Scaling Compute: AI at 1/1000th the cost
Technical Area 4 Benchmarking

Date: 17 February 2025

V1.0

Request for Proposals

Technical Area 4 Benchmarking

SECTION 1: Information and Conditions

Programme Thesis and Overview

This Request For Proposals (RFP) is derived from the programme thesis Unlocking AI compute hardware at 1/1000th the cost and Nature computes better opportunity space.

As described in the solicitation for Technical Areas (TA) 1-3, the programme is designed to demonstrate that:

- It is possible to drop the hardware costs required to train large AI models by >1000x
- It is possible to do this *without* primarily relying on leading-edge fabrication facilities

All programme activities will be anchored around reducing the hardware costs required to train large-scale AI models. The initial **programme targets** are defined below, where we show the targeted time/cost pareto frontier to train three specific workloads from the MLPerf benchmark (to the quality level described in the benchmark).
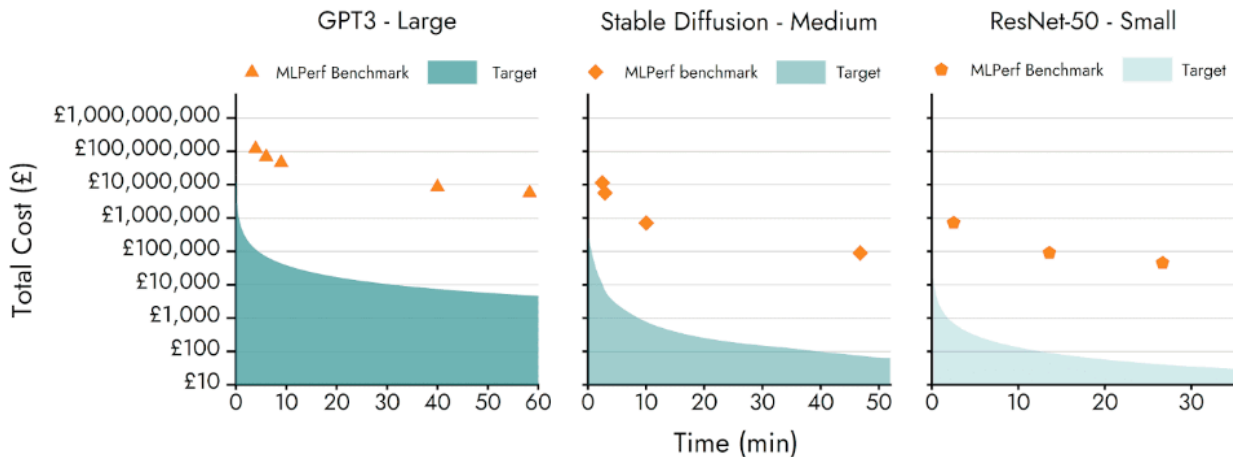
During the delivery phase, **all programme activities in TA 1-3 will be evaluated based on ability to meet these targets.** Creators from TA 1-3 will be asked to estimate the expected manufacturing and energy costs of their proposed solutions, and be tasked with justifying their estimates to ARIA. We acknowledge that both the workloads chosen (individual plots) and the baseline values for the existing state-of-the-art (orange triangles) are moving targets.

This TA 4 RFP is designed to find a team who can help track these moving targets and publish their findings to the research community. The purpose is to provide an accurate (and open) source of ground truth of where the overall programme targets should be in a fast-moving industry.

In this RFP, we are looking for proposals for TA 4 only, applications for TA 1-3 are now closed.

Technical Area Objectives

As outlined above, TA 4 represents the benchmarking component of the programme. Successful applicants for TA 4 will be asked to monitor and select relevant workloads representing the state-of-the-art AI models, profile their performance on commercial AI accelerators, create estimates of cost and runtime, and share their findings (and code) with ARIA and the broader R&D community.

We are looking to fund this TA with up to £2m (inclusive of VAT and all other costs including overheads). We expect to fund one award in this TA. Applicants recognise and accept that it will be at ARIA's sole discretion as to which, if any, proposal is accepted.

What are we looking for/what are we not looking for

We are looking for a team who can periodically poll the current landscape of best-in-class AI models (workloads with available code/data and which can fit on modest-sized commercial hardware). The TA 4 team will be expected to estimate performance/cost of pre-training, fine-tuning, and inference of the selected models. The successful TA 4 team will be evaluated on how best they can: (1) identify models and implement known algorithmic optimisations which are representative of the existing state-of-the-art, (2) perform baseline benchmarking using commercially available hardware and profiling tools, and (3) extrapolate performance for fully scaled-out systems where appropriate.

We are not looking for the TA 4 team to create entirely new models or re-create full-scale training of large LLMs (this will exceed the available budget).

Project Milestones

Activity for this TA will run from May 2025 through the remainder of the Scaling Compute Programme until September 2027. The primary milestones will come in the form of technical reports and usable code, delivered at quarterly intervals, to be provided directly to ARIA and other Creators.

An example of the content/cadence is shown below.

- Month 3 → Model Selection Report
- Month 6 → Initial Baselining Assessment / Code Share
- Month 9 → Model Selection Report (updated)
- Month 12 → Baselining Assessment / Code Share (updated)

Success/pivot/closure criteria for each project will be determined by the applicant's ability to meet these agreed-upon milestones, and the quality of the assessments.

Programme & Project Management

Alongside our standard project management requirements, the ARIA Programme Director (PD) will also monitor progress of each project through a series of 1:1 calls, site visits, and Programme-wide meetings. Project status updates are expected to be shared at quarterly intervals between ARIA and the selected team. Reports and code will also be shared with Creator teams from TA 1-3 for feedback.

SECTION 2: Eligibility & Application Process

Eligibility

We welcome applications from across the R&D ecosystem, including individuals, universities, research institutions, small, medium and large companies, charities and public sector research organisations.

This RFP is not seeking proposals for research projects, the requirement is to provide a service to the programme. Therefore, our review process of applications is different from the other TAs in the programme. For more information on the evaluation criteria for this RFP, see here.

Application Process

The application process for Technical Area 4 consists of one stage.

Applicants are invited to set out how they propose to deliver the services outlined within this RFP. The format below is set out as a guide and represents a maximum length response.

Section 1: Technical proposal

Applicants are required to submit a detailed proposal including:

- **Your Approach -** to help us gain a detailed understanding of your proposal. This should include:

  + A detailed explanation of the proposed solution, how it supports the technical objectives of the requirement. This should be supported by technical information, visual aids and/or data, where relevant.
  + A comprehensive list of the known technical risks/unknowns standing in the way of achieving the objectives of the requirement.
  + Description of the proposed activity of work, key metrics and milestones and any dependencies and assumptions.

- **Background, Expertise and Ability to Deliver -** to help us learn more about the team who will be delivering the project and their expertise. This should include:

  + Details of the project team, including a clear demonstration of experience and expertise in delivering benchmarking projects similar to the requirements outlined in Section 2.
  + If you intend to collaborate with or rely on any third parties, sub contractors/grantees, who they are and which elements of the project they will support/deliver.
  + Provide examples of your ability to provide the operational and executional resources required to deliver.

How to format your technical proposal:

- Page count: 4 pages, (including diagrams, excluding references) single line spacing, standard character spacing (neither expanded nor condensed)
- Font: Arial. Colour: black. Size: 11-point font or larger
- Margins: At least 0.5" margins all around
- File Type: PDF

Section 2: Commercial proposal and administrative questions

In completing your application you must also provide answers to the following questions. Answers to these questions are not included in the 4 page cap. You should complete these questions in the application portal so there is no need to format these in a specific way.

This includes:

+ The proposed cost, including a breakdown of costs. A short table is provided in the application portal (please ensure you account for VAT where applicable). You should base your pricing on a time and materials basis
+ Any reliance on pre existing background IP (this includes third party data you may rely upon to deliver the service - please ensure any cost associated with accessing/purchasing this data is included in your cost breakdown)
+ Any other factors or restrictions that might impact your freedom to operate and deliver the project (such as conflicts of interest with the PD, import/export or security restrictions that you are aware of)
+ Details of any sub-contractors

Key Contract Terms

Any resultant agreement from this RFP will be a contract for services.

The contract will be placed on mutually agreed terms and conditions (T&Cs) provided by ARIA to successful applicants. The proposed terms will include the following principles:

+ Intellectual property created by projects funded in TA 4 shall be:
  ○ Published under a Creative Commons Attribution (CC-BY) licence, if not software
  ○ Dual-licenced under an MIT licence and an Apache 2 licence, if software
  ○ Subject to a patent non-aggression pledge (example), if patented
+ Either party shall have the right to terminate the contract or part of the contract for convenience upon ninety days' prior notice
+ ARIA shall have the right to terminate the contract or part of the contract where the supplier fails to provide the service contracted, upon thirty days' prior notice
+ The supplier shall indemnify ARIA, its employees, officers and agents against the supplier's infringement of third party Intellectual Property Rights
+ All information shared with the supplier shall be subject to confidentiality terms

SECTION 3: Evaluation Criteria

Applicants will be evaluated on the comprehensiveness of the proposed methods to develop hardware cost models. This includes not just processor cores but interconnect subsystems, cooling, energy costs, and other critical aspects of AI hardware systems. As such, proposals will be evaluated against the following criteria:

Section 1: Technical Proposal:
+ **Criteria A - Your Approach** - A clear articulation of what you see as our requirements as set out in the Technical Area Objectives section above, and how you would meet them including proposed technical solution and project plan
+ **Criteria B - Background and Experience** - A clear demonstration of experience in delivering similar projects
+ **Criteria C - Ability to Deliver** - Demonstrate that you have (or have access to) the operational and executional resources required to deliver

Section 2: Commercial proposal and administrative questions
+ **Criteria D - Commercial Proposal** - Commercial terms that demonstrate value for the tax-payer

SECTION 4: Timelines

This RFP for TA4 will be open for applications as follows (we may update timelines based on the volume of responses we receive):

Applications open	17 February 2025 (13:00 GMT)
Application submission deadline	10 March 2025 (13:00 GMT)
Application review	10 March - 3 April 2025

As part of our review we may invite applicants to meet with the PD to discuss any critical questions/concerns prior to final selection — this discussion can happen virtually or we may seek clarification on certain aspects of your proposal via email.

Successful applicants notified	4 April 2025

At this stage you will be notified if you have or have not been selected for an award subject to due diligence and negotiation. If you have been selected for an award (subject to negotiations) we expect a 1 hour initial call to take place between ARIAs PD and your key team members within 10 working days of being notified.

We expect contract signature to be no later than 4 weeks from successful/ unsuccessful notifications. During this period the following activity will take place:

- Due diligence will be carried out
- The PD and the applicant will discuss, negotiate and agree the project activities, milestones and budget details
- Agreement to the set Terms and Conditions of the Contract.

SECTION 5: How to apply

Before submitting an application we strongly encourage you to read this RFP document in full.

If you have any questions relating to the RFP, please submit your question to clarifications@aria.org.uk.

Clarification questions should be submitted no later than 3 days prior to the application deadline date. Clarification questions received after this date will not be reviewed. Any questions or responses containing information relevant to all applicants will be provided to everyone that has started a submission within the application portal.

Please read the portal instructions below and create your account before the application deadline. In case of any technical issues with the portal please contact clarifications@aria.org.uk.

Portal instructions

APPLY HERE

SECTION 4: Conditions of RFP

Confidentiality, Publicity, Conduct and Conflicts of Interest:

The contents of this RFP are confidential and must not be not copied, reproduced, distributed, or passed to any other person at any time, except for the purpose of enabling the applicant to submit a proposal. ARIA may use the information included in an applicant's response for any reasonable purpose connected with this RFP. In particular, once an applicant has been excluded, ARIA reserves the right to use any ideas contained in that applicant's proposal in any ongoing discussions with other applicants but undertakes not to reveal the identity of the provider of these ideas. No publicity regarding the subject matter of this RFP or the award of any Contract will be permitted unless and until ARIA has given express written consent to the relevant communication.

ARIA reserves the right to: a) waive or change the requirements of this RFP from time to time without prior (or any) notice being given by ARIA. In the event that ARIA makes material changes to this RFP, ARIA may choose to extend the deadline for submission; b) reject any applicant that does not submit a compliant Proposal in accordance with the instructions in this RFP; c) disqualify any applicant that is guilty of serious misrepresentation in relation to their Proposal; d) withdraw this RFP at any time, or to re-invite applicants on the same or any alternative basis; e) choose not to award any Contract as a result of the current procurement process; and f) make whatever changes it sees fit to the Timetable, structure, or content of the procurement process, depending on approval processes or for any other reason. ARIA will not be liable for any applicant's costs, expenditure, work, or effort incurred by them in carrying out enquiries in relation to, proceeding with, or participating in, this RFP, sourcing process, or procurement, including if the procurement process is terminated or amended by ARIA.

Suppliers are expected to adhere to the Government's supplier Code of Conduct. This is available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/779660/20190220-Supplier_Code_of_Conduct.pdf.

Any attempt by applicants or their advisors to influence the contract award process in any way may result in the applicant being disqualified. Specifically, applicants shall not directly or indirectly at any time: a) enter into any agreement or arrangement with any other person as to the form or content of any other Proposal, or offer to pay any sum of money or valuable consideration to any person to effect changes to the form or content of any other Proposal; b) enter into any agreement or arrangement with any other person that has the effect of prohibiting or excluding that person from submitting a Proposal; c) canvass ARIA or any employees or agents of ARIA in relation to this sourcing process; or d) attempt to obtain information from any of the employees or agents of ARIA or their advisors concerning another applicant or proposal. applicants are responsible for ensuring that no conflicts of interest exist between the applicant and its advisors, and ARIA and its advisors.

Any applicant who fails to comply with this requirement may be disqualified from the procurement at the discretion of ARIA.

Applicants recognise and accept that it will be at ARIA's sole discretion as to which, if any, proposal is accepted.