MARS Prototype Implementation at the Australian Bureau of Meteorology

T. Le (BoM), B. Raoult (ECMWF), and W. Bourke (BoM)

With the co-operation of ECMWF, the Australian Bureau of Meteorology has been examining a prototype implementation of the MARS software system. In this paper the Bureau's computing environment is outlined, aspects of problems in this implementation of MARS are explored, and finally the actual implementation of MARS in three stages:

1. Proxy MARS servers for model data on either the Bureau's NEC SX-4/32 supercomputer or on HP servers accessing the existing real-time data archive via Neons/ORACLE.

2. A full MARS server on IBM/RS6000 with the accompanying ObjectStore database.

3. In collaboration with NEC the development of a tape handling API to provide the IBM/RS6000 server with access to the Bureau's StorageTek SILO tape system coupled to the SX-4/32.

1. Overview of BoM computing and data management environment

The backbone of the Bureau computing environment is a supercomputer NEC SX4/32 series, on which all numerical models are run both in operations and in research. In operations the output are post-processed on the SX4 and sent to NMOC (National Meteorological Operational Centre), which runs a number of mid-range HP9000 machines, and uses Neons/ORACLE for the storage of model field data as well as observational data obtained through the GTS. On the SX4 all model output are in netCDF format and directly archived to SAM-FS tape system; SAM-FS runs on a Sun E4000 machine with 120 GBytes of Raid disk, two StorageTek 9710 library storage modules and twelve DLT7000 tape drives handling tapes of 35Gbytes capacity. NEC also provides a software called SX-BackStore which migrates files to the directly-coupled SILO tape, the SILO system currently has 4 TimberLine drives, though only 2 are connected to the SX4, (the other two to a Cray J90); these are old drives which handle 3490 tapes with 800 Mbytes capacity, the drives are currently being upgraded to 9840 series. SX-Backstore is currently under evaluation and its use limited.

Most RFCs (Regional Forecasting Centres) use IBM RS6000 machines which access the Bureau's central Neons/ORACLE database for real time data, and NCC (National Climate Centre) ADAM/ORACLE database for climate data. There is little support enabling RFCs to access data stored in the Bureau's deep archive on SAM-FS.

A significant amount of data in Neons/ORACLE is also duplicated and stored separately in McIDAS MD and GRID format, which is used extensively by RFCs and in head office; satellite data is held in McIDAS AREA files.

2. Data Management issues at BoM

When BoM migrated to the UNIX environment in the mid 1990's, NEONS/empress then NEONS/ORACLE was chosen for the storage of model and observational data, however due to capacity limitation only 10 days of post-processed, coarser-resolution recent data are kept online, older data are aged off to tapes and the process of recovering involves manual reloading of data back into the NEONS

database. This architecture has functioned satisfactorily for a number of years in NMOC, although now capacity limited with respect to recent high resolution model implementations. It does not meet the requirement of users working under research and development environment, where data outside the 10 days time window in full resolution is often required; for this reason BMRC staff routinely access SAM-FS for model data while NMOC and other branches of Bureau mostly utilise the rtdb (realtime database).

With the ever increasing computing power, and the accompanying larger amounts of data generated, the current NEONS database configuration is inevitably reaching its capacity limitation, there is clearly an urgent need for a significant upgrade in data management strategy. The Bureau is considering at present alternative software and hardware strategies that address the following issues :

- Seamless access to NWP output for users in all areas of the Bureau.
- Broader and easier access of NWP output.
- A system which can serve the new distributed data paradigm of future AIFS strategies.
- A cost-effective alternative to a significant upgrade of the current systems.
- Improved access to graphical display and visualisation software.

In this endeavour BMRC has identified MARS (Meteorological Archival and Retrieval System) which has been developed in ECMWF as a suitable candidate that has the potential of addressing most of the above-mentioned issues, and of providing much enhanced functionality in overall data management. It is considered that initially MARS can be used to archive model output, which is predominantly the major component of data stored in the current rtdb, thus alleviating the immediate pressure on rtdb, effectively reducing the resources required for the management of rtdb.

With this approach both operational and research staff will share the same MARS database, which will among other things enable a smoother transfer of research products for operational implementation.

3. Porting issues

The whole client/server MARS and Metview source code was provided to BMRC by ECMWF in Nov. 1997 for this prototype project. At BoM it was initially intended to implement the MARS server software on the SX4, taking advantage of the tightly coupled nature of the SILO tape system to the machine. It was, however, discovered that the C++ compiler did not support many of the standard features, in addition negotiation with Object Design Inc. failed to secure a license for ObjectStore database for the SX4. Effort was then diverted to implementing Metview and a netCDF to GRIB conversion interface to enable MARS-style access to the Bureau's model output in its native format.

Porting of the MARS server on a SGI platform though had less trouble, although there were still many C++ compiling problems. BMRC finally negotiated with BoM/COSB for the provision of an IBM/RS6000 machine to facilitate compatibility with ECMWF where the server software was implemented on IBM platforms.

Particular uncertainty was the MARS interface with the Bureau's STK SILO. Specification of an API was developed by B. Raoult in which the functionality of the tape handling of the server software was encapsulated enabling a decoupling of the MARS server from the underlying tape handling software. The development of this API has been undertaken by NEC and provided to BMRC under the auspices of a joint R&D project with the BoM.

4. Implementation of MARS at BoM

4.1 MARS proxy server

Initially, a MARS proxy server was set up on the SX4 utilising the open/read/write/close feature of the MARS client code, descriptions of a MARS request is used to assemble a fixed format file name on SX4, this file in netCDF format is then opened, requested data read and converted to GRIB format prior to return to the requesting MARS client. Here user is able to utilise the content-based retrieval and to access the SX4 file system data without the implementation of the proper MARS server.

The netCDF to GRIB conversion program has been written exclusively to deal with model data in BoM, namely model fields by either spherical harmonic coefficients, or Gaussian/lat-lon/thinned grids, on either sigma or pressure coordinates.

A number of MARS clients have been readily installed on various platforms. Metview has also been set up on a SGI machine for use within BMRC and on HP9000 for use by NMOC staff; the HP implementation also features an interface to the Bureau's NEONS/ORACLE database, enabling the retrieval and displaying of model data in rtdb.

4.2 MARS server with ObjectStore database

In September 1998, BMRC secured an initially evaluation runtime license of ObjectStore, enabling the full installation of the MARS server on a RS6000 platform; few problems have been encountered in implementing the software, apart from some memory initialisation problem which caused the server to crash.

With only 18 Gbytes of disk space and no facility to move data to the deep SILO archive, the main MARS server was initially only used to study how MARS functions with little practical application, apart from some research experiments including experimental ensemble predictions where we found Metview to be particularly useful in displaying results. We had also been able to demonstrate the transparency of data within MARS, if MARS fails to locate data on the SX4 proxy server, it will continue searching other "database" which in this case is the proper MARS server on RS6000.

In March 1999, BMRC secured an additional 27 Gbytes of raid disks, bringing the total of available disk space on the RS6000 to 45Gbytes. It was then decided that selected global model output be routinely archived in MARS, and to make MARS available to a wider community of users including NMOC and COSB. With the disk based archival and retrieval rate of up to four Mbytes/s, users who have used MARS have been impressed by its content-based, transparent and ready of access of data, as well as the efficiency in data retrieval.

A Java-based MARS browser has also been installed enabling users to browse through the content of data being archived under MARS, in addition NMOC is accessing the use of Metview/MARS for generation of BoM's daily prognostic charts.

4.3 Tape handling API

The MARS tape handling functions have been clarified and defined in a specification of an API (Application Program Interface) for the MARS server and the SILO system (as proposed by B. Raoult) and as developed by NEC who supported this work under the auspices of the BoM/CSIRO/NEC HPCCC R&D agreement.

In April 1999, NEC provided BMRC with the MARS/SILO API, called tiAPI with the prefix "ti" being for "tape interface", this tiAPI is client/server based and is written entirely in C. The tiAPI server, with routines to access SILO tapes directly, was installed on the SX4, while the tiAPI client was designed to be implemented on any platforms running UNIX; RPC (Remote Procedure Call) is used for client/server communication; a tiAPI client was installed on the RS6000 where NEC tested with standalone programs writing a disk file to tape, and reading back portions of it in various segments, as required in the API's specifications.

BMRC subsequently has endeavoured to integrate the tiAPI into the MARS server, following broadly the strategies of existing MARS ADSM routines, but with considerable simplification relative to the complex ADSM structures which have been removed. The only major deviation from the original MARS/ADSM implementation is that tiAPI requires a database of NEC's tape names and files stored on tapes, since tiAPI bybasses its SX-BackStore database in any read/write to tape, such information is not available.

The MARS/tiAPI was completed in Jul. 1999 where users can flush data off disk to tape and to read back in a standard MARS request, though this facility does not provide much practical use due to the present limited capacity of the tapes themselves; there are still a number of issues remained to be solved, most importantly these are

- tiAPI lacks the "append" functionality, this was due to the initial misunderstanding of the specs where we believed we needed to write to tape only once, without realising the process involved many disk files, resulting the tiAPI currently can handle only one disk file per tape

- tiAPI does not handle backup, currently backup is done by an offline process once a week

- Querying functions are still yet to be implemented

- A significant difference in the rate of archiving and retrieving is seen in performance testing; the rate is 0.3 to 0.5 Mbytes/s for archiving while in retrieving it drops to a mere 8 Kbytes/s, even when the tape has been pre-mounted; this matter is still under investigation

- It is still yet to test the API particularly the RPC communication channel under multi-threaded applications.

5. Future of MARS in BoM

The prototype has demonstrated both the feasibility of porting MARS to the BoM computing environment and the excellent functionality it delivers. The MARS project has the capacity to impact substantially on data management strategies within various sections of the Bureau, and on data delivery to RFCs.

It is confidently expected that initially the MARS capability is a valuable software system for supporting the BMRC/NMOC data management of research and operational NWP model output. Extending its utility within the BoM environment to handle observational data within the BURF context, as indeed is robustly done at ECMWF, is a strategy that will be considered subsequently.

For a fully operational MARS it is expected that a significant hardware upgrade is required, this will involve the upgrade of tape drives to handle Eagle 9840 tapes which have a much higher capacity of 20 Gbytes, as well as a more substantial front-end server which should be directly coupled to the STK SILO.

Acknowledgements

The authors gratefully acknowledge the technical support provided by NEC in developing the tape handling API under the auspices of a joint R&D project with the BoM.