Computing at DESY — Current Setup, Trends and Strategic Directions

Michael Ernst

DESY Hamburg, Notkestr. 85, 22607 Hamburg

Since the HERA experiments H1 and ZEUS started data taking in '92 the computing environment at DESY has changed dramatically. Running a mainframe centred computing for more than 20 years, DESY switched to a heterogeneous, fully distributed computing environment within only about two years in almost every corner where computing has its applications.
The computing strategy was highly influenced by the needs of the user community. The collaborations are usually limited by current technology and their ever increasing demands is the driving force for central computing to always move close to the technology edge. While DESY's central computing has a multi decade experience in running Central Data Recording/Central Data Processing for HEP experiments, the most challenging task today is to provide for clear and homogeneous concepts in the desktop area. Given that lowest level commodity hardware draws more and more attention, combined with the financial constraints we are facing already today, we quickly need concepts for integrated support of a versatile device which has the potential to move into basically any computing area in HEP.
Though commercial solutions, especially addressing the PC management/support issues, are expected to come to market in the next 2 - 3 years, we need to provide for suitable solutions now. Buying PC's at DESY currently at a rate of about 30/month will otherwise absorb any available manpower in central computing and still will leave hundreds of unhappy people alone. Though certainly not the only region, the desktop issue is one of the most important one where we need **HEP-wide collaboration** to a large extent, and **right now**.
Taking into account that there is traditionally no room for R&D at DESY, collaboration, meaning sharing experience and development resources within the HEP community, is a predominant factor for us.

Key words: mass storage; network, data storage & data access, processor farms, physics data processing;

1 Introduction

Central Computing at DESY has to support a broad spectrum of different user communities having very often very diverse requirements. Concerning our clients one might think immediately of the large HERA experiments, currently three in production (H1, ZEUS, HERMES) and a forth, HERA-B, having very challenging demands, in preparation planning to go 'online' next year. Though these HEP experiments represent a very large fraction of our customers, there is also the many applications in the area of synchrotron radiation experiments, the Accelerator development and operation division, Theory Group and the Administration.
Modern computing concepts require a very high degree in terms of Distributed Services and support for a wide and heterogeneous Hard- and Software spectrum, which is basically impossible without an excellent network infrastructure.

2 Summarizing the topics in Central Computing at DESY

- **Networking**, viewed as the 'Real Backplane' in Distributed Computing. We implemented a 3-level hierarchical network infrastructure ranging from 100 MByte/s links based on HiPPI, mainly serving the demands of the high performance physics data processing and storage applications, and by using Ascend's GigaRouter transparently integrating the middle layer based on FDDI (currently 25 separate rings) connecting shared high performance servers to a broad user community, having their computing environment either directly attached to FDDI or using it as a backbone transport system with routers and bridges transparently translating between FDDI and more than 1000 mostly shared Ethernet segments (10 and 100 Mbps) still making most of the 5000 computer systems network attached.
  Though we implemented a multiprotocol backbone in most areas with routing support for IP, IPX and AppleTalk we are aiming for a limitation to IP whenever possible. With the Internet growing like 'Salmonellae in Tiramisu', this idea is clearly supported very well, leaving all the other protocols in the dust.

- **Data Storage & Data Access:** We in Central Computing decided about 3 years ago that a concept of using huge central fileservers between user processes and the storage repository (disk farms, tape robots, etc) won't scale. Instead we were looking for a system allowing us to make storage devices of any kind network attached using intelligent controllers between them and the network based on commodity workstations running the UNIX operating

system. A prerequisite was the existence of a Storage Management System letting us install the required software component(s) where it's appropriate (e.g. the Storage Server on the machine connecting a tape drive to the network, etc). Part of our initial goals was to let users access the tape inventory the same way they access disk files, meaning they don't need to know 'this is a tape file stored on medium *xyz* kept in robot *abc* and because of that I need to go through procedure *bla* in order to get hold of the data'. Therefore all tape inventory is visible through an ordinary -centrally maintained- filesystem [6]. Internally a database, it is accessed by exporting its contents using standard NFS. The big advantage utilizing NFS for this purpose is obvious: No need for modifications or additional software components on the client side. Making the inventory visible through a standard filesystem also provides for device transparency concerning the data being stored on different storage devices in a multi level storage hierarchy.
Currently installed at DESY is a 'Nearline Capacity' of about 150 TBytes consisting of three different libraries, including four StorageTek ACS 4400 silos (22 Tbytes) with 36 3490 drives, two Ampex DST 800 linear robots (6 TBytes each) with 6 DD-2 helical scan drives and the recently aquired GRAU ABBA/E mixed media library (120 TByts) currently equipped with 10 Ampex DST 312 next generation high performance helical scan drives, the latter handling a cartridge capacity of both 50 GBytes and 150 GBytes (could also handle 330 GBytes/cartridge, however, the large volumes are not supported by the GRAU robot).

– **Physics Data Processing:** Originally initiated to serve the needs of the 'quasi-online reconstruction' DESY started to aquire Symmetric Multi Processor (SMP) machines from SiliconGraphics in 1991. The systems turned out to run very successful for both H1 and ZEUS, successful both in terms of performance and management overhead, the latter being a very important aspect for Central Computing. During '92 we became aware of the excellent I/O (network, disk access) performance of the systems leading to the conclusion that they were potentially the next generation servers for CPU and I/O intensive physics analysis purposes. Hence DESY decided to buy more SGI systems, leading to a present inventory of 11 Challenge XL (compute servers) and 3 Challenge DM (data servers) with a total of 180 CPUs (mixed R4400 150, 200, 250 MHz and R10000s running at 190 MHz). Common to all three HERA experiments being in production is the concept of keeping the annual DST data online, meaning storing them on random access devices (disks) in order to allow concurrent access for about 100 analysis jobs per experiment at any given time. Currently installed is a disk storage capacity of 3 TBytes.

– **Central Data Recording:** DESY has a multi decade history in using Central Data Recording. All major HEP experiments being currently carried out at DESY are running the concept of connecting their third or forth level trigger close to the detector via fibre links to the data logging facilities in the computer centre, with the exception of HERMES sending the RAW data directly from the pit to the central tape archive. By running Central Data Recording, clearly the problem of getting access to the most recent, most interesting data is avoided. Not only for reconstruction, but also for the purpose of an immediately needed detector study, the data is available from the automated tape archive to concurrent processes shortly after it was taken.

- **Interactive Physics Computing:** Interactive services for the physics community are mainly based on Work Group Servers (WGS), usually UNIX boxes from HP, SGI, Sun and IBM. In order to provide the same 'look&feel' (a homogeneous and uniform environment) to the users, we are running a highly AFS (Andrew File System) centric model. All standard applications and librariers, including editors, mailtools, Cernlibs, PAW, etc are stored in AFS volumes maintained on the central binary servers with those volumes being replicated on different servers. Also the users' home directories are kept on a central home directory server with the benifit that if a particular WGS fails, the user just switches to another WGS supplying the same environment. Just to be clear, AFS provides just a fundamental functionality, allowing new workstations to get part of the existing cell and therefore the existing homogeneous (file-) namespace, extremely easily. However, that doesn't free us from preparing all the applications for the various computing platforms, incl. possible concurrent versions of the OSs, and even more important, from providing for a common HEP environment. Wherever on the DESY campus and whatever the physicist decides to use, the ultimate goal is to make the interface between him and the machine as uniform as possible. This has been very well achieved for desktop systems based on X-terminals. We are running a fully centralized boot and configuration management, allowing a new terminal to be used within a couple of minutes following the physical installation. This turned out to be extremely efficient. Less than 1 FTE is required to support about 1000 X-terminals of different kind, including boot, font and all the rest of the X-environment. As there is currently a clear trend towards more versatile devices, e.g PCs and Network Computers (NCs) in the future, revised desktop concepts and support schemes are needed (see chapter 'Hot Topics').

- **Centre for Parallel Computing:** Though there were a lot of activities throughout the HEP community evaluating the use of massive parallel machines, especially in the area of Physics Analysis, basically all the results led to the conclusion that based on the amount of communication needed between processors and a) network and b) storage devices those architectures are neither performing better than workstation clusters or SMPs nor are they cheaper (market niche).
  However, those arguments don't necessarily apply to Theory Computing.

Especially in the area of LQCD applications massive parallel machines are very well suited. Already at CHEP'92 INFN presented a scalable architecture allowing to build machines providing for tens of GFlops at outstanding low cost. No way in getting the same price/performance level with 'general purpose' machines. However, the drawback is that those machines are just performing well in this small application segment, with anything else very likely to fail. On the other hand, as the Theorists never get enough CPU power, all the available resources we have are heavily utilized, up to the impressing figure of 93% of the available cycles devoted to physics from January to November '96. There are currently 2 machines (768 nodes, 50 MFlops each) installed at DESY-Zeuthen, delivering a peak performance of 37 GFlops. Based on an existing contract with QSW, the manufacturer, a staggered upgrade over the next two years yields at a performance level of 320 GFlops.
The installation is part of the German National Supercomputing Initiative and is used by Theorists all over Germany.

3 Hot Topics

- **Desktop Computing:** As described in the previous chapter, recent developments on soft- and hardware on Intel processor based PCs is more than likely to change the entire desktop world dramatically. Perfomance comparable to that of factor 3-5 more expensive RISC processors, mature operating systems (Windows/NT, Linux) and an enormous spectrum of commercial software is raising the question whether rather limited devices like X-terminals are still the right way to go. As industry is currently not prepared to deliver solutions especially in the area of efficient enterprise-wide PC-management, which is only one of many, however probably the most important open issue, the question whether to go for X-terminals or PCs is hard to answer right now. DESY imported NICE from CERN a couple of years ago. The concept is appreciated in many groups very much. By using NICE we are able to support hundreds of *Windows* users easily at the expense of only 2 FTEs and a few inexpensive servers. However, the limited environment, closing out the individual from simply adding what he needs apart from the standard, is forcing more and more people (at DESY especially the Accelerator Development and Operation Groups) to jointly develop a more flexible concept based on Windows/NT [1]. We are currently in the process of finding out whether it's possible to integrate NICE/NT into the existing -single domain- NT infrastructure and thereby maintaining the users' freedom to add and/or modify things according to their needs. There are however serious doubts whether we will ever (meaning 'til the end of the decade) succeed in getting to a single homogeneous solution. Rather than that we feel that there is currently a need for at least two different

PC-based desktop architectures

- **Physicist Working Environment:** 90% Physics and 10% typical office applications likely to require UNIX (Linux, Solaris, etc) in addition to Windows/NT
- **Office dominated Desktop** based on Windows/NT

– **Data Access and Data Storage:** Because of the highly modified Open Storage Manager (OSM) [3], basically removing all the identified deficiencies from the original product, combined with the recent hardware aquisition consisting of the GRAU equipped with 10 Ampex drives, DESY is presently in pretty good shape in this area. The GRAU/Ampex combination is adding 120 TBytes of 'Nearline' capacity with an 'Effective Bandwidth' of about 20MBytes/s in and out of the archive. Our definition of 'Effective Bandwidth' is the sustained bandwidth accessing 200 MByte files at any given time with all the overhead concerning tape mounts, positioning, dismounts, threading (required on helical scan drives only), etc included. Though the streaming rate of recent tape drive technology is often relatively high, very inefficient mechanical and logical media handling is the reason for reduced 'Effective Bandwidth' to a quarter or even less of the streaming rate. When randomly accessing relatively small files on a 50 GByte cartridge out of a multi 10 TByte archive the setup times are highly influenced by factors like the cassette design, in particular whether single or dual reel. Dual reel allows mounts/dismounts wherever the medium currently is, while single reel always requires the medium to be rewound to Begin Of Tape (BOT).
With the GRAU/Ampex configuration we created solid ground on the physical storage side, giving us sufficient storage capacity up to the end of the decade. However, we expect to see bandwidth limitations as soon as HERA-B goes online by mid '98. Based upon preliminary calculations we will at least have to double, better triple the bandwidth to a level of at least 55 MBytes/s. This won't be achievable by just applying brute force, meaning buying as many tape drives as necessary to get to 55 MBytes/s. Much more reasonable is to extend the design by a Disk Layer [5] in between the tape drives and the network. That will allow us to batch a number of small (200 MByte) files to a large chunk copied out to tape in a single action and therefore minimizing the overhead. For reading we will simply start with reading ahead the next file following the requested one.
All the things described above are fairly simple compared to a concept dealing with the data access & data management demands for the next decade. We decided to stop any further development on the OSM in favour of spending our time investigating both existing -hopefully commercially available- systems or alternatively looking into building our own based upon the huge amount of expertise we built up during the past 3 years. Returning to commercial solutions, Mass Storage Management Systems at the scale we need for the next millenium are having an extremely small market niche. Therefore there is currently only a single -semi commercial- product available,

called HPPS (High Performance Storage System) [4], developed by a consortium including a number of National Labs in the US centred around IBM Government Systems providing for coordination of the development efforts and marketing. We carefully looked into the design and capabilities of HPSS and found that HPSS is

· **well designed** in terms of basic architecture and modularity
· **lacking functionality** which is essential for HEP. Missing is the capability to schedule storage requests based on priorities (Data Logging) and integrated user authentication (a storage request has to be processed according to the user's permissions, however, HPPS is not considering the user's attributes. Therefore an external authentication system is required).
· **IBM-Only** on the most critical, most complex part: The Server. Though there are preliminary statements concerning an open server implementation, it is fully unclear to us when the code for other platforms will show up.
· based on **proprietary software components** like Transarc's Encina which is a mandatory prerequisite for HPSS. Encina is offering reliable RPCs. This is from our point of view another very critical dependency, because they rely on the non-predictable future of an extremely weak company (though it's owned by IBM). This is one more conflict with the idea of openness.

According to the published development schedule it is absolutely unclear, whether those issues will ever be addressed by the developers. It is also questionable whether the HEP Labs which have currently committed themselves to HPSS (Fermilab, SLAC, CERN) will provide enough 'critical mass' to convince the consortium in giving HEP related issues a high priority on their long list.
Therefore DESY decided to move to observe HEP related HPSS activities in the Labs mentioned above. Part of our interest in this context is the currently discussed interface between HPSS and Objectivity, the latter a Federated Object Oriented Database, very likely to be the first promising candidate for event stores based on the OO methodology (CERN RD45 [2]).

– **Processor Farms:** As described earlier, Physics Data Processing at DESY is built around large SMPs from Silicon Graphics today. Mainly the CPU and I/O intensive applications, e.g. Physics Analysis and Reconstruction, but also (a decreasing amount of) interactive load is seen as the traditional workload on those machines.
Given the recent developments on commodity computing devices, we see them drawing a lot attraction to the collaborations facing a very tight budget situation. Uni and even better multi processor Pentium Pro systems are currently having a price/performance advantage of 3-5 over RISC systems. ZEUS and HERMES decided in '96 to migrate the reconstruction, which is the most obvious application for a PC farm architecture, away from their SMPs to a collection of 20 and 10 PCs respectively. Both experiments live

in a highly UNIX dominated environment. While HERMES decided to continue using Linux as a UNIX-like operating system on their machines, ZEUS decided to port their reconstruction code to NT. Main reason for the porting effort was because ZEUS sees a large potential for commercial support on Windows/NT both in terms of software components and hardware variety. By gaining experience due to porting the reconstruction code, ZEUS expects to find solid ground for PC farms in other areas including interactive and batch analysis.
The result of those -both collaboration initiated- projects is that

· **HERMES** [**7**] ported their software tree (600.000 lines: 50% C code and 40% Fortran code, rest are shell and PinK (tcl/tk) scripts) in two men-weeks to Linux, while
· **ZEUS** [**8**] spent more than 6 months porting their 1 million line Fortran code and other required assets to Windows/NT. Having only very limited manpower available, ZEUS finally got stuck because of ADAMO problems. Clearly, those issues will be sorted out in the future, however, in order to solve the present resource problem, ZEUS decided to migrate to Linux as well, hoping to get their farm into production real soon.

To conclude on this item, it is our belief that currently Linux has particularly in the area of physics applications a clear advantage over Windows/NT. Hard to predict if and when this will change and what the importance of the OS is anyway. According to our experience Linux runs as stable and as efficient as Windows/NT, and also the software industry has started to realize that there is a market for their products on the Linux platform.

4 Organizational Structure

Rather than having a dedicated person acting as the division leader, we in Central Computing proposed to the DESY directorate to form a steering committee with members from the various sections representing their domain. A spokesperson, being nominated by the directorate for a period of three years, reports to the director of central services and takes responsibility for personnel.

5 Summary

With the current computing, storage and network infrastructure in place, Central Computing at DESY serves the needs of the present HERA experiments very well. However, because of the ever increasing demands in both interactive and batch computing, we have to be prepared to provide for solutions to integrate low level commodity computing products transparently into the existing architecture.

References

[1] Till Poser, et al., Windows/NT at DESY, submitted to this conference

[2] A Persistent Object Manager for HEP - An R&D Project including collaborators from CERN, LHC and other experiments (wwwinfo.cern.ch.asd/cernlib/rd45)

[3] Lachman Technology, Inc., The Open Storage Manager

[4] The High Performance Storage System, see http://www.sdsc.edu/hpss/

[5] Susanne Brand, A Distributed Disk Layer for Mass Storage at DESY, submitted to this conference

[6] Patrick Fuhrmann, A Perfectly Normal Namespace for the DESY Open Storage Manager, submitted to this conference

[7] Wolfgang Wander, Marc-André Funk, The HERMES Data Production PC Farm, submitted to this conference

[8] Tobias Haas, et al., Use of Commodity PCs for Reconstruction and Analysis in ZEUS, submitted to this conference