Remote and GRID computing at D0 and CDF

P.Mättig,

University of Wuppertal

Disclaimer

Not a technical talk!

My objective: to get the most physics out of 10^9 events (current: D0 and in 3 years ATLAS)

➔ a lot of data handling and CPU required!

GRID is needed! But for physics it has to be

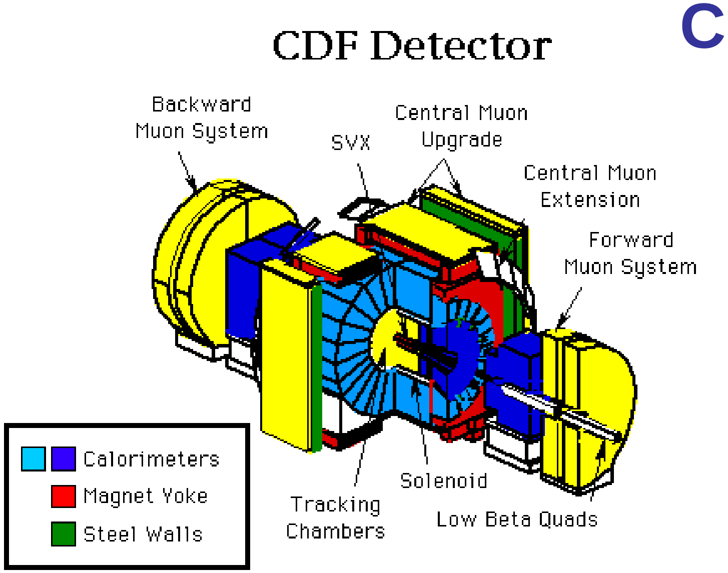
- efficient
- reliable
- easy to use

(Some) people who did the job

I.Bertram, A.Boehnlein, K.Bos, M.Diesberg, G.Garziolio, T.Harenberg, L.Luecking, A.Lyon, W.Merritt, R.StDenis, J.Templon, I.Teranov, V.White, D.Wicke, F.Wuerthwein, W.vanLeeuwen, ...

Thanks for providing me with information

CDF and D0



CDF: ~ 700 physicists,

60 institutions

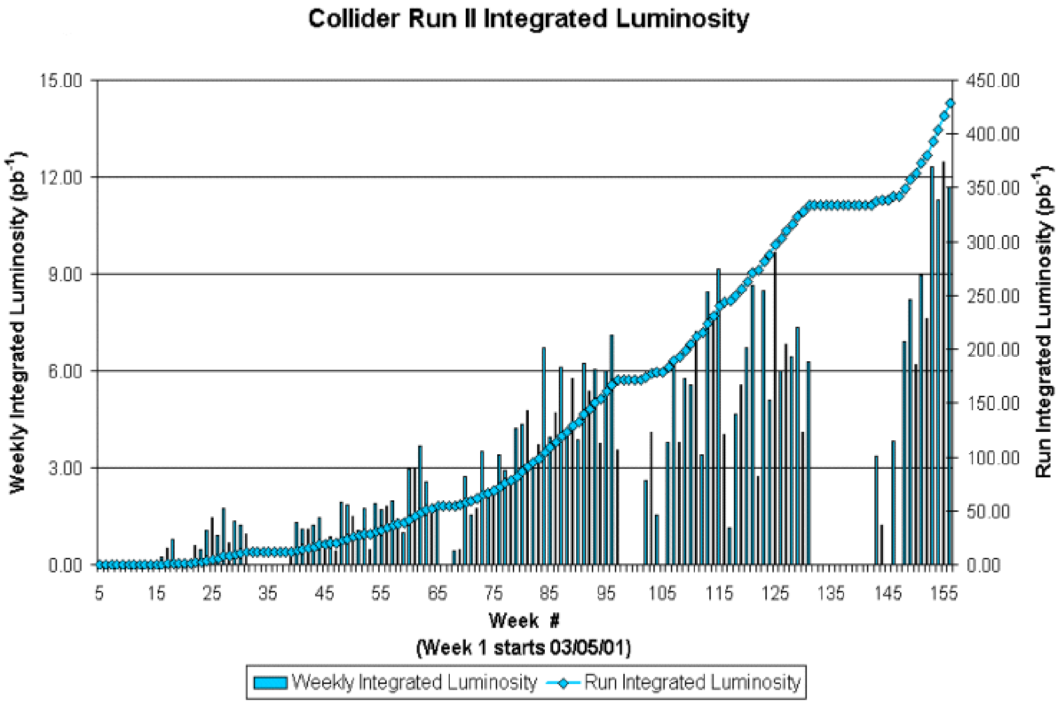
12 countries

D0: 650 physicists,

78 institutions

18 countries

A huge amount of data



600 – 700 TByte

of data

for each CDF + D0

D0 and CDF most similar to the LHC experiments!

FNAL: history of remote cptg.

Collaborations becoming more and more international:

➔ **computing outside FNAL more important**

Tools to submit jobs locally setting up D0 environment

➔ **SAM, runjob, run time environment rte, ..**

Large campaigns: MC production, D0 reprocessing ....

➔ **Millions of events produced outside FNAL**

But: ‚simple‘ remote computing at its limits

➔ ***transition to GRID computing***

Tools @ FNAL

several years development of tools for remote computing

- **SAM:** GRID type data management
- **rte**: tarball to deliver all required executables on remote computer
- **(mc) runjob**: distribute jobs among resources and merge output

**Grew out of experiment specific needs (D0),
now general framework for Fermilab computing**

SAM

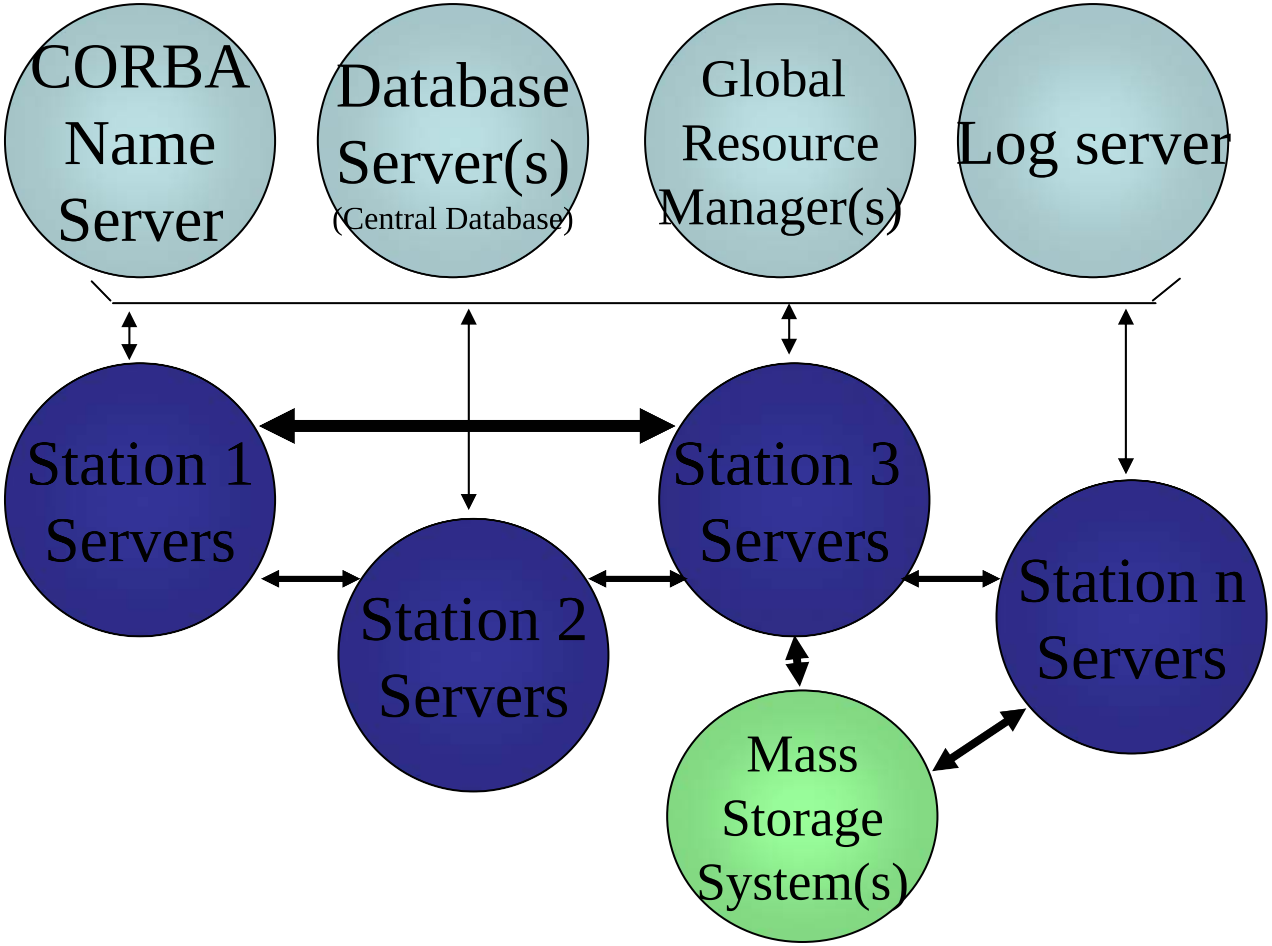
Sequential Data Access via Metadata

World wide data management system

Developed 1999 for D0 ➔ now central FNAL project

- **Data access/catalogue via meta – data. User defines projects instead of file names.**
- **File storage in SAM stations around the world**
- **Managing file delivery from around the world (transparent for user)**
- **Resource optimisation**
- **Substantial bookkeeping and history information**

Dataflow in SAM



CORBA Name Server

Database Server(s) (Central Database)

Global Resource Manager(s)

Log server

Station 1 Servers

Station 2 Servers

Station 3 Servers

Station n Servers

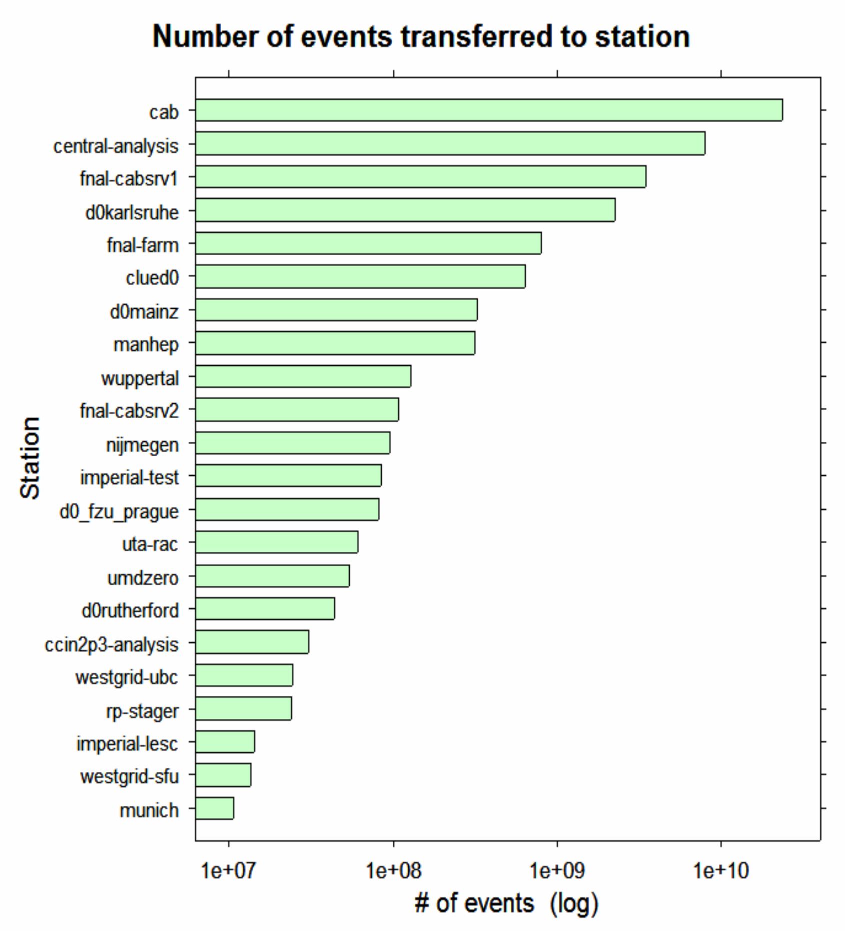
Mass Storage System(s)

Shared Globally

Local To Site

Shared Locally

SAM use in 2003



10s of Billions events,

1 PByte

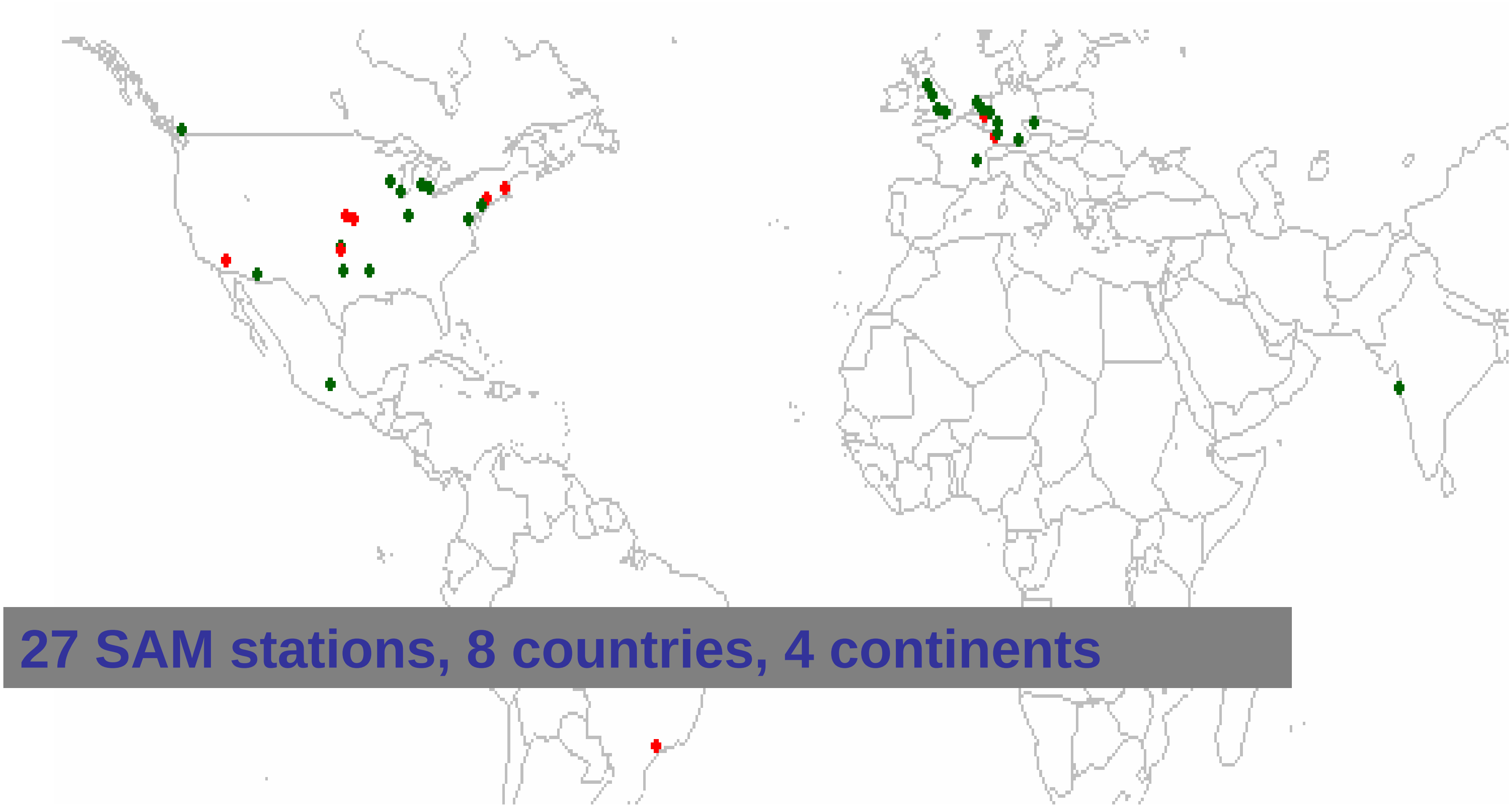
moved in D0 SAM stations!

Very small error rate!

Routinely used for physics analysis

Highly efficient data management even for huge demands

World – wide SAM



27 SAM stations, 8 countries, 4 continents

CDF + D0: different approaches

- CDF: remote computing mainly analysis
- D0 : remote computing also for central tasks

	CDF	D0
MC production	remote	remote
Primary reconstruction	FNAL	FNAL
Re-reconstruction	FNAL	20-50% remote
Analysis	FNAL + remote (20%)	

Remote computing more heavily used by D0!

Use – case I: MC production

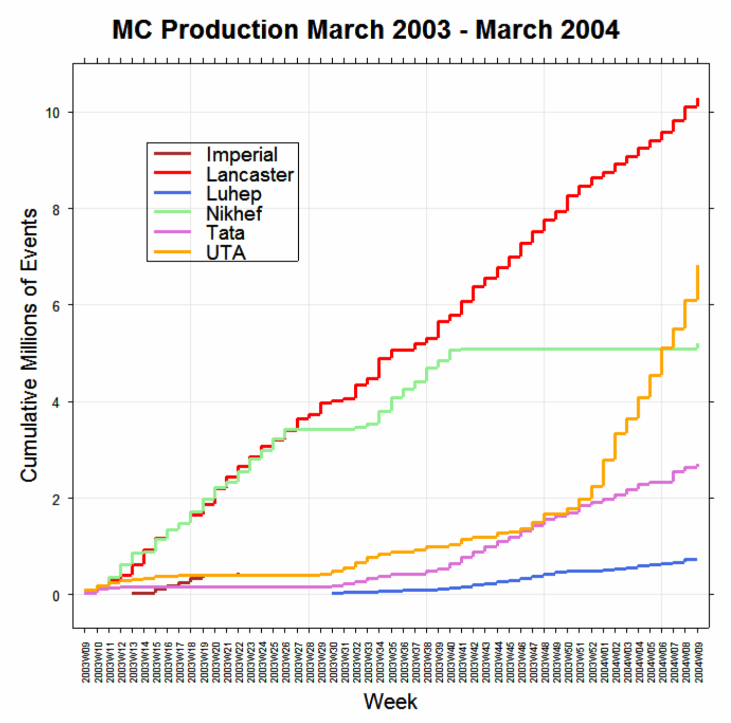
Since three years: all D0 MC generated outside FNAL

D0: UT Arlington, Prague, IC London, Lancaster, Lyon, NIKHEF, Tata, ...

CDF: Glasgow, Karlsruhe, Toronto

Millions of MC events generated on outside farms

stored in SAM ➔ easy use



Use case II: reprocessing

Reprocess all data with up – to date reconstruction

D0: 550 Mio events: Sep – Dec 03

At remote sites: *100 M events over 6 weeks*

➔ adds more than 2000 CPUs !

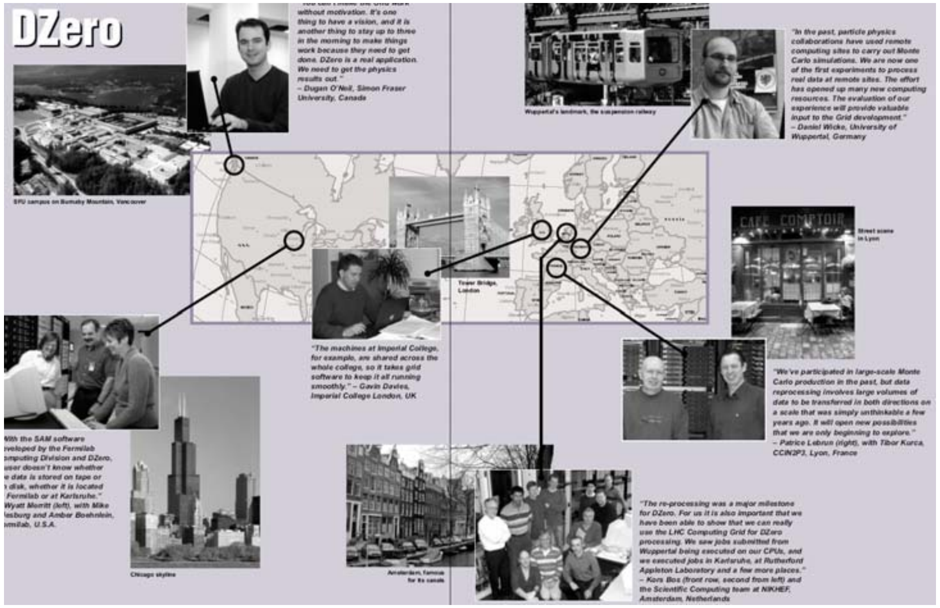
Canada (Vancouver)

France (Lyon)

Germany (Karlsruhe)

Netherlands (NIKHEF)

UK (IC London, RAL, Manchester, Lancester)



DZero

"You can't make the Grid work without motivation. It's one thing to have a vision, and it is another thing to stay up to three in the morning to make things work because they need to get done. DZero is a real application. We need to get the physics results out." – Dugan O'Neil, Simon Fraser University, Canada

SFU campus on Burnaby Mountain, Vancouver

Wuppertal's landmark, the suspension railway

"In the past, particle physics collaborations have used remote computing sites to carry out Monte Carlo simulations. We are now one of the first experiments to process real data at remote sites. The effort has opened up many new computing resources. The evaluation of our experience will provide valuable input to the Grid development." – Daniel Wicke, University of Wuppertal, Germany

Tower Bridge, London

Street scene in Lyon

"The machines at Imperial College, for example, are shared across the whole college, so it takes grid software to keep it all running smoothly." – Gavin Davies, Imperial College London, UK

"We've participated in large-scale Monte Carlo production in the past, but data reprocessing involves large volumes of data to be transferred in both directions on a scale that was simply unthinkable a few years ago. It will open new possibilities that we are only beginning to explore." – Patrice Lebrun (right), with Tibor Kurca, CCIN2P3, Lyon, France

With the SAM software eveloped by the Fermilab omputing Division and DZero, user doesn't know whether e data is stored on tape or n disk, whether it is located Fermilab or at Karlsruhe." Wyatt Merritt (left), with Mike iesburg and Amber Boehnlein, rmilab, U.S.A.

Chicago skyline

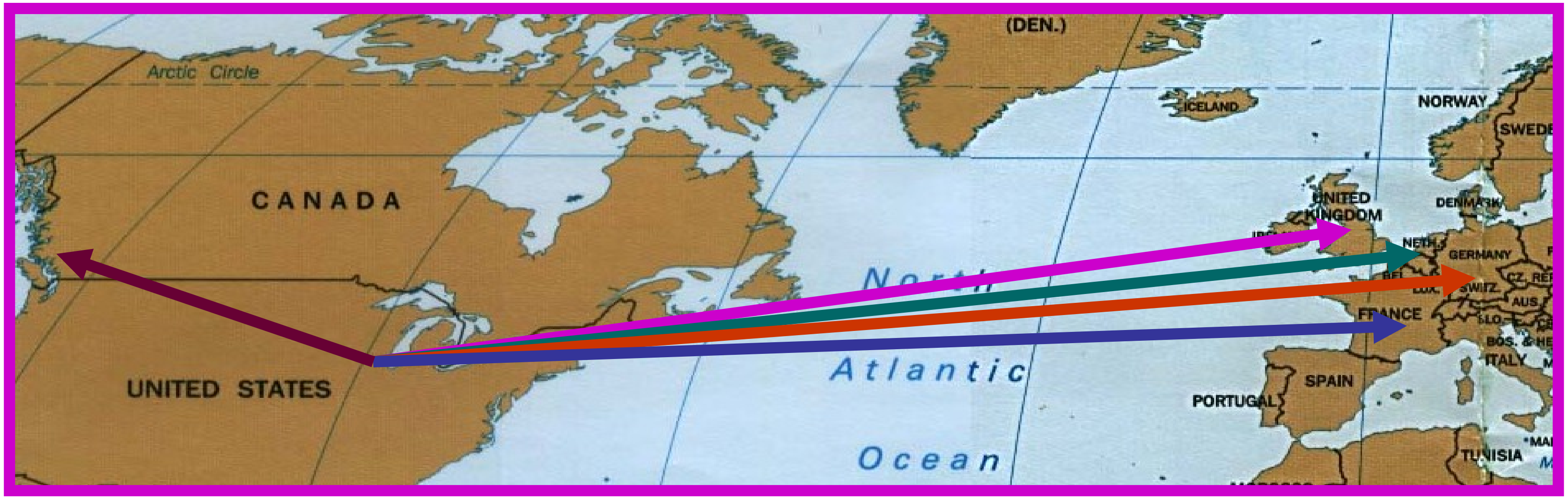
Amsterdam, famous for its canals

"The re-processing was a major milestone for DZero. For us it is also important that we have been able to show that we can really use the LHC Computing Grid for DZero processing. We saw jobs submitted from Wuppertal being executed on our CPUs, and we executed jobs in Karlsruhe, at Rutherford Appleton Laboratory and a few more places." – Kors Bos (front row, second from left) and the Scientific Computing team at NIKHEF, Amsterdam, Netherlands

Data transfer around the globe

Organisation: M.Diesberg (FNAL) + D.Wicke (Wuppertal) + on-site

- **certify sites**: **same sample ➔ same result**
- **Data transfer:** **~ 50 TB to be shipped using SAM**
- **Failed jobs:** **‚manual‘ resubmission per site**
- **Merging of files:** **complicated by job failures**
- **Monitoring:** **ad – hoc at each site**



From current remote computing

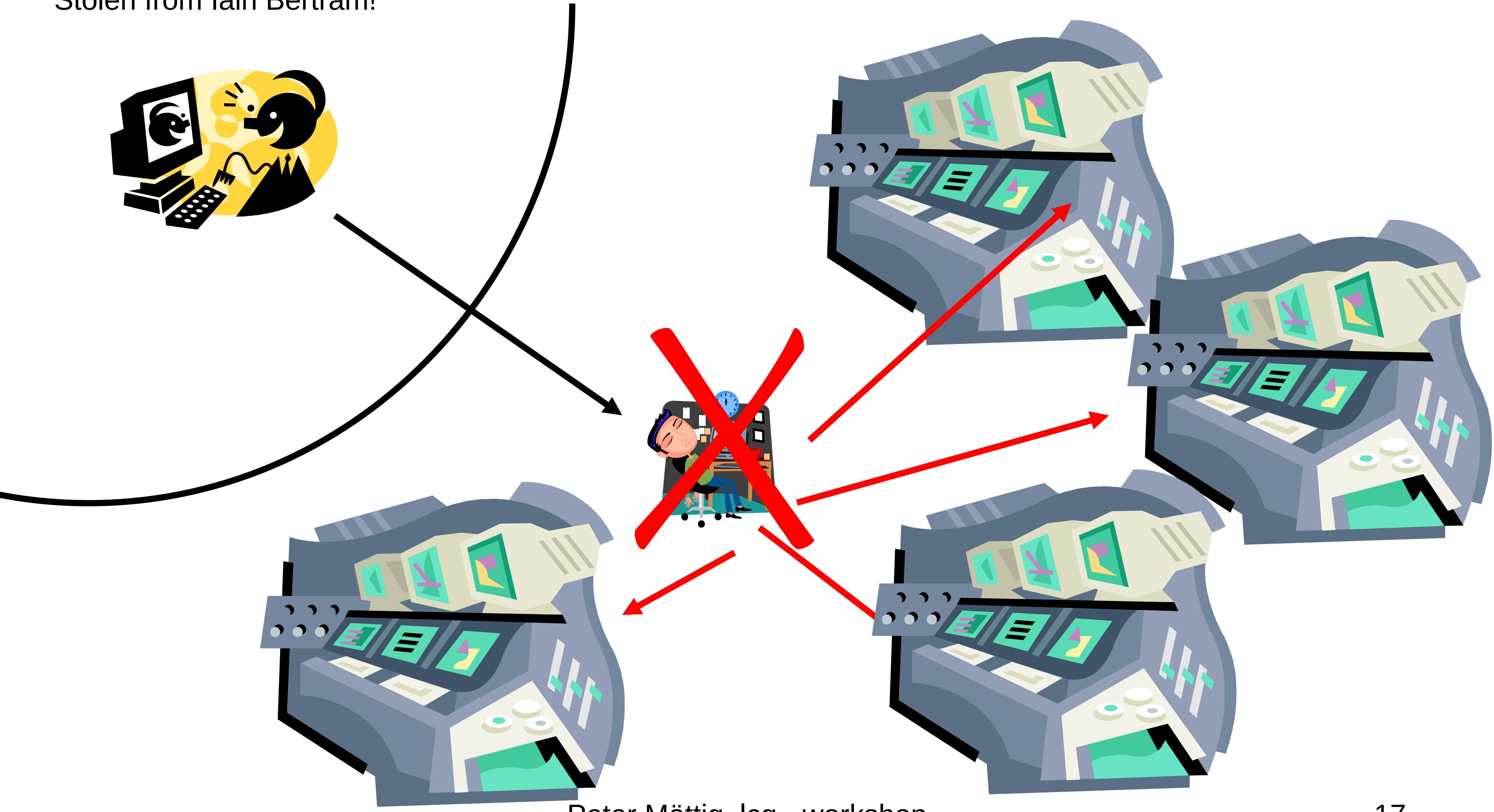
Stolen from Iain Bertram!

Person power intensive!

For imminent larger scale projects not feasible!

to GRID computing

Stolen from Iain Bertram!



Transition to GRID

D0 strategy:

Start with coordinated production:

1. MC production (easy to plan, relaxed reliability,
relaxed stability)

2. Reprocessing (easy to plan, high reliability,
high stability)

both production and test bed

Aim: stable and reliable running in 2004

CDF: plans to use GRID later

GRID platforms @ Tevatron

Fermilab product

SAM –GRID

Add to data management SAM:

-Job submission system

-Monitoring

Common CDF/D0 effort

Europeans (NIKHEF et al.):

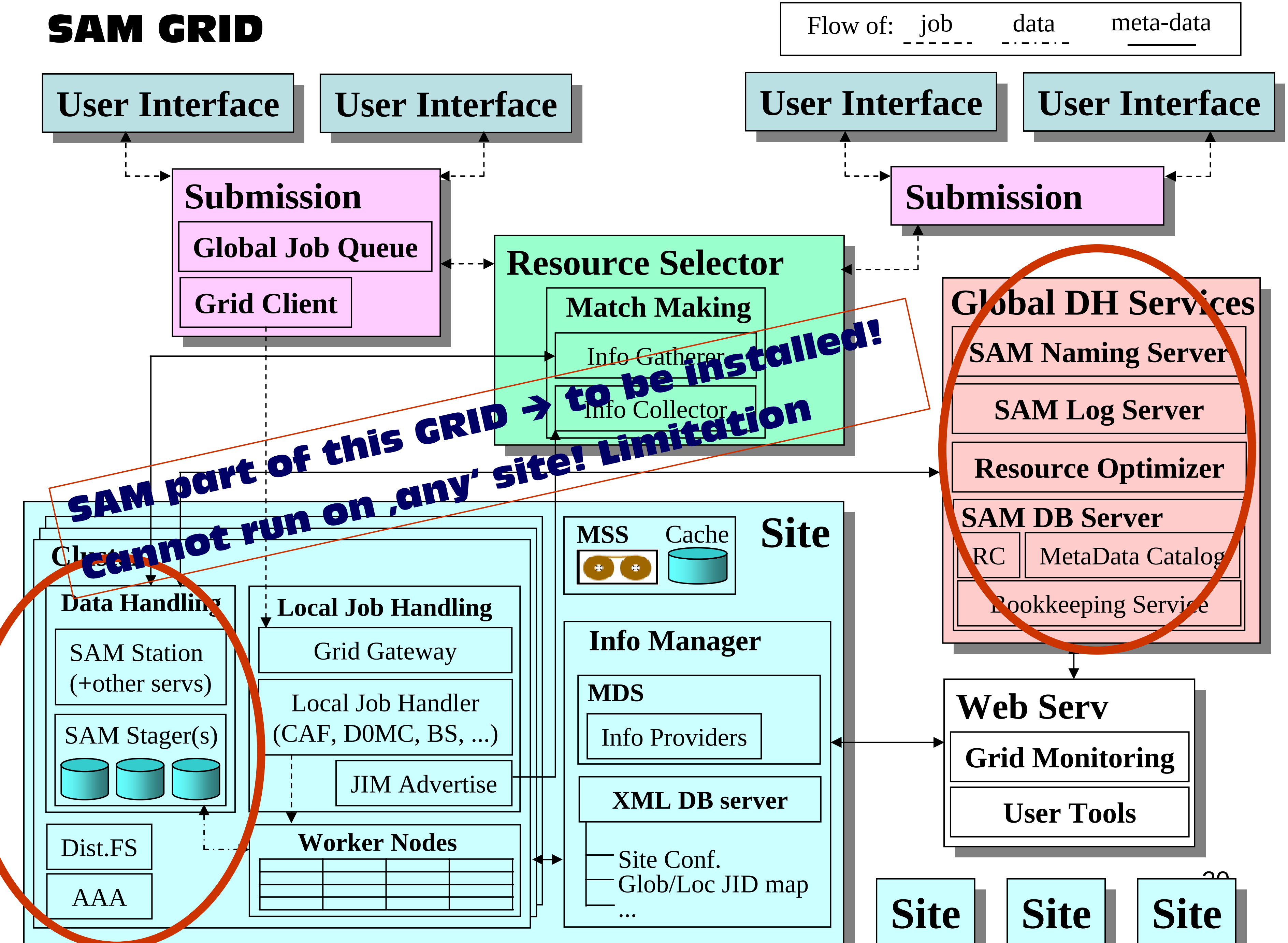
EDG + LCG

interface to SAM data management

and to D0 software

Requires good coordination → interoperability of D0 software/GRID!

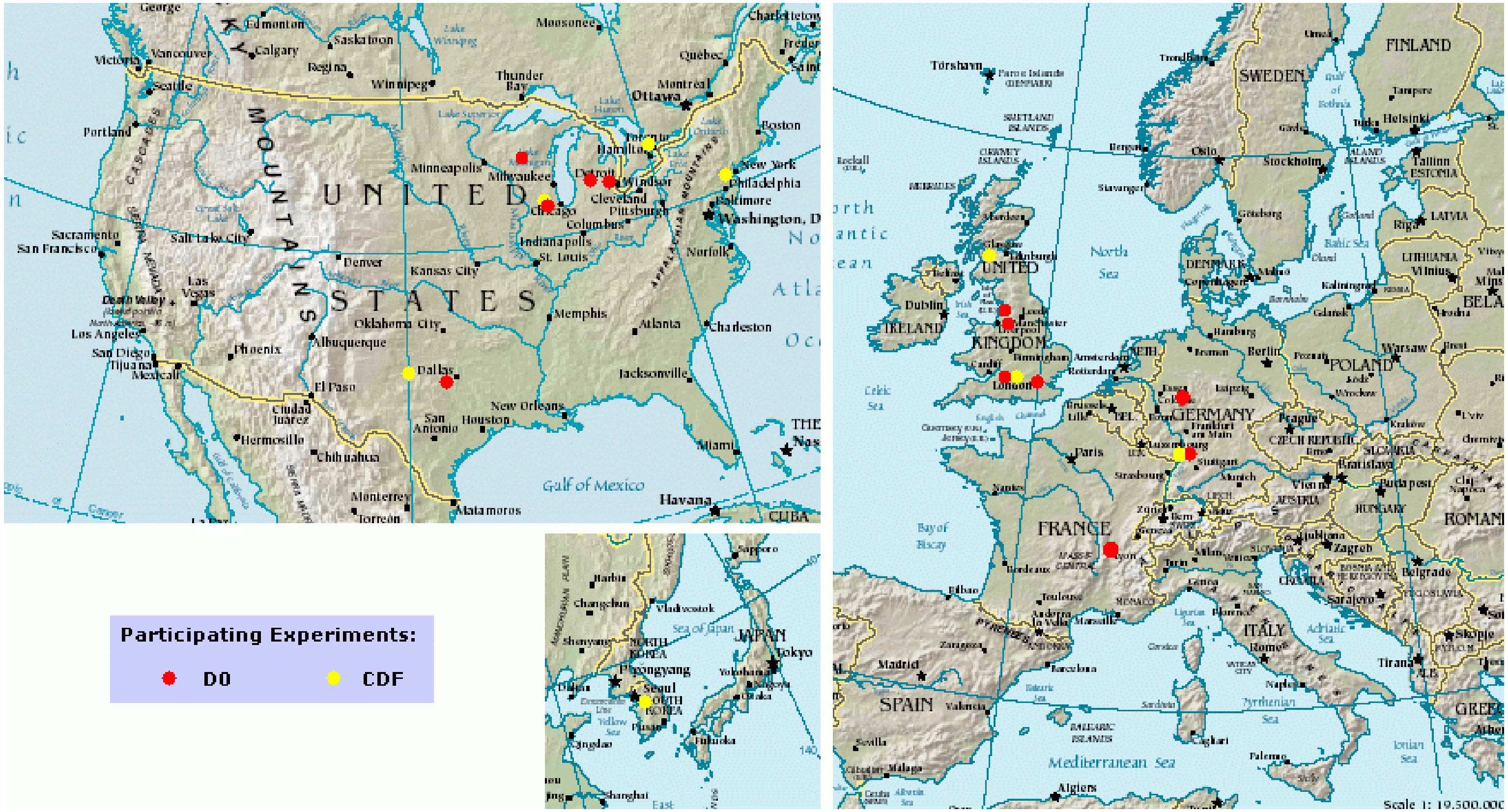
SAM GRID



SAM part of this GRID → to be installed!

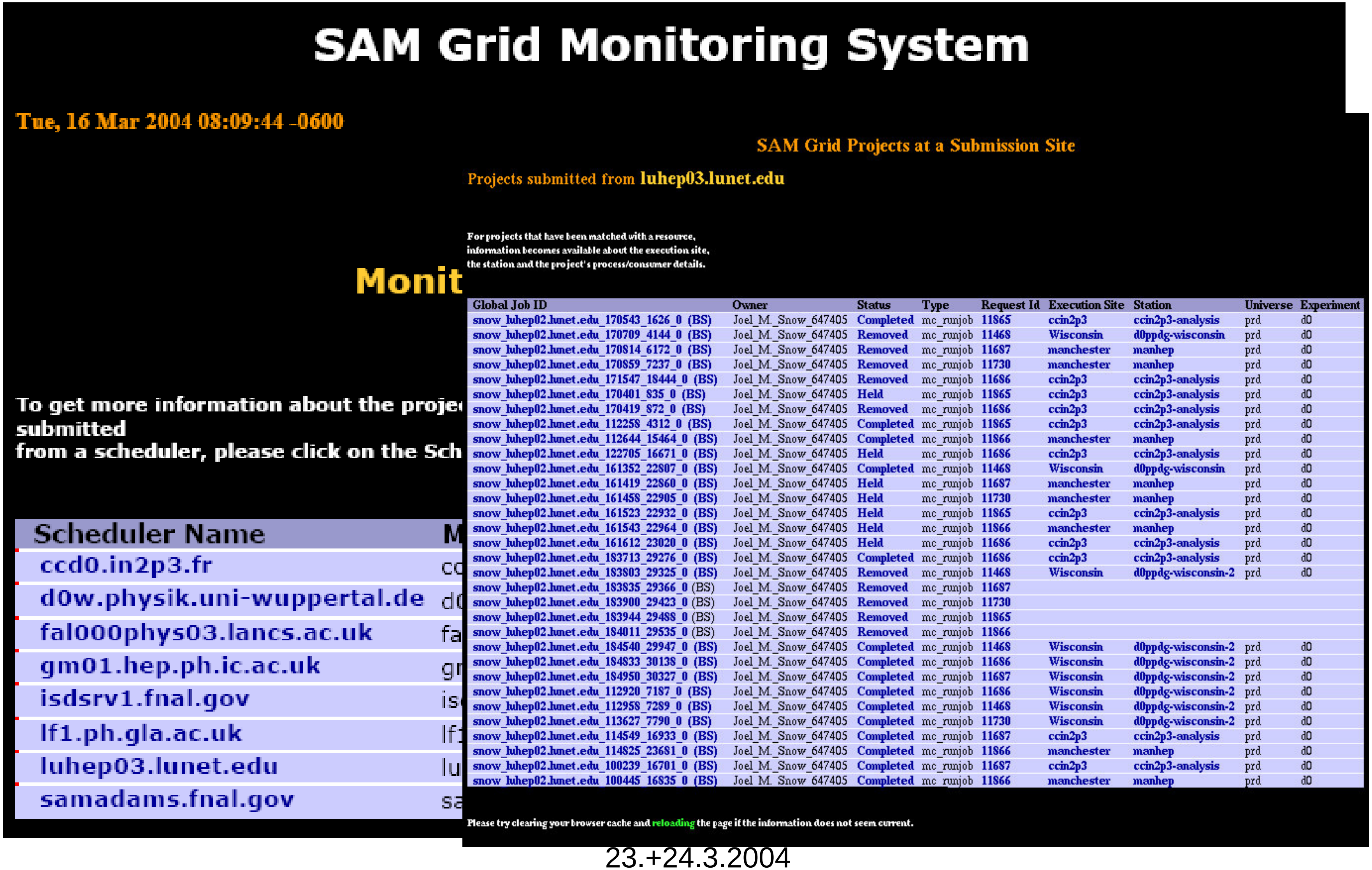
Cannot run on ‚any' site! Limitation

SAM – GRID stations 3 continents



JOINT D0 + CDF PROJECT

Monitoring & Information System



SAM Grid Monitoring System

Tue, 16 Mar 2004 08:09:44 -0600

Monit

To get more information about the projec submitted
from a scheduler, please click on the Sch

Scheduler Name	M
ccd0.in2p3.fr	cc
d0w.physik.uni-wuppertal.de	d(
fal000phys03.lancs.ac.uk	fa
gm01.hep.ph.ic.ac.uk	gr
isdsrv1.fnal.gov	is
lf1.ph.gla.ac.uk	lf1
luhep03.lunet.edu	lu
samadams.fnal.gov	sa

SAM Grid Projects at a Submission Site

Projects submitted from **luhep03.lunet.edu**

For projects that have been matched with a resource, information becomes available about the execution site, the station and the project's process/consumer details.

Global Job ID	Owner	Status	Type	Request Id	Execution Site	Station	Universe	Experiment
snow_luhep02.lunet.edu_170543_1626_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11865	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_170709_4144_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11468	Wisconsin	d0ppdg-wisconsin	prd	d0
snow_luhep02.lunet.edu_170814_6172_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11687	manchester	manhep	prd	d0
snow_luhep02.lunet.edu_170859_7237_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11730	manchester	manhep	prd	d0
snow_luhep02.lunet.edu_171547_18444_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11686	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_170401_835_0 (BS)	Joel_M._Snow_647405	Held	mc_runjob	11865	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_170419_872_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11686	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_112258_4312_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11865	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_112644_15464_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11866	manchester	manhep	prd	d0
snow_luhep02.lunet.edu_122705_16671_0 (BS)	Joel_M._Snow_647405	Held	mc_runjob	11686	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_161352_22807_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11468	Wisconsin	d0ppdg-wisconsin	prd	d0
snow_luhep02.lunet.edu_161419_22860_0 (BS)	Joel_M._Snow_647405	Held	mc_runjob	11687	manchester	manhep	prd	d0
snow_luhep02.lunet.edu_161458_22905_0 (BS)	Joel_M._Snow_647405	Held	mc_runjob	11730	manchester	manhep	prd	d0
snow_luhep02.lunet.edu_161523_22932_0 (BS)	Joel_M._Snow_647405	Held	mc_runjob	11865	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_161543_22964_0 (BS)	Joel_M._Snow_647405	Held	mc_runjob	11866	manchester	manhep	prd	d0
snow_luhep02.lunet.edu_161612_23020_0 (BS)	Joel_M._Snow_647405	Held	mc_runjob	11686	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_183713_29276_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11686	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_183803_29325_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11468	Wisconsin	d0ppdg-wisconsin-2	prd	d0
snow_luhep02.lunet.edu_183835_29366_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11687				
snow_luhep02.lunet.edu_183900_29423_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11730				
snow_luhep02.lunet.edu_183944_29488_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11865				
snow_luhep02.lunet.edu_184011_29535_0 (BS)	Joel_M._Snow_647405	Removed	mc_runjob	11866				
snow_luhep02.lunet.edu_184540_29947_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11468	Wisconsin	d0ppdg-wisconsin-2	prd	d0
snow_luhep02.lunet.edu_184833_30138_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11686	Wisconsin	d0ppdg-wisconsin-2	prd	d0
snow_luhep02.lunet.edu_184950_30327_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11687	Wisconsin	d0ppdg-wisconsin-2	prd	d0
snow_luhep02.lunet.edu_112920_7187_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11686	Wisconsin	d0ppdg-wisconsin-2	prd	d0
snow_luhep02.lunet.edu_112958_7289_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11468	Wisconsin	d0ppdg-wisconsin-2	prd	d0
snow_luhep02.lunet.edu_113627_7790_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11730	Wisconsin	d0ppdg-wisconsin-2	prd	d0
snow_luhep02.lunet.edu_114549_16933_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11687	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_114825_23681_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11866	manchester	manhep	prd	d0
snow_luhep02.lunet.edu_100239_16701_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11687	ccin2p3	ccin2p3-analysis	prd	d0
snow_luhep02.lunet.edu_100445_16835_0 (BS)	Joel_M._Snow_647405	Completed	mc_runjob	11866	manchester	manhep	prd	d0

Please try clearing your browser cache and reloading the page if the information does not seem current.

MC production with SAM - GRID

SAM-GRID: develop towards MC ‚production'

currently: Lyon, Wisconsin, Manchester

Some functionality:

- **deliver needed files via SAM**
- **automatic retries in case of communication failures**
- **file merging being automized**
- **start with on – site submission,**
  proceeding towards central submission

At this stage priority on high efficiency ➔ monitoring!

approaching a stable mode

During last 5 weeks ~ 1000 jobs with a total of 400,000 MC events

Continuous increase of efficiency from ~ 60% → 90%

Detailed bookkeeping of job failures:

-Site specific (exceeding maximum CPU limit, jobs sit idle, ...)

- middleware (Condor client does not work from a lap top, D0 code into infinite loop, ...)

-SAM-GRID (DBS communication, impact of main SAM gridftp server, ...)

Many problems identified and solved

The EDG way

Transfer files and

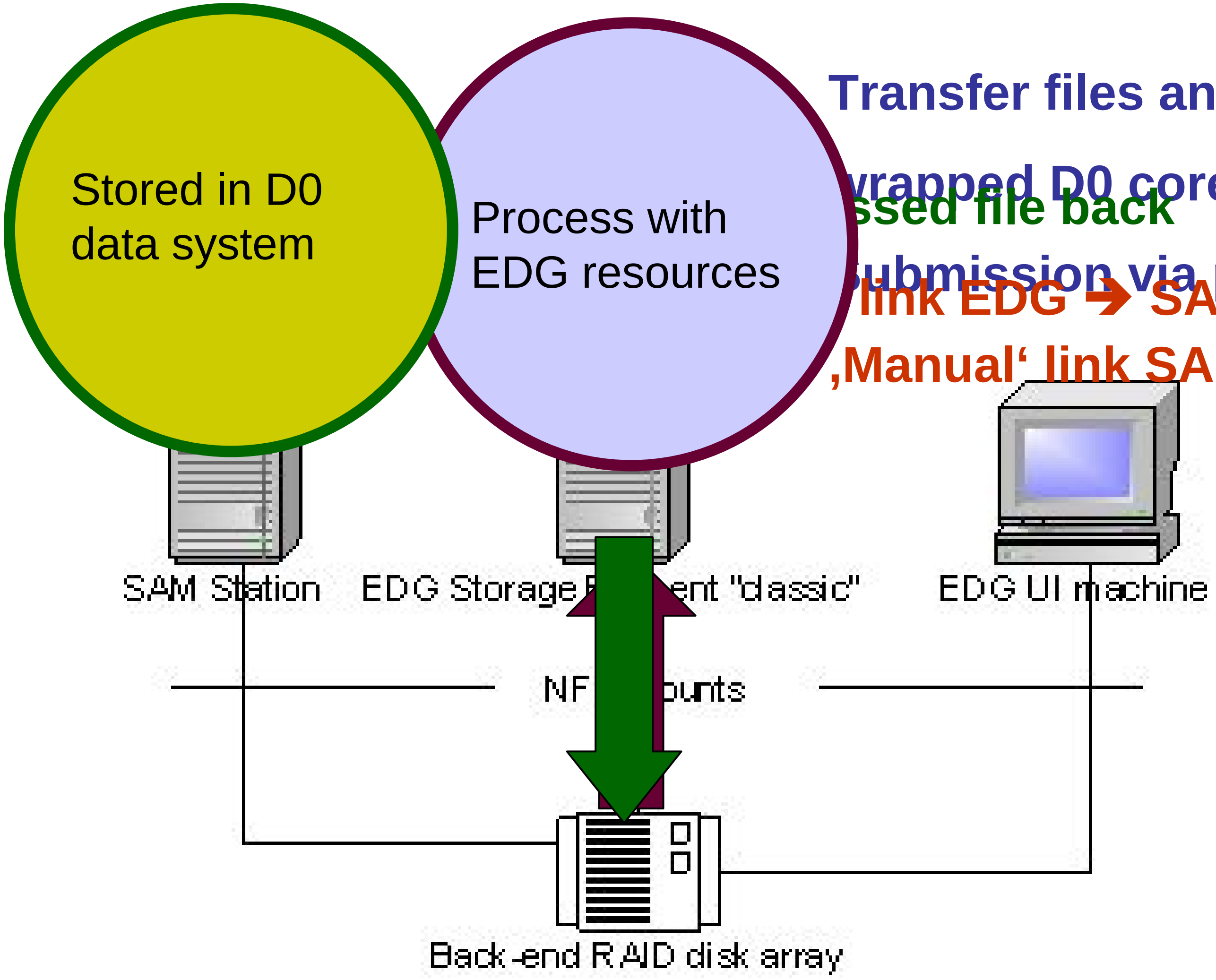
wrapped D0 core software

ssed file back

submission via python script

link EDG ➔ SAM

‚Manual‘ link SAM ➔ EDG



Stored in D0 data system

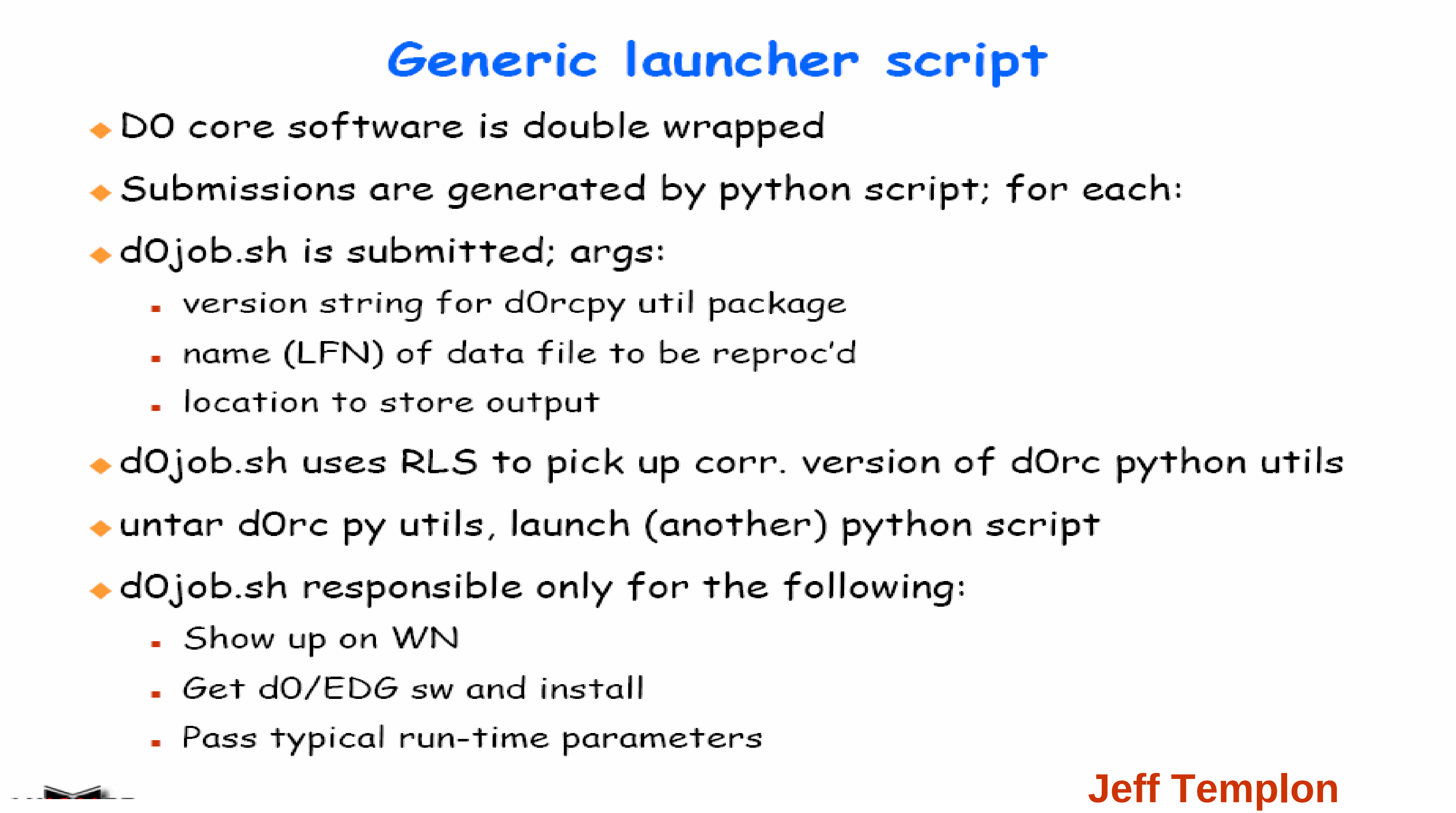
Process with EDG resources

Key point: interface

EDG ←➔ SAM!

Concept NIKHEF

In detail: submission procedure

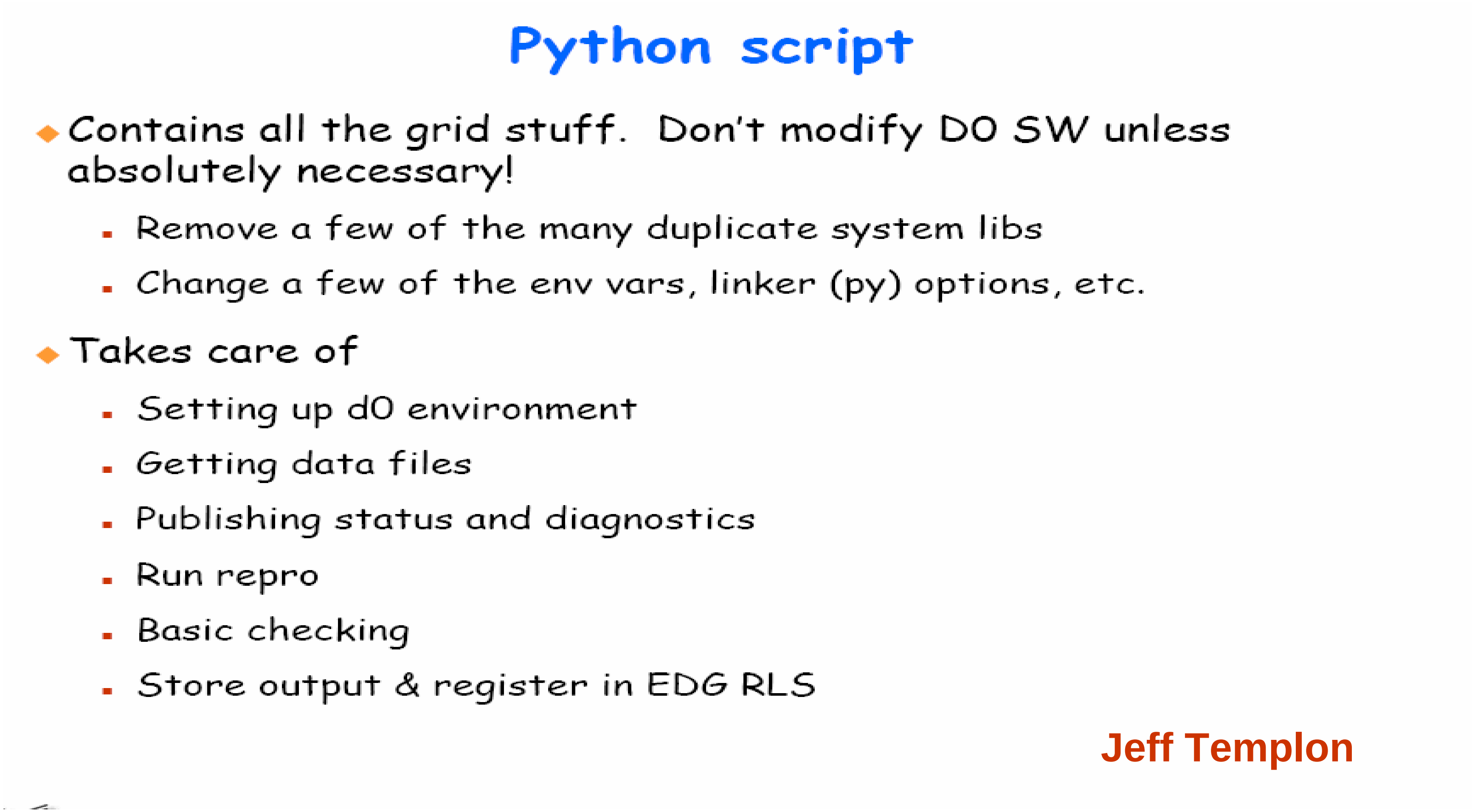


Generic launcher script

- D0 core software is double wrapped
- Submissions are generated by python script; for each:
- d0job.sh is submitted; args:
  - version string for d0rcpy util package
  - name (LFN) of data file to be reproc'd
  - location to store output
- d0job.sh uses RLS to pick up corr. version of d0rc python utils
- untar d0rc py utils, launch (another) python script
- d0job.sh responsible only for the following:
  - Show up on WN
  - Get d0/EDG sw and install
  - Pass typical run-time parameters

Jeff Templon

In the EDG world



Python script

- Contains all the grid stuff. Don't modify D0 SW unless absolutely necessary!
  - Remove a few of the many duplicate system libs
  - Change a few of the env vars, linker (py) options, etc.
- Takes care of
  - Setting up d0 environment
  - Getting data files
  - Publishing status and diagnostics
  - Run repro
  - Basic checking
  - Store output & register in EDG RLS

Jeff Templon

Reprocessing with EDG

End ´03: after 3 months of work – just before Christmas break

Jeff Templon, Dec 19, 23:54 per e - mail

, ... the first successful jobs are coming in now.‘

```
| site                         | cpu_time | wall_time | cpu_freq | success_code
| +----------------------------+-----------+----------+---------+----------------+
| physik.uni-wuppertal.de | 51291     | 57428 | 1792.412 | Job completed OK |
| physik.uni-wuppertal.de | 53958     | 61267 | 1792.409 | Job completed OK |
| in2p3.fr                     | 74107     | 77725 | 996.894   | Job completed OK |
| hep.phy.cam.ac.uk         | 76587     | 81828 | 1139.057 | Job completed OK |
| hep.phy.cam.ac.uk         | 77153     | 82282 | 1139.056 | Job completed OK |
| in2p3.fr                     | 77770     | 82085 | 996.894   | Job completed OK |
```

A proof of principle,
But not set - up for straining long – term production

Major lessons (Jeff Templon)

Note: final challenge for WP8 of EDG

➔ EDG for the first time applied to data taking experiment

- **Single storage machine is bottleneck**

  (Quite a few simultanous jobs trying to pull 2GB files each)

- **Stability of monitoring system, crucial particularly if job fails**

- **Software distribution reliable but inefficient**

- **Some problems could only be detected by D0 reprocessing (misconfigured nodes ➔ D0 much data crunch!**
  r-gma communication ➔ D0: 70 jobs per group!
  problems with production machines ➔ extensive use of management tools)

In preparation: MC with lcg

starting in NIKHEF ...... other sites to follow soon

Major next point:

- **a more automized way to relate to SAM**
- **make sure D0 environment clearly separated from GRID tools**
- **constant and comprehensive monitoring**

Need stable lcg to do stable processing!

... once running stable: more sites

The next year of GRID in D0

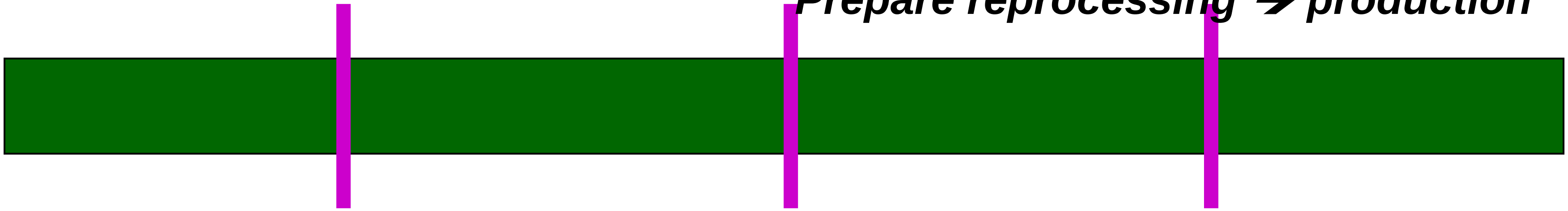
Autumn 03` **Winter – Spring 04** **Autum 04**

Reprocessing **Reprocessing**

MC – Production ==========================

Remote way ***Test SAM-GRID ➔ Production state***

Prepare reprocessing ➔ production



Test LCG ➔ Production state

EDG attempt

Prepare reprocessing ➔ production

The next reprocessing .....

Autumn: next D0 reprocessing

In total ~ 1 Billion events

→ 500 Million outside FNAL

→ 6 months of stable, reliable running!!!

→ No data to lose

A quantum leap → without GRID work intensive!

needed: central submission, monitoring, bookkeeping

A significant production task – a strain test for a GRID!

Beyond 2004

- **Data rate will beat Moore‘s law!**

  → GRID operation more and more important!

  (also CDF intends to use more remote cptg)
- **SAM as a very efficient data management system**

  → make it interoperable for different environments
- **Extend GRID use to more tasks and more users**

  → event selection by physics groups

  → chaotic, individual physics analysis

Tevatron experiments need a production GRID!

Offer insight into GRID performance under live conditions before LHC start-up

Real life always different from simulation!

An almost LHC GRID before LHC

Nothing is as demanding as a running experiment!

D0 and CDF offer environments

which challenges any GRID

100% EFFICIENCY, RELIABILITY, EASY TO USE
→ NO DATA TO BE LOST

PRESENT requirements close to the needs of LHC era

GRID that works for D0 & CDF likely to work for LHC!

→ test tools and system along real physicists needs!

Summary & Conclusions

D0 (and CDF) use extensively remote resources

In transition from remote to GRID computing!

Challenging production tasks

➔ long term strain tests for any GRID

Tevatron can provide invaluable lessons for LHC NOW!