Advanced Research + Invention Agency ARIA

# Unlocking AI compute hardware at 1/1000th the cost

## Programme thesis

v2.0

Suraj Bramhavar, Programme Director

### CONTEXT

This document presents a thesis underpinning a programme that has now launched. Sign up **here** to receive all updates about this opportunity space and see the programme derived from this space **here**.

An ARIA programme seeks to unlock a scientific or technical capability that

+ changes the perception of what's possible or valuable
+ has the potential to catalyse massive social and economic returns
+ is unlikely to be achieved without ARIA's intervention

---

### UPDATE: OUR THINKING, EVOLVED

*A summary capturing the evolution of our thinking since publication.*

Since publishing this thesis in November 2023 we have invited public feedback on our ideas, conducted workshops, and engaged with experts to challenge and refine our thinking. The following key learnings emerged during that process and have evolved our original approach, these have been incorporated into the programme structure:

+ **When pursuing new computing modalities, choosing a workload to focus on is critical.** In choosing a specific workload to focus a programme on, it quickly became clear that the vast majority of AI hardware research is focused on inference workloads (for both technological and economic reasons). However, the capabilities of AI, which dictate much of its societal benefit, are governed by our ability to train AI models. AI research is increasingly governed by the efficiency with which practitioners can access vast compute resources to train new models. Our discovery process surfaced hardware for AI training as both an underserved and important opportunity for funded research.
+ **Technology translation is governed as much by economics as it is by capabilities.** A stronger emphasis will continue to be placed on developing a better economic understanding of the existing state-of-the-art, and using these economics as constraints to spark creative R&D activity.
+ **Systems-level analysis is highly undervalued.** We need to allow R&D Creators, individuals and teams that ARIA will fund and support, to focus on a variety of novel building blocks (all of which are critical to reach our goals), but we also want to create the conditions such that they can track how their specific block fits into a larger system. Systems-level analysis allowing for this (via software simulation) clearly stands out as an underfunded activity, and a critical tool to help prioritise and direct R&D activity.

Having integrated these learnings and incorporated your feedback, we have now launched a live programme in this space: **Scaling Compute — AI at 1/1000th the cost**. Calls for proposals are currently open.

The rest of this thesis document provides an archival deepdive into our journey to this point.

## PROGRAMME THESIS, EXPLAINED

*A detailed description of the programme thesis, presented for constructive feedback, as published in November 2023.*

### Why this programme

The digital computing paradigm as we know it has been an incredibly unique economic engine. Few technologies in human history have provided compounding exponential improvements over fifty-plus years. This growth trajectory has now clearly ended, as traditional methods for improving hardware performance have become economically unviable[1].

Nature has evolved to process complex information in an entirely different way, and has proven incredibly efficient at doing so. Looking to nature provides us with an existence proof that **it is fundamentally possible** to accomplish sophisticated information processing much more efficiently than current computers.

Key principles that distinguish natural systems from existing computers include the facts that:

1. They typically do not distinguish computing elements from memory elements
2. They incorporate noise and are not purely binary
3. They do not operate in discrete time intervals

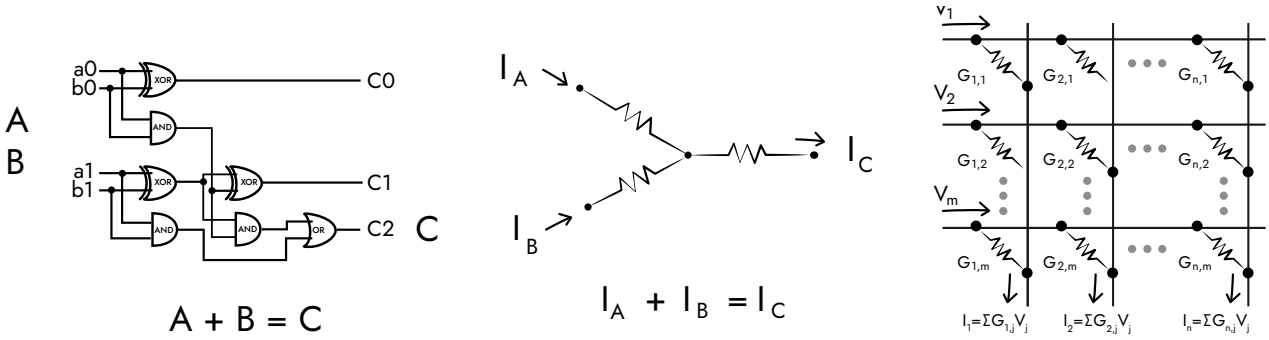
The field of neuromorphic computing was created in this vein, primarily adopting tenet #1 to demonstrate impressive performance[2]. One aim of this programme will be to show that embracing a combination of all three can take this work significantly further and help radically reduce the manufacturing and operational cost of today's AI hardware. AI workloads provide a compelling focal point due to their tolerance to the tenets listed above as well as strong commercial demand for increased performance. If successful, gains realised through this programme will ultimately be large enough to drive commercial adoption, and the techniques will subsequently find utility in a number of fields, ranging from communication systems to weather prediction, each of which relies heavily on advanced information processing.

### What we expect to fund

In order to catalyse meaningful progress, it is critical to pinpoint workloads that are narrow enough to focus the efforts of a research community yet valuable enough to warrant significant investment. Take, for example, the application of cryptographic code breaking as a galvanising force to build quantum computers. Modern AI presents an analogous opportunity, as a few key linear algebraic primitives serve as the bedrock for today's AI hardware infrastructure, and AI as a general workload exhibits such strong demand that it drives economic investment for much of the semiconductor industry.

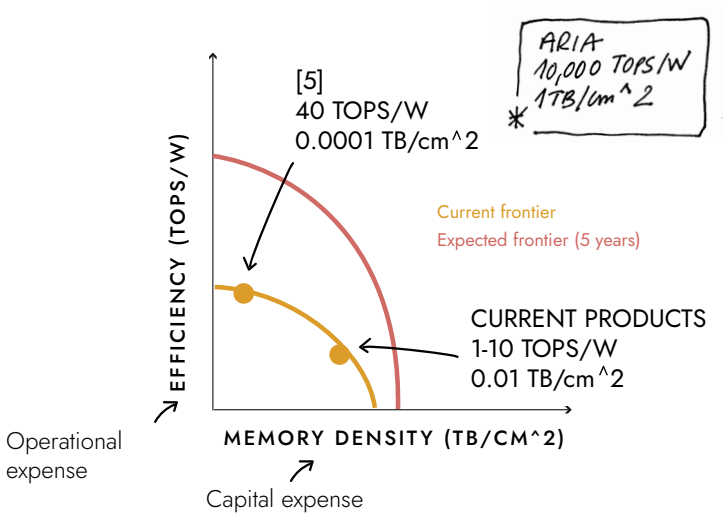
Below we identify three such primitives: **matrix multiplication**, **matrix inversion**, and **monte carlo sampling**, forming a representative (though not exhaustive) set.

As an illustrative example of 'nature's ability to perform computation', take a simple addition operation. A standard way to add two numbers involves a number of transistors tiled up using Boolean logic (shown below using two bits of precision). Alternatively, one could simply combine currents from two wires, and utilise Kirchoff's Law which states that the currents on the third wire will constitute the sum.




'Nature-based' computing example

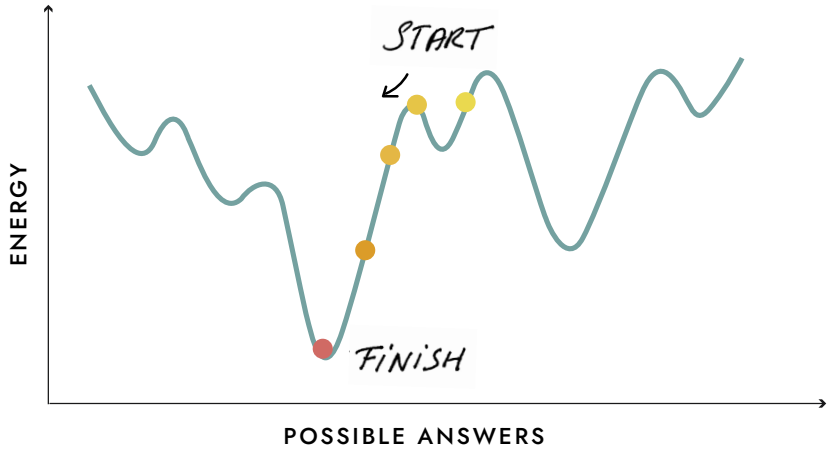
This idea can be taken a step further to perform **matrix multiplication** using a network of resistors and wires (utilising both Ohm's and Kirchoff's laws as shown). Matrix multiplication operations serve as the fundamental primitive for all current AI training and inference accelerators. Alternative physical substrates have been proposed for these kernels with projected efficiency gains exceeding 1000x over the state-of-the-art[3, 4], with the caveat being that the precision of computation is limited. Conventional wisdom has held that this reduced precision would restrict applicability at scale, but the digital AI accelerator community has recently shown that high levels of precision may not be necessary[5], raising the opportunity to challenge convention and realise significant systems-level benefits from low-power mixed-signal matrix multiplication.




Technology Target for AI accelerator, assumption → int8 operations

The energy/time requirements for these workloads are governed not just by the mathematical operations but also the resources required to get information to/from memory and to other processors. Technologies of interest would need to either a) address both the mathematical operation AND data transfer requirements, or b) operate at a large enough scale to effectively amortise the cost of shuttling information around.

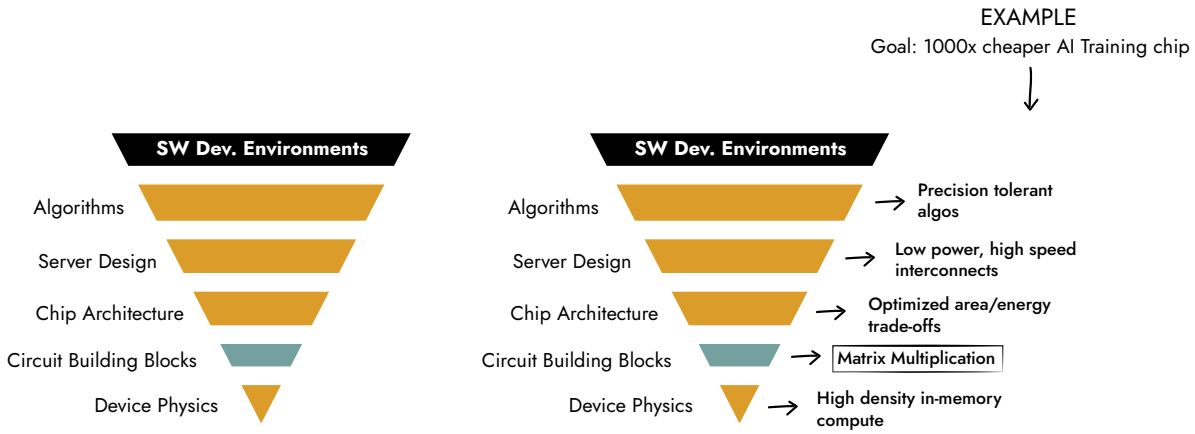
Similar to matrix multiplication, the ability to approximate **matrix inversion** represents another foundational computational kernel, applicable not only in AI but also in myriad other areas including next generation communication systems[6]. Recent proposals suggest that a variety of computational substrates are capable of exploiting energy-minimisation principles to efficiently compute matrix inverses[7, 8], but hardware demonstrations and comparisons against digital alternatives have yet to be shown.




Gradient descent, simulated annealing, and Lagrange multipliers are ALL examples of mathematical optimisation routines designed around finding 'low-energy' energy solutions.

The study of physical systems which perform computation on the basis of minimising energy bears close resemblance to the process of training a large neural network, which itself entails finding specific weights and biases that 'minimise energy' of a simulated objective function (see figure above). The standard way of doing this (stochastic gradient descent and the backpropagation algorithm) has become ubiquitous, but its remarkable performance should not preclude alternatives. It is not yet known whether the brain performs backpropagation, and other under-explored algorithms may exist which perform equally well or better. Examples of these alternative energy-based training algorithms have begun to emerge from the neuroscience community[9, 10, 11], but they have yet to demonstrate their capabilities at scale (and on modern transformer network architectures). This gap is narrowing with time, and could benefit from a large concerted effort to prove (or disprove) merit. Alternative algorithms will offer an entirely new design space for computer architects to exploit in designing new hardware.

Monte Carlo methods represent yet another critical computational primitive underpinning a wide variety of scientific disciplines, from materials discovery to protein folding to weather prediction[12]. The concept of intelligent sampling is again primarily targeted at finding 'low-energy' solutions to complex landscapes, usually in cases where the energy surface is not continuous. Noise plays a critical role in helping sample the widest possible solution space, but this noise is typically introduced artificially. Meanwhile, noise is inherent in every physical system but is rarely utilised as a computational tool, providing an opportunity for large performance improvements.




*Example technology stack*

Many layers exist in the technology stack separating end users from the underlying hardware (shown above). The programme will identify key primitives (dark block in the figure), set ambitious system-level goals which rely heavily upon those primitives, and pursue research in any of the highlighted areas which are best suited to achieve those goals. Take as an example the system-level goal defined in the 'Technology Target' plot above (with matrix multiplication representing the kernel of interest). In this particular case, meeting these targets may require some combination of efforts from research foundries developing novel high-density memory technology, design teams to develop circuit IP, systems engineers to build a scalable/usable server, and algorithms researchers to contribute precision-tolerant AI models. A depiction of different possible technologies for this are shown in in the figure above. Each effort would be evaluated based on its own isolated metrics, but the overall goal would require integration between the layers.

**How we expect to fund**

Regardless of the kernel chosen, well-defined sets of benchmarks will be required to evaluate performance. Teams would be charged with showing that, for a set of target workloads, specific performance/cost metrics can be met representing a radical step-change over the projected state-of-the-art.

Examples of interesting research pursuits include (but are not limited to):

+ Mixed-signal in-memory computational architectures

+ Memory technologies with ultra high density, low-power, or improved input/output

+ Novel quantization or number representation techniques for AI training

+ Low-energy chip-chip communication technologies

+ AI training algorithms utilising energy-based models

+ Probabilistic programming and hardware

+ Optical (or other physics-based) matrix multipliers

+ Neuromorphic architectures or algorithms which break away from purely digital electronics

+ Repurposed 'noisy' quantum technologies for fast/efficient sampling in hybrid architectures

Two distinct types of awards are envisioned:

+ Bold **solutions**: Granted to coordinated teams building integrated systems which can be externally evaluated against target specifications. Evaluation will consider key factors we believe will influence future commercial ecosystem viability (including simulation/design software, licensable circuit IP, compiler technology, scalability, cost, and manufacturability)
+ Bold **ideas**: Granted to radical ideas at the proof of concept stage with little regard for such factors

Creators (recipients of ARIA research funding) will be expected to participate in shared reviews to understand critical roadblocks, discuss progress, and brainstorm paths toward common programme goals.

### What we are still trying to figure out

A number of questions remain in shaping this programme, including:

+ Which workloads do we focus on, and are there other kernels meeting the criteria above we should consider?
+ Are we projecting the right performance targets, and are we sure our targets (e.g. TOPS-W target above) represent a big enough step-change to where industry will be in 3-5 years time?
+ What do we benchmark against, besides MLPerf?
+ Do we have accurate estimates of costs at industrial scale?
+ Are we missing important factors that govern technology translation (other than improved technical specifications)?
+ What are the right funding levels for the two categories of awards described above?

---

## SOURCES

*References cited in this document.*


[1] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The Computational Limits of Deep Learning", arXiv.org, [cs, stat], Jul. 2020, DOI: https://arxiv.org/abs/2007.05558.

[2] D. S. Modha et al., "Neural inference at the frontier of energy, space, and time", Science, vol. 382, no. 6668, pp. 329–335, Oct. 2023, DOI: https://doi.org/10.1126/science.adh1174.

[3] S. Cosemans 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), San Francisco, CA, USA, 2019, pp. 22.2.1-22.2.4, DOI: 10.1109/IEDM19573.2019.8993599.

[4] M. G. Anderson, S.-Y. Ma, T. Wang, L. G. Wright, and P. L. McMahon, "Optical Transformers", arXiv.org, Feb. 20, 2023, DOI: https://arxiv.org/abs/2302.10360.

[5] B. Keller et al., "A 17–95.6 TOPS/W Deep Learning Inference Accelerator with Per-Vector Scaled 4-bit Quantization for Transformers in 5nm", 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 16-17, DOI: 10.1109/VLSITechnologyandCir46769.2022.9830277.

[6] Min Soo Kim, S. Mandrà, D. Venturelli, and K. Jamieson, "Physics-inspired heuristics for soft MIMO detection in 5G new radio and beyond", (Cornell University), Sep. 2021, DOI: https://doi.org/10.1145/3447993.3448619.

[7] M. Aifer et al., "Thermodynamic Linear Algebra", arXiv.org, Aug. 10, 2023. DOI: https://arxiv.orgabs/2308.05660.

[8] S. K. Vadlamani, T. P. Xiao, and E. Yablonovitch, "Physics successfully implements Lagrange multiplier optimization", Proceedings of the National Academy of Sciences, vol. 117, no. 43, pp. 26639–26650, Oct. 2020, DOI: https://doi.org/10.1073/pnas.2015192117.

[9] B. Scellier and Y. Bengio, "Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation", Frontiers in Computational Neuroscience, vol. 11, May 2017, DOI: https://doi.org/10.3389/fncom.2017.00024.

[10] G. Hinton, "The Forward-Forward Algorithm: Some Preliminary Investigations", arXiv.org, [cs], Dec. 2022, DOI: https://arxiv.org/abs/2212.13345.

[11] B. Millidge, Y. Song, T. Salvatori, T. Lukasiewicz, and R. Bogacz, "Backpropagation at the Infinitesimal Inference Limit of Energy-Based Models: Unifying Predictive Coding, Equilibrium Propagation, and Contrastive Hebbian Learning", arXiv.org, Aug. 03, 2022, DOI: https://arxiv.org/abs/2206.02629.

[12] H. G. Katzgraber, "Introduction to Monte Carlo Methods", arXiv.org, [cond-mat, physics:physics], May 2011, DOI: https://arxiv.org/abs/0905.1629.