Update on H1 and ZEUS Computing for HERA-II

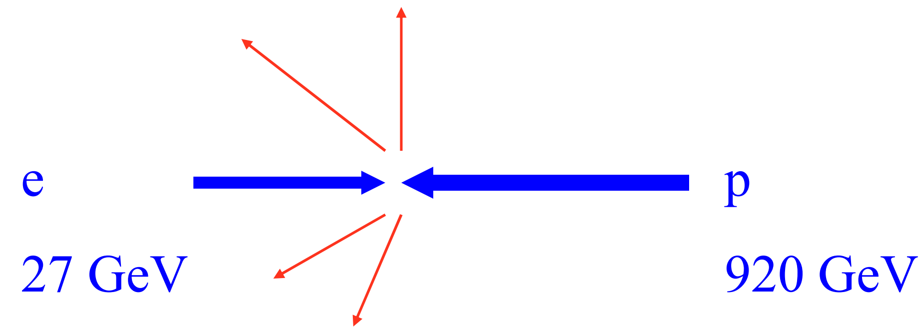
Rainer Mankel
DESY Hamburg

HTASC Meeting, CERN, 2-Oct-2003

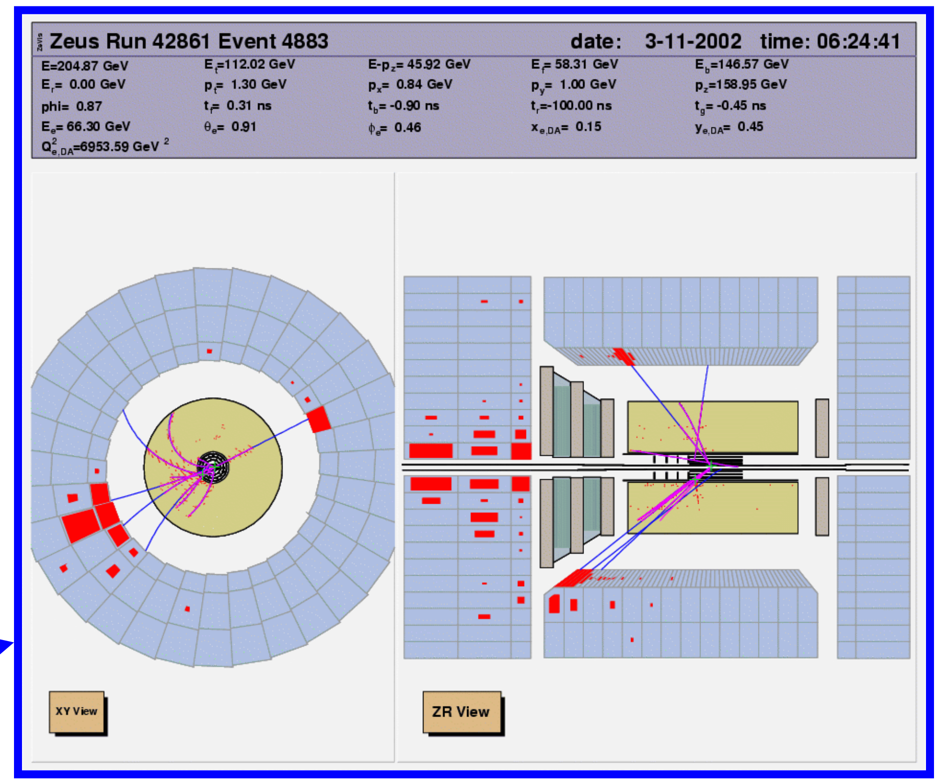
Outline

- HERA-II Computing Challenges
- Key Numbers
- Data Processing
- Mass Storage
- Glance at Application Software

HERA Collision Experiments: H1 & ZEUS



- HERA is currently the only operational collider in Europe
- H1 & ZEUS are general purpose experiments for *ep* collisions
  - HERMES has much smaller computing requirements, at least until 2005
  - HERA-B has finished data-taking
- About 400 physicists per expt
- HERA-II run targets 4-5 fold increase of luminosity compared to HERA-I



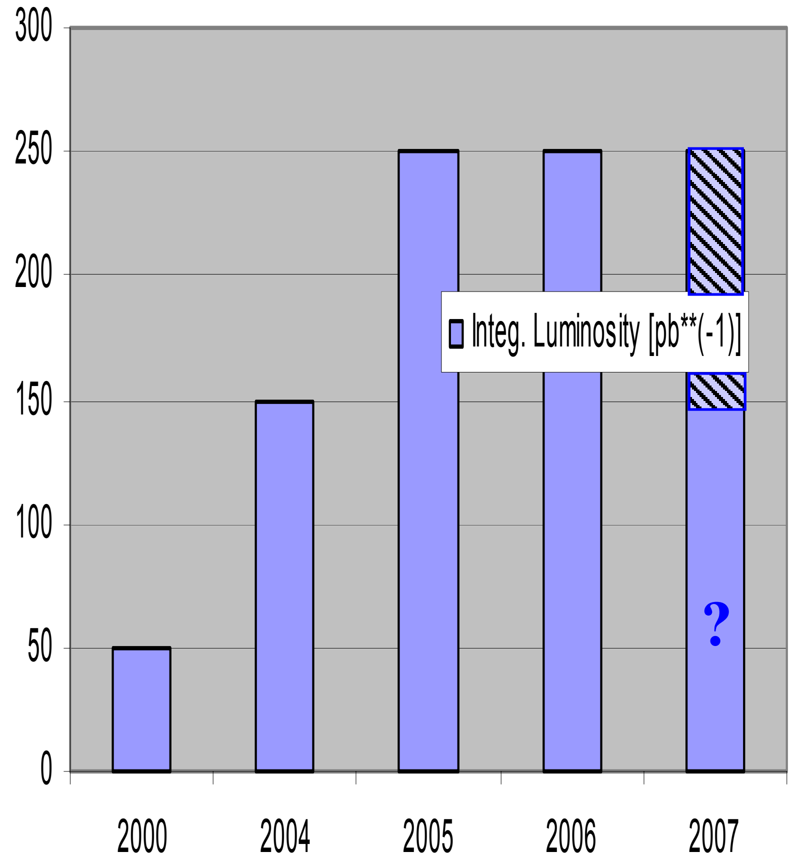
Neutral current event from 2002 run

HERA-II Luminosity

- Luminosity upgrade by increasing specific luminosity at similar currents
- Startup of machine and experiments has been slow because of unexpected background problems
  - proper modifications are in place now
- HERA has already demonstrated a factor of three increase in specific luminosity, as well as positron polarization
- Long runs planned, eg. 10 months in 2004

➔ Considerable new challenges to HERA computing

Delivered luminosity expectations for 2004-2006



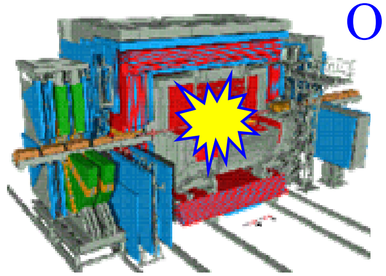
Changing Paradigms

- HERA has never had a multi-year shutdown.
  - Transition from HERA-I to HERA-II (luminosity upgrade) took place during a 9 month break
- HERA is worldwide the only *ep* collider, a unique machine. Data taken during HERA-II run will provide the last statement on many physics questions, at least for a very long time.
- Major paradigm shifts in computing in the last three years
  - transition SMP → Intel based farms
  - transition to commodity storage

[Note: ZEUS-related numbers are very fresh and highly preliminary]

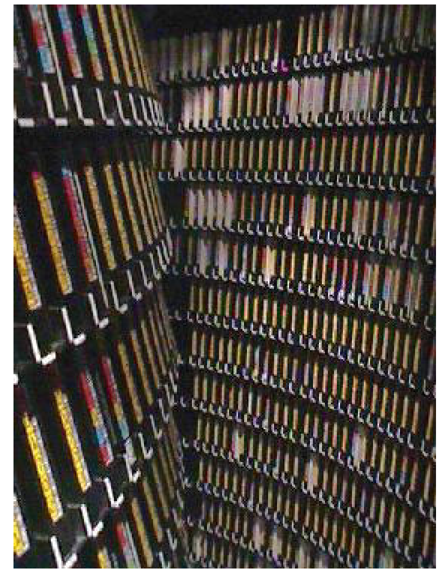
Computing Challenge of a HERA Experiment

50 M → 250 M delivered Events/year



O(1 M) detector channels

Tape storage incr. 30-60 TB/year

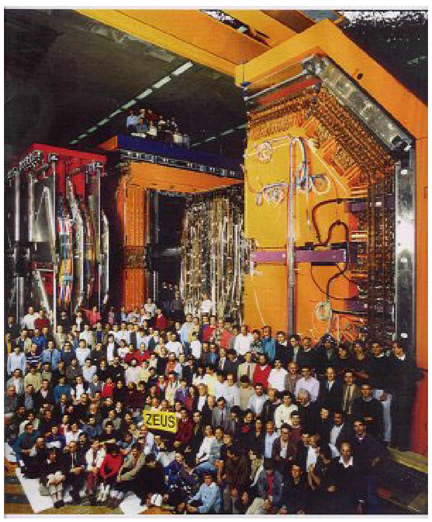


MC production

Data processing/ reprocessing

Data mining

Interactive Data Analysis



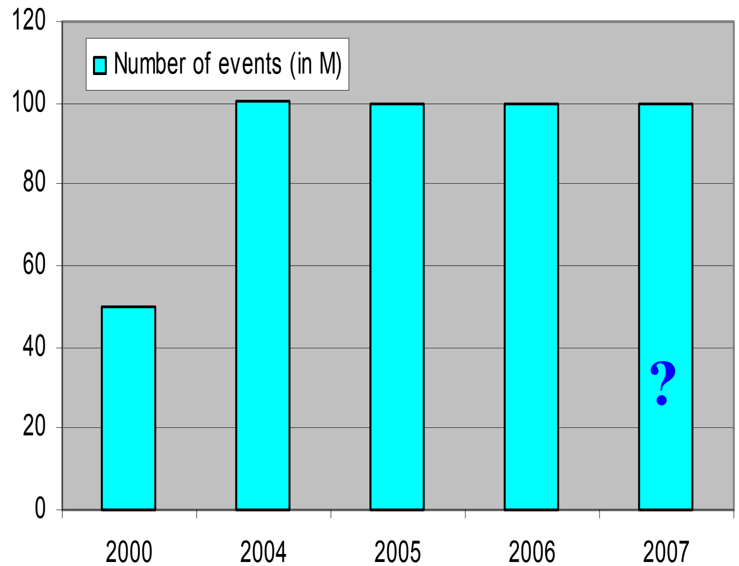
~450 Users

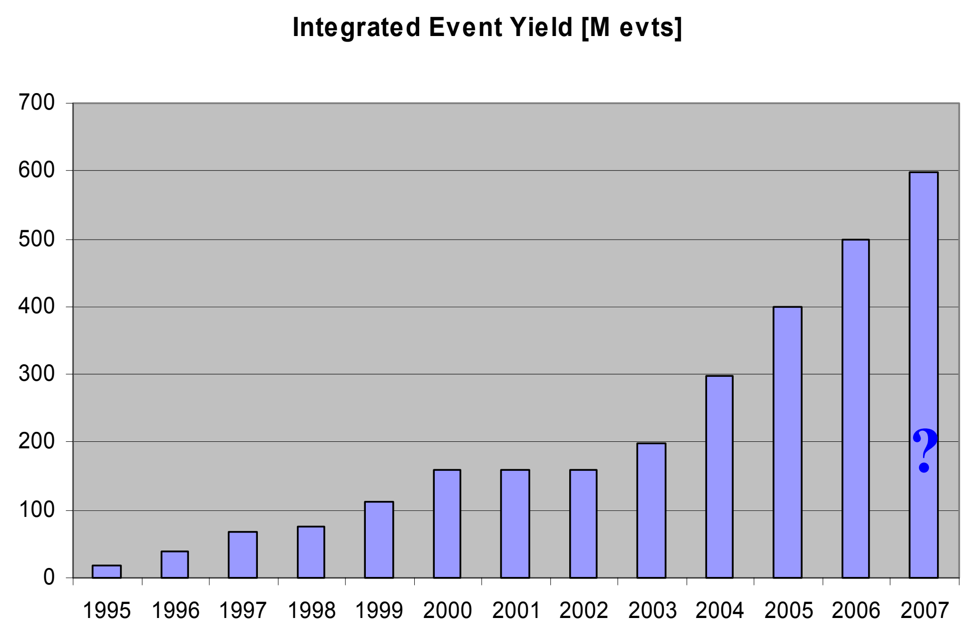
Disk storage

3-16 TB/year

Event Samples

- Increased luminosity will require a more specific online event selection
- Both H1 and ZEUS refine their trigger systems to be more selective
  - aim: reduction of trigger cross section by at least 60%
- Both H1 and ZEUS aim for 100 million events per year on tape





(ZEUS)

Transition to Commodity Components

Lesson learned from HERA-I:

- use of commodity hardware and software gives access to enormous computing power, but

➔ much effort is required to build reliable systems

- In large systems, there will always be a certain fraction of
  - servers which are down or unreachable
  - disks which are broken
  - files which are corrupt

➔ it is impossible to operate a complex system on the assumption that normally all systems are working

  - this is even more true for commodity hardware

- Ideally, the user should not even notice that a certain disk has died, etc
  - jobs should continue

➔ Need redundancy at all levels

Processing Power for Reconstruction & Batch Analysis

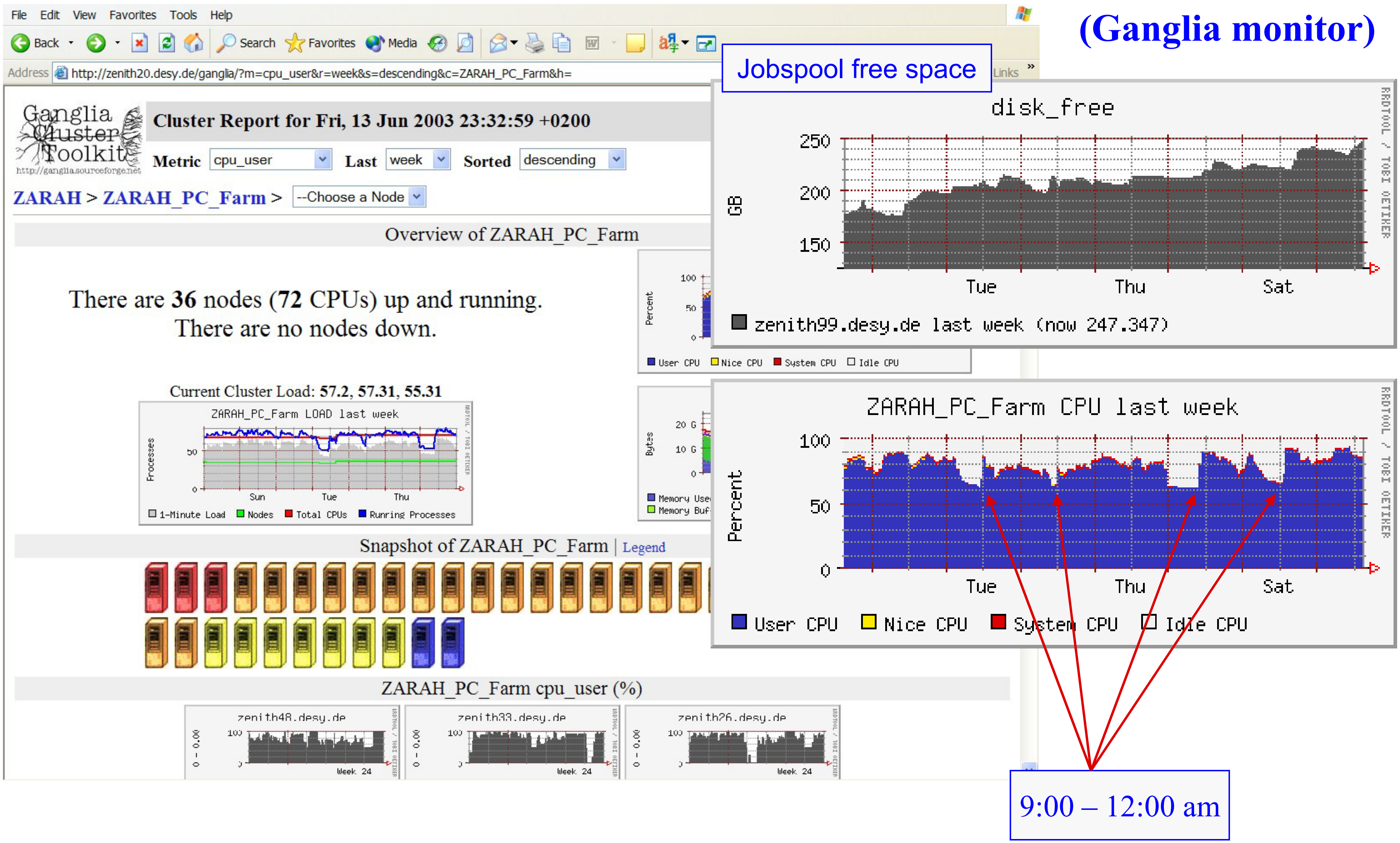
CPU Type	H1	ZEUS
P-III 500 MHz	44	
P-III 650 MHz		30
P-III 800 MHz	40	20
P-III 1 GHz		40
P-III 1.266 GHz	80	
Xeon 2 GHz		40
Total # processors	**164**	**130**
Total CPU power (R4400 units)	**1500**	**1355**

- Linux/Intel strategic platform. Dual-processor systems standard.
- Regular yearly upgrades planned: 10 (H1), 14 dual units (ZEUS)
  - in addition, upgrade of the ZEUS reconstruction farm is envisaged in 2004
- Batch systems in use: OpenPBS (H1) and LSF 4.1 (ZEUS), being upgraded to 5.0

Compute Server Nodes

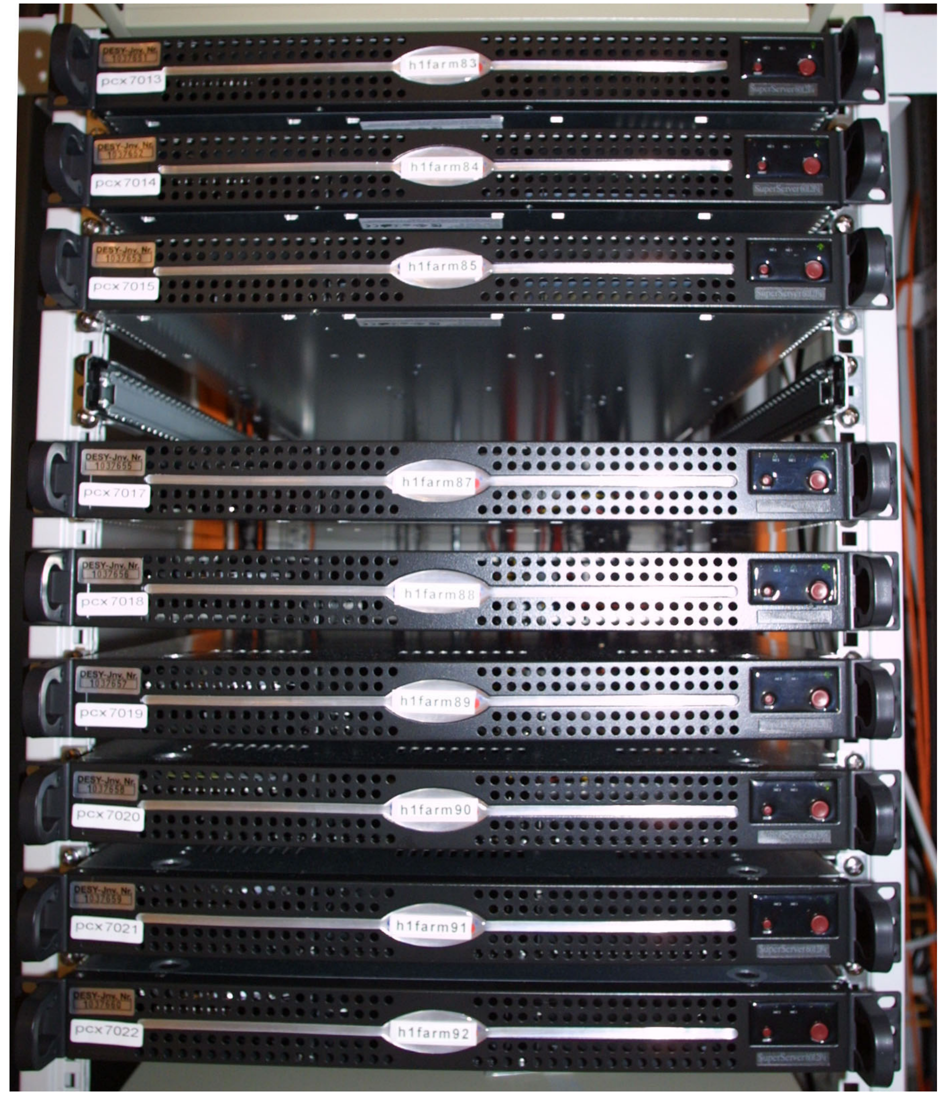
Snapshot of ZEUS Analysis Farm

(Ganglia monitor)



Farm Hardware

- Time for ATX towers is running out
- New farm node standard: 1U Dual-Xeon servers
  - Supermicro barebone servers in production
  - very dense concentration of CPU power (2 x 2.4 GHz per node)
  - installation by memory stick
- Issues to watch:
  - power consumption
  - heat / cooling

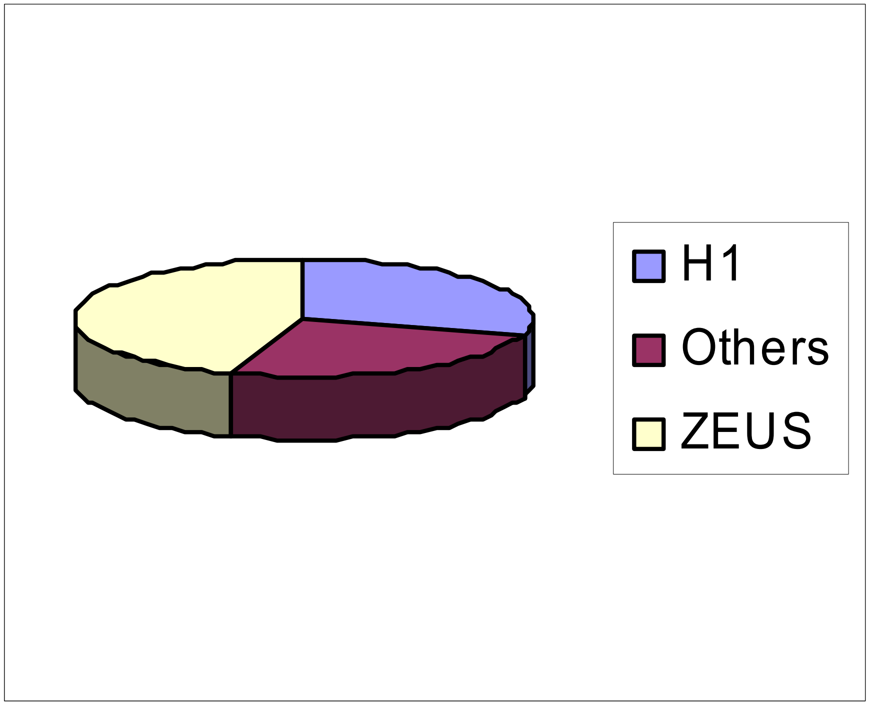


H1 simulation farm

Tape Storage Requirements

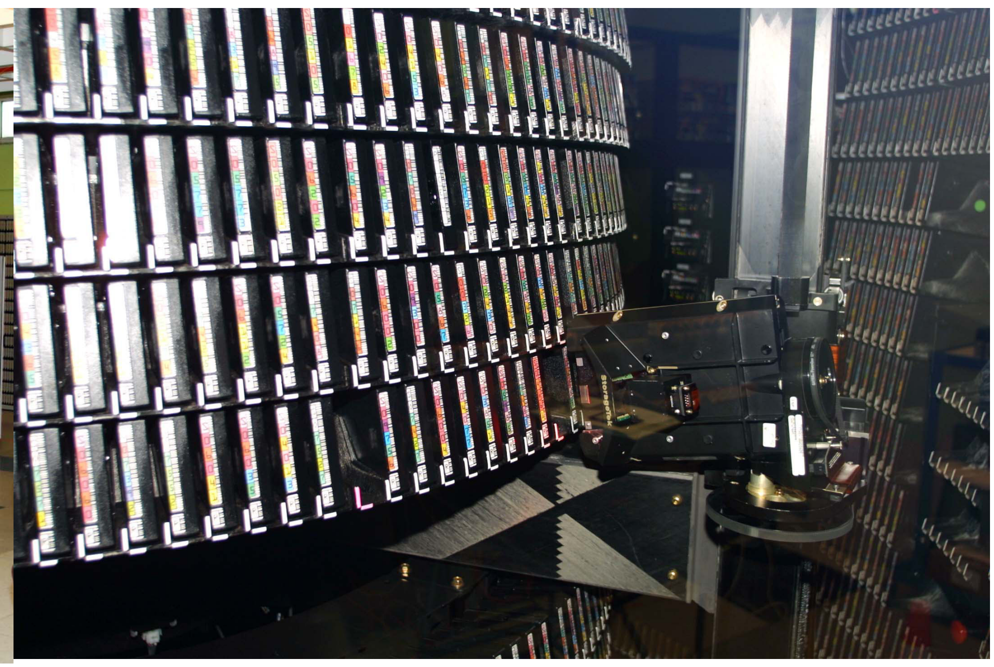
- Main reasons for using tapes:
  - data volumes too large to be entirely stored on disks
  - media cost relation 10:1
  - data safety
  - integral part of a powerful mass storage concept
- expect about 120 TB yearly increase (H1: 34 TB, ZEUS: 52 TB)
  - not including duplication for backup purposes (e.g. raw data)

➔ approaching Petabyte scale with end of HERA-II

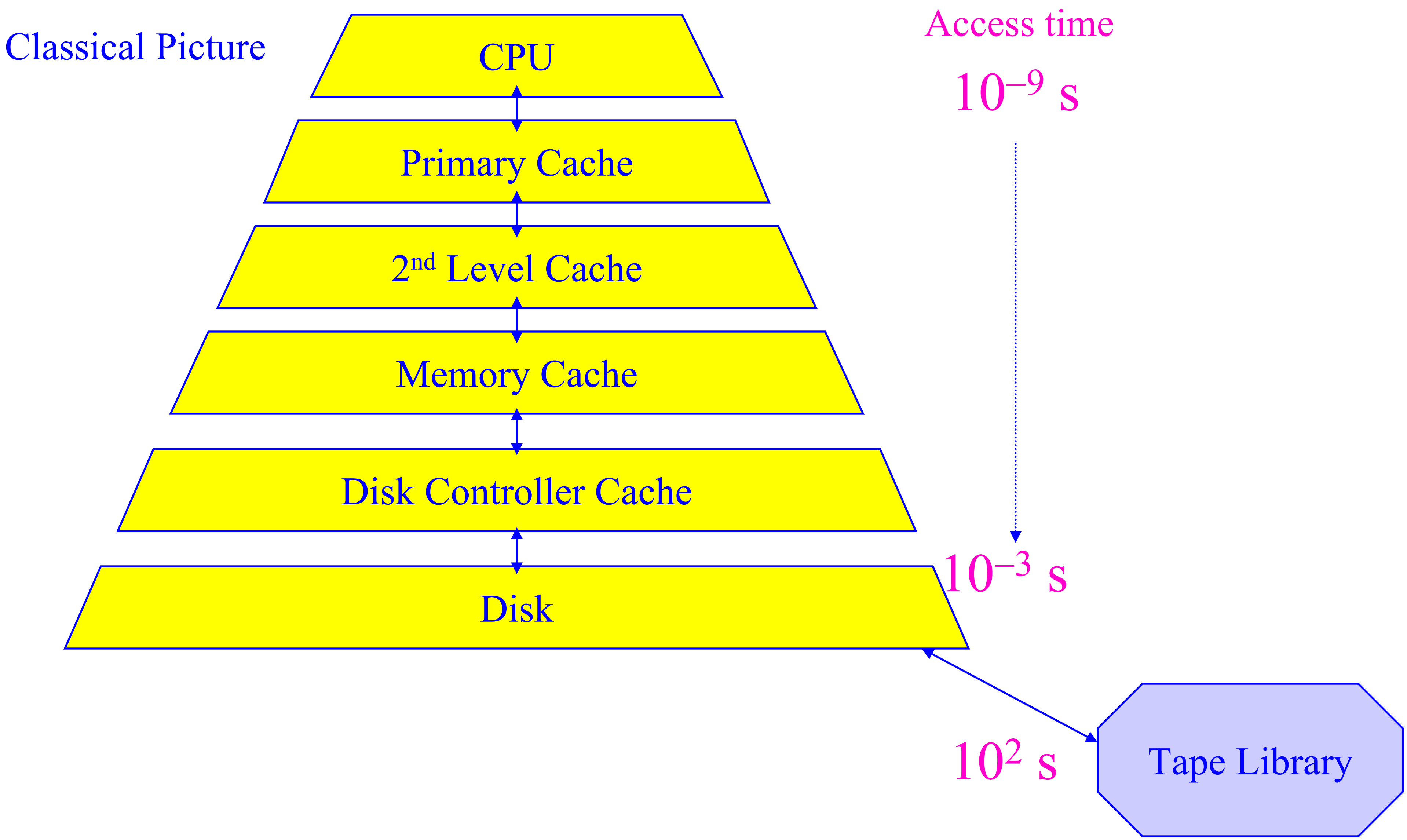


Mass Storage

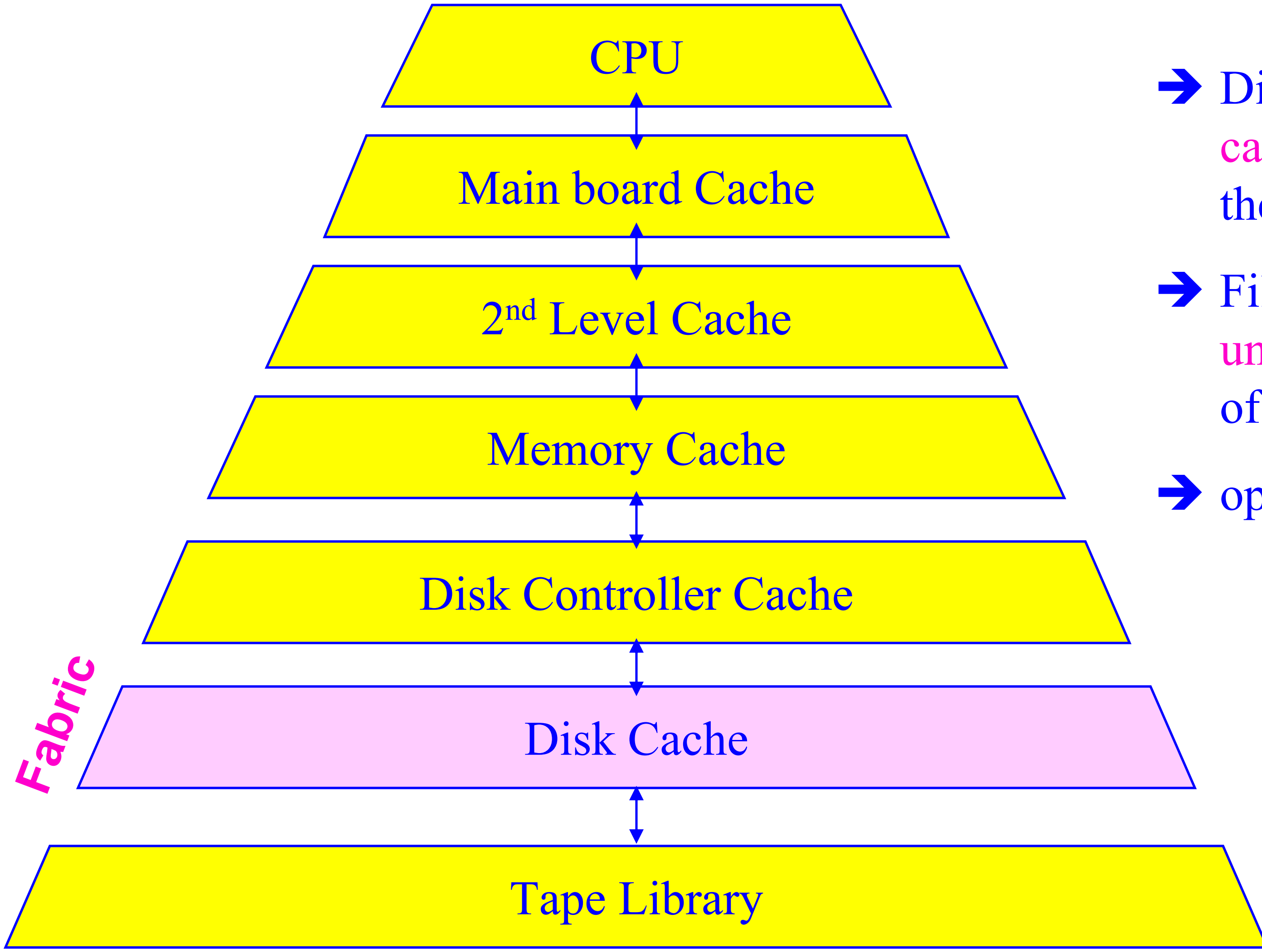
- DESY-HH uses 4 STK Powderhorn tape libraries (connected to each other)
  - in transition to new 9940 type cartridges, which offer 200 GB (instead of 20 GB)
- Much more economic, but loading times increase
- Need a caching disk layer to shield user from tape handling effects
- Middleware is important



Cache within a cache within a cache



dCache Picture



- Disk files are only cached images of files in the tape library
- Files are accessed via a unique path, regardless of server name etc
- optimized I/O protocols

dCache

- Mass storage middleware, used by all DESY experiments
  - has replaced ZEUS tpfs (1997) and SSF (2000)
- joint development of DESY and FNAL (also used by CDF, MINOS, SDSS, CMS)
- Optimised usage of tape robot by coordinated read and write requests (read ahead, deferred writes)
- allows to build large, distributed cache server systems
- Particularly intriguing features:
  - retry-feature during read access – job does not crash even if file or server become unavailable (as already in ZEUS-SSF)
  - “Write pool” used by online chain (reduces #tape writes)
  - reconstruction reads RAW data directly from disk pool (no staging)
- grid-enabled. Also ROOT can open dCache files directly

➔ randomized distribution of files across the distributed file servers

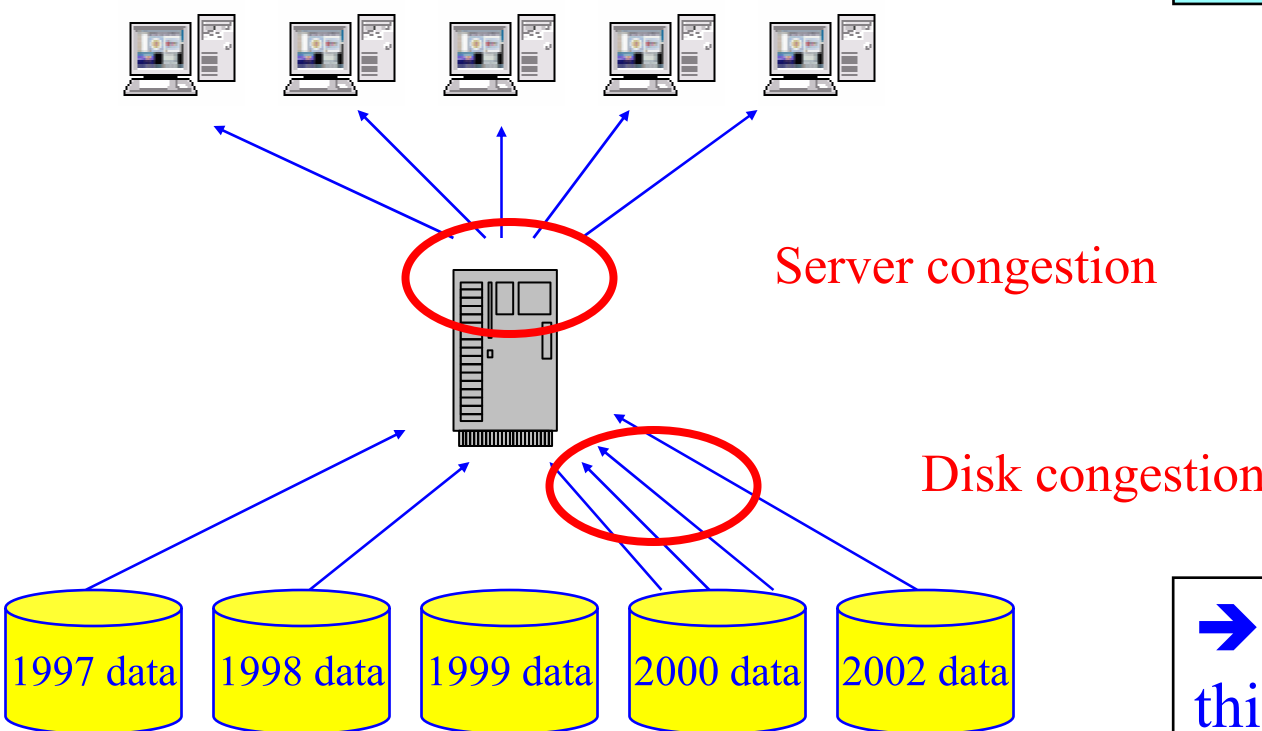
➔ avoids hot spots and congestion

➔ scalable

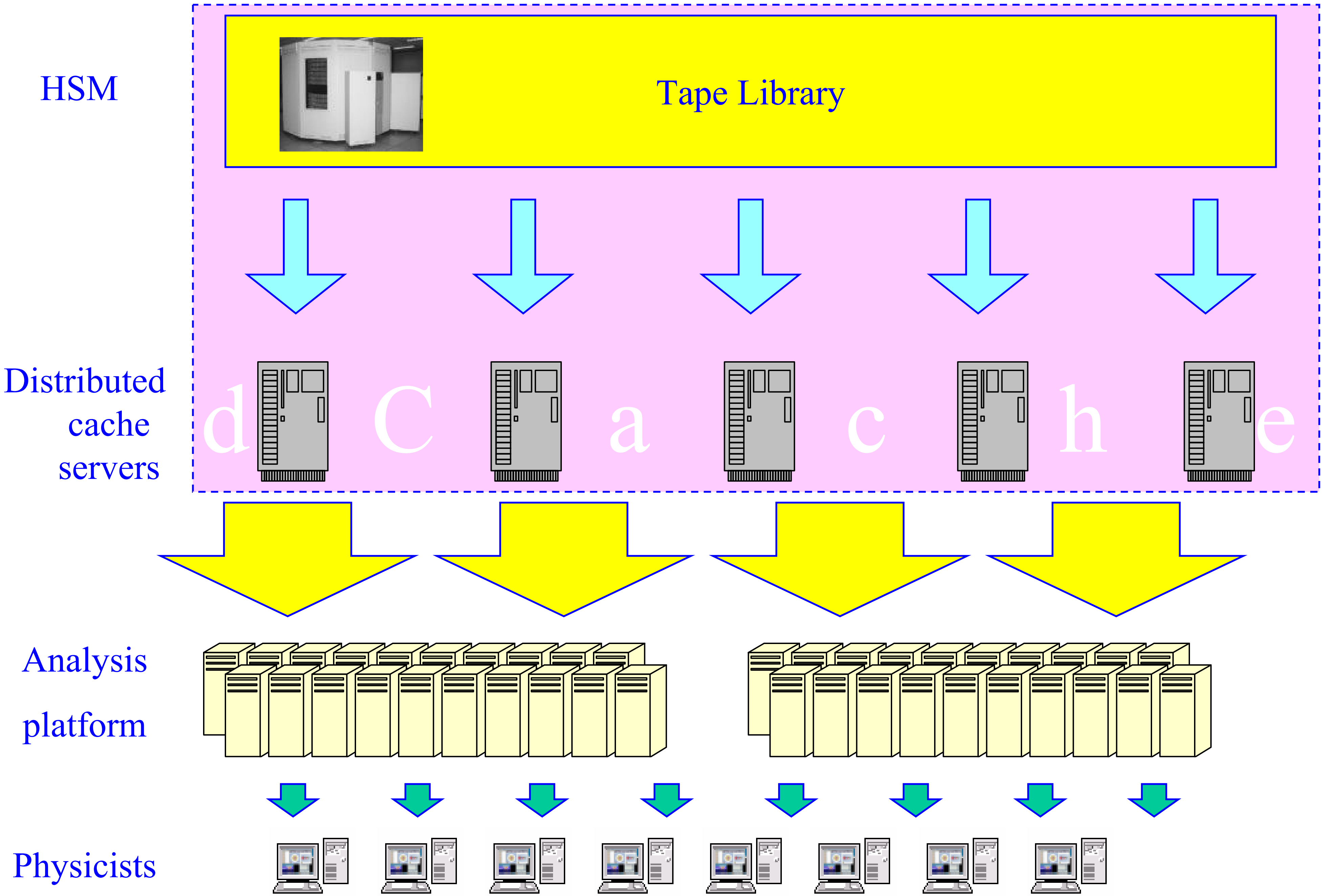
Disk Access Scaling Issues

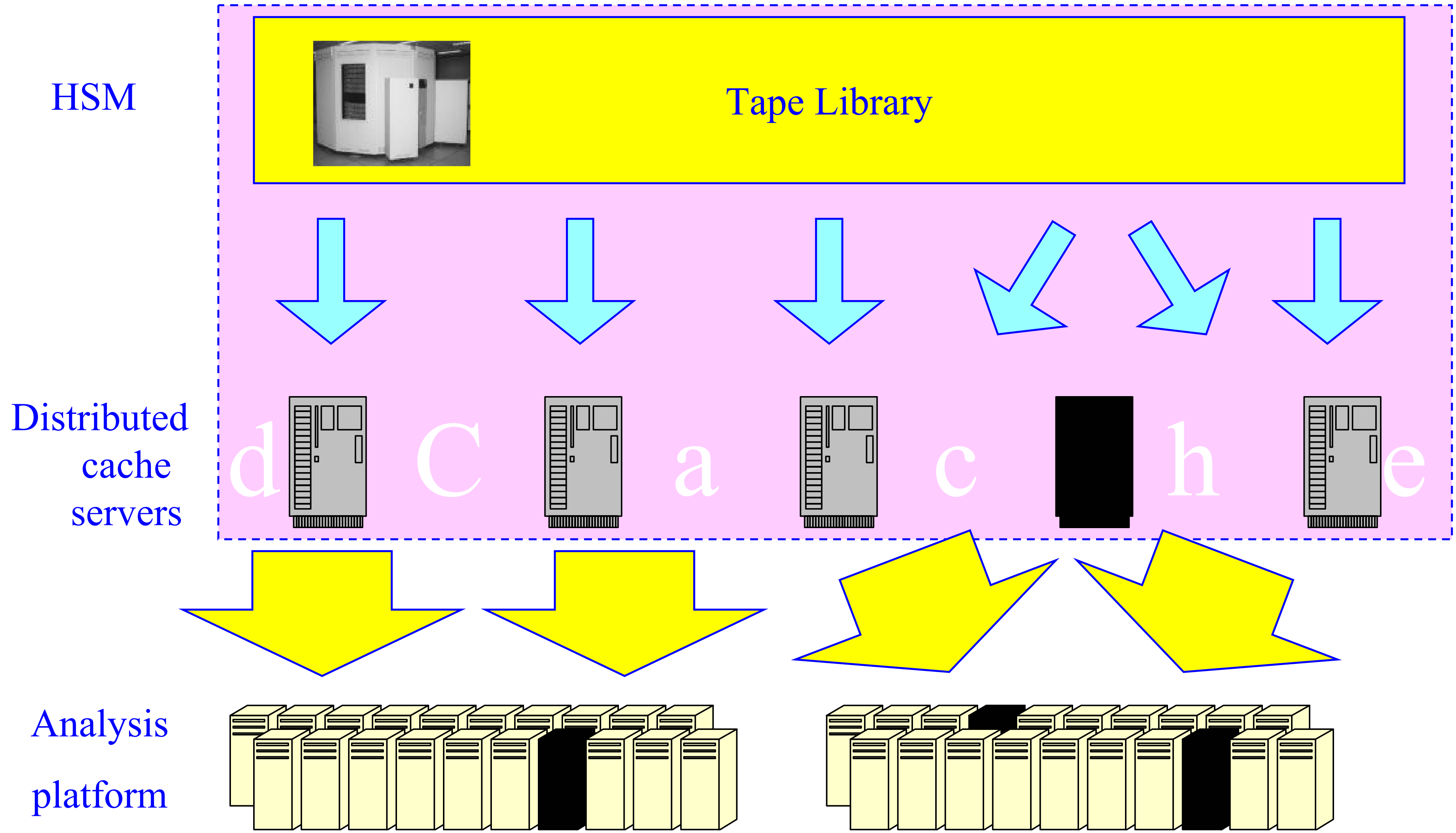
- The classical picture does not scale

➔ need a sophisticated mass storage concept to avoid bottlenecks



➔the dCache solves this problem elegantly


HSM
Tape Library
Distributed cache servers
d C a c h e
Analysis platform
Physicists



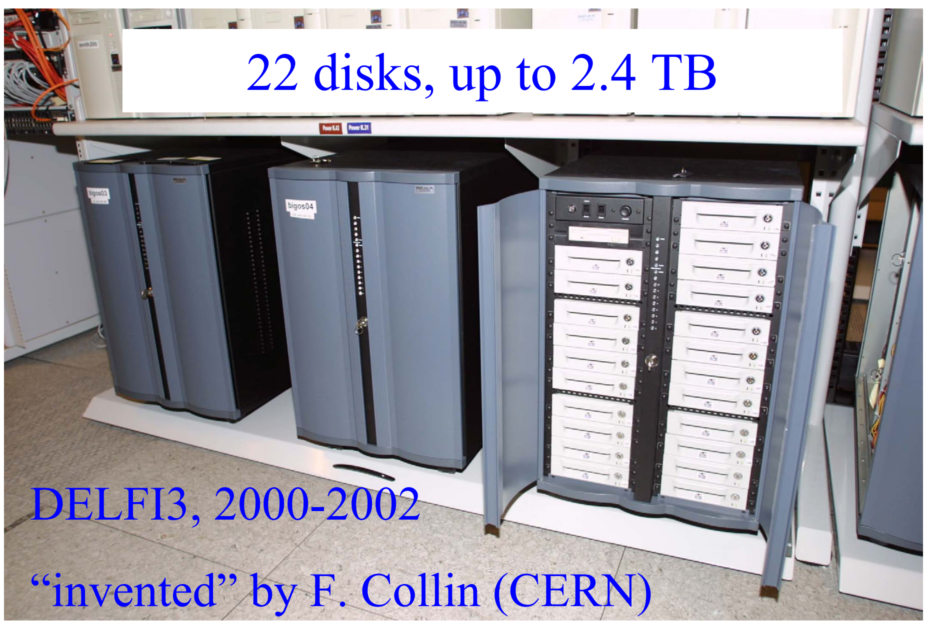
- ➔ Builtin redundancy at all levels
- ➔ On failure of a cache server, the requested file is staged automatically to another server

Commodity File Servers

- Affordable disk storage decreases number of accesses to tape library
- ZEUS disk space (event data):

– begin 2000:	3 TB	fraction FC+SCSI: 100%
– mid 2001:	9 TB	67%
– mid 2002:	18 TB	47%

- necessary growth only possible with commodity components
- commodity components need “fabric” to cope with failures → dCache



22 disks, up to 2.4 TB

DELFI3, 2000-2002

“invented” by F. Collin (CERN)

12 disks, up to 2 TB

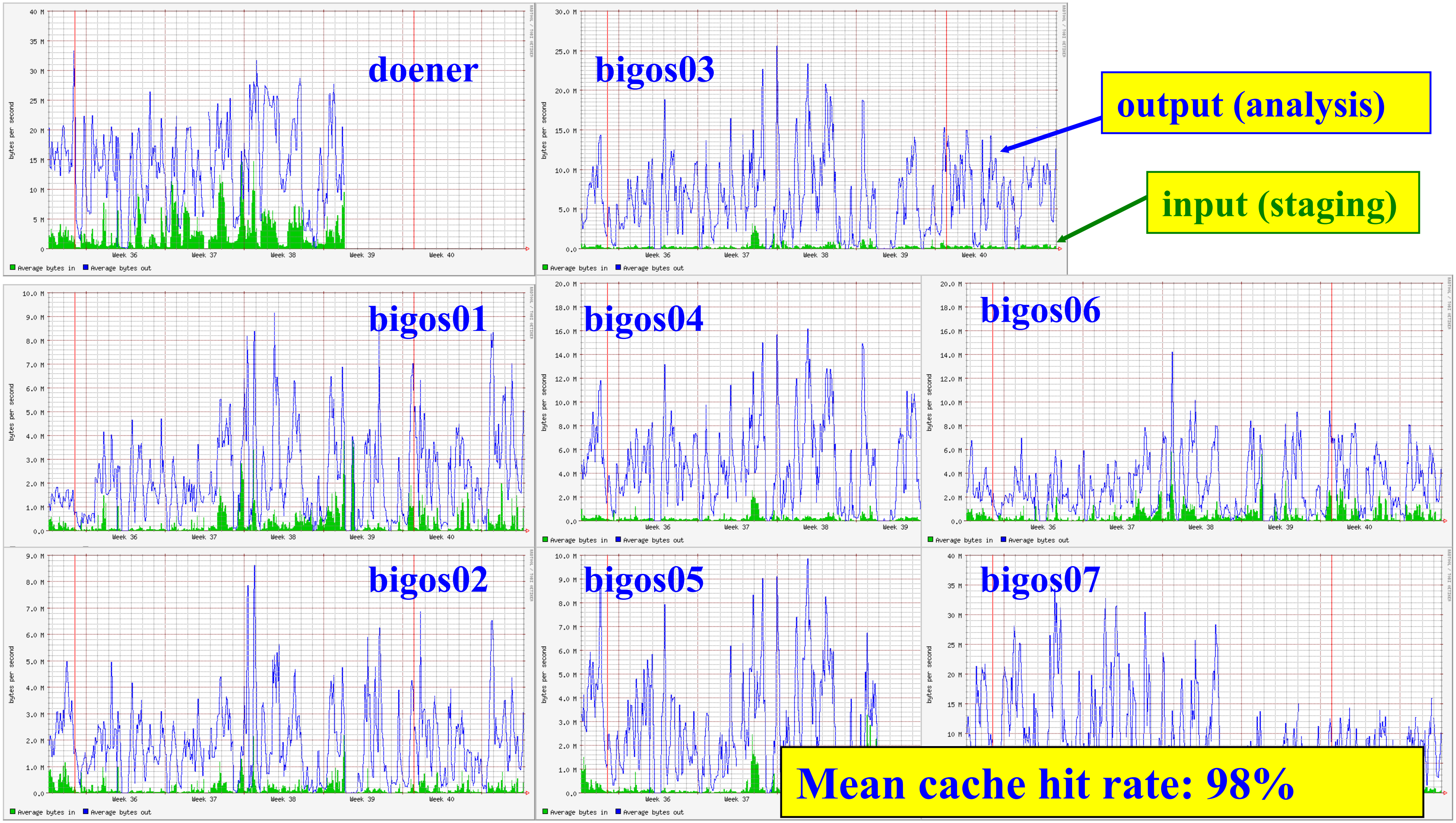
DELFI2, 2002- ?

Commodity File Servers (cont'd)

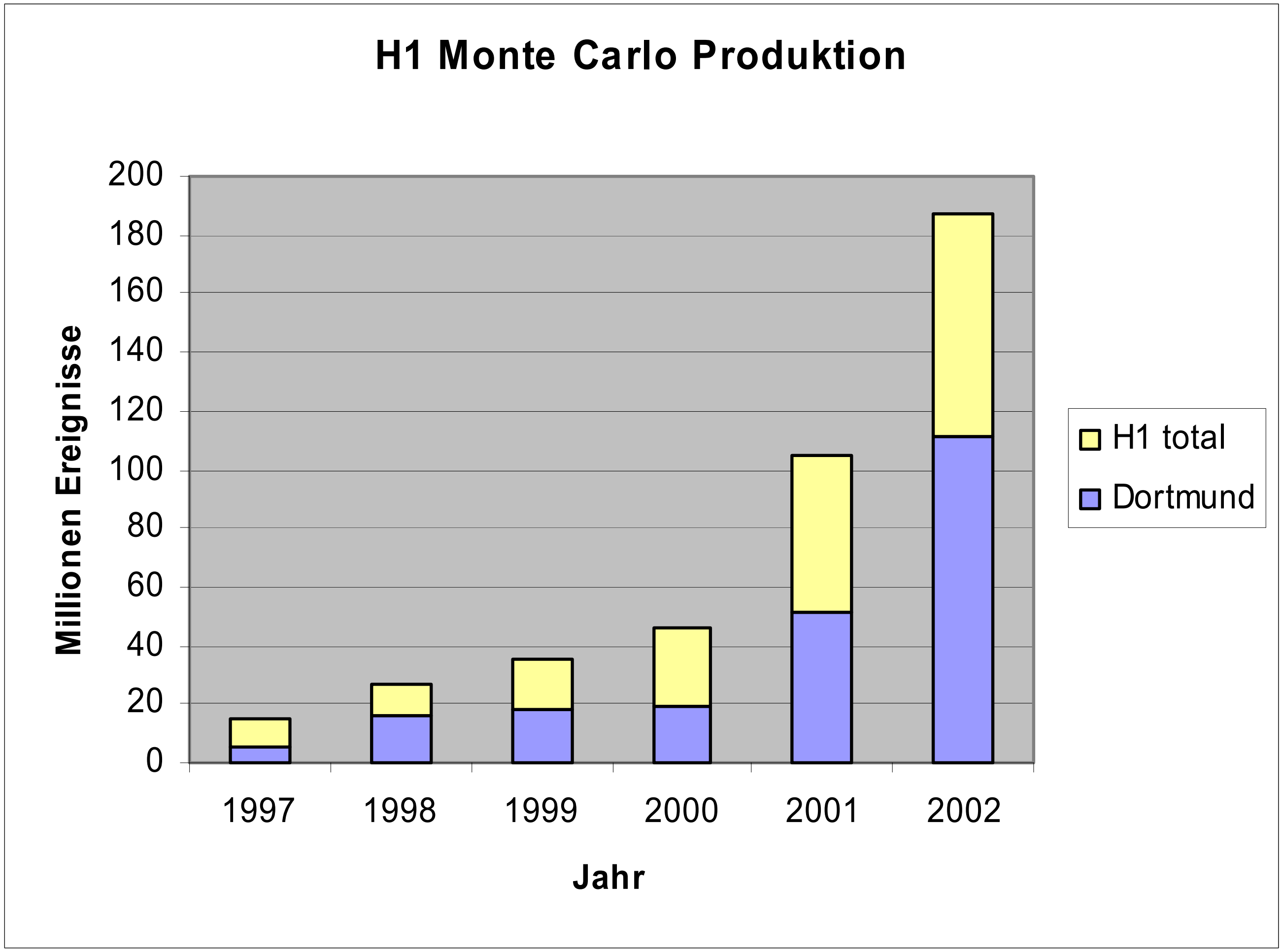
- New dCache file servers hardware
  - 4 Linux front ends connected to 2 SCSI-2-IDE disk arrays
  - SCSI connection to hosts
  - 12 drive bays with 200 MB disks
  - RAID5
  - 2 TB net capacity per array
  - copper gigabit ethernet
- Better modularity than PC-solution with built-in RAID controller
- First system went into production last Monday

ZEUS file server

I/O Performance with Distributed Cache Servers

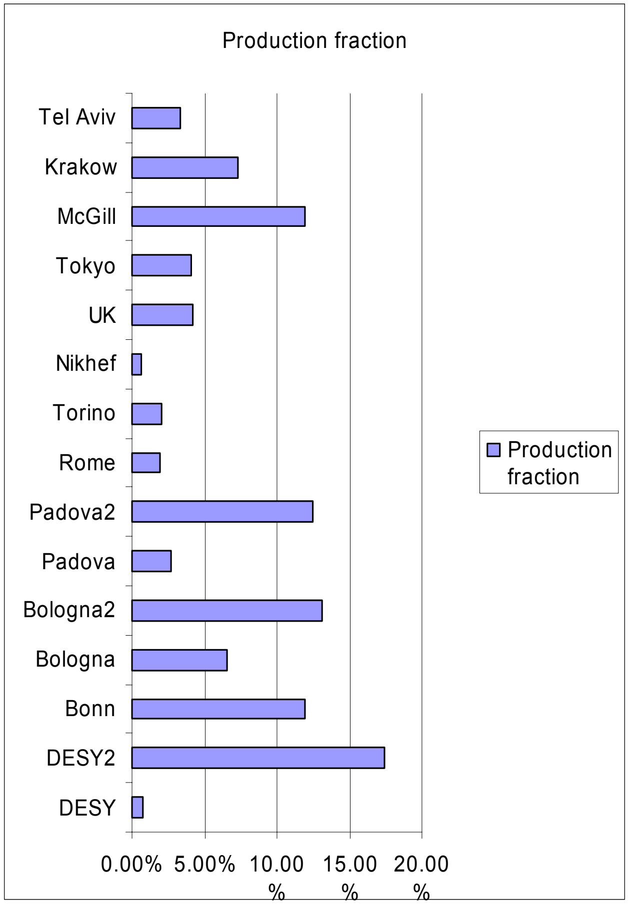


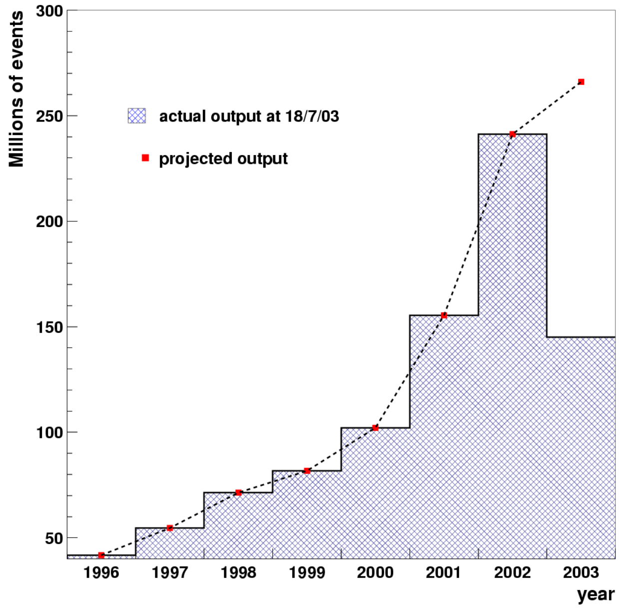
Simulation in H1: Mainly on 2 Sites

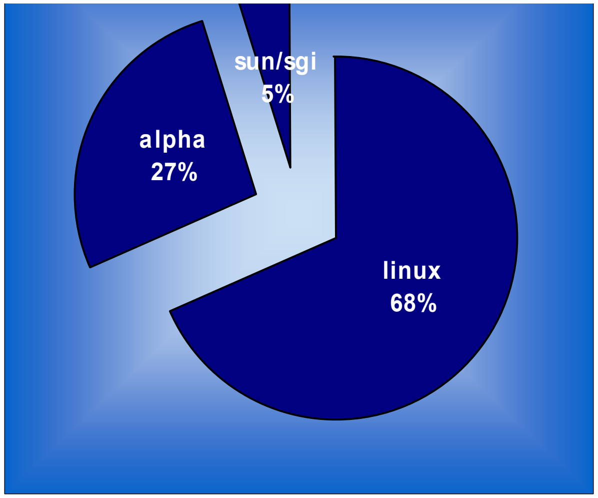


(nach Zahlen von D. Lüke)

Simulation in ZEUS: the Funnel System





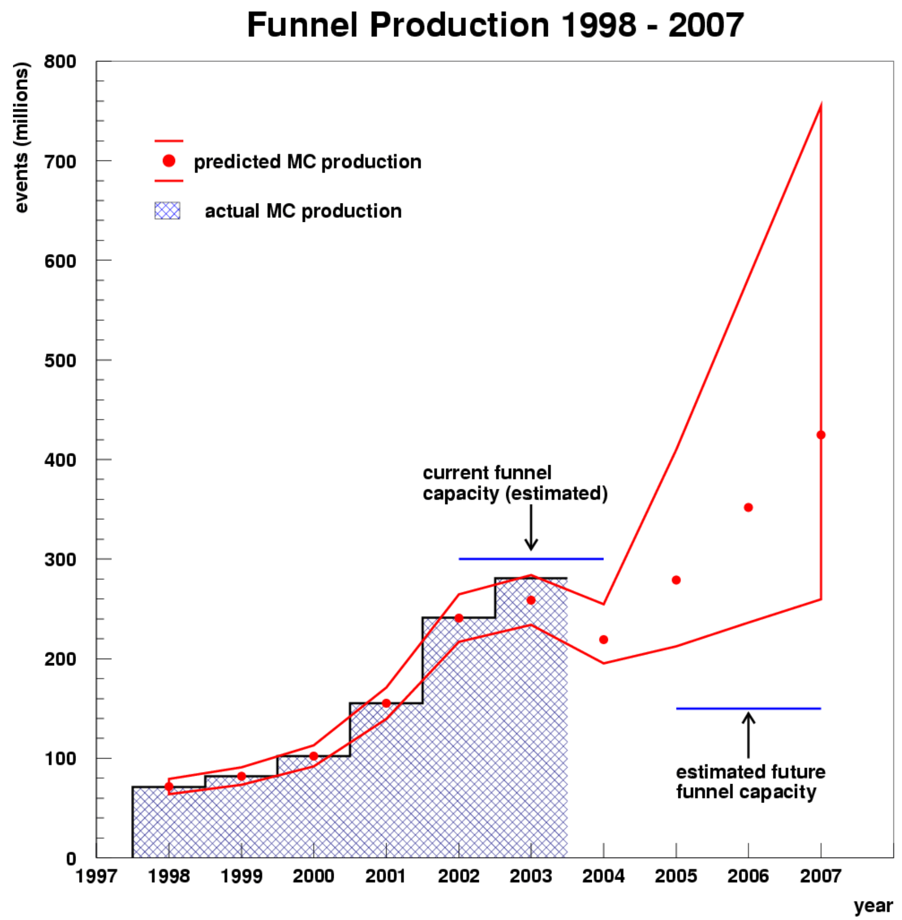


- Automated distributed production
- production reached 240 M events in 2002
- funnel is an early computing grid

Future Need for Monte Carlo

- Modeling is based on our simulation statistics of 1998-2002 in relation to real-data taken
  - on average, 5-10 Monte Carlo events needed for each real-data event
- Additional effects expected from new detector components

➔ considerable increase in simulation demand during HERA-II



ZEUS

Future MC Production Concepts

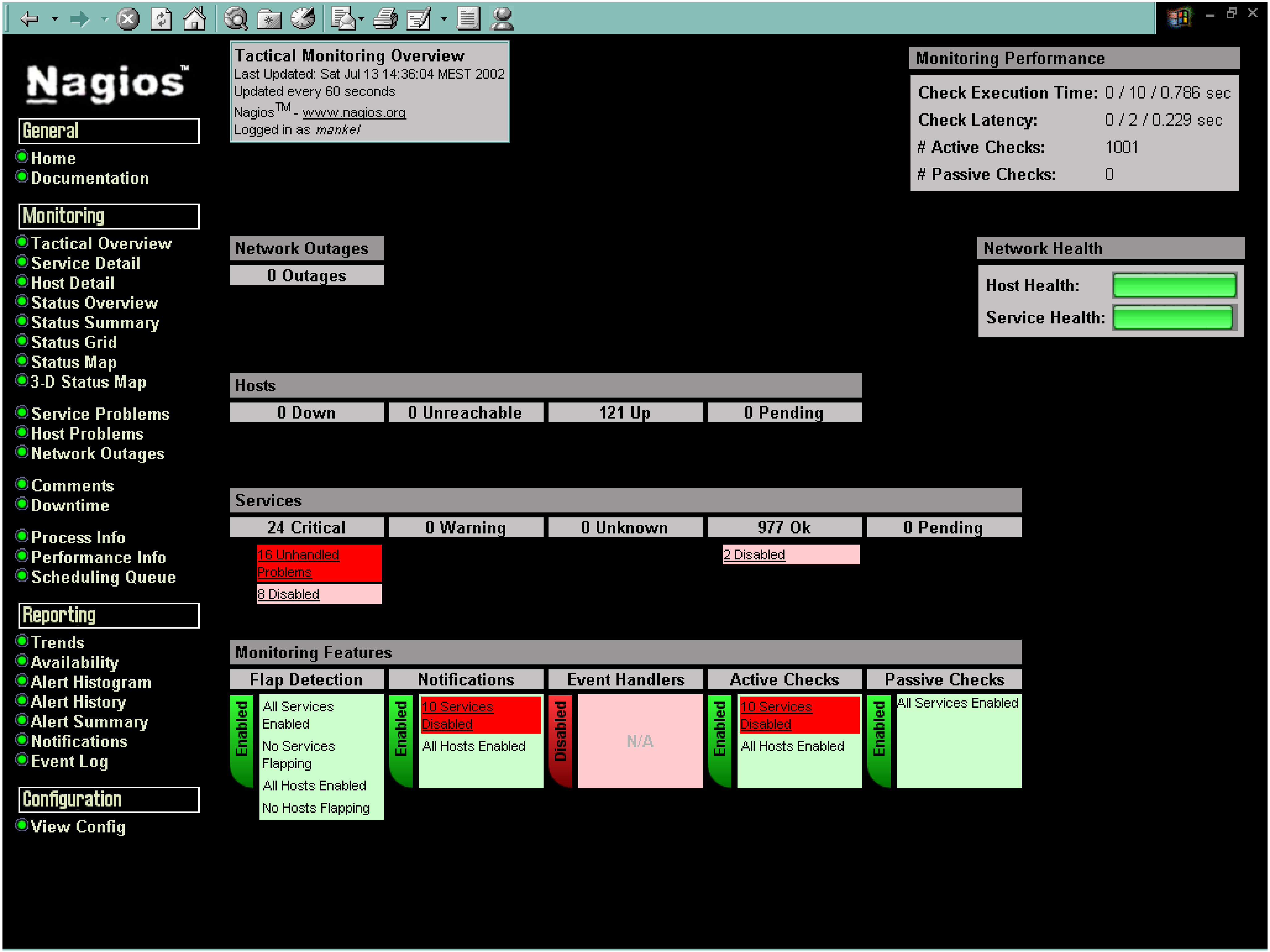
- Both H1 & ZEUS face simulation demands that will not be satiable with their present production concepts

➔ exploration of grid-based production

➔ a grid test bed based on EDG kit is already running at DESY
  - collaboration with Queen Mary college in London (H1)
  - other partners likely to join
  - remote job entry has been demonstrated

Monitoring

- Efficient monitoring is a key for reliable operation of a complex system
- Three independent monitoring systems in ZEUS Computing
  - LSF-embedded monitoring
    - ➔statistics on time each jobs spends in queued/running/system-suspended/user-suspended state
    - ➔quantitative information for queue optimization etc
  - Ganglia
    - I/O traffic and CPU efficiency
    - web interface
    - history
  - NetSaint, now called Nagios
    - availability of various services on various hosts
    - notification (email...)
    - automated trouble-shooting


Nagios
General
Home
Documentation
Monitoring
Tactical Overview
Service Detail
Host Detail
Status Overview
Status Summary
Status Grid
Status Map
3-D Status Map
Service Problems
Host Problems
Network Outages
Comments
Downtime
Process Info
Performance Info
Scheduling Queue
Reporting
Trends
Availability
Alert Histogram
Alert History
Alert Summary
Notifications
Event Log
Configuration
View Config
Tactical Monitoring Overview
Last Updated: Sat Jul 13 14:36:04 MEST 2002
Updated every 60 seconds
NagiosTM - www.nagios.org
Logged in as mankel
Monitoring Performance
Check Execution Time: 0 / 10 / 0.786 sec
Check Latency: 0 / 2 / 0.229 sec
Active Checks: 1001
Passive Checks: 0
Network Outages
0 Outages
Network Health
Host Health:
Service Health:
Hosts
0 Down
0 Unreachable
121 Up
0 Pending
Services
24 Critical
0 Warning
0 Unknown
977 Ok
0 Pending
16 Unhandled Problems
8 Disabled
2 Disabled
Monitoring Features
Flap Detection
Notifications
Event Handlers
Active Checks
Passive Checks
Enabled
All Services Enabled
No Services Flapping
All Hosts Enabled
No Hosts Flapping
Enabled
10 Services Disabled
All Hosts Enabled
Disabled
N/A
Enabled
10 Services Disabled
All Hosts Enabled
Enabled
All Services Enabled

A Glimpse at Software Development

Just a casual glance at two examples...

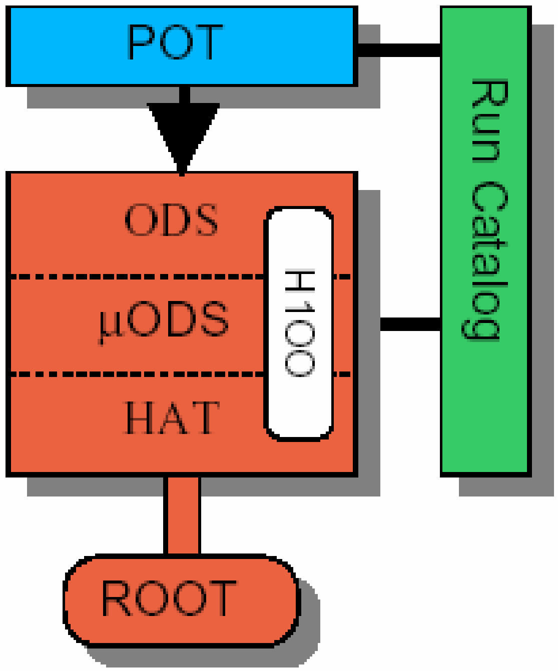
- H1: ROOT-based object-oriented analysis framework
  - new object-oriented data model, based on ROOT trees
  - required redesign of large parts of analysis code
  - in production for normal data analysis since 2002

- ZEUS: ROOT-based client-server event display
  - purely object-oriented event display client
  - retrieve events on demand from central server
  - decided technology after building prototypes with Wired (Java) and ROOT (C++)

H1

Data Storage Model

- Based on ROOT
- Three hierarchical layers
  - Reconstruction: 'ODS'
    (Object Data Storage, 15 kB/evt)
  - Particles: 'μODS' (1.5 kB/evt)
  - Event Tag: 'HAT' (0.5 kB/evt)
- Additional layer: 'User Tree'
  - RunCatalog: Retrieve data by run and event #

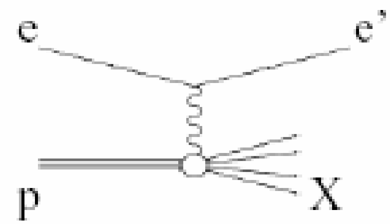


Common environment for both H1 and user data

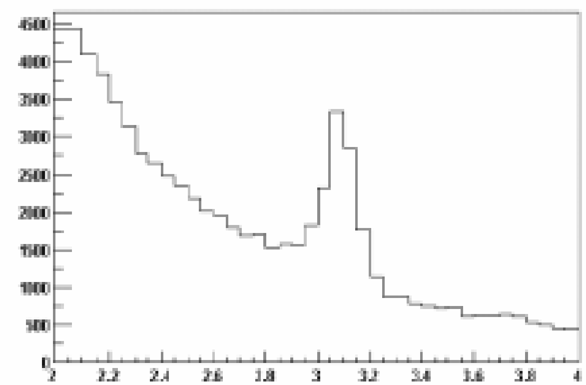
from Andreas Meyer (H1)

H1 Particle Candidates

Persistent storage of physics output on μODS



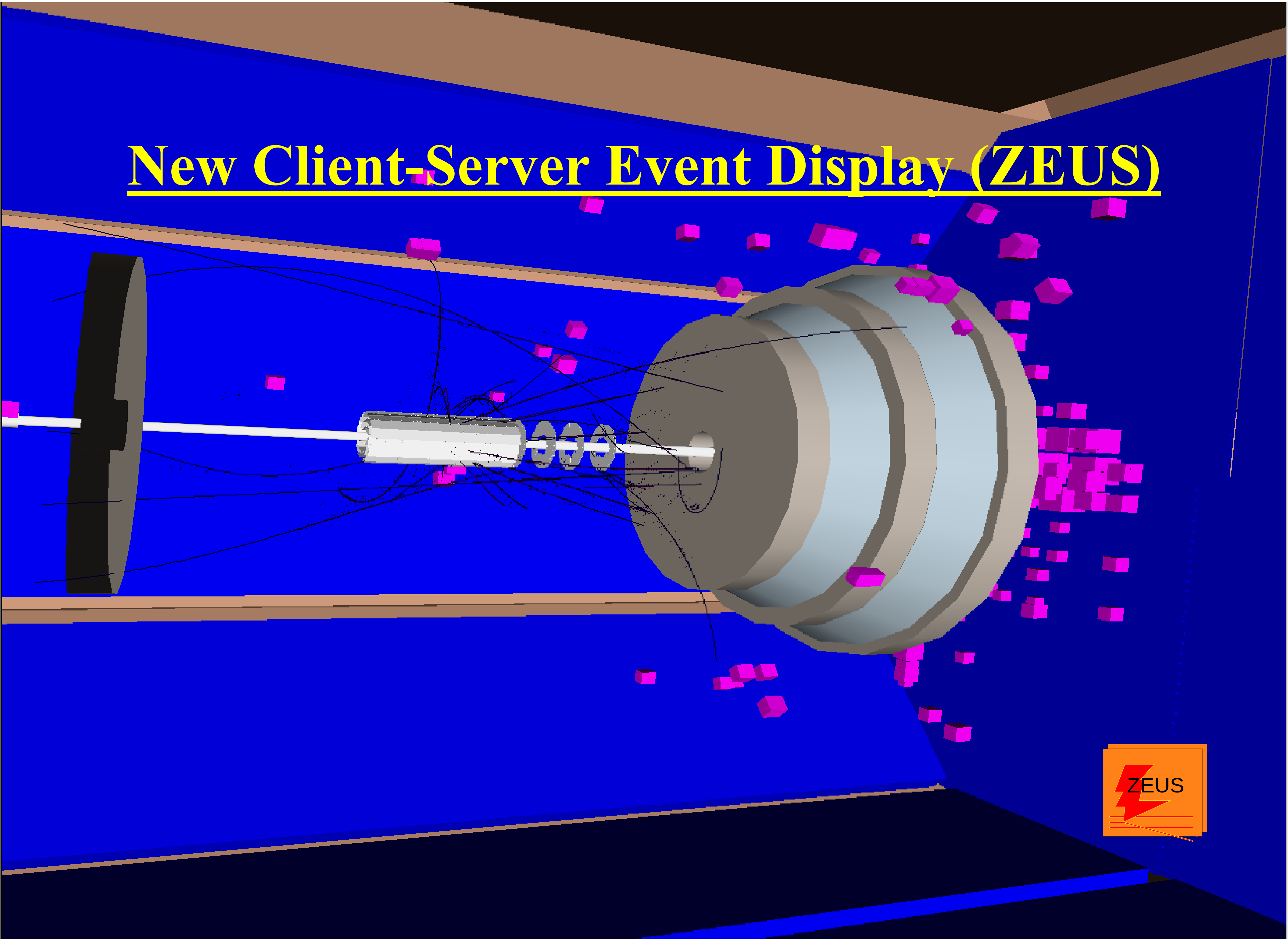
- Charged Particles (Tracks)
- Electrons (incl. scattered electron)
- Muons
- Hadronic Final State Objects
- Jets (K_t, Jade)
- *J/Ψ->ll* / *D*->Kππ*
- … / leading *p* / $\rho^0 \to \pi\pi$ / $\pi^0 \to \gamma\gamma$ / …



Each track and/or cluster object counted only once
with possibly multiple ID hypotheses

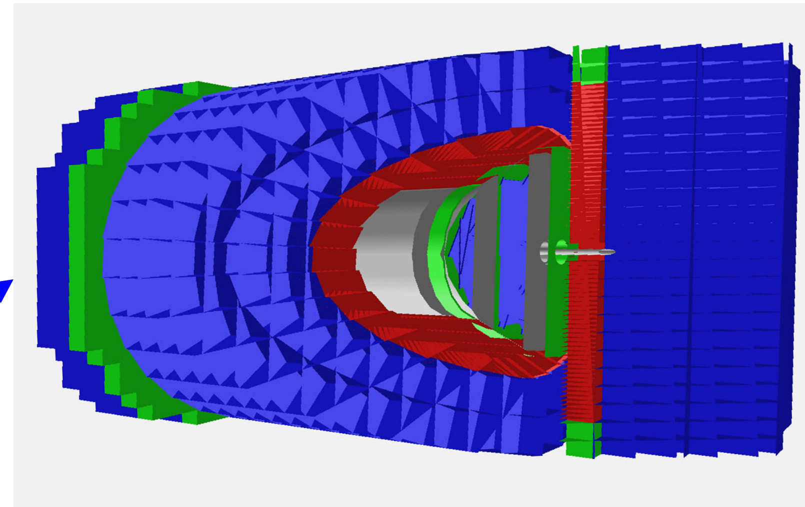
from Andreas Meyer (H1)

New Client-Server Event Display (ZEUS)

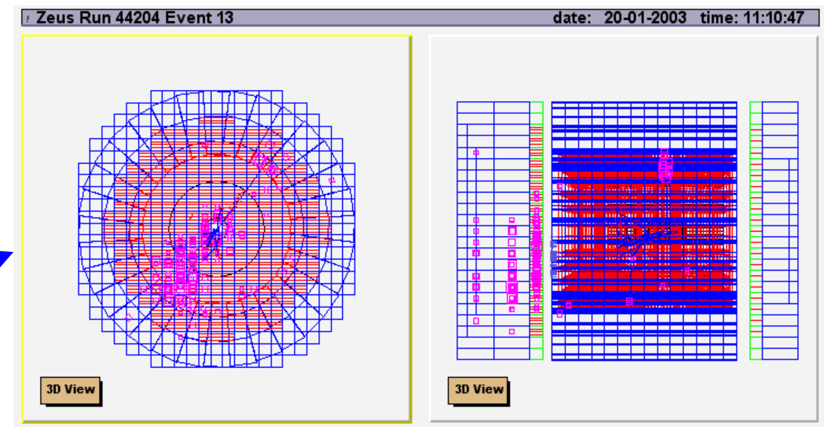


Integration of 2D/3D views

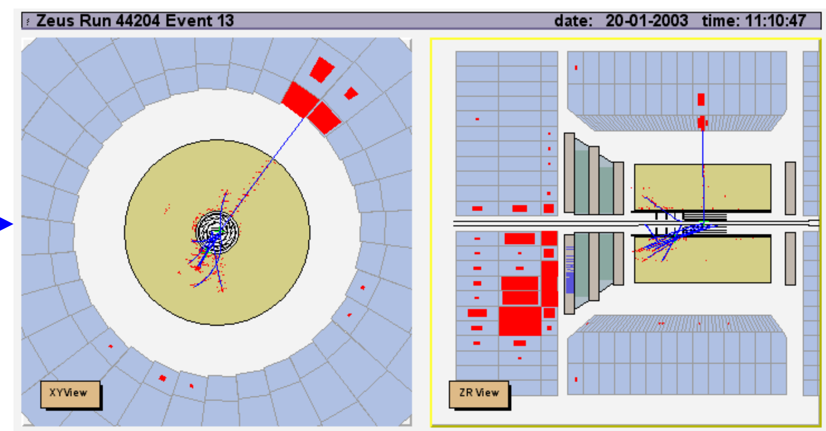
- Perspective views with hidden surface removal are useful for understanding geometry, but do not show the event well



- Straight-forward “projections” of 3D representation can be complex & confusing



- Layered projections with proper ordering allow to really analyze the event

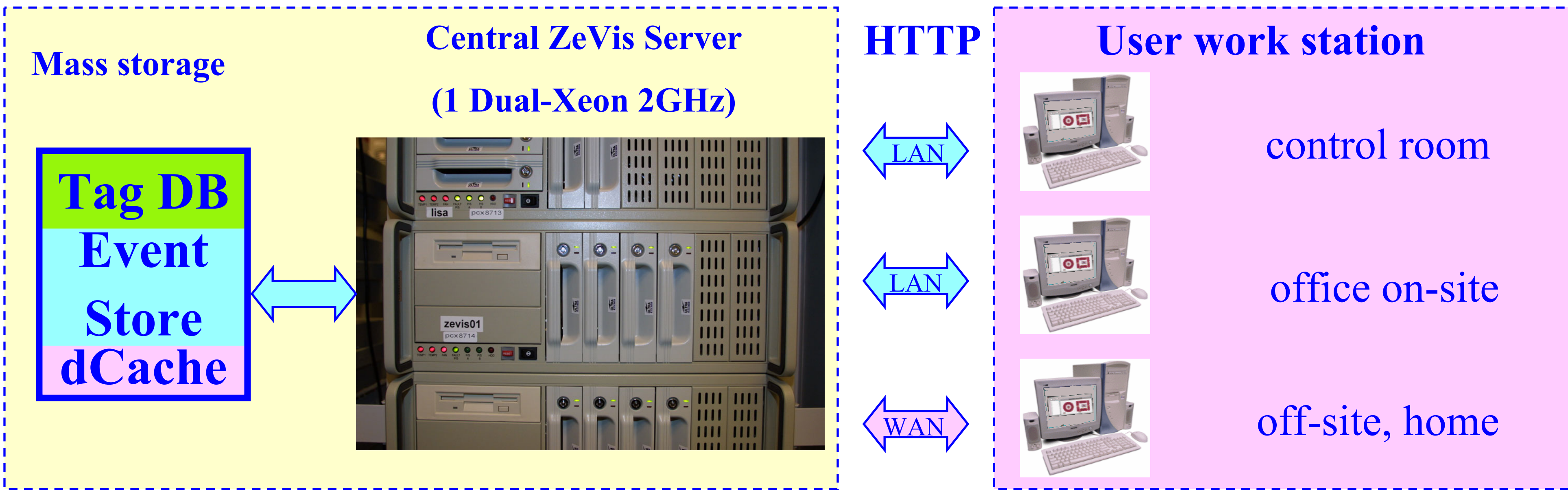


Client-Server Motivation

- Event display needs access to experiment's event store and online system

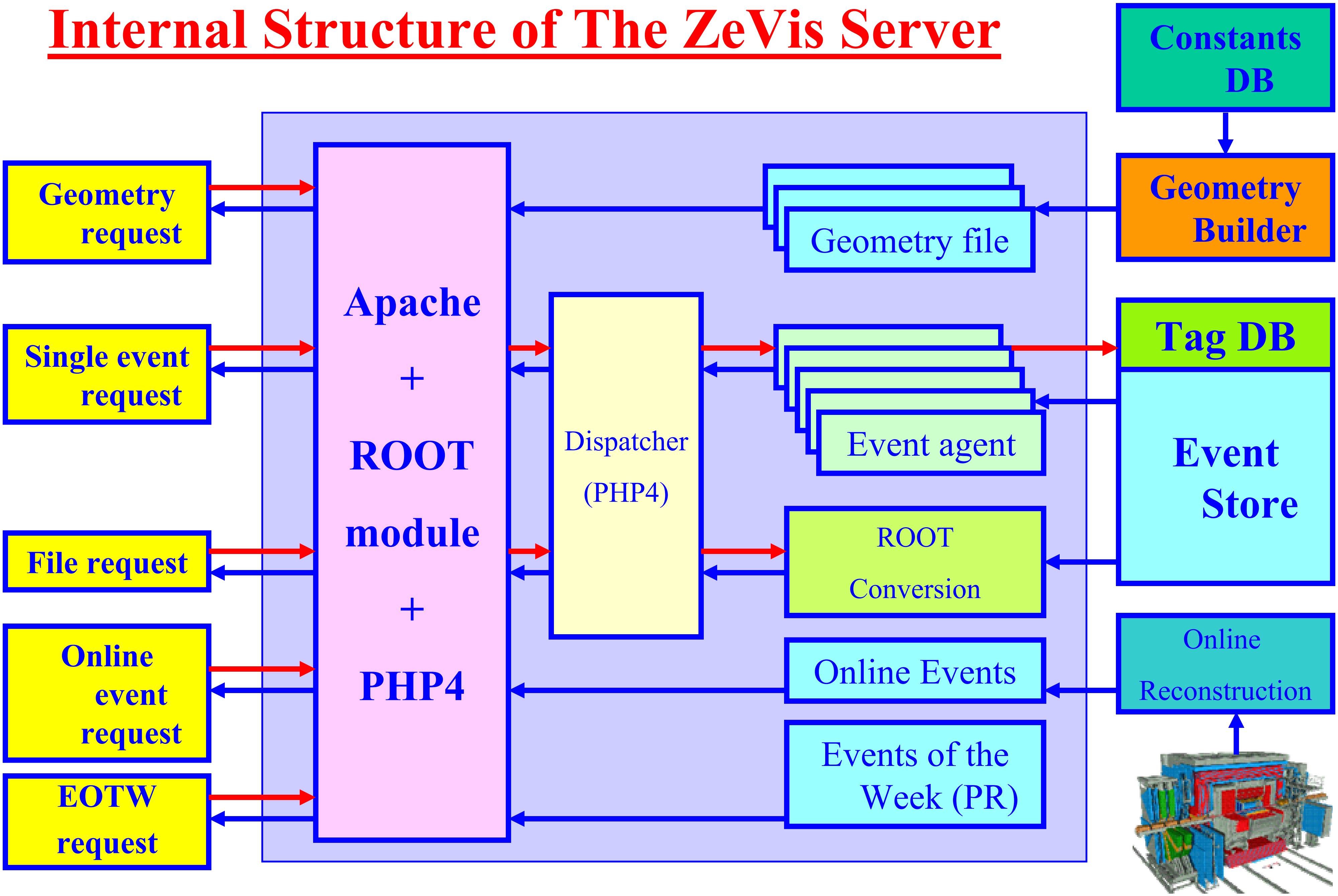
➔ Idea: separation of
  - portable lightweight client, which can run wherever ROOT can run
  - central geometry & event server on DESY site

and connect via HTTP protocol (ROOT's TWebFile) to pass firewall

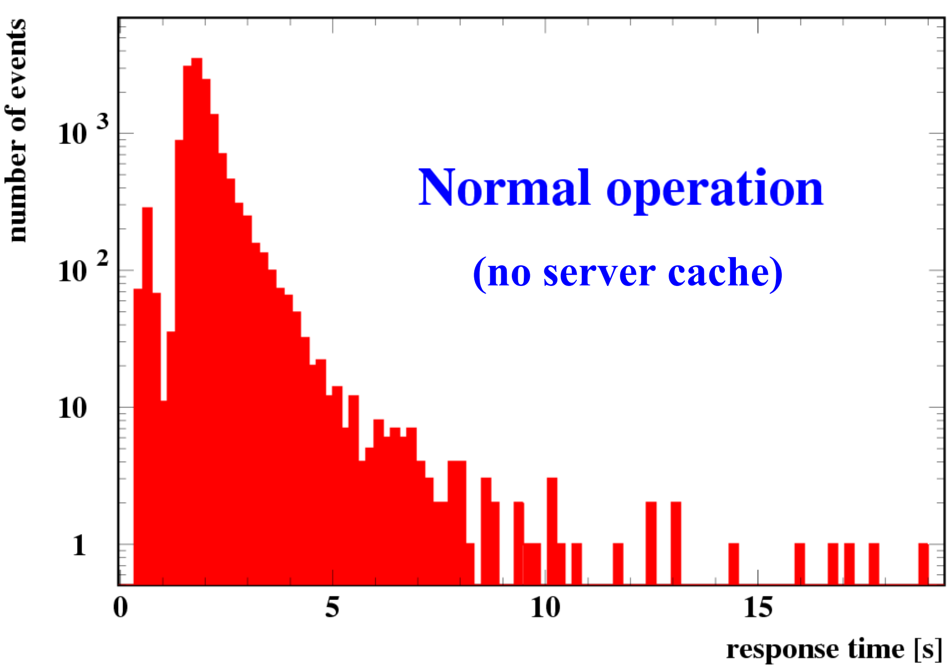


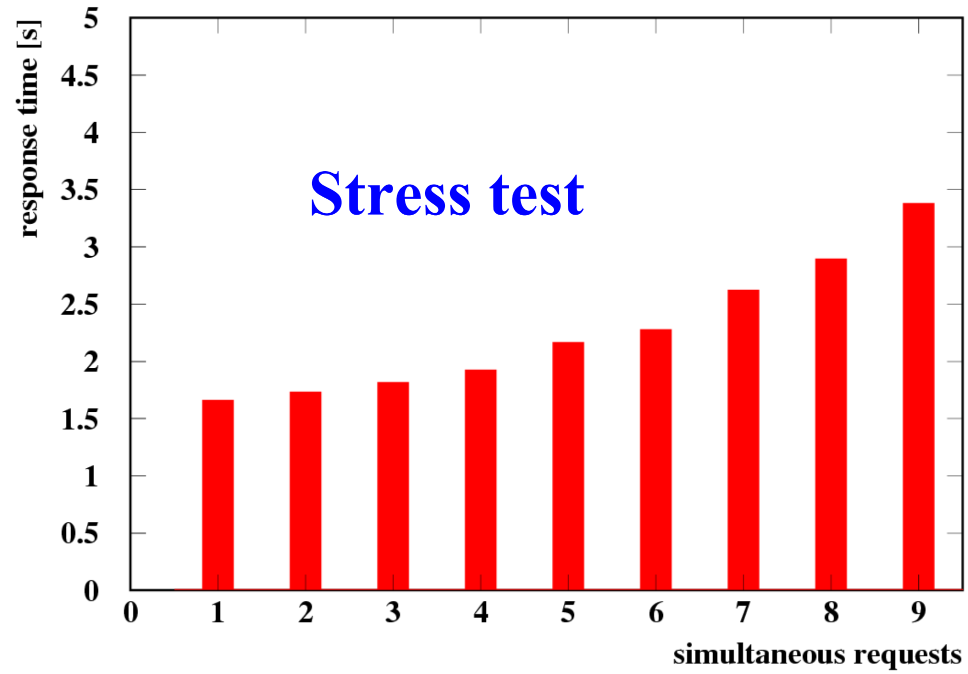
DESY Computer Center

Internal Structure of The ZeVis Server



Server Performance





- Less than 2 s average response time
  - even faster on cached events
- this includes event analysis and conversion to ROOT format, excludes network

➔ In reality, **simultaneous requests will hardly occur at all**

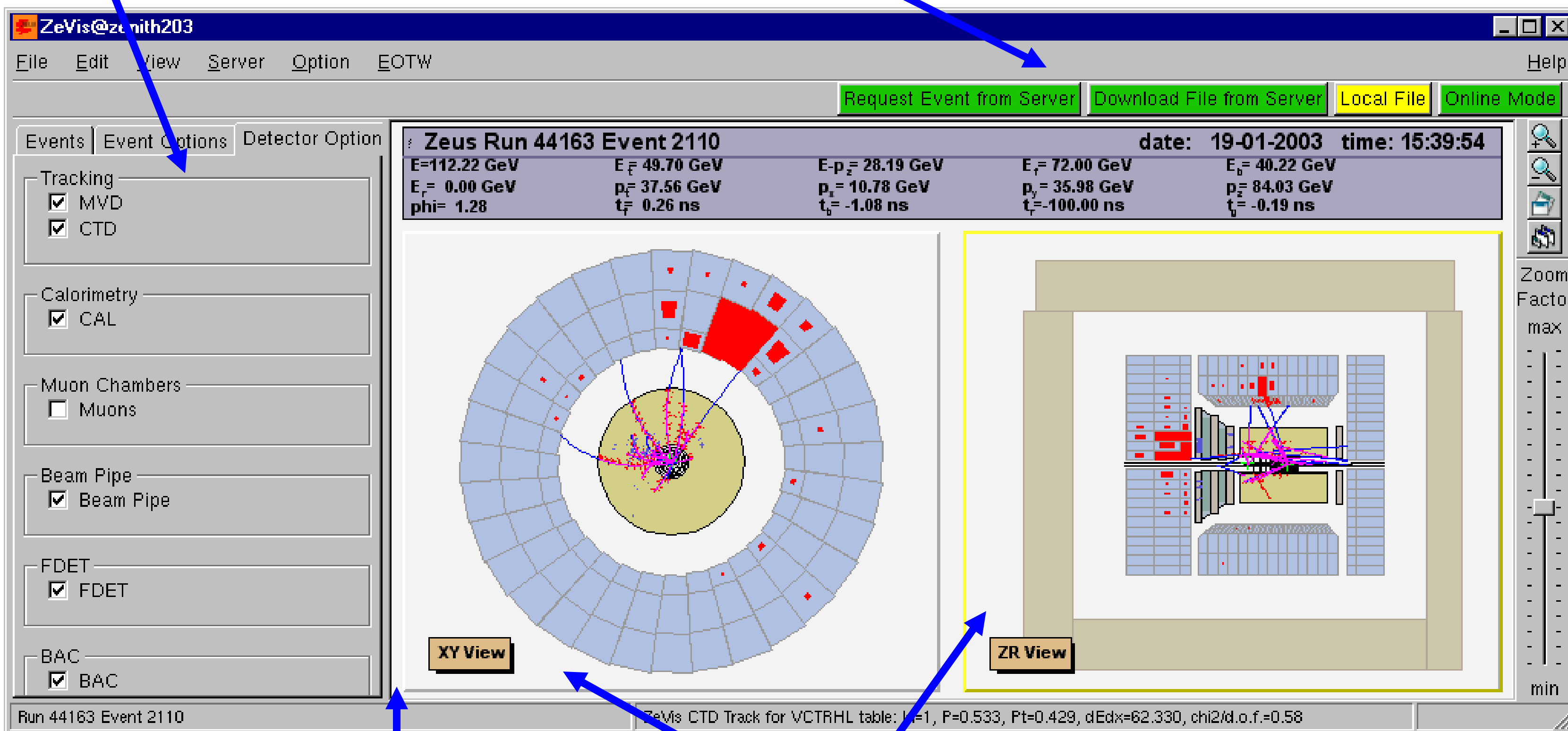
➔ Server can stand up at least **4 requests in parallel** without noticeable performance loss

Graphical User Interface

Option tabs

Input modes

Zoom controls



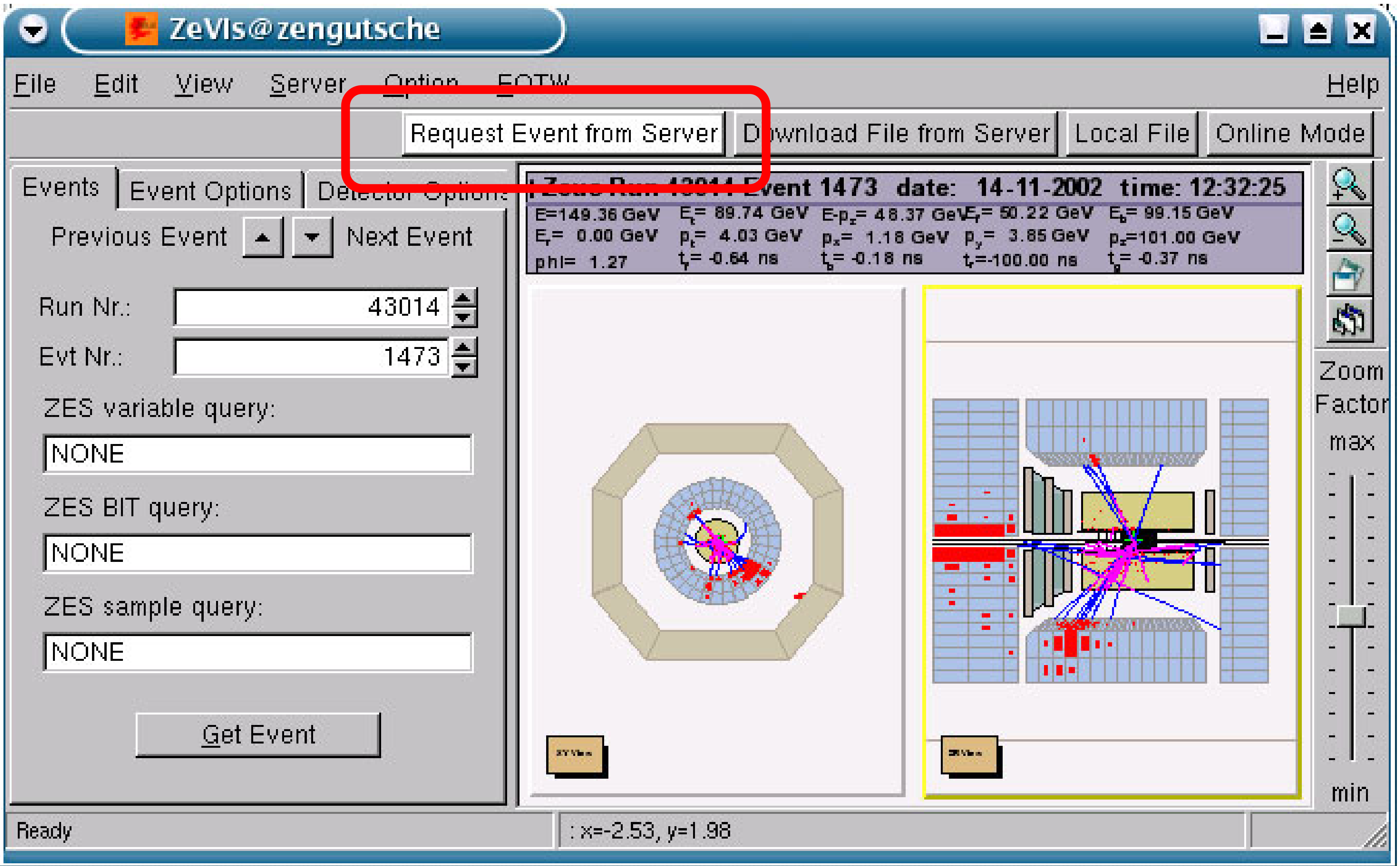
Status information

Canvas

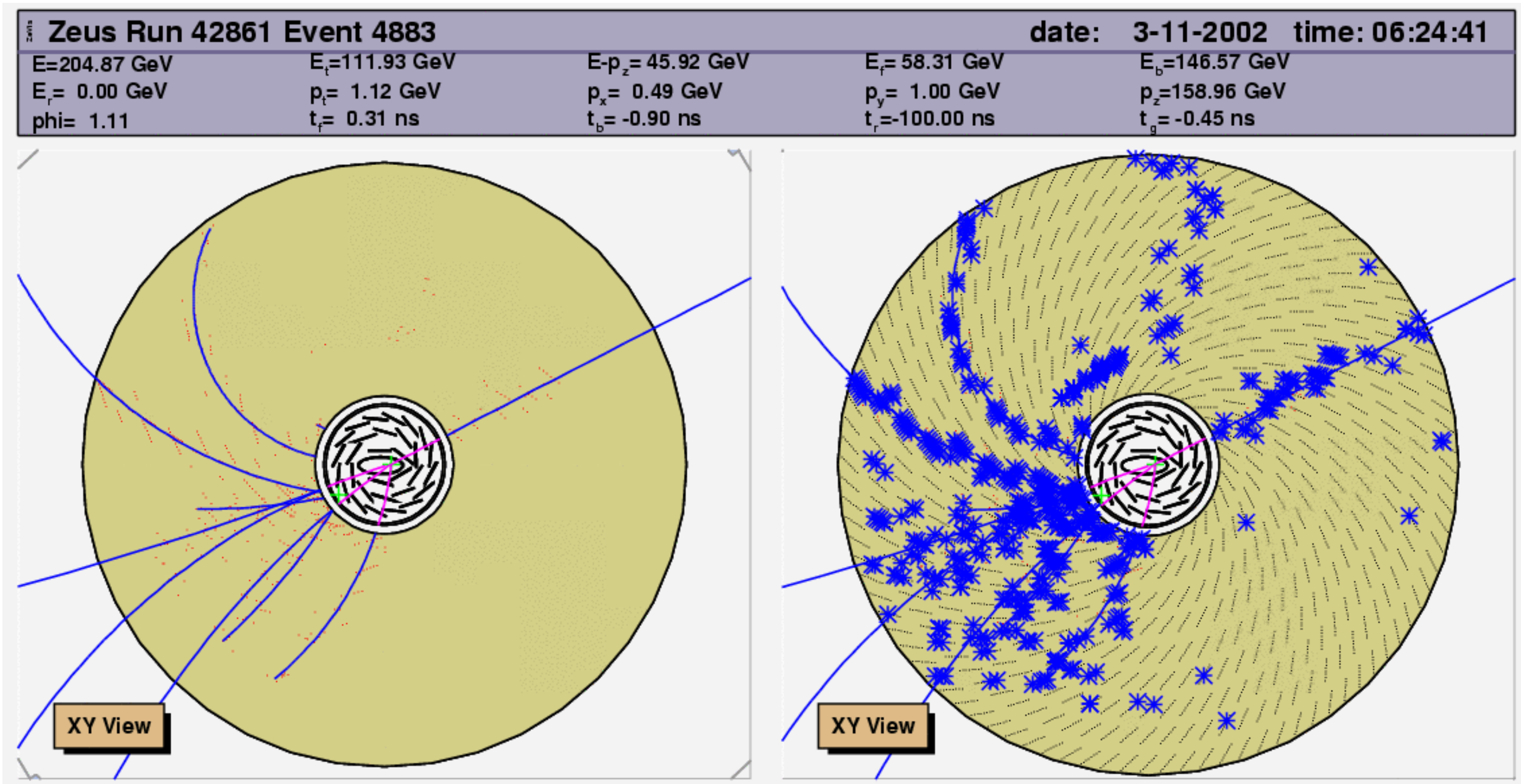
Pads

Object information

GUI: Single Event Server

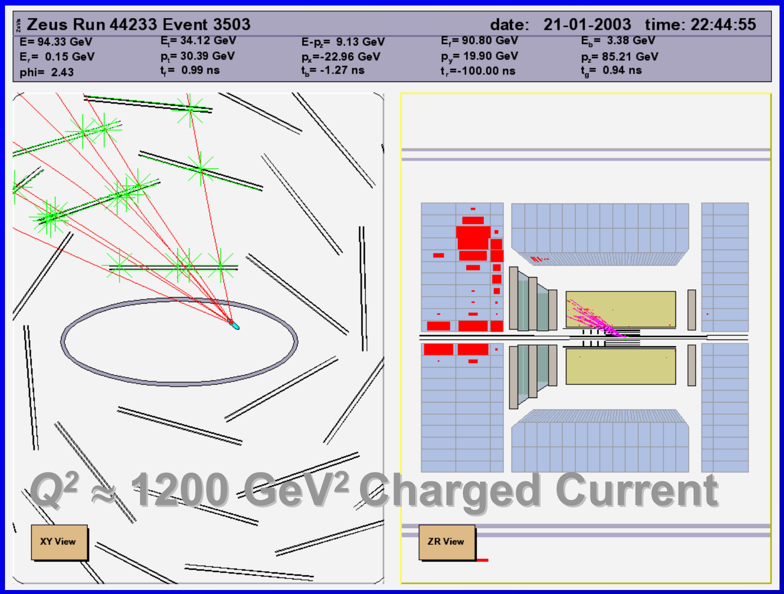


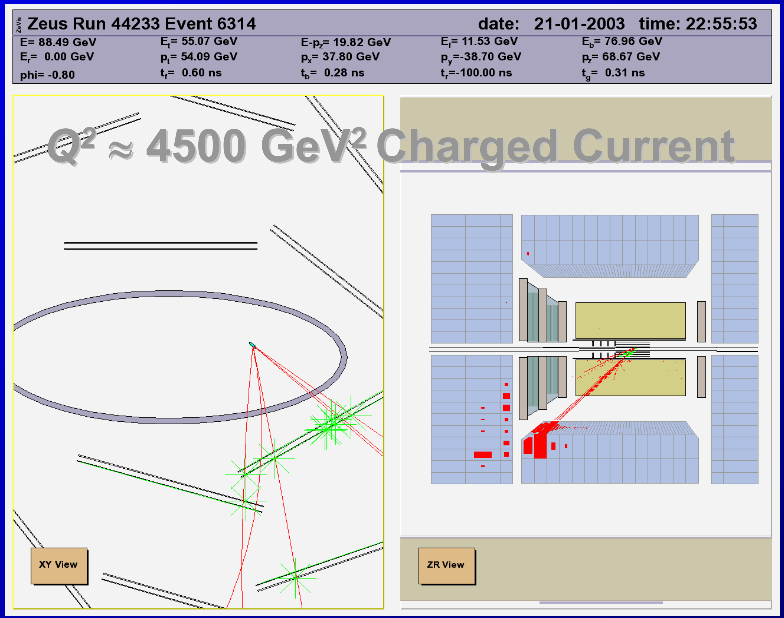
Drift Chamber Raw Hits

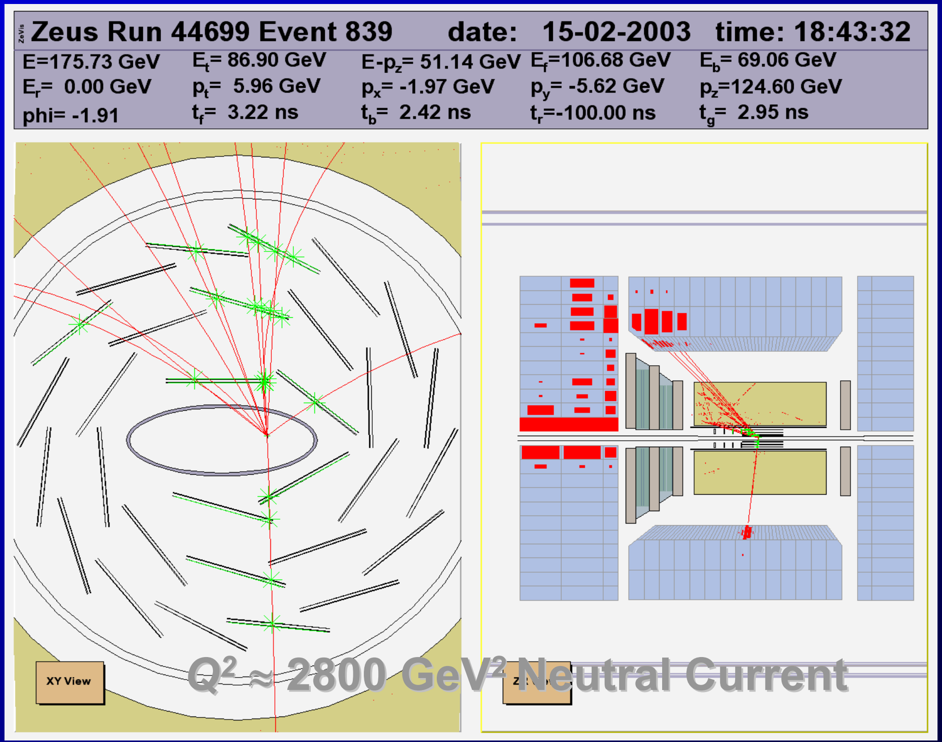


- Normally, CTD display shows assigned hits, taking drift distance into account
- Special mode shows cell structure and raw hits

Visualizing the Micro-Vertex Detector







Summary

- HERA-II poses sizable demands on computing, and they are immediate
  - approaching the Petabyte scale
- Vast increase on standards on turnaround speed and reliability
- Commodity computing gives unprecedented computing power, but requires a dedicated fabric to work reliably
  - redundant farm setups
  - redundant disk technology
- Sophisticated mass storage middleware (dCache) is essential
- New developments in experiments' application software
  - e.g. H1 Root analysis environment, ZEUS client server event display