Access Pattern Analysis in the EOS Storage System at CERN

Olga Chuchuk
CERN, IT department

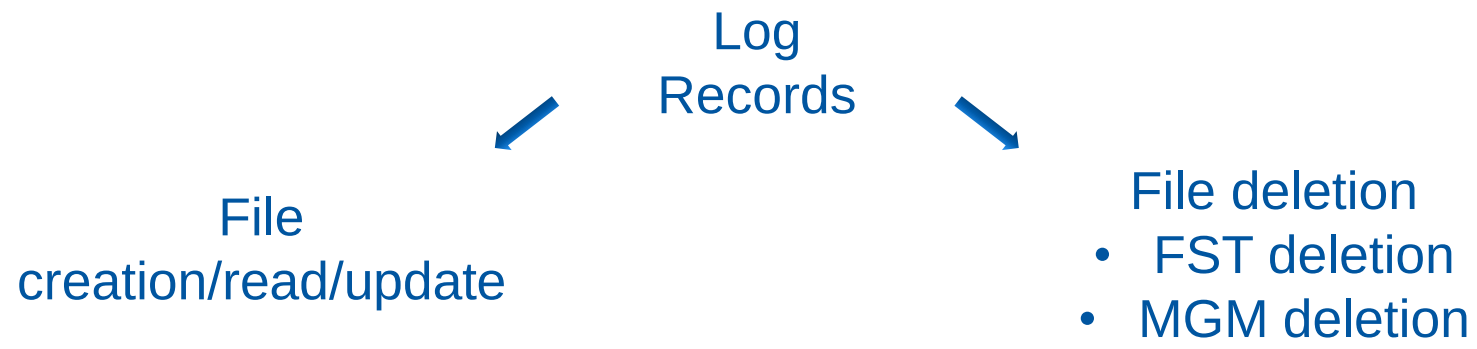
CERN LHC Detectors & EOS

Motivation

- Detailed data about EOS File Access is collected
- Useful for
  - Understanding differences between instances (experiments)
  - Understanding popularity and lifecycle of data
  - To evaluate usefulness of a cache
  - To evaluate the impact of EC
  - To find unexpected uses

How do we do this?

EOS report logs



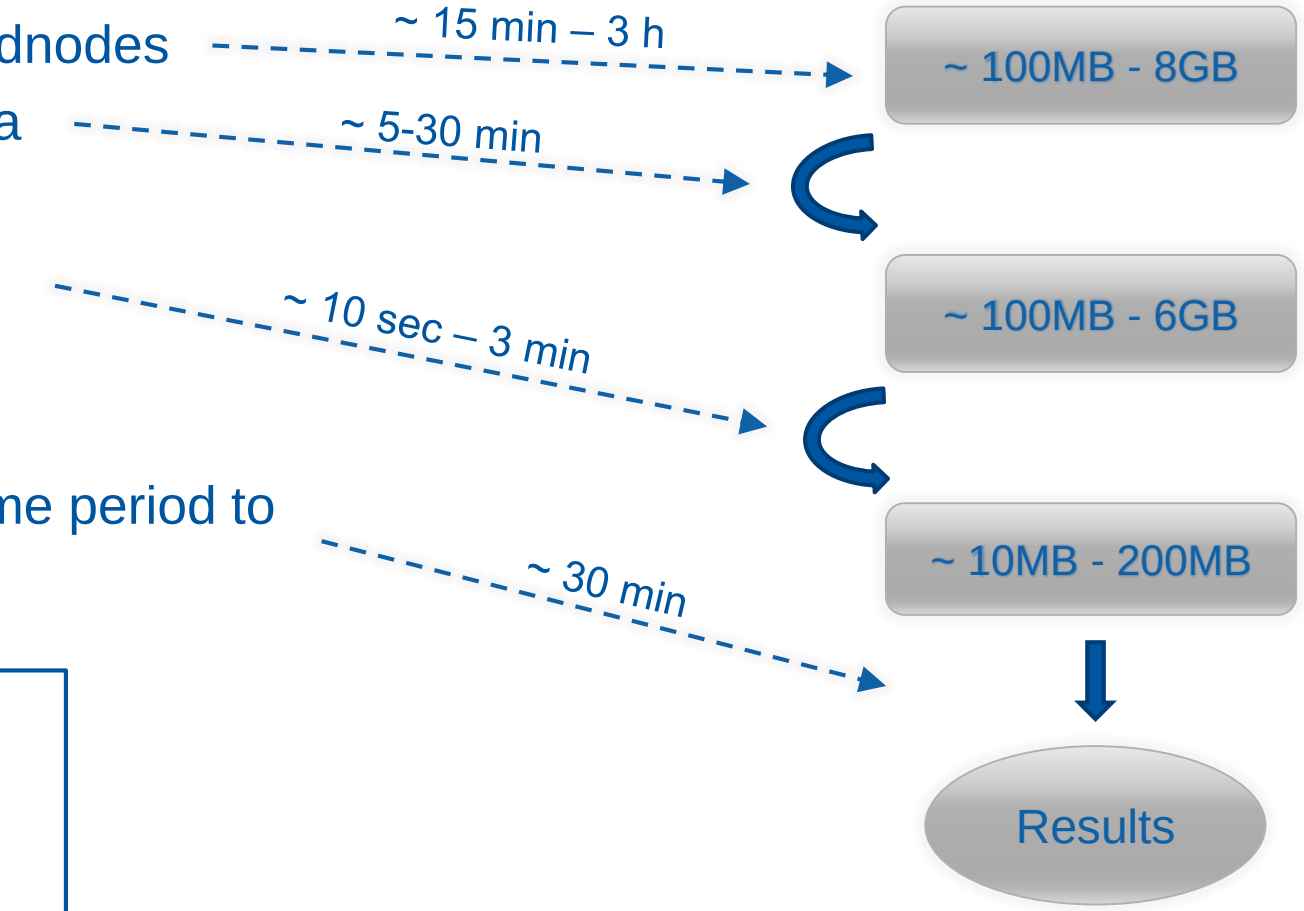
Key Metrics:

fid	osize	csize	rb	wb	ots	cts	ruid	td	…
293031349	0	495399529	0	495399529	2019-01-27 11:29:43	2019-01-27 11:32:45	7947	lhcbprod.45887:92@hlta1013.lbdaq	…
23514666	1744118422	1744118422	1021229	0	2019-01-27 12:06:32	2019-01-27 12:07:30	9801	lblocal.2988:190@lbbuild39	…
…	…	…	…	…	…	…	…	…	…

Analysis Workflow

1 day of logs:

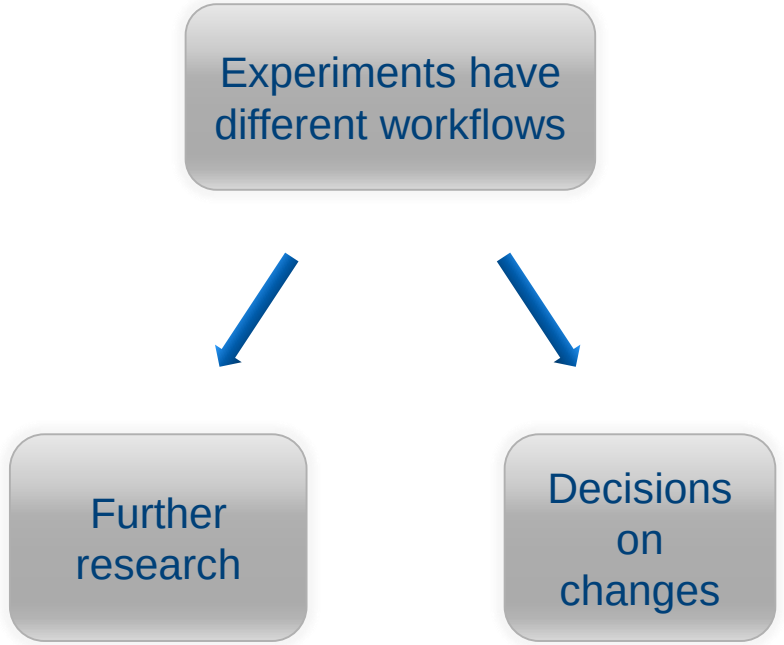
1. Collect raw data from the headnodes
2. Parse and clean the input data
3. Filter and group data
   - Drop system activities
   - Select columns of interest
   - Group by sessions and fids
4. Aggregate data over a long time period to get robust results (3-6 months)



Usually data has defects and/or semantical changes, which we need to compensate by repairing/reformulating across long periods.

Read/Write Workload

Metric	LHCb	CMS	ATLAS
Instance Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Total Accessed	49.2 PiB	67.5 PiB	155 PiB
Total Accessed (% of Total Volume)	328 %	188 %	310 %
Writes	20.0 PiB	28.0 PiB	26.1 PiB
Writes (% of Total Volume)	134 %	77.7 %	52.3 %
Reads	28.4 PiB	39.4 PiB	129 PiB
Reads (% of Total Volume)	190 %	110 %	257 %
Repeated Reads	8.7 PiB	30.7 PiB	108 PiB
Repeated Reads (% of Read Workload)	30.7 %	77.8 %	83.7 %



Read/Write Workload

Metric	LHCb	CMS	ATLAS
Instance Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Total Accessed	49.2 PiB	67.5 PiB	155 PiB
Total Accessed (% of Total Volume)	328 %	188 %	310 %
Writes	20.0 PiB	28.0 PiB	26.1 PiB
Writes (% of Total Volume)	134 %	77.7 %	52.3 %
Reads	28.4 PiB	39.4 PiB	129 PiB
Reads (% of Total Volume)	190 %	110 %	257 %
Repeated Reads	8.7 PiB	30.7 PiB	108 PiB
Repeated Reads (% of Read Workload)	30.7 %	77.8 %	83.7 %

Different in absolute numbers.

Read/Write Workload

Metric	LHCb	CMS	ATLAS
Instance Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Total Accessed	49.2 PiB	67.5 PiB	155 PiB
Total Accessed (% of Total Volume)	328 %	188 %	310 %
Writes	20.0 PiB	28.0 PiB	26.1 PiB
Writes (% of Total Volume)	134 %	77.7 %	52.3 %
Reads	28.4 PiB	39.4 PiB	129 PiB
Reads (% of Total Volume)	190 %	110 %	257 %
Repeated Reads	8.7 PiB	30.7 PiB	108 PiB
Repeated Reads (% of Read Workload)	30.7 %	77.8 %	83.7 %

Different in relative numbers:

- LHCb and ATLAS have more accesses than CMS.

Read/Write Workload

Metric	LHCb	CMS	ATLAS
Instance Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Total Accessed	49.2 PiB	67.5 PiB	155 PiB
Total Accessed (% of Total Volume)	328 %	188 %	310 %
Writes	20.0 PiB	28.0 PiB	26.1 PiB
Writes (% of Total Volume)	134 %	77.7 %	52.3 %
Reads	28.4 PiB	39.4 PiB	129 PiB
Reads (% of Total Volume)	190 %	110 %	257 %
Repeated Reads	8.7 PiB	30.7 PiB	108 PiB
Repeated Reads (% of Read Workload)	30.7 %	77.8 %	83.7 %

Writes:

- LHCb produces more data than other experiments
- LHCb produces more volume than its instance's size.

Read/Write Workload

Metric	LHCb	CMS	ATLAS
Instance Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Total Accessed	49.2 PiB	67.5 PiB	155 PiB
Total Accessed (% of Total Volume)	328 %	188 %	310 %
Writes	20.0 PiB	28.0 PiB	26.1 PiB
Writes (% of Total Volume)	134 %	77.7 %	52.3 %
Reads	28.4 PiB	39.4 PiB	129 PiB
Reads (% of Total Volume)	190 %	110 %	257 %
Repeated Reads	8.7 PiB	30.7 PiB	108 PiB
Repeated Reads (% of Read Workload)	30.7 %	77.8 %	83.7 %

Reads:

- ATLAS reads more data than LHCb and CMS.

Read/Write Workload

Metric	LHCb	CMS	ATLAS
Instance Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Total Accessed	49.2 PiB	67.5 PiB	155 PiB
Total Accessed (% of Total Volume)	328 %	188 %	310 %
Writes	20.0 PiB	28.0 PiB	26.1 PiB
Writes (% of Total Volume)	134 %	77.7 %	52.3 %
Reads	28.4 PiB	39.4 PiB	129 PiB
Reads (% of Total Volume)	190 %	110 %	257 %
Repeated Reads	8.7 PiB	30.7 PiB	108 PiB
Repeated Reads (% of Read Workload)	30.7 %	77.8 %	83.7 %

Repeated Reads:

- Potential for caching
- ~80% for CMS and ATLAS
- Only 30% for LHCb – would profit less from caching.

Operations Classification

$(osize == 0 \ \boldsymbol{and} \ csize > 0 \ \boldsymbol{and} \ wb > 0 \ \boldsymbol{and} \ rb == 0)$ → Create

$(osize > 0 \ \boldsymbol{and} \ csize == osize \ \boldsymbol{and} \ wb == 0 \ \boldsymbol{and} \ rb > 0)$ → Read

$(!create \ \boldsymbol{and} \ !read)$ → Other

Other → Update, Empty, Abnormal

Metric	LHCb	CMS	ATLAS
Other Operations (% of related files)	0.06 %	0.26 %	0.61 %
Other Operations (% of Total Volume)	0.89 %	0.05 %	0.14 %

- Not much influence from abnormal operations.
- *Assumption*: Data is immutable.

Created/Read Volume

Unique data files

Metric	LHCb	CMS	ATLAS
Total Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Created Volume	20.3 PiB	28.0 PiB	26.1 PiB
Created Volume (% of Total Volume)	136 %	77.7 %	52.2 %
Read Volume	22.7 PiB	22.0 PiB	26.6 PiB
Read Volume (% of Total Volume)	151 %	61.1 %	53.2 %
Repeated Read Volume (% of Read Volume)	20.7 %	54.7 %	55.1 %
Average Fraction of File Read	88.8 %	55.1 %	81.6 %

LHCb:

- Produce and read a lot of data
- Well-organized workflow
- Create → Read → Delete
- Don’t have enough space.

Created/Read Volume

Unique data files

Metric	LHCb	CMS	ATLAS
Total Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Created Volume	20.3 PiB	28.0 PiB	26.1 PiB
Created Volume (% of Total Volume)	136 %	77.7 %	52.2 %
Read Volume	22.7 PiB	22.0 PiB	26.6 PiB
Read Volume (% of Total Volume)	151 %	61.1 %	53.2 %
Repeated Read Volume (% of Read Volume)	20.7 %	54.7 %	55.1 %
Average Fraction of File Read	88.8 %	55.1 %	81.6 %

CMS & ATLAS:

- Produce less data
- Use only part of the space
- Higher chance of reuses.

Created/Read Volume

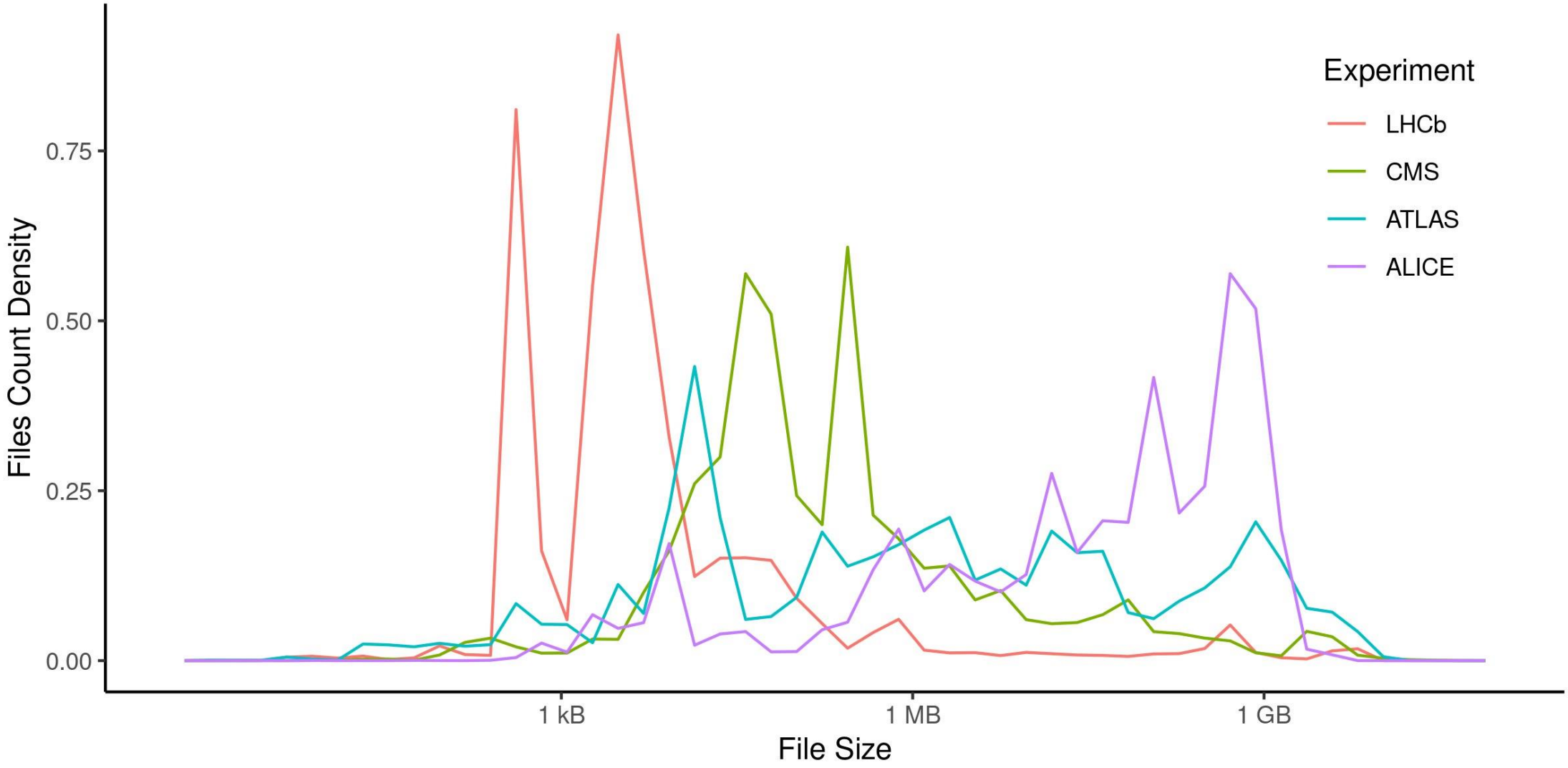
Unique data files

Metric	LHCb	CMS	ATLAS
Total Volume (EOS Control Tower)	15.0 PiB	36.0 PiB	50.0 PiB
Created Volume	20.3 PiB	28.0 PiB	26.1 PiB
Created Volume (% of Total Volume)	136 %	77.7 %	52.2 %
Read Volume	22.7 PiB	22.0 PiB	26.6 PiB
Read Volume (% of Total Volume)	151 %	61.1 %	53.2 %
Repeated Read Volume (% of Read Volume)	20.7 %	54.7 %	55.1 %
Average Fraction of File Read	88.8 %	55.1 %	81.6 %

Fraction of Files Read:

- 80-90% for ATLAS and LHCb
- Only 55% for CMS
- Bad for caching.

Normalized probability distribution of file sizes:



6 Months Period: 01/01/2019 – 30/06/2019

New/Old Files Relevance

Metric	LHCb	CMS	ATLAS
Created and Read	19.2 PiB	16.2 PiB	21.0 PiB
Created and Not Read	1.1 PiB	11.7 PiB	5.1 PiB
Created and Read (% of Created Volume)	94.4 %	58.0 %	80.4 %
Created and Not Read (% of Created Volume)	5.62 %	42.0 %	19.6 %
Old and Read	3.5 PiB	5.8 PiB	5.6 PiB
Old and Not Read	11.5 PiB	31.2 PiB	40.4 PiB
Old Before and Read (% of Old Volume)	23.5 %	16.0 %	11.2 %
Old Before and Not Read (% of Old Volume)	76.6 %	84.0 %	88.8 %

Usage of new data:

- LHCb uses almost all the data they produce
- For ATLAS and CMS the numbers are lower.

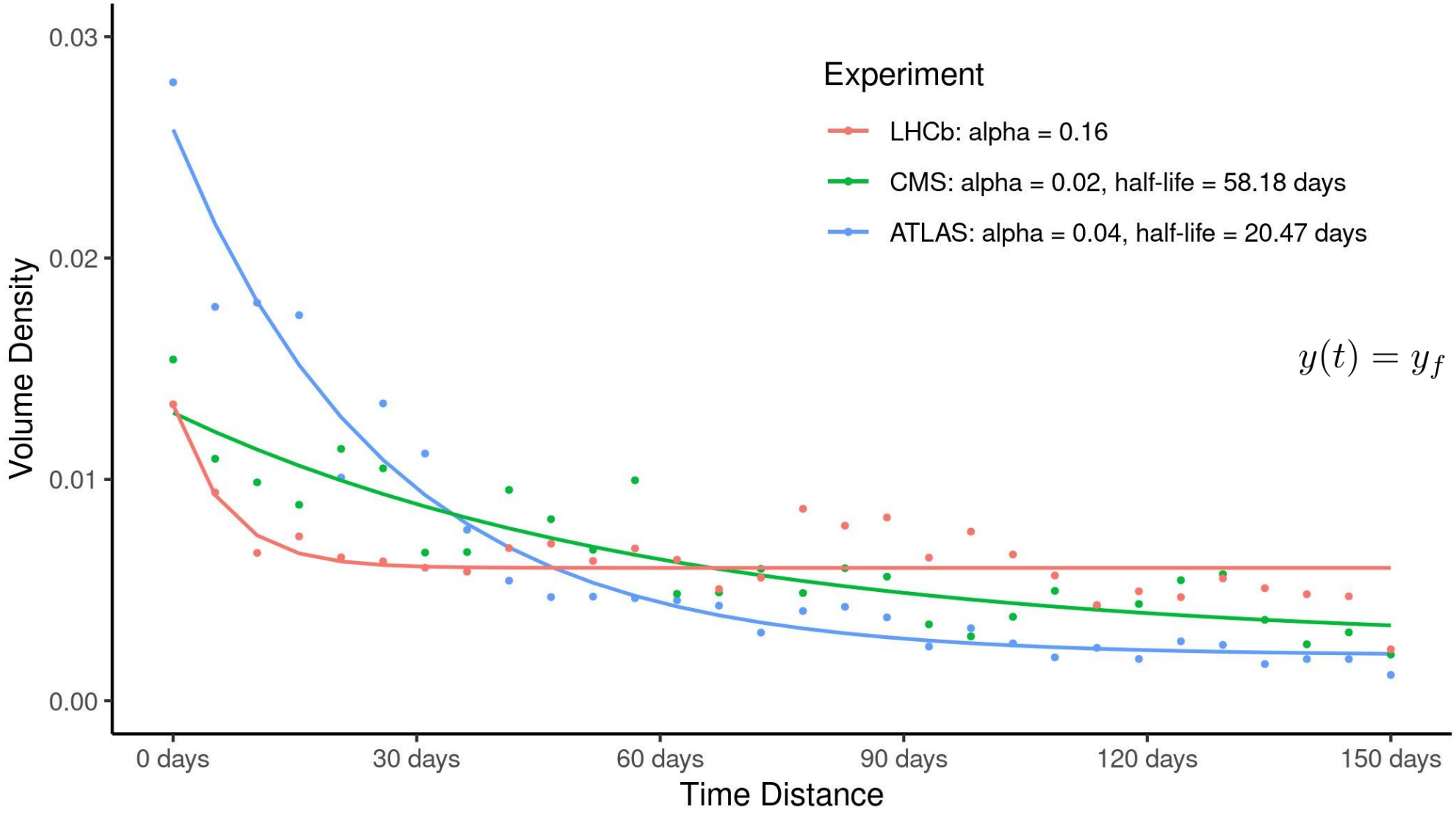
New/Old Files Relevance

Metric	LHCb	CMS	ATLAS
Created and Read	19.2 PiB	16.2 PiB	21.0 PiB
Created and Not Read	1.1 PiB	11.7 PiB	5.1 PiB
Created and Read (% of Created Volume)	94.4 %	58.0 %	80.4 %
Created and Not Read (% of Created Volume)	5.62 %	42.0 %	19.6 %
Old and Read	3.5 PiB	5.8 PiB	5.6 PiB
Old and Not Read	11.5 PiB	31.2 PiB	40.4 PiB
Old Before and Read (% of Old Volume)	23.5 %	16.0 %	11.2 %
Old Before and Not Read (% of Old Volume)	76.6 %	84.0 %	88.8 %

Usage of old data:

- Rarely used in future.

Time distribution of file accesses:

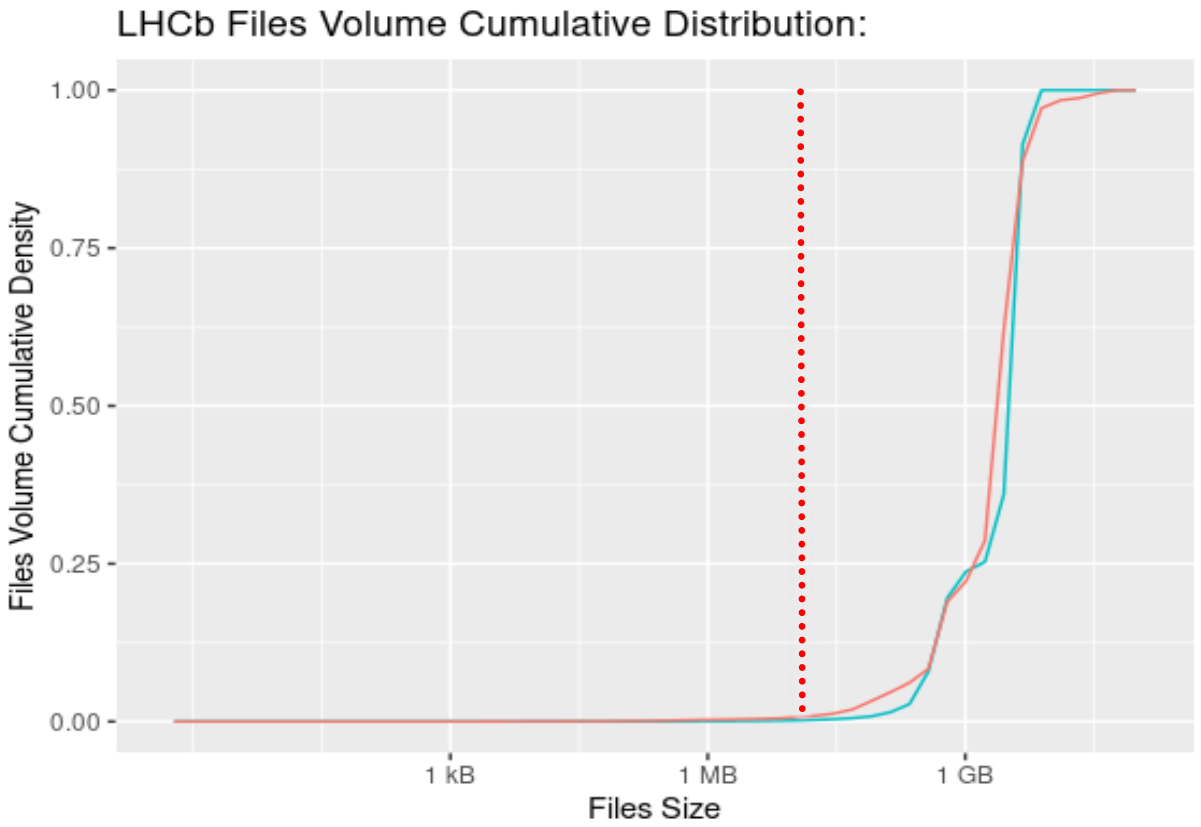


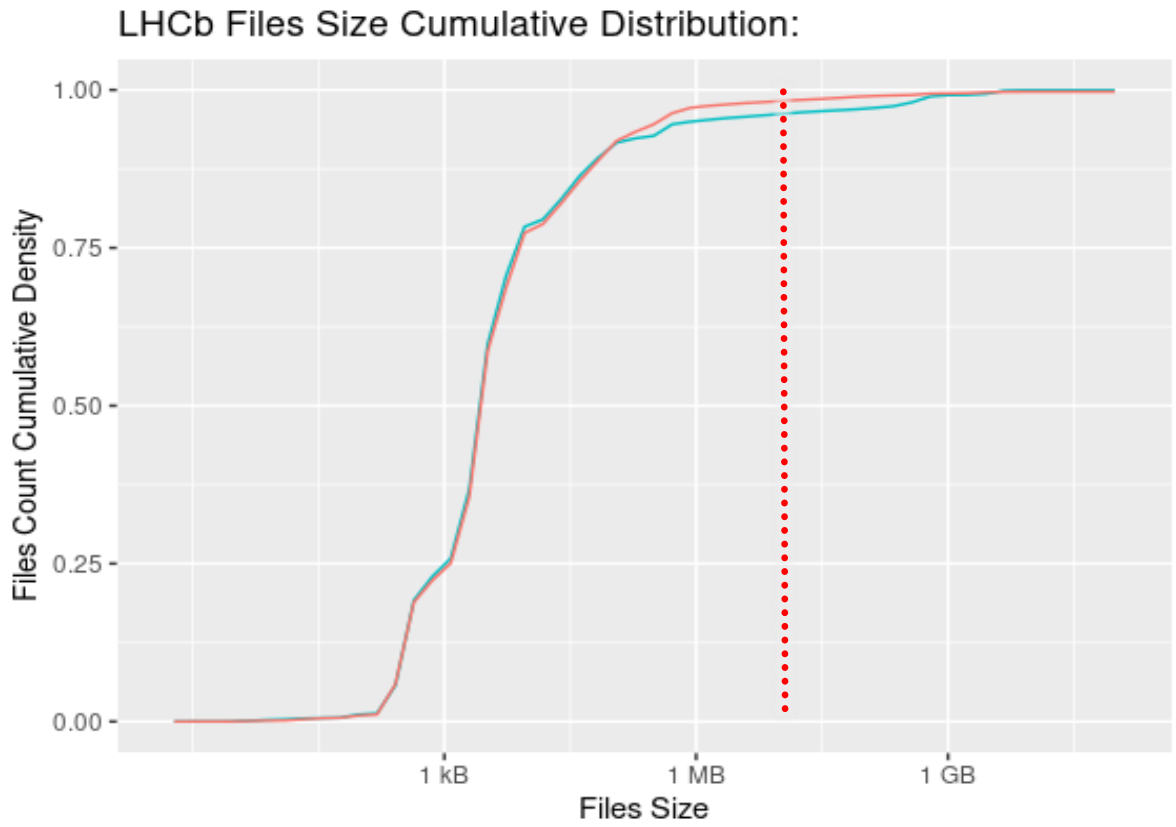
Model:

$$y(t) = y_f + (y_0 - y_f)e^{-\alpha t}$$

Half-life:

$$y(t_{1/2}) = \frac{1}{2}y_0$$





Colour

Current Namespace (all files including old, but excluding deleted)

EOS Logs (only files that were created or accessed)

99% of volume covered by converting <3% of files - 2M files

Input to the future plans

Balance between

Cheaper Storage (Erasure Coding)

More expensive storage (multi-replica, flash-based)

Would we benefit from having several QoS levels?

How do we decide which level to use for each piece of data?

→ Popularity Prediction

Thank you

Any questions?

Olga Chuchuk

olga.chuchuk@cern.ch

CERN, IT-SC-RD, Geneva, Switzerland

Taras Shevchenko KNU, Kyiv, Ukraine

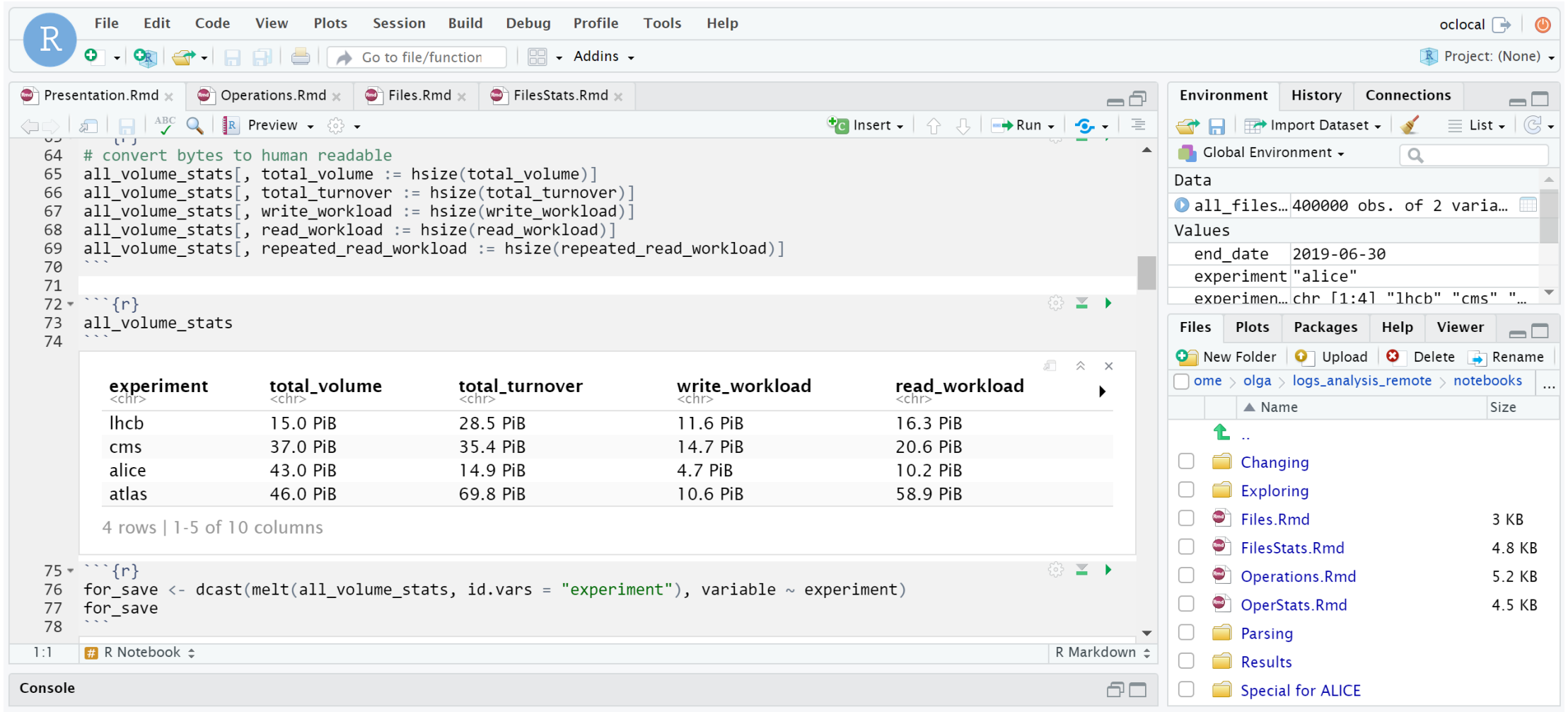
Dirk Duellmann

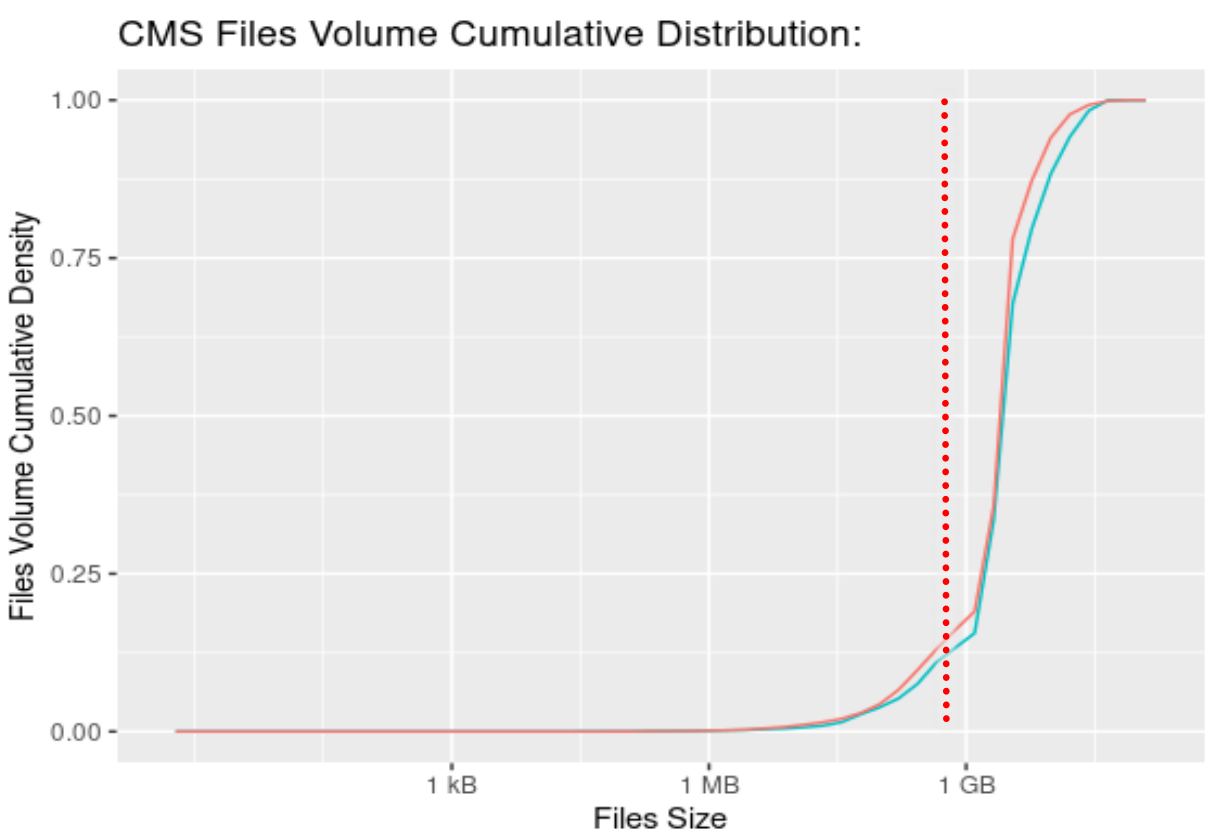
dirk.duellmann@cern.ch

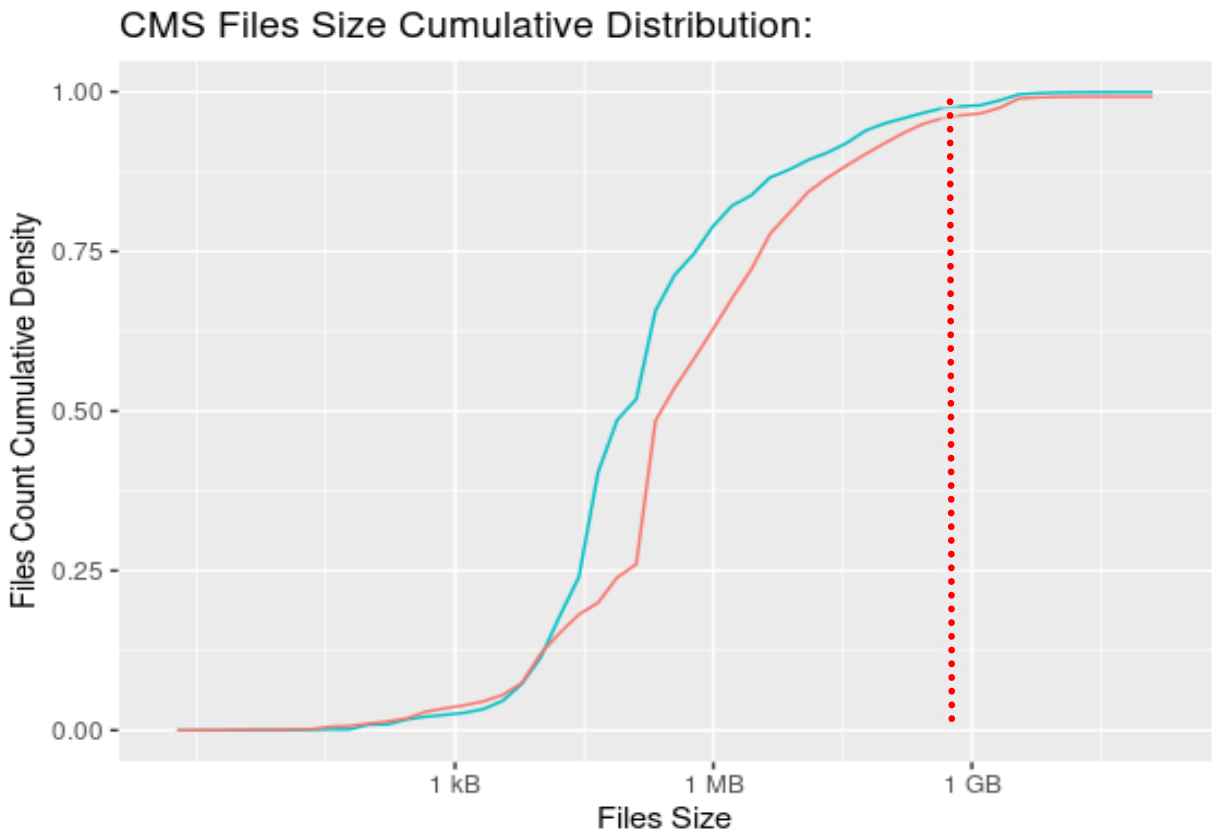
CERN, IT-SC, Geneva, Switzerland

Backup slides

Analysis Environment





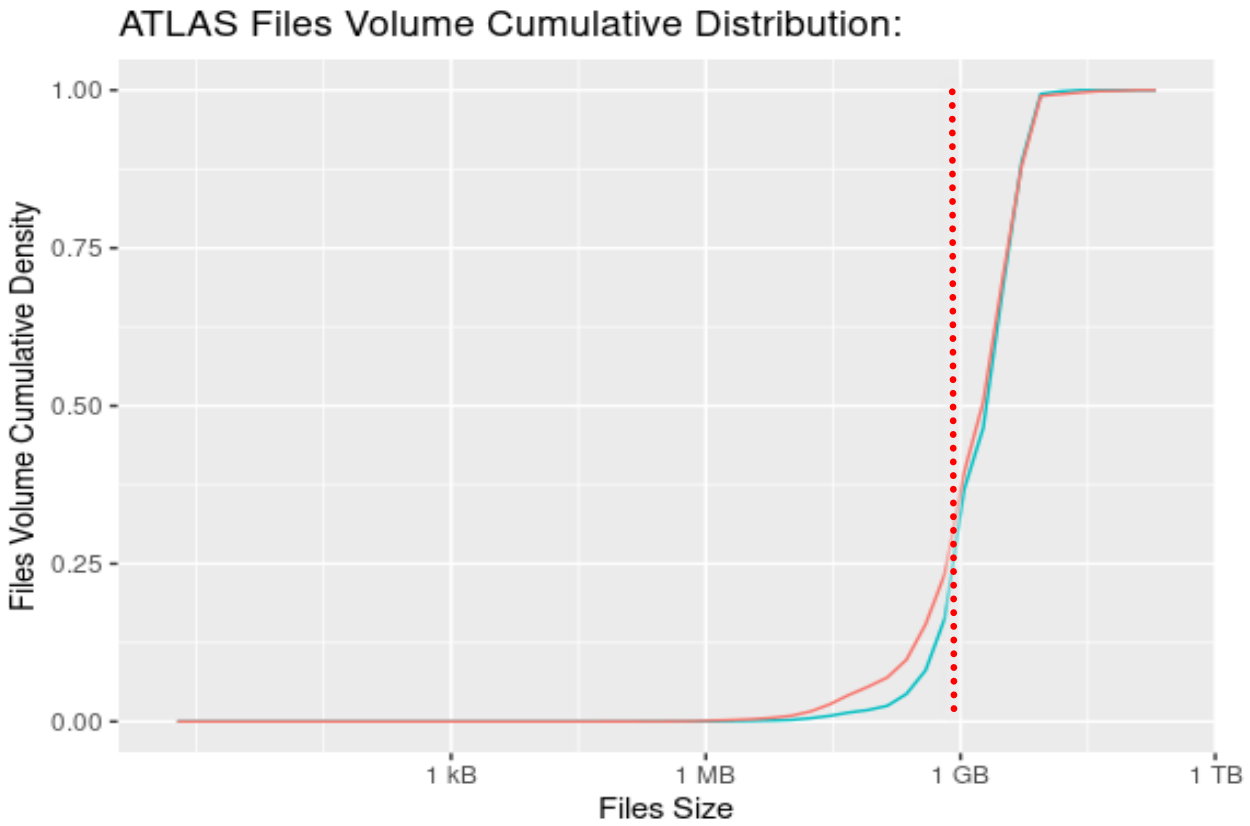


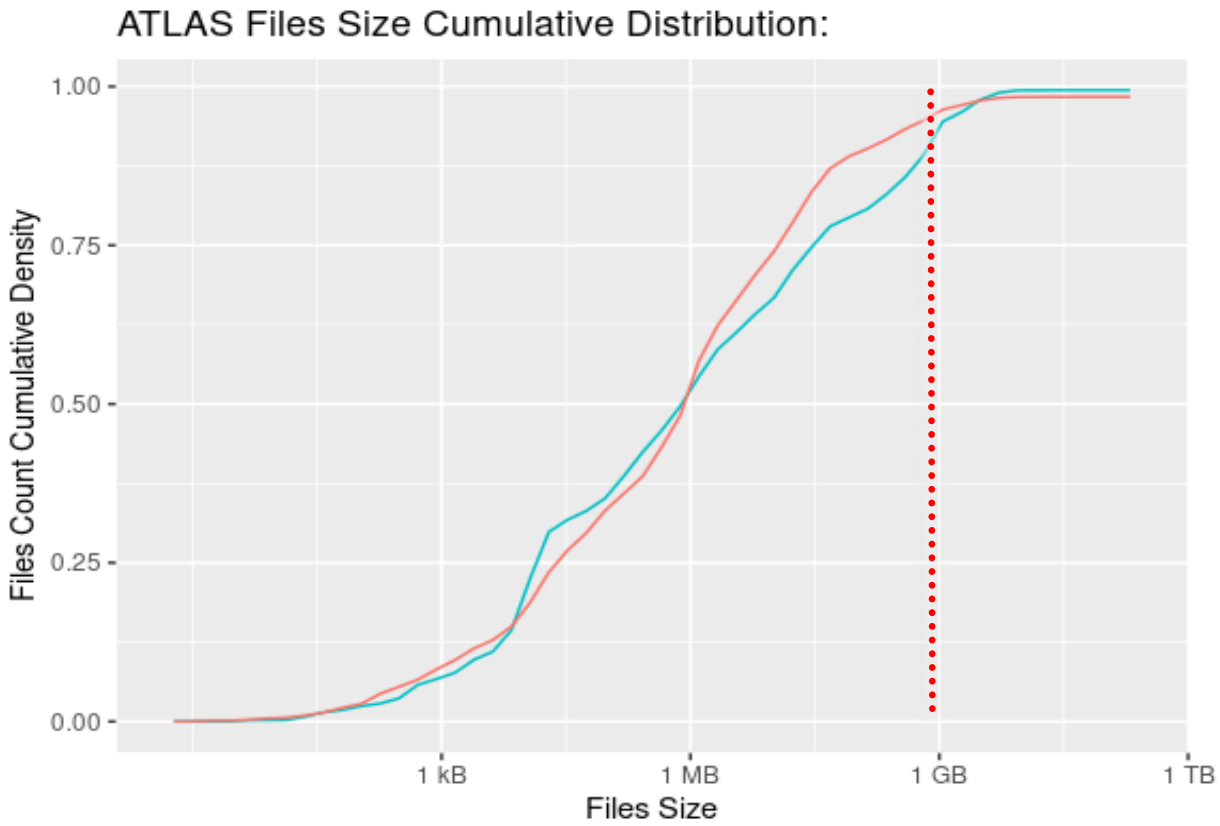
Colour

Current Namespace (all files including old, but excluding deleted)

EOS Logs (only files that were created or accessed)

87% of volume covered by converting <3% of files - 17M files



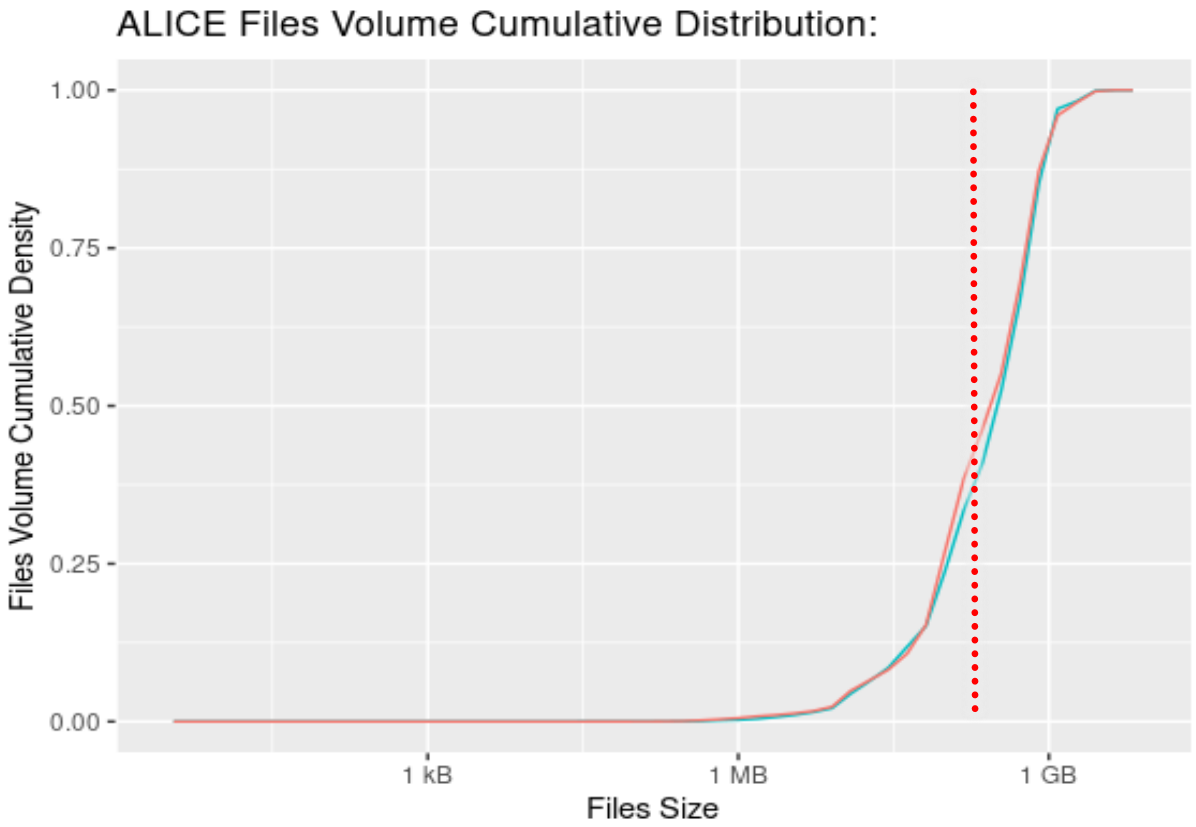


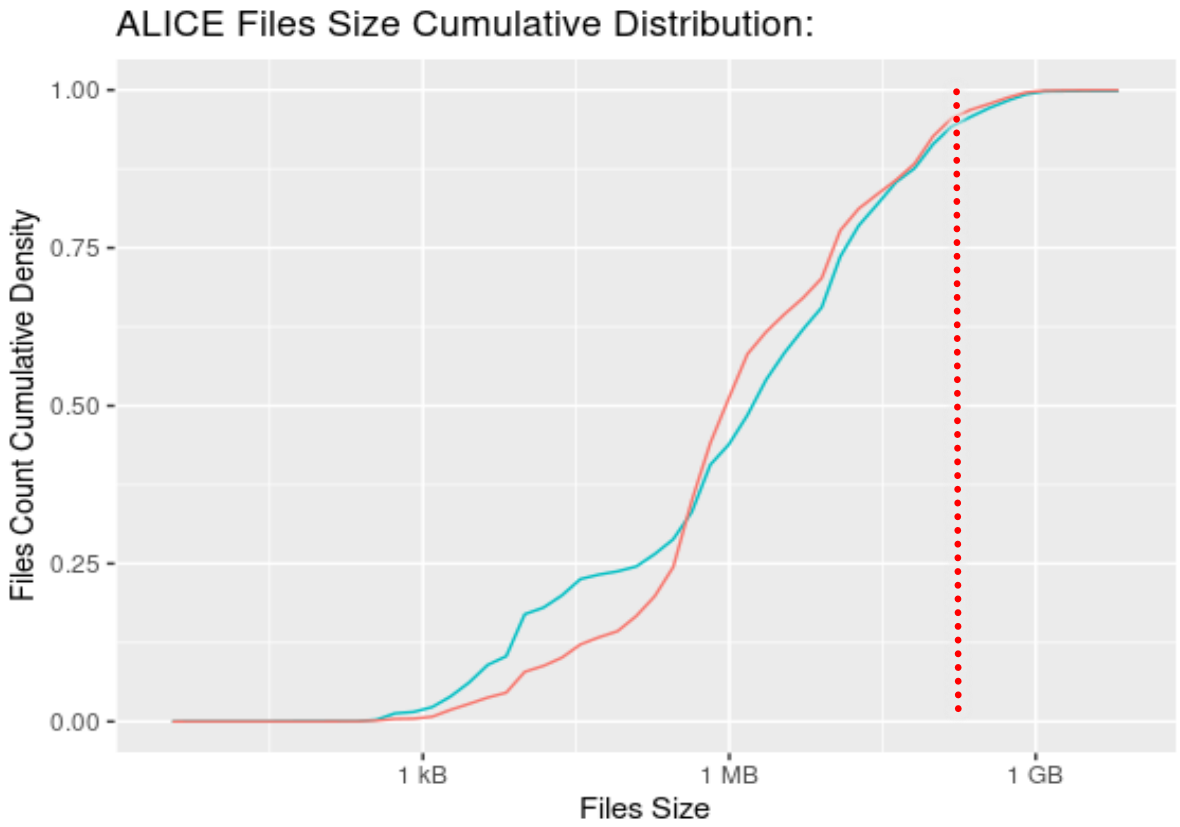
Colour

Current Namespace (all files including old, but excluding deleted)

EOS Logs (only files that were created or accessed)

70% of volume covered by converting <3% of files - 6M files





Colour

Current Namespace (all files including old, but excluding deleted)

EOS Logs (only files that were created or accessed)

65% of volume covered by converting <3% of files - 5M files