A Data Handling Subsystem and a Local Computer Interconnection Subsystem.

A summary of ECMWF's Invitation To Tender, by Dick Dixon, ECWMF.

1. The ECMWF ITT

ECMWF's ITT No. ECMWF/F(82)133 is for a "data handling sub-system and local computer interconnection sub-system". These are known within ECMWF itself as the data handling and data highway systems respectively. What ECMWF requires in response to this Invitation is :

1) A local area network

A high-speed unified local area network able to interconnect ECMWF's existing computers and likely future acquisitions.

2) A general-purpose computer

A small-to-medium-scale general-purpose computer system to be dedicated to centralised file handling for the ECMWF computer service, including:

i) Large-capacity magnetic disk storage.

ii) A bulk data store (i.e. a mass-storage device) capable of holding tens or hundreds of Gigabytes of data on low-cost online storage.

iii) A conventional half-inch magnetic tape system.

iv) An attachment to the local area network.

v) Data management software to provide an integrated file management function for the other computers on the local area network.

This paper describes some of the motivation for requesting these individual functions, and gives some details of the type of system which ECMWF is expecting.

2. The Local-Area Network - Requirements

ECMWF has identified the following requirements which are likely to be best satisfied by a local-area network solution:

i) Simple addition of new computers to the existing system.

For example, when the Cyber 835 (which should have been an easy case, being of a type already well known at ECMWF) was added to the existing configuration in late 1981, a considerable period elapsed before the combined system was performing without unforeseen and unwanted interactions between the existing and the new machines. In future the requirement for new machines is likely to increase as the pace of advance in computer hardware becomes still faster, and ECMWF must have sufficient flexibility in its computer operations to achieve this without disruption.

ii) Simple addition of extra computing power.

Deriving from the same causes as the previous point, this flexibility offers the chance to add extra power in the form of an additional machine of some kind, rather than replacing an existing machine (with consequent disruption of service).

iii) Increased flexibility in operation.

Although existing interconnections between computers have been expensive and difficult to achieve, they do bring an undeniable improvement in operational flexibility and reconfigurability. An example is the use of ECMWF's Cyber 835, which at the present time is dedicated to interactive service during the day. This division of function between the existing machines and the new one was not that foreseen when the new machine was first installed; however, the particular interconnection system used in this case was such as to allow almost complete freedom in allocation of functions to the machines. Such flexibility is not common among existing computer interconnections (it would, for example, be quite out of the question to consider such redistribution between the Cybers and the Cray), but the type of local-area network being requested under this ITT should considerably enhance ECMWF's ability to reconfigure its operations to meet changes in demand.

iv) Availability of special-purpose devices.

ECMWF's existing computer service is based on machines from a large-scale scientific computing background. Such machines are not easily interfaceable to novel peripherals (particularly graphics devices of various kinds, data capture or communications equipment), and such requirements in the past have been either ignored, solved in a 'one-off' fashion at considerable expense (e.g. ECMWF's plotters and Regnecentralen communications front-ends), or handled with manufacturer-specific hardware (e.g. the CDC 2551). What is required is a means of attachment for industry-standard peripheral devices which can be made available for new applications at relatively small cost.

v) Elimination of 'knock on' effects in the replacement of computers.

Where the 'one-off' solution has been adopted to the problem mentioned in (iv), there is then the additional problem that considerable investment exists in the hardware and software of the special-purpose equipment: replacement of any part of the resulting composite system becomes

disproportionately expensive because of the past investment which would be invalidated by the change. By going to a unified system of interconnection, special-purpose subsystems can be built whose existence is not exclusively dependent on a unique environment.

3. The Local-Area Network - Expectations.

ECMWF expects that the local-area network system (LAN) supplied under this ITT will have the following characteristics:

1) Structure

The system is expected to be a bus-structured LAN, using the CSMA/CD access contention procedure (a means of managing access to the interconnection bus without needing a central controller).

2) Speed

The ITT requests a gross network capacity of 25 Megabits/second. The raw hardware throughput of the network is not expected to be a problem.

3) File Transfer

The ITT requires that the system should be able to offer file transfer in both directions between any pair of computers, and that it may be initiated by either the sending or receiving computer. Using ECMWF's existing Cray connection as an example, this requirement exceeds the current capability of that link (since only the Cray may initiate a transfer on the latter), but the user interface might be broadly similar (i.e. ACQUIRE, DISPOSE between the computers on the network).

4) Message Transfer

The ITT requires that the system should be able to offer message transfer ('Interprocess Communication') between any pair of running programs in computers connected to the network. To achieve this requires that both programs should first advise the network software of their intention to pass messages, and requires sophisticated recovery features in the network to cater for the termination of one program while the other is attempting to communicate with it. No system currently in use at ECMWF has this type of facility: it is expected that its availability could have great impact on the next generation of forecast suite systems, for instance.

5) Mainframe-compatible performance.

The system will not be a Cambridge Ring, an Ethernet, or any other of the office-equipment-oriented network systems which have become available in recent years. The performance required and the degree of sophistication in facilities and reliability which is wanted make such systems completely unsuitable.

4. The Data Handling System - Requirements

In installing a data handling system or centralised filing system, ECMWF's requirements are:

1) A way to overcome the expanding bulk of the conventional tape library.

   The ECMWF tape library currently holds around 15000 tapes - not a large number by computer industry standards, but this number has been reached from zero in just 3 years, and the rate of increase is not diminishing. Unless some way is found to economise on the use of tape reels, the tape library will be full (with 35000 reels of tape) at the end of 1986.

2) A way to overcome the problems of mounting large numbers of tapes.

   ECMWF operators currently mount between 2000 and 3600 reels of tape per working week. The tape subsystem (currently supported entirely by the Cybers) is probably not capable of achieving a mount figure in excess of 4500 tapes per week. To go further (and the demand for tape data will inevitably increase beyond this point) demands attention to problems of manpower, software and hardware, and would probably have a severe impact on the service given to interactive computer users.

3) A way to provide extra capacity for the next generation of computers.

   ECMWF will probably install a Cray X-MP in 1984, which will result in an immediate increase in the number and complexity of meteorological experiments being run, and will more gradually increase the complexity and data demands of the operational forecast. To cater for this increasing demand (which will of course increase further with the succeeding generation of back-end processors), a different approach to the storage of meteorological data must be taken.

4) A way to provide better response for system users.

   The time to mount a magnetic tape is of the order of 10 minutes; and the time to read a file from it will be on the average 6-7 minutes. Meteorological researchers have become adept at dealing with these long response times, but a system providing improved response would be of great assistance in aiding the progress of research work.

5) A way to offer long-term continuity of metorological data storage.

   The present format of meteorological data is closely dependent on the hardware and software characteristics of the Cyber computers which handle the magnetic tapes. The lifetime of these machines is not indefinite, and the data formats they use are far from being an industry standard: hence, there is a problem

with the long-term compatibility and supportability of the archived data. A system which uses more commonly accepted data formats, and which can if necessary deliver data in a converted form suitable for any given worker computer, is likely to greatly simplify the task of making the data available over extended periods of time (20 years is a reasonable period to expect for long-term archives, for instance).

5. The Data Handling System - Expectations

The data handling system will be based on a small-to-medium scale general-purpose mainframe computer, equipped with a considerable amount of storage capability, an attachment to the local area network (to enable it to provide service for the other computers at the Centre), and a software package to manage the various types of storage device and provide a consistent and convenient service to users, whichever worker computer they may wish to run from.

In particular, the data handling system will include:

1) General-purpose mainframe computer.

A general purpose mainframe with 8 Megabytes of main memory and a power rating of 1.0-1.5 mips (million instructions per second).

2) Large-capacity disk storage.

The disk storage will be installed in stages: in stage 1, 1 Gigabyte, in stage 2, 6 Gigabytes, in stage 3, 16 Gigabytes will be available. The stage 1 installation will not be capable of running a filing service, and will be used simply to install and check out software, and to gain experience of the attachment to the network.

3) Bulk data store.

The bulk data store (more commonly known as a mass storage device), will be required to hold a large set of data online (i.e. to be available without operator action) with a much lower cost-per-stored-byte than disk storage. The capacity of this device is uncertain, but is expected to be sufficient to hold between 300 and 1000 full tapes of data.

4) Magnetic tapes.

Although magnetic disk and mass storage technology hasa improved dramatically in recent years, there is still none available which appears viable as a means of storing thousands of Gigabytes of data which must remain usable for periods of 20-25 years. Inevitably we are forced to consider conventional half-inch magnetic tape as the only long-term storage medium with sufficient integrity over this period. It is possible that cartridges for a mass storage device might be used to hold such

long-term data, but these have a much higher cost-per-stored-byte. The requirement therefore is to be able to use magnetic tape, but with a higher efficiency than before (by, e.g. using file compaction techniques) for ECMWF's long-term archival storage.

5) Software.

The software system which will combine all these elements to form a complete system is absolutely vital - probably more crucial to the success of the project than any other component. The file management software will:

i) Conceal the characteristics of the various hardware devices from the users: a file will simply be despatched to the filing system and will reappear on demand without requiring any user knowledge of the device(s) on which it may have been stored.

ii) Migrate (i.e. copy) files between storage devices according to the use they have received in the past (and possibly according to the use they are expected to receive the future). The trick here is to ensure that the files which are likely to be used in the immediate future are copied in advance to the fastest available storage medium (i.e. magnetic disk), and that unused files migrate eventually to the lowest cost medium (i.e. magnetic tape). The central filing system installed at Los Alamos laboratories in the USA achieves a remarkable success in this endeavour, satisfying 80% of file requests from disk storage, 18% from a mass storage device and 2% from offline storage. If the system installed at ECMWF can achieve a performance even fairly close to this with its migration algorithm, it will have achieved a major milestone in the progress of the project.

iii) Provide media recovery - i.e. if a disk drive (or any other storage device) should fail, it is possible for the managed storage system to recover from this stiuation with no ill effects other than a deterioration in the performance of the system. The ITT requests extreme resilience in the face of media failure: how much the user can be protected in practice will depend very much on the particular implementation chosen.

iv) Allow flexibility. Since all storage on the filing system is being handled by the data management software, more of any given type of storage can be added to adapt the system to changing requirements, without the characteristics of the system changing as seen by the user. This means that extension of capability can be provided relatively easily, and that advantage can be taken of new technology (e.g. the long-awaited video disks for data storage) as and when it becomes available.

v) Improve efficiency. Since the file is never accessed directly by a user of the system, it may be reformatted, compressed or otherwise processed to maximise the efficiency of its storage, without inconveniencing the file owner. Those

files which must be copied to tape will occupy only as much tape as their data content demands - there will no longer be a need to keep many tapes in the library with only 24 feet (out of 2400) holding useful data. The effect of the file migration algorithm is also to improve efficiency, since files will be migrated to the cheapest form of storage which can reasonably support their expected use.

6) The data highway (LAN) attachment.

Evidently a connection is required between the data handling system and the local area network, simply because the network will be the major means of communication between the computers at ECMWF. Since only the local network concept makes a separate filing computer viable, the connection between it and the network is of considerable importance. This aspect requires careful consideration in the case where the data handling system and the local area network are supplied by different companies: considerable pains must then be taken to ensure that the system and the network are not only compatible, but are constructed with compatible philosophies so that no mutual interference is caused.

7) System software.

The operating system software (as opposed to the data management software) running on the data handling processor will not be a major consideration for users of the system. All filing requests will be filtered through the data management software, which will have its own (hopefully user-friendly) syntax. In particular, it is not planned to run a normal user service on this machine: no job submission services from existing worker computers will exist. This is for three reasons:

i) The data handling processor will not have sufficient spare power to run general data processing work.

ii) ECMWF does not have sufficient manpower to provide software and user support for another general-purpose data processing environment.

iii) Experience has shown that the reliability of system software is greatly improved when it is operated in the relatively restricted environment of a dedicated service.

In practice, a small time-sharing operation will be run on the data handling machine, to permit software maintenance and system monitoring activities etc., so that the normal facilities of a general-purpose operating system must be present, but this will not be supported to the same degree as it would be if generally offered.