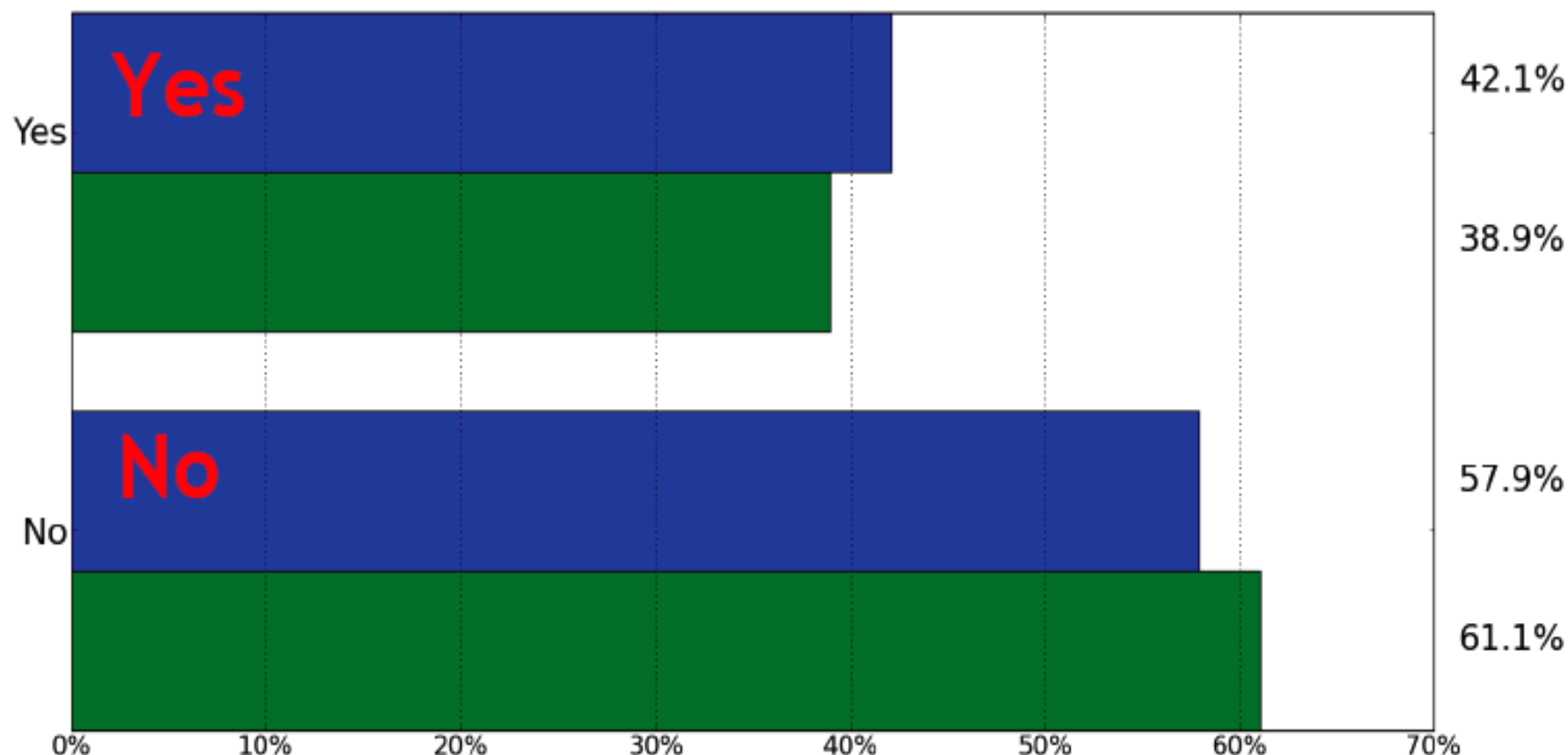# Data Storage and Analysis at
HERMES

Eduard Avetisyan (DESY)

Contents:
- HERMES data model
- Current status
- Perspectives, expectations, preferences
- Outlook

Did anything go wrong so far?

Do you think that in the past important HEP data have been lost ?
(top/blue: theorists, bottom/green: experimentalists)

- **Yes** — 42.1% (blue), 38.9% (green)
- **No** — 57.9% (blue), 61.1% (green)

Over optimistic? Over pessimistic?

Introduction to HERMES

- **HERA lepton beam (long. polarised)**
- **Fixed gas target (long./trans. polarisation)**
- **Variable target density: O(100)**
  - **variable trigger rate (10-300Hz)**
- **Restricted (not 4pi) acceptance**
- **Clean trigger (no need for 2nd level)**

Beam charge	Electron, positron
Beam helicity	Positive, negative
Polarised targets	Hydrogen (Long., Trans.) Deuterium (Long.) Helium3 (Long.)
Unpolarised targets	H, D, N, Ne, Ar, Kr, Xe
Physics topics	Inclusive, semiinclusive, exclusive, nuclear, spectroscopy ...

Analysis Structure

- Major working groups: 5
- Active analysis: 50
- Data productions sharing: 100%
- MC productions sharing: 30-50%
- Common Fortran, C and C++ based analysis frameworks, individual analyses codes
- Software releases in: first 5 years   19
                                            last 8 years   5
- Results published only if a per-mille agreement within two independent analysis is achieved
- Data/MC productions use common chains
- Recent attempts: (successfully) using GRID for MC productions   ~2 billion events in 2 months (~10TB)
- Recent: active ROOT-based analysis

Data Production Model

- Data split to runs (~5 min. , ~500MB raw EPIO)
- First tracking uses primary (test beam) calibrations
- Offline calibration performed using reconstructed tracks
- uDST (ADAMO) used for PID parent distr., then second iteration for analysis
- Rec. data and uDST accessed via NFS
- Homegrown DAD (Distributed ADAMO Database) for calibration/alignment servers
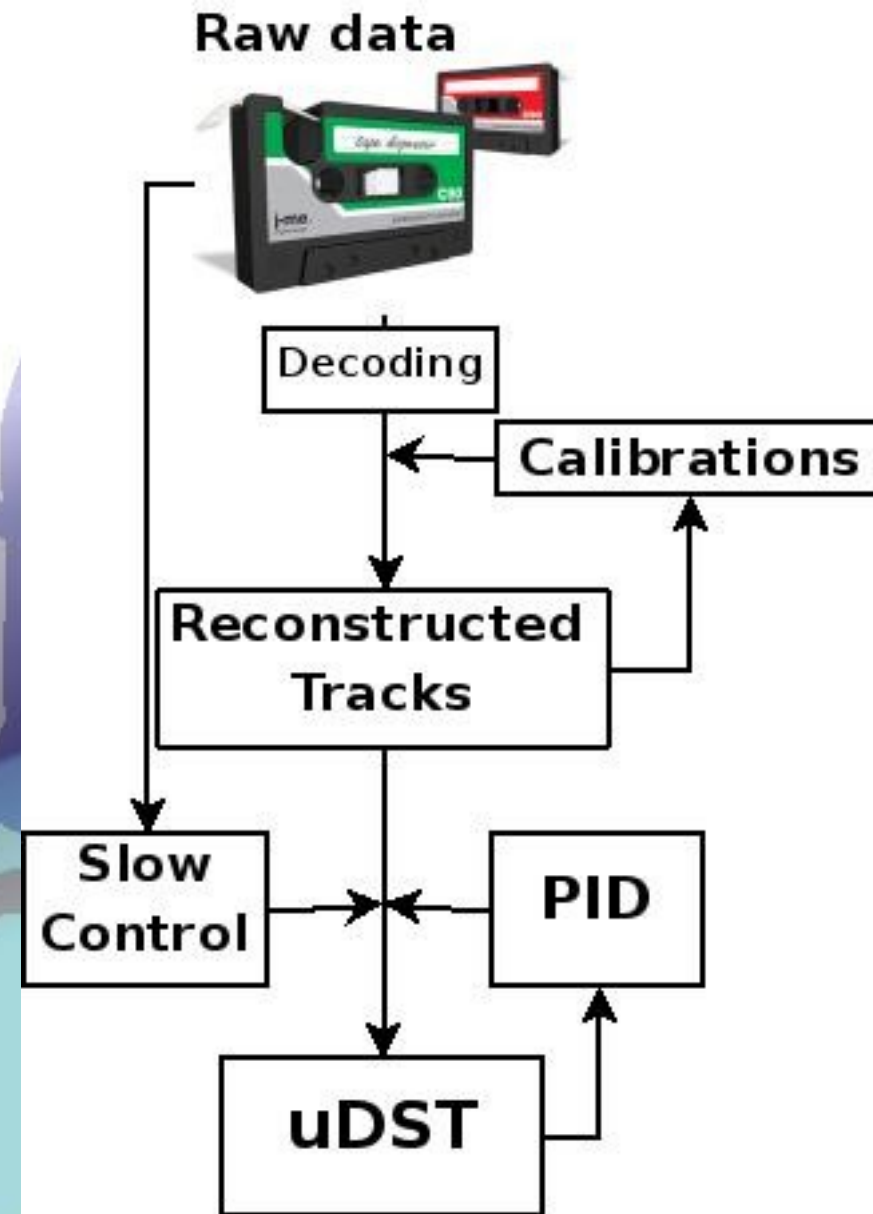
Monte Carlo Structure

- Dedicated physics generators
- Fortran+Geant3 digitization
- Productions on-demand
- uDST only stored
- optional fast smearing generator
- ADAMO database format
- Same tracking code as for data
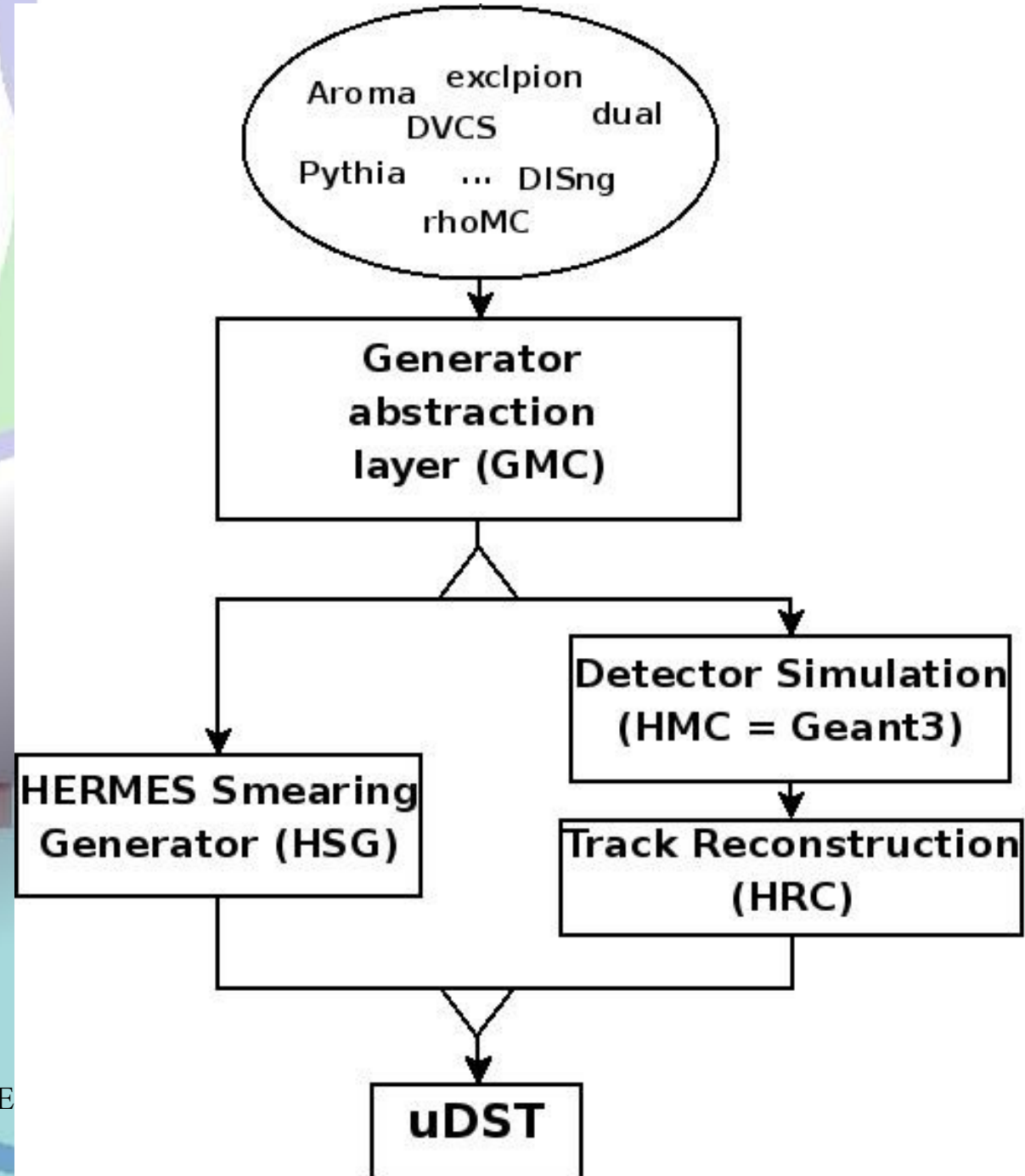
Data Storage and Access

(real) Data:
- **Raw:**   **110TB**   **Tape**
- **Main (trk).:**   **45TB**   **Disc + Tape (backup)**
- **uDST:**   **12TB**   **Disc + Tape (backup)**

up to 5 main and up to 8 uDST prod. per year of data taking (1995-2007)!

Monte Carlo
- **uDST:**   **15TB**   **Tape**
                    5TB   **Disc**

Analysis performed on uDST level, all data and some MC available on NFS file servers (~80TB total), also used as shared user and group space (ntuples etc)

Prospects

Expected to preserve on tape:

- all raw data
- 1-2 best (newest) uDST and main productions per year (~1/4 of the currently used storage)
- all analysis software
- Some group/user data

Software perspectives:

- improved/corrected calibrations, alignment and tracking algorythms for future productions
- Need possibility for major reproduction of all collected data

Possible Scenarios Considered

Transition	Virtualization	As-Is (run static binaries)
Reliable	(Slightly) lower performance	Very low maintenance
Very manpower-intensive (full production chain)	Potentially problematic data access	Not Reliable
Reasonably manpower-intensive (if only ROOT trees preserved)	Version compatibility	
	Low maintenance (freeze once, use forever)	
	Possible coupling with "As-Is" solution	
	Combined effort	

Outlook

- Experimental data highly valuable – preservation needed!
- Multiple scenarios exist – need to explore further
- Current choice: virtualization
  - Explore the effort
  - Data storage and access
  - VM type (Vmware, VirtualBox, KVM, Xen, ...?)
  - Possibility to run in GRID?