MARS, ECMWF's Meteorological Archive: Experience in managing a large archive

Baudouin RAOULT

ECMWF

Shinfield Park, Reading, RG2 9AX, UK

baudouin.raoult@ecmwf.int

Abstract - This paper describes the evolution of ECMWF's archive, as well as the lessons learnt from managing it over the years.

1. Introduction

The European Centre for Medium-Range Weather Forecasts (ECMWF) is an international organisation whose main activity is to develop and run numerical weather prediction systems. They are models of the atmosphere and the ocean running on supercomputers. ECMWF has an operational as well as a research mission.

The Meteorological Archival and Retrieval System (MARS) was born in 1985 to store all meteorological observations used as model input, as well as all model outputs created at the Centre for both operational and research needs. The main concept behind MARS is to hide all underlying data organisation from the user. He or she expresses a MARS retrieval in meteorological terms (date, parameter, level,…) and not in terms of files.

2. Background

Several times a day, the Centre runs a series of numerical models forecasting the weather and the ocean state. The outputs of these models are collections of global meteorological and ocean fields, (e.g. temperature, humidity, wave height…) at different altitudes and at different valid times.

A model run creates thousands of fields, each of a size in the order of 100 Kbytes. This value changes as the resolution of the model increases.

At the same time, ECMWF's researchers carry out experiments that consist of running a variation of the same model with different parameterisations.

All data used as input and the output of these models are archived into MARS.

MARS is also the host of special projects run in collaboration with other research institutes.

3. The first MARS: Using CFS on MVS

ECMWF started its operational work on a Cray 1-A running COS, with a Control Data Cyber under NOS-BE as a front end. The first archive system, named GETDATA, was based on a NOS-BE tape system. It was limited to operational data, and researchers had to make their own arrangements to manage their results.

In 1983, ECMWF started to study the creation of a Data Handling System (DHS). An invitation to tender was issued and a solution based on an IBM mainframe running MVS was chosen. The data storage system was the Common File System (CFS) developed at Los Alamos National Laboratory.

The first version of MARS was designed and developed from 1983 to 1985. The client code was written in Fortran, and the server code in PL1. At that time, client/server architectures were not known, and clients were called "worker machines". Networking was very slow. MARS used CFS to store its data, and managed metadata and users' requests.

The MARS client part analysed the user's request and then submitted it to the MARS server that located the requested data in the data store using its own metadata. The data was then sent back to the MARS client that post-processed it before giving it to the user.

A similar process was involved when archiving data. The data transited through the client that analysed it and sent it to the MARS server; the latter stored it in the data storage system, and updated its metadata.

4. The second MARS: AIX and TSM

Around 1993, MARS had reached its limits:

- it had 5 millions files in CFS, a large number at the time
- the IBM mainframe was running out of capacity
- the MARS metadata could not cater for new data type requirements

ECMWF ran a new invitation to tender to replace its data handling system. The solution chosen was based on IBM RS6000 servers running AIX; the data storage system was TSM (Tivoli Storage Manager, named ADSM at the time).

The major change was the move to a Unix operating system, allowing for a better integration with the Centre's environment: a Cray running UNICOS and desktop workstations running SunOS. With Unix came simpler file management, a better networking environment and easier control over the operating system.

In order to cope with newer data types, and to support future requirements from the operational and research activities, the MARS server was redesigned, to be completely object-oriented and was written in C++.

The new design clearly separated the data semantics from its physical location. The data could then be moved from disk to tape, from tape to disk, reorganised into larger files, or split into smaller files, without any impact on the service. This gave great flexibility in organising the archive to adjust to the ever-changing environment, requirements and access patterns.

The new MARS defined a storage management system as an abstraction in which it could read, write and manage tape files. As a result, it can use simultaneously any storage management system

4.1 Back-archive

Back-archive is the process of copying the data from one archive system to another.

4.1.1 Processing

The back-archive was seen as an opportunity to normalise the data to the latest version of the GRIB 1 format, a standard for meteorological fields. A suite of jobs was set up to retrieve the data from CFS onto a server, processed the data, and then archived it into the new MARS server using the MARS client.

4.1.2 No interruption of service

The client was configured to attempt to retrieve data from the new archive first, and to try the old system if the request was not fulfilled. Data was archived in both systems for a while, in case there was a need to revert for some catastrophic reason.

From the user point of view, there were no changes in the client, and no interruption of service during the back-archive.

4.1.3 Problems encountered

The requirement to terminate the MVS system, due to high maintenance costs and to the need to recover some floor space meant that the back-archive was limited to 18 months. In order to copy 5,000,000 files (about 25 Tbytes) during this period, 6 files had to be processed every minute without any interruption. Knowing that it took in average 2 minutes to mount a tape and position to the beginning of a file, several copy streams were organised, sorted by tape volumes and files position on the tape in order to minimise the number of tape mounts and positioning

Furthermore, not all the tapes fitted in the silo, so the tape librarians could not keep up with the demand for off-shelf tapes. They had to be provided in advance with the list of tapes that would be processed next.

Because the back-archive process involved two robots, two sets of drives, three machines (the old MARS server, a worker machine and the new MARS server), several network adapters and routers for 18 months, the probability of failure of any one of these components was very high. A very resilient suite of jobs was designed, that would recover from any failure and carry on copying.

The 5 million CFS files were copied into less than 200,000 TSM files, a ratio of 25:1.

4.2 Split into several servers

As the contract with IBM specified that more equipment should be delivered as time went by, and as the initial machine was overloaded, the MARS server had to be split into three instances. This was achieved by dividing the MARS metadata.

Nowadays, there are three MARS servers: one serving operational data, one serving research data and one for special projects. Each instance has very different access patterns, and different resources are allocated to each of them. Algorithms have been developed to optimise drive usage between the three servers.

Users don’t need to know which server to contact; the client does this transparently.

5. The third MARS: AIX and HPSS

In 2001, ECMWF ran a new invitation to tender to replace its DHS. The selected solution was once again IBM, using HPSS (High Performance Storage System) as a data storage system. Very large scientific sites, with requirements very similar to the Centre’s, use HPSS. In terms of number of files and size of the archive, ECMWF places itself in the lower half of the HPSS community, so there is great confidence in its ability to suit the Centre’s needs.

Adding support for HPSS in the MARS server was done in a matter of days. The new system is still being tested, and will be used operationally in autumn 2002.

This time the back-archive will be run differently: TSM files will be copied directly into HPSS, using the API of each so data can be moved directly from tape to tape.

The transition should be transparent to the users. The silos will be shared between the two data storage systems and are almost full; therefore some problems are anticipated in the management tapes and drives while the two systems running.

6. Evolution of the MARS client

The MARS system is built on a client/server architecture. The client part has followed its own evolution, in parallel to the server’s development.

6.1 Fortran to C

ECMWF started using Unix when it purchased a Cray YMP running UNICOS and replaced user terminals with Sun workstations. At this time, it made sense to rewrite the client in C to take advantage of the tools and functionality offered by the Unix operating system and this language: better networking with TCP/IP, easier file management, simplified code (LEX and YACC were used to analyses the requests).

At the same time, the client code was also made independent from the MARS server, so it could retrieve data from other sources, preparing the coming migrations.

6.2 Post-processing

The archive contains global fields at full model resolution. Users frequently need the data at lower resolution or on selected areas of the globe. Post-processing facilities were added to the client allowing users to specify such data transformations. This part of the client is still in Fortran.

Later, the facility to perform free-form computation on the retrieved fields was provided. For example, a user could retrieve the two components `U` and `V` of a wind field and compute the wind speed as `sqrt(U*U+V*V)`.

6.3 Multiple data sources

The client was also enhanced to retrieve data from sources other than the main archive. All models write their output to a large Fields Data Base (FDB) on the supercomputer. The MARS client can retrieve data from this FDB, to speed up retrievals and to remove a dependency between the supercomputer and the data handling system. The FDB is large enough so the operational suite can run for at least 24 hours without the archive system being available.

The client can also retrieve data from files, allowing users to use its post-processing facilities on their own data.

Proxies were developed for Member State users that can run a MARS client on their local machines and retrieve data from the Centre.

6.4 Types of client

Most of the MARS activity is from batch jobs running on servers or supercomputers. These requests extract large quantities of data from the archive and may run for many hours.

ECMWF has also developed an application called Metview: a graphical user interface that allows researchers to retrieve data, performs unlimited numbers of calculations on them, and displays the results on their workstations. Metview users work interactively, retrieve small amounts of data and expect a quick response from the system. MARS gives priorities to Metview requests.

A Web interface to the MARS system has recently been developed to allow users to browse the complete archive catalogue, based on the MARS server metadata, as well issuing and monitoring their requests. Providing visibility to the MARS queues has proven to be very useful: users can see the size of the queues and the cost (in term of tape access and volume retrieved) of the active requests. This has led to a better usage of the system by the users.

7. Lessons learned

7.1 Migrating cost

Migrating is expensive. It takes resources and time. It is a complex process to manage and supervise: any part of the chain can break,

In the history of MARS, ECMWF has always run an invitation to tender to replace all components of the system: machines, disks, drives, robots, media and storage software. In the future, only parts of the system will be replaced.

7.2 Quality: don’t trust anything, double check everything

One of the main characteristics of an archive system is that the some of its content may not be accessed for several years. If after a time a user retrieves a piece of data that appears to be corrupted, *confidence is lost in the validity of the rest of the archive*, which therefore loses a lot of its value.

- How many other pieces of data are corrupted?
- Was the data corrupted prior to being archived?
- Was the wrong data archived?
- Did the network corrupt the data?
- Did the disks corrupt the data?
- Was the tape on which the data resided damaged?
- Did a bug in any software modify the data?

It is impossible to investigate an event that occurred several years ago. In the lifetime of MARS, various problems have occurred, such as unnoticed disk and network corruptions due to kernel bugs, tape corruption due to micro-code bugs and various user errors, for example overwriting operational data with test data. Every long-lived system is bound to encounter every possible problem.

The solution is a high degree of paranoia and checking:

- All data archived are self-describing: they contain a header, which uniquely identifies them. This is of paramount importance. If the metadata is lost, it can be rebuild by reading all of the data. This may take years, but everything can be recovered.
- Every file archived is checked on the client side, against an external description, to ensure that the proper data is being archived. Data corruption can also be spotted at this stage.
- The data is scanned once again on the server side.
- All disks are RAID or mirrored.
- “Enterprise quality” tapes are used.

- Multiple copies of important data are made. Copies are made from tape to tape to make sure the tapes are always read at least once.
- Operational data is retrieved after archiving and compared with the original copy.
- Neither the client nor the server modifies the data.
- Every client and server change goes through a series of lengthy regression tests before being installed. These tests exercise every functionality and data type supported by the software.

7.3 Manageability, performance and scalability

To achieve these three concepts, the MARS system:

- Tries to keep the number of data files to a minimum. Most of the storage systems see their performance degrade as the number of files increases.
- Groups related data according to the most likely access pattern, using some means of collocation. This minimises the number of tape mounts during retrievals.
- Decouples the logical view of the data from its physical view, so the data can be reorganised.
- Creates request queues, giving more priority to small interactive requests than to large batch ones, allowing VIP users and favouring retrieval from disk to retrieval from tape.
- Manages its own drive queues, to optimise their use, and minimise mounts and positioning.
- Is only dependent on one commercial software.

8. A few numbers

It has been observed that *the size of the archive is directly proportional to the power of the supercomputer*. This rule of thumb is used to project the growth of MARS.

MARS has been steadily growing at a rate of 60% per annum and nowadays contains more than 400 Tbytes of data, representing over $3x10^9$ individually addressable fields in less than a million files. Everyday, 4,000,000 fields (0.5 Tbytes) are added to MARS and 40,000 requests are dealt with from more than 500 users in Europe.

9. Conclusion

MARS is now a very mature system and its user base is increasing. It has been a very successful system that has evolved without impact on its users: a request written in 1985 is still valid in 2002.

MARS will continue to grow in size and diversity. Developments are underway to make its content more widely available, in particular to other scientific or commercial organisations.

HPSS is expected to scale to meet the Centre’s needs in the long term and there should be no need to go again through a back-archive exercise.