REQUEST FOR INFORMATION

FOR A HIGH PERFORMANCE COMPUTING FACILITY (HPCF) FOR ECMWF

ECMWF/RFI/2024/365

March 2024

TRADEMARKS

All names or descriptions used in this Request for Information (RFI) that are trademarks, trade or brand names, or other references to proprietary products are hereby acknowledged as the property of their respective owners. No entry, term or definition in this RFI should be regarded as having any implication as to the validity or otherwise of any trademark.

The appearance of any proprietary name or reference in this document should not in itself be taken to imply a preference for one product over another unless specifically stated otherwise.

Table of Contents

Tables

1. INTRODUCTION

1.1. Purpose

The purpose of this Request for Information (RFI) is to provide information relevant to the procurement of a future High Performance Computing Facility (HPCF) for the European Centre for Medium-Range Weather Forecasts (ECMWF), so that feedback from potential participants can be considered prior to issuing an Invitation to Tender (ITT).

This RFI has been written with the expectation that responses will primarily come from integrators who would be able to respond to the future ITT for the provision of the complete service. However, the purpose, first and foremost, is to gather information. Responses are therefore welcome from providers of key technology elements, e.g., storage, processors or accelerators.

Specifically, we seek to do the following:

- Identify technologies available in the 2026-2027 timeframe, especially those requiring ECMWF to investigate and test them so that it can exploit them to achieve its performance objectives.
- Determine the necessary level of investment (scope, schedule and budget) required to secure the necessary performance objectives.
- Establish the level of interest and capabilities of providers in working with ECMWF to achieve its goals and identify any barriers to providers responding to a future possible ITT.
- Understand alternatives to ECMWF's current contractual financial model and the benefits to both parties of them.

ECMWF acquired its existing HPCF service under an agreement that will expire at the end of October 2027. The service is delivered by four Eviden BullSequana XH2000 systems, each with 1,920 128-core application nodes and around 100 post-processing nodes, along with Lustre filesystems based on SSDs and HDDs with aggregate bandwidth exceeding 2.7TB/s. For more information, see:

https://www.ecmwf.int/en/computing/our-facilities/supercomputer

The replacement HPCF is currently expected to be installed in our Bologna data centre at the end of 2026 or the beginning of 2027 so that it can be commissioned and the entire workload migrated to it before the existing contract ends.

ECMWF's goals can be found in the strategy. The current strategy for the period 2021-2030 is available here:

https://www.ecmwf.int/en/about/what-we-do/strategy

ECMWF reviews its strategy at least every five years and is aiming to have an update adopted in December 2024 so that it would cover the period 2025-2034.

The purpose of the strategic review is to enable ECMWF to maintain its leading role in Numerical Weather Prediction by responding to fast-paced developments including:

- the evolution of AI/ML models and consequent disruptive changes to the value chain;
- the increasing relevance of environmental monitoring products and services for policy making;
- the development of initiatives to digitally enable environmental programmes such as the European Commission's Destination Earth and the opportunities and challenges for ECMWF and its Member States.

For this RFI, ECMWF is interested in the specification, configuration and indicative costs of building blocks that could be put together to provide resilience and the performance to meet the requirements of the ECMWF strategy.

1.2. Role of ECMWF

ECMWF is a world leader in its field, producing the best available global medium-range weather forecasts and maintaining a comprehensive research programme to continue to improve the quality of these forecasts. Its high-performance computing resources are used both for time critical forecast production and for extensive research and development work as well as providing a shared computing resource for our Member States own usage.

ECMWF is an independent intergovernmental organisation supported by 35 States and is 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.

Information on ECMWF's activities can be found at:

https://www.ecmwf.int/en/about

1.3. Conditions for submission of a response

1.3.1. Disclaimers

This is an RFI issued solely for information and planning purposes and does not constitute a solicitation for a system. ECMWF does not commit to issue a related Invitation to Tender (ITT). ECMWF reserves the right to change the details of this RFI or withdraw this RFI at any time. Respondents are solely responsible for all expenses associated with responding to this RFI.

Nothing contained in this RFI, or any other communication made between the respondent and ECMWF or its representatives shall constitute an agreement, contract or representation between ECMWF and any other party. Receipt by a respondent of this RFI does not imply the existence of a contract or commitment by or with ECMWF for any purpose.

While ECMWF has taken all reasonable steps to ensure, as at the date of this document, that the facts which are contained in this RFI are true and accurate in all material respects, ECMWF does not make any representation or warranty as to the accuracy or completeness or otherwise of this RFI, or the reasonableness of any assumptions on which this document may be based. ECMWF accepts no liability to respondents whatsoever and however arising and whether resulting from the use of this RFI, or any omissions from or deficiencies in this document.

ECMWF may use the information included in a response for any reasonable purpose connected with this RFI or any future ITT.

1.3.2. Timetable

This RFI will close at 14:00 UK local time on Wednesday 1 May 2024

ECMWF envisages the following schedule for this project:

5 March 2024	Issue of this RFI
6 March – 16 April2024	Discussions between ECMWF and vendors to clarify the RFI specification
17 April 2024	Last date for submission of clarification questions for this RFI
1 May 2024 14:00 UK local time	Close of RFI
June – October 2024	Development of strategic options and benefits for review by ECMWF Technical committees
December 2024	Decision by ECMWF Council on procurement of next HPCF
Second half of 2025	Issue of Invitation to Tender for new HPCF
Beginning of 2026	Receipt of tenders
First half of 2026	Evaluation of tenders and negotiation of contract terms
	Selection of the winning tender
Mid 2026	Submission of the contract to ECMWF's Council for approval, followed by signature of the contract.
End of 2026/Beginning of 2027	Start of installation of new HPCF
31 October 2027	End of current HPCF contract

1.3.3. Confidentiality

The contents of this RFI together with all other information, materials, specifications or other documents provided by ECMWF, or prepared by respondents specifically for ECMWF, shall be treated at all times as confidential by the respondents unless it is already in the public domain. Respondents shall not disclose any such information, materials, specifications or other documents to any third parties or to any other part of the respondents' group or use them for any purpose other than for the preparation and submission of a response to this RFI nor shall respondents publicise ECMWF's name or the project without the prior written consent of ECMWF. Respondents shall ensure that all third parties to whom disclosure is made shall keep any such information, materials, specifications or other documents confidential and not disclose them to any other third party except as set out above.

ECMWF reserves the right to retain all documents submitted by respondents in response to the RFI. Any information in such documents that is proprietary and confidential to the respondent will be handled confidentially by ECMWF provided it is clearly and specifically identified as such. Such obligation shall not apply if such information is or was obtained from other sources that do not bind ECMWF as to confidentiality or if the information is in the public domain. ECMWF may make responses available for evaluation purposes to authorised people including its governing body, committees, and professional advisers in addition to ECMWF's own personnel under the same conditions of confidentiality.

Please also note that all personally identifiable information (PII) requested by ECMWF or provided by respondents will be treated in accordance with the ECMWF Policy on Personally Identifiable Information Protection (PIIP). It is available at https://www.ecmwf.int. ECMWF shall process all PII submitted by your response for the sole purposes of assessing your response. In doing so, ECMWF may share such PII with consultants or external advisors.

1.3.4. Enquiries and contact procedure.

In order to be kept up to date with any clarification responses or amendments to the RFI, the invitee is requested to confirm to the email address hpc2024@groups.ecmwf.int whether or not it will be submitting a response and must provide a contact point and contact details to which email notification of the publication of any additional information will be sent. Please give your contact's name, title, address and location, telephone number and email address.

Any other enquiries or requests for clarification of any matters arising from this RFI should also be sought from hpc2024@groups.ecmwf.int at ECMWF and must be made in writing by email, no later than the last date for submission of clarification questions indicated in section 1.3.2

Where ECMWF supplies further information, it will make this information available to all recipients of this RFI who have indicated their intention to submit a response

and provided ECMWF with an e-mail address for communication of additional information.

1.3.5. Format of the response

At the beginning of your response you may provide a short description of your company and similar services that you have provided recently. Please respond to the questions that are relevant to your solution in the sections below, quoting the question before you provide the answer. An Excel spreadsheet is provided for your cost estimates. Please do not provide your company's general advertising material with your response.

1.3.6. How to submit a response

Responses must be written in English.

The respondent must submit their response to

hpc2024-submission@groups.ecmwf.int

as an email with attachments containing its complete response to this RFI. The attachments must contain a printable version of the response in Microsoft Word format, Rich Text Format (RTF) or Adobe Portable Document Format (PDF) and in Microsoft Excel format for any spreadsheets. The email should confirm that the response has been submitted by a duly authorised director or senior officer of the respondent.

The subject of the email must be:

Response to RFI/2024/365 for a High Performance Computing Facility for ECMWF.

2. Background and Scope

This section provide guidance on the specification of building blocks of an eventual High Performance Computing Facility (HPCF) that would meet ECMWF's requirements for installation in late 2026/early 2027.

A building block is a unit of computational or storage performance, bigger than a single node, with all nodes connected by a high performance interconnect. There is no fixed definition of the size of a CPU or GPU building block, respondents are free to choose the most logical size for their architecture; examples might be a full rack of nodes, or all the nodes that can be connected to a single switch. Benchmarks are included in this RFI to determine the performance of a *node* in the building block, and optionally, of the entire building block or part of it, on respondents' new technologies, and compare with results achieved on the current ECMWF HPCF. With this information ECMWF can make a high-level estimate of the likely cost and requirements for options as it develops its next strategy.

Ancillary nodes to provide such services as system management, batch scheduling, network and file-system access and any other equipment/functions required to integrate the building block into a system, for the purposes of this RFI, should not be considered part of the building block. ECMWF envisages the full HPCF being made up of a number of these building blocks, complemented by general purpose nodes, storage systems such as parallel distributed filesystems, I/O gateways and back-end management/administration nodes to resiliently provide the required functionality and aggregate performance.

To cater for the diverse workloads that are anticipated, ECMWF expects the next HPCF to feature a range of different node types. The parallel workload will include applications that will only effectively run on traditional CPUs, but also others such as Machine Learning which have already been enabled and optimised for fully accelerated environments (e.g., GPUs). Furthermore, it is expected that a great number of applications will be able to benefit from running in hybrid configurations using both CPU and GPU resources so that having all the nodes in a building block tightly coupled via a high-performance interconnect with performant, well-integrated access to reliable parallel filesystems is of uttermost importance.

In terms of CPU-time consumed, ECMWF's workload is comprised of predominantly parallel multi-node jobs, but there are also a vast number of single-CPU or partial-node-sized jobs that support the parallel workload by doing batch pre- and post-processing work and I/O-intensive tasks or are assigned to interactive tasks such as development and visualisation. The final configuration of the overall HPC service should be able to cater for this workload, which may also benefit from methodologies closely derived from cloud-orientated technologies such as containerisation and virtualisation.

It is envisaged that all types of such computational building blocks will share access to the high throughput parallel storage pools of the HPCF service at commensurate performance levels.

2.1. Scope

While this RFI includes some background information on ECMWF, its data centre environment and application requirements as in an ITT, the purpose of this RFI is much narrower. The desired outcome is simply to establish indicative proposals for configurations, expected power consumption and outline costs for a scalable block of compute performance of a given hardware type, both what is available currently on the market in 2024 and more importantly what will become available in the 2026-2027 timeframe. The types of hardware architectures that are of primary interest to ECMWF are CPUs, and GPUs although the inclusion of other types of hardware accelerators such as those optimised for AI workloads is also welcome. Separate benchmarks are provided to represent the performance of CPU-only codes, and accelerator-enabled codes (e.g., GPU). Section 0 contains detailed information on the benchmarks.

As well as for compute building blocks we are seeking information on *storage building blocks* that will be used by key workflows. Two types of parallel-accessible storage blocks are of interest, one optimised for high IOPS, and the other with focus on storage capacity, although both aspects might also be addressed by different tiers in a single storage building block design; more details can be found in section 2.5.

2.2. Out-of-Scope

The RFI is not intended to:

- Request a complete sizing and detailed qualitative description of a fully specified HPCF that can meet ECMWF's envisaged application needs;
- Consider the needs of ECMWF's serial or single-node workloads
- Consider aspects of a fully resilient configuration;
- Consider the data centre or data centre infrastructure that is needed to host such HPCF.

2.3. Contract length and timing.

ECMWF has typically procured the HPCF under a service lease for four or five years with the service hosted on-site. Any mid-life upgrades that are required are procured under terms negotiated as part of the contract, so that their value can be assessed as part of the total cost of ownership.

ECMWF would like to understand any alternative scenarios to this approach such as a hosted or cloud-based service, or a hybrid solution combining off-site and on-premises services. In the cases of using off-premises systems either wholly or partially, additional consideration would be required for access to ECMWF's data, both the working data held online in the HPC service, as well as access to the existing archive currently hosted in ECMWF's Bologna data centre, whose data is frequently accessed by the HPCF, and which is expected to hold several exabytes of data by 2027.

A longer contract period could be considered as well, but of key concern would be understanding the benefits of this and how we could demonstrate value when compared to the hitherto standard contract lengths and resulting frequencies of competitive procurements cycles.

2.4. Compute and Memory

For simplicity this RFI is concentrating on understanding the potential performance from two code types, a CPU-only code and two accelerator-enabled codes. For the latter, one of the codes is based on the PyTorch machine learning framework and can thus in theory support a variety of accelerators such as GPUs, TPUs, IPUs etc., while the other can only target GPU-accelerated hardware.

It is expected that the eventual HPCF will run a full range of jobs including many hybrid jobs that benefit from GPU-acceleration while still needing substantial resources from traditional CPUs. It is therefore highly desirable for the proposed configuration of a building block that CPU-only nodes and accelerated nodes are tightly coupled via a high performance interconnect, making it possible to execute efficiently a large communication-intensive MPI program across all or any subset of these nodes.

As well as traditional CPU or GPU-accelerated nodes, or those with CPU and GPU cores, nodes with dedicated AI accelerators would also be of interest if they deliver competitive price/performance.

2.5. Storage

For simplicity, in this RFI, requirements for the integrated parallel storage are only considered at a coarse level. However, ECMWF's HPCF is an exceedingly heavy producer and consumer of data on such storage, and this will increase due to:

- increased use of data driven machine learning;
- increase in forecast model resolution (although, for resolution upgrade, the storage requirements generally grow slower than related computational requirements);
- increase in the number of members in ensemble forecast systems.

ECMWF is keen to see the I/O requirements for each of its dominant workflows implemented with the most cost-efficient storage technologies, as appropriate for the respective use cases. At a high level, in the current HPCF this has been realised via a combination of independent Lustre filesystems with

- OSTs based on solid state technology, to provide resilient responsiveness and performance for highly concurrent I/O-requests from a wide range of sizes, for time-to-solution focused operational forecast suites;

and a number of separate filesystems with

- OSTs based on hard disk technology, for throughput research production, generated by usually many concurrent but independent

experiments and campaigns, with focus on filesystem capacity and more tolerance for higher performance variability in a less strictly-controllable environment.

Key requirements for the storage system will be:

- High-performance parallel storage capacity must be accessible from all nodes in every building block in.
- Resiliency and concurrent serviceability; in particular, there must not be any single points of failure in the storage system.
- A significant part of the storage resources will need to support POSIX-like filesystem semantics, including presenting a consistent global namespace to all nodes at all times, and support consistency for multi-node-concurrent read- and write-access to a single shared file. Respondents are encouraged, however, to include descriptions and performance characterisations of alternative access models to global storage pools, e.g., object storage interfaces.

Respondents are also invited to describe any potential options for providing different implicit or explicit performance tiers within a storage pool and namespace, such as “hot data” and “cold data” tiers, in particular when data placement management is implemented or can at least be assisted by corresponding systems software mechanisms.

To meet the resilience requirements of operational workloads, ECMWF usually has several independent high performance parallel storage pools configured in the HPCF. Furthermore, storage resources supporting time-critical workflows are physically separate from storage resources supporting research workload.

In ECMWF’s current HPCF Lustre filesystems, an OST storage building block, for example, can be seen in a highly-available controller pair with connected assemblies of dual-ported storage devices such has magnetic hard drives or enterprise SSDs, with the following order-of-magnitude characteristics:

OST media in building block	IOR-bandwidth [GB/s]	Net usable capacity [TB] at 80% filesystem utilisation
HDD	~20	~800
SDD (3 drive writes per day over 5 years)	~20	~60

For an overall indication of the order-of-magnitude target size of the storage part for the future HPCF, ECMWF expects this to be some multiple of those of the current HPCF, for which all of its globally accessible parallel storage has been configured as set of ten independent Lustre filesystems:

Main workflows	OST type	Number of filesystems	Building blocks per filesystem	Resulting IOR-BW per filesystem [GB/s]	Resulting net usable capacity at 80% utilisation [PB] per filesystem
Time-critical	SSD	2	25	> 460	>1.45
Time-critical	HDD	2	14	>260	>11.3
Various	HDD	2	6	>110	>4.85
Research/ Member State users	HDD	4	14	>260	>11.3

This presents a grand total of some 3PB on SSDs at ~1TB/s(IOR), and ~80PB on HDDs at 1.7TB/s(IOR). Each of the ten filesystem supports more than 240,000/s creates of 1-byte files.

Respondents are requested to describe some example configurations of storage building blocks and underlying technologies suitable for scaling the indicated order of magnitude for the timeframe of the future HPCF deployment, including indicative costing for such building blocks.

2.6. Support

ECMWF runs its HPC service as a 24-hour, non-stop operation with time-critical operational forecast suites executed several times a day. The input data cut-offs for the principal suites and their configurations are designed to operate at a safety buffer of no more than 30 minutes between the generation of products and their targeted time for dissemination to customers whenever a cycle of such suite executes under normal conditions, i.e., without exceptional systems or application-sided issues.

To support such nearly continuous operational dependency on a certain minimum viable configuration of the HPC service, in addition to redundancies designed into the configuration and its operational use patterns, the HPCF installation itself must be supported via a robust system software support service which provides 24 hours a day, 7 days a week call-out with a commitment to commence competent diagnostics, mitigation and/or remediation, within one hour after initial call-out. This means the support personnel that respond to the call-out within the time allowed are expected to be fully skilled and be closely familiar with the system configuration at ECMWF in order to be able to deal with the issues efficiently and

without additional delays - a simple call-logging service or general level-one helpdesk service will not be sufficient.

It should be possible that urgent hardware issues can be dealt with on the same time scale as software problems, but for failures that do not materially degrade the overall service because of sufficient redundancy, remedial maintenance can be performed during normal working hours. Any resilience on failover and redundancy features will be thoroughly tested as part of an acceptance procedure.

2.7. Financial model

ECMWF has procured previous HPCFs by means of a service contract that includes the provision/removal of hardware together with other items such as software licences and upgrades, hardware and software maintenance, support, training and migration assistance. Previous HPCF contracts have been for an initial four years of operational service (plus a period for set-up, installation and acceptance tests), at times with a mid-term performance upgrade, and the possibility to extend the contract at the end of the initial term.

ECMWF's HPCF contracts are typically based upon a series of equal annual payments, which mirrors the way in which ECMWF's own budgets are agreed and allocated. However, depending upon the specific elements of the successful tenderer's proposal in response to any future ITT, ECMWF may be able to work with the tenderer to develop a series of agreed key targets and milestones for the implementation and acceptance phase to develop a payment schedule which is more closely aligned to the major success factors of the contract and to the cost profile of the tenderer.

ECMWF's preference is for the HPCF to be made available as a service, nevertheless, ECMWF is willing to consider other options where there is a clear advantage to ECMWF in terms of cost and/or performance.

To this end, ECMWF is keen to explore options which:

- reduce the overall financing costs of the contract and thus allow a greater element of the overall contract price to be allocated to the provision of equipment or services;
- optimise depreciation and funding costs whilst maximising technological enhancements during the term, e.g., by varying the length of the agreement;
- more closely align the interests of ECMWF and the tenderer in ensuring the successful delivery of the agreed performance and service levels.

3. Technical Questions

3.1. General

Q1. Respondents are asked to provide a description of a building block or blocks to meet the requirements in section 2. The description of the hardware should include:

- The number of nodes in the building block;
- For each type of compute node pool, a description of the nodes including processor, accelerator and memory technologies;
- The high-performance interconnect. Details should include:
    i. achievable MPI latency and bandwidth;
    ii. topology, routing characteristics and hop counts;
- The maximum expansion capacity (e.g. network switch/topology) within a building block;
- The maximum number of building blocks that can be connected via a single high-performance fabric

Q2. If the solution could be installed in the ECMWF data centre, the description should include for each type of building block:

- Indicative layout drawings, preferably on a 600mm grid.
- Power and cooling requirements for a LINPACK run and additionally for the benchmark workloads

Q3. If the solution is not to be installed in the ECMWF data centre, then the computational building block performance should be fully allocated to ECMWF on a 24x7 basis and the storage capacity and related IOPS should be permanently assigned. The description should include:

- How ECMWF would access and use the building blocks, including read and write access to ECMWF's archive located at its Bologna data centre
- Pricing for the provision of the service, assuming that access to the ECMWF archive for the full service will require a network bandwidth of at least 50GB/s and that four petabytes of data will be read from the archive and another four written every day.

Q4. If the proposed solution uses nodes that have both CPU and GPU cores in the same socket, respondents are asked to explain how the building block could be efficiently used by codes that use only one of the core types. For example, how could the building block be used to efficiently run a mixed workload consisting of CPU only data assimilation codes, and GPU based AI training codes.

Q5. Please describe the expected timeline for the availability of this solution and describe the major risks associated with delivery and performance using Table 1

Table 1: Risk register template

Risk Name	Description	Probability (high/medium/ low)	Impact (high/medium /low)	Mitigations
<hw_feature_1>	<hw_feature_1> is not fit for service by <date>			<respondent> would….
<sw_feature_2>				

Q6. What could be potential benefits to the solution of changing the procurement timeline so that the system installation date is deferred, e.g. by three, six or twelve months or installing in separate tranches over a period of up to one year?

3.2. Compute and Memory

Q7. It must be possible to execute efficiently a single MPI communication intensive program using both CPUs and GPU cores. Describe how the proposed interconnect topology can meet this requirement.

Q8. Indicate how memory configurations impact on node costs and power and cooling requirements. What role is expected for composable memory system architecture?

3.3. Storage

Q9. Give an overview of storage technologies (both hardware and software) expected to be available for installation in the relevant time frame, indicating for each storage building block:

- power consumption and floor space requirements;
- potential options for systems software supported performance tiering within a building block;
- any modes for accessing globally shared storage resources other than as globally accessible filesystems with POSIX-like APIs and semantics;
- how the storage is made globally accessible by the various types of compute resources.

3.4. Hosted services

If you are proposing a service that is not hosted in the ECMWF data centre in Bologna.

Q10. Do you have the ability to host the service in one of ECMWF's member states? What restrictions are there in the hosting, e.g. system size, network bandwidth or level of support. What are the relative advantages or disadvantages of different locations.

- ECMWF Member States are: Austria, Belgium, Croatia, Denmark, Estonia, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Serbia, Slovenia, Spain, Sweden, Switzerland, Türkiye and the United Kingdom.

Q11. What measures do you have to monitor and improve the sustainable operation of the hosted service?

Q12. The ECMWF data centre has been designed with our 24x7 operational requirements in mind and for instance provides resilient cooling and power backed by diesel rotary uninterruptable power supplies cable of supporting the entire service. Bearing in mind the requirements described in 2.6, what features can your hosted service provide that will help ECMWF maintain our time critical operations?

3.5. Support

Q13. Respondents are asked to describe how they could meet the requirements for support described in section 2.6.

Q14. Respondents are asked to estimate the cost of providing the support described in section 2.6.

Q15. How would the software and hardware support costs vary if the number of building blocks were increased?

Q16. Respondents are asked to provide evidence of experience in deploying equivalent services – both in terms of size of building blocks as well as on-going support for environments with time-critical operational requirements

3.6. Physical environment

The high-performance computing facility and its associated cooling requirements account for about 95% of ECMWF's energy consumption. ECMWF strives to be environmentally responsible and optimise energy efficiency.

ECMWF has a centralised chilled water system configured to provide cooling capacity to both data halls in its Bologna data centre via independent secondary chilled water circuits. Additional air cooling maintains the data halls at a nominal temperature of 27˚C and 40%-60% relative humidity. The direction of airflow is from floor to ceiling.

Q17. What cooling options are expected for the proposed system? What are the advantages and disadvantages of each proposed solution?

Q18. Details should include the split between air and water-cooling, inlet and differential temperatures, flow rates and volumes for both air and water-cooling systems.

Q19. If a cooling solution other than air or chilled water is available, e.g., immersion cooling, please provide details and indicative costs related to each building block.

Q20. If appropriate, please indicate any requirements you have for the cooling systems, e.g., quality of water, temperature range, flow rate.

Q21. Please highlight any restrictions on the layout of the system or the distance between connected components.

Q22. Please indicate the size and weight of a full rack of each equipment type (e.g. storage, network, or compute) used in the system.

3.7. Sustainability

Environmental, Social and Governance (ESG) policies are evolving and becoming increasingly important across ECMWF and our member states. Understanding the efficiency and sustainability of our future HPC and what can be done to improve our performance will be a strong element in future procurements.

Q23. What measures are you taking to improve the sustainability of the manufacturing and operations of your systems or services? Can you describe your Net-Zero strategy?

Q24. Can you provide any information on the carbon footprint incurred in the manufacture of the proposed building blocks? Would you be able to measure or estimate the carbon footprint for the installation of a future HPCF service, i.e. including the manufacture and installation of the HPCF itself and any associated infrastructure?

Q25. What options do you have to measure and help improve the sustainability of the system in operation?

Q26. Please provide details of any energy-optimised runtime environments at a software level (tools/BIOS configurations) and evidence of how these reduce the electricity consumption for analogous numerical weather forecasting applications or codes

3.8. Upgrades

To meet its strategic needs ECMWF's requirements will continue to evolve over the contract period. Previous HPCF contracts have been for an initial four years of operational service (plus a period for set up, installation and acceptance tests) with a contracted options for mid-term performance upgrades and the possibility to extend the contract at the end of the initial term.

Q27. For the coming procurement ECMWF expects from the growth of AI/ML and CPU/GPU hybrid computing increasing interest in potentially rebalancing the system during the service life.

Please describe any options and limitations to upgrade a building block in the lifetime of the system.

Note that for a service hosted at the ECMWF data centre a key concern is that generally the power and cooling available to the service is fixed.

- What are the performance improvements that could be expected?
- What are the risks for delivery and performance?

- If the upgrade results in the replacement of nodes, what recognition could one expect to reflect the residual value of the hardware?

4. Financial Questions

Q28. Please complete the "RFI365 cost estimate tables" spreadsheet to indicate the cost of the building blocks and which financial models you would consider in a future ITT.

Q29. If you would consider contracts longer than four years, please state how many years and explain how you would continue to meet ECMWF's requirements for performance upgrades and value for money during this period. ECMWF recognises that due to the size of the initial procurement special pricing may have been obtained from component suppliers that would not be available for later additions and that prises for new services, generations of hardware or new node types cannot be reliably set in advance. What mechanisms could you offer to demonstrate continued good value for money and competitive prices during the life of the contract?

Q30. When considering the "ECMWF purchases system" option (see spreadsheet) please describe the financial benefits that this could bring to ECMWF and describe any buy-back options for equipment that is replaced in an upgrade or removed at the end of the term.

Q31. If there are any other financing options that you would consider please give details including how these options could/would assist in reducing the overall financial costs as mentioned in the first bullet point above, together with the impact of these alternatives on contractual arrangements, residual value risks and other costs (e.g. removal of hardware at contract termination).

Q32. Please state which of the options you believe is preferable and explain the benefits of this option to both parties.

Q33. Please describe how you could see yourselves being incentivised to meet and deliver key targets and milestones, and what actions ECMWF could take to mitigate any risks in this area.

Q34. Whilst ECMWF has a guaranteed 'base' level of funding, it is likely that additional medium-term (3-7 years) funding streams may also become available after the start of the contract period. Please indicate how the contract and pricing model could be constructed to allow ECMWF to utilise these funds to increase its HPCF capacity and how ECMWF could demonstrate value for money in the absence of competitive procurement.

5. Benchmarks

The RFI benchmark is to be conducted by running the following benchmark packages:

- IFS RAPS version 21, which represents the traditional weather forecasting workloads that run operationally on ECMWF's HPC systems daily
- A machine learning (ML) training benchmark using ECMWF's recently released AIFS machine learning model, which represents the ML training workloads that currently consume the majority of cycles in ECMWF's nascent on-premise GPU partition and that are predicted to consume significantly more GPU resources in the future
- The stand-alone ECTrans benchmark that is useful for benchmarking parts of the IFS that are already GPU-enabled

All of the above benchmark packages including relevant documentation packs are available from the ECMWF's FTP site after accepting the "ECBench" license and after attending a virtual briefing meeting with the ECMWF benchmarking team where further instructions on how to build and run the benchmarks will be given. To request the license, please create a user account on https://www.ecmwf.int and then raise a support ticket with the title "RFI2024 ECBench license request <your company name>" by following the link below:

https://jira.ecmwf.int/plugins/servlet/desk/portal/3/create/45

ECMWF kindly asks respondents to also return tarballs containing files generated by the benchmark runs and any relevant source code changes made. These will help the ECMWF benchmarking team to gather useful information regarding the benchmarking environment used by the vendor and provide it with more detailed performance results to the ones submitted in the RFI request. Details on how to generate the tarballs and how to return them to ECMWF will be given at the benchmarking briefing session.

5.1. General considerations

Q35. Respondents are asked to describe the benchmark systems used by completing a copy of Tables 5 and 6. If more than one system is used for either IFS RAPS or AIFS ML training benchmark, please duplicate the appropriate table.

Model	
Processor	
Clock speed (GHz)	
Maximum double-precision/single-precision floating point operations per clock cycle per CPU (e.g., 1TF/2TF)	
Number of CPU sockets per node	
Number of NUMA regions per CPU socket	
Number of physical/logical cores per CPU socket (e.g., 64/128)	
Memory type, speed, size and bandwidth (STREAM) per node	
Cache and register sizes	
CPU-CPU interconnect characteristics	
Peak power under LINPACK load of full node (kW)	
Operating system	
Compilers and runtime libraries	

Table 2: Outline of benchmark system for CPU-only benchmarks

Model	
Processor	
Clock speed (GHz)	
Maximum double-precision/single-precision floating point operations per clock cycle per CPU (e.g., 1TF/2TF)	
Number of CPU sockets per node	
Number of NUMA regions per CPU socket	
Number of physical/logical cores per CPU socket (e.g., 64/128)	
Memory type, speed, size and bandwidth (STREAM) per node	
Cache and register sizes	
Type and model of accelerator	
Clock speed (GHz) of accelerator processing cores	
Number of processing cores per accelerator	
Number of accelerators per node	

Maximum double-precision/single-precision floating point operations per clock cycle per accelerator	
Memory type, speed, size and bandwidth (STREAM) per accelerator	
CPU-CPU, CPU-accelerator and accelerator-accelerator interconnect characteristics	
Peak power under LINPACK load of full node (kW)	
Operating system	
Compilers and runtime libraries	

Table 3: Outline of benchmark system for accelerator-enabled benchmarks

5.2. IFS RAPS benchmark

The IFS RAPS benchmark is split into two components: ENS (ensemble forecast) and EDA (ensemble data assimilation) and shall be run only on CPU-based computational nodes that are part of the building block.

The ENS forecast benchmark comprises 15-day forecasts at three different grid resolutions: Tco159 (~64km), Tco319 (~32km) and Tco639 (~18km). Benchmarks are to be run on a single computational node of the building block and with appropriate instrumentation to record power consumption metrics of the entire node during the benchmark runs. All runs are to be performed in single precision, ensemble configuration (member number 1), without field output, and with and without coupling to the NEMO version 4 ocean model. To that end, TCo159 and TCo319 benchmarks can be coupled with a one-degree eORCA1_Z75 (~110km) grid, and TCo639 to a quarter-of-degree eORCA025_Z75 (~30km) grid.

From our own experience, the TCo639 benchmark cannot be run on a single computational node without at least 512 GB of available system memory. Moreover, it is likely that 512 GB of memory is not sufficient for running TCo639 coupled to the NEMO version 4 eORCA025_Z75 grid and as such, running this particular benchmark configuration is optional.

The EDA data assimilation benchmark is based on a test-of-adjoint (TLADJ) at TCo79 (~125km) and TCo159 (~64km) resolutions, which first runs a 12-hour forecast followed by one round of tangent linear adjoint calculations. All calculations are done in 64-bit double precision and therefore require a double precision build of the IFS. All runs shall use a single computational node from the building block. The TLADJ time is recorded and multiplied by 40, this is a representative timing figure for the costly data assimilation procedures like IFS minimization (IFSMIN).

Detailed information how to build the IFS in single or in double precision, ocean coupled or uncoupled, and how to run all tests described in this RFI will be available

in the documentation pack that accompanies the IFS RAPS tarball and which will be available to download from ECMWF's FTP server.

Q36. The vendor is asked to run at least three different runs of each test in the benchmark package and report the numbers in the corresponding table.

Table 7 provides a summary format for reporting data from the benchmark runs. In the cases where more than one type of CPU-based computational node is used as experimental platform, please replicate the table as often as necessary.

Test	Actual wall clock (seconds)	Measured Power Consumption (kWh)	Predicted wall clock (seconds) for 2026-2027	Predicted Power Consumption (kWh) for 2026-2027
ENS TCo159				
ENS TCo159 + eORCA1				
ENS TCo319				
ENS TCo319 + eORCA1				
ENS TCo639				
ENS TCo639 + eORCA025				
TLADJ TCo79				
TLADJ TCo159				

Table 4: Timings and power consumption measurements for the IFS RAPS tests

5.2.1. Optional higher resolution runs

ECMWF's current operational ENS system is based on 51 members at TCo1279 (~9km) resolution coupled with a NEMO version 3.4 ORCA025 (~30km) grid. We would like to understand how many CPU-nodes from your building block are required to run a TCo1279 benchmark coupled with a NEMO version 4 eORCA025 grid in order to make it run in under 3600 seconds.

Q37. Respondents are strongly encouraged to run a 15-day ENS forecast (member=1) at TCo1279 coupled with the eORCA025 NEMO version 4 grid and report the number of nodes needed to obtain a wallclock time under 3600 seconds as well as details on the configuration used such as number of MPI tasks, number of OpenMP threads and relevant environment tunables. Alternatively, if a wallclock time under 3600 seconds is not achievable using the number of nodes available in the building block, respondents should return the achieved wall clock time and the number of nodes used as well as relevant configuration details.

Q38. ECMWF would like to estimate the cost of upgrading its ENS system in the future from the current TCo1279 (~9km) resolution that is coupled to a NEMO version 3.4 ORCA025 (~30km) grid to a TCo2559 (~4km) configuration coupled to a NEMO V4 eORCA12 (~8km) grid, and, for this, understand how many nodes of the type available in the building block are needed to make it run in under 3600 seconds.

Q39. Respondents are requested to run a 15-day forecast at TCo2559 coupled with the eORCA12 NEMO version 4 grid and report the number of nodes needed to obtain a wall clock time under 3600 seconds as well as details on the configuration used such as number of MPI tasks and number of OpenMP threads and other relevant environment tunables. Alternatively, if a wall clock time under 3600 seconds is not achievable using the number of nodes available in the building block, respondents should return the achieved wall clock time and the number of nodes used as well as relevant configuration details if at all possible, bearing in mind that we require at least 64 compute nodes (with 256GB of main memory) on our current in-house HPC systems to successfully run this configuration.

5.3. AIFS ML training benchmark

ECMWF has been exploring the use of machine learning for improving the skill of its operational weather forecasts for some time now. A culmination of this work has been the release of the AIFS machine learning model. AIFS acts as a companion to the traditional physics-based IFS model where the "I" in AIFS stands for both (artificial) “Intelligence” as well as “Integrated”. The AIFS is an end-to-end machine learning system, taking an atmospheric initial condition, and providing a forecast for the next 10 days. More information about the AIFS can be found on ECMWF's website.

Training the AIFS model consumes by far the largest amount of resources within the ML-based workflows, as the model in inference mode can produce a 10 day forecast with 6 hour time steps in less than 2 minutes using a single NVIDIA A100 GPU. In comparison, the latest AIFS training campaign ran continuously on sixteen accelerated nodes, each with four NVIDIA A100 GPUs (64 GPUs total), for approximately one week. Since training requirements of future campaigns are projected to grow (with increasing spatial resolution and model complexity), we want to assess the performance of AIFS training on current computational nodes featuring accelerators such as GPUs. More information on the requirements for running the AIFS training benchmark is included in the documentation pack that accompanies the sources. However, since AIFS is based on the PyTorch machine learning framework, it should be in theory amenable to run on a wide variety of accelerators.

The training benchmark is based on an N320 resolution model (~30km resolution) and runs for 3 epochs each with 1000 iterations (with 100 additional validation iterations).

Q40. Respondents are requested to carry out at least three different runs of each test presented in the table below and report the relevant throughput and power consumption numbers. In the event that more than one type of accelerated computational node is used, replicate the table as appropriate.

Test	Average training throughput	Measured Power Consumption (kWh)	Predicted average training throughput 2026-2027	Predicted Power Consumption (kWh) 2026-2027
Single accelerator				
Half of available accelerators, data parallel only				
Half of available accelerators, data and model parallel				
Half of available accelerators, model parallel only				
All of available accelerators, data parallel only				
All of available accelerators, data and model parallel				
All of available accelerators, model parallel only				

Table 5: Throughput and power consumption measurements for the AIFS tests

5.4. Further GPU-enabled benchmarks

The IFS is undergoing significant changes to enable it to run on GPU accelerators under the internal Hybrid2024 programme. The goal is to have the majority of the model ported by middle of 2025 (except for the NEMO ocean model) with the help of source-to-source translation tools, directive based programming models such as OpenACC and OpenMP and, in some limited cases and scope, low-level programming models such as CUDA/HIP etc. One component of the IFS that has been ported to GPU accelerators is the spectral transforms library, which accounts for 20-30% of the overall runtime of a typical IFS run. The library has been open-sourced recently and can be run as a standalone benchmark that mimics the way the spectral transforms are called in the main IFS time-step.

More information on how to obtain and run the ECTrans benchmark set will be available alongside the IFS and AIFS documentation packs.

Q41. Respondents are asked to run the ECTrans standalone benchmark on the same GPU-accelerated node that was used for the AIFS training benchmark and report the times of the configurations presented in the Table below.

Test	Measured average time (s) per inverse-direct transform	Measured Power Consumption (kWh)	Predicted average time (s) per inverse-direct transform (2026-2027)	Predicted Power Consumption (kWh) for 2026-2027
TCo79				
TCo159				
TCo319				
TCo639				

Table 6: Wallclock time of running 100 direct and inverse transforms using the ECTrans benchmark at different resolutions.

5.4.1. Optional multi-node runs

Further multi-node runs using both ECTrans compiled for GPU-accelerated nodes and a version compiled for CPU-only nodes would be very useful for ECMWF to understand the scalability and performance of the interconnect used in the building block. As this test requires multiple nodes to be used for running the benchmark, we consider it to be optional (similar to the benchmarks in Section 5.2.1).

Q42. Respondents are asked to perform a weak scaling study of the ECTrans benchmark on both the CPU-only nodes and GPU-accelerated nodes of their respective building blocks starting from a TCo319 configuration on one node. The number of nodes should then be increased by a factor of four as the grid resolution is doubled (e.g., TCo639 on four nodes, TCo1279 on sixteen nodes) until the number of nodes available in the building block is exhausted. The vendor shall report the measured average time in seconds per inverse-direct transform over 100 iterations of each run.

Q43. Respondents are asked to conduct a strong scaling study of the ECTrans benchmark on both the CPU-only nodes and GPU-accelerated nodes of their respective building blocks starting from a TCo639 configuration on one node and then increasing in powers of two the number of nodes until all available nodes in the building block are exhausted. The vendor shall report the measured average time in seconds per inverse-direct transform over 100 iterations of each run.