**REQUEST FOR INFORMATION (RFI)**

**FOR a High Performance Computing Facility (HPCF) for ecmwf**

**ECMWF/RFI/2017/001**

**Response Template**

1. 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. 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

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

1. Questions
2. Respondents are asked to provide a description of a system building block to meet the requirements in section 2 of the RFI document. The description of the hardware should include:
   * Indicative system layout drawings, preferably on a 600mm grid.
   * The high-performance interconnect. Details should include:
     1. Achievable MPI latency and bandwidth;
     2. 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 section 2.3 of the RFI document.
   * The connectivity and bandwidth between the I/O nodes and the pools of compute nodes
3. Please describe the timeline for the availability of this solution and describe the major risks associated with delivery and performance using the table below.

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> |  |  |  |  |  |  |

1. 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?
2. It must be possible to execute efficiently a single MPI communication intensive program of at least the largest routine job size given in section 2.1 of the RFI document, i.e. TCo1999. Describe how the proposed interconnect topology can meet this requirement.
3. 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.
4. Would it be possible to have a subset of the nodes on the same interconnect supported by higher backbone capability?
5. 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.
6. 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?
7. 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?
8. 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?
9. What percentage of the overall system cost does the storage represent?
10. 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.
11. Respondents are asked to describe how they could meet the requirements for support described in section 2.4 of the RFI document.
12. Please identify any impact that the location of the Data Centre may have on your support model.
13. Respondents are asked to describe the training which is available to enable the efficient usage of the HPC.
14. Respondents are asked to describe the support and consultancy which is available to port existing applications.
15. Please provide details of machine power requirements and estimate the total power consumption in kW of the proposed building block.
16. 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.

1. If appropriate, please indicate any requirements you have for the quality of water used in the cooling system.
2. Please highlight any restrictions on the layout of the system or the distance between connected components.
3. Please indicate the size and weight of a full rack of each equipment type (e.g. storage, network, or compute) used in the system.
4. Please provide a cost estimate of the system building block described in Q1 using the “RFI001 cost estimate tables” spreadsheet.
5. How would the costs above vary for a second or subsequent building blocks or parts thereof?
6. How would the software and hardware support costs vary if the amount of hardware is increased from one building block?
7. 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?
8. 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.
9. 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 in section 4 of the RFI document, together with the impact of these alternatives on contractual arrangements, residual value risks and other costs (e.g. removal of hardware at contract termination).
10. 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.
11. Please state which of the options you believe is preferable and explain the benefits of this option to both parties.
12. 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?
13. 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?
14. 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.
15. Please describe how ECMWF could optimise the benefits of any additional funding that may become available during the contract term.
16. 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?
17. Respondents are asked to describe the benchmark system used by completing the table below.

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** |  |

1. 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.

The tables below provide summary formats for reporting data from the benchmark runs.

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** |  |  |  |  |  |

Timings for the IFS tests

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

1. 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.
2. 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. Please provide the results from these additional runs.
3. 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.
4. Respondents are requested to run the Kronos benchmark, which represents ECMWF research workflows, and return the resulting output.