



AI KNOWLEDGE
CONSORTIUM

Locating Equitable Compute for AI

WORKSHOP REPORT AND RESEARCH AGENDA

APRIL 2024

About this Brief

The AI Knowledge Consortium's (AIKC) mission is to elevate AI governance in India, ensuring that the technological revolution triggered by it is inclusive, equitable, and reflective of shared values and aspirations. Towards this, the AIKC shall generate research questions based on multi-stakeholder discussions.

The first such discussion was held on March 7, 2024, and was titled *Beyond Moore's Law: Locating Equitable Compute for AI*. The workshop brought together a diverse mix of stakeholders, featuring leaders from prominent AI firms, start-ups, developers, and practitioners from the AI community, alongside AIKC members.

Introduction


The surge in computing power demands for AI especially with the rise deep learning, underscores the urgent necessity for research and dialogue on optimising compute amidst diverse resource constraints. These include the substantial economic and environmental costs of training large AI models, semiconductor supply limitations, and the concentrated availability of compute resources globally.


India has recently initiated efforts to make publicly funded compute infrastructure available to researchers and innovators. Yet, there remains an absence of public dialogue on the ideal design of such infrastructure to ensure equitable access and sustainable utilisation. Against this backdrop, the inaugural AIKC workshop gathered experts from civil society, government and the private sector to delve into the present state and future needs of compute infrastructure for AI in the country.

The inaugural workshop was structured around three key themes. Firstly it delved into the economic and environmental trade-offs that stakeholders must consider when discussing the design of publicly funded compute infrastructure. This is particularly relevant in light of the widespread attention on large AI models in recent years. Secondly, the workshop explored the various models of shared infrastructure, through comparative assessments. Lastly, discussions centred on legal considerations including access controls, user rights preservation and reforms in public procurement practices.


This workshop summary, compiled by the AIKC Secretariat serves a dual role; it acts as an archive of multi-stakeholder discussions hosted by the AIKC and also outlines key research questions for consideration by AIKC members moving forward. We aim for the preservation of these dialogues to facilitate knowledge sharing, ensuring our collaborative research agenda remains informed, responsive to demand and tailored to context specific contexts


Theme 1: Economic and Environmental Trade Offs


 The exponential demand for computing power in AI, especially with the rise of deep learning since the 2012 launch of ImageNet, has sparked discussions on optimising compute. The compute requirement for the largest AI training models, measured in the number of Floating Point Operations per Second (FLOPs), has doubled every 3.4 months since. Conversely, Moore's Law predicts doubling of transistors on integrated circuits every two years. The availability of compute for developing deep-learning based AI models, including Large Language Models (LLMs) like ChatGPT is undoubtedly important. But resource-scarce developing countries will initially need to confront associated economic and environmental trade-offs.





 Experts debated India's approach to AI-ready compute, discussing capital (capex) model – such as under India's National Supercomputing Mission (NSM), a centralised supercomputing infrastructure – versus the operational expenditure (opex) model – subsidisation of access to scalable and flexible cloud services – to enable researchers, innovators and entrepreneurs.

Historically, countries set aside large capex budgets for high performance computing requirements, in university campuses or scientific facilities, for instance. But this is changing with the availability of low-cost cloud rentals such as Amazon Web Service's Elastic Compute Cloud, Microsoft Azure's N-Series Virtual Machines, and Google Cloud which provide a range of prices and options.

 India has only a few supercomputers ready for GPU-intensive AI. The country has set aside INR 10,372 crore to build specialised compute infrastructure under the India AI Mission . Most of India's existing supercomputing capacity is built on Centralised Processing Units (CPUs).

 Supply-side compute capacities are growing alongside the demand for compute-intensive AI applications. However, the cost of computation in gigaFLOPS has not decreased since 2017, despite rising demand and cloud GPU costs have remained steady.¹ This is due to supply-side shortages in GPU production and incremental computational efficiency gains with each new GPU or CPU chip. This is why subsidisation, through capex or opex remains relevant despite underlying market dynamics.





 Experts find it beneficial to evaluate the need for AI compute based on its end-use applications which can be divided into those for pure science or natural science, applied science, and commercial use. Natural language processing (NLP), a subfield of computer science and linguistics, is an example of an applied science model with commercial applications. NLP stands as the predominant use-case for Param Siddhi, India's fastest supercomputer.² However, the demands for compute in pure sciences like computing possible organic chemicals exceeding 10^{30} can far surpass those in NLP by several orders of magnitude.³




-  During our discussion, academic experts suggested implementing decentralised GPU clusters within institutions could cater to most of their compute needs. However, they also stressed the significance of access to cutting-edge supercomputing resources for highly specialised compute-intensive projects capable of driving significant scientific advancements. Further, the discussion highlighted the importance of ensuring security and sovereignty of compute infrastructure for such use-cases, shedding light on associated legal considerations visited later in this report.
-  India's optimal approach to public funding for AI likely requires a hybrid strategy, consisting of (a) subsidising access to cloud services for the bulk of commercial and research use-cases, (b) ramping up centralised AI-ready supercomputing facilities for strategically vital minority of use-cases, particularly linked to pure sciences research and (c) establishing decentralised GPU clusters to cater to the majority of academic needs. Even so, a comprehensive mapping of compute requirements based on specific use-cases remains an urgent priority for the research community to address.
-  The environmental cost of compute is a crucial dimension to consider alongside specific use-cases. Global AI demand, for instance, are estimated to require between 4.2 to 6.6 billion cubic metre of water, by 2026, exceeding Denmark's total water withdrawal.⁴ Similarly, the energy consumption of generative AI driven searches is about four to five times higher than that of conventional web searches.⁵ Experts present at the discussion emphasised the need to strike a balance between AI model accuracy and energy efficiency. A great deal of AI development has focussed on incremental improvements in accuracy at the cost of the environment. This involves addressing questions regarding the environmental costs associated with legal compliance and the potential liabilities of inaccurate AI in the future.
-  Finally, experts also discussed the need to invest in talent, as a separate and distinct challenge from the development of AI compute. The Department of Science and Technology reports that around "5930 expert users from 100+ institutes are using the NSM facilities routinely".⁶ However, this user base remains relatively small compared to countries like Canada where 16,000 researchers access Compute Canada's resources.⁷ The increasing importance of algorithms in optimising scarce compute resources and enhancing energy efficiency also demands attention.

| Questions for Future Research

- Can end-use applications of AI compute be surveyed and categorised based on tiered needs in the Indian context, considering the balance between pure science, applied science, commercial use?
- What are the advantages and disadvantages of decentralised versus centralised supercomputing in terms of accessibility, security, sovereignty, and environmental sustainability for tiered AI use-cases in India?

Theme 2: Design of Shared Infrastructure




-  Experts unanimously supported the establishment of a shared national computing infrastructure, akin to that envisioned in the India AI Mission. Such a framework would seamlessly connect researchers and innovators providing access to essential resources such as data, models, software, and training resources needed for AI advancement. They emphasised the role of strategic partnerships between the public and private sectors, recognising the potential to leverage the combined strengths of both realms in this endeavour.
-  The public sector offers vital support through funding and access to centralised or vertical computing infrastructure, while the private sector brings flexibility with decentralised or horizontal infrastructure and a pool of skilled talent. This synergy distributes risks and responsibilities, fostering a more sustainable approach to the development and management of compute facilities. Moreover, experts spoke about the creation of a tiered compute architecture to accommodate diverse computational demands, ranging from advanced applications to basic research needs.
-  Public access models in countries like Canada and Taiwan offer valuable insights on tiered compute systems. Canada's Advanced Research Computing Expansion Program, initiated in 2019, and Compute Canada, a cornerstone of the nation's AI strategy, offer researchers access to advanced compute infrastructure, active storage and backup, support services and training, system software, commonly used libraries, privacy and security measures, and high-speed connections. Taiwan Computing Cloud (TWCC), run by the National Centre for High-performance Computing, offers a range of standardised compute services, including container compute services, HPC, virtual compute services and cloud storage. TWCC's high-speed compute capabilities powered by the Taiwania2 supercomputer support over 200 businesses, many of which are start-ups.⁸
-  Similarly, America's National Artificial Intelligence Research Resource (NAIRR), launched as a two-year pilot in 2024, aims to define an optimal design architecture for public compute infrastructure.⁹ Spearheaded by the U.S. National Science Foundation (NSF), NAIRR operates in collaboration with 10 federal agencies and 25 non-governmental entities, including private leaders such as AWS, Anthropic, AMD, Open AI, HP, IBM, Nvidia and civil society organisations. NAIRR includes both on-premise and commercial cloud platforms, featuring dedicated and shared resources equipped with various CPU and GPU options, multiple accelerators per node, high-speed networking, and at least one terabyte of memory capacity per node. The initiative mandates the presence of at least one AI supercomputer capable of training a trillion models, a goal that may be realised by repurposing an existing supercomputer, akin to TWCC, or acquiring a new one through a competitive bid process.


-  Another international lesson learned is the imperative of raising awareness about the availability of public compute infrastructure. Despite Canada's outstanding public compute infrastructure framework, many researchers are unaware of its existence or mistakenly believe it is not accessible to them, or use other non-public compute options.¹⁰ To address this, Canada continually surveys the research community to identify gaps in its infrastructure provision. Similarly, America's NAIRR includes a survey component to survey component aimed at gaining deeper insights into current and anticipated use-cases better. Multistakeholder groups such as AIKC may have a significant role to play here, in the Indian context.
-  Experts also suggested the growth in demand for GPU computing globally and in India, but emphasised the risks of overprovisioning, particularly in academic settings. Many workflows and applications in such settings may not fully capitalise on GPUs leading to underutilisation of public compute resources, or suboptimal utilisation. They suggested that conducting a survey of usage trends within research institutions could be beneficial, particularly to determine the right mix of compute – on-premise, off-premise and hybrid – tailored to different use-cases.
-  Additionally, development of AI compute infrastructure employing both, centralised and decentralised networks highlights the need for holistic design. Experts felt communications linkages were important in this regard. While India's National Knowledge Network (NKN) aims to interconnect all knowledge and research institutions nationwide through a high-speed data communication network, its impact on the research and innovation ecosystem remains unmeasured. Moreover, the involvement of the private sector in providing connectivity between computational nodes and decentralised users is imperative, given the excess bandwidth available with 5G providers.


| Questions for Future Research

- What strategies can be employed to prevent overprovisioning / underutilisation of AI compute, and how can usage trends help determine the optimal mix of compute resources?
- How can awareness and accessibility of public compute infrastructure be improved among researchers and innovators in India, and what role can multistakeholder groups like AIKC play in this?
- How can the private sector effectively contribute to the design, development, and management of public compute infrastructure for AI advancement in India, and what global best-practices serve as guidance in this context?


Theme 3: Legal and Governance Issues

-  Legal experts at the AIKC workshop highlighted concerns regarding public procurement limitations in India's General Financial Rules. They agreed that existing procurement models focussing on cost-competitiveness and defined outcomes may not suit the needs of AI infrastructure. This calls for innovative approaches and a strategic re-evaluation of the public interest in such infrastructure.
-  While global discussions are underway on safeguards and guidelines for AI procurement at the applications-level, there is little to guide the development of underlying compute infrastructure. For instance the IEEE P3119™ Standard for the Procurement of Artificial Intelligence and Automated Decision Systems, aims to assist stakeholders in making “meaningful and accountable choices that are transparent about the socio-technical considerations and impact of AI products, services, and or systems on the public”.¹¹ However, these do not directly address concerns linked to infrastructural innovation which is at the core of supercomputing.¹²
-  It is challenging to predetermine the peak performance of cutting-edge compute infrastructure through requests for proposals. For instance, the world's fastest supercomputer, Frontier, saw an increase in its High-Performance Linpack (HPL) score, which measures the FLOPs needed for solving linear equations, from 1.02 exaflops in November 2022 to 1.194 exaflops in May 2023.¹³ This increment alone surpasses the performance of most supercomputers on the top 500 list. Similarly, energy efficiency is an ongoing area of improvement in compute systems. The most efficient supercomputer in terms of GigaFlops per Watt for instance, is ranked 293rd in terms of HPL score on the top 500 list.¹⁴
-  A public procurement style where parameters are predetermined will not effectively optimise compute performance or efficiency. The evolving landscape of AI presents an opportunity to revisit public procurement practices in India, a topic the AIKC intends to explore through its research efforts. Even though exceptions to lowest bidder or equipment indigenisation mandates can easily be accommodated for specialised procurement such as in the case of software, there is a need to incentivise performance. Experts also debated the merits of a competitive procurement model which could involve small businesses. Research into other agile procurement processes such as stages where the procurer can engage the market, or require proofs of concept as tests before final purchases are also warranted.
-  The State's role in creating or enabling the use of shared compute infrastructure extends beyond its initial availability. Achieving welfare optimised shared use is an ongoing process, which involves legal, regulatory and contractual considerations similar to other infrastructure projects. Experts discussed this in the context of managing rights and ownership in shared AI infrastructure. The potential for fractured ownership obviates the need for establishing clear contractual standards to safeguard user interests.

 Experts discussed the critical need for safeguarding individual ownership rights over their data and algorithms, as well as their privacy, through robust legal frameworks against potential abuses. They described hypothetical scenarios to illustrate the potential harms individuals and firms could face if these rights are left unprotected. For instance, without adequate protection, privately organised data and algorithms on public infrastructure – whether in transit or at rest – could be vulnerable to State expropriation. The creation or development of AI models and user-ready applications involves significant effort and creativity, making the intellectual property (IP) within them important to protect. This is harder to do when the infrastructure owner also regulates its use. Therefore, experts discussed the need to separate the ownership and regulation of compute infrastructure.

 Similarly, there is the issue of continuity of users' rights (including legal persons) vested in data or algorithms stored on shared compute infrastructure managed or owned by an entity that becomes insolvent or ceases to exist. It is also unclear what happens when hardware in public infrastructure fails. Therefore, strong legal and contractual protections are required to ensure that researchers and innovators have incentives to utilise shared infrastructure.

For instance, warranties covering hardware for defined periods, structured backup and restore options, insurance coverage, and clear data usage and transfer agreements are some of the best-practices in this regard. The India AI Mission offers an opportunity to design and implement these protections. Experts also discussed the responsible use of shared compute infrastructure, particularly when it is used for tasks that have a national security implications.

 Indian enterprises must also look beyond traditional security controls, in light of increasing threats of AI privacy breaches and security incidents. For example, companies like NVIDIA are developing solutions¹⁵ to help AI firms secure their IP from unauthorised access or control, even in cases where the physical infrastructure is breached. Secure data enclaves such as Intel's SGX¹⁶ can provide a trusted execution environment, ensuring that sensitive information remains protected against unauthorised access or tampering.¹⁷ These enclaves provide enhanced protections and technical controls for data owners compared to data sharing agreements.¹⁸

| Questions for Future Research

- How can public procurement models in India be innovated to suit the evolving needs of AI infrastructure development, ensuring cost-effectiveness, performance optimisation, and alignment with public interest?
- What specific legal and contractual frameworks are necessary to manage rights, ownership, and continuity of use in shared AI infrastructure, ensuring protection of individual and organisational interests?

Endnotes

- 1 John, Andrew, and Musser, Micah, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?", Centre for Security and Emerging Technology, January 2022
- 2 95% of Param Siddhi's utilisation as per the Annual Report on National Param Supercomputing Systems, issued by the Centre for Development of Advanced Computing in 2021.
- 3 Shankar, Sadasivan and Reuther, Albert, "Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers and Compute-Intensive Applications", paper submitted to the proceedings of IEEE HPEC, 2022
- 4 Jianyi, Pengfei et al, "Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models", arXiv:2304.03271, 2023
- 5 Crawford, Kate, "Generative AI's environmental costs are soaring — and mostly secret", Nature, February, 2024
- 6 <https://dst.gov.in/national-super-computing-mission>
- 7 https://alliancecan.ca/sites/default/files/2022-03/arc_current_state_report_0.pdf
- 8 <https://www.twcc.ai/news>
- 9 <https://new.nsf.gov/focus-areas/artificial-intelligence/nairr>
- 10 https://alliancecan.ca/sites/default/files/2022-03/arc_current_state_report_0.pdf
- 11 <https://standards.ieee.org/beyond-standards/topic/ieeesa-news/process-model-and-requirements-aimed-at-ai-procurement-in-a-new-ieee-standard/>
- 12 Similarly, private international organisations such as the Centre for Inclusive Change have released supplementary literature which can guide governmental departments through the process of procuring AI systems. However, these guidelines mostly focus on AI systems and not specifically on infrastructure; adapting them (especially to the Indian context) would require work.
- 13 <https://www.hpcwire.com/2023/05/22/exascale-frontier-supercomputer-has-passed-formal-acceptance-what-that-means/>
- 14 <https://www.top500.org/lists/green500/2023/11/>
- 15 <https://www.nvidia.com/en-in/data-center/solutions/confidential-computing/>.
- 16 <https://www.intel.com/content/www/us/en/developer/tools/software-guard/extensions/overview.html>.
- 17 Howison et al, Protecting Sensitive Data with Secure Data Enclaves, <https://dl.acm.org/doi/pdf/10.1145/3643686>.
- 18 Howison et al, Protecting Sensitive Data with Secure Data Enclaves, <https://dl.acm.org/doi/pdf/10.1145/3643686>.

Secretariat



Koan Advisory Group, a New Delhi-based public policy consulting firm provides secretarial support to the AIKC

Contact Us

For inquiries, partnerships, or to learn more about our work with the AIKC, please write to:

Secretariat@aiknowledgeconsortium.com

Address: B40, Soami Nagar, New Delhi, 110017

