Increasing Tape Efficiency

HEPiX Fall 2008 Taipei

Nicola Bessone, German Cancio, **Steven Murray**, Giulia Taurelli

Contents

- Tape efficiency project
- Problem areas
- What has been done
- What is under development
- Roadmap
- Summary

Tape Efficiency Project

CERN IT Department

- All functionality dealing directly with storage on and management of tapes
  - Volume database
  - Migrations/recalls
  - Tape drive scheduling
  - Low-level tape positioning and read/write
- Team is from IT/DM
- Contributions from IT/FIO

Problem Areas

Work done

- Write more data per tape mount

Current work

- Use a more efficient tape format
  - The current tape format does not deal efficiently with small files

More ideas

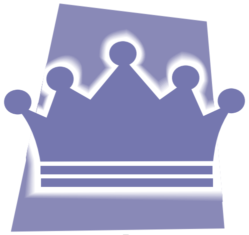
- Improve read efficiency
  - Require modifications from disk to tape

What has been done

Read/write More Per Mount

- Recall/migration policies

  - Freight train approach
  - Hold back requests based on the amount of data and elapsed time



- Production managers rule
  - Production managers plan relatively **large workloads** for CASTOR
  - **Access control lists** give production managers a relatively larger percentage of resources
  - **User and group based priorities** encourage users to work with their production managers

Repack

- Repacks the data from one set of tapes onto another set of tapes
- Repack is used for media migration
- Repack is used to defragment tapes

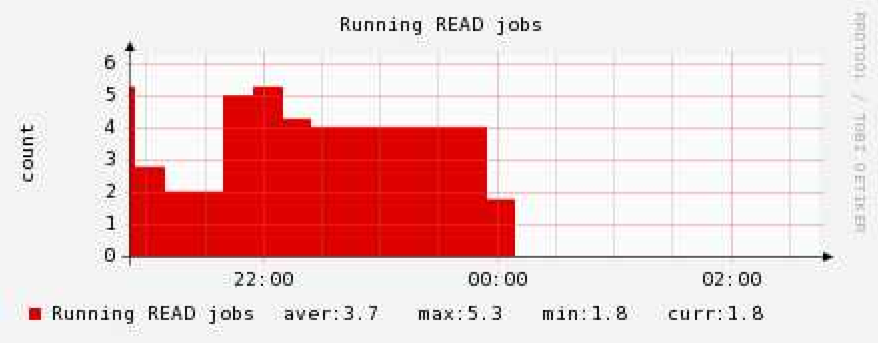
Efficiency and Repack

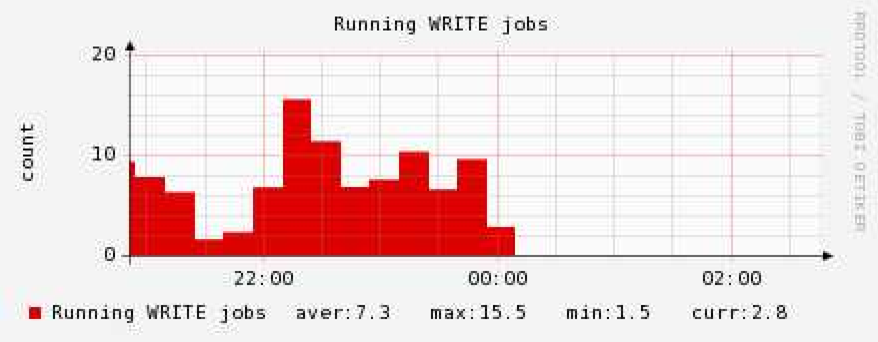
- Reading the current ANSI AUL format is approximately twice as fast as writing
- Repack uses Castor as a cache
- Repack uses the cache to support asymmetric read/write drive allocation
- Repack is equivalent to one LHC experiment and as such is a good test run for Castor

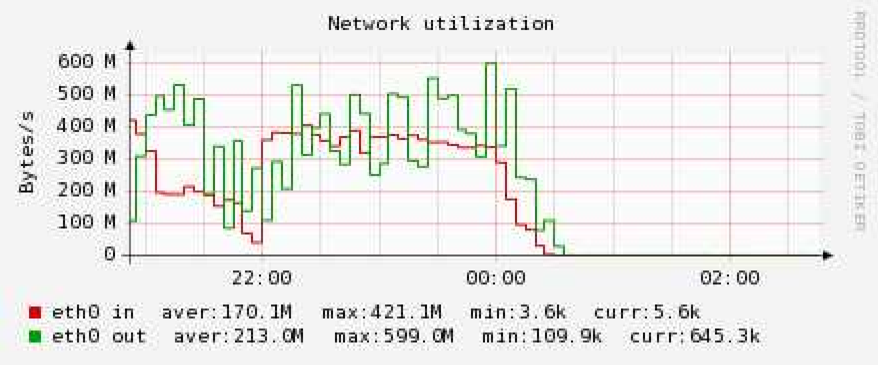
DM

Repack Measurements

CERN IT Department







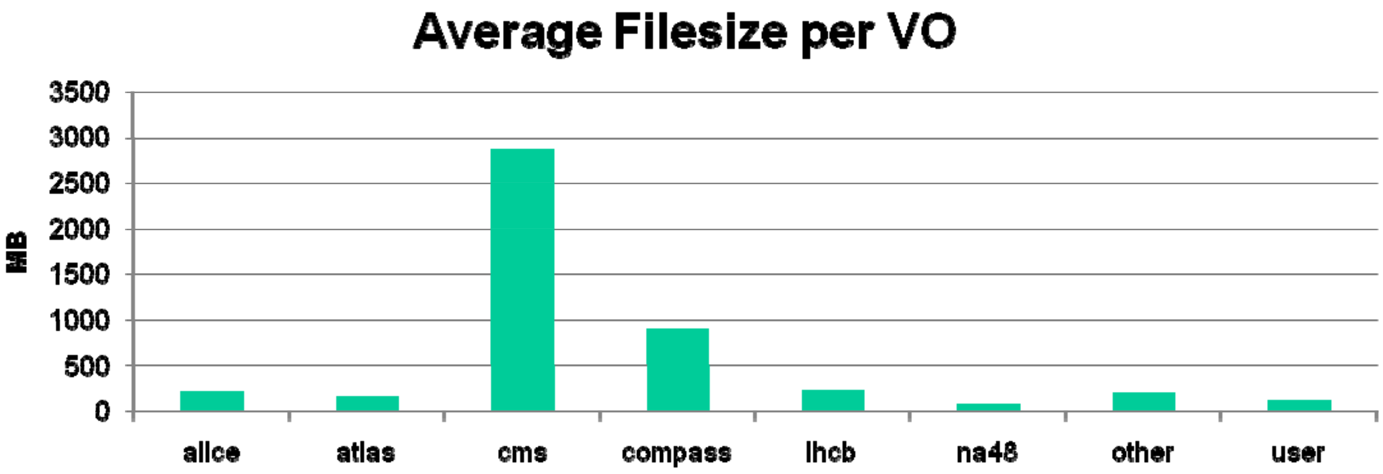
- 4 drives reading
- 7 drives writing
- 400MBytes/s

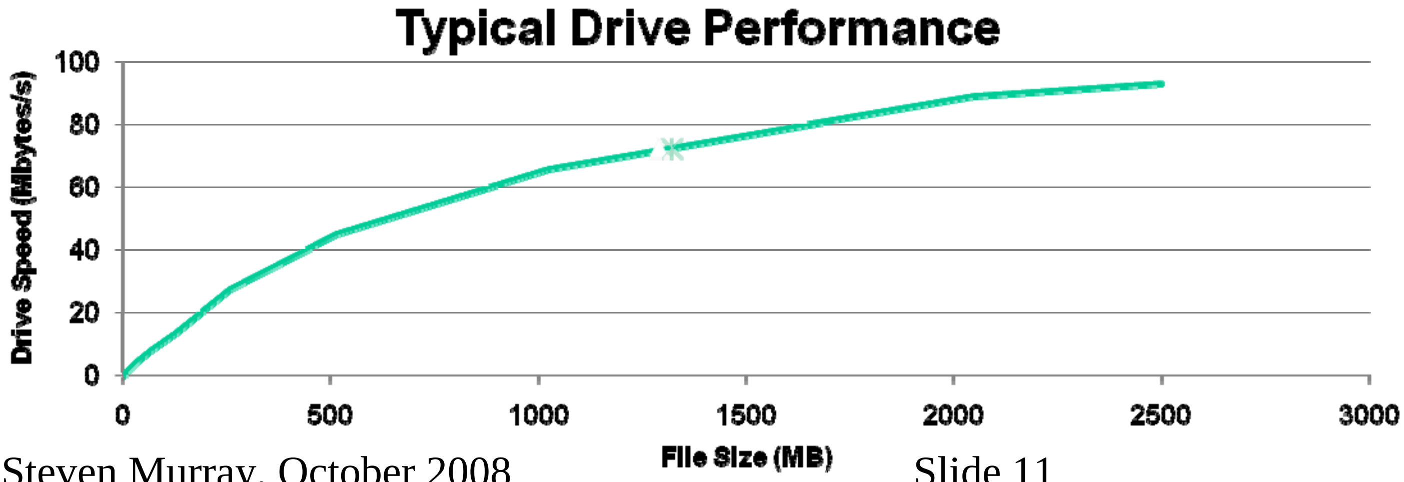
What is under development

Writing Small Files is Slow

CERN IT Department

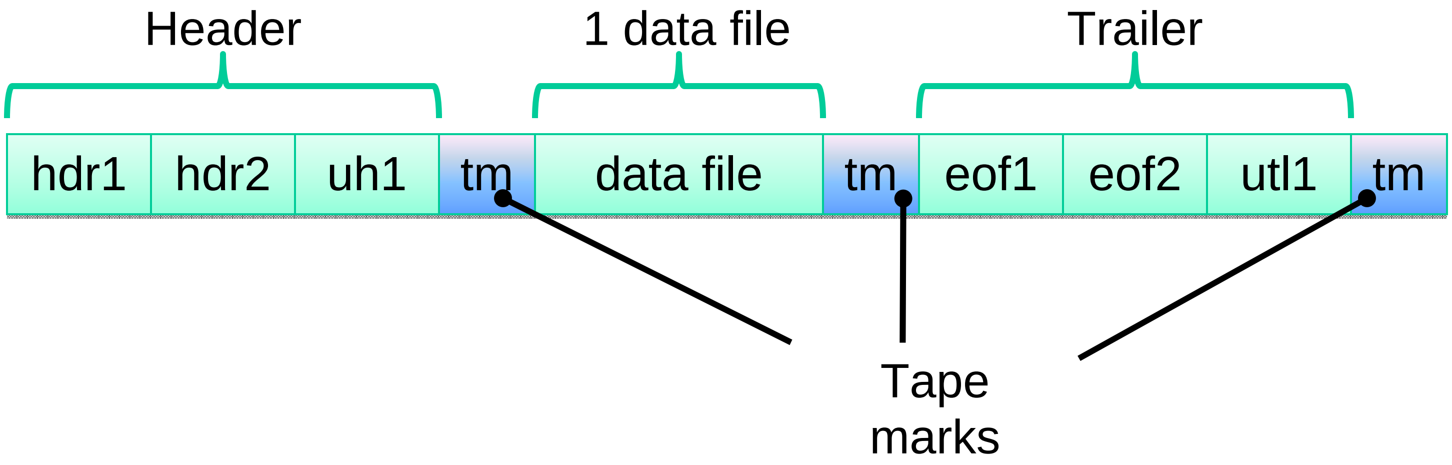
- Users were encouraged to store large files in Castor
- Unfortunately Castor contains many small files





Why Small Files are Slow

CERN IT Department



Header

1 data file

Trailer

hdr1	hdr2	uh1	tm	data file	tm	eof1	eof2	utl1	tm

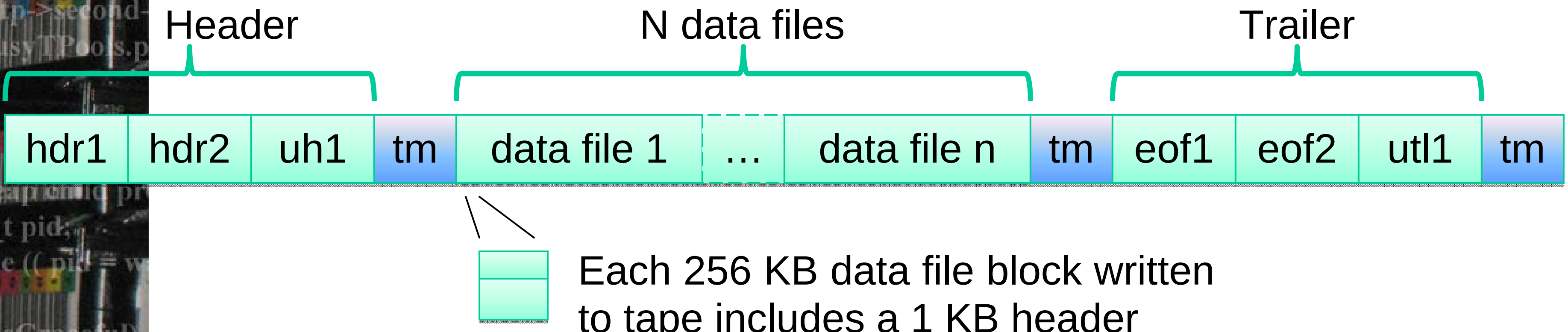
Tape marks

- ANSI AUL format
- 3 tape marks per file
- 2 to 3 second per tape mark
- 9 seconds per data file independent of its size

DM

CERN IT Department

New Tape Format



Each 256 KB data file block written to tape includes a 1 KB header

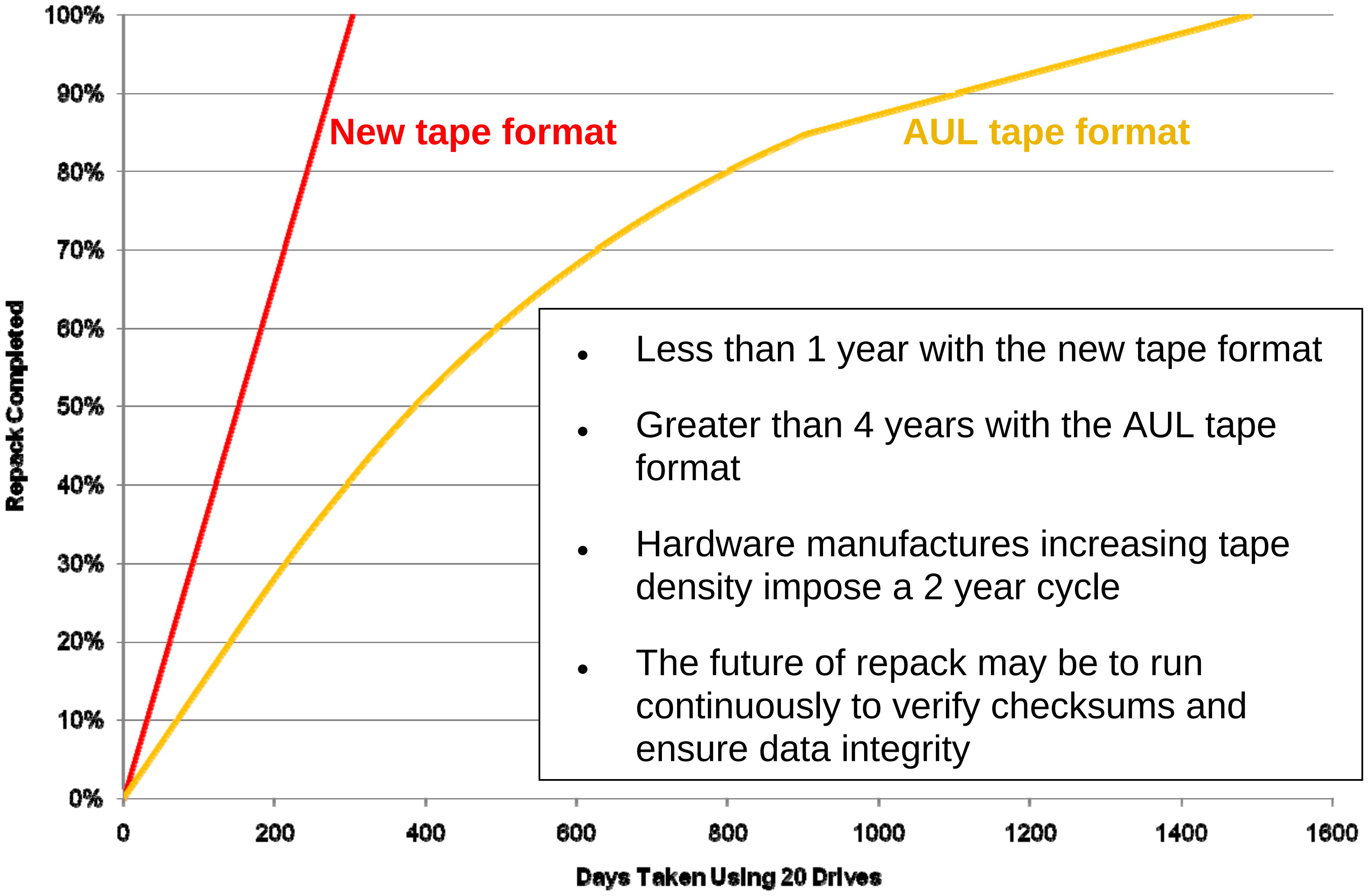
- Multi-file block format within the ANSI AUL format
- Header per block for “self description”
- 3 tape marks per n files
- n will take into account:
  - A configurable maximum number of files
  - A configurable maximum size
  - A configurable maximum amount of time to wait

Block Header Format

#	Meta-data name	Explanation	Examples	Bytes for Data
1	VERSION_NUMBER	The version of the block format	09.13	5
2	HEADER_SIZE	Header size in bytes	01024	5
3	CHECKSUM_ALGORITHM	Name of the checksum algorithm	Adler-32	10
4	HEADER_CHECKSUM	Adler-32 checksum	4146884724	10
5	TAPE_MARK_COUNT	Sequential number addressing the migration-files on the tape	00000000000000012345	20
6	BLOCK_SIZE	Block size in bytes inclusive of header	0000262144	10
7	BLOCK_COUNT	Block offset from the beginning of the tape. Tape marks and labels are included in the count	00000000000000012345	20
8	BLOCK_TIME_STAMP	Time since the Epoch (00:00:00 UTC, January 1, 1970), measured in seconds	1222332810	10
9	STAGER_VERSION	The version of the stager software	2.1.7.18	15
10	STAGER_HOST	The DNS name of the stager host including the domain	c2cms2stager.cern.ch	30
11	DRIVE_NAME	Will be provided by a local configuration file	0003592028	10
12	DRIVE_SERIAL	Will be provided by a local configuration file	00000000456000001642	20
13	DRIVE_FIRMWARE	Will be provided by a local configuration file	D3I0_C90	10
14	DRIVE_HOST	The DNS name of the host including the domain	tpsrv250.cern.ch	30
15	VOL_DENSITY	The storage capacity of the tape	700.00GB	10
16	VOL_ID	Site specific numbering system (the sticker on a tape)	T02694	20
17	VOL_SERIAL	Volume Serial Number	T02694	20
18	DEVICE_GROUP_NAME	The device group name that linked the tape to the drive	3592B1	10
19	FILE_SIZE	The size of the data file in bytes	00000001099511627776	20
20	FILE_CHECKSUM	Adler-32 checksum	1926860616	10
21	FILE_NS_HOST	The DNS name of the host including the domain	castorns.cern.ch	30
22	FILE_NS_ID	The name server ID of the data file	226994274	20
23	FILE_PROGESSIVE_CHECKSUM	Adler-32. Progressive checksum of all the blocks written to tape so far for the current data file	1234567890	10
24	FILE_BLOCK_COUNT	Block offset from the beginning of the data file	00000000000000012345	20
			Header size before file_name :	375
25	FILE_NAME	Last "x" bytes of the filename from the name server. This field acts as a padding to the nearest KiB.		649
			Header size :	1024

VERSION_NUMBER
HEADER_SIZE
CHECKSUM_ALGORITHM
HEADER_CHECKSUM
TAPE_MARK_COUNT
BLOCK_SIZE
BLOCK_COUNT
BLOCK_TIME_STAMP
STAGER_VERSION
STAGER_HOST
DRIVE_NAME
DRIVE_SERIAL
DRIVE_FIRMWARE
DRIVE_HOST
VOL_DENSITY
VOL_ID
VOL_SERIAL
DEVICE_GROUP_NAME
FILE_SIZE
FILE_CHECKSUM
FILE_NS_HOST
FILE_NS_ID
FILE_PROGESSIVE_CHECKSUM
FILE_BLOCK_COUNT
FILE_NAME

Predicted Tape Format Performance

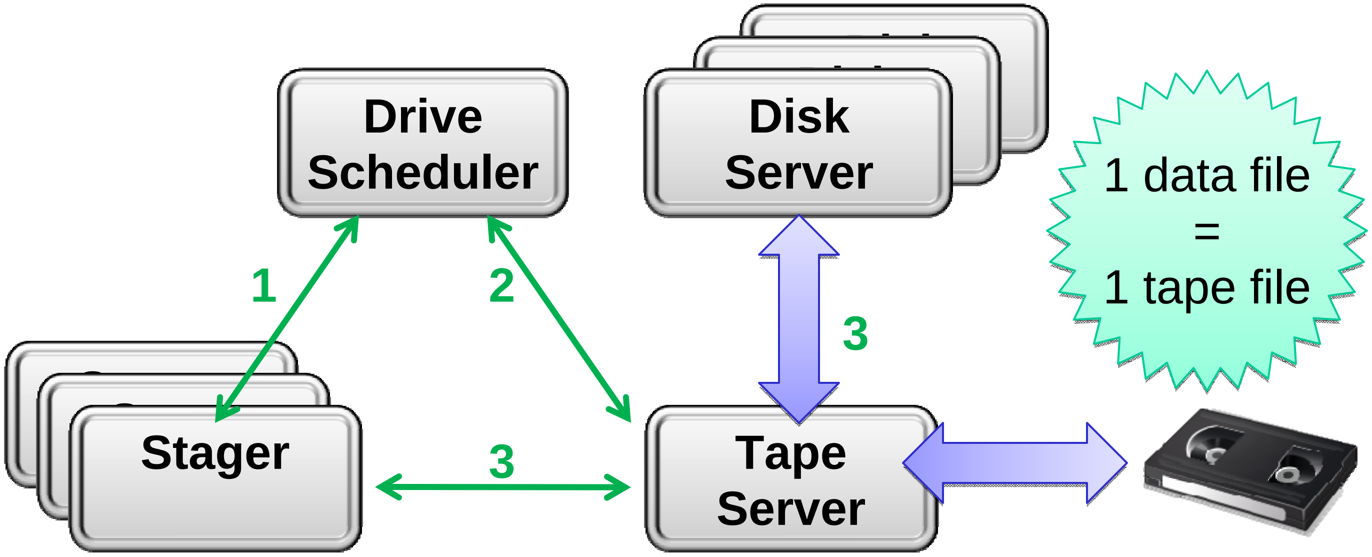


- Less than 1 year with the new tape format
- Greater than 4 years with the AUL tape format
- Hardware manufactures increasing tape density impose a 2 year cycle
- The future of repack may be to run continuously to verify checksums and ensure data integrity

Architecture Needs to Change

- The new tape format is only half of the story
- An aggregator needs to be inserted into the disk ↔ tape data streams
- Anything old that is replaced is an opportunity for code re-use and increased maintainability via the Castor framework

Current Architecture



1. Stager requests a drive
2. Drive is allocated
3. Data is transferred to/from disk/tape based on file list given by stager

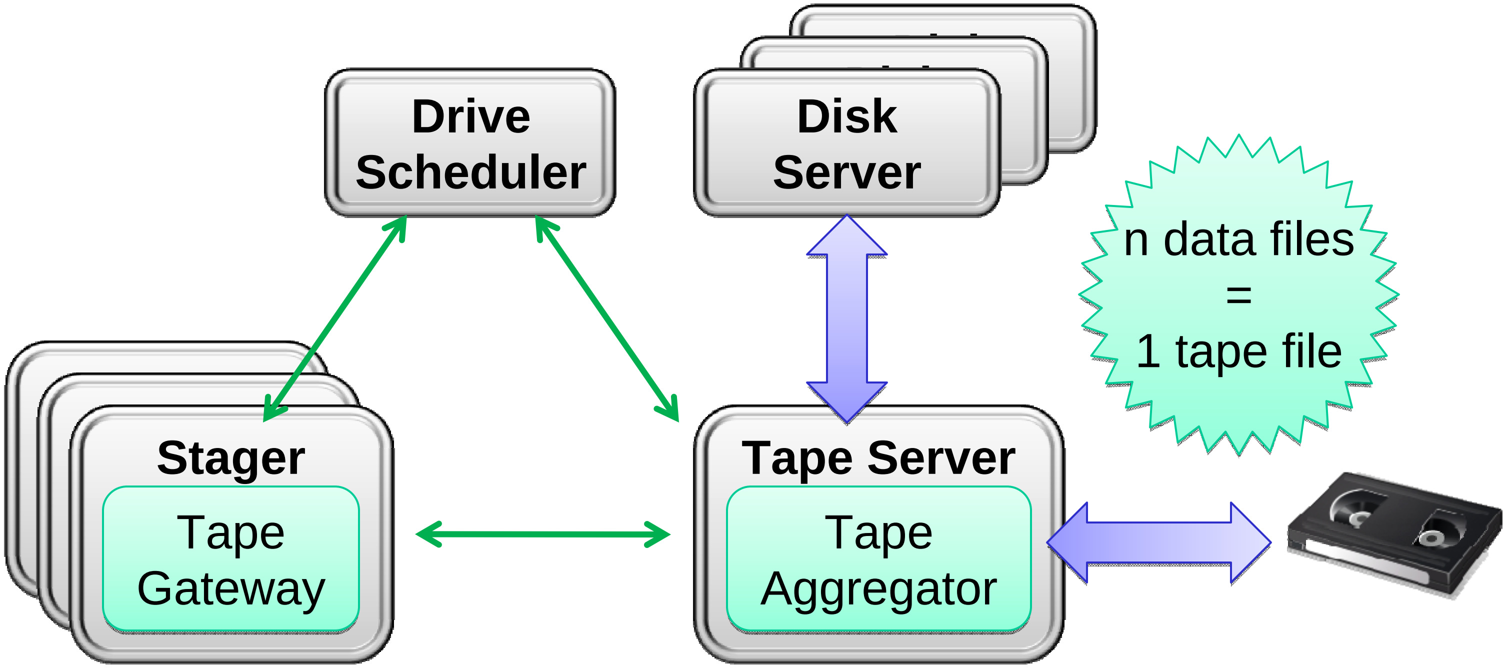
Legend

▭ Host

↔ Control messages

⇔ Data

New Architecture



- The tape gateway will replace RTCPClientD
- The tape gateway will be stateless
- The tape aggregator will wrap RTCPD

Legend

▭ Host

🟩 Server process(es)

↔ Control messages

⇨ Data to be stored

Roadmap

CERN IT Department

Date	Actions
Beginning Q4 2008	Put repack into full production will at least 20 drives. Expecting 700 MB/s. Conclude new tape format architecture.
End Q1 2009	Release first functional prototype of new tape format.
End Q2 2009	Write new tape format with repack only. Read new tape format everywhere.
End Q3 2009	Read and write everywhere
Beginning Q1 2010	Replace RTCPD with tape aggregator

Summary

- We have improved the efficiency of tape by increasing the amount of data we write per mount
- Repack uses Castor as cache to support asymmetric drive read/write allocation
- We are currently developing a new tape format to increase write performance
- The future of repack may be to run continuously to constantly verify data integrity in addition to media migration and tape defragmentation
- We will continue to identify the greatest efficiency improvements that require the least effort