REQUEST FOR INFORMATION

FOR A HIGH PERFORMANCE COMPUTING FACILITY (HPCF) FOR ECMWF

ECMWF/RFI/2017/001

June 2017

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. Background and Scope

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.

Specifically we seek to do the following:

- Identify technologies available in the 2019-2020 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 under an agreement that will expire at the end of September 2020. The final configuration delivered under this agreement comprises two Cray XC40 systems each with 3,600 36-core application nodes. For more information, see:

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

Although ECMWF is based in Reading UK it is unlikely that the next HPCF will be at Reading. ECMWF's Council has identified as a preferred solution a data centre in Italy, however a final decision on this has not yet been taken. It is expected that such a decision could be available before the end of June. However, the contract for the provision of a new data centre has not yet been signed.

The replacement HPCF is currently expected to be installed at the end of 2019 or the beginning of 2020 so that it can be commissioned and the entire workload migrated to it before the existing contract ends at the end of September 2020.

In July 2016, ECMWF published a new strategy for the period 2016-2025. For more information, see:

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

The key goals for computing at ECMWF are:

- to deliver appropriate levels of computational resources to satisfy the requirements of the research, operational and Member State communities,

- to deliver efficient, effective and resilient computational resources, applying industry standard best practice where appropriate,
- to be environmentally responsible and to seek ways to minimise the environmental impact of running such a HPCF.

To further improve our ability to predict high-impact weather, we aim to run a high-resolution ensemble system up to two weeks ahead. An ambitious target that depends on scientific, computing and scalability advances is for this ensemble to have a horizontal resolution of about 5 km by 2025. This represents a much larger numerical and computational task than today by many orders of magnitude and this will be combined with the expected huge increase in observational data volumes.

For this RFI, ECMWF is interested in the specification, configuration and cost of a system that could be used as a self-sufficient building block in a HPCF that comprises two or more of these building blocks to provide resilience and the required aggregate performance.

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 34 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. 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 or contract between ECMWF and any other. 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.

1.3.2. Timetable

This RFI will close at 14:00 UK local time on Friday 4 August 2017

ECMWF envisages the following schedule for the implementation of the project:

19 June 2017	Issue of this RFI
19 June – 21 July 2017	Discussions between ECMWF and vendors to clarify the RFI specification
21 July 2017 at 16:00	Last date/time for submission of clarification questions for this RFI
4 August 2017	Close of RFI
August – October 2017	Development of procurement strategy and review by ECMWF Technical committees
December 2017	Decision by ECMWF Council on procurement of next HPCF
Second half of 2018	Issue of Invitation to Tender for new HPCF
Beginning of 2019	Receipt of tenders
First half of 2019	Evaluation of tenders and negotiation of contract terms
	Selection of the winning tender
Mid 2019	Submission of the contract to ECMWF's Council for approval, followed by signature of the contract.
End of 2019/Beginning of 2020	Start of installation of new HPCF
30 September 2020	End of current HPCF contract

1.3.3. Confidentiality

All information, materials, specifications or other documents 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. In particular respondents shall not publish or make generally available the results of running ECMWF benchmarks on their system(s). ECMWF in turn confirms that it shall treat all information provided to it by the respondent as confidential and further confirms that such information will not be disclosed by ECMWF to any third parties, other than its advisers and consultants.

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 hpc2017@lists.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 hpc2017@lists.ecmwf.int at ECMWF and must be made in writing by email, no later than 16:00 UK local time on 21 July 2017.

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. A Word file containing the questions is available on ECMWF's website for your use. 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 hpc2017@lists.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/2017/001 for a High Performance Computing Facility for ECMWF

2. Requirements

The requirements in this section provide guidance on the specification for a self-sufficient building block of an eventual High Performance Computing Facility (HPCF) that would meet ECMWF's requirements for installation in late 2019/early 2020. ECMWF envisages the full HPCF being made up of two or more of these building blocks to provide resilience and the required aggregate performance. The requested performance of the building block is the performance of ECMWF's current two-cluster HPCF.

ECMWF's workload is predominantly parallel jobs but some serial or single-node workload is run on the system. A building block should consist of one or more pools containing compute nodes to meet the needs of parallel applications, and a number of compute nodes focused on high single-thread performance to support serial or sub-node-sized parallel jobs. Furthermore there should be well-integrated storage systems efficiently accessible in parallel with high throughput by the various node types in a building block. Special nodes to provide such services as system management, batch scheduling, network and file-system access and any other equipment/functions required to enable the building block to operate independently should also be considered as part of the building block.

It is envisaged that building blocks will share their high throughput storage with other building blocks of the HPCF at equal performance levels.

2.1. Parallel performance requirements

In terms of consumed processing capacity, the bulk of ECMWF's HPCF workload is in parallel applications. The largest of these is ECMWF's own Integrated Forecasting System (IFS) that uses a hybrid MPI-OpenMP programming paradigm.

The performance of the building block is defined by running a representative workload in 3600 seconds or less on the technology to be installed in 2019/2020. This can be extrapolated from the results of running the benchmarks on current technologies. The components of the "packet" of workload are benchmark versions of the three most computationally demanding components of ECMWF's current operational IFS configuration, namely Ensemble Data Assimilation (EDA), the High-Resolution Forecast (HRES) and the ensemble of lower-resolution forecasts (ENS), which are part of the IFS RAPS16 benchmark. The benchmark resolutions for the test are the current resolutions for the operational forecasting system and the number of copies represent the balance of work currently seen. The configuration of a packet is:

- 8 copies of HRES - TCo1279L137 double precision forecast
- 90 copies of ENS - TCo639L137 double precision forecast
- 42 copies of EDA - TL399L137 test of adjoint

Details of how to run these tests can be found in section 5.

It is highly desirable that all parallel application nodes in the building block are tightly coupled within a single cluster via a high performance interconnect, making it possible to execute efficiently a single MPI communication intensive program across all or any subset of these nodes. If the interconnect topology organises parallel application nodes into groups, then at a minimum the grouping must still allow efficient scheduling and execution of concurrent jobs of the size of the TCo1999 forecast model (at 5km) also provided in the IFS RAPS16 benchmark.

Table 1 - Summary of parallel performance requirements

Requirement	Goal	Goal, expressed in terms of ECMWF's current Intel E5-2695v4, 128 GiB nodes
Aggregate Memory	1 PiB	8,192 nodes
Largest routine job size	IFS TCo1999 benchmark to run in one hour	1,600 nodes
Building block size	Sufficient to run concurrently, in one hour or less: • 8 copies of HRES - TCo1279L137 double precision forecast • 90 copies of ENS - TCo639L137 double precision forecast • 42 copies of EDA - TL399L137 test of adjoint	• 6,232 nodes (224,352 cores) • 8 x 137 nodes (39,456 cores) • 90 x 37 nodes (119,880 cores) • 42 x 43 nodes (65,016 cores)

2.2. Single node performance requirements

Some of ECMWF's HPCF workload is related to the preparation of input data for the main applications and the post processing of the output data. Such jobs may:

- Use a single or low number of threads, including intra-node MPI,
- Be either I/O intensive or benefit from high single-thread execution performance,
- Not easily exploit vector or accelerator architectures,
- Require a general purpose Linux software environment.

In general, multiple unrelated jobs of this type may be executed concurrently on a single node.

It is highly desirable that these nodes use the same processor or at least processor architecture as the parallel performance nodes.

Though expressed separately for convenience, it is desirable that for flexibility this requirement be met by nodes from the parallel performance pool that can be easily repurposed.

Table 2 - Summary of single node performance requirements

Resource	Requirement, expressed in terms of ECMWF's current Intel E5-2695v4
Memory	2 GiB per core and at least 128GiB per node
Size of pool	Equivalent of 7,500 cores

2.3. Storage performance requirements

Over the last decade, growth in magnetic disk capacity has greatly outpaced growth in bandwidth and IOPS performance. For ECMWF, I/O performance requirements have become the main determinants for sizing storage configurations using only such drives. With NAND-based SSDs now becoming more affordable and novel non-volatile memory technologies surfacing, ECMWF is keen to see the I/O requirements for each of its dominant workflows supported with the most appropriate and cost-efficient storage technology.

A key workflow is the time-to-solution focused operational run of the forecast suite. The size of the storage working-set for a forecast suite (six of such cycles per day) is expected to be in the order of 150-200TB, and the suite's "front-end" performance requirement of some 300GB/s peak might best be implemented via a resilient configuration built around SSDs supported by a slower, higher capacity "back-end", rather than from a single-level parallel file-system that provides both the required front-end performance and back-end capacity.

Shared high performance parallel storage capacity must be accessible from all nodes in the building block.

The storage must be resilient to failures. In particular there must not be any single points of failure in the storage system.

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

Each storage pool must present a consistent global namespace to all nodes at all times and it is desirable that the following "POSIX single-node filesystem semantics" are supported:

- read and write operations from different nodes to a single file can be configured to be atomic and serialised in issuing order;
- POSIX advisory locks.

The building block must provide a general purpose storage pool, used for such things as user home and storing system wide libraries and utilities, this is small block, "N:N" random I/O. The storage must be accessible from all pools in the building block. It is desirable that it supports snapshot functionality and synchronous or asynchronous replication to a general purpose storage pool in another building block.

Table 3 - Summary of aggregate storage requirements, expressed as IOR and MDTEST benchmark performance

Requirement	Goal
Back-end storage bandwidth for time critical workflow	≥150GB/s
Front-end storage bandwidth for time critical workflow	≥350GB/s
Back-end capacity for time critical workflow	≥8 PB
Front-end capacity for time critical workflow	≥150TB
Front-end meta-data create operations per second for time critical workflow	≥200,000
Storage bandwidth for research workflow	≥450GB/s
Capacity for research workflow	≥15 PB
Meta-data create operations per second for research workflow	≥350,000
General purpose storage	≥100 TB

2.4. Support requirements

ECMWF runs its computer facilities as a 24-hour, non-stop operation and runs operational work several times a day to a tight schedule.

To support this non-stop operational requirement the building block can be supported via a service, which provides 24 hours a day, 7 days a week call-out whereby parts, if required, and an engineer can be on-site at the data centre, which is envisaged to be in Italy, within two hours of a request for assistance. Alternatively the building block can be designed with sufficient redundancy that

failures do not materially degrade the service and remedial maintenance can be performed during normal working hours.

In addition to the hardware maintenance service, ECMWF requires full-time system software support during normal working hours for the life of the system and on-call cover for 24 hours a day, 7 days a week with a two-hour response time. It is desirable that this be located at ECMWF headquarters in the UK during normal working hours. This should cover:

- configuration, testing and updating of the supplier's software and any supplied third-party software;
- acting as the point of contact for resolution of such problems;
- provision of effective procedures and monitoring tools for problems with the hardware and supplied software.

ECMWF requires full-time application software support during normal working hours for the life of the system. This service must be provided on-site at ECMWF's headquarters in the UK and it is highly desirable that a single dedicated specialist provides this. The support should assist users in ECMWF and its Member States with migration, optimisation and debugging of their applications and be the point of contact for issues with the vendor supplied programming and development environment.

ECMWF will require e-learning training facilities as well as on site courses to be available for scientists and technicians.

2.5. Software requirements

The main programming language for ECMWF's applications is FORTRAN 90, although some features of FORTRAN 95, 2003 and 2008 are used as well. Increasingly, both C and C++ are used for development of new supporting applications. C11, C++98 and C++11 standards are used. Python is used for building and maintaining the operational suites and for post-processing tasks.

IFS uses a hybrid MPI/OpenMP paradigm. This uses a few MPI-tasks within each node and each task comprises multiple OpenMP threads. It is important that the various libraries provided by the vendor are "thread-safe", the platform's MPI is assumed to support MPI_THREAD_MULTIPLE mode for point-to-point and collective operations. It is desirable that MPI-3 and OpenMP 4.0 features are available.

ECMWF is committed to maintaining the portability of its codes. ECMWF is therefore reluctant to use vendor specific language extensions or run-time libraries that require considerable porting efforts unless very significant performance gains can be achieved.

3. Technical Questions

3.1. General

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

- Indicative system layout drawings, preferably on a 600mm grid.
- The high-performance interconnect. Details should include:
    i. Achievable MPI latency and bandwidth;
    ii. topology, routing characteristics and hop counts;
- Connectivity and bandwidth parallel performance nodes and single-node-job performance nodes.
- for each compute node pool, a description of the nodes including processor and memory technologies
- function, number and type of any ancillary nodes
- Storage that could meet the I/O requirements described in 2.3.
- The connectivity and bandwidth between the I/O nodes and the pools of compute nodes

Q2. Please describe the timeline for the availability of this solution and describe the major risks associated with delivery and performance using Table 4

Table 4 - Risk register template

Risk Name	Description	Pre-mitigation probability (high/medium/low)	Pre-mitigation impact (high/medium/low)	Post-mitigation probability (high/medium/low)	Post-mitigation impact (high/medium/low)	Mitigations
<hw_feature_1>	<hw_feature_1> is not fit for service by <date>					<respondent> would….
<sw_feature_2>						

Q3. What are the consequences to the solution of changing the procurement timeline so that the system installation date is postponed, e.g. by three, six or twelve months or installing in separate tranches over a period of up to one year?

3.2. Parallel Performance

Q4. It must be possible to execute efficiently a single MPI communication intensive program of at least the largest routine job size given in 2.1, i.e. TCo1999. Describe how the proposed interconnect topology can meet this requirement.

Q5. If applicable, characterise the size of sets of nodes sharing common edge switches or similar, and give performance and cost estimates in case the interconnect's backbone bandwidth is configurable.

Q6. Would it be possible to have a subset of the nodes on the same interconnect supported by higher backbone capability?

Q7. Describe support for and characteristics of PGAS models. Any communications runtime must be resilient enough to reasonably support operational workflows, i.e., retransmit or abort affected jobs in case of any communication failures.

Q8. Indicate the possible memory configurations for nodes in this pool. Assuming all nodes in this pool have the same memory configuration, what impact on costs would this configuration have?

3.3. Single node performance

Q9. Describe how the nodes are dedicated to running this workload, e.g. separate sub-system, different operating system or batch scheduler configuration. If applicable, what flexibility and constraints are there to repurposing these nodes to be parallel compute nodes?

Q10. Indicate the possible memory configurations for nodes in this pool. Assuming all nodes in this pool have the same memory configuration, what impact on costs would this configuration have?

3.4. Storage

Q11. What percentage of the overall system cost does the storage represent?

Q12. Give an overview of storage technologies (both hardware and software) expected to be available for installation in the relevant time frame, indicating relative pricing. Describe each technology's performance characteristics, including:

- capacity,
- peak bandwidth,
- latencies and throughput for random I/O at various record sizes (from 4KiB via 128KiB to multiple MiB),
- performances' dependency on span, fragmentation, blocking, consistency semantics,
- metadata performance,
- endurance.

3.5. Support

Q13. Respondents are asked to describe how they could meet the requirements for support described in section 2.4.

Q14. Please identify any impact that the location of the Data Centre may have on your support model.

Q15. Respondents are asked to describe the training which is available to enable the efficient usage of the HPC

Q16. Respondents are asked to describe the support and consultancy which is available to port existing applications.

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.

Q17. Please provide details of machine power requirements and estimate the total power consumption in kW of the proposed building block.

Q18. What cooling options are available for the proposed system? What are the advantages and disadvantages of each proposed solution?

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 appropriate, please indicate any requirements you have for the quality of water used in the cooling system.

Q20. Please highlight any restrictions on the layout of the system or the distance between connected components.

Q21. 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. System costs

Q22. Please provide a cost estimate of the system building block described in Q1 using the "RFI001 cost estimate tables" spreadsheet.

Q23. How would the costs above vary for a second or subsequent building blocks or parts thereof?

Q24. How would the software and hardware support costs vary if the amount of hardware is increased from one building block?

3.8. Mid-term upgrade

To meet its strategic needs ECMWF's requirements will continue to evolve over the contract period. Consequently, an upgrade to the system after two or three years could allow a better match between future technologies and growing performance

requirements. 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 mid-term performance upgrade and the possibility to extend the contract at the end of the initial term.

Q25. When would be the best time for an upgrade you could propose for ECMWF to optimise the mixture of performance, ensuring access to future technologies and cost?

- What leads you to this conclusion?
- What form would any upgrade in this scenario take?
- What are the performance improvements that could be expected?
- What are the risks for delivery and performance?

Q26. For a unit of performance equal to the original building block, purchased at your proposed upgrade time, please describe, using spreadsheet "RFI001 cost estimate tables" for the costs:

- The purchase cost of the compute and storage components
- The electricity consumption
- The annual hardware and software support cost
- The factors that affect scaling of these costs, e.g. amount of equipment, type of equipment, number of building blocks.

4. Financial model

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 and 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 and may include some limited capacity for payments for achieving milestones.

ECMWF has procured previous HPCF 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) with a mid-term performance upgrade and the possibility to extend the contract at the end of the initial term.

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 which 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.

Whilst ECMWF has a guaranteed 'base' level of funding, it is likely that additional medium-term (3-7 years) funding streams may also be available, although these may not be confirmed by the time the responses to any future ITT are submitted. Please indicate how the contract and pricing model could be constructed to allow ECMWF to utilise these funds to increase its HPCF capacity at equivalent or even better terms to reflect the greater buying power that this additional funding could bring. For the purposes of responding to this question, these additional funds could represent an additional 10-20% of the base funding.

Q27. Please complete the "RFI001 cost estimate tables" spreadsheet to indicate which financial models you would consider in a future ITT. 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. If there are any other financial options that you would consider please give details including how these options could/would assist in reducing the overall financing 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).

Q28. If you would consider 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.

Q29. Please state which of the options you believe is preferable and explain the benefits of this option to both parties.

Q30. As ECMWF has income streams in both pounds sterling and euros, we would wish to understand whether the use of either of these currencies in terms of payment for the solution would have any consequences on your bid, and if so, how could these be mitigated?

Q31. As it is anticipated that the HPCF will not be sited in the UK, but potentially in Italy, would this have any financial consequences on your offer, and if so, how could these be mitigated? Does the location of the HPCF have any implications for the originating country of your contracting or invoicing process?

Q32. 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.

Q33. Please describe how ECMWF could optimise the benefits of any additional funding that may become available during the contract term.

Q34. ECMWF has previously contracted via a service agreement and insured the HPCF directly. Would there be benefits for the insurance of any system provided under a service agreement to be arranged by the provider and would you be willing to undertake this?

5. Benchmarks

The RFI benchmark is to be conducted by running ECMWF's latest benchmark set IFS RAPS16 released in June 2017. The benchmark package is available from ECMWF as a download from our FTP site after accepting the "ECbench" licence. To obtain the licence please contact the ECMWF Data Services team by email to dataservices@ecmwf.int.

5.1. General considerations

Q35. Respondents are asked to describe the benchmark system used by completing a copy of Table 5.

Table 5 - Outline of benchmark system

Model	
Processor	
Clock Speed (GHz)	
Maximum double-precision floating point operations per clock cycle	
Total number of nodes	
Number of CPU sockets per node	
Number of processor cores per socket	
Memory speed, size and bandwidth per node	
Cache and register sizes	
Interconnect characteristics	
Storage configuration	
Compilers and runtime libraries	

5.2. IFS work packet benchmark

As described in 2.1 the packet of IFS work comprises 8 copies of a TCo1279 (~8km) 10-day forecast, 90 copies of a TCo639 (~16km) 10-day forecast and 42 copies of a 12-hour TL399 (~50km) test of adjoint, which mimics data assimilation operations very closely. All these operate at double precision i.e. using 64-bit floating point arithmetic.

The two forecast components of a packet are run as 24-hour forecasts, requiring less than 500 seconds to run. No field output is requested. The estimated elapsed time for the 10-day forecast therefore has to be extrapolated. Furthermore, the TCo1279 benchmark is run without the WAve Model (WAM) subcomponent, requiring an additional multiplier of 1.1 to reflect the full cost of the operational configuration. Similarly the TCo639 benchmark component is run without WAM and NEMO ocean subcomponents, which means the estimated 10-day forecast time has to be multiplied by 1.5.

Section 5.2.1 gives details on how to estimate the runtime of a 10-day forecast figure based on a 24-hour run for both components.

The EDA benchmark component is the test of adjoint, which first runs a 12-hour forecast followed by one round of tangent linear and adjoint calculations i.e. TLADJ. The TLADJ time is recorded and multiplied by 35, this is a representative timing figure for the costly data assimilation procedures like IFS minimization (IFSMIN). Section 5.2.2 gives details on how to estimate the IFSMIN runtime from the 12-hour test-of-adjoint TLADJ-time.

Q36. The vendor is asked to run at least three different runs of each test in the benchmark package to calculate the number of nodes needed to get each of the extrapolated benchmark times to be 3600 seconds or less.

Table 6 and Table 7 provide summary formats for reporting data from the benchmark runs.

Table 6 - Summary of the individual IFS test results

Test	Number of Nodes	Number of Cores	Number of hardware threads per physical core	Number of MPI tasks	Number of OpenMP Threads per MPI-task
TCo1279					
TCo639					
TLADJ					

Table 7 - Timings for the IFS tests

Test	Actual wall clock (seconds)	Extrapolated runtime (seconds)
TCo1279		
TCo639		
TLADJ		

5.2.1. Procedure for estimating the 10-day forecast (FC) runtime from the 24-hour FC runtime

As described in sections 7.4 and 8.6 of the IFS RAPS16 benchmark document, the *jobinfo* script can be used to extrapolate from a 24 hour forecast run to a the runtime of a 10-day forecast. For a single IFS 24h forecast, you:

- Should set FCLEN=24h
- Should set the GENFACT environment variable to 1.1 for TCo1279 to account for the missing WAM component.
- Should set the GENFACT environment variable to 1.5 for TCo639 to account for the missing WAM and NEMO components
- **May** adjust NPROMA & NRPROMA
- **MUST NOT** adjust UTSTEP (time step length) or NRADFR (frequency of solar radiation steps)

Adjust the number of nodes (or MPI-tasks/OpenMP-threads) until the extrapolated elapsed runtime is just below the required 3600 seconds.

5.2.2. Procedure for estimating the EDA IFSMIN runtime from the 12-hour test-of-adjoint TLADJ

The *jobinfo* script can also provide an estimate of the time to complete the 12 hour IFS minimization. Sections 7.2.4 and 8.7 of the IFS RAPS16 benchmark document provide more details. For a single IFS 12h forecast followed by one iteration of test-of-adjoint, you:

- Should set resolution to TL399
- Should set FCLEN=12h
- **May** adjust NPROMA & NRPROMA
- **MUST NOT** adjust UTSTEP (time step length) nor NRADFR (frequency of solar radiation steps)

Adjust the number of nodes (or MPI-tasks/OpenMP-threads) until the extrapolated elapsed runtime is just below the required 3600 seconds.

5.3. Additional runs

Respondents are invited to report the results of further runs to assist ECMWF in the assessment of future performance:

5.3.1. Single precision

Q37. We ask the vendor to run similar and additional TCo1279 & TCo639 tests using IFS in a single precision mode with the node counts that were just below the 3600 seconds barrier described earlier.

We expect 1.4 to 1.6 times performance improvements against double precision runs.

5.3.2. 1-hourly output

Q38. Respondents are asked for additional runs of the TCo1279 and TCo639 forecast resolutions, with 1-hourly field output activated. Enabling I/O requires collection from all MPI ranks and encoding of field data as well as subsequent writing to the filesystem.

Two different sets of runs per resolution are requested:

1. using 2 extra nodes for I/O-server MPI-tasks (so called I/O quilting)
2. allowing one compute MPI-task per node to act as an output writer task with no extra MPI-tasks.

The output can be redirected to a special fast disk via the environment variable FDB_ROOT_DIRECTORY with the default being the normal run directory -- potentially a less performing disk. All I/O-tests are to be carried out in double precision only.

5.3.3. Higher resolution

We would like to estimate the cost of our future high resolution model TCo1999L137 (5km) both in double and single precision mode and understand how many nodes and cores are needed to make it run in under 3600 seconds.

To give an idea about the cost and resource requirements of TCo1999, here are some data points, all without field output that have been derived from experience on our current system:

- A minimum of 100 nodes and 128GiB per node is required to run TCo1999. Single precision may fit into 50-70 nodes.
- TCo1999L137 double precision version is 4 to 4.5 times more expensive than the corresponding 10-day run with TCo1279L137 (8km).
- TCo1999L137 single precision version is 1.6 times less expensive than its double precision version.
- In the order of 800 nodes are needed to complete in under 3600 seconds.
- Adding WAM and NEMO requires 1200 nodes to complete in under 3600 seconds.
- Enabling one-hourly field output, with I/O quilting, requires around 1,600 nodes to complete in under 3600 seconds.

Q39. Respondents are requested to run just the 24-hour forecast and use the provided extrapolation scripts to estimate the runtime for 10-day forecast. Field output, WAM and NEMO should not be activated.

5.3.4. Kronos workload simulator

ECMWF is a partner in the European Commission funded NEXTGenIO project, as part of this project ECMWF is developing a tool named Kronos. The aims of Kronos

are to generate, run, and benchmark simulated representations of I/O, MPI communications and computation of a real life workload but in an environment that is controlled and easily portable. At this stage, Kronos is still a prototype, however, it is expected that it will play an integral part in describing and measuring the real I/O and interconnect requirements in the next HPCF procurement.

Q40. Respondents are requested to run the Kronos benchmark, which represents ECMWF research workflows, and return the resulting output.