ALEPH 88– 13
NOTE 88-3
J.J. Aubert et al.
13.2.1988

A Proposal for the Facility for ALEPH Computing and Networking

J.J. Aubert, W. Blum, G. Bonneaud, C. Bowdery, J. Boucrot, M. Delfino, P.J. Dornan, E. Fernandez, J.F. Grivaz, G. Kellner, J. Knobloch, P. Laurelli, D. Levinthal, J.F. Renardy, F. Ruggieri, D. Schlatter, H. Videau, H. Zobernig

ABSTRACT

We describe the proposal for FALCON, the Facility for ALEPH Computing and Networking, including a functional specification and a model implementation based on existing hardware. The proposal consists of three parts: a DST-production facility based on DEC equipment to be located at the experiment; an IBM-compatible mainframe with large disk capacity dedicated to analysis to be located at the barn, next to the CERN Computer Center; and a data network linking the previous two systems to each other and to the workstations and computers located in the ALEPH office buildings 2 and 32.

1. Introduction

In the ALEPH meeting in Munich we had a presentation, followed by a discussion, of the ALEPH computing requirements and the CERN computing capacity expected at the time of the start of LEP. These issues were summarized in our report of December 1986 [1]. The main conclusions of that report were that:

1. The computing required by ALEPH at CERN for general work would amount to about 9 units. In addition 6 units of computing power dedicated to Data Processing (event reconstruction) would be needed for about 6 months each year, the time the experiment should be running at full efficiency on average. It was also noted that these requirements were only estimates, especially the amount of computing required for analysis.

2. The computing power offered by CERN to ALEPH would amount to about 7 units, distributed over three kind of computers (3 IBM-compatible, 3 CRAY, 1 VAX). These numbers were also estimates, there was no written statement by CERN on this.

As a consequence of the above we concluded that:

(a) To meet the Data Procesing requirements ALEPH should have a privately-owned computing facility at CERN with a computing power of about 6 units dedicated to event reconstruction, which we called FALCON (Facility for ALeph COmputing and Networking).

(b) For economic reasons FALCON should consists of a standard (e.g. IBM-compatible or DEC) host computer with attached processors. The function of the host would be to run the attached processors on the one hand and on the other to supplement the shortcomings of the analysis capacity offered by CERN, in particular computing power, storage capacity and access to this storage.

(c) It was also found that the storage capacity offered by CERN (5 to 10 GB, mainly on IBM-compatible disks) was insufficient to store the large data base required for adequate analysis. We concluded that ALEPH should also have privately owned disks of large capacity (about 20 GB). We implicitly assumed that this disk should be IBM-compatible. The reasons for this were that the private disks should be an extension of what we already have available from CERN, and that the disk should be accessed from where we have most computing power available. Also, the main tape storage at CERN is expected to be in IBM-type 3480 cartridge tapes.

(d) To cope with the uncertainties of the computing requirements it was concluded that the facility would be expandable, especially the host.

(e) Fast access of the data base (on disk and tape) from FALCON would require an IBM-compatible host. This was possible by having a dual ported controller to the disk (one from FALCON the other from the IBM-compatible computers in the CERN Computer Center), by a channel to channel connection between both computers or by both.

(f) Access of the ALEPH (IBM-compatible) disk and the CERN tape library from a VAX host was found slow under all possible pathways. This was essentially the main difference (other considerations aside) between the so called VAX and IBM solutions described in the report.

Further analysis of the two possibilities mentioned in the report lead us to a dilemma. To summarize simply the main points:

- The IBM solution was preferred because of the accessibility to the disk, but the only serious possibility for attached processors were the 3081E emulators, which do not have commercial support and the collaboration did not favour.
- The VAX solution had the attraction of using integrated processors, namely the Zodiak boards or VAX workstations, but the accessibility of the large disks remained a problem.

This was basically the situation presented at the Copenhaguen meeting [2], together with specific preliminary offers from IBM and DEC. The decision in Copenhaguen was to wait until November for a final proposal.

2. New developments

One of the preliminary offers from DEC was to use workstations as attached processors, namely 24 VS2000 workstations to make six units of computing power. A specific study was made this summer of the possibility of producing the DST with a battery of such workstations [3], with favorable results. The dramatic drop in prices that occured in the last months, and the announcement of more powerful VAX workstations, makes the price/computing-power ratio of workstations competitive with that of emulators.

The fact that the VAX workstations run a complete operating system, including input-output, also makes it possible to configure the DST production facility with a very minimal

host computer. Therefore, it is not necessary to use the large expandable analysis computer in a dual role, of not only providing interactive analysis but also of controlling the attached processors. This allows a separation of the two main functions of FALCON, namely the DST production and the analysis, which were linked in the past due to the lack of sophistication of the attached processors. This separation allows the design of a system which uses the best hardware solution for each function.

The low price of workstations will also affect the way analysis is done. It is expected that a large fraction of the people doing analysis will be using workstations. However, these workstations do not solve all the analysis problems. In particular, the speed at which they can access the large disk and the tape library is very limited. It is necessary to match the large disk with enough CPU capacity and input-output bandwidth to process files (such as the DST) and produce the information (such as n-tuples) that the workstations will be demanding. Thus, the use of workstations does not do away with the need for large compute power with fast access to a large amount of data, but in fact enhances that need.

As has been mentioned in the introduction, our estimates indicate that the compute power and storage provided to us by CERN will fall short of those needed to make ALEPH a competitive experiment. In spite of the introduction of workstations into the analysis process, we still see a need for additional mainframe power, closely coupled to our planned large disk and to the CERN mainframes. This additional capacity could be provided by an expansion of present CERN facilities or by adding equipment owned by ALEPH. Given our present knowledge of our likely future allocation of CERN computer resources, we propose here to acquire ALEPH owned equipment.

3. The proposed FALCON system

Based on the arguments just given, our proposal for FALCON consists of the following three parts (see Figure 1):

(*a*) A DST production system, based on DEC equipment, located at the interaction region. The system will obtain the raw data files directly from the online system through a network connection.

(*b*) An analysis system, including a large storage disk, based on IBM-compatible equipment, located at the "barn" next to the CERN computer center. The large disk should be accessible, at channel speeds, from both the analysis system and the CERN mainframes.

(*c*) A strong data network, linking both parts of the system with each other and with the rest of the ALEPH facilities, such as the workstations and computers located in Buildings 2 and 32. This network will support the transmission of the DST from the interaction region to the CERN computer center.

We now explore the three parts in more detail. In addition, proposed functional specifications for FALCON are included in Appendix A, and specific configurations using presently available hardware and fulfilling the functional specifications are included in Appendix B and are used to arrive at cost estimates. It should be understood that the example configurations in Appendix B will most likely not be the final ones, since the hardware is in constant evolution. The proposed timetable for FALCON implementation is given in Appendix C.

3.1 DST PRODUCTION SYSTEM BASED ON DEC EQUIPMENT

DEC VAX computers can be connected together into VAXclusters, sharing system resources such as disks and tapes. Such a configuration is already being used for the ALEPH online system. The Local Area VAXcluster architecture, implemented over an Ethernet network, makes it possible to add CPU power in the form of inexpensive disk-less workstations, which have a price/computing-power ratio as good or better than 3081E emulators and adequate input-output capabilities. The newly announced series 3000 VAXstations have a power of about 0.5 units per node (yet to be benchmarked with JULIA). Therefore a VAXcluster of about 12 such stations and one small control node to host disk and tape units and manage connections with the online and offline computers is a possible solution available today which will deliver the 6 units of compute power required for DST production. It is also likely that more powerful VAX workstations will be available before startup of production and that they will fit into the VAXcluster architecture. Therefore the system will be flexible and expandable.

Detailed in Appendix B is a configuration of the DST production system based on presently available hardware. This configuration can be used for price estimates and would be refined after further tests and consultation with DEC engineers. Our cost estimate for the DST production part of FALCON is 0.7 MSF.

3.2 Analysis system based on IBM-compatible equipment

Most of the interactive computing power with fast access to data, as well as most of the disk storage made available to us by CERN is in the form of IBM-type equipment. The most effective way of supplementing these resources with ALEPH-owned equipment, to cover the foreseen shortcomings for the analysis, is to integrate the ALEPH-owned equipment as closely as possible with the CERN IBM-type facilities. Therefore our proposal is to acquire an IBM-compatible mainframe with a large disk subsystem which can be shared, through a dual-ported controller, with the CERN IBM-type computers.

As has been mentioned in the introduction, our original estimates were that we would be short by a minimum of two units in compute power. Furthermore, indications from measurements by a MUSCLE subgroup using PAW are that such interactive analysis creates heavier demands on the mainframes than previously estimated. In addition, it is expected that by the startup of LEP there will be a very large number of interactive users so that the response of the CERN computer center could be considerably degraded. Therefore we propose that FALCON provide the support for about 50 users doing interactive analysis, the corresponding CPU power being estimated at a minimum of 3 units.

A large disk will also be essential for fast effective analysis. The proposed storage capacity is a minimum of 20 GB. To give an idea of what this size means, it corresponds to about 10^6 DST events, or 10^7 MDST events plus several different selected DST event files.

In Appendix B, we have included a rough parametrization of the cost of a mainframe installation as a function of CPU capacity and disk space, obtained from a survey of about a dozen configurations of DEC and IBM mainframes. Our estimate for the analysis part of FALCON is 2.0 MSF.

3.3 Networking

The network part of FALCON is proposed to be based on the Ethernet standard, which is supported by CERN. It will provide file transfer, remote login and electronic mail capabilities. It is also desirable to have support for remote procedure calls.

The network has two geographical domains:

(a) Link from the experimental pit to the CERN computer center:

This link must support the transmission of the DST files from the DST-production farm to the analysis mainframe, as well as miscellaneous traffic such as terminals, constant

files and job output. The total throughput necessary is estimated at 100 KB/s.

Part of the LEP infrastructure plan already provides for transparent bridging between Ethernet segments at the experimental halls and the CERN site, using optical fiber and coaxial cable installed in the ring and devices called T.D.M.'s. The speed of this bridging is expected to be 200 KB/s or more. A dedicated link of this type (additional to the one provided by LEP) between the DST-production and the analysis parts of FALCON will satisfy the requirements. The cost of this connection is estimated at 0.15 MSF.

(*b*) Link from the computer center to ALEPH office buildings (2, 32, etc.):

It is clear that the traffic between workstations and computers located in the ALEPH office buildings, and the mainframes used for analysis located at the computer center, will overwhelm the present CERN general-purpose network. Therefore it is proposed to isolate the ALEPH Ethernet traffic from the rest of CERN using LAN bridges. This has already been done around buildings 2 and 32. The equipment in the barn would form a separate Ethernet segment and be connected to the office buildings over the CERN optical fiber backbone. To maintain access to the CERN network, this segment would be bridged to the general CERN Ethernet.

The connection of the analysis mainframe to the Ethernet could be done either through a native interface, such as the recently announced IBM channel attached Ethernet port and IBM TCP/IP software, or through a gateway device, such as the Interlink box used at CERN. This would remove the bottleneck of access through the CERN Interlink box, which is already saturated and shared by all the CERN users.

The Ethernet bridging between ALEPH offices and the computer center should be done over the CERN optical fiber backbone network which we assume will be provided by CERN. In this case, the cost of isolating the traffic with LAN bridges and adding a dedicated mainframe Ethernet interface would amount to about 0.1 MSF.

Our estimate for the cost of the networking part of FALCON is 0.25 MSF. An additional amount of 0.1 MSF is needed for the infrastructure of the endpoints of these links, such as conditioning the space at the "barn".

4. Manpower requirements

The Barcelona group proposes to take responsability for all aspects of the DST-production part of FALCON. Some work such as running JULIA with optimized system parameters on the ALEPH workstation cluster has already been done [3].

The final configuration of the analysis mainframe and the purchasing negotiations would be handled by the present ALEPH Data Analysis Facilities Group (i.e. FALCON group) with possible help from the CERN DD division. It is estimated that operation and maintenance of the analysis mainframe and the network would require two trained computer specialists, one mainly for software and the other for hardware. These two people should work in close contact and collaboration with the DD division. The Barcelona group is looking into the possibility of providing the software specialist, at least for the initial two years of running. A possibility for the second specialist would be a rotation of experts from the home institutes with large installations. In addition to the specialists, there probably will be need for some help from general members of the collaboration to run shifts as operators. This need may disappear as we gain experience running the system.

5. Conclusions

In this note, we have made a detailed proposal for the computing facilities which we believe will give us adequate support for the analysis and keep ALEPH in a competitive situation with respect to the other LEP experiments. These facilities include an autonomous DST-production system, an analysis system, and proper interconnections between them, the CERN computing facilities and the ALEPH analysis equipment in Buildings 2 and 32. The detailed implementation of this plan should begin as soon as possible in order to be ready at the start of LEP operation.

APPENDIX A

FALCON FUNCTIONAL SPECIFICATIONS

A.1 FALCON DST Production System

The FALCON DST production system is a computer system whose purpose is to process files of events and associated information produced by the ALEPH online system (which will be called "Runs") through the ALEPH reconstruction algorithm (known as JULIA) and output a summary file (known as a DST file).

The DST files should ultimately appear on IBM 3480 cartridge tapes in the central CERN tape library.

The system should have a speed such that incoming data can be processed at the rate at which it is collected by the online system; it should be noted, however, that this rate is the average rate over a long period of time (i.e. one day) and not the instantaneous real-time rate.

The system should also be capable of using as input Runs which have been archived by the online system onto mass-media.

CPU capacity The system should have a minimum total CPU capacity of 6 IBM-168 units, or about 7.2 MFLOPS, when running the ALEPH reconstruction program JULIA. It should have an expansion capability up to a minimum of 9 IBM-168 units.

Program memory The system should make available a minimum of 5 MB of memory to the JULIA program.

Input-output throughput The minimum input-output throughput should be such that it can support the following rates:

a) Input event rate of 1 Hz with an average event size of 100 KB.

b) Output event rate of 1 Hz with an average event size of 20 KB.

c) Control and summary information (estimated negligible with respect to the sum of *a* and *b*).

Disk Storage Aside from the necessary storage for operating system and program files, disk mass-storage is needed to buffer the online Run files. The minimum space required is the one to hold one Run's worth of information, assumed here to be 4 hours at the rates above, therefore 2 GB.

Secondary Storage The system should have access to secondary storage media identical to those used by the online system to archive Run files. This can be done either through a secondary storage system directly on the system, or by providing access to the online secondary storage system through appropriate connections. The minimum, average, input-output throughput rate for accessing this storage should be the same as specified above for real-time operation.

CERN Tape Library The system should be able to deliver the DST files on IBM 3480 cartridge tapes to be stored at the CERN Tape Library. These tapes should be readable directly from the CERN IBM-type computers. The system can either write the tapes on directly attached cartridge drives or transmit the DST files over a network to the CERN Computer Center, where they would be written to tape using that facility.

Operating System Environment The system should fit within the established rules for ALEPH programming environments. In practice this means that:

(a) The executable code and numerical representations should be compatible with either IBM System/370 or DEC VAX.

(b) The system should be controlled, managed and programmed from either IBM VM/CMS or DEC VAX/VMS operating systems.

Network connections The system should provide network connections as specified in the FALCON network functional specification.

Maintenance There is strong preference for industrially supported hardware.

A.2 FALCON Analysis System

The FALCON Analysis System is a computer system for ALEPH data analysis, to serve in a complementary fashion to the computer resources provided by the CERN computer center.

CPU capacity The system should be capable of supporting a minimum of 50 interactive physicists-users, as defined by CERN resource-demand studies. A batch system will use the CPU capacity when the interactive demand falls from the peak.

Disk Storage The system should provide a minimum of 20 GB of user disk space, aside from the disk necessary to contain the system files, editor, compilers, etc. This disk space should be accessible at channel speeds (> 3 MB/s agregate rate) from the CERN IBM-type computers.

CERN Tape Library The system should provide access, through the shared disk subsystem or through direcly attached drives, to the CERN IBM 3480 cartridge tape library.

Operating System Environment The system should fit within the established rules for ALEPH programming environments. In practice this means that:

(a) The executable code and numerical representations should be compatible with either IBM System/370 or DEC VAX.

(b) The system should be controlled, managed and programmed from either IBM VM/CMS or DEC VAX/VMS operating systems.

Network connections The system should provide network connections as specified in the FALCON network functional specification.

Upgradability The system should be upgradable to a minimum of double the initial CPU capacity and disk space.

Maintenance Maintenance of the system should be commercial and of the same level of service as the CERN mainframes.

A.3 FALCON Network

The FALCON network is a data network linking three geographical sites:

1. The ALEPH experimental hall, where the DST-production system (specified in Appendix A.1) will be located.
2. The CERN computer center "barn", where the analysis mainframe (specified in Appendix A.2) will be located.
3. The ALEPH office and laboratory buildings, around Building 2 and 32 of CERN, where terminals, workstations and other computers are located.

The distance between sites 1 and 2 is about 24 Km; that between sites 2 and 3 is about 2 Km.

The network should be based on the Ethernet standard, with suitable, transparent bridging over other media where necessary. Networks using standards other than Ethernet would be considered if they provided substantially higher functionality and/or speed.

The network should provide the following facilities:

File transfer: The actual throughput transfer rate for two collaborating processes, transmitting a file between the DST-production and the analysis systems, disk-to-disk with normal load, should be a minimum of 100 KB/s. The actual throughput aggregate transfer rate between the analysis system and the workstations and computers in site 3 should be a minimum of 200 KB/s, under normal load conditions.

Remote login: The network should provide remote login (character-oriented or full-screen, as necessary) between any two machines connected to it, and should allow connection of Ethernet Terminal Servers.

Electronic mail: The network should allow integration into the CERN electronic mail system.

Remote job submission: The network should allow for remote job submission from any machine on the network to any other, as long as the target machine supports this activity.

Access to the CERN general purpose network: The network should provide transparent access to the CERN Ethernet-based links.

Maintenance and availability: These should be of the same quality as the CERN general purpose network.

Desirable features are:

Higher speeds.

Support for remote procedure calls.

Higher level of functionality (or "user-friendliness") in the mainframe-workstations link.

APPENDIX B

MODEL IMPLEMENTATION USING HARDWARE AVAILABLE IN OCTOBER 1987

B.1 FALCON DST Production System

Based on the functional specification, the FALCON DST production system could be implemented as described below. It is noted that the type of hardware necessary for this system is in a state of continuous change, and therefore a different and better implementation might be possible in the future. For the same reason, the system has not been described to the last detail. However, here we describe the situation utilizing existing announced hardware, with a few exceptions as noted below, in order to obtain an estimate of the cost to be used in financial planning for FALCON. A full, detailed configuration of the system will be arrived at later, with the help of DEC engineers; however a target cost figure is absolutely necessary for this type of work to proceed.

The system is based on a DEC Local Area VAXcluster architecture. A Local Area VAXcluster (LAVC) is a set of DEC VAX CPU's connected through an Ethernet, all sharing access to the same disk-space. It is noted that a commercial IBM-type solution was explored, but the DEC solution is chosen for price/performance and integration reasons.

The system consists of a LAVC with:

- 12 VAXstation 3200 workstations used as compute servers (0.5 IBM 168 units of CPU each, yet to be benchmarked with JULIA). Each workstation has 8 MB of memory, of which approximately 7 MB are available to programs. An additional 12 workstations can be added to the LAVC. . . . . 490 KSF
- 1 control node, to serve as the boot-member of the LAVC and provide access to peripherals and the online VAX cluster. Here there are several choices depending on how the connection to the online system is made. We use here a VAX 8250. 195 KSF
- Disk space for buffering incoming data, 2.5 GB (4 RA81 disks). . . . . 150 KSF
- Tape equipment to write DST files on 3480 tapes. . . . . 80 KSF
- Contingency for cables, installation, etc. 10%

The prices above (except the 3480 tape equipment) are derived from specific quotations from Digital. The tape equipment cost is estimated as an addition to the precently planned tape system for the online cluster. The total list price, including contingency, adds up to 1.0 MSF. A substantial discount, on the order of 30%, can be expected.

The total cost estimate for the FALCON DST system is 0.7 MSF.

B.2 FALCON Analysis System

A survey of several mainframe configurations, ranging in power from 2 to 6 units and in disk space from 10 to 20 GB has been made, using approximate prices and discounts provided by IBM and DEC. From this survey, an approximate parametrization can be made of the cost of a mainframe as a function of CPU power and disk capacity. This parametrization is given by

$$\text{Cost} \approx 0.8 \text{ MSF} + 0.5 \text{ MSF}(\frac{\text{Disk space(GB)}}{20 \text{ GB}}) + 0.3 \text{ MSF}(\text{CPU Capacity (168 units)})$$

and applies to cost in second quarter of 1987. A system with 3 units and 20 GB, for example, is seen to cost about 2.2 MSF. It is also noted that the cost of computer equipment tends to decrease about 20% per year. Taking this into account, we estimate a cost of 2.0 MSF for a system to be installed in early 1989. A final configuration of an optimal system which fulfills the functional specification would be arrived at later, with the help of CERN experts and manufacturers of IBM-type equipment.

B.3 FALCON Network

No detailed model implementation has been made yet. The long distance link from the experimental hall to the CERN computer center will be based on a coaxial cable and optical fiber link using TDM devices, providing transparent Ethernet bridging; this system is being developed as part of LEP infrastructure. The link from the analysis part of FALCON in the barn to the office buildings 2 and 32 will be based, if possible, on the CERN backbone network.

In order to isolate the ALEPH network traffic to assure adequate bandwidth, but also maintain access to the rest of the CERN network, LAN bridges would be used. These are already in common usage at CERN and are manufactured by DEC and other companies.

The type of interface and software for access to the analysis mainframe must be specified as part of the machine itself. In addition, devices from third-party manufacturers, such as the Interlink interface currently used at CERN, can provide expanded functionality.

APPENDIX C

PROPOSED SCHEDULE FOR IMPLEMENTATION OF FALCON

All parts of the FALCON system should be installed, tested and ready for event processing by "day 1", which we assume to be sometime in Fall 1989. Work can proceed in parallel on the three parts, i.e. the DST-production farm, the analysis mainframe and the network. Following is what we believe is a realistic schedule for implementation, based on the assumption that FALCON is configured along the lines proposed in this document:

- Technical proposal . . . . . ALEPH meeting November 1987
- Discussions and clarifications of proposal . . . . . Spring 1988

- Begin development of DST-production software on the ALEPH workstation cluster . . . . . Spring 1988
- Begin negotiations for purchase of partial DST-production system (i.e. without all the CPU power) . . . . . Summer 1988
- Install partial DST-production system, begin tests of file transfer with the online system, develop run strategies . . . . . Fall 1988
- Purchase full CPU power for DST-production system and move it to experiment area; tests using Monte Carlo events . . . . . Spring 1989

- Begin development of analysis software to exploit specific features such as the large disk and the mainframe-workstations connection . . . . . Spring 1988
- Begin negotiations for purchase of analysis mainframe . . . . . Summer 1988
- Delivery and installation of analysis system . . . . . Spring 1989

- Start specific discussions about network with CERN DD and LEP groups . . . . . Spring 1988
- Install network from computer center to Buildings 2 and 32 (temporarily connected to CERN mainframes) . . . . . Summer 1988
- Begin tests of Ethernet mapping from the pit to the computer center Summer 1988

- Install dedicated DST transport link . . . . . Spring 1989
- Install dedicated mainframe Ethernet interface . . . . . Spring 1989

- All parts of FALCON available for general work . . . . . Summer 1989
- Begin production . . . . . Fall 1989

REFERENCES

1. Report of the ALEPH Analysis Facilities Group, J.J.Aubert et al., ALEPH-NOTE 170, 1986.
2. W. Blum presentation to ALEPH Plenary Meeting in Copenhaguen, May 1987.
3. Measurements of the Performance of a Local Area VAX Workstation cluster, Manuel Delfino and Andreu Pacheco, ALEPH 87-84, 1987.

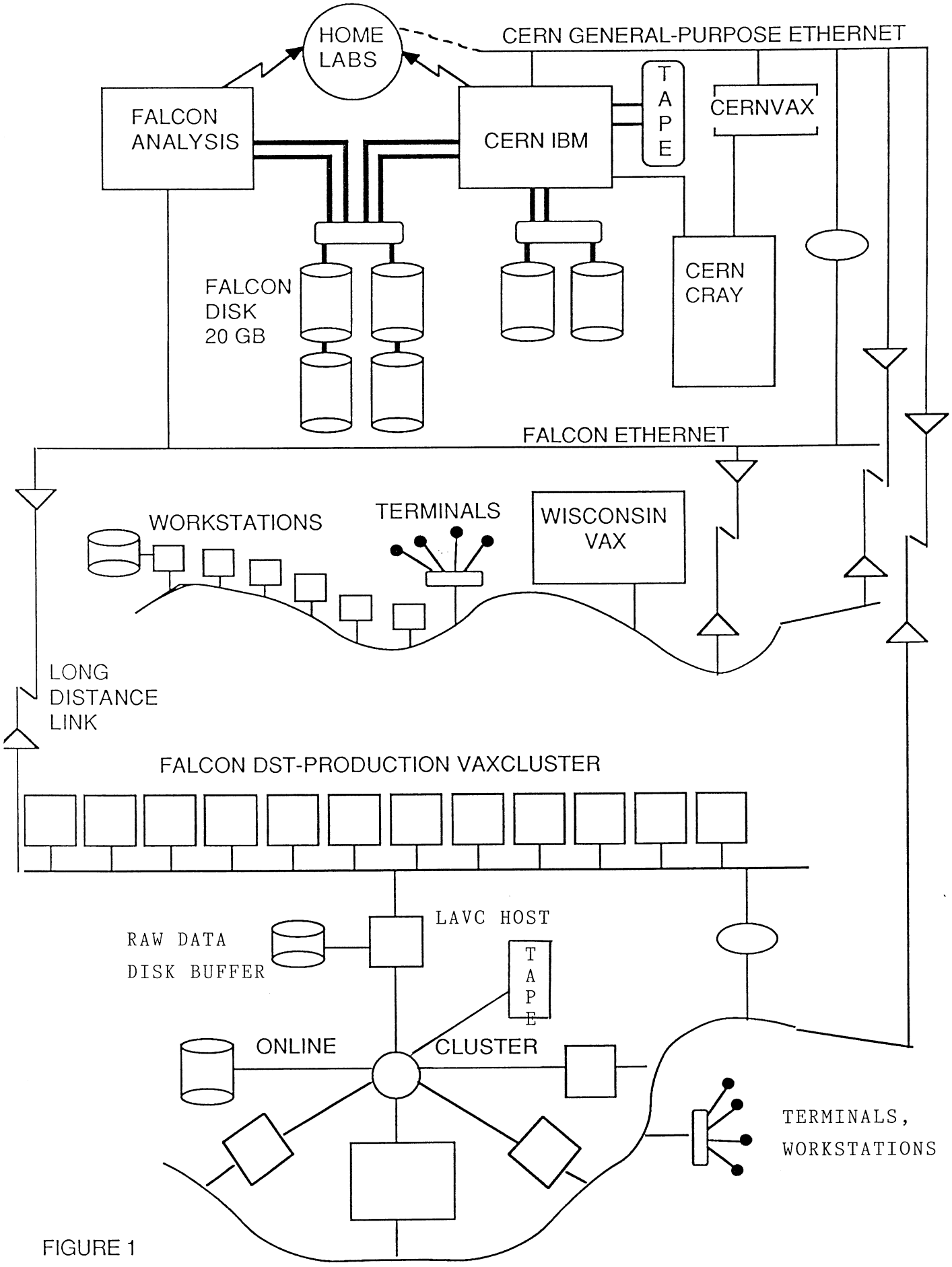

FIGURE 1

PRELIMINARY SKETCH OF THE PROPOSED FALCON SYSTEM AND ITS CONNECTIONS TO OTHER RESOURCES.