CLARIFICATIONS

ECMWF/RFI/2017/001 Request for Information for a High Performance Computing Facility (HPCF) for ECMWF

Clarifications issued 24 July 2017

This document will be updated as additional clarifications are published. See version history below.

Ref: ECMWF/RFI/2017/001
ISSUED BY: ECMWF Administration Department Procurement Section
First issue: 29 June 2017 Second issue: 4 July 2017 Third issue: 19 July 2017 Fourth issue: 24 July 2017

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

First issue clarifications:

1 Ref: C1_RFI 001

Q: In Section 2.4 - Support Requirements, Para 5 ECMWF request a full-time application software support service. Could you please clarify which applications you require support on and if home grown code such as IFS what training will you provide and in what time frame? Do you need this service, along with the requirement for a full time software support resource requested in Para 3 costed within the spreadsheet?

A: The role of the "full-time application software support" service is not to support any particular application, but rather to assist developers with migration, optimisation and debugging of their code. A person involved in providing the support service is expected to be an experienced scientific software developer with knowledge of the vendor's hardware and software who can work closely with an ECMWF or Member State developer to get the most out of their codes or diagnose a particular issue. Prior familiarity with IFS or Member States' applications is not necessary.

Training will not be provided. However, it is expected that an ECMWF or Member State developer will work closely with the application support person to understand the issue with the application.

Please include the costing for this support in the spreadsheet under the "Annual software support" column.

2 Ref: C2_RFI 001

Q: Are the storage performance metrics provided for the slowest of either Read or Write performance (i.e. both separate 100% Read and 100% Write performance tests must be > performance requirement)?

A: The figures in table 3 should be achievable as a minimum for the sustained aggregate bandwidth of concurrently executing IOR kernels in read and write mode, for all ratios of read:write bandwidths in the range from 2:1 to 1:2. Respondents can select suitable transfer sizes and alignments.

Second issue clarifications:

3 Ref: C3_RFI 001

Q: Does ECMWF have a completion date for the newly proposed Data Centre in Bologna and what mitigations are in place if this build slips?

A: As recently announced, the new ECMWF Data Centre will be located in Bologna, Italy. Contractual arrangements are currently under discussion, but with a view to having the Data Centre completed by late summer 2019.

The project plan contains some contingency for slippage. Respondents can assume that the Data Centre in Italy is available. Should anything happen to change this assumption, it will be dealt with in the Invitation to Tender.

4 Ref: C4_RFI 001

Q: Please could you provide details on how to access KRONOS?

A: All the RFI benchmarks, including Kronos are available for download from the ECMWF ftp site.

Third issue clarifications:

5 Ref: C5_RFI 001

Q: Could you clarify what counts as a “good” correctness check for the test-of-adjoint benchmark?

A: This test serves two purposes: to measure the accuracy of the compiler and math libraries and to estimate the cost of the minimisation process in the IFS 4D-Var data assimilation application. When the test completes, near the bottom of the output can be seen some text as follows:

```
 TEST OF THE ADJOINT
                   12345678901234567890
 < F(X) , Y >  = -.13405644731607836206E+02
 < X , F*(Y) > = -.13405644731674996706E+02
 THE DIFFERENCE IS 22562.436      22562.4359033512     TIMES THE ZERO
OF THE MACHINE
```

If the difference in the zero of the machine is "excessive", meaning greater than 1000 for the moment, then examine the figures. Within a run the figures for < F(X) , Y > and < X , F*(Y) > should be exactly the same up to and, preferably, beyond 10 decimal places (the lower the resolution the more decimal places should be exact). If they are only accurate to nine decimal places, there is an issue somewhere (with the Compiler being most likely). If they are only accurate to eight decimal places, then there is a serious error somewhere and the run should be considered invalid.

For runs using the same inputs, number of nodes and cores/threads then the values should be consistent across multiple runs.

Fourth issue clarifications:

6 Ref: C6_RFI 001

Q: The endurance of flash and other non-volatile solid-state media is limited and can vary significantly for different technologies. The endurance rating also has a significant impact on the price point for the media. Please can you provide an estimate of the amount of data written per day to the front-end time critical storage? For example if the whole 150TB capacity were re-written for each of the six cycles per day the media would need to sustain 900TB of writes per day.

A: Each forecast cycle is independent and will produce a full set of new files. Therefore, the front-end storage tier should be able to cope with a full write on each cycle (i.e. 900TB of writes per day).

7 Ref: C7_RFI 001

Q: In section 2.3. Storage performance requirements, it is stated “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.” Our understanding is that this

corresponds to an independent time-critical pool and research pool per building block, with a minimum of two building blocks, resulting in at least four pools. Is this correct?

A: Yes, each independent building block, of which there will be more than one, should be self-sufficient and have both a time-critical storage pool and a research storage pool, resulting in at least four pools in total. Building blocks should be able to share their high throughput storage with other building blocks of the HPCF at equal performance levels.

8 Ref: C8_RFI 001

Q: Please can you confirm that nodes used to meet the single node performance requirements described in section 2.2 (equivalent of 7,500 current cores with 2GiB/core and 128GiB/node) are in addition to any nodes required to meet the parallel performance requirements in 2.1, even if all nodes used are identical?

A: We can confirm that the nodes required to meet the single node performance requirements are in addition to the nodes needed to meet the parallel performance requirements. Special nodes to provide such services as system management, batch scheduling, network and file-system access should also be counted separately.

9 Ref: C9_RFI 001

Q: In paragraph 2.1, Table 1 of the RFI document, is there a typo for the number of nodes needed to run TCo1279 under an hour?

A: The number of nodes needed to run a copy of the "HRES - TCo1279L137 double precision forecast" is incorrect. The value should be 8 x 173 nodes (49,824 cores) which increases the size of the building block to 6,520 nodes and 234,720 cores.

We apologise for this mistake.

10 Ref: C10_RFI 001

Q: The text in section 2.1 of the ECMWF RFI defines a packet to be completed within one hour. The table in that section sizes this to be 6,232 nodes of your current capability. However, the table also lists as a goal, a total size of 8,192 nodes & memory size of 1 PiB. Please could you clarify which system size you would like to see costing for.

A: The values in the "Goal, expressed in terms of ECMWF's current Intel E5-2695v4, 128 GiB nodes" column are only intended as a guide. Please size the system according to the information in the "Goal" column.

11 Ref: C11_RFI 001

Q: The final line of Section 2 (page 9) states: "It is envisaged that building blocks will share their high throughput storage with other building blocks of the HPCF at equal performance levels." Would ECMWF anticipate this applying to the front-end tier of the time-critical storage system?

A: For this RFI we expect that the front-end tier of the time-critical storage is local to the building block. If there is potential for this storage tier to be shared between building blocks then we would be interested in understanding the options available.

12 Ref: C12_RFI 001

Q: Could you give us some clarification on how to interpret Q36 in the RFI document please. The question says: “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.” What do the “three different runs” refer to?

A: Where a particular test requires a wall-clock runtime of 3600 seconds (or less), the vendor is kindly asked to run the test three times, with each run being on a different number of nodes. This will enable a curve fitting or graphical plot inspection method to be used to extrapolate or interpolate the number of nodes needed. The wall-clock time that is used should be after the application of GENFACT via the jobinfo-script.