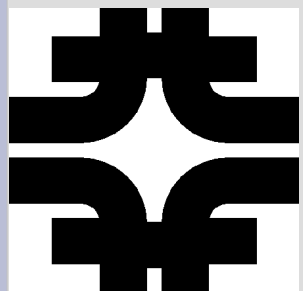
# Automated recovery of data-intensive jobs in D0 and CDF using SAM

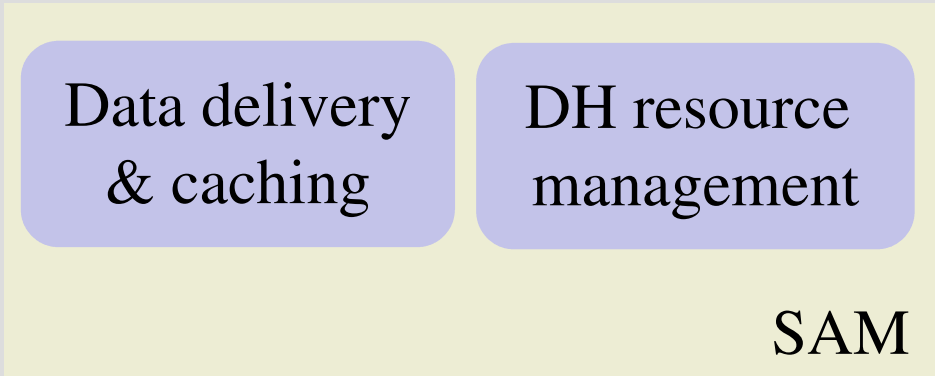
A. Baranovski[1] (primary author)
V. Bartsch[1,2] (presenter)
E. Lipeles[3], A. Lyon[1], I. Sfiligoi[1],
D. Benjamin[4], K. Genser[1]

[1] Fermi National Accelerator Laboratory,
[2]University College London,
[3]University of California, [4]Duke University



# SAM
# data handling system

- allows distributed data handling for CDF, D0 and Minos
- creates basis for easy GRID extensions (job handling, authentication, monitoring, brokering)
- GRID extension diverse for CDF and D0



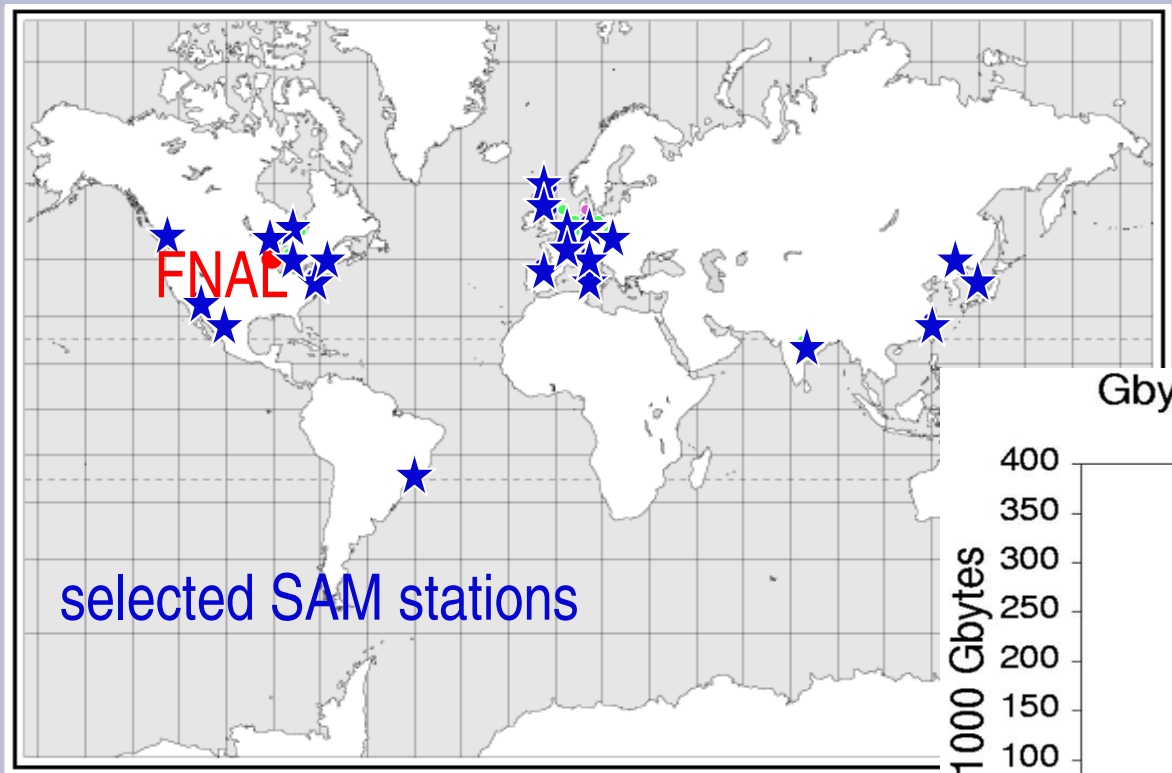
# SAM
# data handling system

requirements:

- scalability to transfer data
- robust transfers with different types hardware over long distances
- monitoring and book keeping
- efficient use of remote resources
- easy to handle system

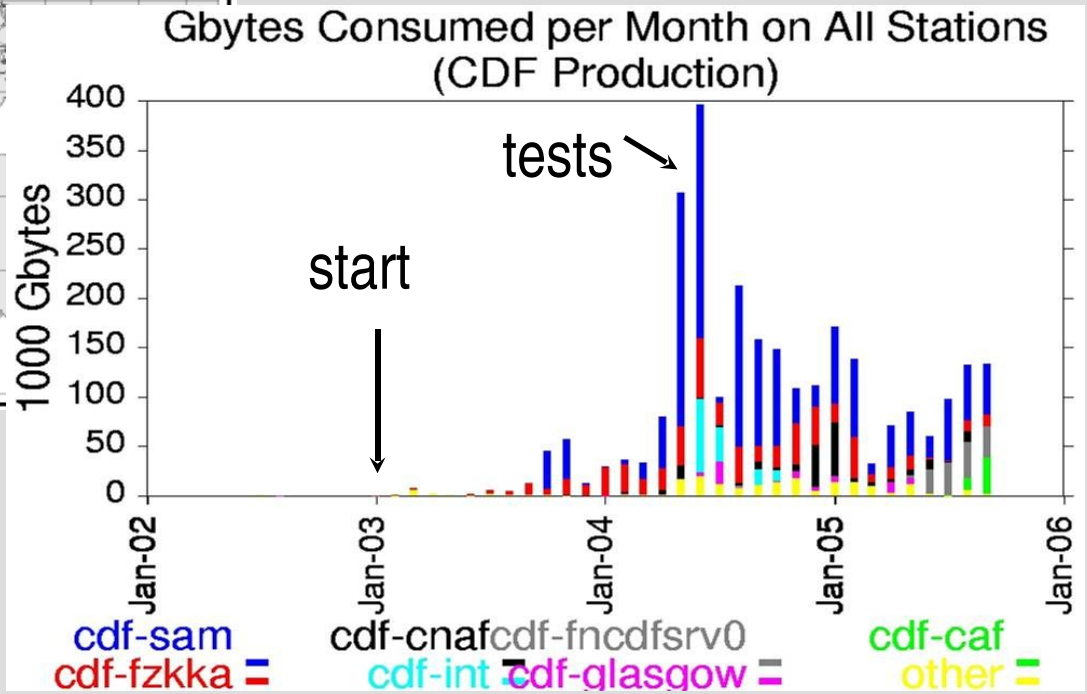
=> solution for Fermilab experiments: SAM

# SAM distributed data handling



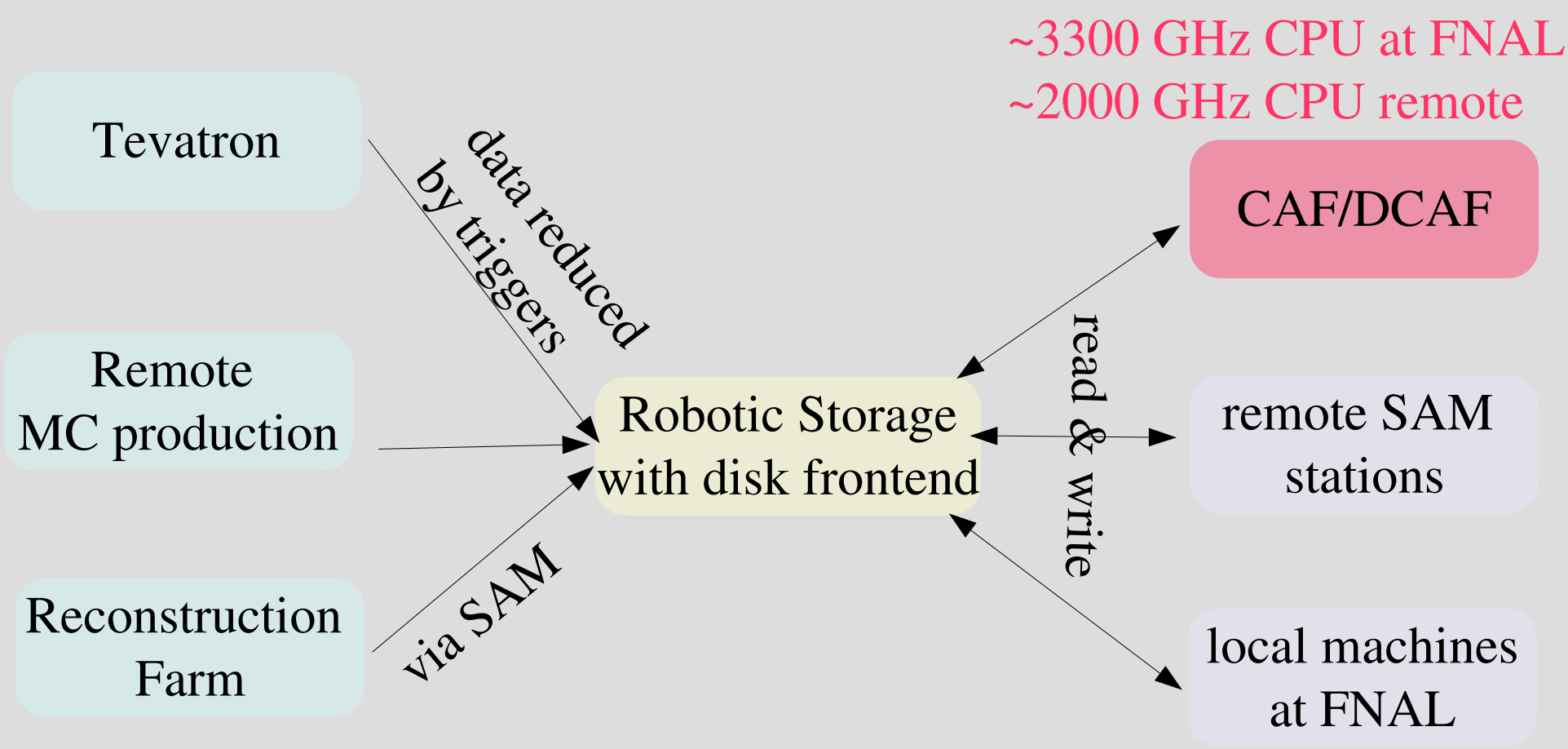

DZero and CDF:

- 10k/20k Files declared/day
- 15k Files consumed/day
- 8 TByte of Files cons./day

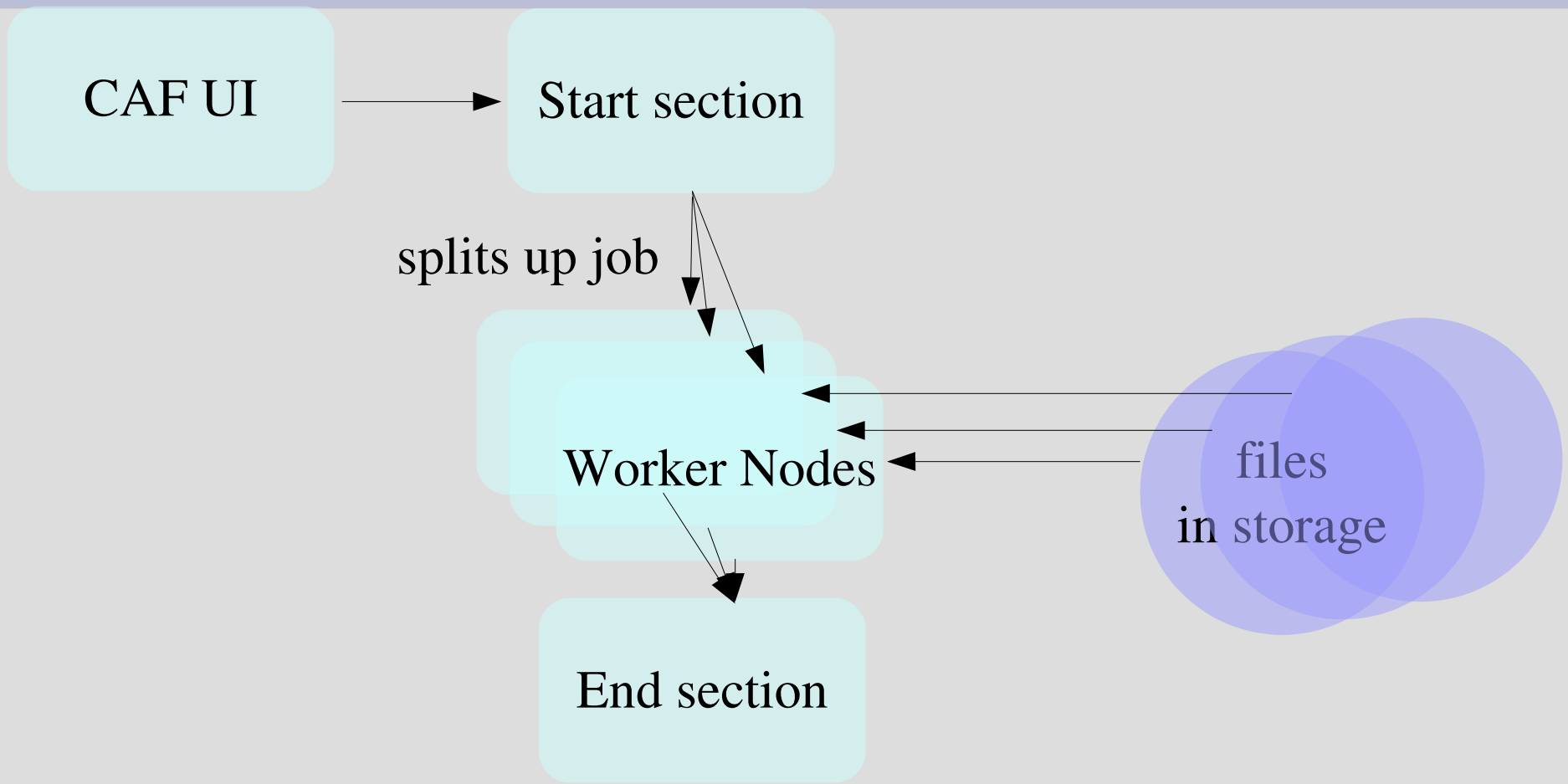



- main consumption of data still central
- remote use on the rise

# CDF DAQ/Analysis Flow

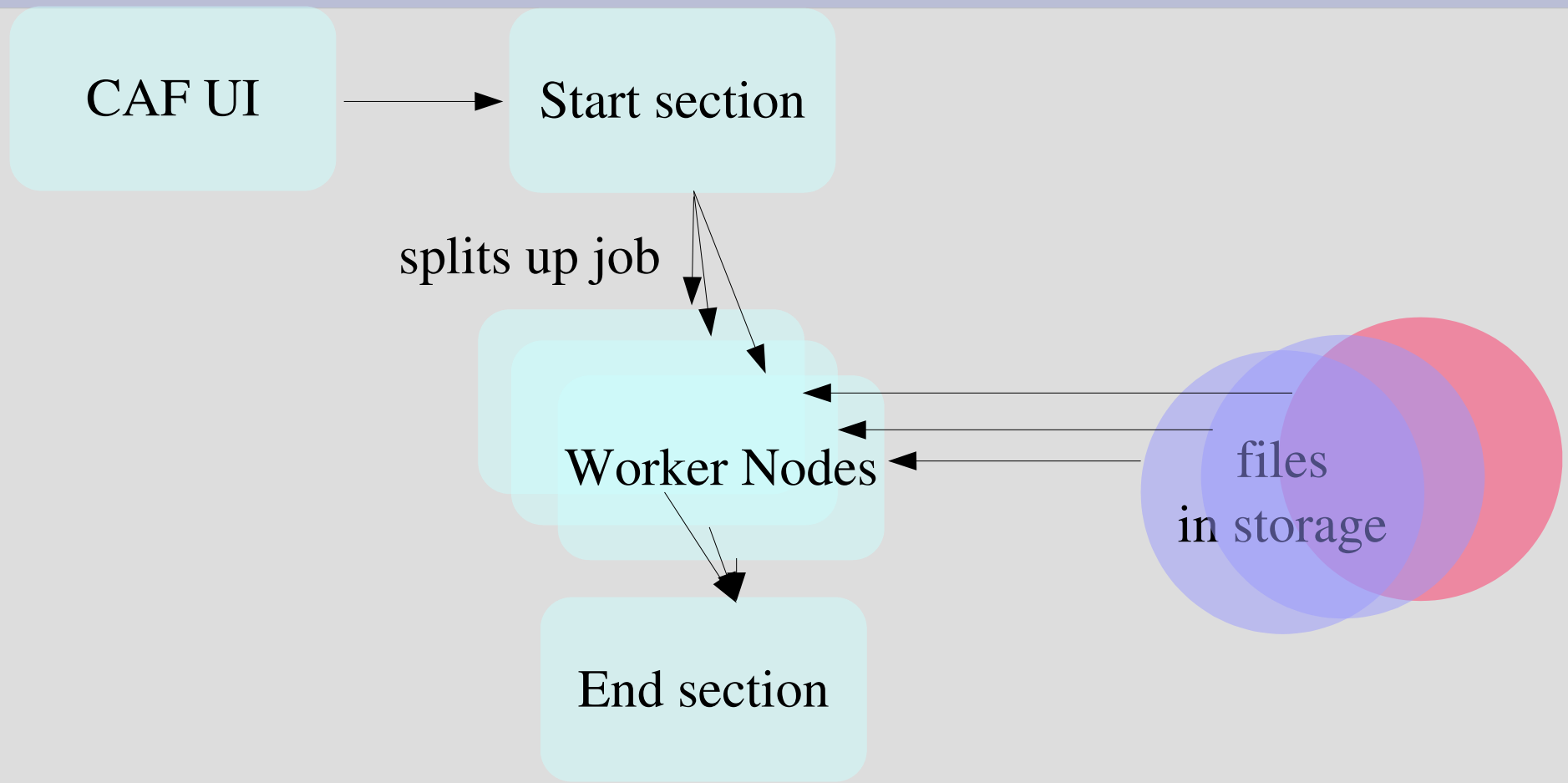


# (D)CAF: job handling

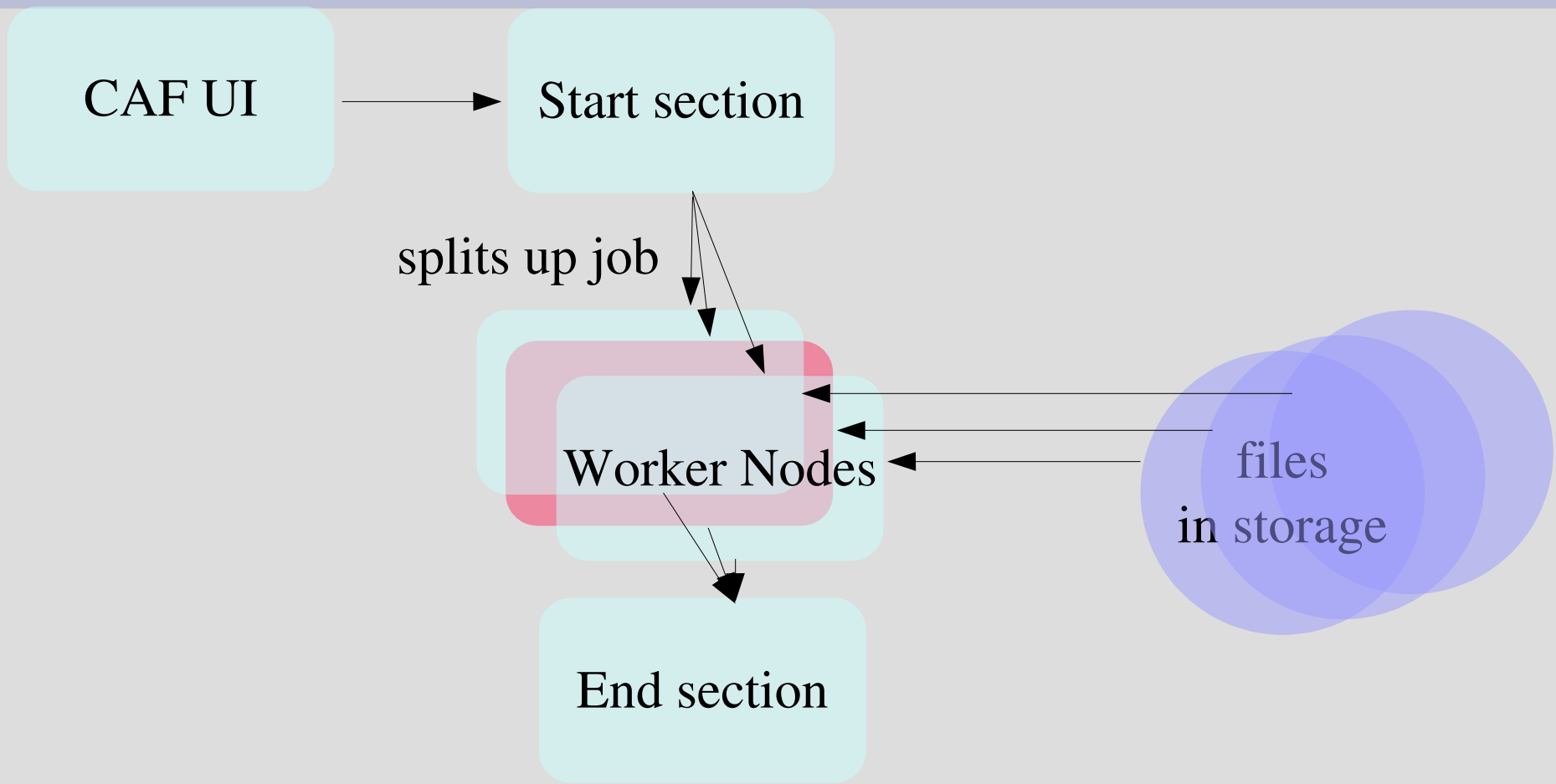



a typical analysis job consumes terabytes of information during several days of running at a job execution site

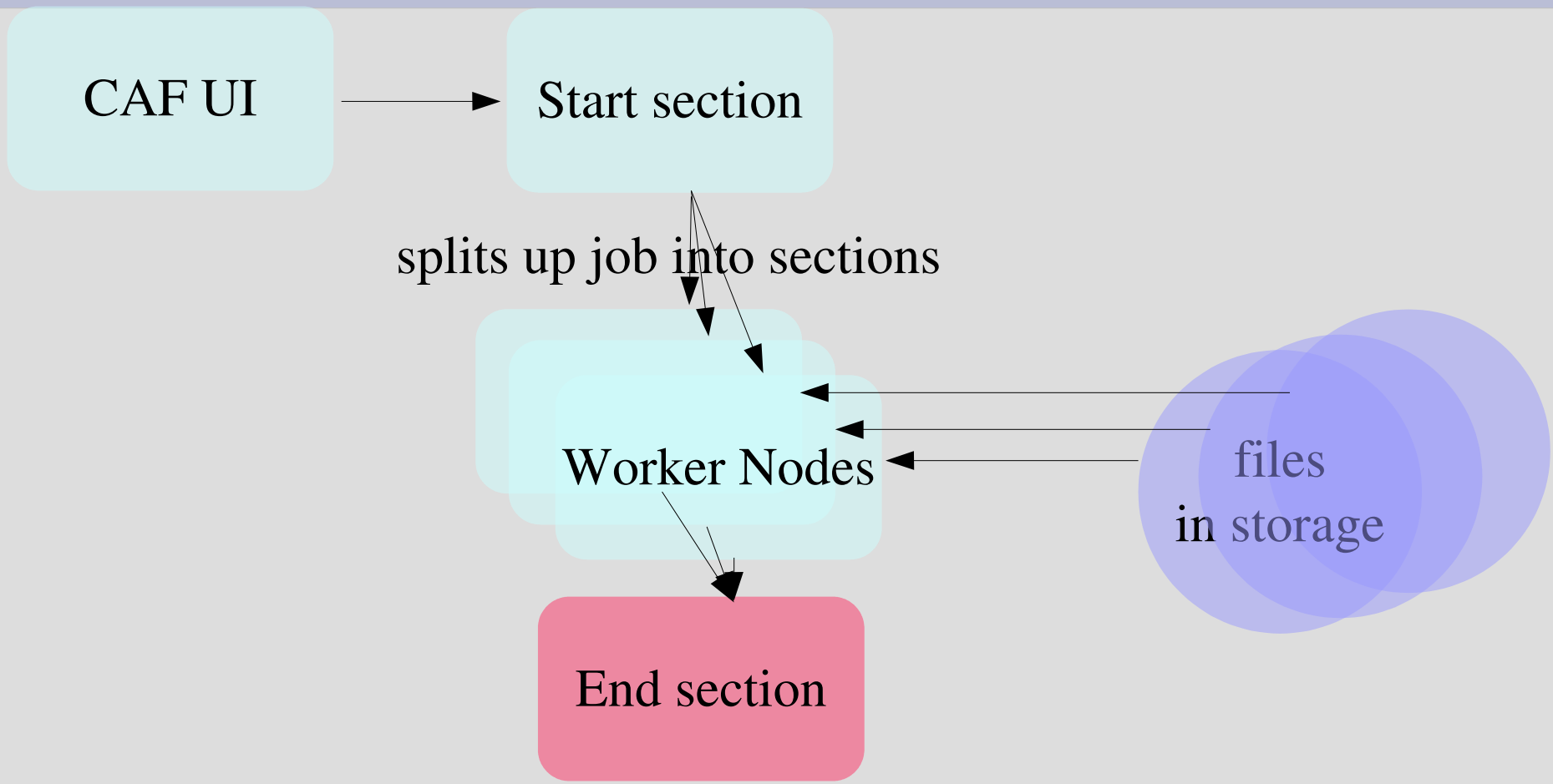
# Error cases: file delivery



# Error cases: section restart

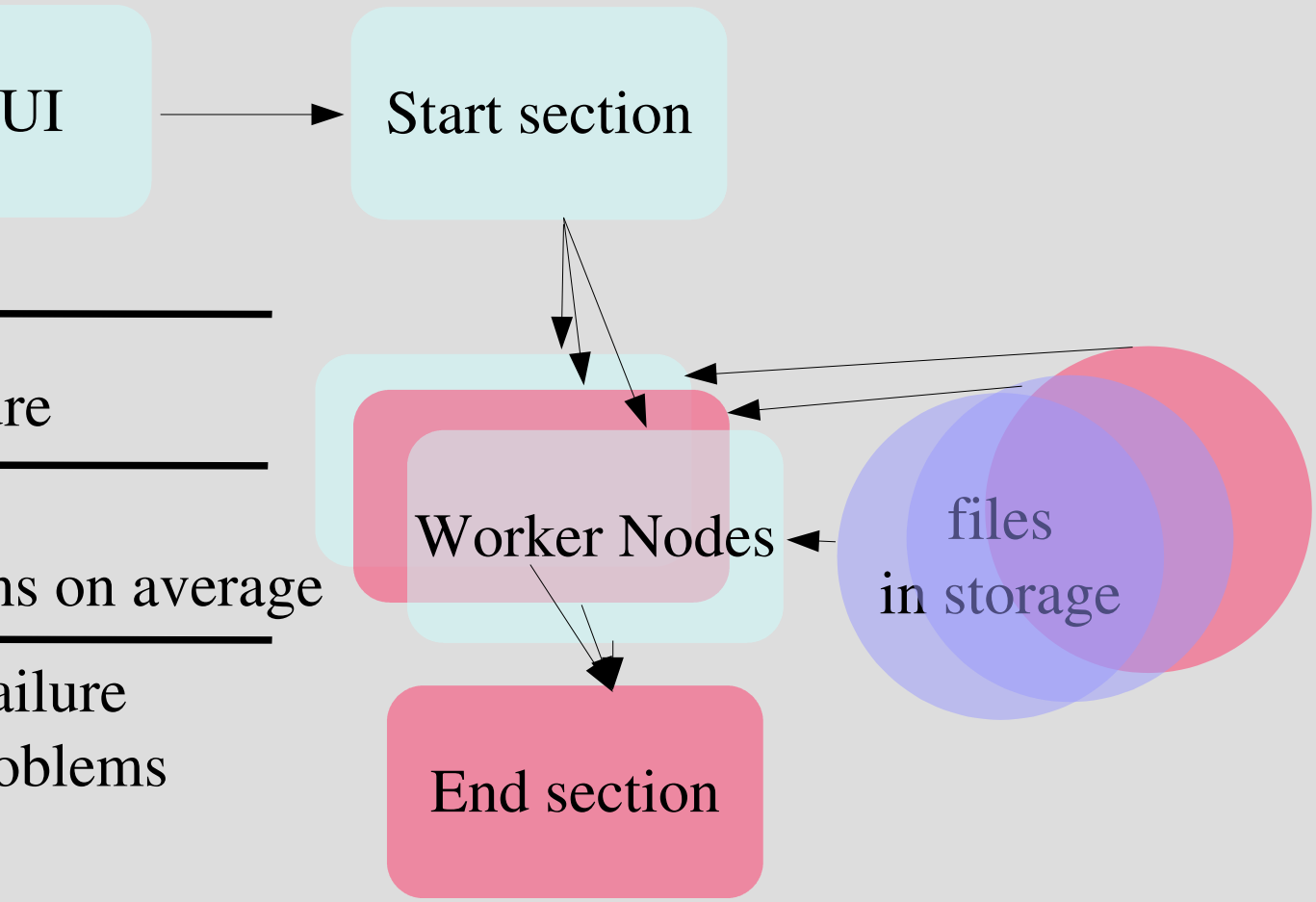


# Error cases: output retrieval



# Error cases: statistics

| failure rates | causes |
|---|---|
| 3% | file retrieval failure |
| 0.4% | segments restarts<br>job has 70 sections on average |
| 3% | output retrieval failure<br>including user problems |




status before SAM: analysis of the output files of the job
=> time consuming and depending on the output format

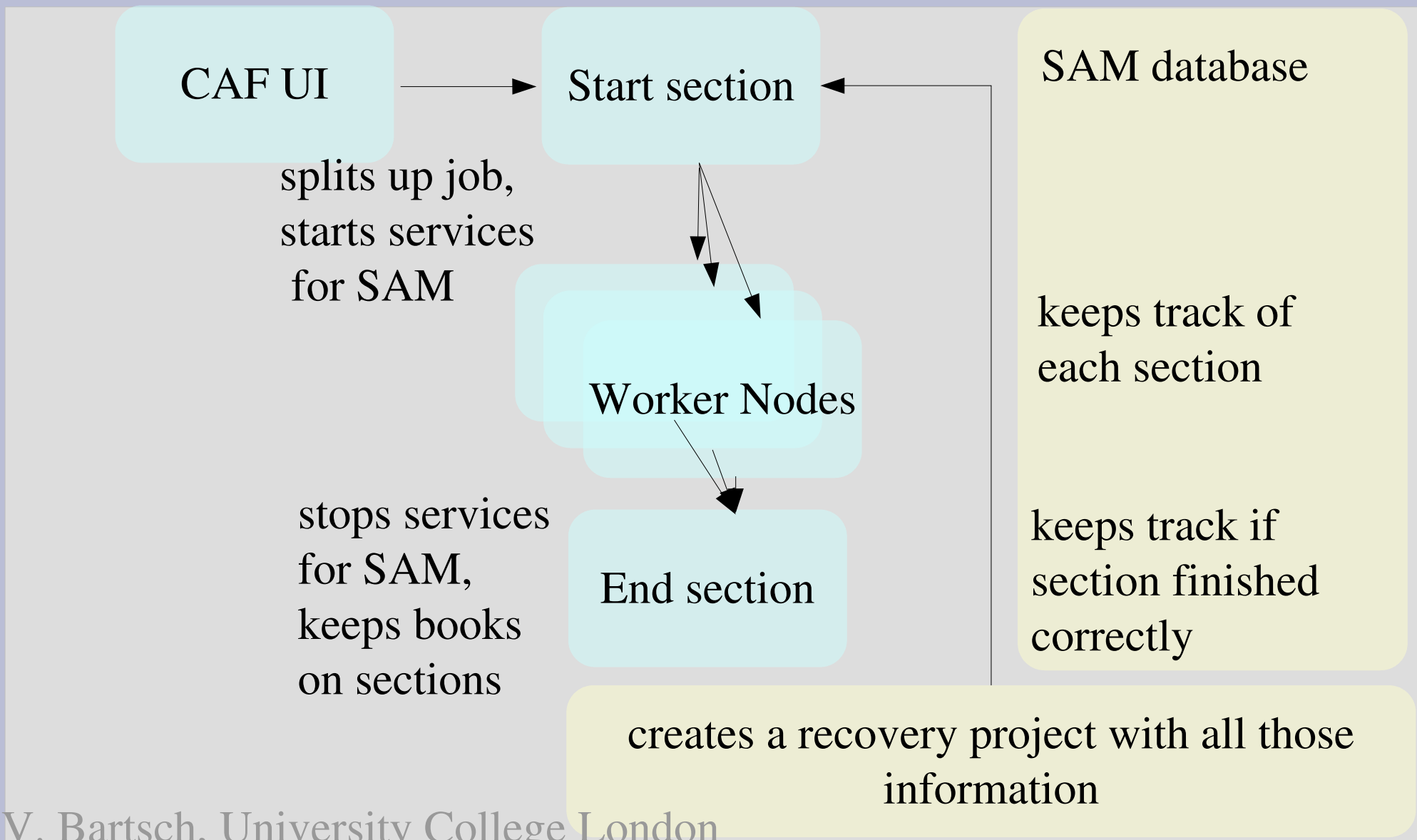
# Error handling approach

Failure at any point of the CAF job:
=> time consuming parsing of the output files
=> whole file list, dataset, being reprocessed or lot of fixes by hand.

Approach:

- job is monitored as a SAM project in the database
  => monitoring of file delivery errors
- interface to experiment specific software (CAF) to decide on success of a job / sections
  => monitoring of section failures, output storage / certification failures, etc

# Automatic recovery



# status of the automatic recovery

- file delivery problems are already caught
- interfaces to the CAF built and currently under test
- decision about the level of optimization  not yet made
- CAF deployment can be handled fast after the testing phase

# Automatic recovery: conclusion

- SAM takes care of the bookkeeping and allows to create recovery projects
- interfaces between SAM and experiment specific software, e.g. CAF, can be built in order to incorporate the knowledge of the software into the recovery

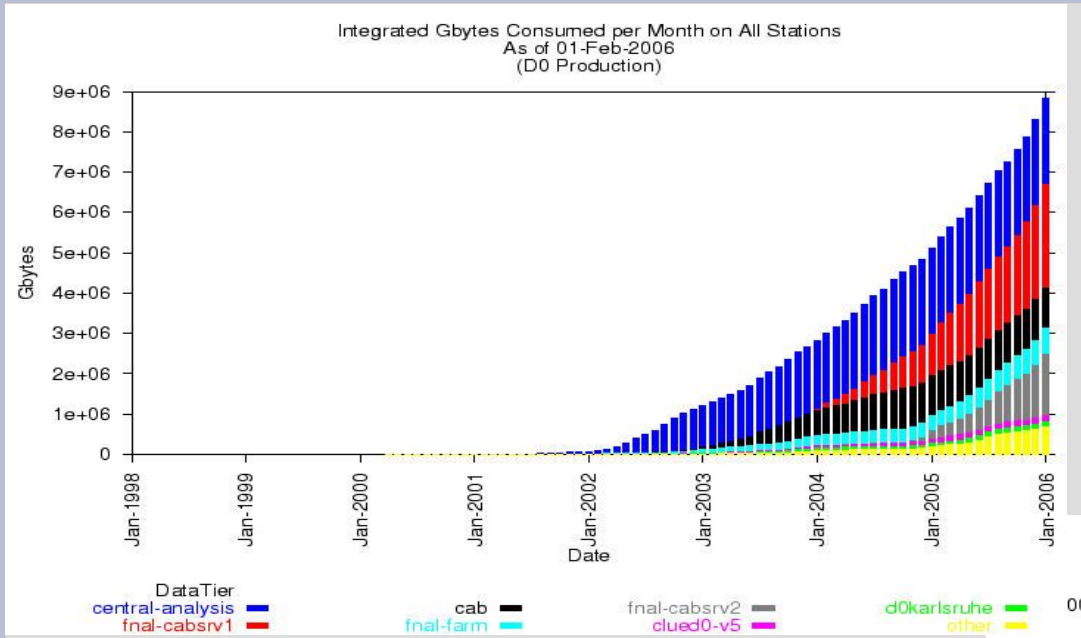
=> error handling for the users much more transparent and effective

# other SAM related contributions

- Interface between the data handling system (SAM) and CDF experiment software
- Experiences with operating SamGrid at the GermanGrid centre for CDF
- The SAM-Grid / LCG interoperability system: a bridge between two Grids
- Lightweight deployment of the SAM grid data handling system to new experiments
- SAMGrid Peer-to-Peer Information Service

no claim on completenes

# backup slide




D0 usage of SAM