# INVITATION TO TENDER

# ECMWF/2025/380

# PROCUREMENT OF A HIGH PERFORMANCE COMPUTING FACILITY (HPCF)

## CLARIFICATIONS
## issued on

## 5 August 2025

**TRADEMARKS**

All names or descriptions used in this document which 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 document 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.

**Disclaimer**

The information contained herein refers to the benchmarking activities undertaken by prospective tenderers for ITT380 and is provided solely to assist prospective tenderers in this process.

ECMWF expressly reserves the right to change, modify, or revise some or all of the aspects of the benchmarking methodology and related requirements presented herein at any time prior to the release of ITT380.

ECMWF will not reimburse any expenses, costs, or charges incurred by prospective tenderers or any third parties in connection with: (a) the preparation for and undertaking of the benchmarking activities for ITT380; (b) the preparation and submission of responses to the ITT380; (c) any preliminary work undertaken based on this information; or (d) any other activities related to ITT380 process.

ECMWF accepts no liability whatsoever, whether in contract, tort or otherwise arising from or in connection with any preparatory activities for and the actual ITT380 process and documentation, including but not limited to (a) the benchmarking methodology or activities for ITT380; (b) any acts, decisions, or commitments made in reliance upon the information, guidance, or specifications contained in this document; (c) any costs, damages, losses or expenses (including consequential losses) incurred by the prospective tenderers or any third party in this context.

The information contained herein will be superseded by the documentation issued as part of the ITT380 procurement exercise, as required.

**Definitions**

In these clarifications the following words or phrases have the meaning ascribed to them:

| | |
|---|---|
| Cluster | A collection of compute nodes and their shared high-performance interconnect. A Cluster may comprise SIM-, GPIL- and ML-nodes, or just SIM- and GPIL-nodes without ML-nodes, or just ML-nodes. Note that the Clusters comprising the System are not expected to be identical. |
| Partition | All the nodes of the same type in a cluster, e.g. MLT-Partition, GPIL-Partition. |
| Nodeset | All the nodes of the same type in the Works. E.g. the MLM-Nodeset is all MLM-nodes in both MLM-Partitions. |
| general-purpose and interactive login (GPIL) workload; GPIL-Partition | This workload can consist of interpreted code such as Python or shell scripts as well as C, C++ or Fortran compiled code and has typically been developed and deployed under a Linux environment.<br>A key function of this workload is pre/post-processing and staging of data to and from ECMWF data handling system external to the HPC system, and as such it can be very filesystem-I/O and TCP-network-transfer intensive.<br>General-purpose or interactive workloads can be small scale parallel applications, using OpenMP and/or intra-node MPI, but often cannot exploit a vector architecture and do not require inter-node MPI communications.<br>This workload runs on general-purpose and interactive login nodes (GPIL-nodes). GPIL nodes on a common interconnect form a GPIL-Partition. The System shall have four GPIL Partitions separate from each other. |
| MLM-node; MLM-Partition | A GPU machine learning compute node that is intended to run a mix of GPU applications including time-critical and well as research-use inferencing, as well as small and medium-scale training. All MLM-nodes will be installed in the Bologna Data Center. The main data-driven modelling workload runs on such nodes. All MLM-nodes on the same high-performance interconnect supporting collective communication backends for PyTorch Distributed such as NCCL/RCCL or MPI form one of the System's two MLM- Partitions. |
| MLT-node; MLT-Partition | A GPU Machine Learning node primarily intended for large-scale training workloads. All MLT-nodes must share a common interconnect supporting collective communications backend for PyTorch Distributed such as NCCL/RCCL or MPI, forming the single MLT-Partition. This MLT-Partition may be installed in the Bologna Data Centre, or off-premises. If installed in the Bologna Data Centre, the MLT-Partition will |

| | |
|---|---|
| | also serve as a further backup option for time-critical inferencing usually carried out on MLM-Partition(s). |
| ML-node; ML-Partition | An MLM-node or MLT-node; an MLM- or MLT-partition |
| SIM-node parallel Application node; SIM-Partition | A compute node that is dedicated to running physics-based simulation applications such as IFS. It must support CPU intensive applications requiring OpenMP and inter-node MPI communications. SIM-nodes sharing a low latency interconnect form a SIM Partition. The System shall include four identically sized SIM-Partitions. |

We are pleased to provide the following clarification responses to questions received:

## C1_ITT380

**Question:**
What are the policies for accepting optimisations in the pre-ITT period?

**Answer:**
A list of permitted types of changes will be included with the benchmarks and is more generous than in previous ITTs of ECMWF. Early interactions are encouraged for review of suggested optimisations and acceptance by ECMWF. The main criterion for acceptance of modifications is that forecast scores do not degrade and that changes do not break bit repeatability. That is, consecutive runs with the same number of MPI tasks and OpenMP threads must produce bit identical results across all runs.

## C2_ITT380

**Question:**
Will support for RAPS be available until the ITT release?

**Answer:**
Yes, limited support will be provided via email. If required and if time permits, ad-hoc meetings for specific issues may be possible as well. However, while ECMWF will be able to advise on some specific issues, we do not have the resources for in-depth investigations.

Regular meetings with vendors are unlikely to be possible as they would need to be offered to all tenderers and would likely be too time consuming for all involved. All meetings will be recorded and minuted. If any issues of general relevance arise, they will be published as further clarifications.

All questions relating to the ITT380 benchmarks must be directed to itt380-benchmarks@groups.ecmwf.int.

## C3_ITT380

**Question:**
Will there be a cut-off date for any changes and optimisations to be integrated into the final ITT benchmark release?

**Answer:**
Yes, there will be a cut-off date after which code optimisations and changes meant to improve the performance of benchmarks will not be accepted. The cut-off date is two weeks prior to the release of the ITT. ECMWF's intention is to publish the ITT on the 15th of October 2025. If this remains to be the case, the cut-off date will be the 1st of October 2025.

The cut-off date will apply to both IFS and AIFS, and any other code controlled by ECMWF. After the cut-off no changes to any ECMWF code will be allowed except for those necessary to allow for the application to run on the new environment.

To be clear, no optimisations will be accepted after the cut-off date, and reported improvements from further optimisation will not be considered in the evaluation.

## C4_ITT380

**Question:**
Will there be further changes to the scripts and input data in subsequent RAPS releases until the final release for the ITT?

**Answer:**
Yes. As documented in the benchmarking methodology presentation slides, there will be further releases of the RAPS benchmarking package which will provide new features such as the data assimilation benchmark, increased GPU coverage of IFS components, optimisations from ECMWF and vendors as well as bug fixes where necessary. Changes between RAPS releases up to the final ITT release will be documented as Release Notes in the RAPS documentation. However, the underlying IFS cycle will not change nor do we plan to change at this time the architecture and model of the AIFS benchmarks.

## C5_ITT380

**Question:**
Will code changes to ECMWF code in non-ECMWF branches of packages be valid to use in the procurement?

**Answer:**
No, all changes classed as optimisations or meant to improve the performance of the benchmarks will have to be already present in the final ITT release of the RAPS benchmark package. This means that these changes would have been shared with ECMWF and approved before the aforementioned cut-off date **(C3_ITT380)**.

## C6_ITT380

**Question:**
Is help available to merge branches back into the ECMWF repositories? What happens with if there are conflicts between different vendor optimisations?

**Answer:**
ECMWF will offer support under the conditions already specified in **C2_ITT380** with merging vendor changes and optimisations into the main RAPS benchmarking repositories prior to the cut-off date. Vendor changes will have to be guarded by *ifdefs* and be implemented in such way that either the original code path or the optimised code path can be turned on at compile time via the use of the *ifdefs*. This means that, as long as all changes are valid, different optimisations from different vendors can co-exist at the same time even if this leads to code duplication in the benchmark code.

It should be noted that all accepted changes will be made available to all tenderers, and a tenderer will then be allowed to use any code path that produces valid results. Further information on how code changes and optimisations are to be implemented by vendors will be released in due course together with the RAPS benchmark code.

## C7_ITT380

**Question:**
What is the mechanism for handing back optimisations?

**Answer:**
Code changes must be sent to ECMWF via the **itt380-benchmarks@groups.ecmwf.int** email address in the form of separate patches for each distinct optimisation and change. ECMWF will then verify the changes and report back to the vendor whether the changes are valid and can be integrated into the next RAPS benchmark release or not.

## C8_ITT380

**Question:**
Are changes to build and run options allowed after RAPS ITT release?

**Answer:**
Yes. However, ECMWF reserves the right to reject any changes of this type if it feels that they contravene the spirit of the benchmarking methodology or if they run the benchmark code in a way in which ECMWF would not be able to replicate when using the new system.

## C9_ITT380

**Question:**
Is the current release's AIFS code the same code as in the RFI issued in 2024?

**Answer:**
No. AIFS' architecture has evolved and the code is now hosted in an open-source codebase. The model architecture is now a mix of GraphAttention and windowed attention (using the flash attention library) blocks.

## C10_ITT380

**Question:**
Do versions of libraries such as PyTorch or FFTW as used in the response to the ITT's benchmark release have to be available in a public release?

**Answer:**
Yes, but versions are not limited to those publicly available at the time of ITT release. Any libraries and library versions publicly available at the time of submission of an ITT response are permitted.

## C11_ITT380

**Question:**
Do the libraries have to be stable releases or can nightly builds be used?

**Answer:**
Please see **C10_ITT380.**

## C12_ITT380

**Question:**
Are the benchmarks going to be released with an output correctness test?

**Answer:**
The methodology for verifying output correctness is still being developed for both IFS and AIFS. Some correctness check mechanisms are already available in both IFS and AIFS. A guide on how to use these will be available in the RAPS benchmark documentation.

## C13_ITT380

**Question:**
Will the AIFS training benchmark have any quality metrics or just performance metrics?

**Answer:**
The most important metrics are the performance metrics. However, we still need to ensure that we are producing sensible results. We are currently working on a mechanism to verify this after a benchmark run and will provide more details in the RAPS benchmark documentation.

## C14_ITT380

**Question:**
What are the values of n, m and l in the IFS benchmark runs?

**Answer:**
The final values of these are not set at this time and will only be done so once all code optimisations are accepted, merged and the ITT release version of the RAPS benchmarks are finalised. At this point all the benchmarks will be run and the final values of n, m and l calculated for ECMWF' current HPC system.

## C15_ITT380

**Question:**
Do I need a benchmark machine that is able to run n, m and l copies of the benchmarks?

**Answer:**
No, each copy is being designed to be standalone and non-interfering. Therefore, on a system with a good fabric it is expected each copy will have a similar runtime. Only during acceptance is it required to run the n, m and l copies at the same time.

## C16_ITT380

**Question:**
Do the recorded times for IFS include set up costs?

**Answer:**
No, we are contemplating completely removing all setup and initialisation costs from the recorded IFS benchmark times.

## C17_ITT380

**Question:**
Is it allowed to run multiple instances of AIFS inference on one GPU if future technology supports this?

**Answer:**
Yes.

## C18_ITT380

**Question:**
How many parameters does the current AIFS model have?

**Answer:**
The model makes forecasts for approximately 100 variables and has 1 billion trainable parameters for the large-scale run, 240 million for the throughput run.

## C19_ITT380

**Question:**
What is the size of AIFS training data?

**Answer:**
The largest 4km data set will be a few tens of TB in size. However, this will be generated locally by the vendor from a much smaller 150 GB data set using utilities provided by ECMWF which will also perform the necessary validations to ensure that the correct data set has been generated.

## C20_ITT380

**Question:**
Is the AIFS parallelism based on samples?

**Answer:**
AIFS supports both data and model parallelism, this is described here:
https://anemoi.readthedocs.io/projects/training/en/latest/user-guide/distributed.html

## C21_ITT380

**Question:**
Are you expecting to use node-local SSDs (directly, or based on an intermediate software layer for the aggregation of the local SSDs of the nodes, for the duration of the job, to a single namespace accessible by all the nodes in the job) in the AIFS training?

**Answer:**
This is still under investigation and node-level storage may be preferred. Further clarification and guidance will be issued at a later date.

## C22_ITT380

**Question:**
Can you share log files from running on other non-ECMWF systems?

**Answer:**
Yes, log files and example configurations for a number of HPC systems where ECMWF ran IFS and AIFS as part of RAPS will be made available to all vendors.

## C23_ITT380

**Question:**
Will IO be part of the benchmarking tasks?

**Answer:**
Yes, but as a separate benchmark kernel focussed on IO only, targeted mainly at storage system benchmarkers; its configuration is still being finalised. The only IO actually needing storage hardware included in any of the IFS or AIFS inference benchmarks is reading-in initial data, but please see the C16_ITT380; however, AIFS training benchmarks will read the supplied, or constructed by cloning, training data.

## C24_ITT380

**Question:**
Will there be any high-performance interconnect network related benchmarks?

**Answer:**
General evaluation of partitions' interconnect performance is included through the technical requirements as to how the successful tenderer is required to run the benchmarks as part of the Acceptance Tests for the verification of their performance commitments, but there may be some small benchmarks to run as part of the system side of things to sanity check aspects of the machines. These will be outlined in the full ITT documentation.

## C25_ITT380

**Question:**
Do you expect to inject inference into the main physics models?

**Answer:**
No, not for this ITT.

## C26_ITT380

**Question:**
Do you expect the SIM and MLM partitions to be separated systems?

**Answer:**
Requirements will be finalised in Volume 2 of the specification, but it is currently expected that a flexible approach will be taken: An MLM partition may share a common interconnect with one SIM/GPIL partition, or, alternatively, MLM partitions may each be stand-alone. For resiliency considerations, the two MLM partitions must not be on the same interconnect. If the MLT-partition is provided on-premises, it may share a common interconnect with one SIM/GPIL partition (from the data hall without the MLM partitions), or it may be stand-alone. It should be noted, however, that all on-premises partitions will be required to have access to all global parallel filesystems installed in the Bologna Data Centre; if the MLT partition is provided off-premise, it is only expected to have access to a co-located storage pool, not the storage in the Bologna Data Centre.

## C27_ITT380

**Question:**

Is it required to have homogeneity in the architectures and manufacturers within and between the partitions?

**Answer:**
We expect partitions of the same type to be identical. It is desirable but not required that the SIM-nodes and ML-nodes have the same CPU architecture.

## C28_ITT380

**Question:**
Will the benchmarks, particularly for AIFS, be architecture independent?

**Answer:**
Yes, it is ECMWF's intention to ensure that IFS and AIFS will work with as many hardware options as possible, in order to give tenderers flexibility in their approaches to optimised responses.

## C29_ITT380

**Question:**
Do you have any requirements regarding the power consumption of the system?

**Answer:**
The Bologna data centre's environmental constraints for the installation will be detailed in ITT documents.

## C30_ITT380

**Question:**
Will electrical power consumption come into the benchmarking evaluation?

**Answer:**
No, Energy consumption is not part of the benchmark evaluation but will be evaluated under other areas in the ITT.

## C31_ITT380

**Question:**
Can the MLM partitions include both CPU-only nodes for inference and GPU nodes for training?

**Answer:**
Both training and inference will be performed on the MLM-partitions, and benchmarks will also assess the performance of training on the MLM-partitions. Moreover, ECMWF prefers to have the same environment and software stack for training as for inference to avoid problems it has encountered in the past where certain packages were only supported on GPUs and where support for CPUs has taken significant amount of time to arrive (e.g., flash attention). Consequently, all nodes in the MLM partition need to include GPUs.

## C32_ITT380

**Question:**
How important is double precision for the MLT- and MLM-partitions?

**Answer:**
The MLM- and MLT-partitions are primarily designed for running data-driven models such as AIFS in both training and inference configurations. AIFS does not make use nor require hardware support for double precision. Consequently, hardware support for double precision is not mandatory for the GPUs hosted in the MLM- or MLT-partitions.

## C33_ITT380

**Question:**
Given that CPU-based inference has been excluded from the current ITT scope, we would like to understand how this option could be reconsidered or integrated into the evaluation process.

**Answer:**
ECMWF requires additional time to consider this request.

## C34_ITT380

**Question:**
Do you foresee IFS SIM work ever moving onto the MLT/MLM partitions if the successful tender is based on CPU-only nodes for the SIM partition?

**Answer:**
Yes, we expect that the IFS in single precision may also use MLM- partitions and perhaps even the MLT partition as well. However, these partitions are not designed primarily for IFS, but for AIFS. At the same time, the GPUs in the MLM- and MLT-partitions will be required to be able to support the programming models used in the IFS for running on GPUs, which is either OpenACC or OpenMP target offload, although no performance commitments will be required.

## C35_ITT380

**Question:**
Will ECMWF consider MLM-partitions comprised of heterogeneous nodes (say TRAIN nodes and INFERENCE nodes) to serve the different AI challenges of training and inference, respectively? And hence split the scoring of $F_{train}$ and $F_{inf}$ into $F_{train}$ scores on TRAIN nodes and $F_{inf}$ scores on INFERENCE nodes?

**Answer:**
No. ECMWF does not want the MLM partition to contain different node types.

## C36_ITT380

**Question:**
Will ECMWF consider evaluating $F_{inf}$ on both a SIM partition and an MLM partition?

**Answer:**
ECMWF requires additional time to consider this request.

## C37_ITT380

**Question:**
The ITT is stated to be released on 15th October, is this correct?

**Answer:**
Yes, this is the aim and earliest it will be released. Any substantial delays will result in an updated PIN and emails to vendors.

## C38_ITT380

**Question:**
Will the benchmarks be sent to component manufacturers?

**Answer:**
All component manufacturers will be encouraged to ask for the benchmarks, but any communications between them and ECMWF will not be shared with potential tenderers. Tenderers may contact component manufacturers directly if they wish to do so.

## C39_ITT380

**Question:**
Are there any expectations of reporting results back to ECMWF before the ITT?

**Answer:**
The benchmark releases prior to the official release of the ITT are intended to allow all potential tenderers and OEMs to make improvements to the code bases, and for ECMWF to validate and incorporate such improvements. They are not intended to provide early feedback on performance measurements.