COMPASS and HARP
Objectivity to Oracle Migration (2002-2004)

Andrea Valassi (CERN)

Third Workshop on Data Preservation and Long Term Analysis in HEP
CERN, 7th December 2009

References

- **M. Lübeck, D. Geppert, K. Nienartowicz, M. Nowak, A.Valassi,** ***"An Overview of a Large-Scale Data Migration"***, **IEEE/NASA MSST 2003, San Diego**
  http://storageconference.org/2003/presentations.html

- **M. Nowak, D. Geppert, M. Lübeck, K. Nienartowicz, A.Valassi,** ***"Objectivity Data Migration"***, **CHEP 2003, La Jolla**
  http://www.slac.stanford.edu/econf/C0303241/proc/cat_8.html

- **V. Duic, M. Lamanna,** ***"The COMPASS Event Store in 2002"***, **CHEP 2003, La Jolla**
  http://www.slac.stanford.edu/econf/C0303241/proc/cat_8.html

- **A.Valassi, D. Geppert, M. Lübeck, K. Nienartowicz, M. Nowak, E. Tcherniaev, D. Kolev,** ***"HARP Data and Software Migration from Objectivity to Oracle"***, **CHEP 2004, Interlaken**
  http://indico.cern.ch/contributionDisplay.py?contribId=448&sessionId=24&confId=0

- **A.Valassi,** ***"HARP Raw Event Database"***, **LHC DB Developers Workshop 2005, CERN**
  http://indico.cern.ch/conferenceDisplay.py?confId=a044825#11

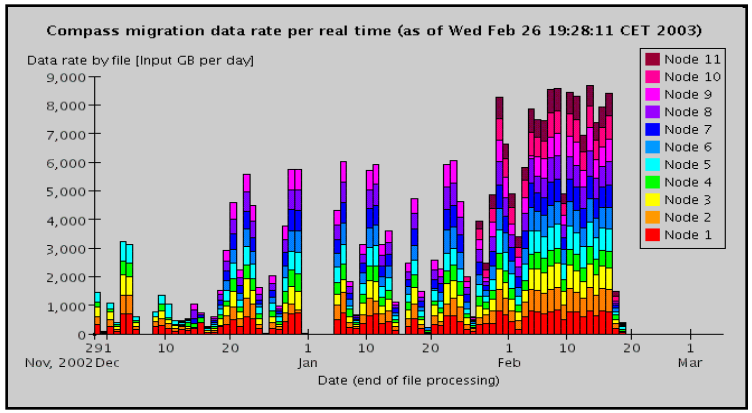
Thanks to Jamie Shiers for many useful comments about this presentation

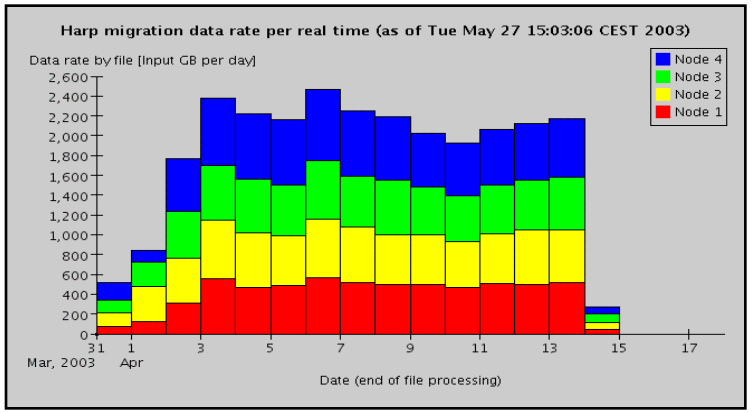
Overview

- **Main motivation: end of support for Objectivity at CERN**
  - The end of the object database days at CERN (July 2003)
  - The use of relational databases (e.g. Oracle) to store physics data has become pervasive in the experiments since the Objectivity migration
- **Double (or triple!) migration**
  - Data format (and software!) conversion from Objectivity to Oracle
  - Physical media migration from StorageTek 9940A to 9940B tapes
- **Data sets involved**
  - **COMPASS** raw event data (300 TB)
    - Data taking continued after the migration, using the new Oracle software
  - **HARP** raw event data (30 TB), event collections and conditions data
    - Data taking stopped in 2002, no need to port event writing infrastructure
  - Objectivity used by LHC experiments too, but with no production data

Migration history

- **Summer 2002: start preparing the migration (team of 5 in CERN IT)**
- **Dec 2002 to Feb 2003: COMPASS raw event data migration**
  - 300 TB in 3 months at 100 MB/s and 2000 rows/s peak rate
  - New storage system validated before the 2003 data taking
- **Apr 2003: HARP raw event data migration**
  - Fewer nodes but much higher efficiency, thanks to COMPASS experience





- **Summer 2003: HARP event collection metadata migration**
  - Longest phase (most complex data model) in spite of low data volumes
- **End 2003: HARP conditions data migration**
  - Jan 2004: final validation of new storage system for HARP data analysis

COMPASS & HARP raw event data

CERN IT Department

- **Both experiments used the same model in Objectivity**
  - Raw data for one event streamed as one binary large object (BLOB)
    - Using the "DATE" format - *independent of Objectivity*
    - Each such BLOB is encapsulated in one "event object" in Objectivity
  - One 'database' file contains all events in a partial run (subset of a run)
    - COMPASS: 200k files (300 TB) archived on 3400 CASTOR tapes
    - Objectivity 'federation' (metadata of database files) permanently on disk

- **Migrate both experiments to the same 'hybrid' model**
  - Migrate all raw event BLOB records to flat files in CASTOR
    - *Treat BLOBs as black boxes – no need to decode and re-encode them*
    - *This was possible because DATE format is independent of Objectivity*
  - Migrate BLOB metadata (file offset and size) to Oracle database
    - Large partitioned tables (COMPASS: 6×10^9 event records)

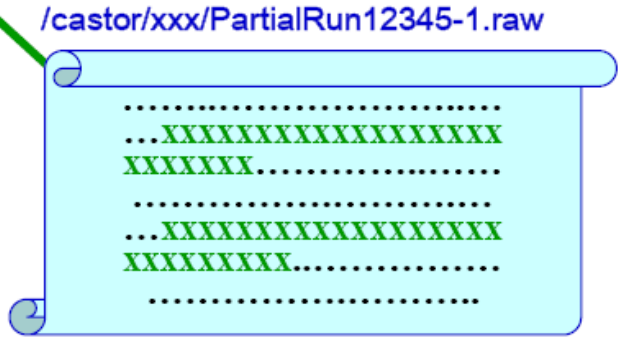
CERN

Raw events in Oracle plus CASTOR

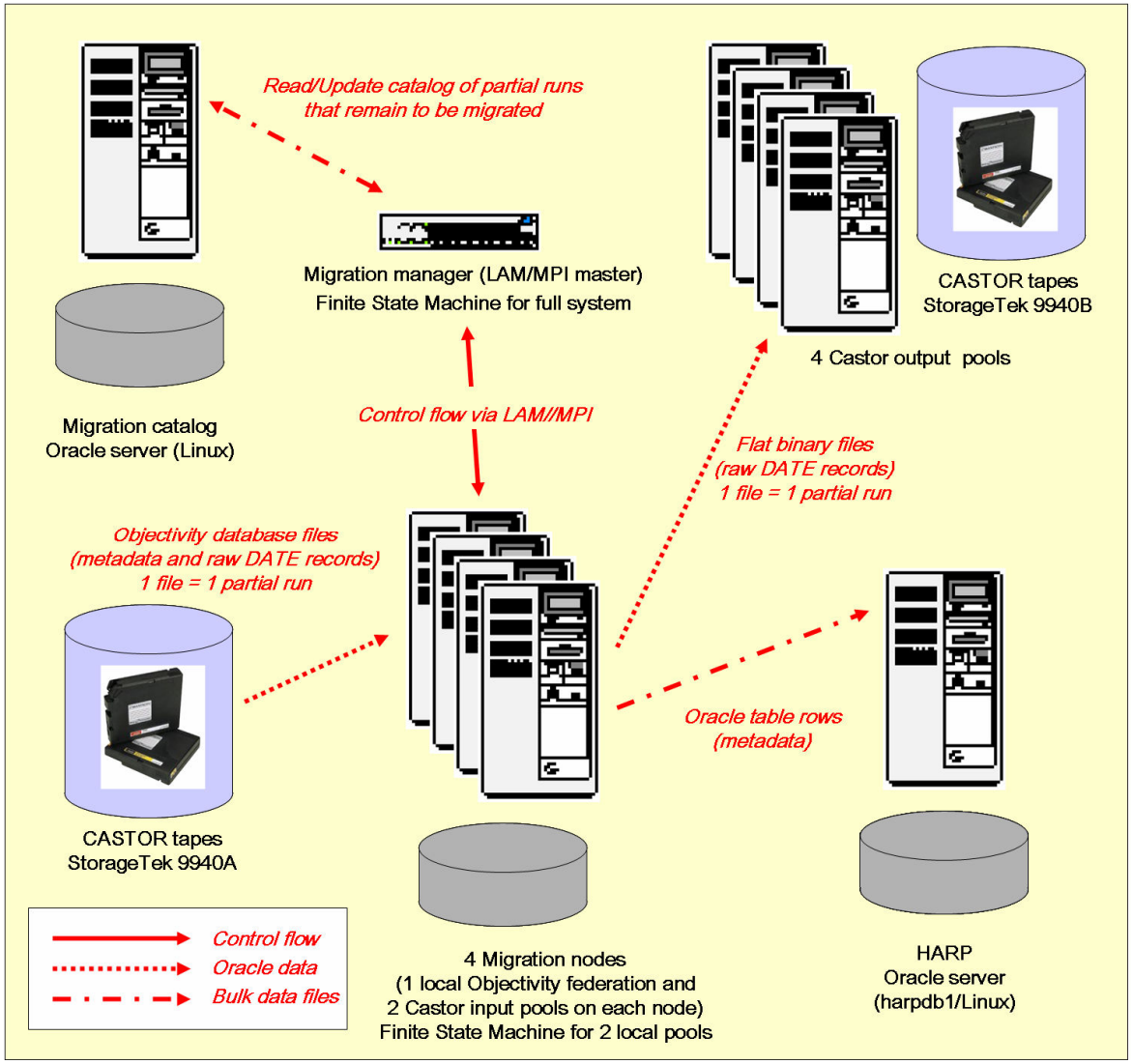
CERN IT Department

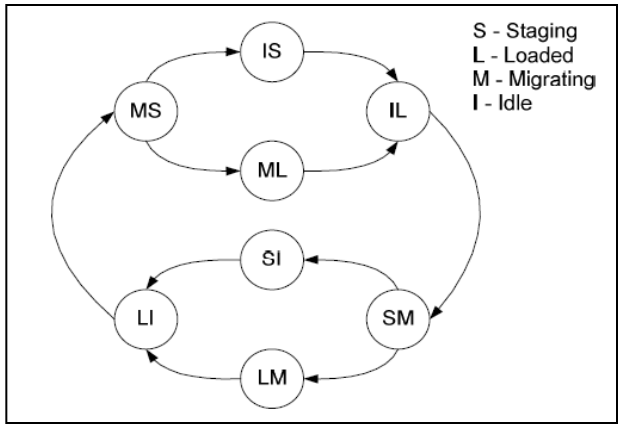
- **Main table in Oracle: the "event table"**
  - Metadata:100 GB (HARP)
  - Offset and size of the corresponding event record in the CASTOR file for that partial run
- **BLOBs in flat files**
  - Raw data: 30 TB (HARP)
  - Could have stored them inside Oracle, but saw no obvious advantage in this
    - No need to query BLOBs

RAW_EVENT		
PK,FK1,U1,I1	RUN_ID	N-Decimal(5,0)
PK,FK1,U1	EVB_ID	N-Decimal(1,0)
PK,FK1	SPILL_ID	N-Decimal(10,0)
PK	EVENTINPARTIALSPILL_ID	N-Decimal(10,0)
U1	EVENT_TIME	N-Decimal(10,0)
FK2,I2	EVENT_TYPEID	N-Decimal(1,0)
I1	EVENT_ID	N-Decimal(10,0)
	EVENTINSPILL_ID	N-Decimal(10,0)
	DHA_ATTTYPE0	N-Decimal(10,0)
	DHA_ATTTYPE1	N-Decimal(10,0)
	DHA_DETMASK0	N-Decimal(10,0)
	DHA_DETMASK1	N-Decimal(10,0)
	DHA_DETMASK2	N-Decimal(10,0)
U1	DHA_TIMEMICROSEC	N-Decimal(10,0)
	EVENTRECORD_FILEOFFSET	N-Decimal(10,0)
	EVENTRECORD_SIZE	N-Decimal(10,0)
	EVENTRECORD_OBJYOID	C-Variable Length(30)
	ORACLE_INSERTIONTIME	T-Auto Timestamp
	ORACLE_INSERTIONHOST	N-Decimal(3,0)
	ORACLE_LASTUPDATETIME	T-Auto Timestamp
	ORACLE_LASTUPDATEHOST	N-Decimal(3,0)



Raw event migration infrastructure





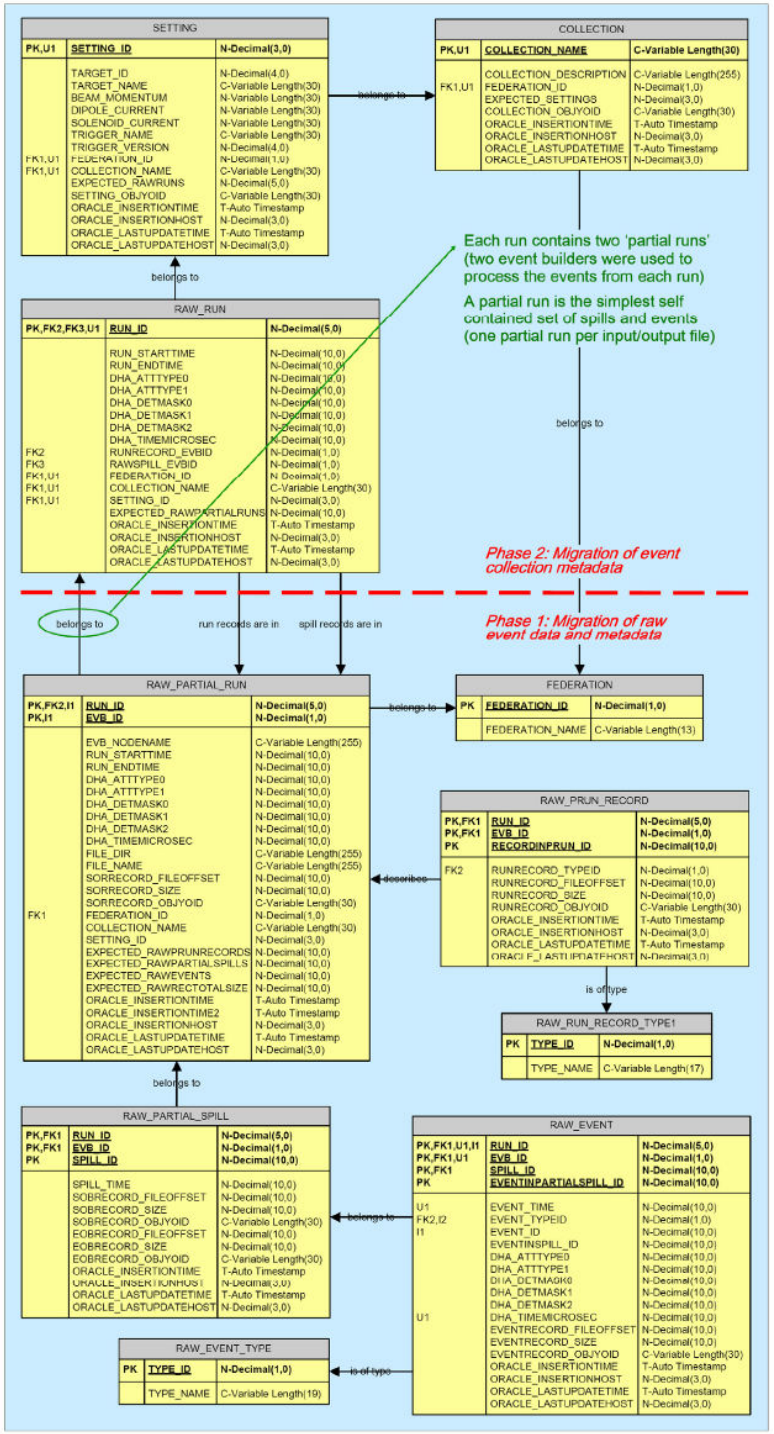
Two jobs per migration node (one staging, one migrating)

Setup to migrate the 30 TB of HARP (4 migration nodes)

[A similar setup with a larger number of nodes (11) had been used to migrate the 300 TB of COMPASS]

HARP event collections

- **Longest phase: lowest volume, but most complex data model**
  - Reimplementation of event navigation references in the new Oracle schema
  - Reimplementation of event selection in the Oracle-based C++ software
    - Exploit server-side Oracle queries
  - *Completely different technologies (object vs. relational database)*



HARP conditions data

- **Stored using technology-neutral abstract API by CERN IT**
  - Software libraries for time-varying conditions data (e.g. calibrations)
  - Two implementations already existed for Objectivity and Oracle
- ***This was the fastest phase of the migration***
  - *Abstract API decouples experiment software from storage back-end*
    - *Almost nothing to change in the HARP software to read Oracle conditions*
    - *Actual migration partly done through generic tools based on abstract API*
- **Compare to LHC experiments using CORAL and/or COOL**
  - Abstract API supporting Oracle, MySQL, SQLite, Frontier back-ends
    - Strictly nothing to change in the experiment software to switch back-end
    - Used now for data distribution, DB failover… later for data preservation?

Outlook - lessons learnt?

CERN IT Department

- **Data migrations are unavoidable**
  - To preserve the bits (e.g. end of support for tape hardware)
  - To preserve the ability to use the bits (e.g. end of support for software)
    - But you must also preserve people's expertise to make sense of the bits!
- **Data migrations have a cost**
  - In this case: several months of computing resources and manpower
- **Layered approach to data storage software helps**
  - Software decoupling makes it easier to replace backend technology