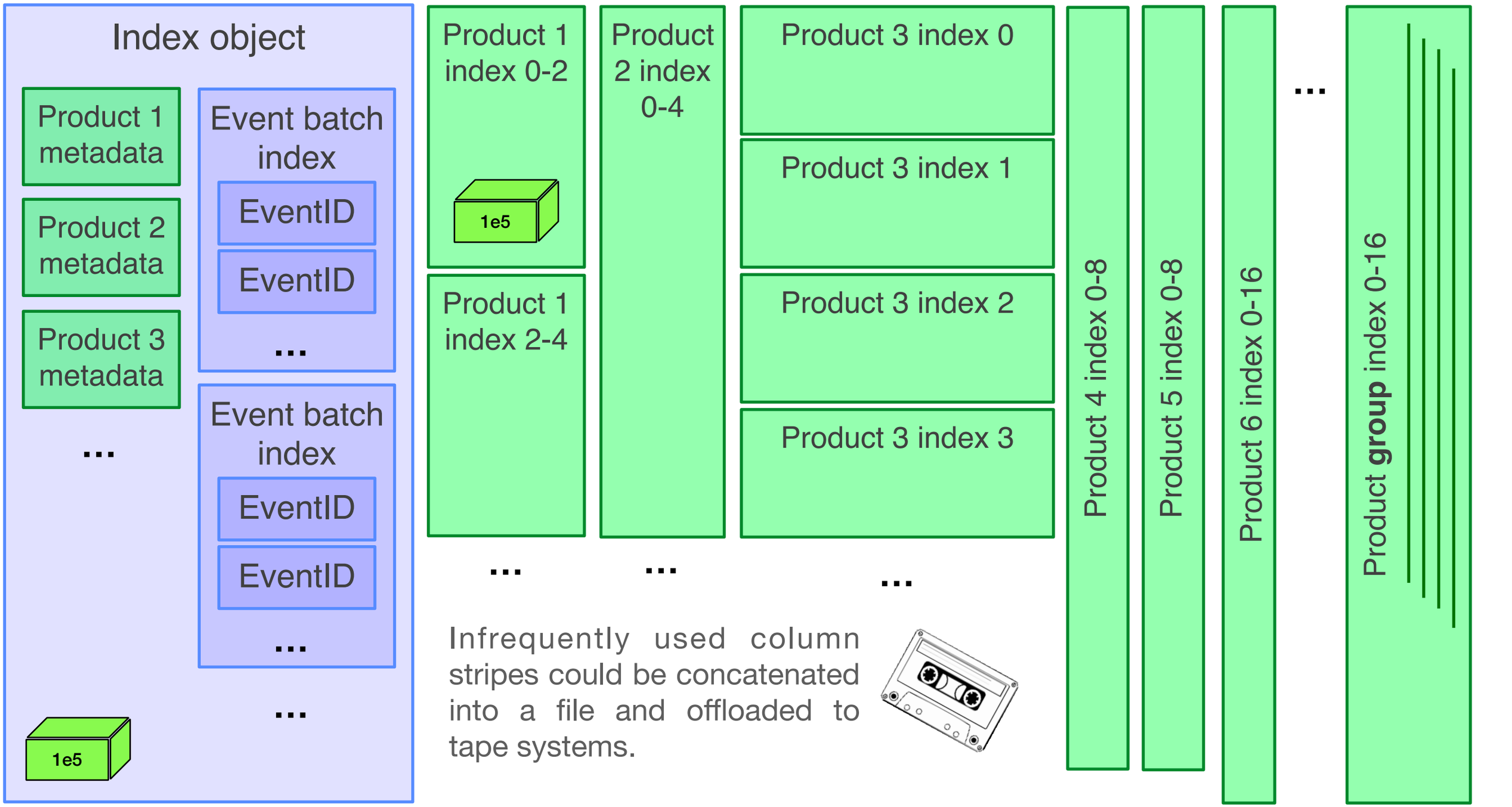




Not read!

Fermilab

Object data format



Fermilab

Strawman

Tier-based scheme

MiniAOD Data product	KB per event	
	v1	v2
packed+pruned genParticles	5.7	5.7
slimmedElectrons	1.3	1.3
Others	48.7	48.7
Total	55.7	55.7

Object store scheme

MiniAOD Data product	KB per event	
	v1	v2
packed+pruned genParticles	5.7	-
slimmedElectrons	1.3	-
Others	48.7	-
Updated slimmedElectrons	-	1.3
Total	55.7	1.3

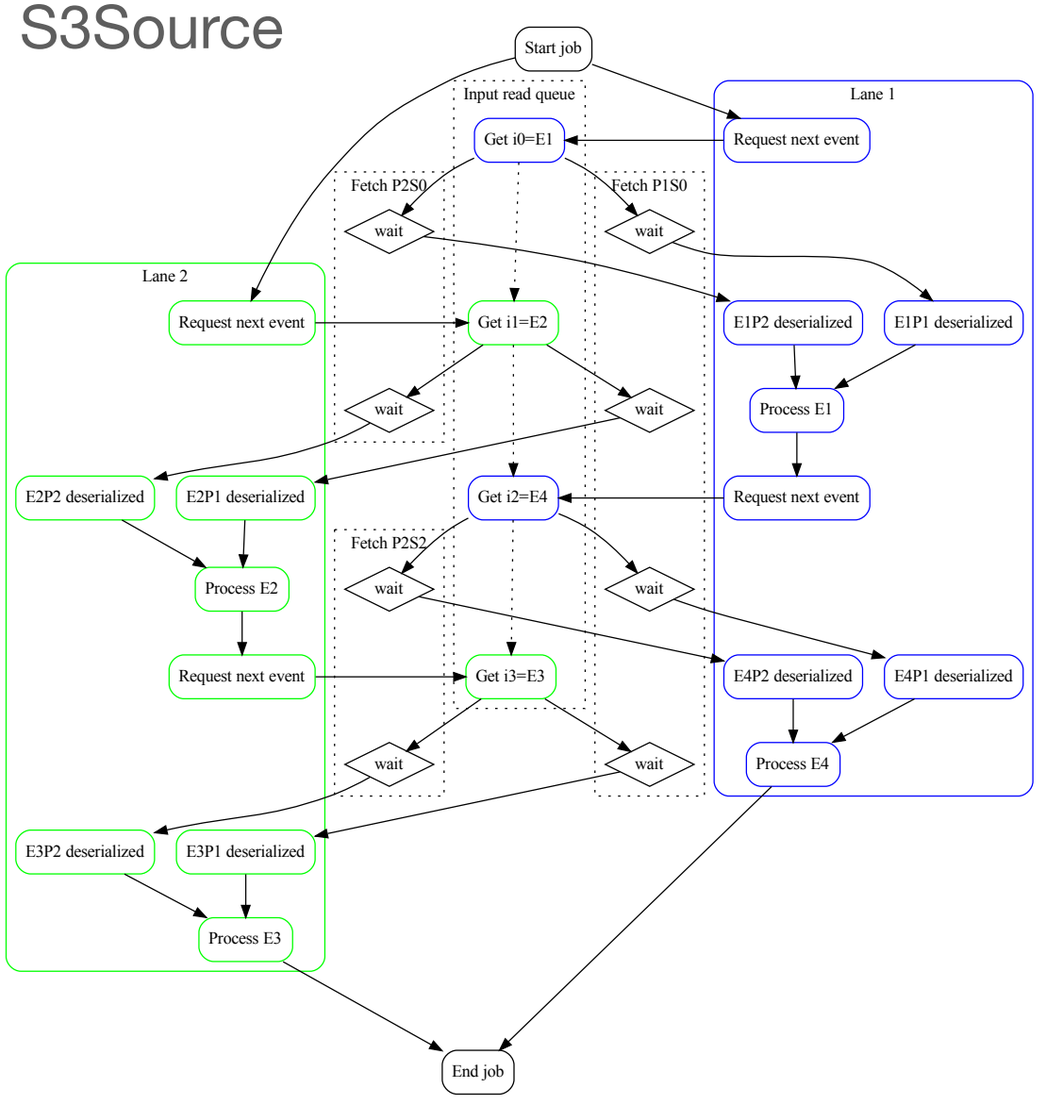
What to prove

- Read/write performance

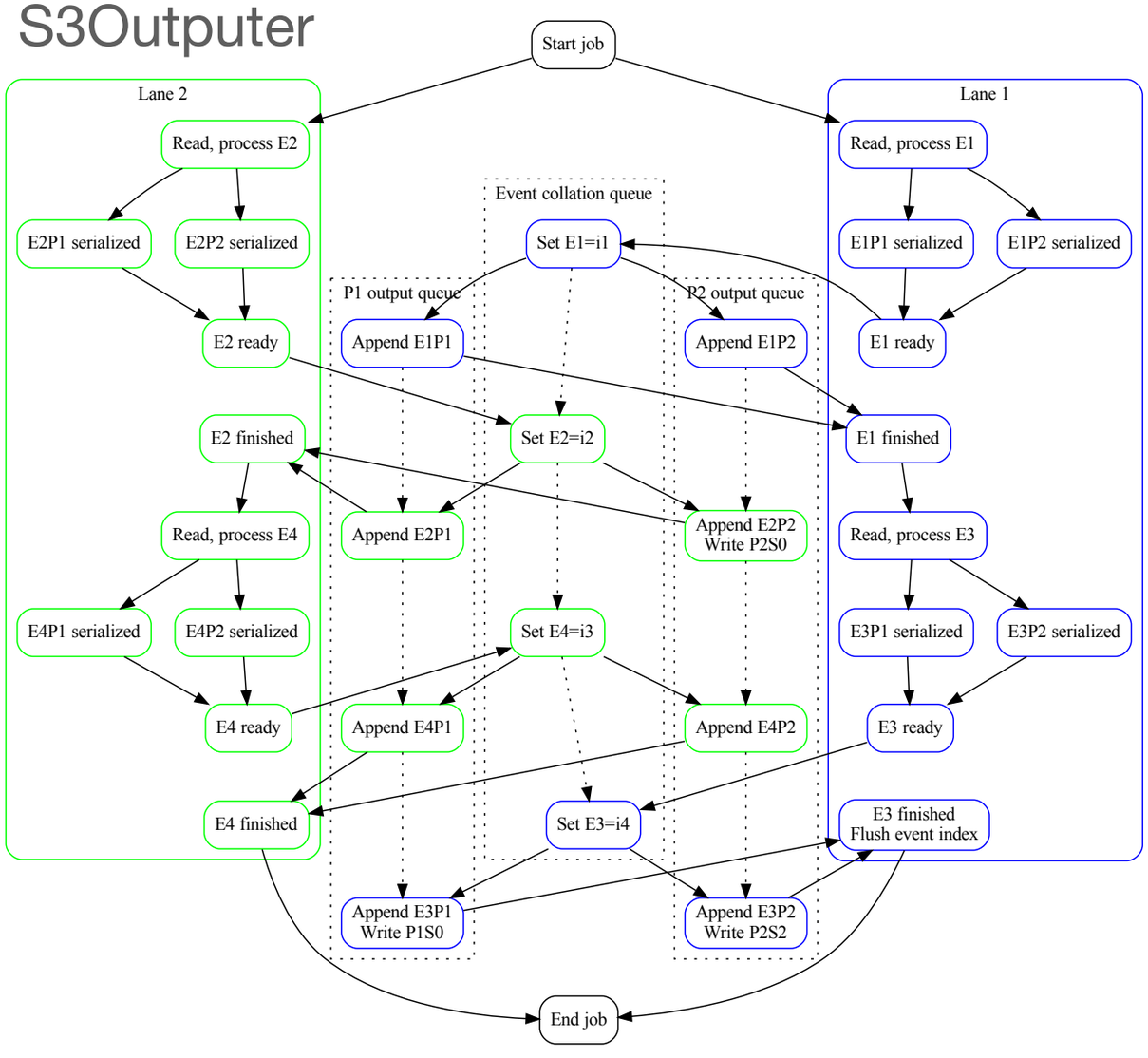
- Metadata scaling
  - Next time

Prototype framework

S3Source



S3Outputer



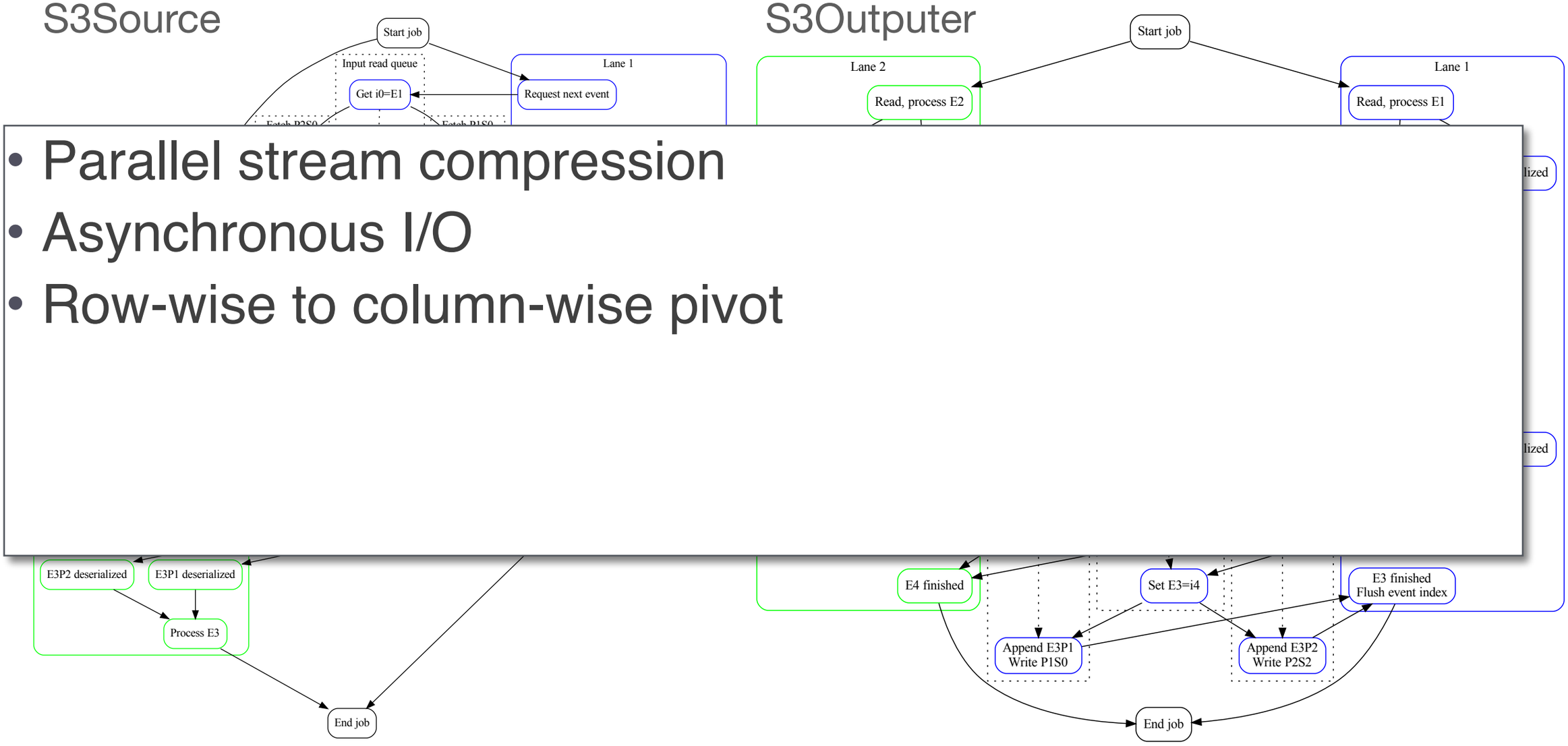
Fermilab

Prototype framework

S3Source

S3Outputer

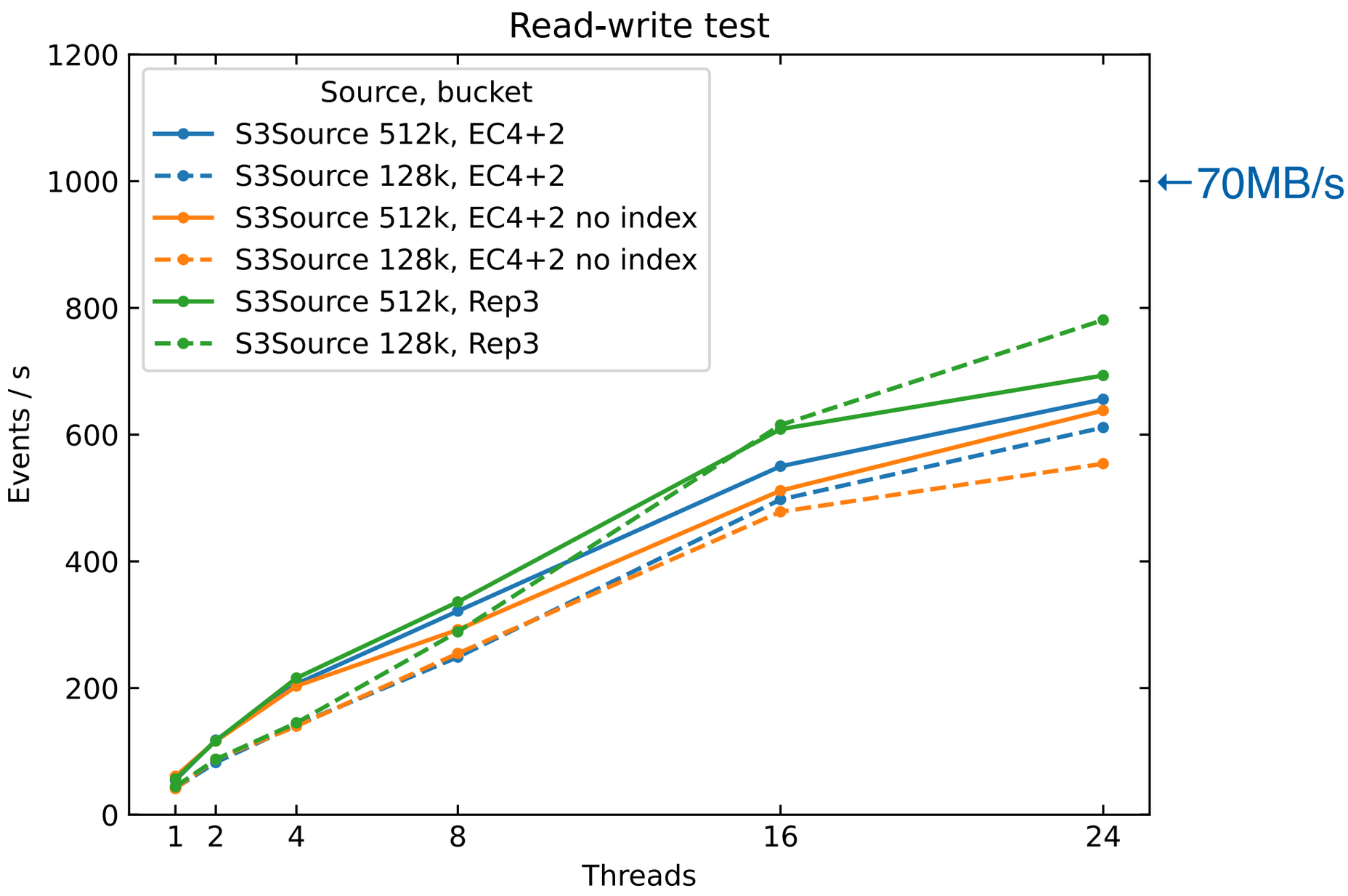
- Parallel stream compression
- Asynchronous I/O
- Row-wise to column-wise pivot



Fermilab

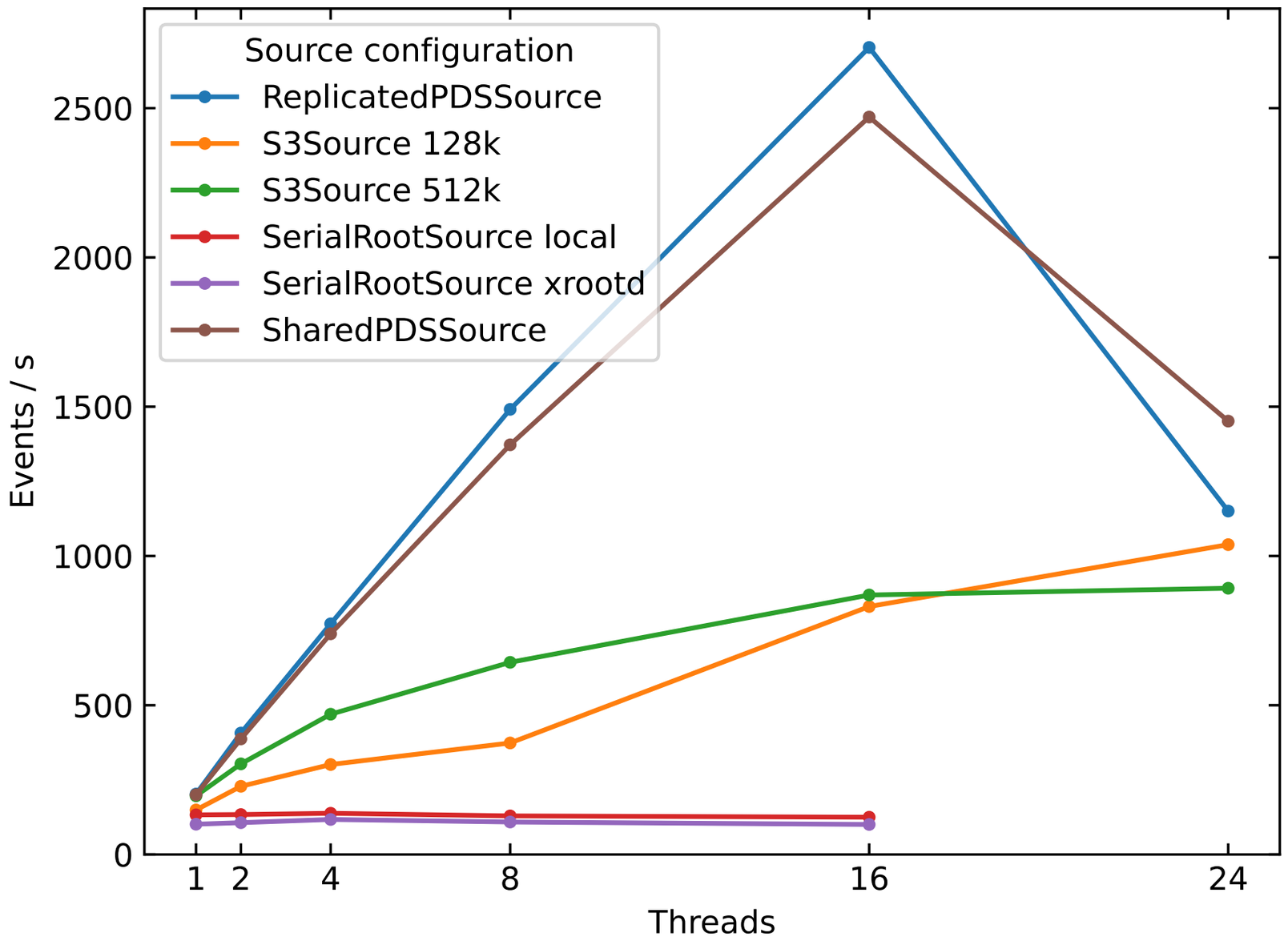
Prototype read-write performance

- Requirement: 1-10 event/s per thread
  - MiniAOD processing



Fermilab

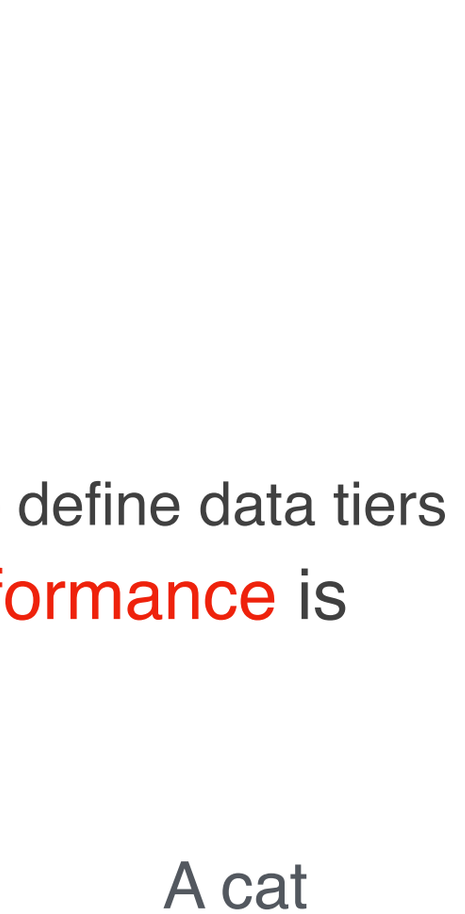
Read-only performance comparison



Fermilab

Conclusion

- Object data formats provide new data management capabilities
  - Compared to current tier-based EDM file model
  - Reduce disk storage requirements for re-processing, obviate the need to define data tiers
- In a prototype framework accessing a Ceph S3 service, I/O performance is excellent
  - On-disk data and metadata volume is as expected
- To fully utilize, more software development will be needed
  - New data management service requirement: column tracking
  - Full data format requires provenance and auxiliary data handling



P.S. for analysis facility use case, check out RNTuple: G. Miotto et al