INVITATION TO TENDER

ECMWF/ITT/2025/382

Acquisition of

An Advanced Hierarchical Data Storage Management System

Volume II:

Specification of Requirements

ISSUED BY: ECMWF Administration Department Procurement Section
Date: 03 December 2025
Version: Final

Table of Contents

Introduction – Context and Background

Introduction

This Invitation to Tender (ITT) has been prepared by the European Centre for Medium-Range Weather Forecasts, (governed by its Convention and associated Protocol on Privileges and Immunities which came into force on 1 November 1975, and was amended on 6 June 2010) ("ECMWF") for the purposes of obtaining proposals from Tenderers for licensing and support services for an advanced hierarchical data storage Management system and related consultancy services.

ECMWF Data Centre

ECMWF has one of the largest IT data storage facilities in the world comprising its main Data Handling System (DHS), which primarily stores observational data and data generated by its High-Performance Computer system, and various ancillary storage systems including network attached storage. The equipment is currently from diverse manufacturers, including but not limited to Brocade, Dell, HPE, IBM, NetApp, Spectra Logic and Western Digital. Following sections present a brief introduction of ECMWF computing facilities, which can be complemented with further information online at: https://www.ecmwf.int/en/computing/our-facilities

ECMWF's HPC Facility

ECMWF's High Performance Computing Facility is based on an Atos Sequana XH2000 based service. This is used to run large mathematical models allowing the Centre to predict the weather worldwide over periods of several weeks. *A replacement for this system is in the process of being tendered*.

ECMWF's Storage Systems

ECMWF maintains a large digital archive of weather-related information which holds over an exabyte of data, the bulk of it being pre-compressed and scientific in nature. This data is stored in a tiered environment controlled by High Performance Storage System (HPSS). Most data are stored on tape media, with only 9% of the data residing on disk. HPSS is highly distributed software, making use of many servers to transfer data between disks or tape drives, with users archiving or accessing the archives, via a high-performance network.

HPSS is a high-performance, hierarchical storage management software solution offered by IBM Consulting. HPSS was developed jointly by IBM and several US Government sponsored laboratories, with IBM Consulting having rights to license HPSS and provide related services.

HPSS is primarily used by two applications, both developed at ECMWF; MARS and ECFS. Users are sheltered from the complexities of the underlying storage by interacting with the two applications:

- MARS, the Meteorological Archival and Retrieval System, provides access to a powerful abstraction engine that allows the thousands of registered users to access the meteorological data that has been collected or generated at ECMWF for more than 40 years. MARS stores GRIB and BUFR data, abstracting from its users all the details concerning the physical location and internal organisation of this data. It manages its own set of disk caches for managing data that has been recently archived or retrieved.
- ECFS, the ECMWF Common File System, provides users with a logical view of a seemingly very large file system and is used for data in formats that are not suitable for storing in MARS, eg NetCDF checkpoint files of log files. UNIX-like commands enable users to copy whole files to and from any of ECMWF's computing platforms.

MARS data represents about 73% of the volume of data, with ECFS representing the remaining 27%. In a typical day the archive grows by 620 TB.

Description of existing system

The existing Data Storage Management System (DSMS) is based on IBM's HPSS (High Performance Storage System). A very scalable storage system implemented on a set of x86 based servers running Linux (RHEL 8 for the Core and Rocky Linux 8 for the data movers). A single large server acts as the HPSS Core, with hot spares available in both data halls (Data Storage 1 (DS1) and Data Storage 2 (DS2)), see Figure 1: Hardware Layout of Current System. For access to the disk and tape devices, a set of around 20 HPE DL360 Gen10 and Dell R630 servers are used as HPSS disk movers and ~100 HPE DL380 Gen10/11 and Dell R440 servers as HPSS tape movers. A mover is a server managing the I/O traffic to a storage devices, either disk or tape.

The disk is currently provided by block storage from IBM FlashSystem 5035 and 5045 disk systems. Currently 50 arrays each with 120 hard disks, spread over both halls. Most being 1PB in size but some 1.6PB providing higher capacity for cached field data already written to tape. A further 24PB of Ceph storage is provided to MARS as disk cache for its operational data. ECMWF is trailing a Dell PowerVault disk system to test its characteristics in production. Although HPSS manages most of the disk space for ECFS, it does not for MARS. MARS is allocated disk file systems directly, managing their usage by migrating file objects to tape freeing up disk space. The reading of MARS data from HPSS is direct from tape, back to its own disk cache file systems. Normally MARS targets only discrete areas within a file object to read and only partially reads the files on tape. This efficiency gain, will need to be replicated in some form as retrieving whole file objects to disk to later hand to MARS will reduce both performance and disk cache efficiency.

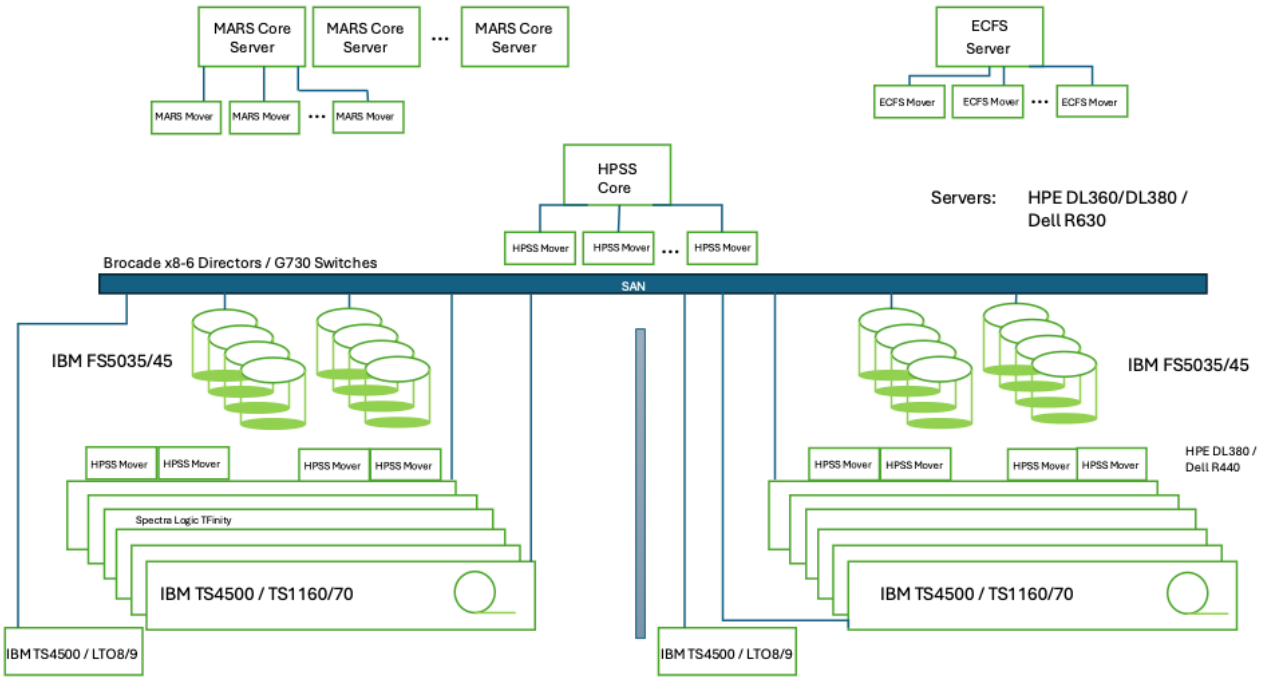


Figure 1: Hardware Layout of Current System

The tape environment is eleven IBM TS4500 and one Spectra Logic TFinity tape libraries, together acting as the primary tape storage layer, see Figure 2. Each of these libraries is made of 10 frames with at least 60 tapes drives (either IBM TS1160 or TS1170). This gives a total slot count for 90,000 tape volumes. Beneath this are two IBM TS4500 8 frame libraries providing space for LTO 8 and 9 drives and media to maintain a secondary copy of critical data. Most of the tape volumes (around 30,000) in the secondary system are outside of the libraries on shelves and are tracked by HPSS, if ever needed to be read.

The Centre's usage of HPSS matches a similar pattern to other large meteorological sites. The disk layer is very busy but the random nature of access across all the data means that the tape drives and libraries are very heavily utilised. It has been observed by IBM that these might be the busiest libraries in world.

Figure 2: IBM TS4500 Tape Libraries in ECMWF Data Centre

The archive grows at around 620TB per day of primary data. This is written through disk and onto tape media. Currently ECMWF has 740 TS1160 and TS1170 drives and 80 LTO-8 and LTO-9 drives. The high TS11xx enterprise drive count is caused by a current media transition to TS1170 and should return a more normal 400 primary drives, once this migration completes. Most of the drives are occupied 24 hours per day, with around 60% of them being used to read data, which is approximately 200TB/day. On average the 12 libraries achieve around 20000 mount cycles per day.

The LTO drive and library is much lighter. Not only are there fewer drives (80), but the workload is also almost exclusively the writing of around 36% of the primary critical data, ~220TB/day. The only reading of this secondary data is to recovery damaged primary media, migrate to a new LTO generation or a real disaster recovery. As the read expectation on tape volumes is relatively low, most of the media is exported and held on shelving, rather than having to maintain them all within automatic tape libraries. This model will probably have to change if this secondary copy of data is to be moved off-site.

Alongside this tender, ECMWF is looking at moving this secondary copy of data off-site to improve the overall safety of the archive. Part of this tender will examine the options available against each tendered system. In

particular, to see how easy it would be to design and operate the system remotely without ECMWF staff. The repository at an off-site location would primarily be the secondary media completely contained automated libraries, together with the control and I/O servers to access them. It may also contain some online storage assets as an initial storage layer to provide some buffering and allow for network outages.

Each MARS service and ECFS share a similar scaling model to HPSS. All of them use a single host as the Core server and farm the I/O workload out to a pool of mover hosts. Technically the HPSS Core can be split over multiple hosts, but this has not been necessary so far with ECMWF's implementation. The IBM database system DB2 which underpins the HPSS metadata, is partitioned but these are currently running on the same server. ECFS also differs slightly in that it mirrors its metadata over two hosts, both running postgres. All MARS and ECFS services are running Rocky Linux 8. MARS using around 100 and ECFS 20 servers.

The plot below, Figure 3: Estimate of Primary and Secondary data holdings 2022 – 2035 shows the expected growth profile up to 2035. The two inflection points are the result of changes to the HPC systems in 2027 and 2032. This are only a guide, as the phasing in of HPC generations can introduce unexpected data growth.

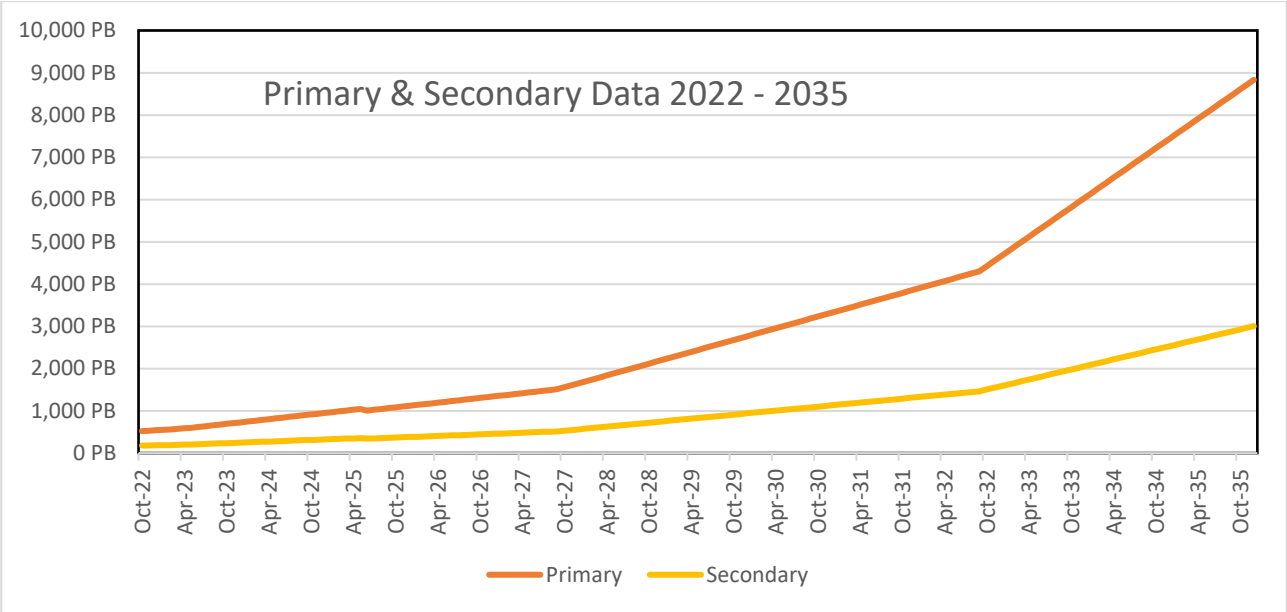


Figure 3: Estimate of Primary and Secondary data holdings 2022 – 2035

Table 1: Expected End of Year Data Capacities (2026 – 2035) shows the numeric expected values for the end of each of the coming years. This may be useful in calculating any licensing and support costs.

Year End	Total Primary Data (PB|PiB)		Total Secondary Data (PB|PiB)	
2026	1342	1221	456	415
2027	1649	1500	561	510
2028	2207	2007	750	682
2029	2765	2515	940	855

2030	3323	3022	1130	1028
2031	3881	3530	1319	1200
2032	4648	4227	1580	1437
2033	6043	5496	2055	1869
2034	7438	6765	2529	2300
2035	8833	8034	3003	2731

Table 1: Expected End of Year Data Capacities (2026 – 2035)

Objectives

- This ITT is expected to result in the setting up of a contract to license and support a multi-tiered storage solution, to underpin ECMWF's in-house storage applications MARS and ECFS for a period of at least ten years.
- Additional support with initial installation, configuration and possibly migration from HPSS will be required.
- Achieving the best value in the long-term for the Centre to continue its storage growth and usage expectations. This will be measured over the period of the contract.
- Primarily looking for a software solution, however;
    - There may be recommendations or requirements from tenderers of hardware where this is seen to optimise the system as a whole.
    - If a tenderer does propose hardware, they will need to provide their commitments on performance targets based on this hardware (in replacement or combination with the model hardware, as detailed in Annex 3 Requirements of Acceptance. Explain the reasoning for not using commodity hardware. Explain how the new DSMS hardware will co-exist with current hardware.
    - Provide suggestions on how, in the future, the hardware should be extended or replaced to accommodate the expected growth in activity and capacity.
    - Whether the tenderer provides hardware or not, any new DSMS will predominately run initially on new hardware. The primary exceptions to this will very likely be the SAN infrastructure and the IBM and Spectra Logic tape libraries. These would need to be shared. However, it is envisaged that during any migration, hardware that was used by HPSS would be put back into use again under the new DSMS. Whether that be NVMe appliances, disk systems or servers. Annex 3 Requirements of Acceptance will outline the model hardware that ECMWF is considering as the benchmark hardware to run acceptance tests upon.
    - If there are other recommendations or requirements, these may be accommodated. A form of loan or purchase of such equipment would need to be discussed closer to any acceptance test.

The software will be running on equipment installed in the ECMWF data centre, at:

ECMWF

Tecnopolo di Bologna,

Via Stalingrado 84/3,

40128 Bologna,

Italy

The contract award is subject to approval by ECMWF's governing body, its Council will meet mid-June and December in 2026.

Tenders are welcomed from suppliers based in any country. However, a preference will be given to tenderers from any ECMWF Member State or Co-operating State country[1], (subject to supply and support requirements).

Term of the License and the Support

Tenderers are asked to provide ECMWF with a license and support for the term from the contract. The Tenderers are not expected to have an option for termination for convenience during this term.

Preferably the proposal should include options for ECMWF to terminate the license and/or the support within this term. If termination before any particular date would result in termination charges these must be fully explained together with the timescales.

The proposal should include options to extend the term beyond the initial ten years. Tenderers should specify the notice period required to extend the support, the duration they are prepared to extend the support by and any limits on the number or length of any extensions.

Tender requirements

Points of specification are categorised by the bold notations M, H, D or R to the left of the pertinent section:

Points of specification are categorised by the bold notations **M, H, D** or **R** to the left of the pertinent section:

M	denotes a MANDATORY requirement: a requirement that must be adhered to, or a performance requirement that must be met in order that the tendered solution can be accepted by ECMWF. ECMWF will not consider a tendered solution that fails to meet a mandatory specification requirement (marked M) unless the tenderer offers valid reasons why the feature in question is either unnecessary for, or irrelevant to the tendered solution.
H	denotes a HIGHLY DESIRABLE requirement. The degree to which a tender meets the highly desirable requirements (marked H) will be a key factor that will be taken into account in selecting the winning tender. If offered, the feature must be included in the overall price for the tendered solution.
D	denotes a DESIRABLE feature. The extent to which any tender offers features listed as desirable (marked D) will be one of the factors taken into account in selecting the winning tender. If offered, the

[1] Those Co-operating States that afforded privileges and immunities to ECMWF are Latvia, Bulgaria, North Macedonia and Lithuania.

	feature must be included in the overall price for the tendered solution.
R	denotes a REQUEST for information. A response must be given to all such requests. Requests for information (marked R) are intended to provide a description of the construction, philosophy, operation and the cost implications of the tendered solution in areas that are regarded as being of particular importance. A clear response to such requests will be of assistance to ECMWF in the tender evaluation process.

Tenderers should note that, where relevant, when responding to points of specification that are met, tenderers must give sufficient detail to explain the way in which the requirement is met - a simple expression, such as “compliant” or “agreed”, will not normally suffice.

Any additional features not listed in the ITT as requirements, but which the tenderer feels may be relevant, should be supported by descriptive material.

Operational Context

At a high level, ECMWF runs two large, persistent data archives. Here we give an overview of the “structured” archive (MARS), in which all data is self-describing according to carefully curated data governance and its resultant schemas, and handled in a consistent manner, followed by an overview of ECFS which provides the “unstructured” archive.

The interfaces to HPSS by the applications are POSIX-style API calls, with extensions to allow it to build access strategies to read tape volumes efficiently and improve parallelisation of writing to tape.

The MARS Application

MARS manages its own disk cache, into which data are archived and then reorganised before being stored on tape, or into which data is retrieved before being sent to users. Some fundamental functionality is required from the DSMS in order to provide this service:

1. It is helpful if the interface to the DSMS supports direct reading and writing of data from tape media. At the least it is important to be able to manage in which tier data is held in the system, and how and when data is moved between tiers, such as between disk and tape layers.
2. It is necessary to be able to determine where in the storage system a specific object is held, and in particular to identify the tape and the tape library in which a data object is held. This information is used to optimise the order in which data is retrieved, minimise tape mounts and minimise seeking movements performed on the tape. It is also used to help manage queues, and track when data is unavailable.
3. Partial retrieval API support. MARS data is typically stored in large objects on tape. MARS relies heavily on the ability to read back multiple small subregions, sparsely, without retrieving the entire data object. Typically, only a small percentage of each object is retrieved. This is a requirement only for the primary copy of the data.

MARS also tries to minimise the number of tape mounts required to process a user request by sorting data by type into separate sets of tapes (also known as pools or families). As a result, a user request that needs access to hundreds of objects is likely to require access to only a few tapes storing data of the same type,

rather than requiring a large set of tapes where different types of data are all mixed but data per tape are all of a similar age. As an example, all reanalysis data over a set time being placed together on a few tapes.

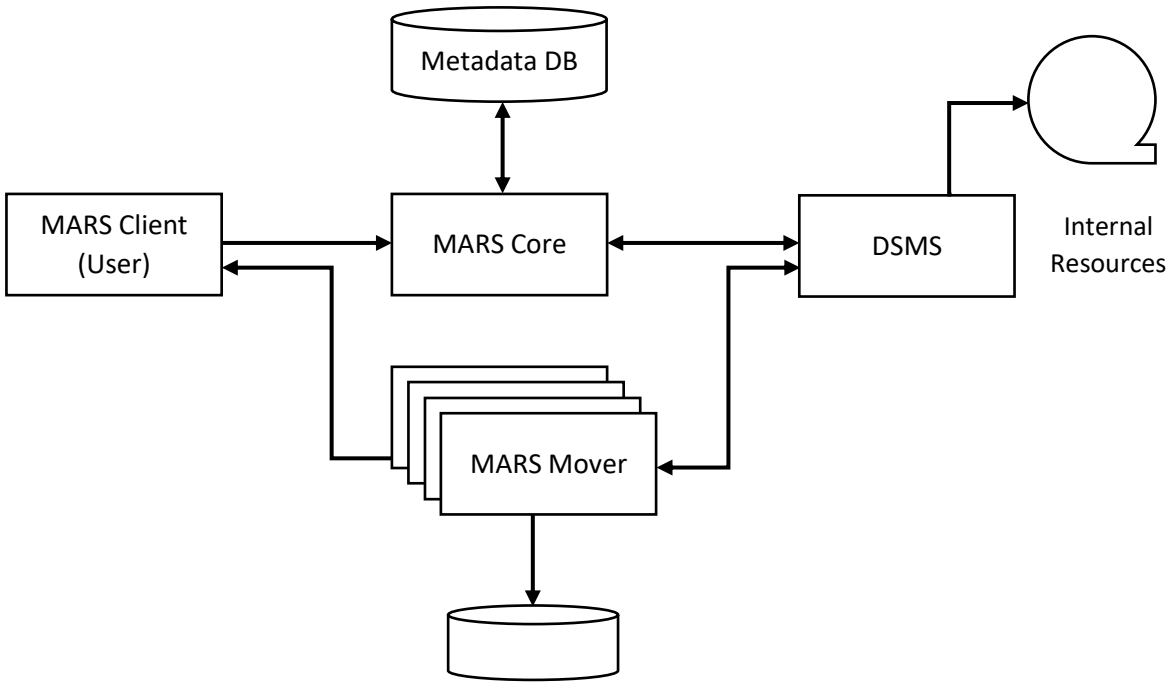
Note that whilst MARS currently uses a file system paradigm to store its data, where billions of meteorological fields are grouped into a relatively small number of files for storage, it could be straightforwardly adapted to make use of an object storage system, in which data are stored as binary objects referenced by unique identifiers. Under an object storage paradigm, the ability to retrieve (many) subranges from a single object would still be required.

Within the MARS system three layers of cache are explicitly used:

- The incoming cache (prearc), where data created on the HPC is kept until all data gets aggregated, such as a complete experiment or a meteorological forecast
- The outgoing cache (cache), where data read from tape is placed, before being streamed to the user client
- A long-term cache (locked), an outgoing cache layer with administratively controlled ingestion retention behaviour, meant to keep very popular data online without eviction policies.

This caching is managed and controlled by the MARS software and is not part of the underlying DSMS. In contrast to other systems (including ECFS) this sharply reduces both the need and the value of internal caching within the DSMS from the perspective of the MARS service.

Please note that externally to the MARS system, ECMWF operates a caching layer within the HPC. This layer absorbs recent output from the forecasting model, typically about 3 days of production. These data are also the most desirable for downstream users. As such, most requests for recently produced meteorological data never reach the MARS system, which is more oriented towards access to older archived data and research data. This strongly impacts the structure of the caching used in MARS.



The MARS Service in the DHS is currently split in 6 instances, each of them with a Core server and a number of data movers. The cache capacity is directly attached to these data movers. The MARS instances are sized according to their load and access patterns. In total they include 69 data movers, managing 15 PiB of disk space together with 24PB of Ceph cache storage for the operational MARS system. MARS holds 750 * 10^9 meteorological fields, comprising 650 PiB of primary data across 25.6 million files.

Users of the MARS system run approximately 490,000 requests per day against MARS in the DHS, of which 35% are for storing data and 65% are retrievals. Around 340 TiB are retrieved per day, while the overall archive grows at an average of 375 TiB per day.

The ECFS Application

ECFS provides ECMWF's unstructured archive. In this capacity, ECFS contains a wide variety of data types, with wildly varying access patterns. Data are owned by a large variety of users and held for lifetimes varying from hours to indefinite and permanent archival. As a result, there is a lot of potential scope for intelligent cache behaviour.

From the user's perspective, ECFS provides functionality similar to a POSIX filesystem accessible only via command line tools. The service provides users with functional analogues of the ls, cp, mv, rm, touch, chmod, chown, chgrp, cd and pwd commands. Data are owned by unix users and groups and accessed according to a standard POSIX filesystem permissions model.

This functionality is provided by a set of ECFS servers, maintained in the DHS, and separate from the systems operating the DSMS. ***Figure 4*** presents an overview of the structure of the ECFS service. All the metadata about user-visible files and paths is maintained in a high-availability Metadata Database, which is used by the ECFS Core server to present the service. The core server handles incoming requests from clients and all purely metadata operations. Transfers of data are delegated to specialised data mover nodes. Both the core server and the data movers interact with the DSMS to query and modify state, and to transfer data.

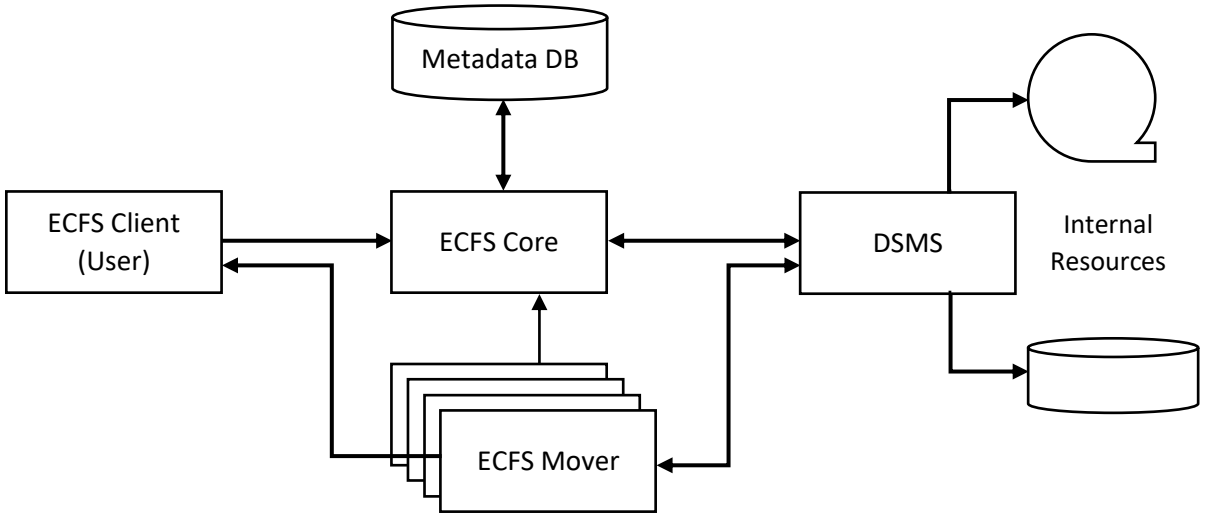


Figure 4: Structure of the provision of the ECFS service at ECMWF

The current operational setup includes one core server and 8 data mover nodes. Hot spares and synchronised replicas are available.

Internally the DSMS is ultimately used as an object store. All data belongs to the service user which operates the ECFS service and owns the unix processes running on the ECFS Core and Movers. Each user file is stored as one (or more) objects within the DSMS, referenced by a unique URI which can be constructed from information in the Metadata Database. Notably, if a user-visible path is reused (for instance if a file is overwritten) this results in an object being stored with a different internal URI, such that ECFS is able to

maintain a short-term versioned history of its contents and support recovery of inadvertently removed user files.

In the current operational setup, the URIs used to store the objects are derived from the user path combined with a unique numerical object identifier and an identifying “family” (which is used to partition our data into broad classes, such as data from the operational forecast, research, member states or external projects). The family is used to identify groups of tapes, such that related data and data with similar access patterns are grouped together on tape. This approach means that in the extreme case of the loss of the Metadata Database, its contents could be (partially) reconstructed by analysing the URIs in the DSMS to minimise loss of important data. Whilst attractive, having unix path-like URIs is not required of the DSMS.

ECFS uses the DSMS for the vast bulk of its storage. The choice of storing data on disk or tape is delegated to the DSMS, along with the responsibility for choosing the number of copies of the data and how data is copied, migrated and cached. ECFS is able to provide hints to the DSMS at file creation time to indicate the priority of the data, whether a secondary copy is required and the likelihood of data being retrieved. In particular, a significant slice of the data has only a very limited lifetime in the system (typically 90 days) and this can be indicated up-front to support appropriate storage choices.

ECFS also has the capability to manage disk-based storage outside of the DSMS, and this is used (in addition to caching within the DSMS) to manage long-term, locked caches of popular data to reduce predictable load on underlying removable media-based storage.

ECFS currently contains 750 million user supplied files, for a total of 1 billion inodes, aggregating to 250 PiB. These are being written at a rate of 300,000 files (200 TiB) per day. As data written to tape are migrated efficiently internally to the DSMS, the read workload has a larger impact on the DSMS capabilities. Currently we serve 100,000 files (100 TiB) per day to users. The daily data volumes being stored and retrieved have typically grown at approximately 40-45% per year.

For ECFS which writes data to HPSS, again using the POSIX-style API, HPSS manages the capacity levels in the disk cache by migrating to tape as files become available as candidates to move and separately purge these files from the cache as it maintains its warning and critical thresholds on the disk pools.

Specification of Requirements

Overview of a Data Storage Management System (DSMS)

The ultimate purpose of the Data Storage Management System is to provide a solid foundation on which efficient and reliable operation of the ECFS and MARS services can be based. This demands that the Data Storage Management System proposed in response to this ITT should be able to support an application-specific data management service in an efficient fashion via published and supported interfaces.

The DSMS will be responsible for maintaining the central data resource of ECMWF. It must take all necessary steps to preserve the integrity of the data, despite the inherent unreliability of the media and devices on which the data is recorded. It must continue to offer services through foreseeable equipment and media failures as can reasonably be circumvented.

The DSMS will receive data from and deliver data to MARS and ECFS server processes situated on the LAN network. Moving data between the media and devices at its disposal to optimise the use of resources and provide a reliable, efficient and consistent service to the client processes. It also allows the progress of data and requests through the system to be monitored and controlled by operators and systems analysts.

M(1): Tenderers must explicitly undertake to provide, or to arrange provision of software maintenance, including any third party software, for a period of not less than 10 years from the date of acceptance. Responsibility for the provision of such maintenance and support rests solely with the tenderer

If a tenderer proposes to sub-contract another company to provide software maintenance services, this must be clearly stated, and full details of the company shall be given.

Support requirements

Tenderers shall provide responses to the following requirements in the context of the complete System. This includes information related to the operating system, Data Storage Management System (DSMS), and the disks, removable media drives and robotic subsystems.

Tenderers are required to provide support with the following functions:

1. Support of software
    a. Provide fixes for software defects
    b. Provide patches and enhancements
    c. Explain any scheduled maintenance which would be needed whether the DSMS needs to be shut down or not. For example, maintenance on the metadata of the system, either in normal operation or during upgrades to the software.
2. Help with problems, upgrades, planning, status, questions
3. Problem reporting, determination, and resolution,
4. Status calls as mutually agreed
    a. During the status calls, review the state of the operational system and work with ECMWF to provide early identification of problems before they become severe.
5. Analysis of the configuration to optimise system performance.
6. If any hardware is provided as part of this solution, (non-critical) hardware support should include:
    a. Providing hardware break/fix with 24/7 4hr response time support coverage for all of the equipment,
    b. Providing firmware upgrades and patches to address defects,
    c. Expected average repair times for each hardware system supplied ,
    d. Explain any preventive maintenance which would need to be scheduled or other mandatory shutdowns of the hardware,
    e. Cover any other unexpected hardware intervention.

Provide proactive support as follows:

1. Assign a primary contact and support representative to ECMWF.
2. Provide remote monitoring by logging on to the system and checking status and settings. This monitoring will be for the purposes of problem resolution.
3. Work with ECMWF to plan upgrades to new releases as they are issued and assist with their installation.
4. Review ECMWF plans for changes in operating systems/distributions, processors, network interfaces, storage devices, and configuration for impact on system stability and participate in configuration reviews.

Problem resolution method should consist of the following:

1. Respond to software problem reports submitted by ECMWF or potential problems identified within two hours of receipt.
2. Log onto the ECMWF system and investigate problems in depth.

3. Recommend and implement or help implement changes that will restore the system to useful operational service, including a change in procedure or recommend the move to a more current version of software levels, including prerequisites, to provide a fix.
4. Provide fixes for mission-critical software defects.

R(2): Describe your approach to providing support, including:

- how incidents and issues are logged,
- requirements for information that needs to be provided to initiate a call, and how log and diagnostic information is gathered,
- which days and hours this standard support is available,
- types of calls dealt with by themselves, and which will be escalated,
- details of how an escalation will work,
- size and experience of their support team

R(3): Describe the levels of support offered, and the costs associated with these levels of support.

M(4): The Tenderer is required to provide 365x24x7 support for critical issues of their tendered solution.

Critical issues are limited to situations where, in the judgement of ECMWF, the DSMS is unavailable, damaged or degraded such that it is unable to support our operational services. The priority in addressing severe problems with the DSMS is to take corrective action so that the system becomes operational. Temporary workarounds are an acceptable method of resolving severe problems but work to finally resolve the issue must be started as a priority during the next standard support hours.

M(5): As part of the critical issue support (whether software or if tendered, hardware) the Tenderer should:

1. Provide remote telephone support twenty-four hours a day, seven days a week.
2. Respond to a service call and commence resolution of the problem within two hours. A response is defined to be a call back to the authorized caller who originally reported the problem.

R(6): If hardware is provided as part of the Tendered solution, detail the support options available for non-critical support and their costs.

R(7): It is vital to ECMWF that an escalation route exists to discuss problems with the support team. Tenderers should describe how their proposal enables this requirement and supports problem solving in critical contexts.

R(8): Describe procedures for handing over calls and issues to ensure continuity at the end of support analysts working period. Describe particularly any procedures for ensuring effective handover from staff involved in out-of-hours critical issue response.

Consultancy Support

ECMWF expects to require assistance from a small group of experts to plan and implement major system upgrades or configuration changes. It is also anticipated that help will be needed during the early phase of any migration to a new software solution, which will require considerable time to establish in operations.

R(9): Describe what types of assistance and support will be made available to ECMWF during the installation/acceptance stages and during the early operational stages. State what is available and if any additional cost would be incurred with any element of this support.

M(10): Provide priced options for both on-site and remote consultancy, to support and assist during the migration of ECMWF systems to a new software solution.

These options should describe the experience and abilities of the consultants and include rates for a standard eight-hour day. On-site consultancy will be at one of the ECMWF sites; currently in Bologna, Reading or Bonn. Remote support would be required from 09:00 to 18:00 Central European Time.

M(11): A need for similar consultancy is expected to be needed throughout the contract period. Provide a pricing mechanism for arranging future, or ongoing consultancy support. Include any variations for different levels of support within the mechanism.

ECMWF will require access to core members of the development team. Access to the source code of the solution may be useful but is not essential.

R(12): Explain what access there would be to the product development team and whether access to the current running source code would be available.

R(13): There may be cases where a bespoke feature may be requested by ECMWF. Explain how you would address this.

R(14): Explain what options remain available for ECMWF to read the data from the DSMS if the license and/or support contract expires or is terminated. ECMWF would not expect to be able to write to the system if this were to occur.

Training

M(15): Tenderers must include in the quotation the cost of an initial training programme, preferably given at ECMWF's premises, which shall provide at least:

a) Training for 6 analysts, to instruct and prepare them with sufficient understanding of the internal working of the software and possibly hardware being tendered to enable them to provide effective day-to-day support and emergency support;

b) Training for 2 separate groups of 6 computer operators, to instruct and prepare them with sufficient understanding of the software systems being tendered to enable them to effectively monitor the normal operation of the systems, to identify and report abnormalities or inefficiencies in the running of the systems and to initiate corrective action in cases not normally requiring the support of an analyst;

c) Training for 6 developers, to instruct and prepare them with sufficient understanding of the software being tendered to enable them to develop applications software that will run efficiently on the tendered system.

R(16): Describe the content of these training programmes.

R(17): Provide details of other relevant training courses and their cost.

Documentation

M(18): A complete set of the DSMS documentation must be made available.

M(19): Tenderers must undertake to provide updates to the documentation of the tendered systems for a period of not less than 10 years from the date of acceptance of the system.

Technical Requirements

Data management and migration

To achieve cost-effective data storage, the tendered system might include multiple classes of storage device, which typically will range from faster SSD / nearline disk storage to slower tape / optical storage. The software system managing the data and the devices should ensure, by means of configurable procedures, that inactive data is moved onto low-cost media, while active data is brought onto rapid-access devices (referred to as cache storage).

M(20): Following the initial writing of a data object, the tendered DSMS must, using parameters such as time, size, likelihood of retrieval, etc to be able to resolve where to place each object in a hierarchy of data storage resources. From this, it must migrate the objects accordingly.

H(21): Describe how the tendered DSMS supports the following concepts:

a) The grouping of storage devices allocated to the DSMS into multiple data pools, each separately managed, with its usage characteristics. These structures may be hierarchically related.

b) The allocation of data between these pools, and the movement between these pools.

c) The control of policies (both automated and manual) driving the allocation of data amongst these pools, based on past patterns of usage and future predictions.

R(22): Describe how the tendered system manages storage devices and/or data pools, stating in particular how many classes of storage device and data pools can be managed, and the degree of flexibility in the rules that may be used to migrate data between them.

R(23): Describe the migration processes and what flexibility any rulesets would have in defining the migration processes. Further, describe the configurable migration parameters and the processes involved in changing them; e.g., whether the DSMS must be halted/restarted/rebuilt to incorporate any new migration rules.

H(24): It is highly desirable, under DSMS control, to allow multiple streams of data concurrently in migration, to and from multiple devices. Tenderers shall describe any limitations of their systems in this respect.

H(25): It is highly desirable that the tendered DSMS take automatic action to migrate instances of data objects between the available pools, to optimize the use of system resources. Data objects migration should take place as a background activity which should not interfere with delivery of service to the MARS and ECFS applications.

R(26): Describe any circumstances in which such migration would impact production activities (MARS and/or ECFS jobs), *excluding simple media drive resource contention.* Further, describe the mechanisms available to mitigate the above-mentioned impact of resource contention between migration and production activities.

H(27): It is highly desirable that application requests and internal housekeeping management do not abort as a result of device contentions but suspended instead. Client requests should normally take priority and housekeeping continues once client read or write activity completes. It should be

possible to suspend or halt the process of migration, and later to resume it without undue overheads.

H(28): It is highly desirable to group removable media in families of media; each family being used to keep a specific subset of the stored data. The boundaries between subsets should be site-defined. Judicious usage of such a feature would minimise the number of mounts required to respond to requests where multiple data objects from a given subset of data need to be read to satisfy application requests. The relation between data pools and media families does not need to be one-to-one, as multiple media families could be stored in the same data pool, provided that segregation between data belonging to separate families is maintained.

Explain how your proposed solution allows the management of families of media, and what impact such management might have, e.g. by introducing a limitation on the number of parallel migration streams that can be performed from/to pools where media are grouped in families.

R(29): This a series of questions over the performance of the DSMS and any upper limits the system might have. The performance questions here assume that the metadata of the DSMS is based on dedicated NVMe appliances, with no external usage beyond the system. The expectation is to have at least two independent copies of the DSMS metadata available, but how this is achieved is implementation dependent:

- How many data object creations can be achieved per second,
- How many data object deletions can be achieved per second,
- The total number of data objects per instance.

R(30): Describe what parameters of a data object may be automatically taken into account in data cache management decisions, for example data pool residency, migration and eviction. This should include parameters such as:

- Timestamps taken at create/write/read/cache/migrate times
- The total number of times the data objects have been accessed
- The recent frequency of access
- The size of the data object
- The available space in the data pool
- The identity of the object's owner
- The client system on which the data originated;
- The identification of the object (e.g. fully qualified path name)
- Any user-generated advisory information
- All current residences of the data

H(31): It should be possible to override the normal decisions taken on migration of data object within the storage hierarchy managed by the DSMS, so that single data objects, directory trees or families may be given different hierarchy residence from that which would normally be given to them by the migration algorithms.

R(32): Describe what capability is present for data to be read directly from underlying storage tiers, including removable media, bypassing layers of caching within the system.

D(33): It is desirable for the DSMS to monitor migration activity in order to detect potential bottlenecks or situations where the data lifecycle workflow is affected by unexpected issues, such as unusually large migration backlogs or data objects that cannot be migrated for any reason.

Ability to reconstruct primary data using redundant data elements

The tendered system should be able to upon request make multiple copies or encode any data object so that the primary copy can be reconstructed from remaining elements stored in the DSMS.

It must be possible to stipulate a configurable period following the creation or last modification of a set of data objects so that the DSMS makes multiple independent copies of the data blocks of the object(s). The locations and media on which such copies are stored must also be configurable, considering such parameters as data object size, age, etc. The secondary version could be located at a separate data hall or data centre.

It also needs to manage removable media volumes that have been ejected from the robotic *library* and are stored in off-line racks, *and/or off-site*

H(34): Where such additional copies are required, the process creating them should always complete without the need for a utility to be run to ensure such completion.

H(35): When any media or region of media, either primary or secondary, is damaged or incomplete, for example as a result of media loss or incomplete creation, then this media must be marked accordingly. It must then be prevented from being used, unless its use has been specifically allowed by a system administrator.

D(36): When multiple copies of a data object reside on multiple media or devices, the retrieval actions for this data should take into account this multiple residence (e.g. if one copy is stored on a volume which is currently in use, and a second copy of the data exists on another volume, it may be optimal to use the second volume to read the data object).

M(37): At ECMWF our storage is distributed across multiple libraries and multiple data halls. It must be possible to set policies such that redundant copies of data are stored in physically distinct locations to safeguard against loss in catastrophic scenarios.

R(38): Tenderers shall describe how the DSMS provides policies to place different copies of data in different locations (media, libraries, data halls, data centres). Assume networking is provided in the latter cases to provide the necessary connectivity (both IP and potentially Fibre Channel.

D(39): The ability to designate some removable media, containing secondary copy elements of data, as being destined for storage in a vault or off-site facility is desirable.

H(40): The number of redundant copies of a data object which are known to the system may become less than the number required. This may happen because of a volume, device or subsystem being made unavailable (either by manual intervention or by the system's own determination of its status), or as a result of the required number being increased by user or system administrator action. In such a situation, the system should automatically detect the issue and notify system operators who can allow the creation of such additional copies.

H(41): If a volume is damaged and an operation is started to restore this volume then, prior to commencement, the tendered system should provide a list of secondary copy volumes that are required.

H(42): As long as there is at least one viable copy of data in existence, then in any situation where data retrieval is initiated, this should proceed without incident (other than impacting response time) as far as the client is concerned. It is preferable that any automatic recovery is configurable so that operator approval can be sought, depending on the likely cost of retrieving this viable copy.

D(43): Either MARS or ECFS, while in the process of creating a data object for which multiple copies are required, should have the option of waiting for all requested instances of the data object to be created before returning to the application, and returning immediately after writing the first copy.

H(44): It should be possible to query the number of copies of any data object, and their location (media volume(s) and position(s) on these) both programmatically and through administrative tooling. Tenderers shall describe the process of recover/restore from the secondary copy of:

- A data object
- A set of data objects possibly aggregated in sections/segments of a removable media
- A complete removable media

D(45): Describe the configurable automated operations that the DSMS can initiate in the event of:

- unavailability of the primary copy
- unavailability of the secondary copy

Data repacking and copying

It is expected that from time to time it will be necessary to transfer data from one media volume to another, e.g. to retire a volume which has reached it's end of life. It is also expected that as new media generations are released the entire archive will be copied from one set of media to this next media generation. It is possible that the new media may be of a different technology.

H(46): It is highly desirable that the DSMS is able to recover unused space on data volumes which has resulted from the deletion, expiry or other obsolescence of data objects on those volumes ("repacking").

R(47): Describe the mechanisms that can be employed to recover unused space, in particular trigger mechanisms, and whether they involve human interactions.

H(48): Repacking operations should be done as background activities. In any case, it should be possible to suspend or halt the process of repacking, and later to resume it, without undue overheads.

H(49): In case that some or all of the content on a removable media are unavailable, it is highly desirable that the system allows operators to initiate repacking data to different media. Tenderers shall describe the granularity of the repacking operation, such as if it is possible to move single data objects, aggregated objects or the remaining available content of the removable media. Where necessary skipping damaged segments of the media.

R(50): Provide an estimation of the time required to provide a list of all data objects stored on a given volume, assuming that 1 million data objects have been stored in the volume.

Equipment Supporting the Data Storage Management System

This is generally trying to obtain a baseline of what hardware flexibility there is in running the tendered solution. Using the Existing Hardware Section as a reference of what hardware ECMWF is currently running its HPSS system upon. Explain whether there are any restrictions in what hardware systems could be deployed or what recommendations could be made to guide future hardware purchases of storage and server systems.

Where capacity, transfer rates or other performance figures are quoted, GiB/TiB/PiB indicates $2^{30}/2^{40}/2^{50}$ bytes and GB/TB/PB indicates $10^{9}/10^{12}/10^{15}$ bytes, etc. Please refer to Table 5: Byte Scale for Data capacity and transmission rates for details on the numbering scales used. Tenderers must not take into account any potential gain provided by any hardware or software data compression techniques.

R(51): Tenderers shall detail any restrictions they may have over what types of systems are not compatible with their solution, whether that be storage systems or server types. As this list could be very long it may be better to express this as what is supported rather than what is not.

R(52): Tenderers shall detail any storage system, SAN interface or server type recommendations they may have in running their solution.

Removable media volumes

ECMWF expects that the bulk of the data managed by both MARS and ECFS will be stored on removable media. Far from being inactive once written to media, this data is likely to be read multiple times. Likewise, it is likely that primary media will be emptied and reused several times during their lifetime at ECMWF.

Due to the cost involved in keeping multiple copies of the data stored in a DSMS, ECMWF has a policy of not taking secondary copies of large parts of its non-critical data. This does not, however, mean that the loss of data, for example, as a result of media loss, is considered lightly, and such events, while unavoidable, must be kept to a minimum.

Whilst ECMWF expects that magnetic tape is the technology most likely to be tendered for removable media, it does not rule out alternative media solutions that may become available during the lifetime of the tendered system. Within this ITT wherever reference is made to tape drives and tape media, tenderers may provide descriptions and responses that might embrace other types of removable media.

ECMWF wishes to minimise the delays and manpower requirements necessary to manage the very large number of media required to store this data. To achieve this the complete copy of the primary data will be kept in a robotic library.

R(53): Provide a list of the robotic libraries with a capacity of at least 1000 slots which are compatible with the tendered DSMS.

R(54): Describe any features of the DSMS, which allows the time taken to load a volume to be kept to a minimum, e.g. by ensuring that the media is mounted on the closest available drives.

R(55): Explain how multiple generations of a media (e.g. LTO-9 and LTO-10) resident in a single library are managed intelligently. How drives are selected, taking into account the read-write and read-only combination rules of the media and drive technologies.

H(56): It is highly desirable that removable media can be inserted into and removed from the robotic libraries without interrupting the service it provides. Describe the procedures to follow in these

situations, with reference to any procedures required to inform the DSMS of this change in status of the media in the robotic library. Include any means of specifying Volume ID's directly or indirectly by requesting the DSMS to select a specified number of volumes base on criteria, for example, last recently used.

R(57): Describe the impact on normal operation when maintenance of any kind must be carried out on drive or a whole robotic library.

R(58): Provide the following limits to the DSMS.

- State the maximum number of removable media drives that may be connected to the DSMS
- State whether there is a maximum number of removable media libraries that can be defined to a single instance of the DSMS? If there is please provide the limit and reasoning.
- State whether there is a maximum number of I/O mover servers/services that can be defined to a single instance of the DSMS? If there is please provide the limit, reasoning and whether this limit applies to fixed or removeable media movers or both.
- State whether there is a maximum number of removeable media volumes that can be held within a single instance of the DSMS? If there is please provide this limit and the reasoning behind it.

H(59): If it is required to specify the capacity of media volumes either in a configuration file or via administrator commands, it is highly desirable that this information should be used only as a guide by the system, and not as an upper limit to the amount of data that can actually be written on the volume. Tenderers shall explain how their system behaves in this respect, in an environment where data compression can be employed when data is written to a volume.

R(60): For data written to removable media, explain what data formats and data protection mechanisms are available with the tendered solution. Include any methods to validate and recover data where possible.

Media Equipment Management

R(61): Provide details of how their system manages media volumes which are not resident in the robotic library (offline volumes). They shall describe whether the system provides some means whereby location information can be assigned to offline volumes, the level of granularity of such information (e.g. room, shelf, box, cell) and whether this information can be entered or updated by administrative staff.

R(62): Describe the facilities that are available to manage demands for media volumes that are not currently held in the robotic library and which therefore must be mounted manually or imported into the robotic subsystem. They shall describe:

- if it is possible for an operator to be notified automatically to import a volume stored outside a robotic library;
- if a site-written procedure can be executed automatically in order to prompt operator intervention.

R(63): Describe what parameters of a media volume and its data objects activity are available to algorithms moving volumes into and out of attached robotic libraries, and what process normally makes the decision.

D(64): It should be possible to manually introduce batches of volumes in response to, or in anticipation of a system request to do so.

H(65): All of the information regarding media volume activity and performance collected by the DSMS should be available via some standard interface, for example, a REST API to ECMWF-developed or third-party packages, in order to generate reports, summaries, cross-references and projections.

H(66): In the context of an apparent failure of a volume, tools should be provided, preferably independently of the DSMS software, for verification of the integrity and readability of any media volume. Ideally, such tools should not require the total volume content to be restaged to cache.

R(67): Describe how the DSMS solution accommodates the automatic cleaning and media verification features of common robotic library systems.

M(68): Assuming that recommended procedures have been followed to create duplicate copies of data objects, it must be possible, either via the DSMS software or otherwise, to reconstruct the contents of a media volume which has become damaged.

R(69): Describe the options and processes are necessary in the event that recovery of a damaged media volume is necessary.

H(70): It should be possible to audit the contents of the media volume and produce a list of any data objects which are irretrievable.

R(71): Tenderers shall describe the steps taken by the DSMS when it needs to allocate devices to mount removable media. Mention should be made of any features which would allow the priorities of different type of requests (such as 'read' versus 'write' or requests from the storage applications versus requests generated internally by the DSMS) to be specified or modified 'on the fly'.

D(72): It should be possible to ensure that a given number of media drives are reserved exclusively for read or for write operations.

R(73): Describe the general procedure for initialising volumes for entry into the DSMS, including any provisions which would allow the recording medium to be certified as part of the initialisation process.

R(74): Describe the procedure necessary for removing or unallocating a logical or physical device from the DSMS configuration and describe the impact on the service. The situations in which the device is to be made read-only or totally unavailable to the DSMS shall be covered for the following devices:

- Robotic library
- Disk volume
- Removable media volume
- Media drive

D(75): It should be possible to reconfigure, rename, and change attributes of an attached device, such as a media drive. For example, when replacing a drive in a library position with a drive of the next generation the DSMS can be reconfigured to accept the drive, recognising its new media type.

D(76): It may be necessary for ECMWF to rebalance the media in the robotic libraries used in the DSMS. In such a situation it may become necessary to transfer some of the media volumes residing in one *library* to another for reasons of capacity, or to balance the load. Provide a summary of the

procedures to follow in order to perform such an operation, with particular reference to any procedures required to inform the data management software of the resulting state.

D(77): State whether the RAO (Recommended Access Order) feature of the IBM TS11xx drives is supported. If it is, describe how the feature can be used by clients. In particular reference to specialised API calls or techniques of how to read data objects or regions of a data object.

Host operating system and Installing / Upgrading DSMS.

M(78): The host operating system of the platforms must be a distribution of Linux or another implementation of Unix.

R(79): Detail the supportable operating systems that their solution could run on, both for the Core server and mover components as well as possible platforms for the client systems using the client API or interface.

R(80): Describe the general procedure for, and usual frequency of, installation of new versions of the host operating system supporting the Core server. In particular, the need to minimise exposure of the production service to down-time as a result of such installation should be borne in mind. The situation in which a new version is installed and must then be backed out with minimum disruption shall be addressed.

R(81): Software installation tools must be provided to allow the easy installation of patches and major upgrades. They should also provide good visibility of installed software, down to the level of individual components and allow the easy removal of unwanted software and the reversal of unwanted updates. Tenderers shall describe the process used for major upgrades and patches.

R(82): Tenderers should describe any facilities allowing network installation and managing the software remotely.

R(83): During the life of the new DSMS system it will be necessary to move it from one server platform to another, for example, to provide increased CPU performance and memory. Tenderers shall describe the expected impact on the service and the process to achieve this.

R(84): Likewise, it will be necessary to upgrade the DSMS to newer versions. Tenderers shall describe the expected impact on the service and the process to achieve this. In particular, any need to convert or modify the metadata used by the DSMS shall be described.

H(85): Since it is not operationally acceptable to halt operation of the system while such backup copies are made, it is highly desirable that mechanisms should be provided so that internally consistent backups of the changing system data can be constructed 'on the fly' without interfering with normal system operation. As an alternative, it may be acceptable that data storage in the DSMS is prevented or constrained during the taking of backups, provided that data retrieval can continue as normal.

Networking

H(86): If the tendered solution includes server hardware, they should be connected to the internal ECMWF Network (currently based on bonded 25Gb/s connections) and the Centre's Management network (1Gb/s copper). However, any servers should be adaptable to future network requirements.

Resilience, availability and recoverability

Because ECMWF runs its computer facilities as a 24-hour, non-stop operation, the DSMS must be able to fulfil its role in such a critical environment. Further, since ECMWF runs a managed operational environment with full cover, the DSMS must interact effectively and correctly with system operators and administrators.

Most available hardware and software components cannot in themselves offer the continuous fault-free operation that is desired. It is therefore important that the design of the DSMS software should take into account the inherent unreliability of the facilities it is built upon, and that it should seek to minimise the duration and impact of any failures in servers, storage devices or media.

If storage media becomes lost or damaged, it is essential that operations staff should be able to take prompt and appropriate action to minimise any consequent loss of data. Utilities and procedures should exist to facilitate this.

Despite the use of high-quality components and sub-systems, and despite the observance of proper design considerations in its construction, situations will inevitably arise in which the operation or administration of the DSMS is disrupted. The degree to which proper provision is made for restoring normal operation in these situations is one of the most important characteristics of the system.

H(87): The overall function of the DSMS should continue uninterrupted in all normally foreseeable operational situations short of fundamental equipment or software malfunction. Human intervention to maintain primary system function should preferably be limited to the authorisation or denial of system-initiated recovery.

H(88): The failure of any single drive, volume or other storage device should only affect data objects on that drive (etc.), and any requests directly related to it. An alarm should be raised, but other operations should continue as normal.

H(89): It is highly desirable that the design of internal administrative systems and update mechanisms within the DSMS should be such that it is impossible for an unscheduled or unclean shutdown event to cause the system, its metadata or any of its subsystems to be left in an unresolved or inconsistent state. Any tools needed to recover from such a state should be supplied with the DSMS. It is acceptable that certain situations left by shutdowns of such a nature may require explicit recovery actions to be taken before the system is again available for normal operation, provided that:

- The recovery procedure does not normally require expert intervention;
- The recovery procedure does not normally occupy more than two (2) hours elapsed time;
- The status of any transactions, in progress at the time of the event, is resolved unambiguously;
- Data objects and metadata that were not involved in any on-going transaction at the time of the event are invariably preserved intact.

R(90): Describe the steps required to close down the DSMS in an orderly and systematic fashion, and provide estimates of the time required to go from active to closed-down in the two cases of systems containing one and five billion data objects.

R(91): Describe the steps required to start up the DSMS, from the initial situations of:

- A DSMS closed down in accordance with the supplier's recommendations, but the underlying computer system remaining active;
- Start from and unscheduled or unclean shutdown event, especially after significant power or network failure.

R(92): Provide estimates of the time required to become fully active on start up in the two cases of systems containing one and five billion data objects.

M(93): The tendered solution must be designed in such a way that, when correctly operated, no foreseeable event can result in metadata corruption or rendering the system unusable.

The degree of exposure of the DSMS to loss of data, following from the restoration of metadata from backups and journals, is a matter of great concern.

R(94): Tenderers shall describe their suggested operational schedule, including any necessary manual intervention, for backing up and journaling an DSMS and its support structures. The tenderer shall provide estimates of the time required to perform this in the two cases of a DSMS which contains one and five billion data objects.

H(95): Since it is not operationally acceptable to halt operation of the system while such backup copies are made, it is highly desirable that mechanisms should be provided so that internally consistent backups of the changing system data can be constructed 'on the fly' without interfering with normal system operation. As an alternative, it may be acceptable that data storage in the DSMS is prevented or constrained during the taking of backups, however ensuring that data retrieval can continue as normal.

R(96): Describe how the metadata of the system is protected to avoid leaving the tendered solution into a unusable state.

H(97): At some level such provision will consist of the periodic taking of backup and journal copies of the DSMS metadata. Restoring the system from metadata backup and journal logs should take into account the situation where the copy no longer reflects the current state of the system. Some means should be provided to detect and determine the extent of loss of system function or data resulting from such a mismatch.

D(98): Following an incident in which the metadata of the DSMS are only partially destroyed, it is advantageous if the mechanisms for restoring the system can verify which elements of the metadata are intact and which are destroyed and recover only the intact elements.

R(99): Describe any provisions which would allow the isolation of corrupted portions of the system for repair work, while normal operation continues on the remainder of the system.

R(100): Tenderers shall describe the ability of their systems to recover from, and shall outline the recovery procedure for each of the following exceptional events:

- Host operating system kernel panic;

- Unscheduled loss of power;
- Total loss of one of the volumes holding a metadata database;
- Total loss of one of the volumes holding cached data;

R(101): As fault conditions and maintenance activity are inevitable on the hardware underpinning the DSMS, explain what measures are in place to aid with the failover of the system to other hardware components or how operations are continued without the failing component. With these failure scenarios, explain any consequential effect on other production activities in case of failure of:

- a server crashing which is running all or part of the Core metadata services to the DSMS,
- a server crashing which is acting as a disk, tape or SSD I/O data mover,
- a block storage LUN becomes unavailable,
- all contact to a robotic library is lost,
- a media drive becomes unavailable,
- network connectivity between the elements of the DSMS and client sessions are lost.

R(102): It should be possible to restore the components of the DSMS from backup onto similar hardware and operating system in a straightforward manner to bring it back into operational use.

R(103): The impact on service to clients shall be described and the tenderer shall provide estimates of the time taken for each recovery procedure in the two cases of a system which contains one and five billion data objects.

R(104): If any part of the operation of any component of the DSMS is dependent on the periodic running of system utilities, then describe:

- What functions the relevant utilities perform, and what resources they consume;
- The impact on normal operations such as read/write access and performance;
- How such periodic running is achieved;
- What (if any) human involvement is required;
- How any problems in the running of these utilities are communicated to operators;
- What the consequences may be if one or more of these utilities is not, or fails to run for a protracted period.

H(105): It is highly desirable that the Tendered solution has as few single points of failure as possible, this preferably being zero. Describe how the DSMS continues to provide the full complement of its services, possibly at reduced performance, after any hardware or software failure has been experienced.

R(106): Describe any features allowing the system to remain operational after failures, as well as the interaction of such features with the operation of the DSMS in the event of an unscheduled stop, hardware failure or panic.

It will be necessary during the lifetime of the system to ensure that the DSMS metadata is consistent, and reflects the data saved in the multiple data pools that the DSMS manages. Housekeeping jobs may have to be run to reconcile information relating to different aspects of this metadata. There may be occasions where hardware or software errors will result in inconsistencies that need to be rectified.

It should be possible to perform such consistency checks on interrelated sub-sets of all data pools managed by the DSMS. Examples of interrelated sub-sets of data pools include:

a) In the case of hierarchical storage managed filesystems, all data pools that include copies of data objects stored in a filesystem;
b) In the case of a relational database, all data pools which include data related to a specific table.
c) In the case of a DSMS composed of multiple components, each with its own metadata, the metadata for each individual component.

H(107): The DSMS should provide utilities that can be used to verify the internal consistency of the system metadata on the entire DSMS. These utilities should be able to identify potential inconsistencies, and possibly automatically repairs any such inconsistencies.

H(108): It should not be necessary to halt the normal function of the DSMS in order to perform consistency checks, nor should it create a significant impact on the performance of the DSMS when it runs. The routine operation of the DSMS should not depend on such a global check being run periodically.

H(109): Metadata corruption introduced by software errors may exist unnoticed for some time, leading to increasingly pervasive effects as multiple, or possibly all backups of the metadata information become corrupted. In this context, utilities should exist that could be run at low priority, to scan through the metadata structures of the DSMS while the DSMS is fully functional.

H(110): The tenderer shall describe the scope of such metadata and system verification utilities. Provide estimates of the time for the tendered systems to perform consistency checks as described in requirement for the two cases of a DSMS containing one and five billion data objects.

Housekeeping Tasks

At the larger scales envisaged for the tendered system, it is likely that many housekeeping and administrative functions (e.g. fsck-like activities; directory and system table backups, accounting information extractions, etc.) will run for a considerable elapsed time. Such operations then run the risk of interfering with, or of being interfered with by, other scheduled activities of the system. The risk of a malfunction preventing the completion of the administrative action also increases the longer the elapsed time.

R(111): Describe the approach to minimising, mitigating and controlling the impact of scheduled housekeeping functions on the behaviour and performance of the system.

H(112): It is highly desirable that housekeeping and administrative functions do not result in unavailability of the system.

D(113): The administrative processes in the system should be designed so that any wasted system resources, (e.g. caused by the need to abandon partially-completed work at a recovery point), are minimised so far as it is consistent with their efficient operation in other ways. In furtherance both of this objective and that of providing maximum resilience in the overall system, it is desirable the DSMS metadata can be divided into logically self-contained domains, along directory-tree or other lines, which although sharing a common set of physical resources are otherwise independent. This would permit the backup, restoration, verification, reporting and other processes in the system to be more limited in scope and therefore more rapidly (and reliably) completed.

R(114): If it is possible to divide the DSMS metadata up into domains as described in **D**(**113**), explain the general process and overhead cost of moving data objects from one domain to another. Provide time estimates for transferring one million data objects, each of one megabyte in size, from one domain to another.

Recoverability of Media Volumes

The vast majority of ECMWF's archived data, and all of the standard meteorological output, are in self-describing, streamable data formats. This supports recovery processes and investigations of what are stored on different media.

D(115): It should be possible to read the data stored on removable media independently from the DSMS. In this context, the inclusion of appropriate metadata on the removable media at creation time is a further significant advantage.

R(116): Describe the amount and contents of metadata information recorded on removable media.

Testing and development facilities

It must be possible to install, operate and support new software effectively in ECMWF's production environment. This implies that procedures for developing, installing, testing and upgrading software must be devised with the intention of minimising the risk to production services.

It is expected that ECMWF will run at least three instances of the DSMS. A development, a pre-production and production instance of the software. It is likely that at times a testing DR instance of production environment will also be needed.

M(117): When updates are to be made to the DSMS, it must be possible to run tests of new software versions, support packages, and supporting operating system (new releases, bug fixes, changes of configuration, etc.) in parallel on a test instance of the software on independent server hardware, without interfering with the operational DSMS service.

H(118): ECMWF would prefer this to be provided by separate test and pre-production systems, compatible with the production system and sized adequately to enable realistic testing to be performed. Tenderers are requested to recommend their preferred environment and methods of testing updates on to the system and explain their rationale and support coverage.

Operator and system administrator facilities

An intuitive and consistent interface should be provided for the management and operation of the DSMS.

D(119): The provision of a web-based administrative interface is desirable.

M(120): Regardless of the provision of a GUI interface, it is essential to allow operation and administration via a Command Line Interface (CLI) or through scripts, such that tasks can be integrated into cron jobs.

R(121): Provide a list of functionalities that are not provided by a CLI, but only through a GUI or web-based interface.

An API should be provided to enable system monitoring and management to be performed.

H(122): When manual intervention is required, notification of the requirement should be presented clearly and unequivocally to operations staff.

H(123): If requested human intervention is not forthcoming in any situation, those parts of the DSMS which are not affected should continue normally. Such situations include: requesting authorisation to proceed with some action; request the importing of a shelved media volume; requesting a utility to be run; requesting minor maintenance activity, etc.

R(124): Tenderers shall describe the general provisions made for management of queued requests, including any limitation on queue sizes, the priority scheduling scheme used and abilities to manually intervene in the scheduling or processing of a request.

H(125): Tools must be provided to facilitate management of users and all aspects of security, resource limits, and access control associated with users. It should be possible to change these dynamically without needing to interrupt the normal operation of the system.

H(126): Describe from the quality of service point of view how system resources of the DSMS can be controlled and allocated to individual client users of the system.

Observability

Operators and administrators require an extensive set of tools allowing them to evaluate the health of the system, ascertain if any part of the system is down, malfunctioning or overloaded, react to abnormal situations, and schedule preventive actions on equipment or media starting to deteriorate. Logs must be provided to allow the analysis of problems and the determination of most likely causes. Administrators must be provided with sufficient information that they can analyse the system usage, in order to evaluate usage patterns and trends, and evaluate the quality of the service, through the use of; disk systems, robotic libraries, media volumes, storage pools, client activity, the DSMS itself etc.

Monitoring

Monitoring tools should be provided to give full visibility of DSMS activity; all requests being processed and their related jobs, any DSMS housekeeping, the status of all data pools, the software and hardware systems. Usage and performance metrics shall be reported to the Analysts. Such system components shall include, media drive status, main DSMS system daemon status, media storage pools status, etc. Usage and performance metrics shall include, system resource utilization (disk system, robotic library, network and media volume usage, etc), job throughput, queue lengths and transfer bandwidth, etc. Metrics should be structured and made accessible in real-time via standard protocols for integration with external observability systems.

H(127): Describe which DSMS usage and performance metrics are available and how these are exposed to the operational staff.

H(128): It should be possible for an alarm or error indication to be raised whenever critical system parameters fall outside an acceptable defined range. When error conditions arise with the system, or whenever manual intervention is required, notification must be presented clearly and unequivocally to staff. Describe how error conditions and system alarm workflows are managed by the DSMS, including how the Operators & Analysts interacting with the system are notified about any problematic condition.

H(129): It is highly desirable that monitoring system can be customized by the activation of scripts written by ECMWF to raise alarms, to retrieve usage information not directly exposed by the system or to

activate contingency procedures meant to deal with any problematic situation. Describe how this is implemented in the tendered DSMS.

R(130): List any third parties available monitoring packages, through which monitoring of the tendered solution could be implemented. The use of open observability platforms such as Grafana, OpenTelemetry, ELK, etc for visualization and correlation of performance and error data should also be described where applicable.

Logging and log analysis

H(131): It is highly desirable that all transactions, errors and exception conditions, auditable events, configuration changes and other relevant information are logged, with an appropriate timestamp and message identifier, to files or databases.

R(132): Describe the logging mechanisms and the utilities and methods that can be used to process this information. Log access via open interfaces (e.g. syslog, REST APIs) and integration with centralized log management systems (e.g., Fluentd, Logstash, or Splunk) should be described.

R(133): Describe any available tools for error log analysis or for correlation of error log entries to hardware diagnostics.

M(134): Logs must be structured, persistent, and searchable.

H(135): It should be possible to classify and filter logs by system component, severity, and time range, and to correlate logs with relevant performance metrics and traces for diagnostics and trend analysis.

Tracing/Debugging

H(136): The system should support tracing capabilities to provide visibility into the sequence and timing of operations across different components of the DSMS stack. Tracing should allow correlation of system events and user jobs across subsystems such as clients, data movers, metadata managers, and archive services, to support bottleneck identification and performance tuning. Traces should be exported in a structured format and be linkable to logs and metrics. Describe any such tracing features provided by the tendered system, together with the interfaces and protocols used to expose trace collection.

Diagnostics

Persistent error states can easily be missed if they are only reported once to a transient log.

H(137): It is highly desirable that tools be provided for dynamic and/or periodic error log analysis, which would report persistent errors and could be integrated into a hardware diagnostic facility. Hardware diagnostic tools should integrate with error log analysis and should provide facilities such as device testing and reconfiguration, control-ware/microcode management, and associated device service aids. Tenderers shall describe the diagnostics mechanisms implemented if available.

Impact of partitioning

As the scale of capacity and workload increases partitioning of resources may provide noticeable benefits, such as;

- reducing the impact of consistency validation checks,
- dedicating storage space to specific subsets of important data,
- increasing overall system availability.

However, these benefits could introduce problems related to load balancing between the multiple partitions, resulting in potentially serious underuse of the total resources allocated. This is a particular concern in ECMWF's environment, where load patterns vary during the day, and changes from month to month, both in volume and frequency of access, often in an unpredictable way.

The tendered systems might make use of resources as a whole or they might partition them into multiple entities. For example:

a) One or more servers, or a cluster of nodes may be deployed to support only one of the data applications (MARS or ECFS) or both;
b) Robotic libraries partitioned to guarantee separation of workloads or scope compatible drives with media volumes.

ECMWF considers these issues to be very important.

R(138): For each type of DSMS resource (robotic libraries, disk systems, removable media drives, servers, network connections), tenderers shall provide the following information:

a) Whether the solution tendered is designed in such a way that the resources can be partitioned. If so, the nature and extent of this partitioning shall be described;
b) The rationale behind this partitioning and any benefits that it might confer;
c) The boundaries at which partitioning can be established, e.g. at the subdirectory level within a filesystem, or at the level of media drives connected to a given robotic library etc.;
d) The potential impact of such partitioning on resource usage, how this could be minimised, and the provisions that have been made in the tender to overcome this;
e) How short-term re-balancing of the load between the multiple partitions could be performed (e.g. reallocation of removable media drives to another partition). Whether this could be done dynamically, semi-statically using scripts or commands without interruption to the service, or whether it requires an interruption to the service;
f) How long-term re-balancing of the load could be performed (e.g. redistribution of the data objects in one pool to another, transfer of media from one robotic library to another);
g) How new resources could be added to a partition;
h) How a new partition could be added efficiently to the system;
i) How a partition could be removed from the system.

R(139): Explain the thresholds that would guide how the tenderer would decide to suggest partitioning of the solution and the rationale behind these decisions. If any level of hardware partitioning it suggested, then either reference the suggested model hardware as detailed in Annex 3 Requirements of Acceptance or the Tenderers required or recommended hardware.

Information security

Ensuring adequate level of security for ECMWF's DSMS is essential as many critical services rely on uninterrupted and trustworthy access to these resources. Any compromise in security could disrupt these vital operations, potentially impacting downstream services relying on accurate and timely meteorological information, as well as jeopardizing the integrity and confidentiality of data.

R(140): Tenderers shall describe the approach to align to industry best practices and standards to ensure compliance with security principles such as network segmentation, least privilege, RBAC, multi-factor authentication, user management, data integrity protection mechanisms, real-time cybersecurity event monitoring, etc.

M(141): The system must provide an access control mechanism that restricts access to data objects.

R(142): Tenderers should describe the data object access control mechanisms supported by the proposed solution.

R(143): Indicate whether the System can perform LDAP requests for authentication and access control.

D(144): A facility is desirable whereby tasks requiring additional privileges (such as operator, super-user and administrator functions) can be separated so that one individual does not necessarily get all privileges at the same time.

R(145): Tenderers should describe the administrative security model, including the different classes of administrative roles and considering the scope and traceability of actions involving elevated privileges.

H(146): Keys, certificates, and credentials should be securely stored and rotated using a secrets management solution.

D(147): The System should adopt a Zero Trust model (all requests shall be authenticated, authorized, encrypted, logged).

D(148): Tenderers shall provide information about EDR (Endpoint Detection and Response) or similar protection mechanisms for servers' protection, including detailed resource cost information.

D(149): Facilities to extract summary reports from security logs should be available.

H(150): The system should provide the capability of rate limiting, and traffic shaping to prevent workloads from monopolizing certain system resources and thereby effectively denying service access.

H(151): The system should validate all input requests to prevent command injection and/or unauthorised manipulation of data.

R(152): Tenderers shall describe their approach to the management of information security. The description shall include:

a) whether they have an Information Security Management System in place.

b) whether they have a documented information security policy.

c) the time interval at which security policies are reviewed and updated.

d) whether a qualified third-party audit the Information Security Management System, indicating how often audits happens.

e) whether a qualified third-party review the security policies, standards, and procedures, indicating how often audits happens.

f) whether a complete set of their security policies is available for review.

R(153): Tenderers shall describe the methodology for vulnerability management. The description shall include:

a) whether they have a documented process for technical vulnerability management.

b) how they will inform ECMWF about potential security vulnerabilities in the System.

c) how they will provide security fixes and workarounds, including SLA details.

R(154): Tenderers shall describe their security incident management process. The description shall include:

a) whether they have an Information Security Incident Management process which clearly identifies responsibilities, procedures and communications.

b) whether they have a process for timely reporting to ECMWF and acting on information security events and incidents.

c) whether their process reflects the classification and severity of information security incidents.

d) whether, in the event of an information security incident, relevant data is collected in a manner which allows it to be used as legal evidence.

e) whether their organization has a process or framework which allows the organisation to learn from information security incidents and reduce the impact and/or probability of future events.

f) whether they have a process to detect and prevent infection by and the spread of malware.

g) whether they have a process to recover from a malware infection.

R(155): Tenderers shall describe the aspects of their Business Continuity Plan that are relevant to the delivery of support for the tendered solution. The description shall include whether contingency arrangements are in place for the supply of hardware (where this may be necessary) and the provision of software support.

R(156): Tenderers shall describe how they secure their endpoints used for remote access, monitor for unauthorised use and their procedures in the event of misuse or security breach.

R(157): Tenderers shall describe their approach to security for any personnel involved in the provision of services under this contract. The description shall include:

a) whether they conduct formal information security awareness training.

b) how they inform personnel of any information security requirements, including non-disclosure provisions.

c) whether they have formal procedures dictating actions that must be taken if an employee has violated any information security policies.

R(158): Tenderers shall describe their approach to information security management for any sub-contractors or suppliers that will be involved in the delivery of this service. The description shall include:

- whether information security is included in contracts established with their third-party suppliers and service providers.
- whether subcontractors and suppliers are provided with documented security requirements.
- whether subcontractors and suppliers are subject to regular review and audit.

Transition between DSMSs

It is expected that the volume of data stored in ECMWF's *MARS and ECFS services* will grow at the rate as described in Figure 3: Estimate of Primary and Secondary data holdings 2022 . It will become increasingly difficult to back archive data from one DSMS to its successor if this involves copying all of data from one set of media belonging to the old system to a set of media belonging to the new one. To migrate 1.5EiB over three years, it is estimated that ECMWF will need to move almost 66TiB/hour (51 streams of continuously flat out data transfer from 51 TS1160 to 51 TS1170 drives) It would be useful if a successor system could access and adopt the data on the media belonging to the old system. In essence, a successor system would need access to elements of the data object / tape placement metadata of the old system.

ECMWF wishes to preserve its ability to deploy alternative archiving solutions in the future. To this end tenderers will be asked to provide the facilities necessary to allow a simple data-reader application to be created, by ECMWF or by the tenderer on behalf of ECMWF. At the time of migration from the tendered system to its successor system, this data-reader would provide read-only access to the data stored by the DSMS in an environment outside of the tendered system.

M(159): It must be possible to migrate MARS and ECFS data, in some form, from the current HPSS system to the tendered solution. Either through metadata ingest or straightforward data copying.

R(160): Describe any tools or processes which could facilitate the migration of data from the current DSMS, based on HPSS, to the tendered DSMS.

H(161): It should be possible to read data held within the tendered solution without actively running the DSMS. With accessible metadata gained earlier from the system, be able to reconstruct the data objects found on each volume.

H(162): It is highly desirable that sufficient portions of the DSMS metadata can be made visible to an application outside the DSMS environment to enable it to read data objects stored on removable media within the DSMS. This requires that:

- It is possible for an application to create a list of the data objects stored in the DSMS and to map these to physical volumes and to their location on these volumes. It is acceptable to limit the scope of this mapping to data objects stored on removable media.
- The list must provide sufficient information that it is possible to clearly identify each of the data objects as seen from applications point of view. For example, to simply provide a list of internal object IDs without mapping these IDs to meaningful data object identifiers is not

useful. Metadata such as ownership, permissions, creation, access and modification dates should also be included.

- It must be possible to retrieve this information for all copies of the data objects.
- Should a data object be segmented onto multiple volumes, e.g. because it is too large for a single volume, then the locations of all of the segments must be provided, as well as the order of these segments.
- It must be possible to create such list in a reasonable time. It should (at least) be possible to run the DSMS in such a way that read access to the data is maintained while the list is being created, and that no updates to the removable media are performed.
- Should the data be encrypted in any form by the DSMS, then tenderers must provide a tool allowing this data to be decrypted.

R(163): Provide estimates of the time required to create such a list of data objects in the two cases of a DSMS containing one and five billion data objects.

R(164): Comment on any issues that would prevent or unduly complicate the creation of the desired simple data consuming application from the removeable media volumes.

R(165): Describe any thoughts on how both the Tendered solution could run in parallel with HPSS for a limited period before taking on the full responsibility of providing the total service. Allowing HPSS to be a read-only resource and using the new solution for new data. Later adopting the HPSS data once the new DSMS is operationally well established.

There is a general need to be able to modify disk storage to evolving requirements. Management tools should be provided to allow easy expansion of disk capacity and filesystems without any interruption to the normal operation of the system. Similarly, there is a need to be able to locate and relocate, or migrate, disk volumes, without interrupting normal system operation.

D(166): Describe the management of disk capacity within the DSMS. Covering how the capacity is added, migrated away from and removed. Give the maximum size a single fixed online data volume, for example HDD, that can be defined to the DSMS. In addition, state whether there is a limit to the number of such volumes that can be allocated to an instance of the Tendered solution.

Application Service Requirements

ECFS and MARS make use of knowledge of the sizes, locations, media identifiers, and availability information supplied by the system to determine how requests are queued, and when they are submitted to the DSMS.

R(167): Describe the available programmatic interfaces available for applications to query information describing how many underlying objects the data of the users' object is stored in, how many copies of the data are present, where these copies are in the system and whether they are available to the application.

D(168): It should be possible for an application to obtain the last access and modification date/times of any data object.

H(169): It is advantageous if the DSMS is able to accept indications, or hints, from the application of a likely usage timeframe for data which is being stored. This can help optimise against unnecessary read/write cycles to lower-performance storage tiers.

H(170): The system should be able to give a clear indication of the number of removable media drives, or other run-time resources, available or allocated to application workloads, to support managing of workload, queueing and the submission of requests from the application to the DSMS.

R(171): Describe the ability of the DSMS to offer an interface to the applications to allow them to gain performance or functional advantages of any parallel capability in the system. For example, I/O transfer rates or optimisations of data objects on removable media.

M(172): Applications must be able to store data objects in the system according to application-determined identifiers. It is helpful, but not required, if these identifiers take the form of paths. Applications must be able to retrieve data objects from the system according to these identifiers.

M(173): Applications must be able to efficiently read sets of partial byte ranges from within stored data objects. This may take the form of a set of (potentially hundreds of) offset-length pairs, or equivalently a sequence of seek and read operations. These byte ranges are typically of the order from 0.1 to several hundreds of MiB each.

M(174): The data objects written and read are typically too large to be held in memory by the application server processes. As such, the DSMS and the application facing APIs must support writing and reading data objects in the form of a continuous stream of data, or a sequence of read/write calls.

R(175): Describe the performance overhead of handling objects as streams of data, or sequences of read/write operations, rather than as bulk transfers.

H(176): For data security purposes, MARS writes and persists a selection of important data to tape immediately on receiving it. This data is then later aggregated with other data for long-term archival. It is highly desirable to support direct archival of a selected set of data to removable media or blocking functionality to determine when such data has been written.

H(177): As many of our retrieve requests are small relative to the data objects archived, it is highly desirable to be able to efficiently read sets of partial byte ranges from within stored data objects directly from their storage on removable media, without staging the entire data objects on intermediate cache filesystems.

H(178): Applications should be able to write data objects into the DSMS without the need to wait for removable media availability and movements, unless this blocking behaviour is explicitly requested.

R(179): Please describe the aggregate application write and read bandwidth that would be expected from the system under concurrent use. Further describe the bandwidth that would be expected per data stream to or from an application. For these numbers, please assume that the data is larger than the memory available for any single process, such that the data needs to be streamed rather than being read or written as a single large block.

R(180): A considerable proportion of the files stored in ECFS are deleted. In the case of ECFS about a third of files are deleted within 6 months of their creation or last modification. Comment on the effect that this has on the tendered solution. In particular, describe the effect of such deletions on the metadata used to describe the data objects, and on how resources allocated to deleted data objects can be reclaimed and reused.

R(181): If the DSMS is able to perform asynchronous I/O operations, tenderers are asked to describe the details of the API.

R(182): If the DSMS uses file systems in its higher storage tiers, please state which file systems are supported.

R(183): State the maximum number of clients that can be simultaneously connected to a single instance of the DSMS?

Existing Hardware

Disk subsystems

Currently the existing disk layer of the archive is mostly made up of;

- IBM FS5035 and FS5045 disk systems (~50 independent disk systems with a total usable capacity of around 50PB).
- a Dell PowerVault ME5084 disk system (1.5PB)
- a Ceph cluster, built upon Dell servers and Dell/EMC ME484 disk enclosures (24PB).
- Two IBM FS7200 NVMe flash appliances holding the metadata and backups of HPSS and the two applications MARS and ECFS (170TB).

The disk systems use conventional block storage whereas Ceph is advertising CephFS file systems to MARS mover servers. Most disks in the disk layer are performance based, which means they house 8 or 12TB NL-SAS disk drives. The remaining systems are capacity based, made using between 16 to 24TB NL-SAS hard drives. HPSS uses raw LUNs presented from the disk systems of which ECFS is a heavy user, whereas MARS uses ZFS or CephFS file systems. The ZFS file systems being built from LUNs presented from block storage disk systems.

Each of the flash appliances are 85TB of NVMe and hold multiple copies of the HPSS and application metadata.

R(184): Confirm whether these existing disk and NVMe systems are supported by the proposed solution. If not, name which alternative systems are covered under your support.

R(185): Indicate any performance or management benefits if recommending any alternate system.

Storage Area Networks (SAN)

ECMWF operates a high-performance SAN infrastructure designed for resilience, scalability, and high availability. The architecture is built around four Brocade X6-8 SAN Directors, with two directors deployed in each data hall (DS1 and DS2). Each hall hosts two independent SAN islands, providing fault isolation and operational flexibility.

Dual SAN Islands per Hall: Each data hall (DS1 and DS2) contains two SAN islands, ensuring that workloads can be distributed and isolated for performance and fault tolerance. Each of the Brocade X6-8 directors offer 384 ports at 32Gb/s, ultra-low latency, and advanced fabric services. Currently these directors are 75% populated. Over 50 IBM FlashSystem storage arrays and 400 servers are connected across the SAN islands, supporting a wide range of workloads from high-throughput MARS and ECFS workloads to latency-sensitive database applications.

Four Brocade G720 64 port switches, two in each hall provide the connectivity for the tape library control paths to the IBM TS4500 and Spectra Logic TFinify libraries.

R(186): Confirm whether the proposed solution supports the Brocade SAN environment of X8-6 directors and G720 switches that the Centre currently deploys. If not, explain which alternative device I/O connections would be recommended and covered under your support for both the disk and tape environments.

R(187): Indicate any performance or management benefits if recommending an alternative to the Brocade SAN.

Automated Tape Libraries

ECMWF presently makes use of eleven IBM TS4500 and one Spectra Logic Tfinity tape library based on TS1160 and TS1170 for the primary copy of data and two IBM TS4500 with LTO-8/9/10 libraries for the secondary copy. Many of the Centre's drives are now SAS connected and not Fibre Channel connected.

R(188): Describe how this equipment could be integrated into the proposal.

R(189): Indicate any performance or management benefits if an alternative tape library type is recommended.

R(190): State whether the tendered DSMS is able to make full use of all of the features of ECMWF's existing IBM TS4500 and Spectra Logic Tfinity robotic libraries.

R(191): State any restrictions on media compatibility beyond those imposed by the tape drive manufacturer.

Servers

The servers currently used are x86 servers, either from HPE or Dell. The Core servers are all HPE Proliant DL380 with 24 Core Intel CPU and 1TB memory. They each have access to two NVMe based IBM FS7200 flash arrays, one it each hall for redundant protection. The mover servers are a mix of HPE Proliant DL360 and Dell R630 servers. Each with a single or dual socket 8-core CPU and 64GB of memory. The current tape movers are a mix of HPE DL380s and Dell R440s. Either with two quad port or four dual port cards to gain direct access to the tape drives.

R(192): Describe how this equipment could be integrated into the proposal.

R(193): Indicate any performance or management benefits if recommending or requiring an alternate server type.

Integration with ECMWF's Recovery System

In the current HPSS configuration, the secondary copies of MARS and ECFS data are stored in the *alternate Data Storage* Hall. Primary data held in DS1 will have its secondary copy in DS2 and vice-versa.

The new DSMS will also need to write its secondary copies of data in the alternate hall. Although this does not mean that it needs to make a mirror copy of this data, it must however have a means to reconstruct the data from an independent source completely available in the alternate hall. Tenderers are asked to explain how the system will be able to integrate with the two IBM TS4500 libraries running with LTO-9 and LTO-10 drives. It could possibly be based on an alternative solution proposed by the Tenderer, in which case the costs involved in deploying such solution will have to be provided. No matter what solution is proposed, it must be noted that the service will need to provide a recovery service on the basis that one hall is completely unavailable.

R(194): Tenderers shall describe how they could integrate with the existing recovery service or how they would recommend constructing an alternative over the two data halls. Include any business continuity benefits as well as it's resistance to more catastrophic failure of a single data hall.

R(195): Tenderers shall describe any method allowing their DSMS to generate a secondary copy of critical data (using either the existing hardware or suggest an alternative option).

R(196): ECMWF may decide to hold its secondary copy of data off-site with all removeable media volumes being held within automatic tape libraries. Part of the DSMS would run at this alternate site to support the reading and writing to data to the secondary media. In such a case, Tenderers shall describe whether there would be any cost implications, effecting licencing, support or other costs. If other costs are implicated, no estimates of costs are needed at this stage, however please details the any increases above those quoted above and what they would be based upon.

R(197): Tenderers shall describe any implications of keeping only 10 percent of all volumes created in the Recovery System under robotic control in the Bologna Data Centre. They shall describe licence costs involved, and potential management issues that would result from not having direct access to the offline tapes.

Evaluation Criteria

Evaluation Method and Selection Criteria

Tenders will be evaluated based on the evaluation criteria and weights shown in the table below.

Evaluation criteria	Weight
Tenderer's Credentials - Financial standing and Legal organization - Track record and references including depth of experience with large-scale storage solutions	10%
Quality of Proposal - Performance/Scalability - Resilience/Recoverability of solution - Management/Operation/Configuration - Maintenance/Support - Hardware/Software compatibility - Usability for System Administrators and Operators - Security - Applications (Client interface)	55%
Cost and Prices - Price quoted and long-term costing based on the TCO of the solution - Level of price guarantees through the entire contract - Prices of consultancy support	35%

Table 2: Evaluation Criteria

Tenderers must achieve a mark of at least satisfactory (i.e. 60%) for each high-level criterion. The preferred Tenderer will be selected based on obtaining the highest overall score following the evaluation and which offers the best value for money to ECMWF considering the most advantageous offer overall regarding total cost of ownership, technical value, quality, and other characteristics, including implementation risks and affordability.

Licensing of Data Storage Management System

ECMWF holds a very large volume of data in its archive and migrating to a different product would be a multi-year effort, therefore it is of great importance that the tenderer can commit to licensing and supporting for the full period as specified in 'Term of the License and the Support'

The *licensing and support* fees should be payable either annually or quarterly. To assist ECMWF with future planning ideally the cost of support for each year in this period would be fixed as part of the proposal.

Pricing and pricing mechanism

Table 3: Example hardware available for DSMS

Hardware Equipment	Servers dedicated to the DSMS	Online storage (Disk/NVMe)	Automated Tape Libraries	Primary Tape Drives	Secondary Tape Drives	SAN Infrastructure Ports[2]
2026	**120**	**15PB**	**14**	**740**	**80**	**FC 32Gbs 1800**
2027	**120**	**20PB**	**14**	**740**	**80**	**FC 32Gbs 1800**
2028	**140**	**28PB**	**14**	**700**	**80**	**FC 32Gbs 1800**
2029	**140**	**39PB**	**14**	**600**	**64**	**FC 32Gbs 1800**
2030	**180**	**55PB**	**14**	**500**	**64**	**FC 32Gbs 1800**
2031	**180**	**76PB**	**14**	**400**	**64**	**FC 32Gbs 1800**
2032	**220**	**107PB**	**16**	**550**	**64**	**FC 64Gbs 2000**
2033	**220**	**150PB**	**16**	**650**	**64**	**FC 64Gbs 2000**
2034	**240**	**210PB**	**16**	**750**	**64**	**FC 64Gbs 2000**
2035	**240**	**290PB**	**16**	**750**	**64**	**FC 64Gbs 2000**

R(198): The Tenderer must explain the basis by which they plan to charge for the licensing and support of their product. Explaining how the mechanism will be used to derive future prices and how any inflation linked increases are factored in. Whether the licensing is based on the capacity of the system, the hardware used to run the system or a site wide license. Figure 3: Estimate of Primary and Secondary data holdings 2022 – 2035 shows the expected growth over ten years and Table 3: Example hardware available for DSMS shows the expected hardware to be used over the initial period.

[2] The current tape drive infrastructure is partially SAS based and the Fibre Channel based drives are connected directly to mover hosts and not the SAN environment.

R(199): Tenderers must provide fixed prices for the licensing and support of their product and consultancy for at least the next five years, based on the growth expectations above and explain the mechanism that will be used to derive future prices. The mechanism should include any inflation linked increases.

R(200): If tenderers are prepared to use the pricing mechanism beyond the initial ten years. Tenderers should specify the notice period required to extend the support, the duration they are prepared to extend by and any limits on the number or length of any extensions.

R(201): Only for Tenderers who are requiring additional equipment to run their solution, provide a detailed quote including breakdown of components and services. Delivery and installation will be to ECMWF's Data Centre in Bologna, located at Tecnopolo di Bologna, Bologna, Italy. Please detail the equipment recommended, as this may be available to ECMWF via alternative procurement arrangements. At the end of year 5, estimate the cost of replacing this required hardware, record any changes to the licensing costs.

R(202): Only for Tenderers who are recommending alternative equipment to run all or part of their solution. Provide a detailed quote including breakdown of components and services. Delivery and installation will be to ECMWF's Data Centre in Bologna, located at Tecnopolo di Bologna, Bologna, Italy. Please detail the equipment recommended, as this may be available to ECMWF via alternative procurement arrangements. At the end of year 5, estimate the cost of replacing this required hardware, record any changes to the licensing costs.

The prices shall be quoted in Pound Sterling (£) or Euro (€). For the purposes of comparison for this ITT, prices will be converted into a single currency at a conversion rate to be established as the average ECB exchange rate for the calendar month prior to the closing date of this ITT.

Schedule

Following receipt of replies to this ITT by the date and time specified in Volume IA, ECMWF envisages, based on current Volume IA details, the following schedule for the implementation of the project:

05 March 2026: Closing date of the ITT

By mid-April 2026 following initial evaluation: potential clarifications, demonstrations or site visits with shortlisted tenderers

May and June 2026: Negotiation with preferred tenderer

July 2026: Signature of the contract

September/October 2026: Installation of tendered solution ready for acceptance testing

Presentations and demonstration

- If requested by ECMWF, tenderers must give a presentation at ECMWF of the tendered system during the period stated in volume IA or as otherwise may be requested. The date of the presentation will be agreed with tenderers following initial evaluation. Costs for presentations cannot be reimbursed.

- If requested by ECMWF, tenderers must arrange for demonstrations to be given of the software representative of that tendered. Arrangements of the place, time, and exact content of the demonstrations will be made known following initial evaluation.
- Tenderers shall provide a list of reference sites that would be willing to discuss with ECMWF matters related to the tendered software.
- If requested by ECMWF, tenderers should facilitate site visits to other organisations using the software being tendered. The timing and organisation of such visits would be made following initial evaluation.

PROCEDURE FOR THE SUBMISSION OF APPLICATIONS

Presentation and Order of the Tender

The Tender response shall be presented as separate documents, which are to be uploaded to the respective question on the eProcurement Portal, as follows:

- Completed Volume IIIA (Template for Tenderers – Administrative Information);
- Completed Volume IIIB (Template for Tenderers – Response to Specification of Requirements);
- Completed Volume IIIC (Template for Tenderers – Response to Specification of Requirements – Pricing Elements
- Attachments and Annexes, as requested in the Volume III documents.

Note that for any information that has been provided as part of the Tender, but not specifically requested by ECMWF, then ECMWF shall, at its sole discretion, decide whether to utilise that further information within its evaluation process.

Volume IIIA

Volume IIIA (Template for Tenderers – Administrative Information), which can be found under the ITT Online Questionnaire should be completed for the following information:

- Details about your organisation:
  Information on the legal, commercial or professional status of the Tenderer, as well as contact details of the person who can be contacted by ECMWF in relation to the Tender. The Tenderer should also attach a copy of the official Company Registration Document and provide complete and accurate information on the Tenderer's shareholding structure and, if applicable, full details of its parent organisations up to and including the ultimate parent organisation.
- Economic and financial capacity:
  Financial information on your organisation to enable us to evaluate your financial status.
- Staff resources.
- Experience and references.
- Additional questions:
  This *section contains a set of questions which seek either information or confirmation from the Tenderer.*
- Confirmation of agreement to Volume IV of the ITT (Draft Contract).

Volume IIIB

Volume IIIB (Template for Tenderers – Response to Specification of Requirements) shall contain the Tenderer's response to the requirements specified in Volume II. The document lists the requirements and

provides a structure for the Tenderer's response. This is the minimum information requested; Tenderers can provide any additional information or documents as necessary. Some of the requirements make reference to various sections in Volume II and do not provide the full description of the requirement; Tenderers are advised to formulate their response based on the description of the requirement provided in the respective section of Volume II and touch upon all the elements described or requested therein. The ***Figure 4: Structure of the provision of the ECFS service at ECMWF*** can be used to cross-reference which section in Volume IIIB to answer individual questions. The questions are in sequence in Volume III but should be answered in their functional sections in Volume IIIB.

ECMWF seeks focused responses, rather than responses which include a significant amount of standard marketing material. If you wish to include marketing material in your proposal documentation set, it should be provided as discrete documents and limited to only marketing material which is directly relevant to the response and marked as "Marketing Material", however ECMWF may, at its sole discretion, not evaluate any such marketing material.

Questions in Volume II	**Answer in Volume IIIB Section**
M(1), **R(2)**, **M(4)** - **R(8)**, **R(12)** - **R(14)**, **R(16)**, **M(18)**, **M(19)**	Maintenance and Support
R(3), **R(9)** - **M(11)**, **M(15)**, **R(17)**	Pricing Requirements
M(20) - **D(33)**	Management, Operation and Configuration
H(34) - **D(45)**	Resilience and Recoverability
H(46) - H(49)	Management, Operation and Configuration
R(50), **R(54)**, **R(58)**	Performance and Scalability
R(51) - **R(53)**, **R(55)**, **H(56)**, **R(67)**	Hardware/Software Compatibility
R(57), **R(60)**, **M(68)** - **H(70)**	Resilience and Recoverability
H(59), **R(61)** - **D(64)**	Management, Operation and Configuration
H(65), **H(66)**,	Usability For System Administrators and Operators
R(71) - **D(77)**, **R(80)** - **H(85)**	Management, Operation and Configuration
M(78), **R(79)**, **H(86)**	Hardware/Software Compatibility
H(87) - **H(89)**, **M(93)**, **H(95)** - **R(100)**	Resilience and Recoverability
R(90), **R(92)**, **R(94)**	Performance and Scalability
R(91), **D(113)**, **R(114)**, **H(118)**	Management, Operation and Configuration
R(101), **R(102)**, **R(104)** - **H(109)**	Resilience and Recoverability
R(103), **H(110)**	Performance and Scalability
R(111), **H(112)**, **D(115)**, **R(116)**	Resilience and Recoverability

M(117)	Hardware/Software Compatibility
D(119) - **H(137)**	Usability For System Administrators and Operators
R(138), **R(139)**	Management, Operation and Configuration
R(140) - **R(158)**	Security
M(159) - **H(162)**, **R(164)**, **R(165)**	Hardware/Software Compatibility
R(163)	Performance and Scalability
D(166)	Management, Operation and Configuration
R(167) - **R(183)**	Applications (Client Interface)
R(184) - **R(195)**	Hardware/Software Compatibility
R(196) - **R(202)**	Pricing Requirements

Table 4: Index of Questions to Volume IIIB

Volume IIIC

Volume IIIC (Template for Tenderers – Response to Specification of Requirements – Pricing Elements) shall contain the Tenderer's response to the requirements specified in Volume II related to the pricing and optional costs of the tender. The document is an Excel spreadsheet which lists the specific requirements and provides a structure for the Tenderer's response. Within the excel spreadsheet, a tab has been provided for each cost type, with space provided for cost estimates and the Tenderer's detailed explanations. The summary tab is then linked to each cost type tab to bring together the main cost information in one place. The Tenderer may add additional lines as needed on the individual tabs, and amend the summary tab formulae to include them. This is the minimum information requested; Tenderers can provide any additional information or documents as necessary.

Volume IV

Tenderers should note that the successful Tenderer is expected to sign up to the T&Cs attached to this Volume IV. The T&Cs for Services will be applicable and if hardware purchases will be acquired as part of this Tender, the T&Cs for Goods will additionally be applicable.

Tenderers are requested to confirm their acceptance of the terms and conditions by checking the corresponding box in Volume IIIA. Should any of the clauses in Volume IV pose difficulties, Tenderers should list such clauses and explain why they would impact their ability to deliver the services in this contract. A Tenderer's acceptance of these terms and conditions form part of the evaluation process and hence will contribute to the success of a Tenderer's bid. ECMWF will consider any issues that are raised in a Tenderer's response and may agree to changes in these terms and conditions. Requests for changes identified later in the process, which have not been identified here, may not be considered.

Please note that as a consequence of ECMWF's extraterritorial status and immunities, any contract, including any sub-contract and any ensuing contract between ECMWF and a third party (e.g. maintenance/support/license agreements), resulting from this ITT must contain an arbitration clause which is offered by ECMWF to all contracting parties. Tenderers are requested to explicitly confirm their willingness

and ability to comply with this requirement and describe their approach in ensuring that such contracts comply with the above requirement. Further information may be found at https://www.ecmwf.int/en/about/suppliers in the documents under "ECMWF's status: Arbitration and VAT".

Acceptance of the System

Tenderers should note in particular Annex 3 which contains the requirements for testing and accepting the System as it is installed in its phases. In outline, the acceptance sequence is as follows:

• At least 30 days before installation (anticipated September/October 2026), ECMWF will advise the Contractor of the detailed specifications of the acceptance tests and of the test data, together with the expected results of processing the data.

• Following installation, a 28-day period during which inter alia the test data and scripts are prepared by ECMWF with help from the Contractor during which staff familiarize themselves with the System.

• A 5-day Functional Test to ensure that the System provides all the functionality and performance as described in the ITT specification of requirements and as responded to by the Contractor in his tender. Thus, the Test will inter alia check that: the System meets the contracted performance criteria; the System operates properly under normal and stress loads; the integration of the various sub-systems is harmonious and effective; and the System is able to respond efficiently to simulated incidents and to mitigate adequately any adverse impact of such. The Test will simultaneously check the System's support for both the ECFS and MARS applications if such a tender is accepted. Further details of the performance testing for each application are contained in Annex 3 Requirements of Acceptance;

• A 30-day Reliability Test to check that the System has adequate reliability and availability. Flexibility is provided within the draft agreement to allow some re-testing, but tenderers should note that continued failure of testing could lead to termination of the contract and the return to ECMWF of all monies paid under such contract. The System will be tested as a whole system.

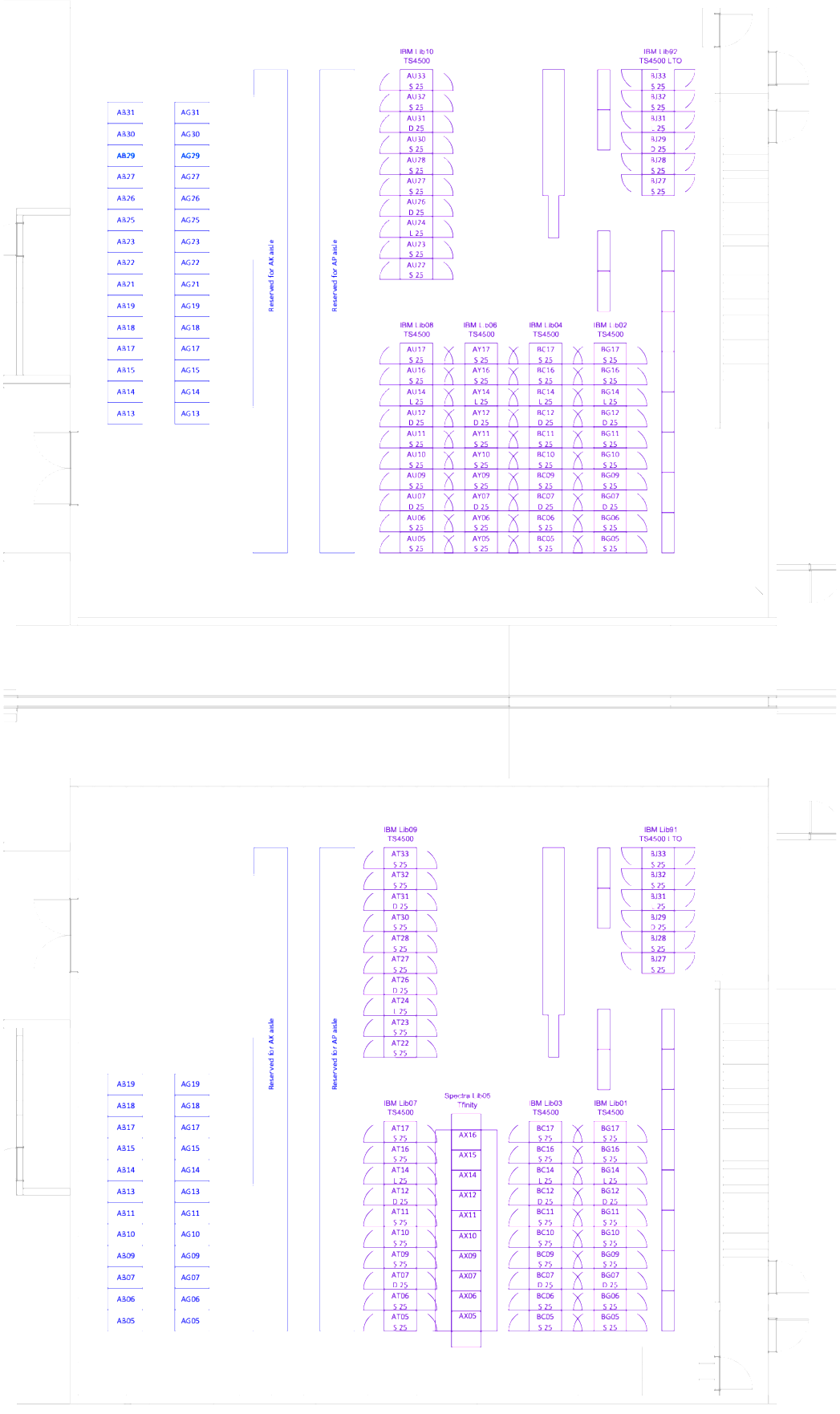
The payment schedule for the procurement is linked to the acceptance of the system.

Assistance with the system will be required during the period of acceptance. If any hardware component is supplied, maintenance will also be required.

Tenderers to provide diagnostic assistance where necessary.

Annex 1 – No longer used

Annex 2 Data Hall layout for Data Storage 1 and Data Storage 2 halls



Annex 3 Requirements of Acceptance

Specification of Hardware

To run the acceptance a model set of hardware will be used to provide a benchmark of performance and reliability. This will consist of the components listed below. However, as it may be necessary for the Tenderer to require or recommend alternative hardware this can be accommodated for and the testing done on recommended hardware instead.

The performance metrics given as responses to some of the specifications in Volume II will be validated against the baseline hardware, or the agreed hardware for the acceptance tests. Failure to meet the metrics given in response to the specifications, will require additional or alternative hardware, including any software changes and covering related costs, which would need to be borne at the Tenderers expense.

Servers

Two types of servers will be used: A single metadata core server and 10 I/O mover servers. Some of the I/O mover servers will be used for online disk storage and some as tape drive movers. This may not suit all configurations and adaptations can be made to suit the solution and testing environment.

Server Type	No. used	CPU	Memory	Network	HBA
Metadata Core Server	1	HPE DL380 Gen 10 2 socket, each Intel E - 18 core	1 TB DIMM	Bonded 2 x 25Gb/s interfaces	HBA HPE SN1100Q dual port FC – to disk SAN, HBA HPE SN1100Q dual port FC – to library control SAN
Disk based I/O Mover Server	4	HPE DL360 Gen 10 2 socket, each Intel - 8 core	64 GB DIMM	Bonded 2 x 25Gb/s interfaces	HBA HPE SN1610Q dual port – to disk SAN
Tape based I/O Mover Server	6	HPE DL380 Gen 11 2 socket Intel – 12 core	64 GB DIMM	Bonded 2 x 25Gb/s interfaces	HBA Lenovo ThinkSystem 440-8e dual port SAS – direct connect to tape drive

Metadata (NVMe) system

Used to house the metadata of the DSMS. This will be an IBM FS7300 system, based on Flash Core Module 4 will ample space to accommodate the metadata of a configured acceptance system.

SAN

The SAN will be shared with other operational work but independently zoned to only the fibre channel components of the acceptance system. The disk and metadata system will be separately zoned the appropriate servers using Brocade X6 directors. All paths will be made available between the host and the devices, with the paths being spread over different blades in the directors and over two directors.

Disk systems

The disk systems will be two IBM FS5045 systems, each with one controller and 9 enclosures. Each system will contain 120 12TB NL-SAS disk drives. These systems will be exclusively available during the acceptance.

Tape libraries

As it is not straightforward to isolate a library for exclusive use for the acceptance. Each library in these tests will be allocated with a separate logical partition to act as an independent library. Two partitions will be defined in an IBM TS4500 and one in a Spectra Logic Tfinity library.

- A logical library in an IBM TS4500 will be made up of 20 tape drives (IBM TS1170) and 1,000 slots and access to a VIO area of 250 slots.
- A logical library in the IBM TS4500 will be made up of 8 tape drives (LTO-9) and 300 slots and access to a VIO area of 100 slots.
- The logical library partition for the Spectra logic Tfinity library will be made up of 20 tape drives (IBM TS1160), 1,000 slots and an Entry/Export area size of 250 slots)

Define Model Hardware to perform tests on

- If the Tenderer is not able to run on the equipment above or has recommendations of other hardware, these should be stated in Vol IIIB (Requirements 10 and 11). List which hardware components they would like to replace and the reasons for believing this would be a preferred choice.

Define performance metrics

- Refer to Vol II perf. Metrics

List of tests to setup and run

Acceptance setup

The DSMS will be primed with an adequate number of data objects to test the assumptions made in the performance tests, with objects being of varying sizes. All data will be test data with the possibilities to delete and replace as appropriate.

Running the acceptance tests

An set of tests will be prepared once a storage solution is selected. These tests will reflect the questions laid out in Volume II and any reasonable verifications of commitments made in response to this ITT. The acceptance tests will be agreed up on by both parties and will be performed in two stages; a functional test and a resilience test. The acceptance tests will become part of the contract between ECMWF and the Tenderer. Staff of the Tenderer will be allowed to be present during this testing phase and may well be useful to quickly adapt to a allow retesting to particular elements of the tests.

Functional Tests

The functional tests will be run over five days to validate the features of the DSMS and claims made over the performance of the system. If any part of the tests fail, they can be retested, but this may at ECMWF's discretion cause the whole five-day testing cycle to be repeated.

Resilience Tests

Following the successful completion of the functional tests, the testing immediately starts on the resilience tests. This will consist of running of 30 days continuously, where workloads are placed on the system and behaviour observed. If at any point during those 30 days components of the system fail, the whole 30 days will need to be restarted. This is to demonstrate that the solution can run consistently without administrative input to maintain operational use. If the cause of any failure is determined to be as a result of a starvation of media resources then this error lays with ECMWF and not the vendor, so would not cause the test to be repeated.

If a form of storage solution migration from HPSS is suggested, this could be tested partially in the functional tests but also an actual migration would be attempted during the resilience test period.

Glossary

Item	Meaning
Data object or file	A collection of data that can be created, read from, written into, and deleted from a storage system, and which is treated as a string of bits without any particular structure. Such an object may correspond to a file within a client's application view, but not necessarily so. A file can be interpreted as an object depending on the storage solution offered.
Data pool	A collection of similar storage media (for example, a group of disk or tape volumes) managed as a single entity, for example, a set of volumes found in a robotic library or those managed by a specific server.
Media family	A collection of removeable storage media which allows an application to direct certain data objects at a discrete set of media volumes. Thus attempting to improve the number of read requests on a single volume and reduce the mount workload on the automated libraries.
DSMS migration	The process by which data stored in the current HPSS system for MARS and ECFS are transferred to a new DSMS system.
Media drive	Currently this can be read as a tape drive. However, this could also be interpretted as any other device drive that could read media, be that optical, magnetic, or SSD.
Robotic library	Equipment to contain a number of removable media and their drives, such that the process of mounting and dismounting is automatically controlled, this being driven by the DSMS.
Removable media	Media for storage and recall of data in a media drive, such a tape or optical. Within this ITT wherever reference is made to tape drives and tape media, tenderers may provide descriptions and responses that might embrace other types of removable media
Secondary copy of data	A second copy of data held in the primary layer of the DSMS. In ECMWF's case only the critical data is held with a second copy. This copy is separated from the primary, either in robotic libraries in the alternate data hall or potentially at an alternate site.
HPSS	High Performance Storage Systems, a collaboration between IBM and US Dept of Energy laboratories to produce a very scalable data storage archive system.
DB2	IBM's database system, used by HPSS to manage its metadata
Data Halls	Data Storage Hall 1 & 2 at the Tecnopolo in Bologna as detailed in 'Annex 2 Data Hall layout for Data Storage 1 and Data Storage 2'
Endpoint Detection and Response (EDR)	Monitor client activity for signs of suspicious activity.
RBAC	Role Based Access Control
RAO	Recommended Access Order – feature on enterprise tapes for optimising tape file access and recall times in the Glossary table.

When referring to data storage capacity (e.g. disk/tape/memory sizes), multipliers are indicated in powers of 2. In all other contexts (e.g. data rates), they indicate powers of 10. See Table 5: Byte Scale for Data capacity and transmission rates below.

Table 5: Byte Scale for Data capacity and transmission rates

Name	Kilo	Mega	Giga	Tera	Peta	Exa	Zetta
Capacity	1 KiB	1 MiB	1 GiB	1 TiB	1 PiB	1 EiB	1 ZiB
	2^{10} bytes	2^{20} bytes	2^{30} bytes	2^{40} bytes	2^{50} bytes	2^{60} bytes	2^{70} bytes
Data Rates	1KB/s	1MB/s	1GB/s	1TB/s	1PB/s	1EB/s	1ZB/s
	10^{3} bytes	10^{6} bytes	10^{9} bytes	10^{12} bytes	10^{15} bytes	10^{18} bytes	10^{21} bytes