

Distributed Computing Concepts in DØ



Daniel Wicke
(Bergische Universität Wuppertal)



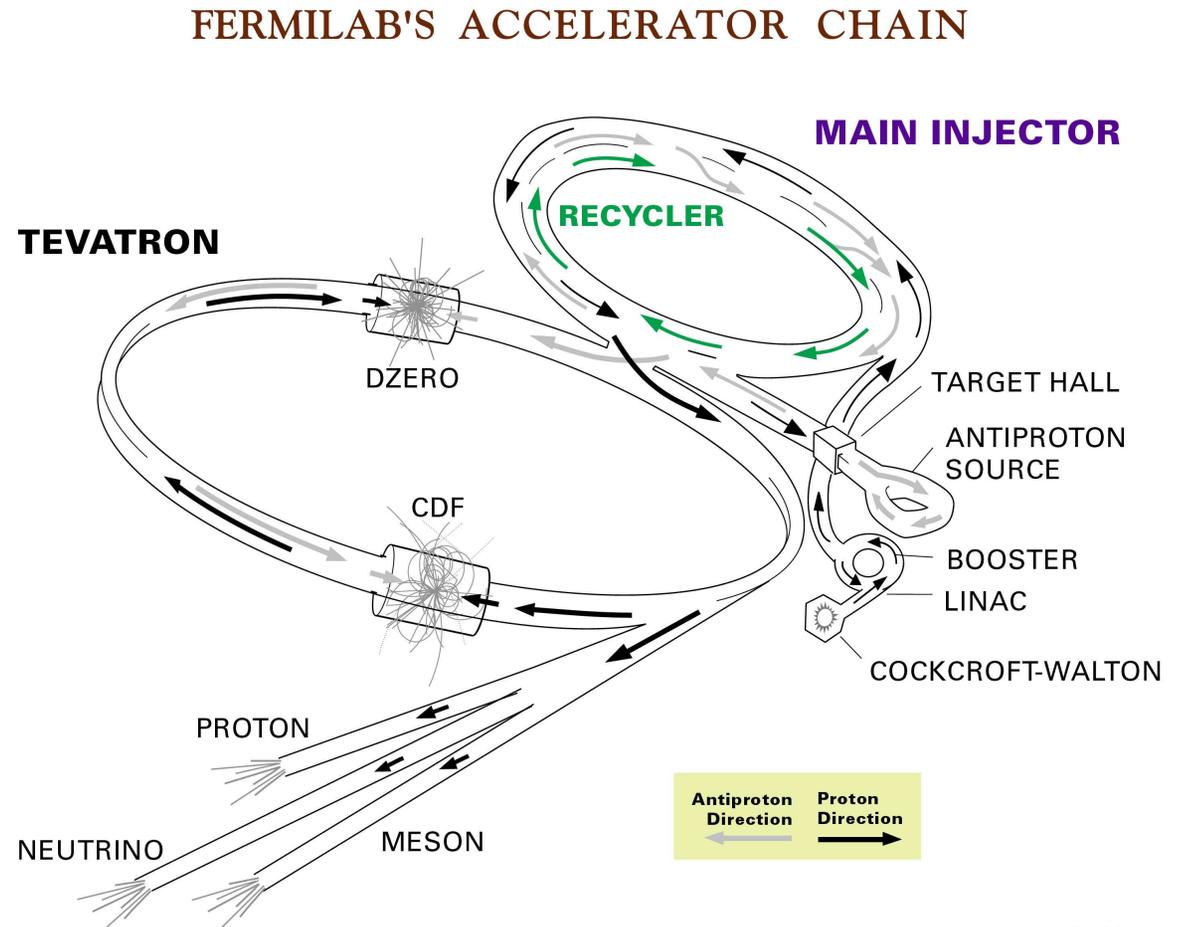
Outline

- Introduction: Computing Demands
- Computing Concepts
(Workflow Management, Local Resource Optimisation,
Distribution of Data and Jobs, GRID)
- Summary

Introduction

The $p\bar{p}$ Accelerator Tevatron

- Circumfence 7 km.
- Run I (1987-1995)
- Run II (since 2001)
- $p\bar{p}$ collisions
- 2 experiments, CDF and DØ, record events.



Fermilab 00-635

Physics: Probing the Several 100 GeV Range

Example: Top and Higgs

Top-signature: $2b$ -jets + at least 4 jets
 $2b$ -jets + missing E_t + lepton(s)

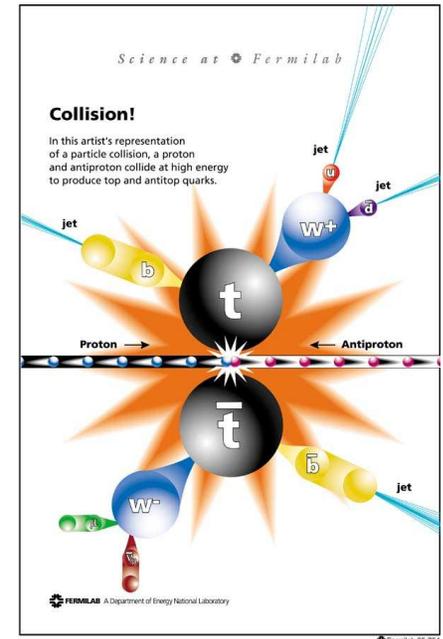
Higgs-signature: $2b$ -jets
 $2W$ Bosons

Background: QCD

- Production of light quarks ($udscb$) results in 2 or more jets.
- Gluon radiation may add more jets.
- These events can't be fully suppressed in realtime.

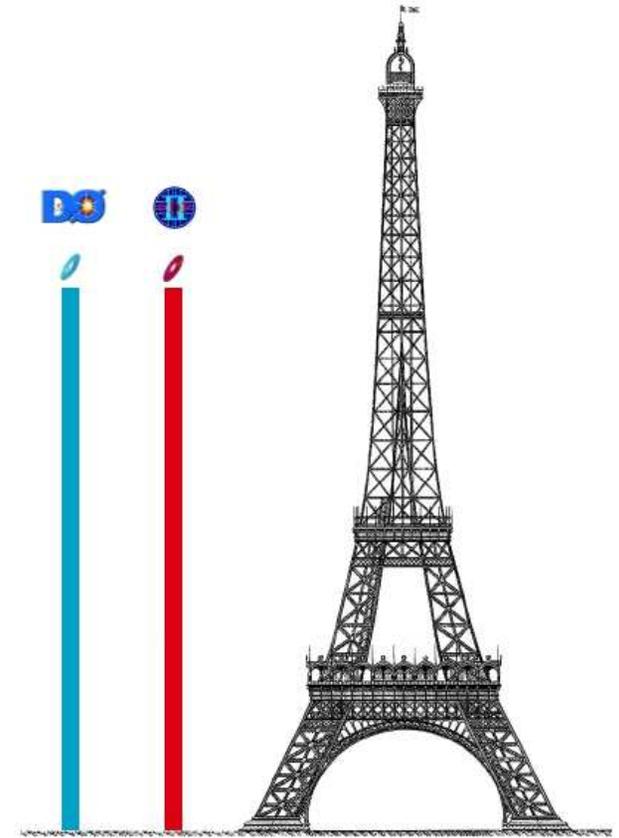
$$\sigma_{\text{tot}}^{p\bar{p}} \simeq 70\text{mb}, \quad \sigma_{t\bar{t}}^{p\bar{p}} \simeq 6.6\text{pb}, \quad \sigma_H^{p\bar{p}} \simeq 0.7\text{pb} \quad (\text{for } M_H = 116 \text{ GeV})$$

$$\sigma_{t\bar{t}}^{p\bar{p}} / \sigma_{\text{tot}}^{p\bar{p}} \simeq 10^{-9}$$



Computing Demands

- Top and Higgs are difficult to recognise in realtime.
- To find sufficiently many top- and Higgs-events a large number of reactions must be recorded.
- CDF and DØ record 500GB/day each.
- Another 1100GB/day come from processing the raw data.
- This sums up to 200m of CDs per year.



Providing facilities to analyse these data is a big challenge.

Estimated CPU need: 1.8MSpecInt2000 (4000 1GHz PIII Computers).

Management of Large Batches of Jobs

Problem 1

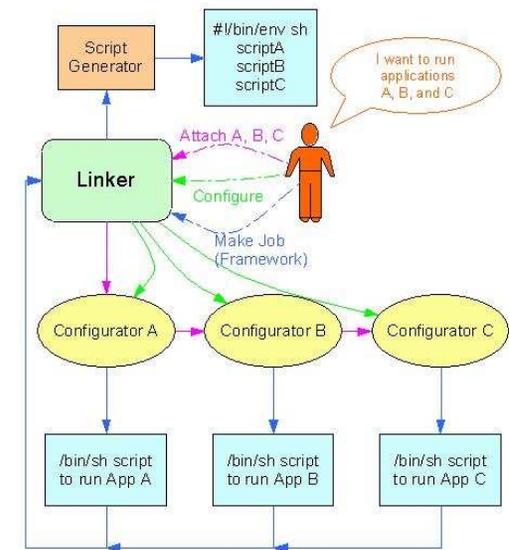
Handling MC production chains with ever changing programm versions/parameters is very person power intensive.

Base of improvements

Automate linking for several programs into a single job based on a job description (configurator).

Automatic job management: Runjob

- handles chains of executables
- passes output of a programm as input to a following
- track metadata
- separates calling details from organisation
- allows for automatic parallelisation of jobs



Optimised Use of Local Resources

Problem 2

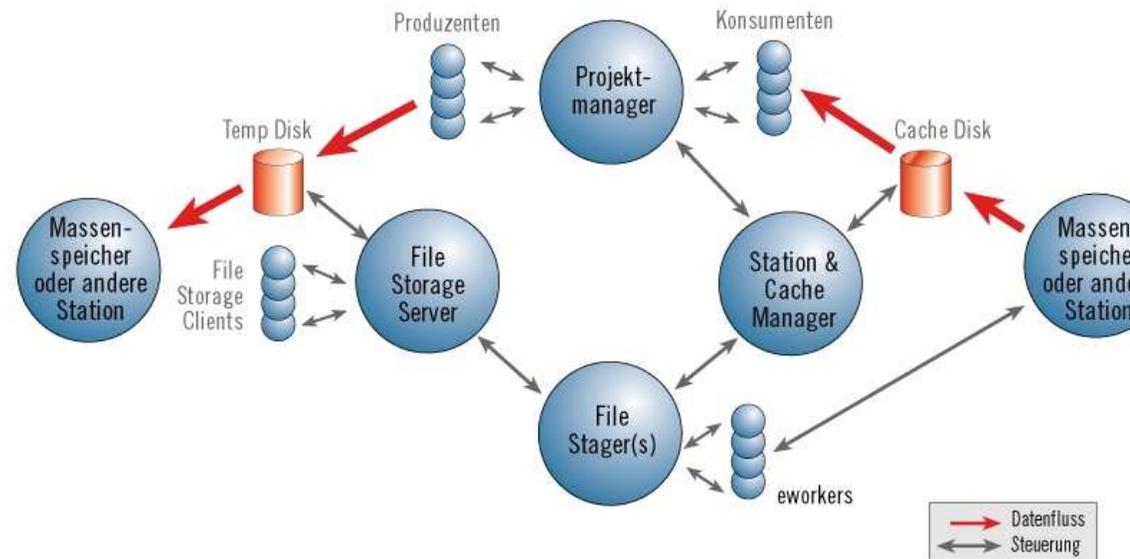
It is impossible to store all data on disks:
Tape access is the major bottleneck.

Base of improvements

The order of events in the dataset
has no meaning.

Optimisation

- Don't loop through file lists.
- Request **datasets**.
- The order in which files corresponding to a dataset are processed may change.
- The system optimises the order to minimise tape access and tape mounts.



Sequential Access through Metadata: SAM

Worldwide Distribution of Data

Problem 3

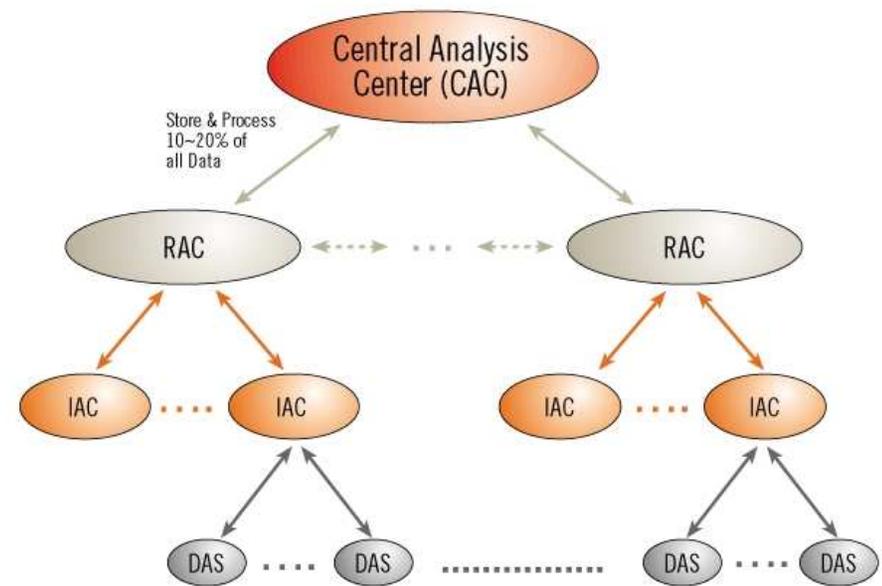
For the most frequently used data the I/O rate to disks limits performance.

Base of improvements

Setup copies of these most frequently used data.

Regional Analysis Centers (RACs)

- Allow full physics analysis.
 - ⇒ Hold all Thumbnails.
 - ⇒ Provide computing power to process these.
- Allow distributed reprocessing.
 - ⇒ Hold 10% of full DST.
- Serve institutes.
- As such a major building block for the DØ Grid.



RAC Prototype

Goals

- Proof of principle for RAC concept.
Provide a working analysis environment for DØ-Germany.
- Check needed resources and its scalability (mostly person power and network).
- Check DØ-software compatibility.

Specifications

The prototype should implement the following major features of a full RAC:

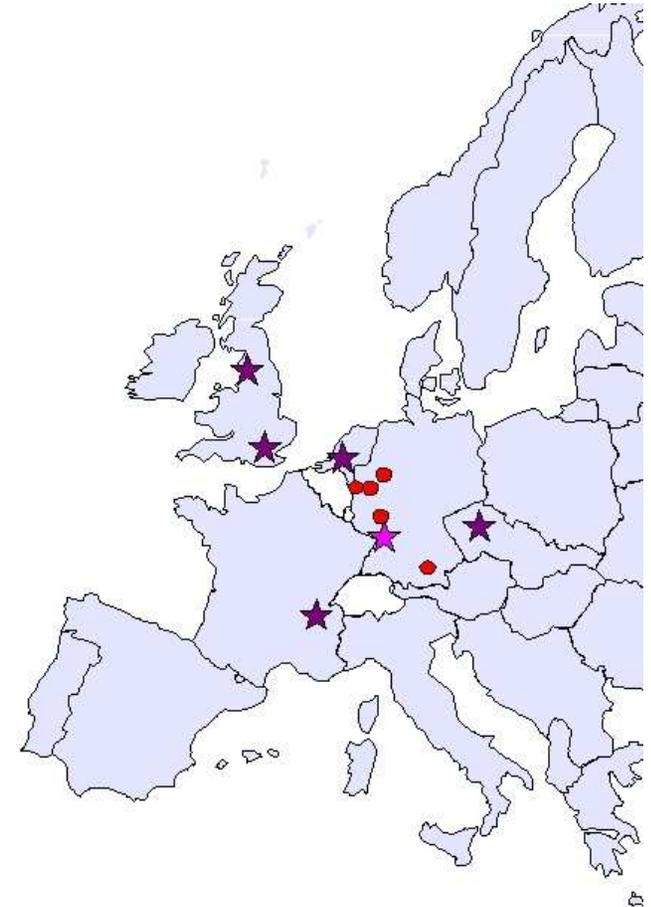
- Continuously and immediate transport of thumbnails from Fermilab to GridKa disks. Working
- Fetching files available at the prototype from associated institutes. Working
- Automatic installation of DØ-software updates. Being tested

Prototype Location



Grid Computing Centre Karlsruhe: GridKa

- located at Forschungszentrum Karlsruhe (FZK).
- established in 2002.
- centre for Grid development.
- regional data and computing centre.
- 8 HEP experiments:
Alice, Atlas, Babar, CDF, CMS,
Compass, DØ and LHCb.



Resources

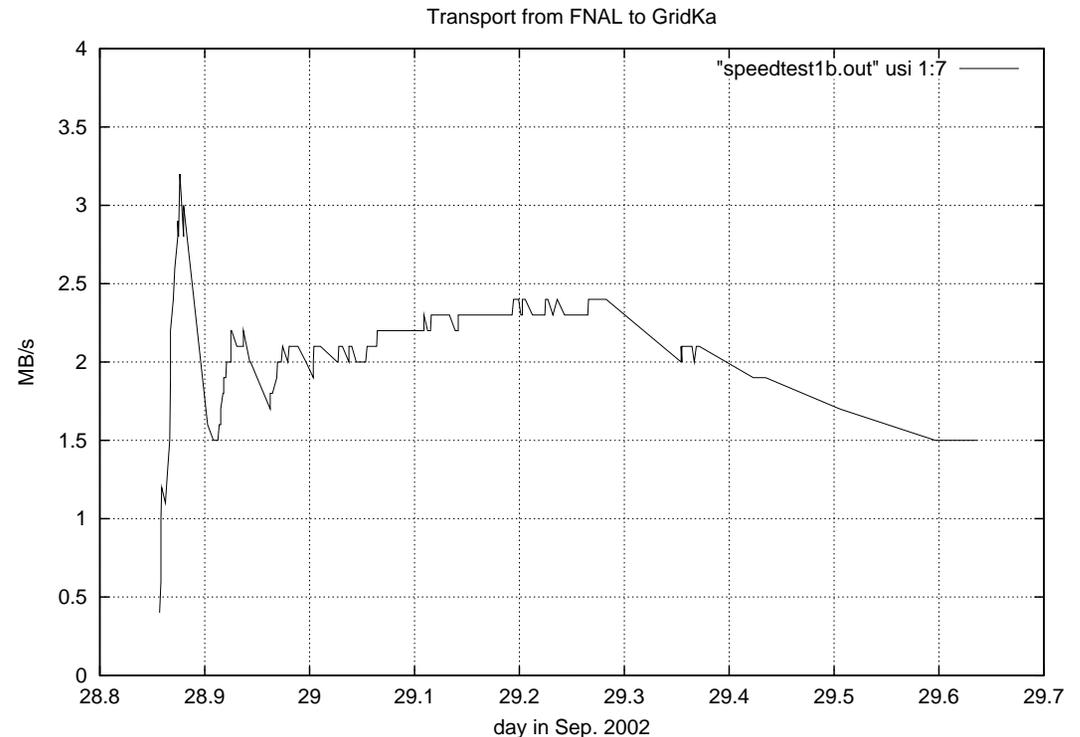
160 dual CPU nodes, 44 TB disks, > 100 TB tape

Results: Transfer Speed

Integrated size of arriving files over time:

2–3MB/s = 15–25MBit/s
(averaged over 7 hours).

larger gaps in the transport decreased effective speed thereafter.



⇒ (At that time) limited by FZK connection (32MBit/s)
Currently upto 3.3MB/s (or 8TB/month) is observed.

Internet bandwidth to FNAL is sufficient for our current demand.

Software Problems Lead to Person Power Problems

DØ software originally written for the FNAL computing environment and is still tested in this environment

Deploying this software at remote sites

- requires site specific adaptations
(hardware setup can't be changed everywhere, needs repetition for each new version)
- requires site specific configurations
(sometimes quite fancy and possibly fragile)
- introduces correlations between **all** sites.

Person power becomes a problem when running many systems

Example: CAB induced transport problems

The launch of a new linux cluster (CAB) at FNAL induced unexpected problems at GridKa:

Symptom

- Large number of transfer errors/undelivered files.

Causes

- GridKa sam-station tried to get files directly from CAB.
- This failed because the CAB-nodes aren't known to our firewall.

Solutions

- Opening the firewall isn't feasible.
- Route files from CAB through central-router (d0mino).

Person power required to run the systems needs to be reduced!

⇒ We should aim to remove/avoid dependencies between sites:
Think globally, avoid site specific paths, code and libraries.

The GRID

Problem 4

Network, CPU and Diskspace availability will be constantly changing. Communication lines need to be configured explicitly.

Base of improvements

Dynamic adaption to actual situation.

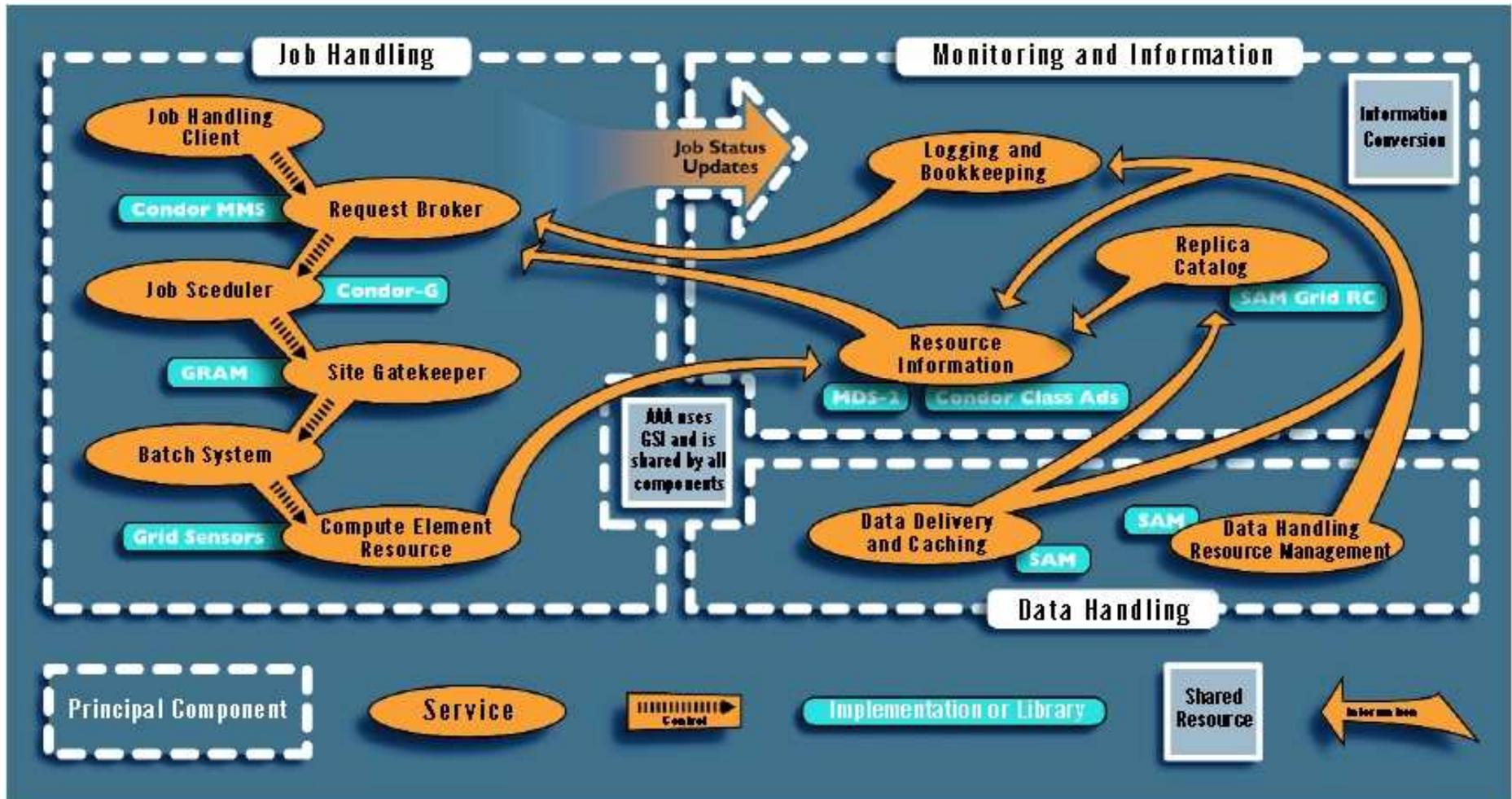
Optimised use of globally distributed resources: The GRID

- Retrieve data from remote disks with best availability.
- Submit jobs to centres for which they're best suited.
- Base these decisions on **current** status of the systems.

Requires

- Common protocols for data and status information exchange.
- The GRID

SAM Grid



Common CDF-DØ effort to finalise this for Run IIb (2005).

Summary

- DØ computing strategy is driven by physics needs.
- Computing demands are met by using
 - automated workflow and metadata management (runjob)
 - local optimisation of data access (SAM)
 - global distribution of data (Regional Analysis Centres)
- Moving towards GRID technologies we hope to
 - reduce the person power needed to run the clusters
 - add global resource optimisation

Outlook

- Perform distributed data rereconstruction.
- Test GRID requirements in a running experiment.
- Evolve towards full GRID standards.