

MARS: ECMWF ARCHIVE

Baudouin Raoult

Meteorological Applications

European Centre for Medium-Range Weather Forecasts

HPSS User Forum, Indianapolis June 2002



MARS: A managed archive

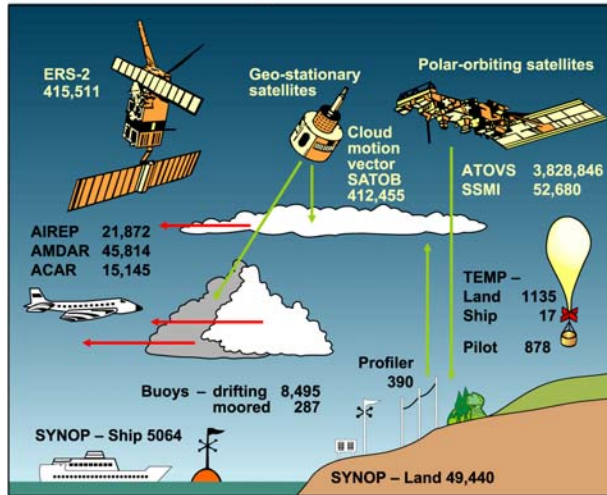
- **Meteorological Archival Retrieval System**
- **16 years of existence**
- **Retrievals expressed in meteorological terms**
- **Post-processing facilities**
 - Interpolation between various data representation
 - Interpolation on coarser grids
 - Sub-area extractions
- **Data in GRIB and BUFR format according to WMO standards**

HPSS User Forum, Indianapolis June 2002



Observations

Example 24 hour summary of observations received at ECMWF

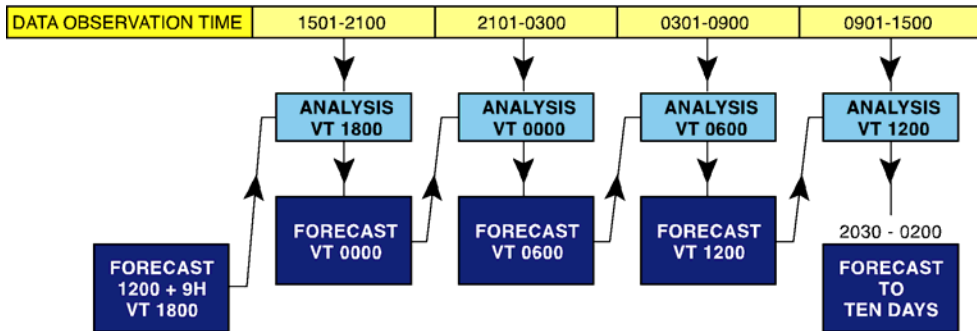


HPSS User Forum, Indianapolis June 2002



Assimilation

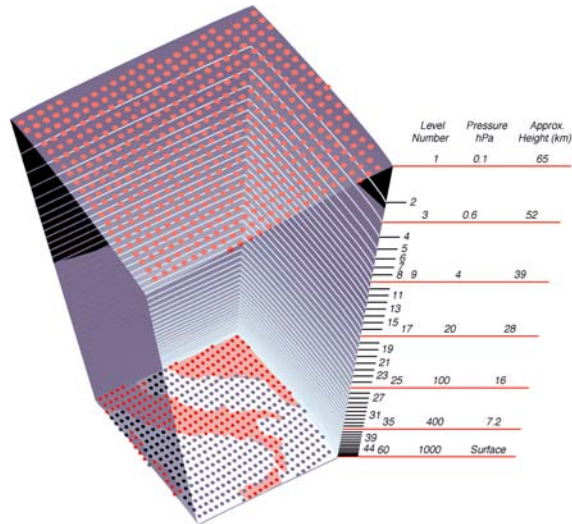
Operational Data Assimilation - Forecast Cycle



HPSS User Forum, Indianapolis June 2002



What is a field



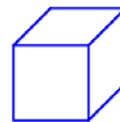
HPSS User Forum, Indianapolis June 2002



MARS Design: Field

A meteorological field

- The smallest addressable object
 - Class (Operational)
 - Version (1)
 - Stream (Daily archive)
 - Parameter (Temperature)
 - Level (1000 hPa)
 - Date (1993-08-10)
 - Base Time (12Z)
 - Time Step (120 H)
 - Member (42)

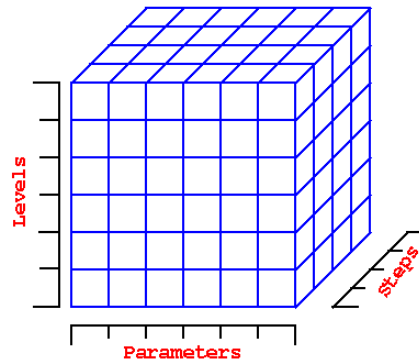


HPSS User Forum, Indianapolis June 2002



MARS Design: Cubes

Archive objects are hypercubes

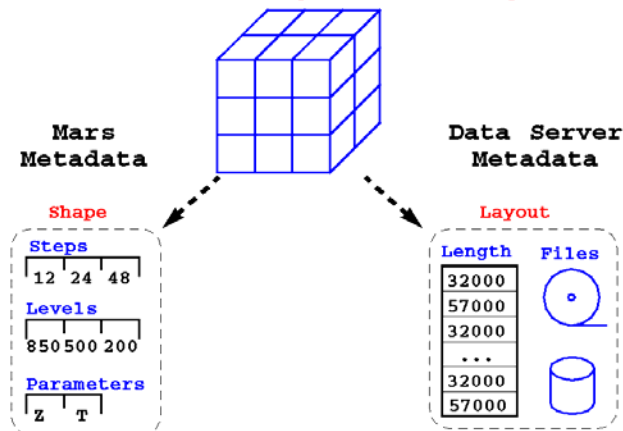


HPSS User Forum, Indianapolis June 2002



MARS Design: Metadata

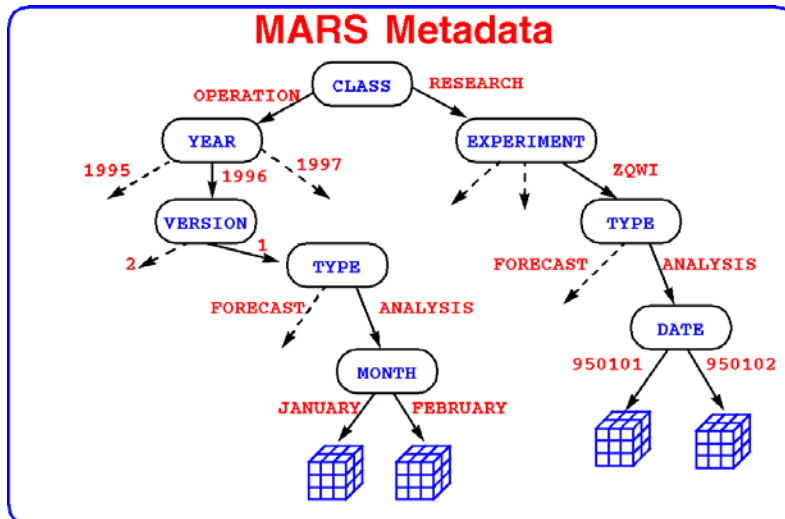
Metadata: Shapes and Layouts



HPSS User Forum, Indianapolis June 2002



MARS Design: Tree



HPSS User Forum, Indianapolis June 2002



A meteorological language

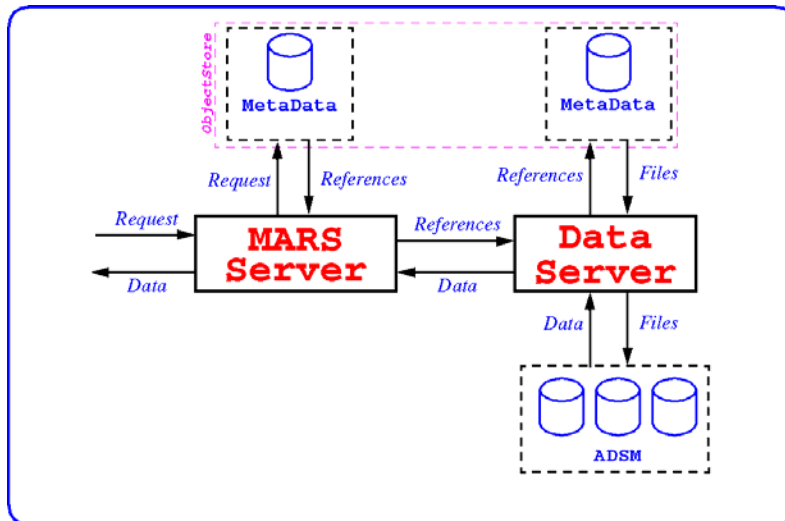
Retrieve,

date	=	20010101/to/20010131,
parameter	=	temperature/geopotential,
type	=	forecast,
step	=	12/to/240/by/12,
levels	=	1000/850/500/200,
grid	=	2/2,
area	=	-10/20/10/0

HPSS User Forum, Indianapolis June 2002



MARS Design: Architecture



HPSS User Forum, Indianapolis June 2002



MARS manages its own caches

- **“Pre-archive” space**
 - Data is first stored there as produced by the model
 - Efficient archival
 - Allow incremental archiving
 - Data is then sorted and aggregated into large tape files
 - Efficient retrievals
- **Retrieval caches**
 - Field level caching, only small parts of files are cached

HPSS User Forum, Indianapolis June 2002



MARS manages its own queues

- **Three queues:**
 - **User requests queues**
 - **Tape read queues**
 - **Tape write queues**
- **One user request can create several tape read requests**
- **Read requests are sorted according to volume and position**
 - **All possible requests for a volume are processed**
- **Write requests are sorted by families**

MARS manages its own queues – cont.

- **A fixed number of tape drives is allocated for reading or writing**
 - **Queues and disk spaces are monitored**
 - **Results are fed into a fuzzy logic algorithm**
 - **Drive allocation is adjusted accordingly**
- **Better control**
 - **Minimise tape mounts**
 - **Optimise tape drive usage**
 - **Priorities (serve VIPs first)**

MARS/HPSS first impressions, pros

- **HPSS fits the MARS design**
 - **Direct writes to tapes**
 - **Partial get from tapes**
 - **Families**
- **Simple and comprehensive POSIX API**
- **Same concepts as TSM**
 - **Storage Pool = Storage Class**
 - **Management Class = Family**

MARS/HPSS first impressions, cons

- **No high level transactions**
 - **Need garbage collection to delete incomplete files**
- **API not resilient to errors**
 - **Cannot have non-stop processes**

MARS/HPSS Code migration

- **MARS is highly object-oriented, based on three abstract classes:**
 - **Tape manager (create, remove, exists, location, ...)**
 - **Tape file (size, last access, ...)**
 - **Data stream (open, read, write, close, partial reads)**
- **Done in a week**
 - **Implemented the three subclasses for HPSS**
 - **No other code change**

MARS/HPSS Back-archive (Data Migration)

- **MARS will use TSM and HPSS simultaneously**
 - **Some users requests will access both storage systems**
 - **No interruption of service (no data unavailable)**
- **Copy will be done file by file**
 - **Using TSM API and HPSS API in the same process**
 - **Several streams in parallel**
 - **Sorted by volumes and families (optimise reads and writes)**
- **Migration is expected to last two years**

MARS/HPSS Wishes

- **Tape to tape backups**
- **Possibility to query on which volume a file offset resides**
- **No DCE token expiry**

MARS: A Few Numbers

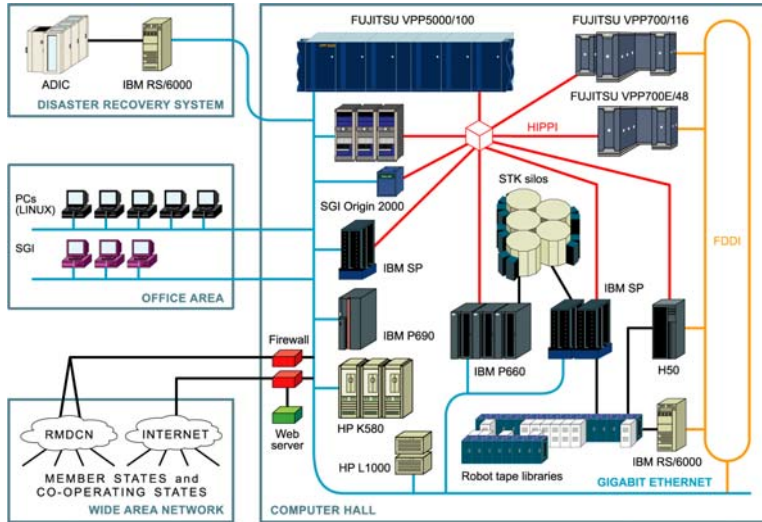
- **560 active users, at ECMWF and in the Member States**
- **40 000 retrieval requests a day, 2 000 000 fields**
- **4 000 000 fields added daily (0.5 Terabyte)**
- **About 400 Terabytes**
- **More than $3.5 \cdot 10^9$ meteorological fields**
- **Analysis from 1980, Forecasts from 1985**
- **After ERA40, analysis and observations since 1957**

Francis...

Some changes that we would like to see...

- HPSS will be used to archive data which does not fit in the MARS model. This will be done through the use of an home written interface which will “mask” HPSS from the end-users.
- We would like to propose some Change Requests accordingly:
 - Ability to define some servers user-exits to authenticate and user transactions outside the supported schemes.
 - Full guaranteed support for the HSI interface.
 - The ability to Aggregate and manage small files collections, especially when these are stored on tape.

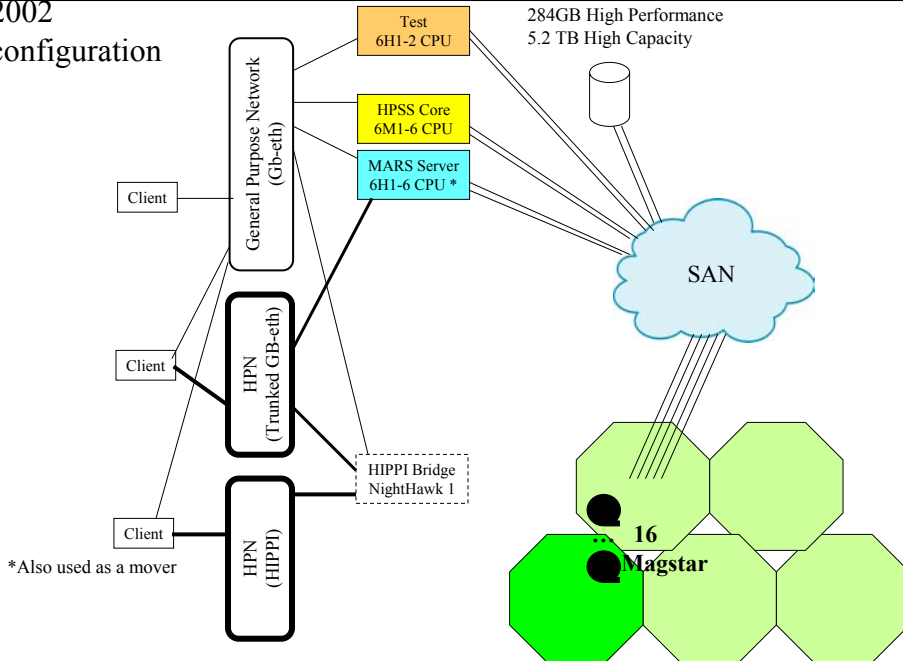
Computer hall



HPSS User Forum, Indianapolis June 2002

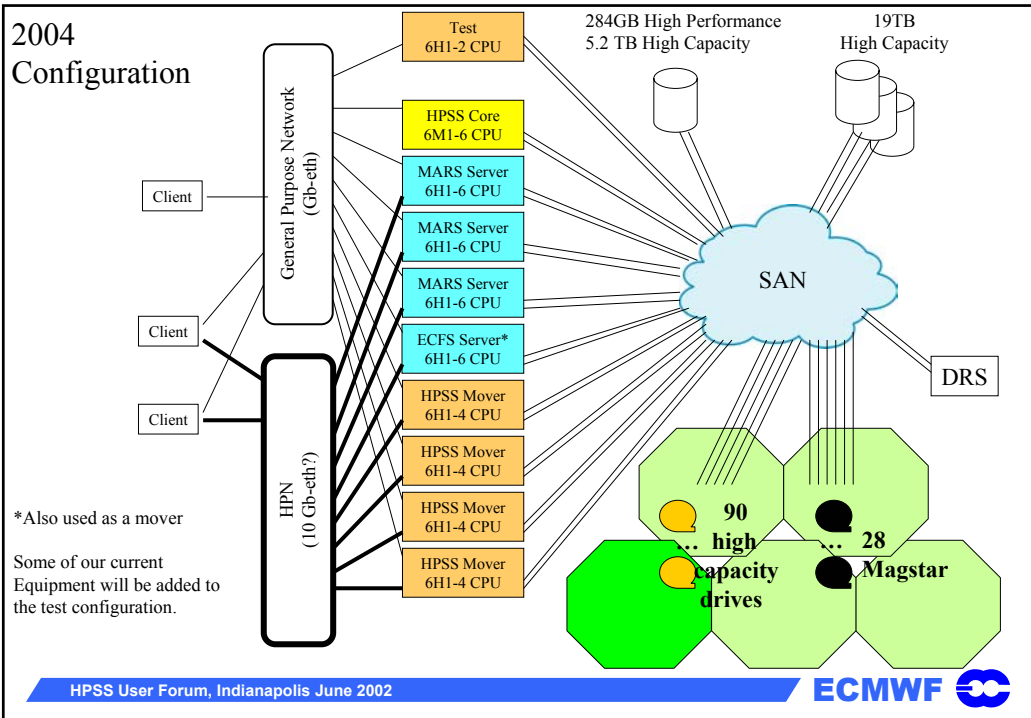


2002 configuration



HPSS User Forum, Indianapolis June 2002

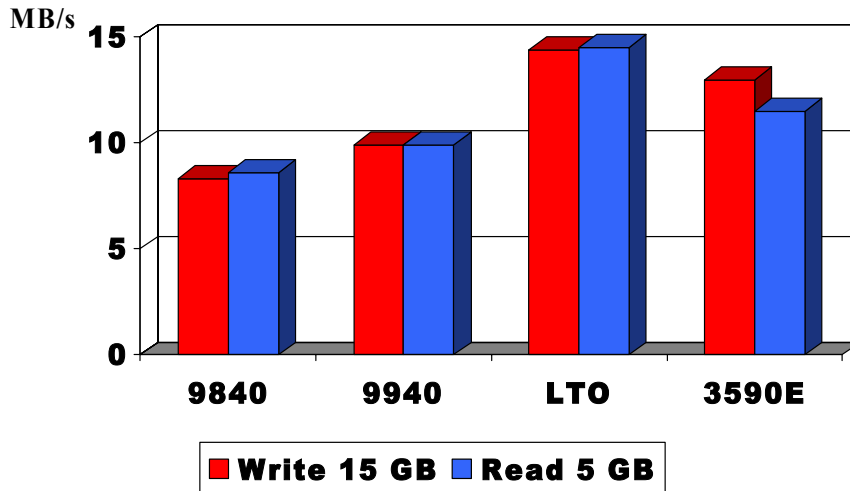




Tape drives Positioning Tests

- Done last year during evaluation of our data handling replacement.
- How do various tape drives behave in a positioning intensive environment.
- Tests done through home written C program, not necessarily optimised for high bandwidth read and writes.
- Test does not take in account the mount, dismount, load and unload drive characteristics.

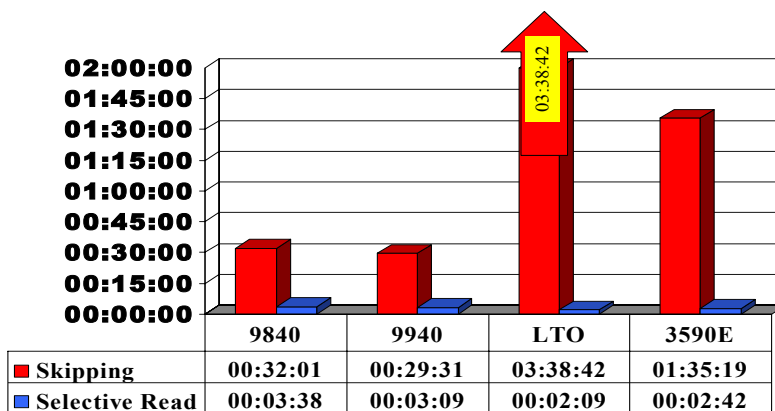
Sequential I/Os.



HPSS User Forum, Indianapolis June 2002



Retrieve fields separated by short intervals.

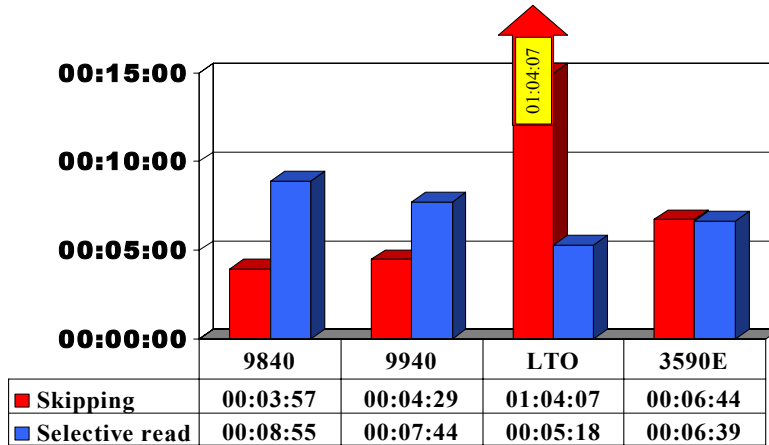


Retrieve 6000 130 KB fields, separated by 130 and 300KB intervals.

HPSS User Forum, Indianapolis June 2002



Retrieve fields separated by long intervals.



Retrieving 60 130 KB fields, separated by 50 and 90 MB intervals.

HPSS User Forum, Indianapolis June 2002



Conclusions.

- 3590 and LTO tape drives are the best suited for long sustained I/O operations.
- LTO unsuitable for positioning intensive operations.
- STK provides the positioning intensive access tape drives.
- One needs to investigate when it becomes more efficient to “read and discard” unnecessary information, instead of skip it.

HPSS User Forum, Indianapolis June 2002

