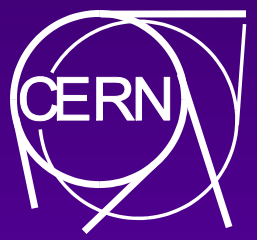


An overview of a Large-Scale Double Data Migration

From ODBMS to LHC-like solution

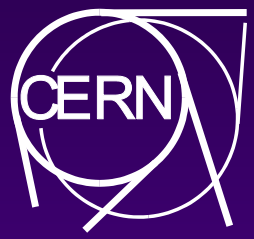
IEEE Conference on Mass Storage Systems and Technologies 2003

April 8, 2003

Magnus Lübeck

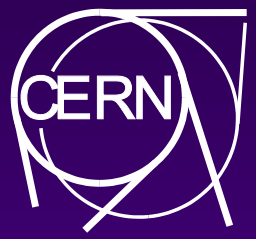
CERN Database group, http://cern.ch/db/

CERN – The European Laboratory for Particle Physics



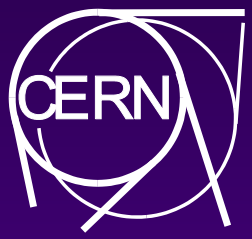
Hundreds of TB

- Double migration
  - Media migration
  - Database migration
- The migration was a non trivial task
  - The amount of data is (today) considered to be large
  - Plenty of technological obstacles
  - Had to be done during:
    - Regular CERN production
    - LHC Data challenges



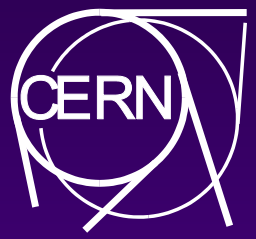
Migration is part of operation

- For long term projects (10 – 20 years) there are only two choices for long term storage
- If you keep a lot of data you better plan for migrations!
- Main issues this time



Migration is part of operation

- For long term projects (10 – 20 years) there are only two choices for long term storage
  - Either find technology that lasts forever
  - Plan for migrations
- If you keep a lot of data you better plan for migrations!
  - Decouple dependencies
    - Database technology
    - Storage
  - Plan for changes
- Main issues this time
  - Moving from an Object Database to Oracle 9i
    - Decision of LHC experiments to change persistency baseline
      - Based on Risk Analysis for 15 year lifetime of LHC
    - Next datataking for COMPASS start early spring 2003

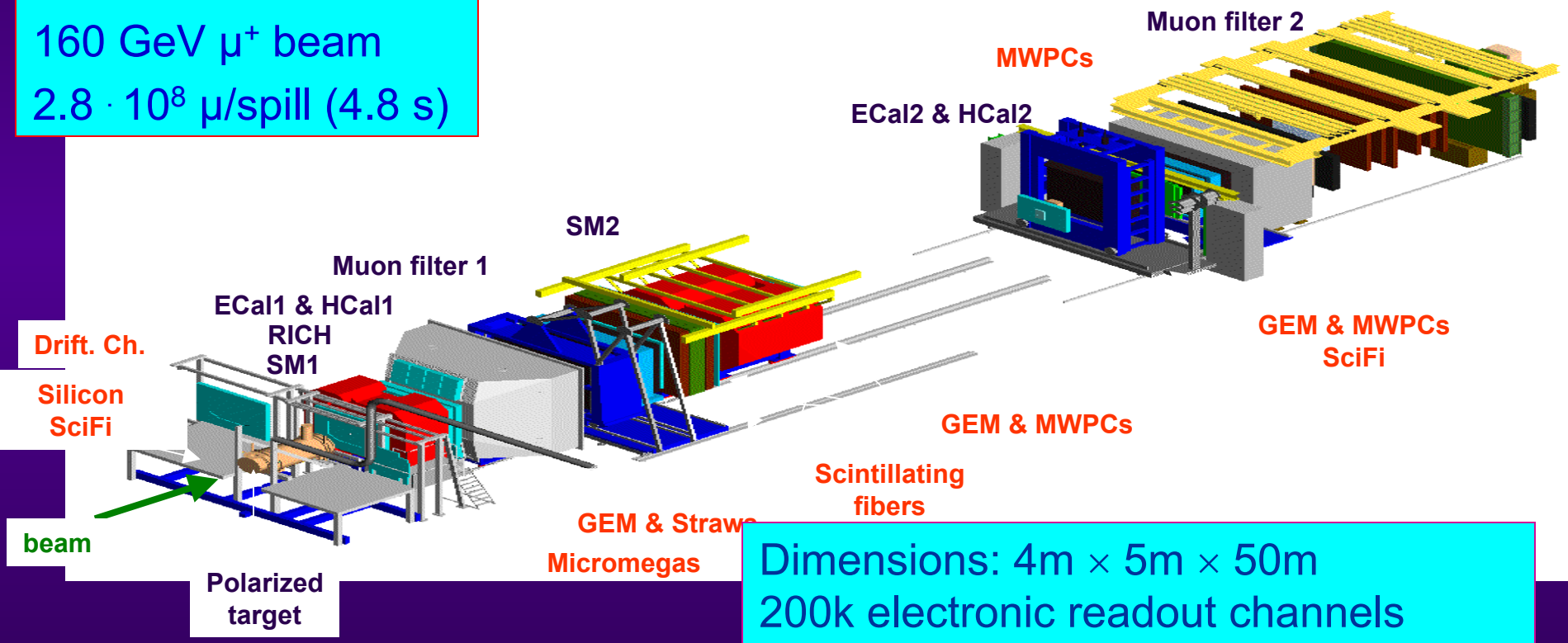


The COMPASS experiment

- COmmon Muon and Proton Apparatus for Structure and Spectroscopy
- More than 200 physicists from 26 institutes
- A High Energy Physics experiment
  - Hadron structure and hadron spectroscopy with high intensity muon and hadron beams.
  - Fixed target experiment
  - High datarate
- http://wwwcompass.cern.ch

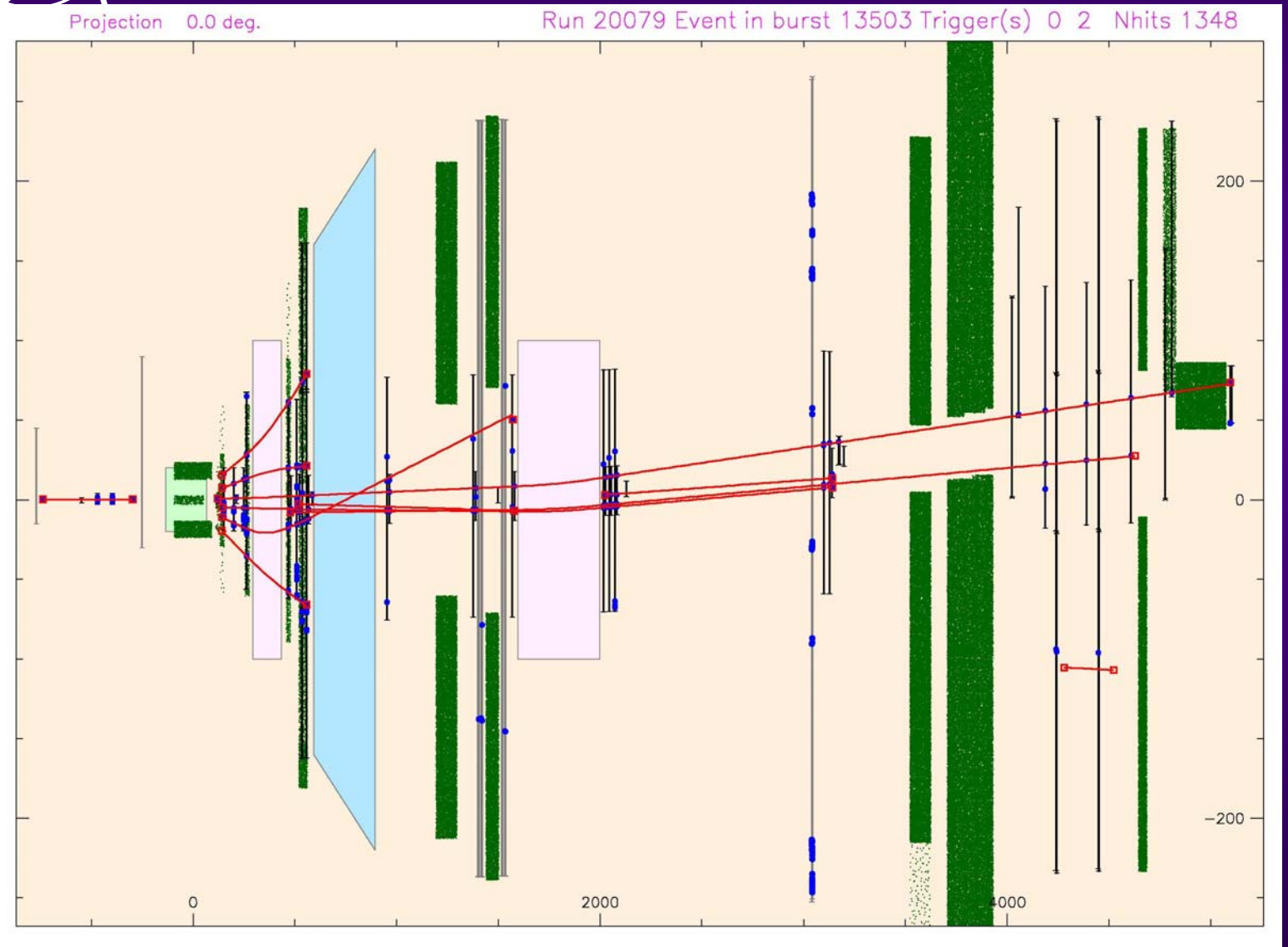
The COMPASS experiment

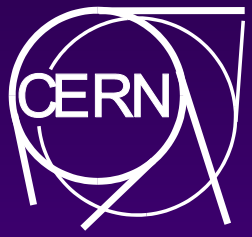
2002 run:
160 GeV μ^+ beam
$2.8 \cdot 10^8$ μ/spill (4.8 s)



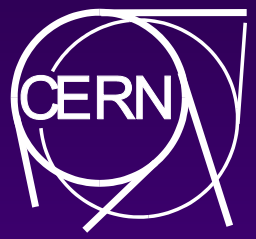
Dimensions: 4m × 5m × 50m
200k electronic readout channels
40kB event size
Trigger rate: 20k ev/spill = 4kHz
Datarate is 35 - 60MByte/s

A typical event



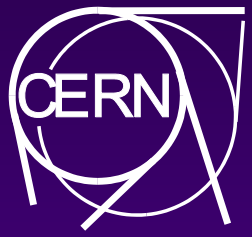


- **Migrating 300TB of data**
  - **300.000 files**
  - **Data stored on ~3500 tapes stored in a Mass Storage System**
  - **Assumed Constraints**
    - One node - 10MB/s -> one year
    - Available time: ~2 months -> MUST do processing in paralell
      - 2 months -> migrating 70+MByte/sec
      - Need at least 10 conversion nodes
      - Need 9 input and 5 output drives
  - **All hardware resources available inhouse**
    - no purchases – low lead time
    - Resources allocated when needed
  - **5 people working in the project for 8 months**



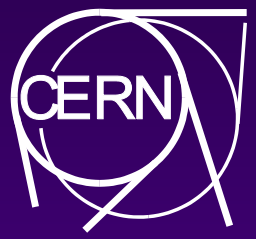
Planning the migration

- **How to do it?**
  - **Analyzing the environment**
    - How to control the migration system?
    - How do we monitor the progress?
    - How do we catch possible errors?
    - What about consistency?
  - **Discussing the metadata**
    - Should we use clustered databases?
    - What setup meets up with the requirements?



Technologies used in the migration

- Conversion
  - State machine descriptive language
  - LINUX/RedHat 7.3
  - LAM/MPI
  - C++
  - OCCI (Oracle C++ Interface)
- Monitoring
  - Apache
  - Java servlets
  - OC4J
  - Perl
  - DBI/DBD
- I/O
  - CASTOR Mass storage - POSIX like system calls for mass storage
  - Gbit networks
  - Tape silos
- RDBMS/Databases
  - Objectivity
  - Oracle 9i
  - SQL



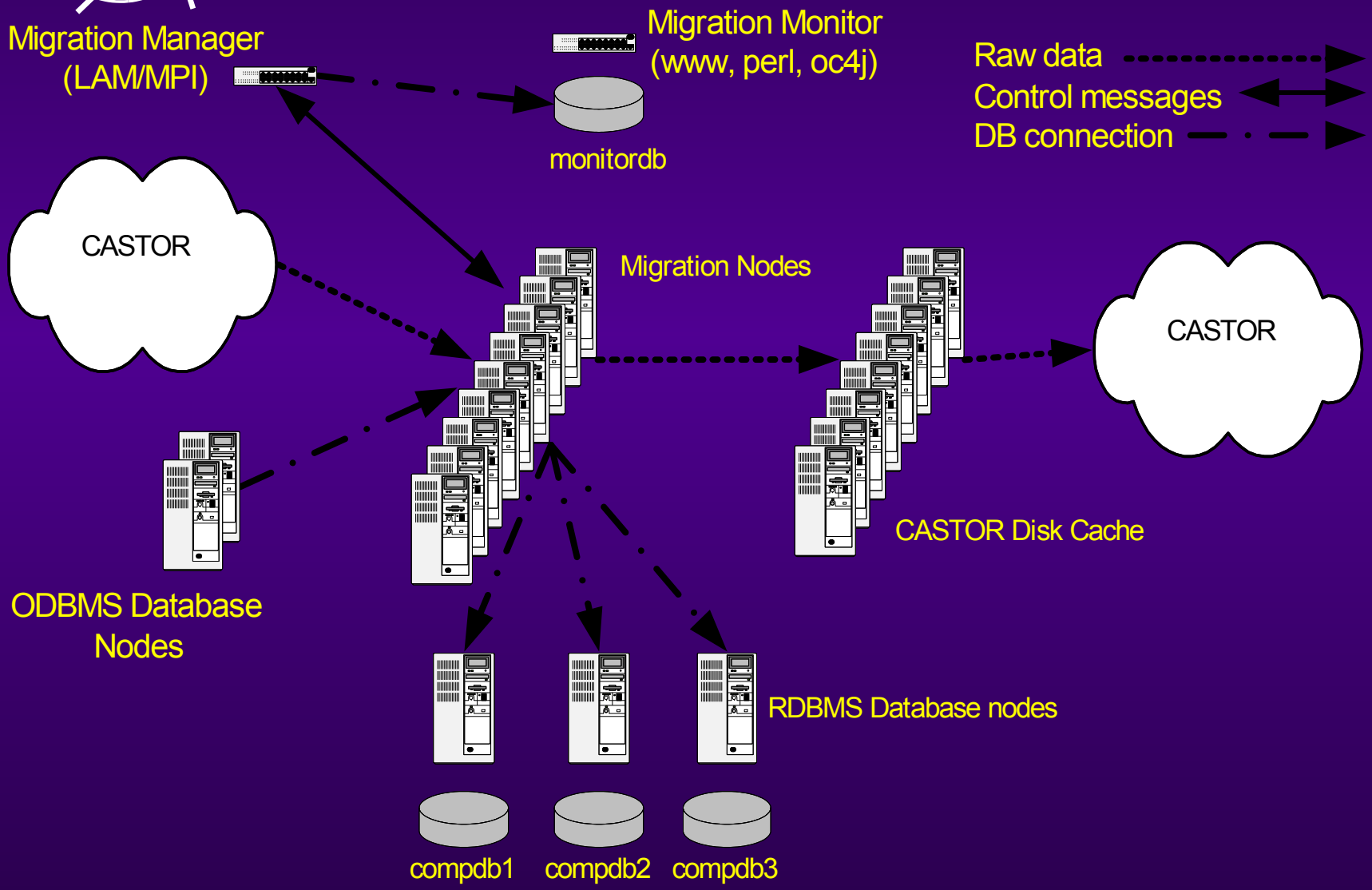
CASTOR – Mass Storage at CERN

- CERN Advanced STORage system
- HSM system built at CERN to meet the physicists' need for scalable/guaranteed bandwidth to storage
- POSIX like API (open(), write(), close()) to files
- Command line tools also available
- LINUX disk servers used for caching
- Tape drives in STK silos for offline storage

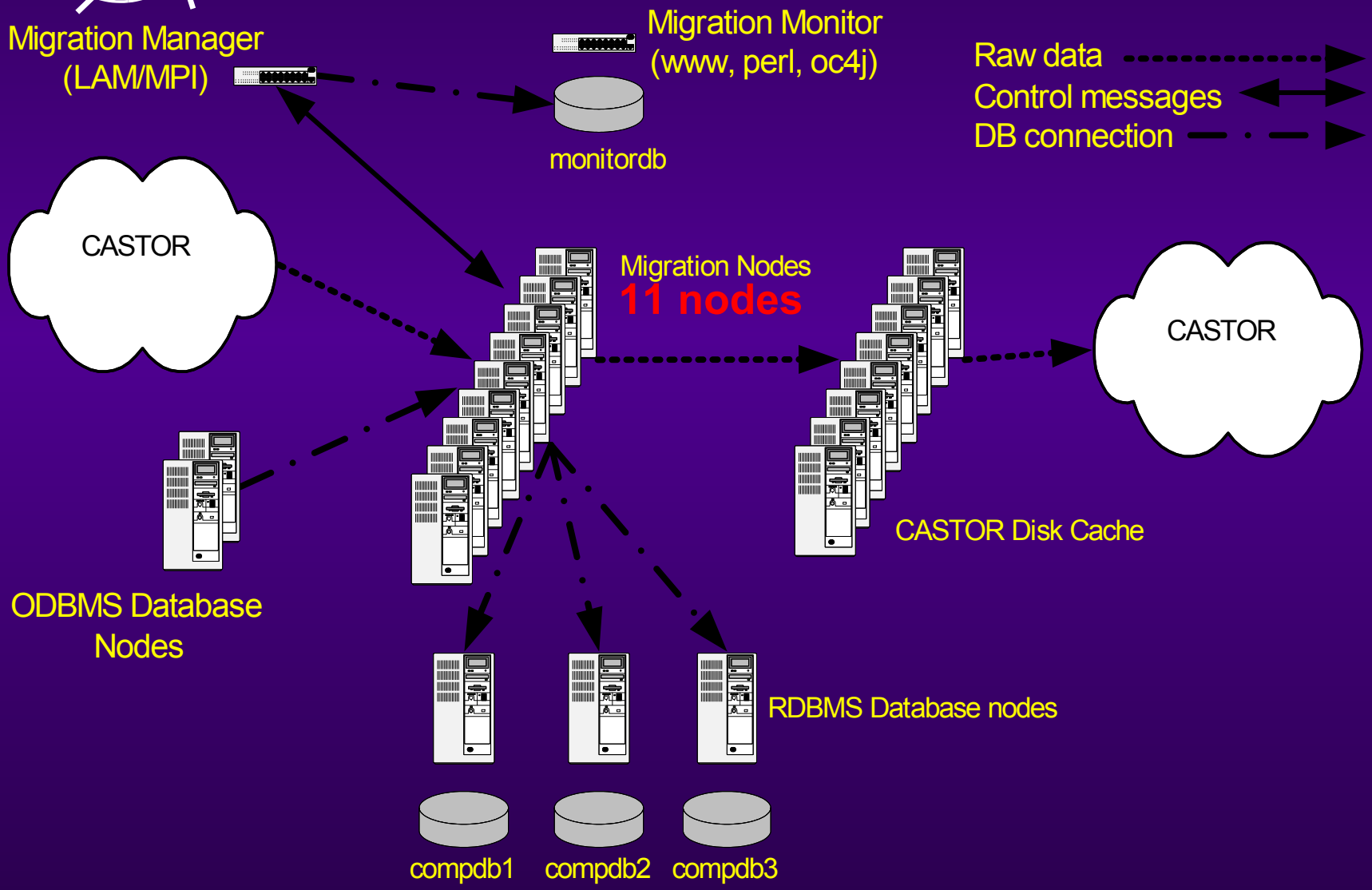

EMERGENCY EXIT


COMPASS
STORAGE TEK
POWDERHORN TAPE LIBRARY
UP TO 300TB/SILO
IBM 3590E tape drive
40 GB cartridge – 14 MB/s

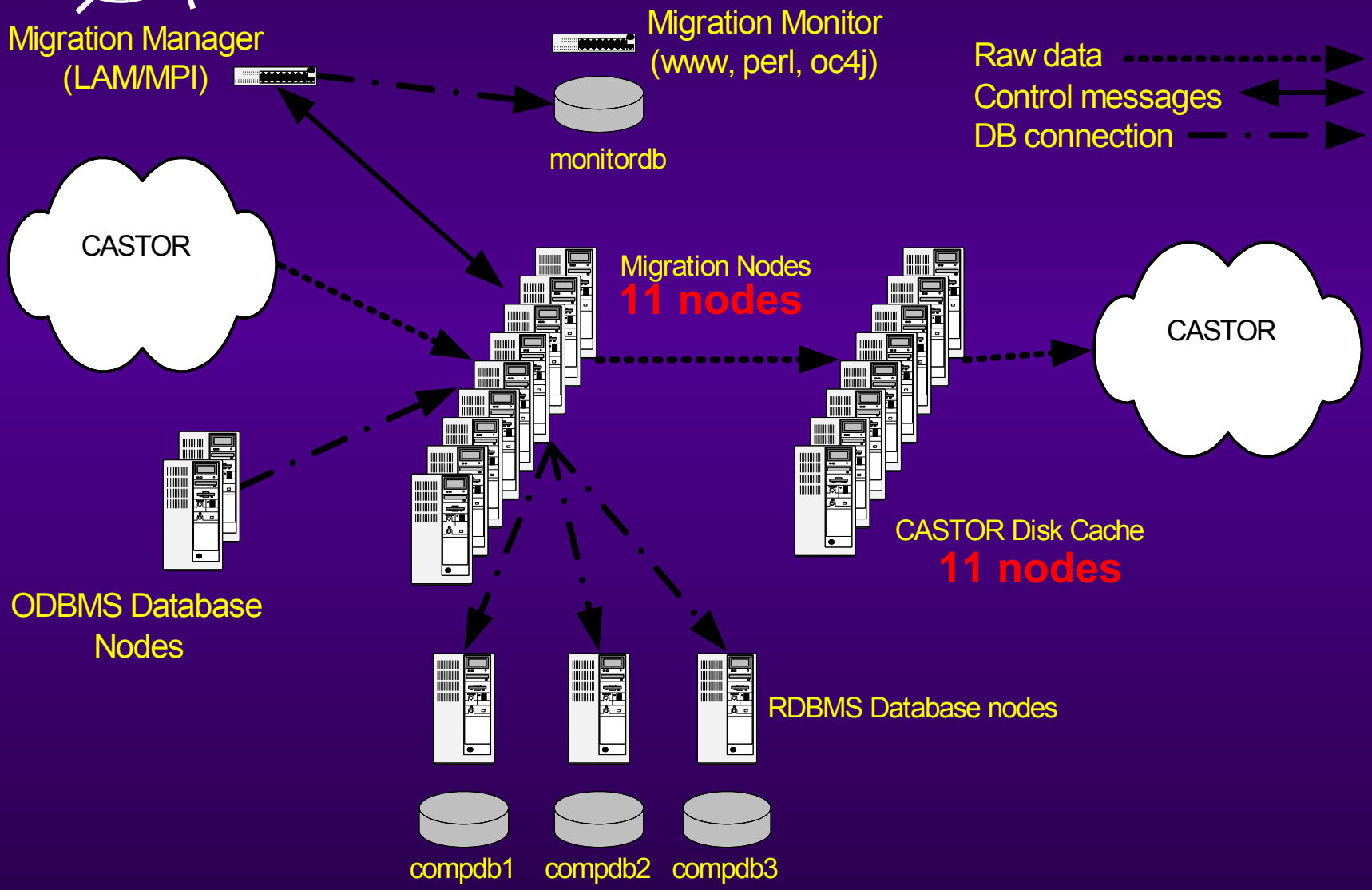
Conceptual view



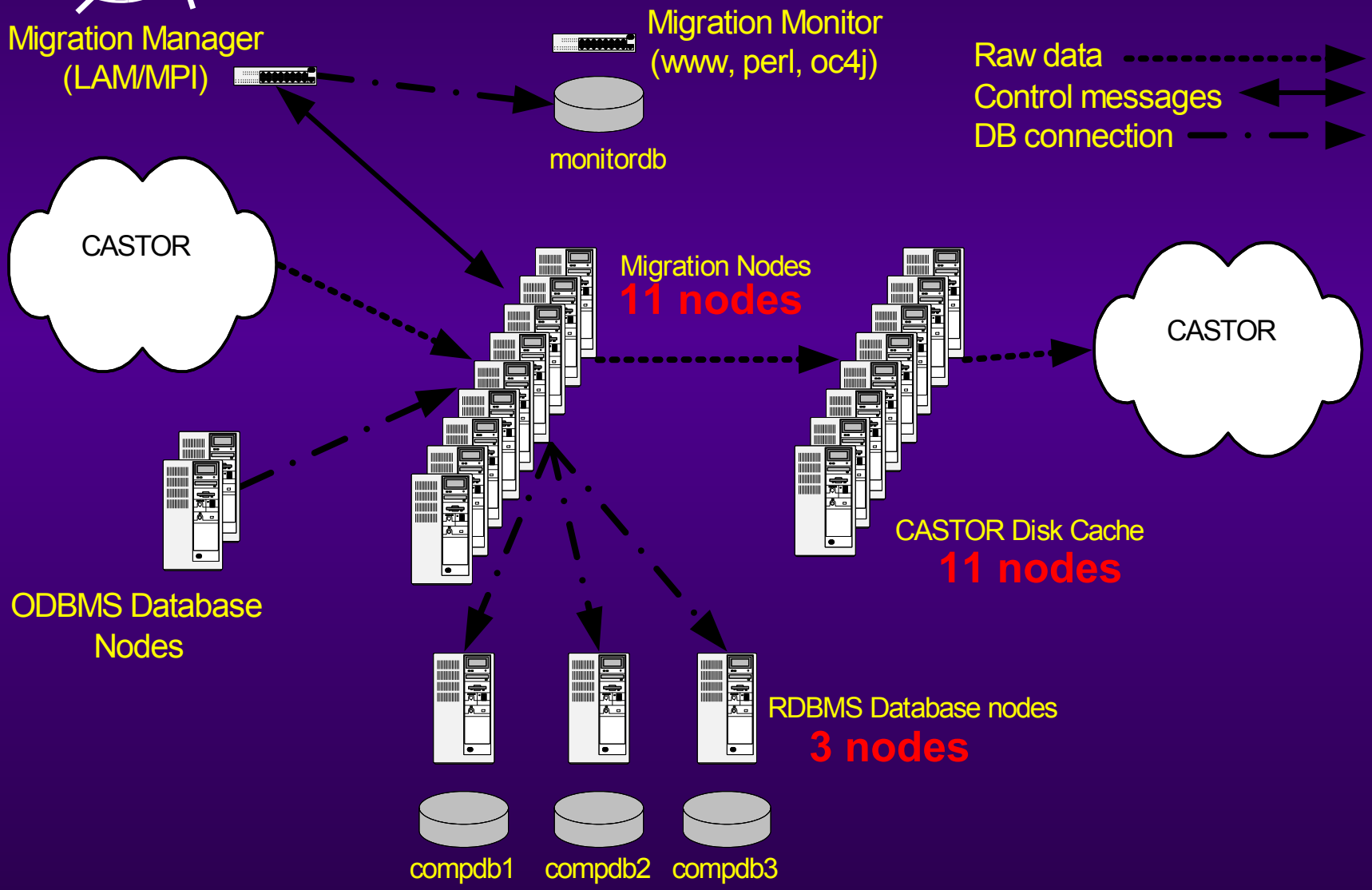
Conceptual view



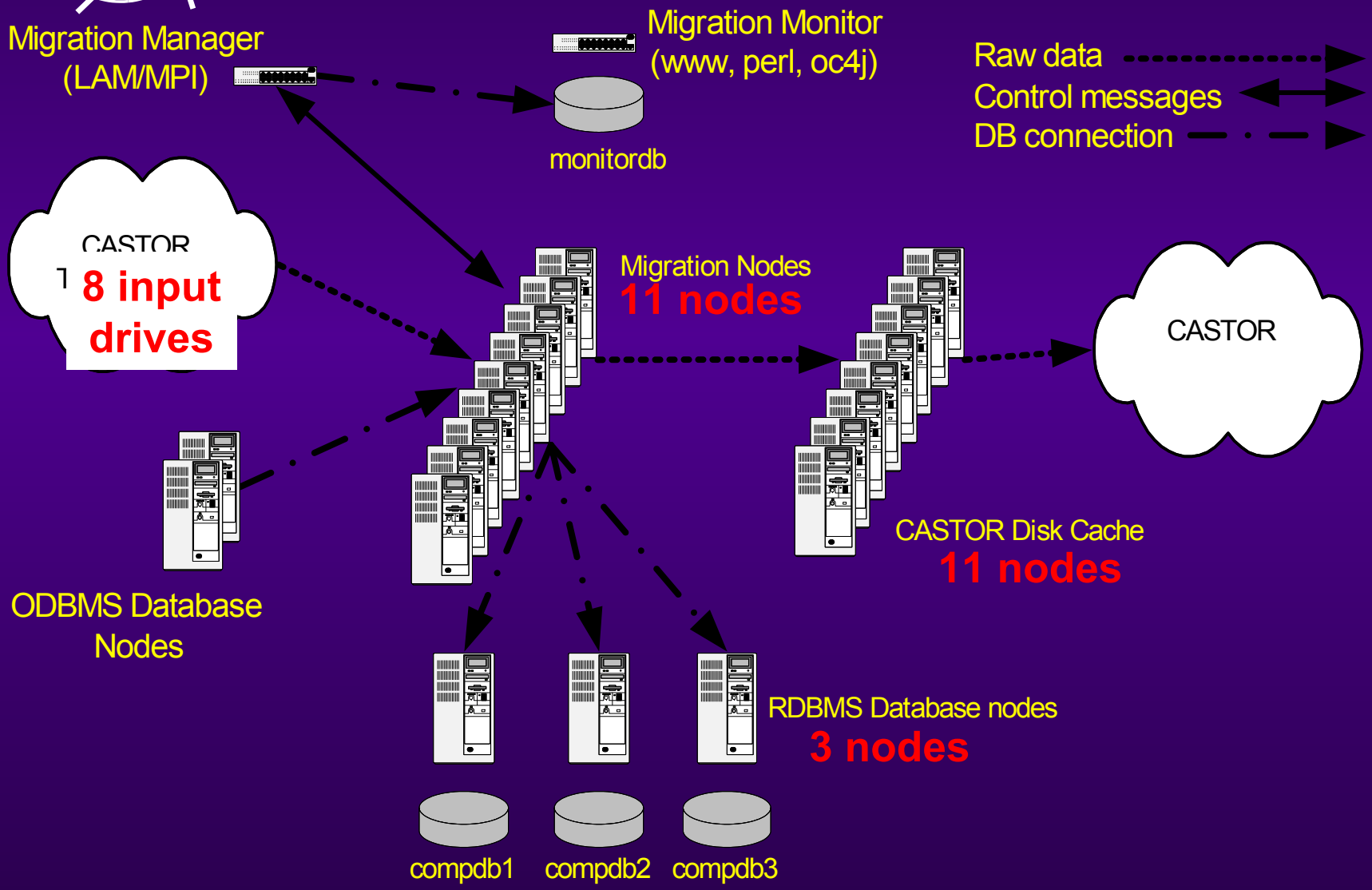
Conceptual view



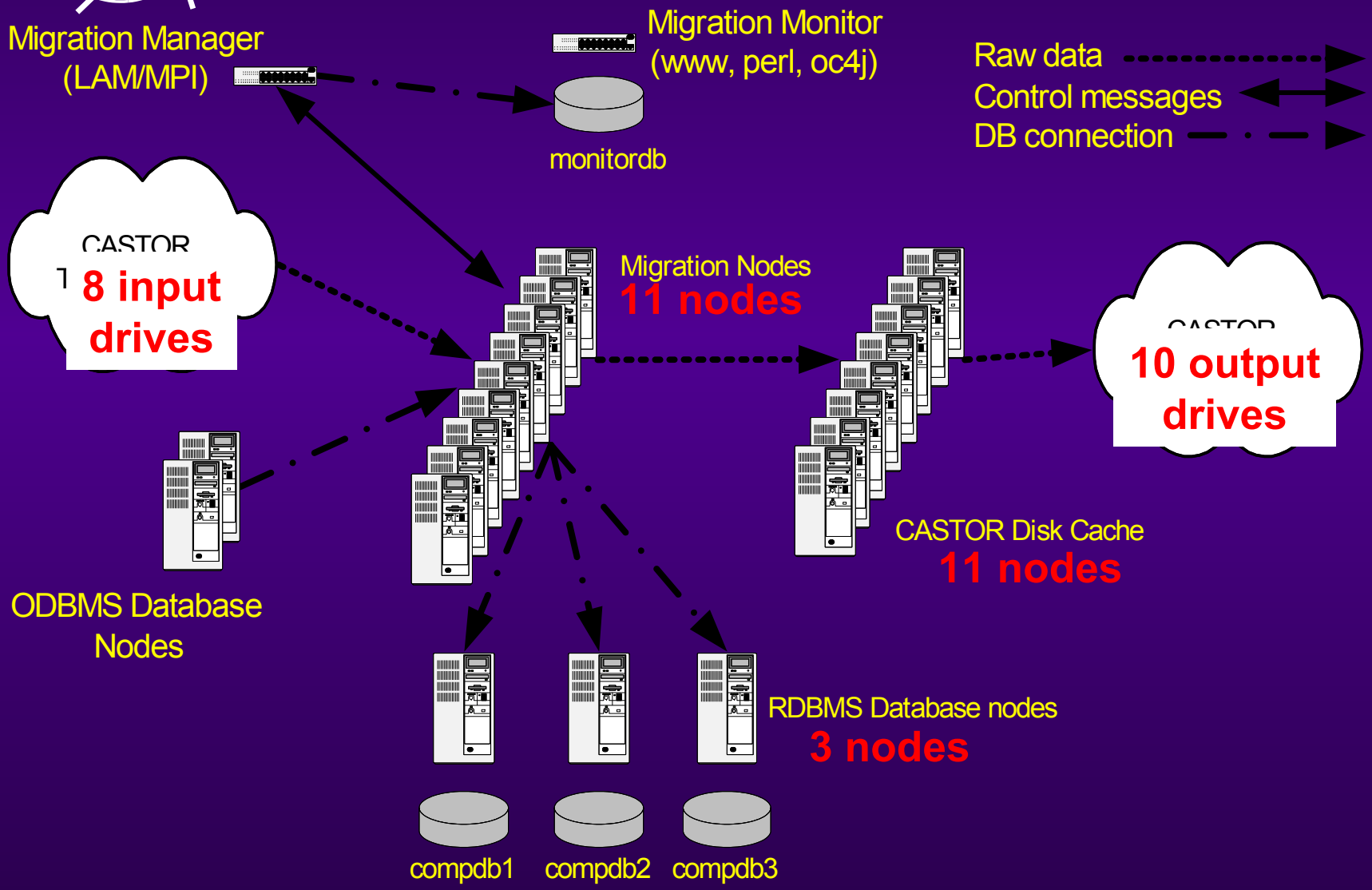
Conceptual view

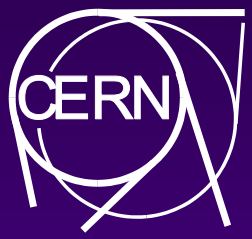


Conceptual view



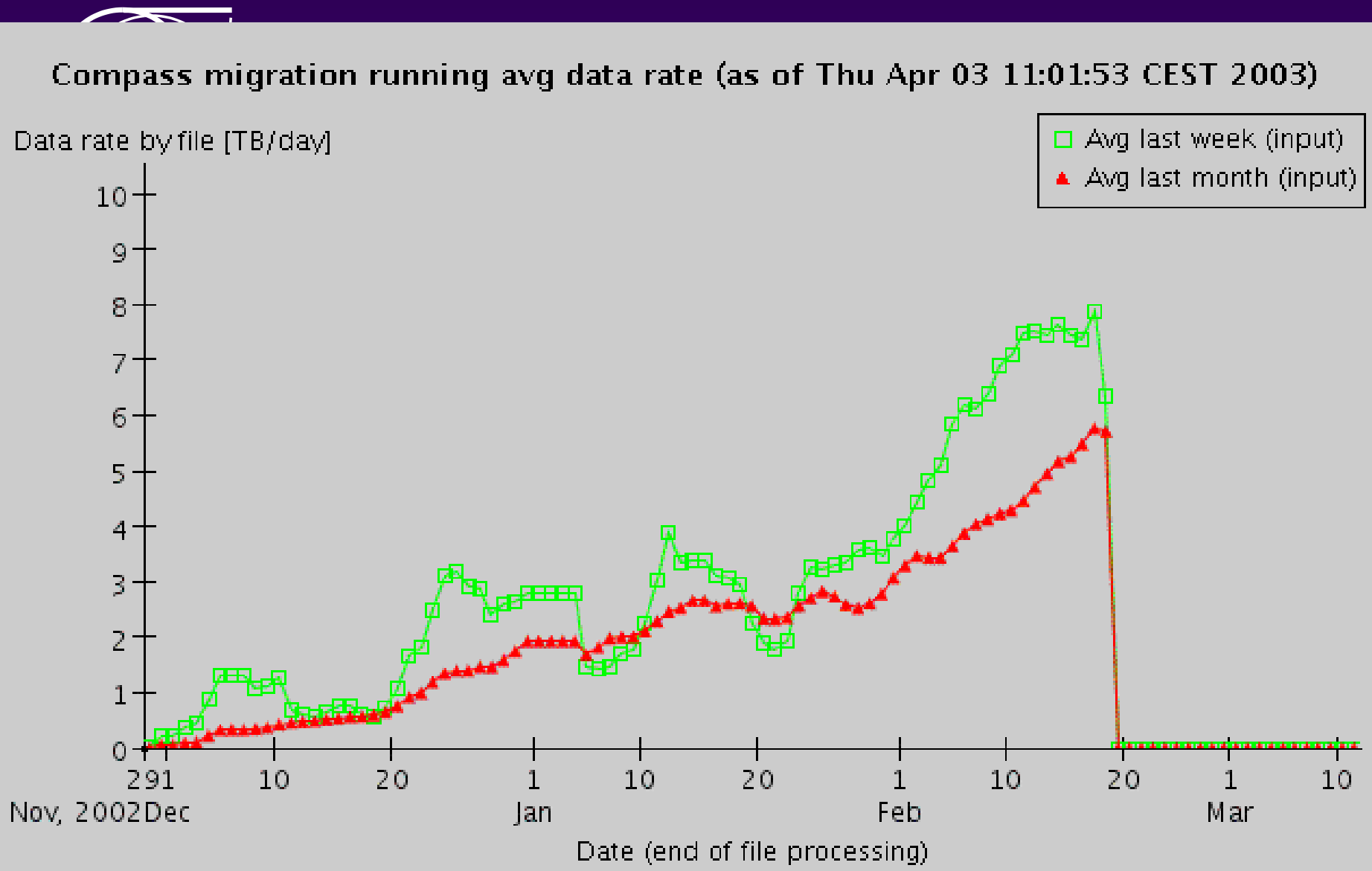
Conceptual view

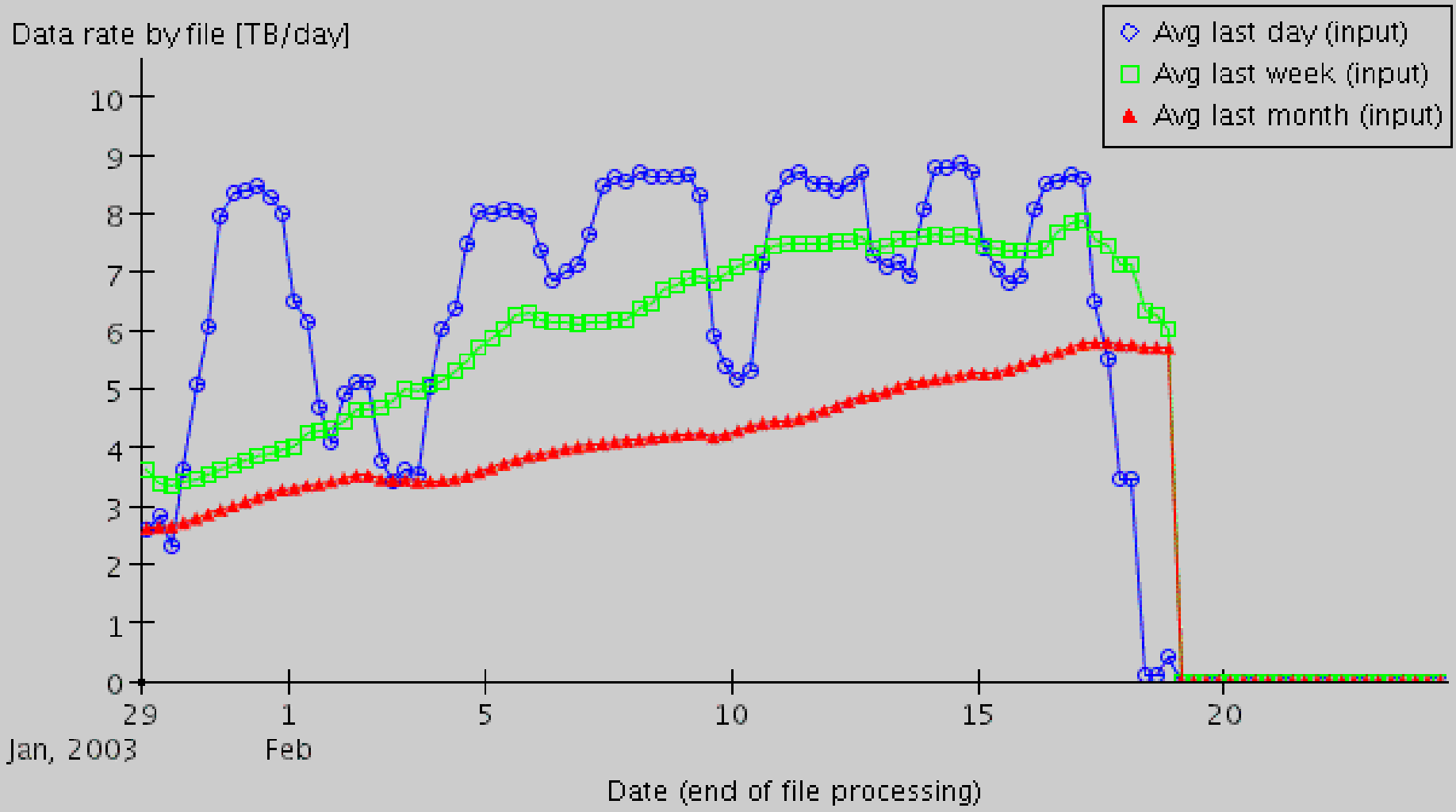


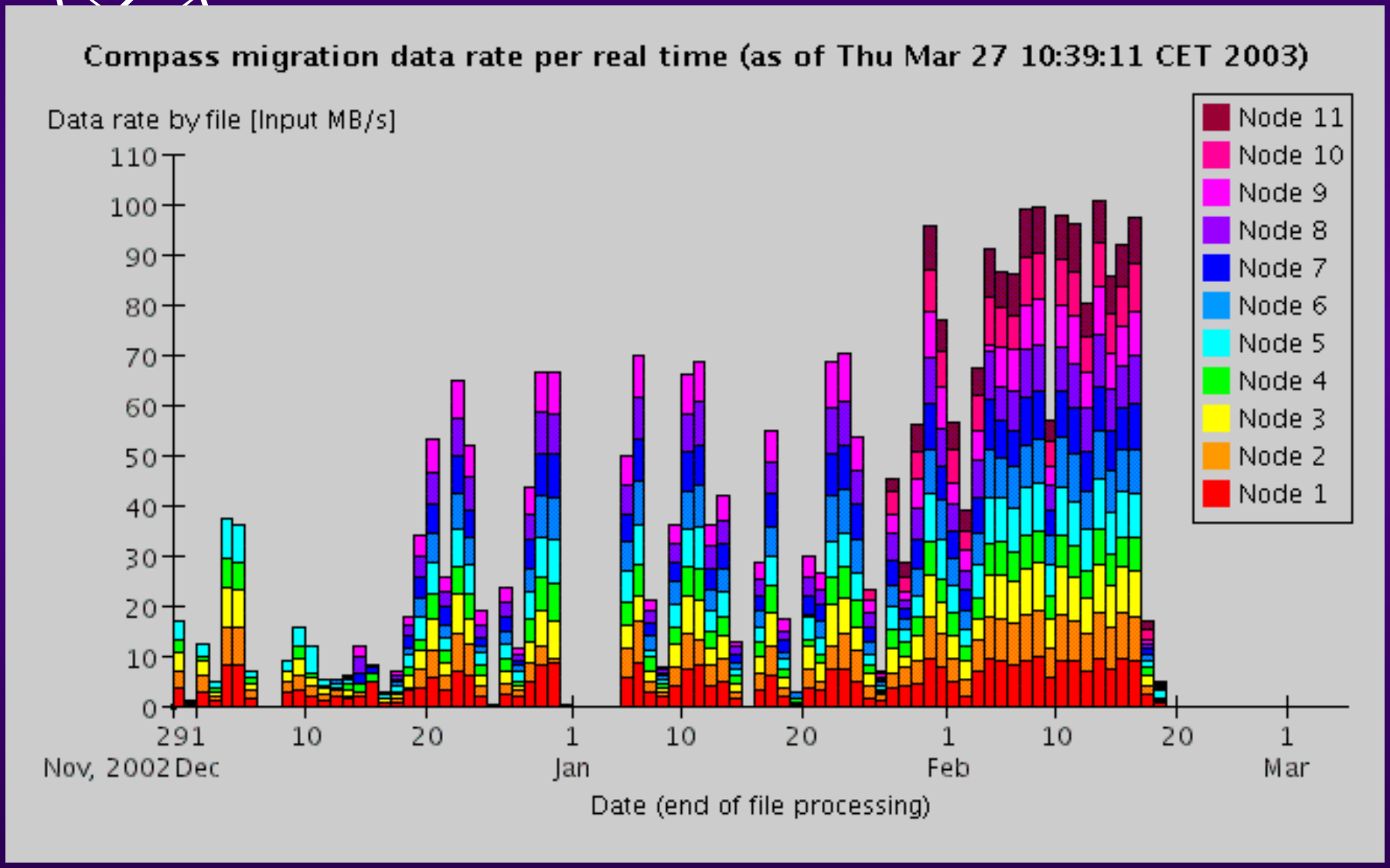


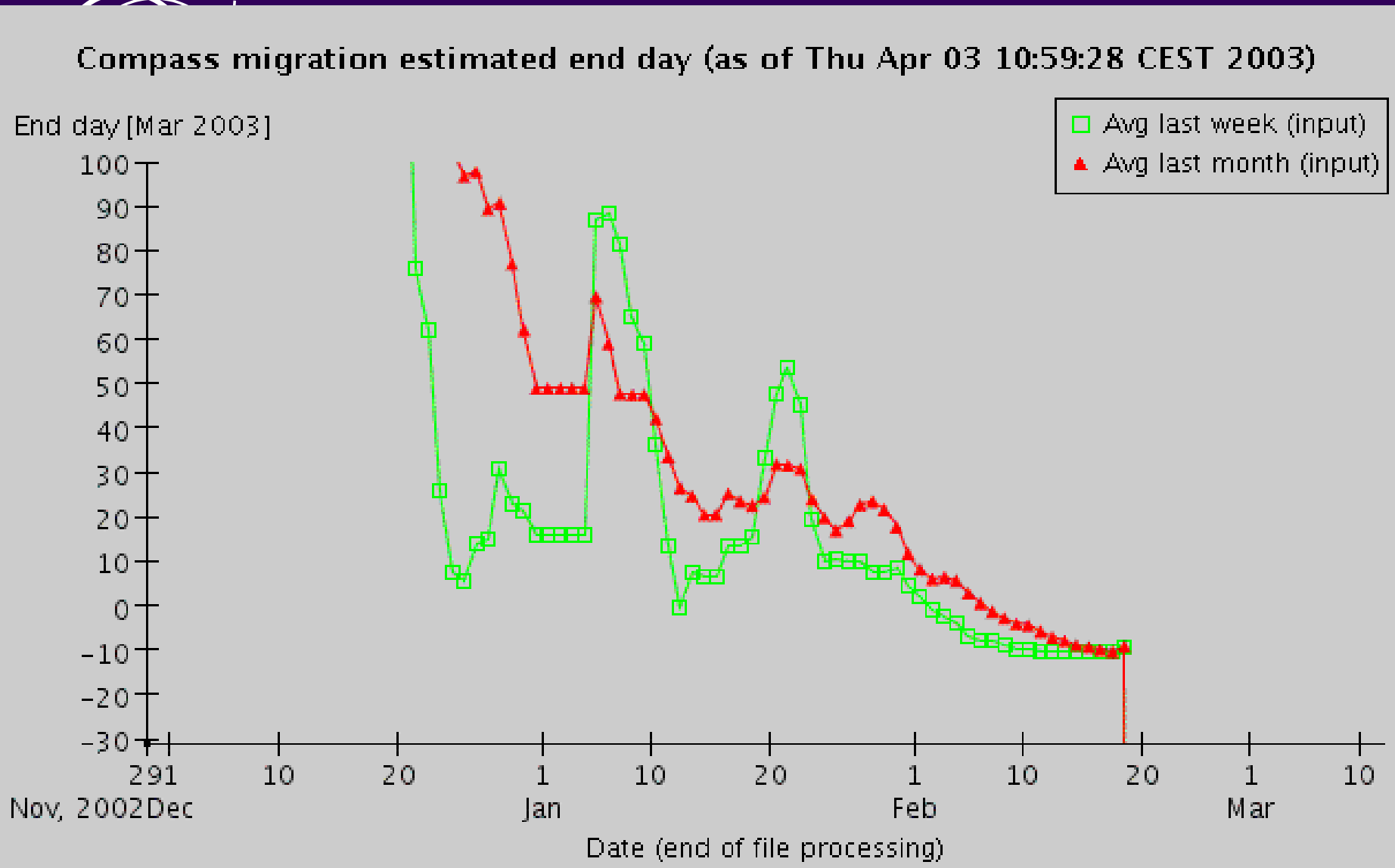
Project Timeline

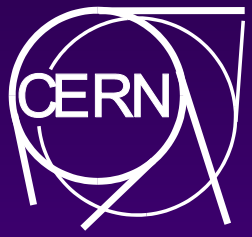
- Summer 2002 – initial ideas about migration
  - Designs, testing various Oracle features
    - VLDB, R.A.C., partitions, Linux, C++ binding
- Fall 2002 – implementation
- Winter 2002 – testing and integration
- Mid December – migration start
- Christmas CERN closure (3 weeks) – running unattended/remote monitoring
- Early 2003 – achieved full planned speed of 100MB/s sustained
- 20th February 2003 – migration completed
  - Ahead of schedule
- Middle March 2003 – migrated data and databases system achieves production status


Compass migration running avg data rate (as of Thu Apr 03 11:01:53 CEST 2003)
Data rate by file [TB/day]
Avg last week (input)
Avg last month (input)
0
1
2
3
4
5
6
7
8
9
10
29
1
10
20
1
10
20
1
10
20
1
10
Nov, 2002
Dec
Jan
Feb
Mar
Date (end of file processing)


Compass migration running avg data rate (as of Mon Mar 24 10:43:26 CET 2003)
Data rate by file [TB/day]
Avg last day (input)
Avg last week (input)
Avg last month (input)
10
9
8
7
6
5
4
3
2
1
0
29
1
5
10
15
20
Jan, 2003
Feb
Date (end of file processing)

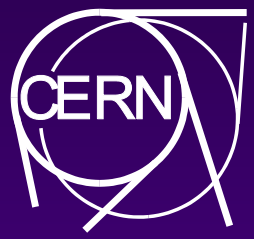

Compass migration data rate per real time (as of Thu Mar 27 10:39:11 CET 2003)
Data rate by file [Input MB/s]
110
100
90
80
70
60
50
40
30
20
10
0
Node 11
Node 10
Node 9
Node 8
Node 7
Node 6
Node 5
Node 4
Node 3
Node 2
Node 1
291
10
20
1
10
20
1
10
20
1
Nov, 2002
Dec
Jan
Feb
Mar
Date (end of file processing)


Compass migration estimated end day (as of Thu Apr 03 10:59:28 CEST 2003)
End day [Mar 2003]
Avg last week (input)
Avg last month (input)
100
90
80
70
60
50
40
30
20
10
0
-10
-20
-30
291
10
20
1
10
20
1
10
20
1
10
Nov, 2002
Dec
Jan
Feb
Mar
Date (end of file processing)



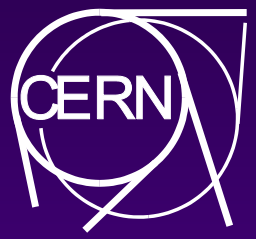
Experiences

- **The outcome**
  - **We successfully migrated all the data**
    - We migrated around 15% of all the stored data at CERN
    - Only 80 files failed due to tape media failure
  - **6.1 x 10^9 rows in Oracle (335GBytes)**
    - The validation/start of DST production was done using some 400 clients
  - **Some minor differences in the file names**
    - A few of the destination files was renamed after the migration
  - **The total amount of output data stored on tape is 220TByte (20% datareduction)**
  - **Peak throughput around 120MByte/s**
    - The system scaled in correspondence to the plans
    - Last month throughput averaged around 70MByte/s



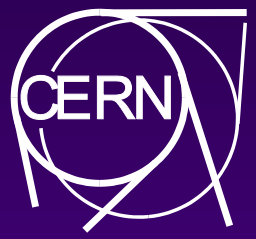
Experiences

- Converting a datamodel is not trivial, no matter how well documented it is.
  - Object model in
  - Relational model + flat file datastructure out
  - Minor misconceptions could have disasterous results
- Project management/Coordinating the resources is crucial for success
  - The system is complex
- Backing up metadata is also non trivial



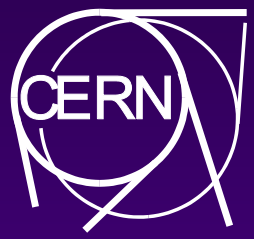
Future of COMPASS

- Expect to run around 12 Oracle database nodes
  - RAC/Non RAC?
  - Monitoring
  - Backup
- Expect around 1TByte of data in Oracle
- Expect around 1PByte of data in mass storage



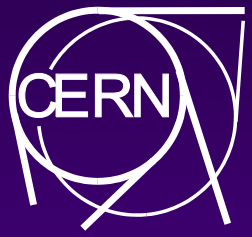
Questions?

- Thank you!



The hardware setup

- Off the shelf PC's
  - 2 Objectivity servers
  - 11 conversion nodes (dual cpu machines)
  - 11 output disk cache servers (4.5TByte total diskspace)
  - 3 Oracle servers (dual cpu, 1GB RAM, ½ TB disk)
  - 1 management node
  - 1 monitoring node
- Peripheral
  - GBit Network
  - 8 input tape drives
  - 10 output tape drives



Trivia/COMPASS detector

Data rates

- 200k electronic channels
- **22k ev/spill**
- **40kB event size**
- **220MB/s from detector**
- **1200 ev to tape =** 34.5 MB/s

Run 2002

- 5Gev events
- 260k files, 3000 tapes
- ~300 TB/yr

2003, 2004 and plans for >2006

- **~1PB on tape**
- **O(10TB) in Oracle**

Data processing

- **100 Intel PIII dual processor, running Linux**
- **700ms/ev**
- **5Gev × 700ms/ev = 3.5Gs**
- **200 CPUs**
  ⇒ 17.5Ms/CPU
  ⇒ 200 days on 200 CPUs
  @100% efficiency