Aggregation to Improve Tape Efficiency

Architecture Meeting

Wednesday 16th July 2008

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Contents

- The problem
- Timelines
- Vision
- Proposed solutions
- Current status
- Tasks to be done
- Tape capabilities

The problem

- Writing small files to tape as individual files is dropping the performance of the tape system to an unacceptable level
- The reason for bad performance is the fact that each individual file written to tape is surround by its own set of tape marks
- It takes approximately 7 seconds to write the tape marks of a file to tape. With a top drive speed of approximately 120 MB/s, files less than 840 MB (7s 120MB/s) will reduce the performance of the tape storage system to less than 50%.

Vision

- The tape system will be a STORAGE POOL that provides AGGREGATION for CASTOR.
  - Well defined interfaces for
    - Efficient file storage and retrieval
    - Aggregation retrieval
  - Optimised management of volumes and tape drives
  - Support for efficient media migration

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Abbreviations and acronyms

- **IL** Internal block-oriented Labels
- **AUL** Ansi User Labels
- **FSEQ** File SEQuence number
- **NL** No Label
- **NS** Name Server
- **RTCPD** Remote Tape CoPy Daemon
- **TAR** Tape ARchive
- **VMGR** Volume ManaGeR

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Jargon

- "More work" protocol
  - The part of the RTcopy daemon protocol that allows addition files to be appended to the work to be done whilst the tape is being used
- RTcopy daemons
  - RTCPD and RTCPClientD

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Proposed solutions

- NL tapes
- Aggregation
  - IL + NL tapes
  - TAR + AUL tapes

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

NL tapes

- What is it?
  - One tape mark per file, no tape labels
- Advantages
  - Already implemented
  - Reduces number of tape marks by 3
- Disadvantages
  - No metadata on tape (label files)
  - Performance gain is not enough

IL + NL tapes

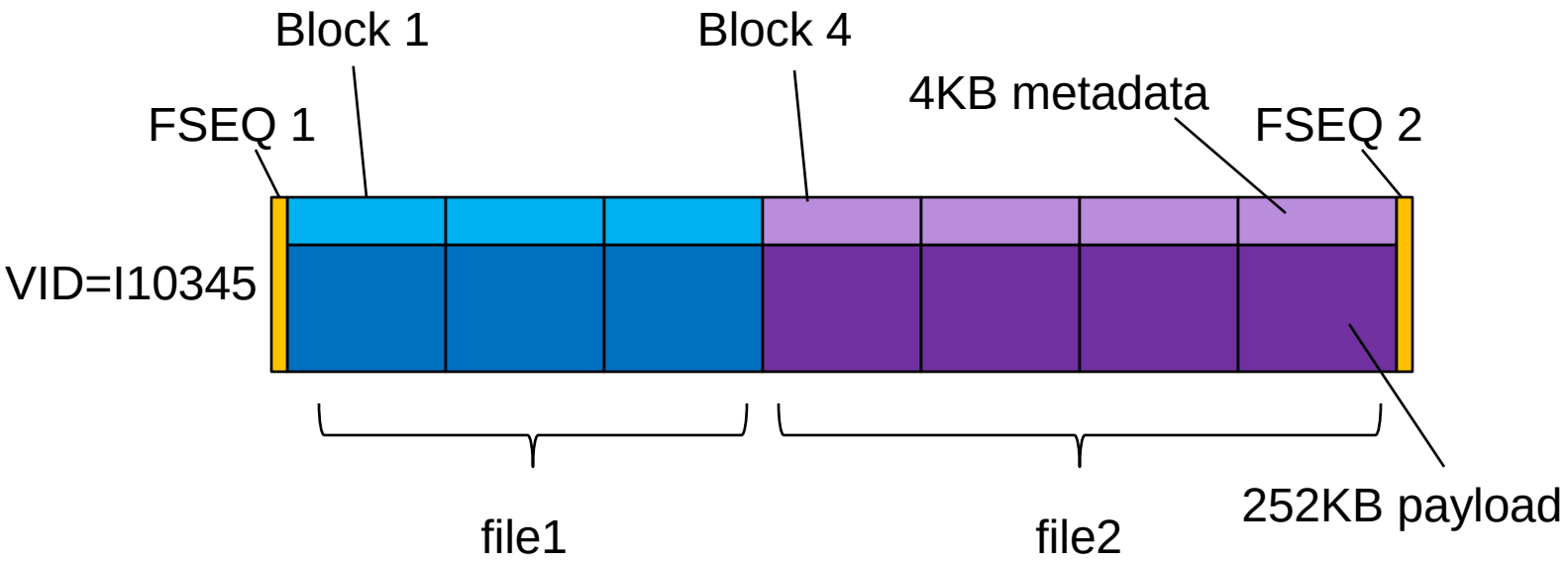
- What is it?
  - In-house block-based tape format enclosed by tape marks
  - Each 256KB block has a 4KB metadata header and 252KB payload
- Advantages
  - Simple design
  - Metadata redundancy
  - We control the metadata specification
  - Configurable number of tape marks per tape
  - Easy to inspect tapes for repair / problem diagnosis
  - Naturally fulfils write position verification requirement
  - Compatible with the rtcopy “more work” protocol
- Disadvantages
  - In-house format

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

IL + NL tapes

NS entries

- **Filename=file 1, VID=I10345, FSEQ=1, block ID=1**
- **Filename=file 2, VID=I10345, FSEQ=2, block ID=4**



TAR + AUL tapes

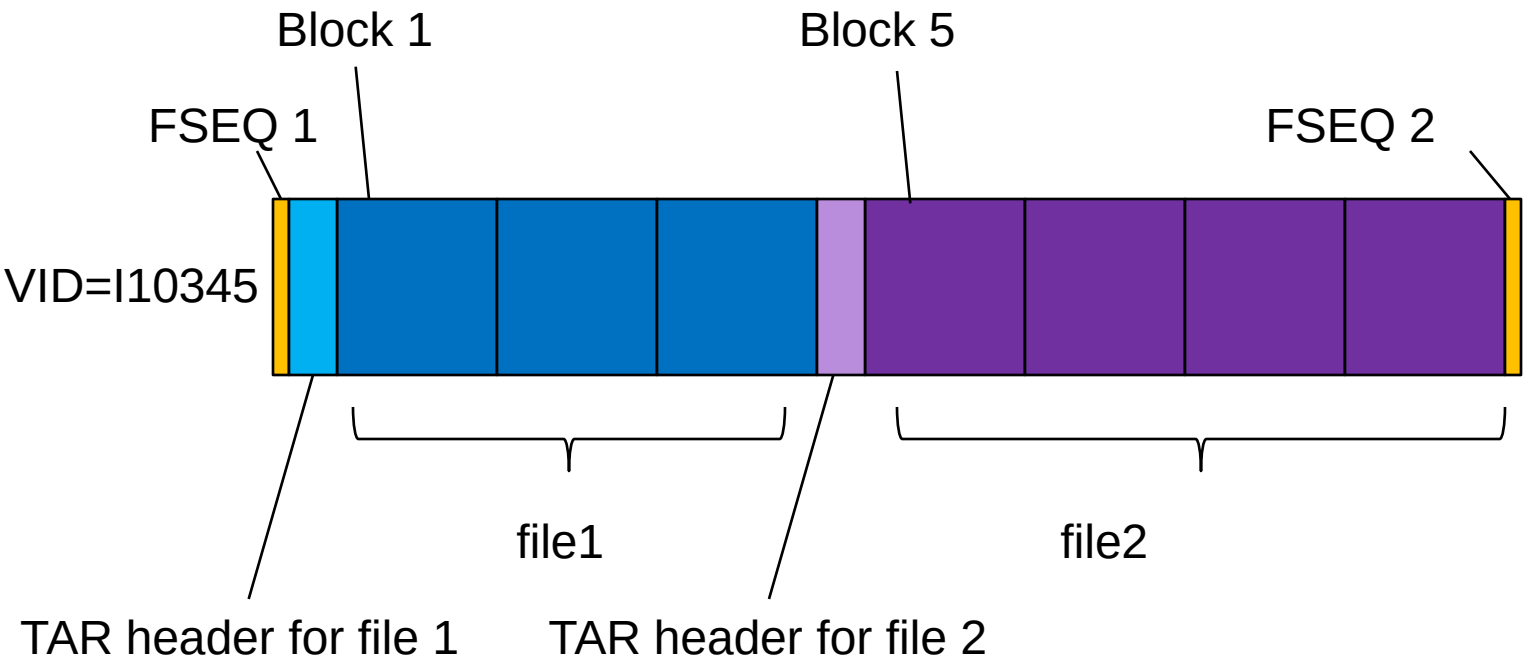
- What is it?
  - Tape software layer aggregates on-the-fly small files into tar files as they are written to tape
  - Tar block size is set to that of the tape
  - Tar files are written as files are today, using the AUL format
  - AUL format required to meet write position verification requirement
- Advantages
  - Configurable number of tape marks per tape
  - Industry standard format
- Disadvantages
  - Tape inspection requires knowledge of tar
  - NS needs to be modified
  - Not compatible with the rtcopy “more work” protocol

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

TAR + AUL tapes

NS entries

- **Filename=file 1, VID=I10345, FSEQ=1, block ID=1**
- **Filename=file 2, VID=I10345, FSEQ=1, block ID=5**



Current status

- Aggregation is the way to go
- More investigation required into format
  - IL + NL
  - TAR + AUL

Timelines

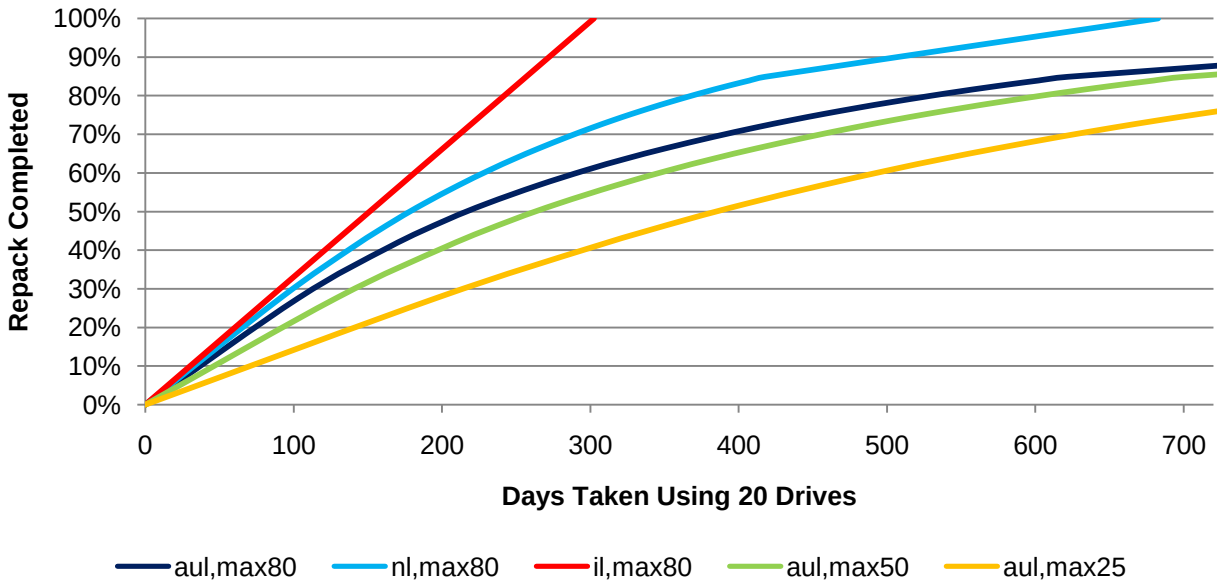
- Goal is to minimize media expenditure in 2009 by taking advantage of higher density writing on existing tapes.
  - Savings of ~ 3.2M CHF
- Software should be available by end Q1 2008 to avoid additional investment

Why are we copying ?

Vendor	Current	Future	At CERN	Delta Capacity	Cost to purchase
IBM	700GB	1000GB	9692	2.9PB	0.5MCHF
Sun 513	500GB	1000GB	14890	7.4PB	1.3MCHF
Sun 613	500GB	1000GB	15408	7.7PB	1.4MCHF
Total				18.0PB	3.2MCHF

- Cost to purchase is the additional media and slots required if we write at new densities but do not copy and recycle old tapes
- Adds up to a saving of 3.2M CHF
- With higher density and repack, media requirements for 2009 are covered
- Without higher density, 2 new 10,000 slot robots would be required in 2009

Repack using 20 drives



- Approach to take easy tapes with large files first
- Repack using aul tapes would take over 3 years to complete
- Max80 figures reflect the performance if engine is able to sustain reading at 80MBytes/s. Max50 for 50MBytes/s and Max25 for 25MBytes/s
- The 'to migrate' queue would be around 400,000 files at the end of processing if 20 drives are used.

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Tasks to be done

- Investigate hardware capabilities and evolution
- RTCOPY
  - Reverse engineer RTcopy daemons
  - Develop proof of concept to show that data write rate can be sustained within rtcpd with thousands of small files
  - Modify RTcopy daemons to be aggregate-based
  - Possible re-write of RTCPClientD
  - Modify RTCPD to write new format (IL or TAR)
- Modify VMGR to identify tapes using TAR / IL
- Modify tpread and tpwrite if necessary
- Develop test harness, test suite and production time verification software to validate new format

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Disk cache modifications

- Modify NS to support aggregate / bulk updates
- Modify NS and STAGER to store new information
  - Aggregations, their sizes and checksums
  - File uniqueness still to be decided (VID + FSEQ + block ID)

Tape capabilities

- **Could an IL format in an NL tape be read out by tpread, what modifications if any would need to be made to tpread?**
- **Do we need to have the containers providing aggregation, e.g . tar files, recorded as first-class objects within the NS?**
- **Does a write to an AUL tape check its position before writing?**
- **Can all drives do FSEQ + block offset?**
- **Do the ANSI tape header labels contain the size of the file?**
- **Does a SCSI sync cost as much time as a file mark?**
- **What does a file mark give you beyond a sync (very open ended)?**
- **Can you go to the end of data on a tape without a file mark?**
- **Do the AUL tape labels contain a checksum?**
- **How and where could the work on tar or IL be cleanly divided between Steve and Arne, being fair to both?**
- **Why was the IL block format chosen, was it to facilitate fault diagnosis for example?**
- **For fault diagnosis, is IL better than TAR + AUL**
- **Can you seek to an absolute block id on an IBM (+ LTO) drive?**
- **Can you tell which absolute block id you are at on an IBM (+ LTO) drive?**

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it