Advanced
Research
+ Invention
Agency

ARIA

# Scaling compute: AI at 1/1000th the cost

## Call for proposals

**Date: 13.03.24**

v1.0

## Table of Contents

## SECTION 1: Programme Thesis and Overview

This solicitation is derived from the programme thesis Unlocking AI compute hardware at 1/1000th the cost and Nature computes better opportunity space.

The digital electronics industry that has transformed our lives in immeasurable ways is defined by the simple fact that, for 60+ years, we have benefited from **exponentially more computing power at lower cost**. This fact is no longer true and has coincided with an explosion of demand for more compute power driven by AI.[1]

The mechanisms used to train these AI systems utilise a surprisingly narrow set of algorithms (gradient descent) which iteratively call an even narrower set of hardware building blocks, and much of the industry focuses its efforts towards:

1. Cramming more components into each building block at lower costs
2. Devising new architectural arrangements of these building blocks
3. Developing algorithmic advances when operating these blocks at scale

While items (2) and (3) are being pursued vigorously by the world's largest firms, physical limits have made progress for item (1) economically unfavourable. This trend has manifested itself in a growing gap between the demand for more AI processing power and the supply of compute resources (shown in Figure 1). The world's leading AI models now cost upwards of **£100M to train**, and this combination of technological significance and scarcity have far-reaching economic, geopolitical, and societal implications.

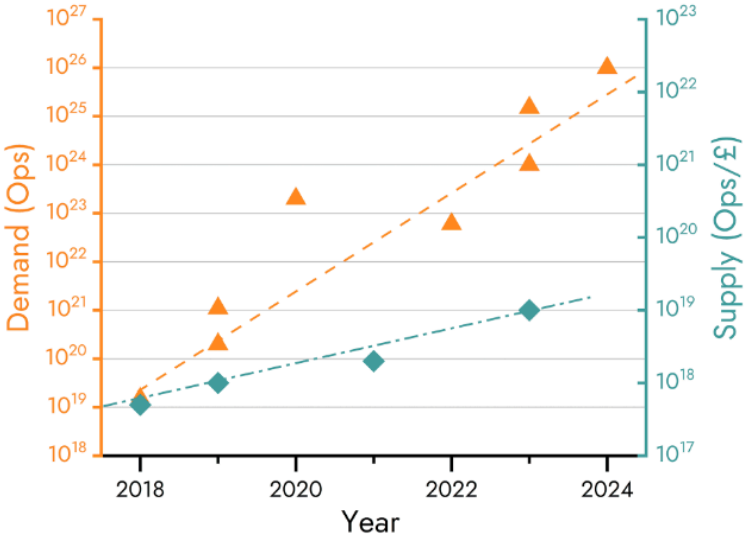


*Figure 1. Trends in demand for AI processing power and the availability of compute resources*

Nature provides us with at least one such existence proof that it is fundamentally possible to accomplish sophisticated information processing much more efficiently than today's large AI systems. The phenomenal scale and capabilities of our digital universe mask inefficiencies buried in the decidedly unnatural characteristics of discrete, clocked signals, and an opportunity now exists to rethink our existing paradigm in an effort to open new vectors of progress in the field of computing.

Combining this insight with the underlying economic trends, this 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).[2]

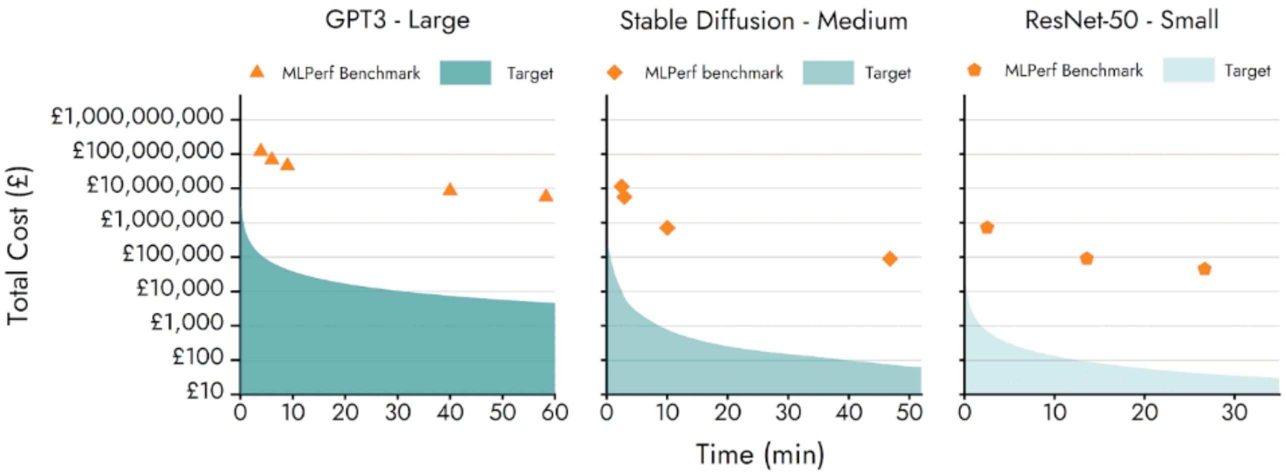


*Figure 2. The current cost of training large AI models and the programme target for training workloads from the MLPerf Benchmarks*

As AI becomes a primary economic driver for the semiconductor industry, capabilities proven out in the AI domain can more easily cascade into numerous disciplines beyond AI where information processing is critical, from scientific simulation to communication systems.

## SECTION 2: Programme Objectives

The programme seeks R&D Creators, individuals and teams that ARIA will fund and support. They will focus on three key tasks:

1. Develop software simulation tools capable of estimating performance of novel compute hardware focused specifically on training large-scale AI models
2. Demonstrate working hardware meeting specified building block targets
3. Demonstrate through systems-level analysis that if these building blocks were developed at scale, they could meet disruptive programme targets (Figure 2)

In order to realise these targets, we're looking to fund **three Technical Areas (TA)** detailed below, applicants must choose **at least one** path (and are free to choose more than one). Applicants will be tasked with describing how their proposed technology can meaningfully contribute to achieving the system-level targets in Figure 2.

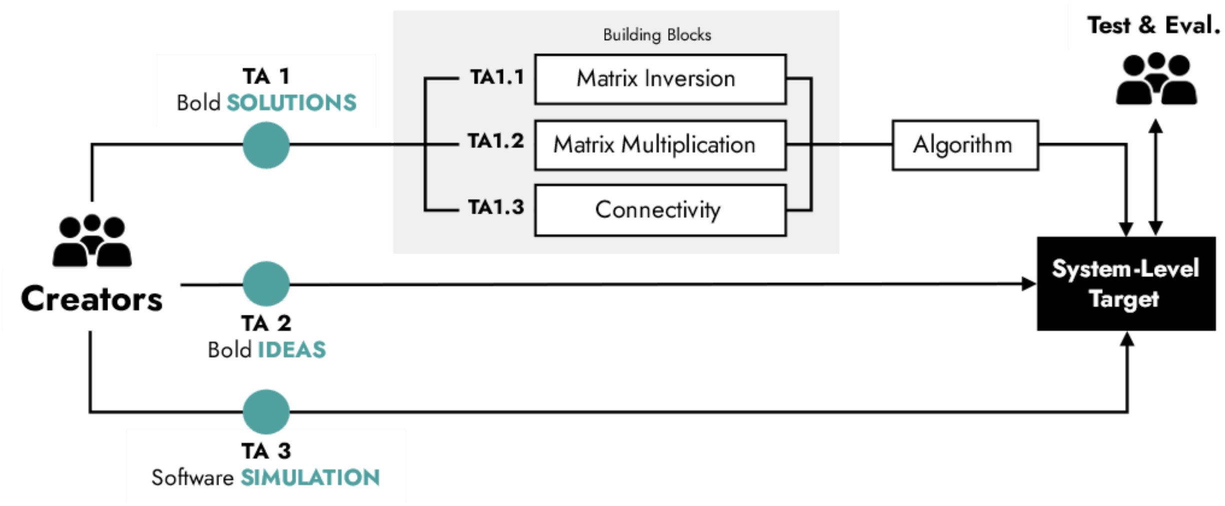


*Figure 3. The three programme technical areas, all anchored by the system-level targets that programme is shooting for*

### Technical Area 1 - Bold Solutions

Applicants choosing this TA will be tasked with combining specific building block hardware and unique algorithms which demonstrate feasibility towards achieving the system-level targets. These key building blocks will be:

- **TA 1.1 - Matrix Inversion**
  Matrix inversion represents a foundational mathematical kernel used in applications ranging from communications to scientific simulations, and represents a key tool in a variety of alternative AI training techniques. Meeting the specified targets for energy/time required to calculate the

approximate inverse of a matrix could allow researchers to revisit a wide array of optimisation algorithms which have not yet shown significant advantage over stochastic gradient descent due in part to the computational expense of this particular kernel.

- **TA 1.2 - Matrix Multiplication**
  Matrix multiplication represents a fundamental kernel underpinning ALL modern AI algorithms. Demonstrations of the performance targets shown in Section 3 would dramatically alter the current technological landscape for AI hardware accelerators.

- **TA 1.3 - Connectivity**
  All chip-scale building blocks will bump against fundamental size limits, at which point scalability will be determined by efficiently sending information between the blocks. Large gains in connectivity, either through new devices, topology-aware algorithms, or switching technologies, will profoundly impact the systems-level trade-space.

Depending on the building block chosen, Creators will be tasked with demonstrating working hardware which meet certain target specifications (described below). For TA 1, each building block should be manufacturable using existing CMOS processing facilities.

In combination with these demonstrations, Creators can then choose to rely on existing algorithms used for training large AI models, augment these algorithms to accommodate the novel hardware, or design entirely new algorithms uniquely suited to their building block. Creators will also be tasked with making the case, through a combination of hardware demonstrations and scaled-up systems analysis, the feasibility of meeting the overall system-level targets of the programme.

We are looking to fund this Technical Area with up to £26M. We expect to fund two to four awards in this TA.

### Technical Area 2 - Bold Ideas

This category is reserved for radical concepts which do not fit neatly into the structure of the Bold Solutions category. Creators will be free to propose any combination of algorithmic and/or hardware innovation manufacturable in **any** environment (not restricted to CMOS manufacturing) and will be asked to make the case that the proposed strategy can meet the programme targets.

The end deliverables will be:

- **Laboratory-level demonstrations of critical components for the proposed approach**
- **Scaling analysis that clearly demonstrates how these critical components can help realise the overall programme targets in Figure 2**

We are looking to fund this Technical Area with up to £8M. We expect to fund three to five awards in this TA.

### Technical Area 3 - System-Level Software Simulation

We will fund teams to develop software capabilities which will allow us to answer the following question:

> *"If we use the building blocks to build a fully functional system at scale, how well will it work and how much would it cost?"*

In the digital electronics industry it is possible to use software to make these performance assessments very accurately prior to fabricating larger chips. Software systems capable of achieving similar estimates when the underlying hardware modality deviates from standard digital architectures are much less mature. Packages which currently exist [3-5] have primarily targeted AI inference workloads. Applicants choosing this pathway will be tasked with building out a similar software ecosystem targeted towards large-scale AI training workloads. The ultimate goal of this pathway will be to develop software capable of accurately estimating system-level performance targets identified above. The software simulation tool should be open-source and available for other Creators to use in their efforts to perform systems-level analysis.

We are looking to fund this Technical Area with up to £5M. We expect to make one award in this Technical Area.

### Test & Evaluation (T&E)

In addition to the Technical Areas identified above we plan to appoint a test and evaluation team in order to ensure that the programme targets remain relevant throughout the programme. This external organisation will continuously evaluate both the target models that are chosen as well as the target cost/performance metrics to ensure relevance and accuracy. Selection of this T&E team will be subject to a separate competitive solicitation due to be released in the week commencing 3rd June. **Organisations interested in applying for the T&E component should not submit a proposal in response to this call**, instead applicants interested in participating in this element should register their interest by sending an email to clarifications@aria.org.uk and we'll notify you when the T&E solicitation goes live. The indicative value of this award is £3M.

Applicants can submit applications for Technical Areas 1, 2 and 3 above and the T&E component; however, Creators awarded funding under Technical Areas 1, 2 and/or 3 will not be permitted to fulfil the role of the T&E team.

## SECTION 3: Technical Metrics

During the delivery phase, **all programme activities will be evaluated based on ability to meet the targets from Figure 2.** Throughout the project duration, Creators will be asked to estimate the expected manufacturing and energy costs of their proposed solutions, and be tasked with justifying their estimates to the ARIA T&E team.

In addition to evaluating Creator outputs, the ARIA T&E team will be tasked with continuously validating Figure 2 throughout the course of the programme, ensuring that baseline targets stay relevant. Initial benchmark workloads will come from MLPerf, which is an industry organisation chartered to maintain relevant data, integrity, and relevance. The models and cost estimates will be re-evaluated by the T&E team every six months, and the results of these evaluations will be published by ARIA in an online newsletter. Based on the findings of the T&E team, systems-level targets will be subject to change, and may be re-adjusted at the end of each phase of the programme.

### Cost Estimations

Creators will be tasked to make cost estimations for their proposed technology on a best effort basis and to justify these estimations to ARIA (with the understanding that these estimates contain many assumptions). ARIA T&E team will be tasked with validating these assumptions and calculations.

An example calculation (used to determine costs for Figure 2) is described below:

Using the following MLPerf benchmark result:

| ID | Workload | Accelerator | No. GPUs | Time |
|---|---|---|---|---|
| 3.1-2057 | GPT | Nvidia H100-SXM5-80GB | 512 | 58.3 min |

$$\text{Capital Costs} = \frac{\text{Cost/Server}}{\text{GPUs/Server}} \times \text{No. GPUs} = \frac{\text{£60,000}}{8} \times 512 = \text{£3.84M}$$

***Assumption:*** *Cost/server is defined by an estimate of the cost of goods for Nvidia (NOT the price that Nvidia charges customers), assuming use of TSMC 5N process.*

$$\text{Operational Costs} = \frac{\text{No. GPUs}}{\text{GPUs/Server}} \times \frac{\text{Power}}{\text{Server}} \times \text{£/kWh} \times \text{PUE} \times \text{Utilisation\%} \times \text{total hrs}$$

$$= \frac{512}{8} \times 11{,}3\text{kW} \times \text{£}0.02 \times 1.1 \times 80\% \times 43{,}800 = \text{£}557{,}500$$

***Assumption:** Total hours assumes a 5 year life-cycle. Data centre PUE (power usage efficiency) is assumed to be close to optimal.*

**Total Cost = (Operational Cost + Capital Cost) = £4.4M**

*Note that many capital expenditures are NOT included in this cost calculation (switch ASICs, optical cables, cooling, etc..), leading to an underestimate of real-world costs.

Additional TA-dependent programme targets to be delivered throughout the projects are described below.

### TA1 - Bold Solutions

Creators in this TA will be tasked with demonstrating individual prototype building blocks meeting the specifications described below by the end of Phase 2 of the programme. By the end of Phase 3, Creators should demonstrate that these devices are amenable to scalable manufacturing processes and provide systems-level analysis showing how the particular approach could lead to realisation of the programme targets from Figure 2.

+ **TA 1.1 - Matrix Inversion**
  Matrix inversion represents a key computational principle in a multitude of applications. To evaluate baseline capabilities for Creators choosing this building block, we will look for technologies which can *approximately* compute (to within 1% accuracy) the inverse of random, positive semi-definite matrices. These computations must be accomplished in 10x less time and 1000x lower energy than the best known alternative. An initial target is shown in Table 1 (and will be re-evaluated by the T&E team throughout the programme).

| Matrix Inversion | Target Size | Time to Solution | Energy to Solution |
|---|---|---|---|
| TODAY<br>*Conjugate Gradient on Nvidia A10* | 1000 x 1000 | 10 ms | 1 J |
| Programme Target | 1000 x 1000 | 1ms | 1 mj |

*Table 1. Programme target performance for matrix inversion*

- **TA 1.2 - Matrix Multiplication**
  Matrix multiplication represents a key primitive underlying modern neural networks. Creators choosing this building block will be tasked with demonstrating in-memory computing technologies capable of meeting the requirements described in Figure 4. Note that the 'Compute Efficiency' metric below must describe the measurement at the **system-level** (not simply the arithmetic efficiency). Precision of the operation is left for Creators to define, bearing in mind that the technique (in combination with algorithmic innovations) must ultimately meet the quality targets set forth in the MLPerf benchmark (from Figure 2).

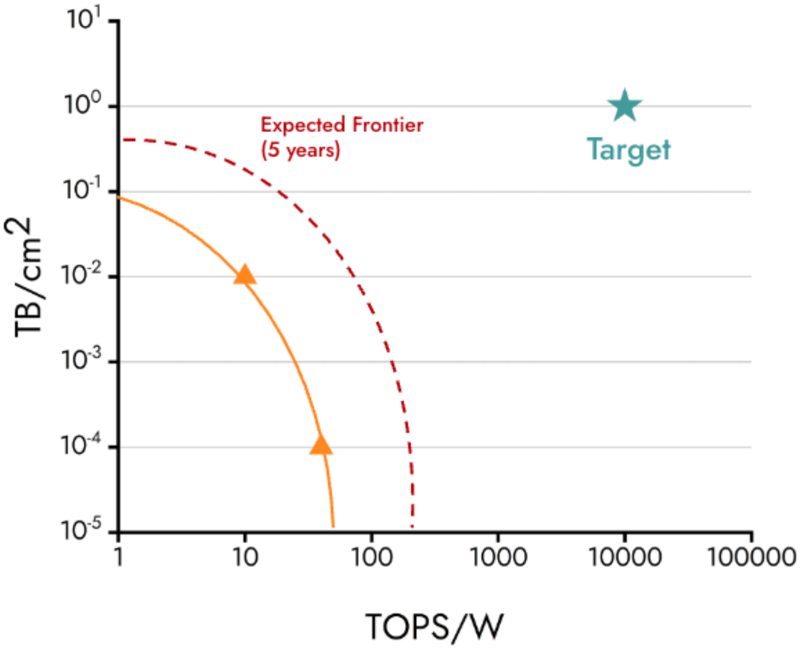


*Figure 4. Programme performance target for matrix multiplication*

- **TA 1.3 - Connectivity**
  Connectivity represents another critical piece of any system-level architecture. Creators choosing this building block will be tasked with developing technologies capable of connecting the **most chips** with the **highest bandwidth** at the **lowest cost**. Target metrics are displayed in Figure 5 (Point-point link bandwidth refers to the minimum bandwidth between any two points in the network). Proposed approaches can include any combination of device-, architectural-, or algorithmic-level innovation.

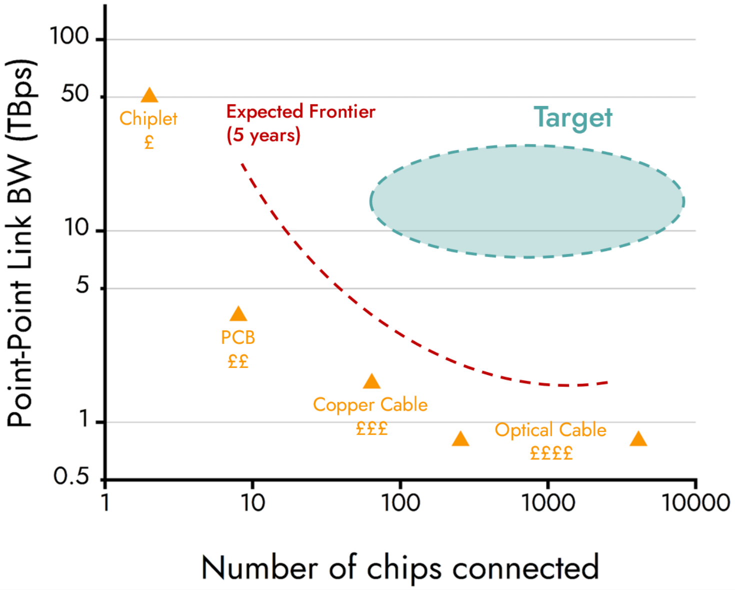


*Figure 5. Programme performance target for connectivity*

### TA2 - Bold Ideas

Creators in this Technical Area will be required to:

- Define their own building blocks and associated performance metrics
- Explain why these blocks/metrics are meaningful measures towards meeting the programme targets in Figure 2
- Demonstrate working building blocks in a laboratory environment
- Perform the systems-level analysis to show how these blocks can be scaled up in order to meet programme targets

Creators in this TA will be tasked with demonstrating working laboratory prototypes by the end of Phase 2 of the programme, and providing thorough systems-level analysis showing the impact of their proposed technology by the end of Phase 3. Throughout the project Creators will be primarily evaluated on whether or not their proposed solution is able to help meet the targets described in Figure 2.

### TA3 - Systems-Level Simulation Software

The primary metrics in this Technical Area are represented by whether or not Creators can use the software to simulate the performance of AI training workloads (specifically for the three workloads defined in Figure 2).

Creators in this TA will be tasked with presenting a live software demo displaying the capabilities of the software to the larger Programme community at the end of Phase 2, and making a well-documented open-source GitHub repo widely available at the end of Phase 3.

## SECTION 4: What are we looking for/what are we not looking for

We expect to fund a variety of technologies with diversified risk profiles. Examples of interesting technologies include (but are not limited to):

- (TA1.2) High-density, low-energy memory technologies (and associated in-memory computing architectures) [6]
- (TA1.2) Analog in-memory compute architectures based on reduced precision or noisy arithmetic [7-8]
- (TA1.1 / TA2) Novel mathematical frameworks which allow for a *fundamental change* in the underlying hardware [9-10]
- (TA1.3) Electrical interconnects capable of connecting more chips at longer reach
- (TA1.3) Optical interconnects technologies with dramatically lower costs
- (TA1.3) Novel switch architectures or sparsity-aware algorithms
- (TA1.1 / TA2) Energy-based AI models [11-13]
- (TA2) Neuromorphic computing techniques capable of training large-scale models
- (TA3) System simulation environments capable of estimating performance of novel computing HW blocks [3-5]
- Any other technological approach that can conceivably meet the desired programme targets

Examples of technologies we do **not** expect to fund include:

- Pure algorithmic advancements whose primary benefits can be realised using commercially-available hardware

## SECTION 5: Programme Duration and Project Management

The maximum term of the programme is 4 years, though applicants are encouraged to consider plans which may reach success (or failure) on faster timelines. Teams selected at the full proposal stage will enter into a contracting phase with ARIA where the specific scope of work will be finalised. This phase will require updated and more accurate cost assessments for the proposed project.

### Project Milestones

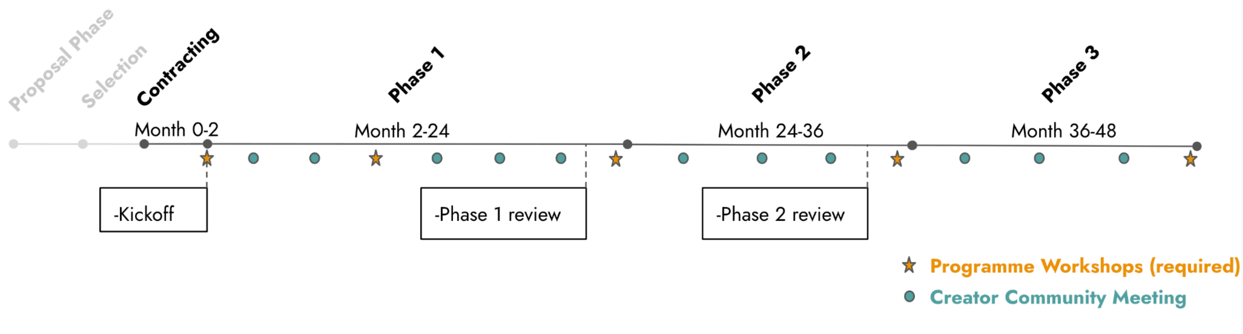
Each project's progress will be monitored using clearly defined milestones. Milestones will be defined by the applicant prior to the start of a project, be agreed upon by ARIA, and should be designed to easily convey progress to a third party. In order to do this, milestones should:

- Be specific, measurable, and signify a meaningful step towards reaching the overall programme goals.
- Include details on methods used for measurement and evaluation
- Be defined on a quarterly cadence for all phases of the Programme
- Include major "Go / No-Go" decision points

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

### Programme & Project Management

Alongside our standard project management requirements, the ARIA Programme Director 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 regular intervals between ARIA and each Creator. Additionally, ARIA will visit Creators once per quarter to discuss project status.



During each quarterly site-visit, Creators and the ARIA Programme Director will review the agreed upon Milestones, and discuss further details of each project. As part of that discussion, Creators will be encouraged to think through the following questions as they execute on their plan:

- What is(are) the target deliverable(s) for each phase of the programme?
- What are the top 3 risks identified at this stage of the project?
- What are the first 3 experiments required to overcome each risk?
- What are the expected outcomes/learnings from these experiments?
- How long will these experiments take and how much will they cost?
- What are the dependencies from prior activities/phases of the Programme?

Upon completion of each experiment, questions we will look to answer are:

- What new information has been gleaned?
- What (if any) risks have been overcome? What new risks have emerged?
- Did we learn what we thought we would learn?  If not, why not?
- Is there anything we can do to learn more or faster?
- Is there still a path towards the target? Are we heading towards any dead ends?

### Community events

In an effort to foster a collaborative research environment, ARIA will host regular Creator community events to allow all participants to exchange updates, ideas, and feedback on best paths forward. ARIA will also host annual in-person workshops at which Programme Creators can showcase their work to a wider research community.

## SECTION 6: 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.

### Application Process

The application process for Technical Areas 1, 2 and 3 consists of two stages:

### Stage 1 - Concept paper

Concept Papers are designed to make the solicitation process as efficient as possible for applicants. By soliciting short concept papers (no more than three pages) ARIA reviewers are able to gauge the feasibility and relevance of the proposed project and give an initial indication of whether we think a full proposal would be competitive. Based on this feedback you can then decide whether you want to submit a full proposal. **If you miss the deadline for submission of concept papers you can still submit a full proposal.** You can find out more about ARIAs review process here.

To ensure the process is quick and open we do not require your organisation's consent prior to submission of a concept paper.

You can find guidance on what to include in a concept paper here.

Following review of concept papers applicants will either be encouraged or discouraged from submitting a full proposal. For more details on the evaluation criteria we'll use, click here.

### Stage 2 - Full proposals

This step requires you to submit a detailed proposal including:

- **Project & Technical information** to help us gain a detailed understanding of your proposal
- **Information about the team** to help us learn more about who will be doing the research, their expertise, and why you/the team are motivated to solve the problem
- **Administrative questions** to help ensure we are responsibly funding R&D. Questions relate to budgets, IP, potential COIs etc

You can find more detailed guidance on what to include in a full proposal here. **You can submit a full proposal even if you did not submit a concept paper.**

For more details on the evaluation criteria we'll use, click here.

### Non-UK applicants only

Our primary focus is on funding those who are based in the UK. However, funding will be awarded to organisations outside the UK if we believe it can boost the net impact of a programme in the UK. If you are a non-UK applicant, you must therefore outline any proposed plans or commitments that will contribute to the programme in the UK within the project's duration (note the maximum project duration is 4 years).

If you are successfully selected for an award subject to negotiations this proposal will form part of those negotiations and any resultant contract/grant.

More information on the evaluation criteria we will use to assess your answers can be found later in the document here.

If you are a non-UK applicant we have provided some additional guidance in our FAQs including available visa options.

## SECTION 7: Timelines

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

<table>
<tr><td><b>Applications open</b></td><td><b>12.03.24</b></td></tr>
<tr><td><b>Concept paper submission deadline</b></td><td><b>27.03.24 (12:00 GMT)</b></td></tr>
<tr><td><b>Concept paper review & notification of encouraged/not encouraged to submit full proposal sent</b></td><td><b>28.03.24 - 11.04.24</b></td></tr>
<tr><td colspan="2">At this stage and based on your concept paper, you will either be encouraged/ discouraged to submit a full proposal. If you receive feedback indicating that you are not encouraged to submit a full proposal you can still choose to submit a full proposal. You should note that this preliminary assessment/encouragement provides no guarantee of any full proposal being selected for award of funding.</td></tr>
<tr><td><b>Full proposal submission deadline</b></td><td><b>07.05.24 (12:00 BST)</b></td></tr>
<tr><td><b>Full proposal review</b></td><td><b>21.05.24</b></td></tr>
<tr><td colspan="2">If you are shortlisted following full proposal review, you will be invited to meet with the Programme Directors to discuss any critical questions/concerns prior to final selection — this discussion can happen virtually.</td></tr>
<tr><td><b>Successful/Unsuccessful applicants notified</b></td><td><b>07.06.24</b></td></tr>
<tr><td colspan="2">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 lead researcher within 10 working days of being notified.<br>We expect contract/grant signature to be no later than 8 weeks from successful/ unsuccessful notifications. During this period the following activity will take place:<br>● Due diligence will be carried out<br>● The PD and the applicant will discuss, negotiate and agree the project activities, milestones and budget details<br>● Agreement to the set Terms and Conditions of the Grant/Contract. You can find a copy of our funding agreements here</td></tr>
</table>

## SECTION 8: Evaluation Criteria

### Concept Paper and Proposal Evaluation Principles

To build a programme at ARIA, each Programme Director directs the review, selection, and funding of a portfolio of projects, whose collective aim is to unlock breakthroughs that impact society. As such, we empower Programme Directors to make robust selection decisions in service of their programme's objectives ensuring they justify their selection recommendations internally for consistency of process and fairness prior to final selection.

We take a criteria-led approach to evaluation, as such all proposals are evaluated against the criteria outlined below. We expect proposals to spike against our criteria and have different strengths and weaknesses. Expert technical reviewers (both internal and external to ARIA) evaluate proposals to provide independent views, stimulate discussion and inform decision-making. Final selection will be based on an assessment of the programme portfolio as a whole, its alignment with the overall programme goals and objectives and the diversity of applicants across the programme.

Further information on ARIAs proposal review process can be found here.

### Proposal evaluation process and criteria

Proposals will pass through an initial screening and compliance review to ensure proposals conform to the format guidance and they are within the scope of the solicitation. At this stage we will also carry out some checks to verify your identity, review any national security risks and check for any conflicts of interest. Prior to review of applications Programme Directors and all other reviewers are required to recuse themselves from decision making related to any party that represents a real or perceived conflict.

Where it is clear that a proposal is not compliant and/or outside the scope, these proposals will be rejected prior to a full review on the basis they are not compliant or non-eligible.

Proposals that pass through the initial screening and compliance review will then proceed to full review by the Programme Director and expert technical reviewers.

**In conducting a full review of the proposal we'll consider the following criteria:**

1) **Worth Shooting For** – The proposed project uniquely contributes to the overall portfolio of approaches needed to advance the programme goals and objectives. It has the potential to be transformative and/or address critical challenges within and/or meaningfully contribute to the programme thesis, metrics or measures.

2) **Differentiated** – The proposed approach is innovative and differentiated from commercial or emerging technologies being funded or developed elsewhere.

3) **Well defined** – The proposed project clearly identifies what R&D will be done to advance the programme thesis, metrics or measures, is feasible and supported by data and/or strong scientific rationale. The composition and planned coordination and management of the team is clearly defined and reasonable. Task descriptions and associated technical elements provided are complete and in a logical sequence with all proposed stage-gates and deliverables clearly defined.

4) **Responsible** – The proposal identifies major ethical, legal or regulatory risks and that planned mitigation efforts are clearly defined and feasible.

5) **Intrinsic motivation** – The individual or team proposed demonstrates deep problem knowledge, have advanced skills in the proposed area and shows intrinsic motivation to work on the project. The proposal brings together  disciplines from diverse backgrounds.

6) **Benefit to the UK – Applicable to non-UK applicants only.**

   There is a clear case for how the research will benefit the UK. Proposals originating from applicants outside the UK who seek to establish operations inside the UK, perform a majority of the research inside the UK and present a credible plan for achieving this within the programme duration will be deemed 'UK Applicants' (note this will be reflected in your contract terms).

   For all other non-UK applicants we will evaluate the proposal based on its potential to boost the net impact of the programme in the UK. When considering the benefit to the UK, the proposal will be considered on a portfolio basis and with regard to the next best alternative proposal (from a UK organisation/individual).

## SECTION 9: How to apply

Before submitting an application we strongly encourage you to read this call in full, as well as the general ARIA funding FAQs.

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

Clarification questions should be submitted no later than 4 days prior to the relevant 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. We'll also periodically publish questions and answers on our website, to keep up to date click here.

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.

Application Portal instructions

APPLY HERE

## SECTION 10: References

[1] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso,
"The Computational Limits of Deep Learning", arXiv.org, [cs, stat], Jul. 2020.
[2] ML Commons, MLPerf Training Benchmarks. Available at:
https://mlcommons.org/benchmarks/training/.
[3] Sandia Labs, CrossSim (2023), GitHub repository
https://github.com/sandialabs/cross-sim
[4] K U Leuven-MICAS, ZigZag (2023), GitHub repository,
https://github.com/KULeuven-MICAS/zigzag.
[5] IBM, AI HW Kit (2024) GitHub repository, https://github.com/IBM/aihwkit.
[6] Manipatruni, Sasikanth, et al. "Scalable energy-efficient magnetoelectric spin–orbit logic." Nature 565.7737 (2019): 35-42.
[7] Ankit, Aayush, et al. "Panther: A programmable architecture for neural network training harnessing energy-efficient ReRam." IEEE Transactions on Computers 69.8 (2020): 1128-1142.
[8] Cosemans, Stefan, et al. "Towards 10000TOPS/W DNN inference with analog in-memory computing–a circuit blueprint, device options and requirements." 2019 IEEE International Electron Devices Meeting (IEDM). IEEE (2019).

[9] Ceyhan, Ozgur. "Algorithmic Complexities in Backpropagation and Tropical Neural Networks." arXiv preprint arXiv:2101.00717 (2021).
[10] Martens, James, and Roger Grosse. "Optimising neural networks with kronecker-factored approximate curvature." International conference on machine learning. PMLR, (2015).
[11] Geshkovski, Borjan, et al. "A mathematical perspective on Transformers." arXiv preprint arXiv:2312.10794 (2023).
[12] Laborieux, Axel, and Friedemann Zenke. "Holomorphic equilibrium propagation computes exact gradients through finite size oscillations." Advances in Neural Information Processing Systems 35 (2022): 12950-12963.
[13] Aifer, Maxwell, et al. "Thermodynamic Linear Algebra." arXiv preprint arXiv:2308.05660 (2023).