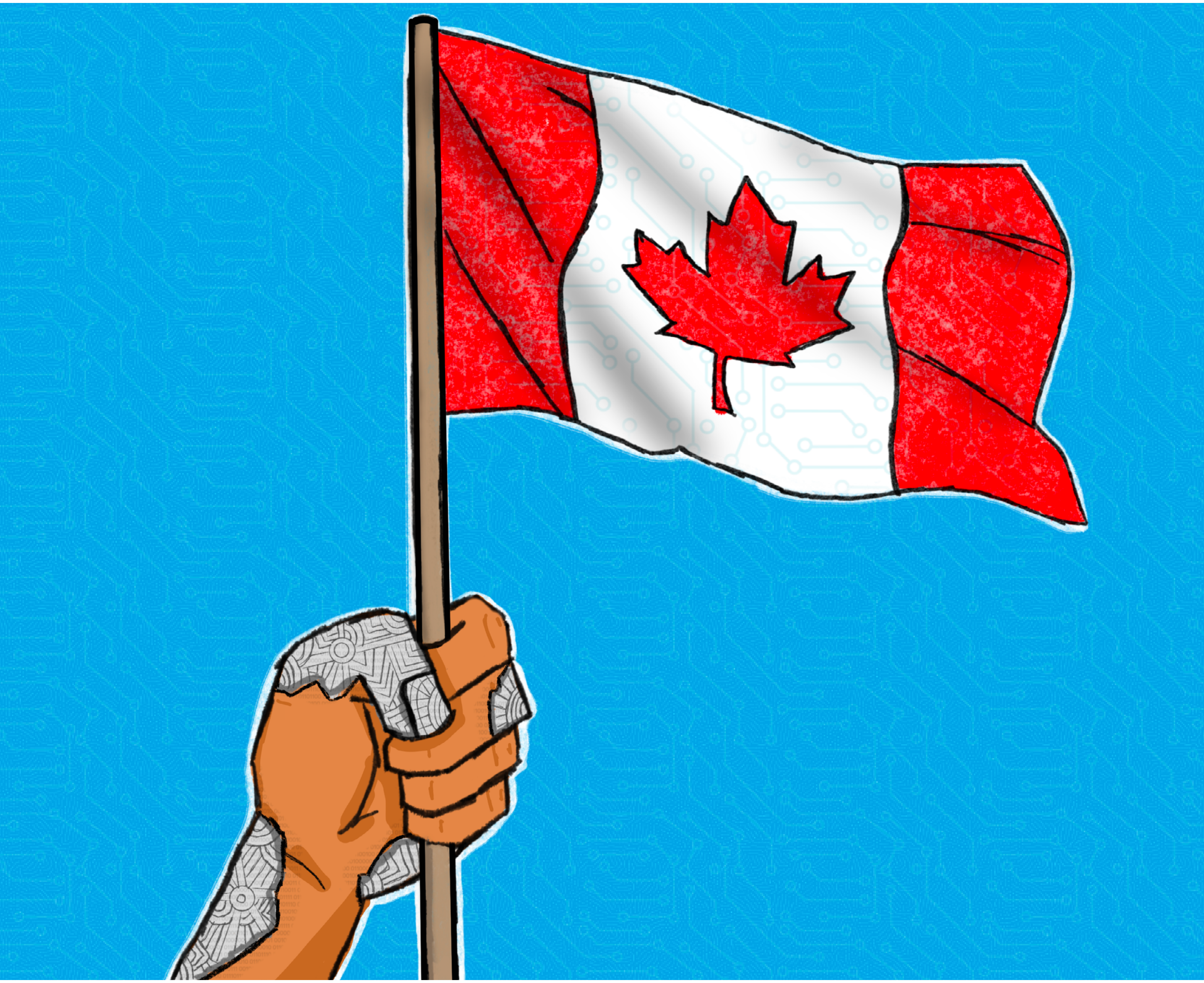
Can Canada Compute?

Policy Options to Close Canada's AI Compute Gap

Graham Dobbs and Jake Hirsch-Allen | March 2024



Toronto Metropolitan University

the dais

Acknowledgements

The Dais is Canada's platform for bold policies and better leaders. We are a public policy and leadership think tank at Toronto Metropolitan University, connecting people to the ideas and power we need to build a more inclusive, innovative, prosperous Canada.

For more information, visit dais.ca
20 Dundas St. W, Suite 921, Toronto, ON M5G 2C2

@daisTMU

How to Cite this Report

Graham Dobbs and Jake Hirsch-Allen, "Can Canada Compute? Policy Options to Close Canada's AI Compute Gap.", the Dais, March 2024.
https://dais.ca

ISBN 978-1-77417-086-1

The Dais proudly engages a diverse group of funders to support and catalyze our work, consistent with our values, and subject to a thorough internal review. As a non-partisan, public-interest institute, we only accept funds from organizations that support our mission and enable us to undertake work independently, with full editorial control. The names of all of our financial supporters are publicly and transparently displayed on all online and printed material for each project or initiative.

Design and Illustration
Mariana Rodrigues

Copy-Editing
Suzanne Bowness, CodeWord Communications

Contributors
Nina Rafeek Dow
André Côté
Viet Vu
Karim Bardeesy

Authors

Graham Dobbs
Senior Economist

As a senior economist, Graham Dobbs (he/him) is a part of the research team at the Dais at TMU. Graham explores educational and technological innovations in the Canadian labour force and its impact on occupational distributions and transitions.

In the past, Graham developed tools and insights designed to make labour market information more accessible and navigable for Canadians. His previous research topics include following taxable earnings of post-secondary graduates and Red Seal journeypersons to provide credible wage information to students, mapping the access of career service use among Canadian adults, and broadening access to skills and job demand through online job-posting dashboards.

Jake Hirsch-Allen
Senior Advisor

Jake Hirsch-Allen is a senior advisor at the Dais, focusing on inclusive technology and skills. He builds public-private partnerships in workforce development and higher education for LinkedIn. The views in this paper are his own and those of the Dais, but not of LinkedIn or Microsoft.

Jake advises impact investors, public sector leaders and start-ups, including on responsible tech and ethical AI. He is a director on the boards of Ontario Tech Talent and the Canadian Club and founded Lighthouse Labs. A former intellectual property and international criminal lawyer, Jake was also chair of the Technology Committee of the Global Education Platform, taught Global Health at McMaster University and clerked at the Supreme Court of Israel.

The Dais is a policy and leadership think tank at Toronto Metropolitan University, dedicated to taking a tech-first approach to some of the most challenging economic, education and democracy policy challenges that can promote shared prosperity and citizenship for Canada. This report, which draws on the Dais' expertise in AI policy, was initiated in mid-December 2023, at the suggestion of our advisory board and after many conversations with industry and policy leaders about the need for an urgent national public policy response to the issue of scarce domestic AI Compute capacity in Canada.

Table of Contents

1 Executive Summary

Canada is well positioned to play a leading role in the future of artificial intelligence (AI). In 2017, Canada was the first nation to launch a national AI strategy, establishing itself as a home for trailblazing academic researchers. Other standout features include a strong AI start-up ecosystem, and one of the fastest-growing and most skilled AI communities worldwide. By embracing broader adoption, the country could leverage AI to remedy its longstanding productivity challenges.[1]

However, there is growing concern that despite possessing the talent and the algorithms, Canada faces a shortage of the third critical input in advancing national AI ambitions: AI Compute (AIC) infrastructure (the physical infrastructure including servers and computing processing units to train and develop AI systems). Compared to its G7 peers, Canada holds the lowest amount of publicly-available computing infrastructure and performance. Canada's current computing performance potential is half that of the United Kingdom's, the next lowest G7 nation. In a highly competitive global environment, this AI Compute gap could hinder Canadian innovators from growing successful AI firms, slow down businesses' adoption of AI technologies, and limit researchers' advancement of the next generation of AI inquiry. This gap poses a risk to Canada's economic and data sovereignty in a changing global landscape.

Still, the resources required are significant to immense, and benefits for the broader economy, let alone society, remain speculative. To quote a recent critic of a US public AIC investment strategy: "We need to know that these investments will meaningfully benefit society at large, broadening the horizon for innovation in ways that will accrue to the many and not just the few."[2]

Compared to its G7 peers, Canada holds the lowest amount of publicly-available computing infrastructure and performance.

Through a review of secondary research evidence, and extensive off-the record discussions with experts on computing capacity and AI ecosystem in Canada, this policy options report:

- **sets the context for AI in Canada and why AI Compute capacity matters;**
- **summarizes the state of AI Compute in Canada** in relation to peer jurisdictions, as analyzed through the lens of the OECD's national AIC framework; and
- **outlines three scenarios and public policy-informed approaches**, including the resulting trade-offs, for addressing and bridging Canada's AI Compute gap.

This report provides key insights about the current state of Canada's AIC in three major areas:

- **AIC capacity:** Canada lags behind all other G7 countries in AIC infrastructure. Its available AIC infrastructure is insufficient for frontier research.
- **AIC effectiveness:** Canada primarily uses its AIC capacity for research, while private companies have to rely on hyperscaler (large tech multinationals such as Amazon and Microsoft) cloud-based solutions outside of Canada. Price-related pressures restrict their use.
- **AIC resilience:** in contrast to many peer jurisdictions, Canadian public policy relies on broad programs that encourage the adoption of technology in general, with no policies specifically targeting AIC access.

We presume that the Government of Canada and other AI ecosystem stakeholders will want to embrace this public policy objective to grow access to AIC, close the AI compute gap, and therefore help retain AI firms and talent and promote greater AI development and adoption. In turn, this objective can promote a larger public policy goal of greater national productivity and competitiveness. Like the Internet before it, AI capacity could power a wide variety of areas of economic growth that require immediate, medium, and long-term solutions. AIC has the potential to be treated as a utility-like infrastructure.

With this background in mind, we present three policy scenarios as possible approaches:

- **Scenario 1 | Centralize and subsidize AIC via federal procurement or direct subsidies from existing enterprise cloud computing providers**
- **Scenario 2 | Work with key trade partners to jointly purchase and access AIC infrastructure and cloud computing at scale**
- **Scenario 3 | Build domestic AI supercomputing capacity via first- or third-party vendors in multi-year, multi-vendor partnerships**

The network of Canadian actors requiring AI Compute access includes researchers, not-for-profits, governments, AI-focused firms training AI models, and companies that need access to computing power to run AI applications. While we need AI Compute for these players, we can either build such capacity domestically, or facilitate accessing international capacity, as domestic AI Compute in Canada simply won't be available in sufficient quantities on the time scale that users need access. So, we need near- (next six months) and medium-term (next two years) strategies to open access up, while we work on achieving more AI Compute in Canada.

Even in the long term, AI Compute sovereignty may not be fully achievable, and it may not be worth the cost—we will continue to be interdependent with supply chains and providers housed outside Canada. If market access can be secured, private sector players in particular should expect to continue to look outside as well as within Canada for AI Compute access.

The technology needed for AI Compute, and the needs for future models and future applications of AI, could look very different to what we have today. In every possible case, Canada needs to act now to regain some of its standing, while being nimble and ready to adapt to changing circumstances.

BOLD IDEA

Canada needs to urgently tackle the AI Compute gap to retain and grow our business and talent investments, and improve productivity.

2 AI in Canada and Why Compute Matters

Canada is recognized internationally for its contribution to Artificial Intelligence (AI), as it was among the early jurisdictions to introduce legislation, and to establish a national policy and regulatory framework for AI, with efforts now underway to develop internationally-aligned industry standards for AI. Canada was also the first to launch a national plan for AI development, the Pan-Canadian AI Strategy (PCAIS), led by the Canadian Institute for Advanced Research (CIFAR). These and other initiatives have contributed to Canada's position as a leader in AI research, talent, and start-up growth.

The country is home to trailblazers in the AI field, notably Geoffrey Hinton and Yoshua Bengio, who worked extensively on neural network algorithms,[3] and Richard Sutton in the field of reinforcement learning. These names are among the most high-profile in a robust ecosystem of researchers, which leads in AI research publications per capita among G7 nations as of May 2023.[4] Canada's domestic AI workforce is estimated to be between 200,000 and 220,000, composed of data science, software engineering, and development professionals, or one percent of Canada's workforce in 2023 (compared to just 0.5 percent of the workforce in the US in 2019).[5] [6] Other studies also suggests Canada's leading position in AI talent.[7] [8]

Canada also boasts a vibrant AI start-up ecosystem, including over 1,500 AI companies, with over 150 firms raising a total of $2.5 billion in venture in 2023.[9] [10] Canada ranks fourth among the top ten nations in venture capital investment, total funding raised, Generative AI (GAI) companies founded, and AI patent filing growth per capita as of mid-2023.[11]

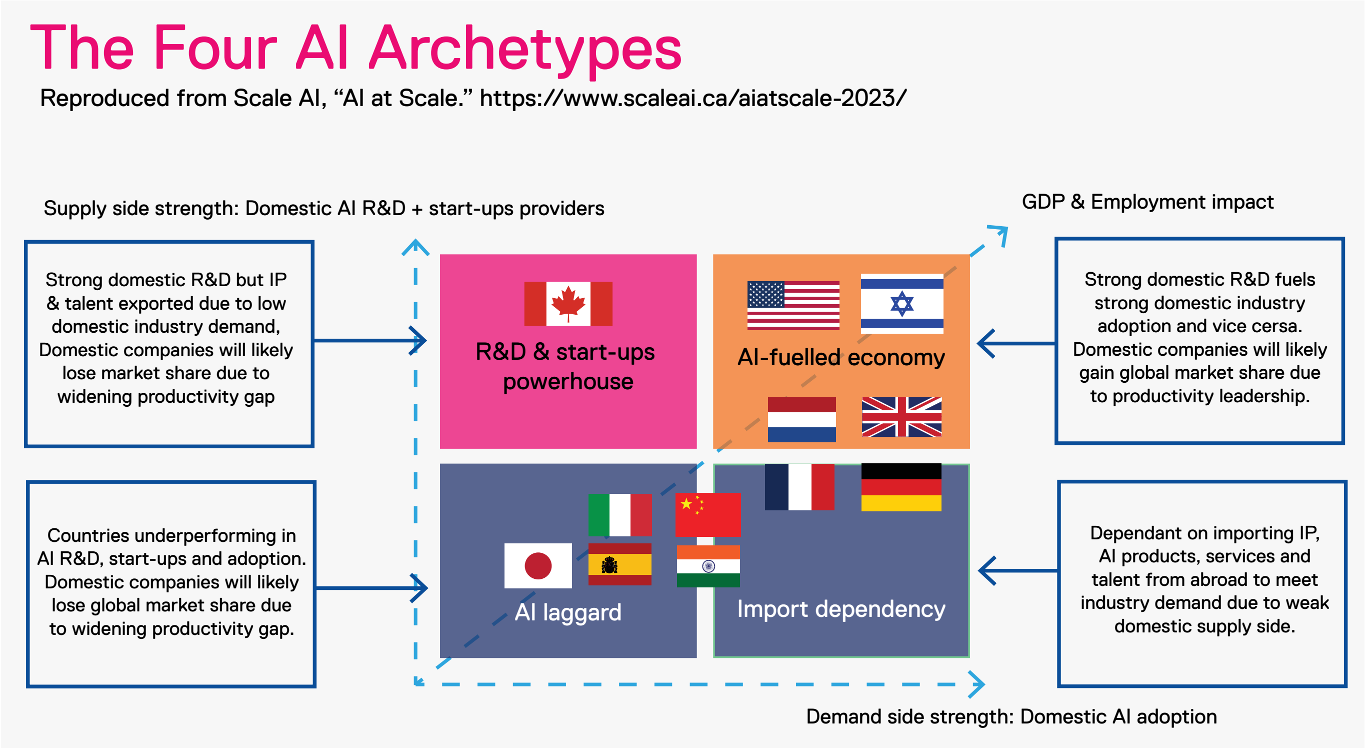
Canada's AI Strategy is funded by the federal government's Innovation, Science and Economic Development (ISED) department under the Pan-Canadian Artificial Intelligence Strategy (PCAIS, or the Strategy). Established in 2017, the Strategy has allocated over $573 million to support AI organizations, focusing on commercialization, standards, talent, and research until 2031. CIFAR, supported by ISED, administers PCAIS with $160 million since 2017, extended with an additional $208 million in 2022. CIFAR supports AI research institutes like Alberta Machine Intelligence Institute (AMII), Montreal Institute for Learning Algorithms (Mila), and Vector Institute, along with Canada's Global Innovation Clusters (CGIC), namely Scale AI, and the Standards Council of Canada in developing AI adoption standards.

As Canadian researchers and firms are faced with a race to develop and adopt AI in a rapidly advancing field—and policymakers seek to strike the right balance between enabling conditions and guardrails against harm—one critical issue has received relatively little attention in this country: Canada's "AI Compute" (AIC) capacity.

As Canadian researchers and firms are faced with a race to develop and adopt AI in a rapidly advancing field—and policymakers seek to strike the right balance between enabling conditions and guardrails against harm—one critical issue has received relatively little attention in this country: Canada's "AI Compute" (AIC) capacity.

Canadian AI start-ups face challenges in scaling innovation to commercialization.[12] Previous research found that despite a strong start-up ecosystem, adoption of AI by general businesses is low, with fewer than 1 in 25 Canadian firms reporting AI being used in their business, among the lowest of 35 OECD countries, and the second lowest in the G7.[13] A recent report by Scale AI, Canada's AI Global Innovation Cluster, finds that Canada is at a critical juncture: despite strength in domestic research and R&D and AI start-ups, Canada is missing key ingredients to allow AI to responsibly offer benefit to the economy. Canada risks falling behind leading international peers that are translating R&D and industry adoption of AI into productivity improvements, weakening Canadian competitiveness.

Figure 1. AI at Scale report graphic depicting Canada's AI ecosystem trajectory[14]



The Four AI Archetypes

Reproduced from Scale AI, "AI at Scale." https://www.scaleai.ca/aiatscale-2023/

Strong domestic R&D but IP & talent exported due to low domestic industry demand, Domestic companies will likely lose market share due to widening productivity gap

Countries underperforming in AI R&D, start-ups and adoption. Domestic companies will likely lose global market share due to widening productivity gap.

Strong domestic R&D fuels strong domestic industry adoption and vice cersa. Domestic companies will likely gain global market share due to productivity leadership.

Dependant on importing IP, AI products, services and talent from abroad to meet industry demand due to weak domestic supply side.

Supply side strength: Domestic AI R&D + start-ups providers
GDP & Employment impact
R&D & start-ups powerhouse
AI-fuelled economy
AI laggard
Import dependency
Demand side strength: Domestic AI adoption

Recent reporting suggests that AIC resources form a significant share of an AI firm's capital cost, as high as half of many companies' annual capital spending.[15] [16] Even the largest companies, despite spending tens of millions on computing infrastructure, cannot afford, let alone access, the fastest AI-specific computing infrastructure and cloud computing.[17] In February 2023, the OECD highlighted the lack of targeted plans for national AI Compute capacity in Canada's national AI strategy as a "policy blind spot [that] may jeopardize domestic economic goals."[18] While policymakers at all levels have started focusing on the issue, Canada's domestic supercomputing infrastructure lags in the number of compute units available (capacity) and performance (effectiveness) in AI training and inference.

What is Artificial Intelligence, and Artificial Intelligence Compute?

Before we delve further, we discuss the definition of Artificial Intelligence (AI) and Artificial Intelligence Compute (AIC):

What is Artificial Intelligence (AI)?

According to the Organisation for Economic Co-operation and Development (OECD)[19], artificial intelligence (AI) *"is a machine-based system for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical real or virtual environments. Different AI systems are designed to varying levels of autonomy and adaptiveness after deployment."*[20]

AI systems require computing power to generate inferences from their inputs, which are generally conducted by one of two types of systems:

(a) parallel computing infrastructure such as supercomputers, including publicly-owned supercomputers used primarily for academic and scholarly research purposes; or

(b) cloud-based computing systems, typically operated by large publicly traded US-based private companies such as Microsoft (Azure), Google (Cloud), and Amazon (Web Services). To combine these multiple contexts, we rely on the OECD's working group definition[21] of Artificial Intelligence Compute (AIC):

What is AI Compute (AIC)?

AI Compute is a specialized stack of hardware and software involving processors or chips, servers, storage, software, and networking, all designed to support AI-specific workloads and applications.
AI Compute covers a range of different technologies from computing chips to data servers to cloud computing.

AI development has two phases: training and inference. In the first phase, an AI model is "trained" on data. In the second phase, a trained AI model is deployed and then "infers" (i.e., makes decisions or takes actions) in the field based on new data. Training and inference phases run on AI Compute infrastructure in a private data centre or in the publicly accessible cloud infrastructure (such as Amazon Web Services).

We will expand on this definition in the subsequent sections.

Public funding for Canadian AI Compute

Only $40 million of the $573 million of the Pan-Canadian AI Strategy's funding is earmarked for AI Compute initiatives directly. The majority of AI computing infrastructure is the responsibility of the Digital Research Alliance of Canada (DRAC) and its research partners. DRAC is the successor to the high-performance computing ecosystem that began in the 1990s and evolved through research-focused initiatives like SHARCNET and Compute Canada. DRAC is now funded through PCAIS, and collaborates with the national research and education network operators, CANARIE (formerly the Canadian Network for the Advancement of Research, Industry and Education), to provide computing resources to researchers, start-ups, and educational institutions.

Investment in Canada's computing infrastructure has been largely limited to government and research institutions. Moreover, the current infrastructure's lacklustre computing performance effectively inhibits the ability to handle AI workloads within domestic borders.

There are no Canadian programs that focus on the AI Compute needs of private-sector actors, especially companies seeking to scale, or to dramatically grow public AI Compute infrastructure. Canada is already behind, as we'll explore in the next section, so this is a policy gap that needs to be addressed urgently.

There are no Canadian programs that focus on the AI Compute needs of private-sector actors, especially companies seeking to scale, or to dramatically grow public AI Compute infrastructure.

3

The Current State of AI Compute Worldwide and in Canada

Canadian AI firms and researchers face an AI Compute gap, and there are insufficient public policy measures to close it.

The lack of access and shortage of AIC-specific hardware, notably graphics and tensor processing units (GPU/TPUs[22]), limits the potential of Canada's AI ecosystem to thrive in a highly competitive global landscape.

While the specific compute need for AI may change in the future,[23] a growing chorus of business leaders, innovation experts, and AI researchers have raised the lack of access to AIC in Canada as a key business constraint.

To assess the state of AIC in Canada, we consider the OECD's policy framework for building national computing capacity, which expands upon the OECD definition presented in the previous section.[24] The framework identifies three main dimensions of national AIC:

OECD's national AIC dimensions

Capacity

- What is the availability and use for national AIC?
- How much national AIC is being used, by whom and in which sectors?

Effectiveness

- How effectively is the national AIC capacity being used?
- Is there sufficient skilled labour, R&D, affordable access, and an enabling policy environment?

Resilience

- How resilient is a country's compute capacity (e.g., secure, sovereign, sustainable)?
- Who owns the capacity and where is it located? Are supply chains secure?

While the specific compute need for AI may change in the future, a growing chorus of business leaders, innovation experts, and AI researchers have raised the lack of access to AIC in Canada as a key business constraint.

The OECD policy framework provides a rubric for identifying the key gaps, strengths, and opportunities for Canada's future AI Compute ecosystem. Our analysis employs the following approach to applying the OECD framework dimensions:

- ***AIC capacity*** is the computing infrastructure available to research and industry as measured by the number of computing cores and computing power available for use.
- ***AIC effectiveness*** examines who, where, and how domestic AI Compute infrastructure is used for industry and research purposes.
- ***AIC resilience*** is the current policy strategy, international and/or industry partnerships, and ensures consistent and reliable access to computing resources in disruptive scenarios (including economic, geopolitical, and natural disasters).

We explore each of these three aspects in Canada.

AIC capacity in Canada

The core measure for assessing and benchmarking national computing capacity (AIC infrastructure and supercomputers) is floating-point operations per second (FLOPS) performance.[25] As of November 2023, Canada's capacity is 41 petaFLOPS (PFLOPS, or 1 Quadrillion FLOPS) or 0.7 percent of global compute performance. By comparison, the US has 3,700 PFLOPS or 53 percent of global compute capacity, with Japan in second position at 670 PFLOPS or 10 percent, followed by China at 407 PFLOPS or six percent.

To better identify gaps in computing capacity and performance among nations, we calculate the per-capita and GDP-adjusted ratio of Canada to our G7 counterparts. We then multiply the adjustment ratios by Canada's AIC metrics to compare the total performance metrics of international comparators' compute capacity more effectively. The AI Compute gap persists despite controlling for international market and population differences. In comparative terms, the US leads Canada in adjusted AIC capabilities by a factor of 8 to 11 times, Japan leads by approximately eight times and France and Germany lead by two to three times. In practical terms, this means the Canadian AI ecosystem has less access to computing power domestically than its international peers.

Given Canada's relative strength compared to OECD and G7 peers on other AI-related metrics, this AI Compute gap creates special challenges for Canada's high-growth domestic AI firms and workers with AI skills. The gap is not just a competition gap between Canada and its peer jurisdictions—it is a gap within Canada, with high-capacity talent, firms, and research encountering low-capacity computing infrastructure. Firms in need of AIC buy these resources in the open market from primarily US-based cloud providers. Canadian AI firms and researchers are forced to seek alternative jurisdictions for innovation and commercialization incentives, leaving Canada further behind in the fiercely competitive market for AI productivity and economic growth potential.

Given Canada's relative strength compared to OECD and G7 peers on other AI-related metrics, this AI Compute gap creates special challenges for Canada's high-growth domestic AI firms and workers with AI skills.

Table 1. Canada's Compute capacity vs peer jurisdictions[26]

Country	Unadjusted Compute Performance Capacity	Per Capita Performance	GDP Adjusted Performance	Core Performance (PFLOPS)	Population Ratio	GDP Ratio
USA	**90.4**	**10.6**	**7.6**	3725.85	8.56	11.90
Japan	**16.3**	**7.6**	**8.5**	669.83	2.15	1.91
Italy	**8.5**	**5.6**	**8.9**	351.76	1.51	0.96
Germany	**6.2**	**1.9**	**3.1**	256.27	3.21	1.98
France	**4.2**	**2.4**	**3.2**	173.23	1.75	1.30
UK	**2.0**	**1.2**	**1.4**	81.71	1.72	1.44
Canada	**1.0**	**1.0**	**1.0**	41.21	1.00	1.00

Notably, this constraint is further emphasized for top talent working on frontier models. Only three of the five Canadian university-based supercomputers available have the specific hardware required for effective AI computation. With limited capacity, performance, and specialized AI hardware (GPUs), Canadian AI ecosystems face longer training times and tighter resource allocation measures when exploring frontier AI innovations as compared to other leading AI innovation nations.[27]

Canada's public AIC infrastructure needs to improve its performance by an order of a magnitude to meet the needs of frontier AI innovations, such as OpenAI's generative pre-trained transformer version 3 (GPT-3). As an example, GPT-3 requires 3,640s PFLOPS of computing performance to train in a single day.[28] To put this in perspective, combining all Canadian computing resources would allow that model to be trained in 100 days. If one can only use the five university-based supercomputers, the training would take six months. In comparison, it would take the US Frontier research supercomputer less than four days and privately-owned Microsoft's Azure Eagle Supercomputer roughly six-and-a-half days to complete the same workload.

Industry-owned infrastructure also remains scarce for commercial deployment of AI in Canada. There are signs this is growing, though the profit will accrue primarily to US-based firms.[29] Amazon Web Services plans to build over $24.8 billion dollars in cloud infrastructure in Canada through 2037, already having invested $2.57 billion domestically. This is in comparison to the $40 million contribution of AIC infrastructure earmarked for DRAC in the 2022 PCAIS funding.

Table 2. Domestic supercomputers and locations[30]

Domestic Supercomputers & Hosts	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
Underhill - Shared Services Canada	148,320	7.76	10.92	1,295
Robert - Shared Services Canada	148,320	7.76	10.92	1,295
Narval - Calcul Québec/Compute Canada	76,320	5.89	12.17	311
Niagara - SciNet/University of Toronto/Compute Canada	80,640	3.6	6.25	919
Cedar (GPU) - Simon Fraser University/Compute Canada	67,584	3.37	5.83	310
Banting - Shared Services Canada	53,200	2.68	4.09	
A14A - Software Company MCA	78,336	2.66	5.26	
Cedar (CPU) - Simon Fraser University/Compute	67,584	2.61	4.9	792
Daley - Shared Services Canada	53,200	2.6	4.09	
Béluga - Calcul Québec/Compute Canada	72,480	2.28	7.49	240
Total	845,984	41.21	71.92	5,162
Government	403,040	20.8	30.02	2,590
Research	364,608	17.75	36.64	2572
Private	78,336	2.66	5.26	N/A

Yet, private investment in building AIC in Canada continues to lag behind the US. Amazon plans to spend over $47 billion on new data centers in Virginia alone by 2040.[31] While Microsoft, Google, and Amazon are beginning to expand Canadian cloud computing capabilities, Canadian domestic early-growth firms cannot afford to access them.

The nascent domestic AI supply chain also constrains building AIC capacity in Canada.

Domestic AIC supply chain

The nascent domestic AI supply chain also constrains building AIC capacity in Canada. Canada hosts more than 100 companies conducting semiconductor R&D for vehicle and generative AI. Major AI industry players, like Radical Ventures, Ada, and Cohere, exemplify the substantial growth in the deployment of venture capital into Canadian AI companies. AI infrastructure and compute hardware firms in Canada, like Ranovus, Tenstorrent, Untether, Tartan AI, and Epic Semiconductors manufacture cloud infrastructure, server compute cores, and AI-specialized inference hardware.[32] Recently, Tenstorrent, an AI chipset design maker, has licensed their intellectual property (IP) to Japan's publicly backed AI chip start-up Rapidus, as they have a $2 billion investment for chip manufacturing in Hokkaido.[33] While these investments are laudable, mirroring trends seen in private investments in AIC above, larger competing economies have invested more heavily in AI supply chains from chipset design, distribution, infrastructure, and commercial cloud services, providing larger market opportunities and cost-effective AIC infrastructure.[34]

Internationally, we are seeing some signs of a consolidation of AI Compute supply chain companies under and into the large cloud-based AIC providers. See Table 4 for national policy examples of this consolidation.

Canada's most powerful research supercomputers, Narval and Beluga, containing 76,000 compute cores, are dwarfed in comparison to the 8.7 million compute cores US supercomputer Frontier hosts or the one million core count of Microsoft's Eagle supercomputer.

AIC effectiveness in Canada

Research effectiveness

Significant AIC infrastructure in Canada is directed towards academic and government researchers in biomedical science, climate science, and aerospace engineering.[35] Canada currently hosts 10 large-scale computing sites, with four being deployed by Shared Services Canada (SSC, 21 PLFOPS or 48 percent of total capacity), one privately operated by engineering consulting firm MCA (nine percent), and five sites among regionally dispersed universities (18 PFLOPS or 43 percent).[36] The majority of the Canadian AI industry and early adopters access AIC through US-based cloud computing service providers.

Despite having access to nearly half of the domestic AIC infrastructure, Canada's research ecosystem faces a shortage of computing resources compared to international ecosystems. Canada's most powerful research supercomputers, Narval and Beluga, containing 76,000 compute cores, are dwarfed in comparison to the 8.7 million compute cores US supercomputer Frontier hosts, or the one million core count of Microsoft's Eagle supercomputer.[37]

Industry effectiveness

As the majority of domestic supercomputing capacity is for research, Canada's AI industry, other small- and medium-sized enterprises (SMEs), and larger enterprises seeking to adopt AI, are primarily reliant on renting limited cloud-based AI Compute from large vendors like Amazon Web Services (AWS), Google Cloud Platform and Microsoft Azure outside of Canada's borders.[38] By comparison, the United States industry-owned Compute holds over a quarter of total computing (27 percent vs Canada's 9 percent) and performance capacity (31 percent vs Canada's 7 percent). This presents concerns for domestic scalability and economic and data sovereignty.[39] The lack of AI-specific domestic enterprise cloud infrastructure incentivizes early-growth firms to operate outside of Canadian borders, posing challenges to data sovereignty and talent retention.

The high costs of AI Compute remain a barrier to early AI firms. OpenAI's founder, Sam Altman, recently estimated that the cost of training the AI model behind GPT-4 was over $130 million.[40] While there are alternative, less computationally intensive forms of frontier AI models, the current prices of AIC remain untenable for the majority of domestic AI firms. Domestic early-stage AI firms face trade-offs in exchanging equity with cloud computing incumbents to acquire the resources needed to achieve effective and commercially-viable AI offerings.[41] The winners are AI industry giants like Microsoft, Google, and Amazon, which are paid for the growth in almost all computing needs.[42]

The lack of public AIC leaves domestic AI ecosystems and industry with no other viable options except to outsource their business operations and data to US private industry AIC infrastructure.[43]

Table 3. US-based cloud service provider infrastructure and investments in Canada

Cloud Infrastructure Locations	Investment since 2014	Toronto	Québec City	Montréal	Calgary
Microsoft[44]	~$677Million[45]	x	x		
Google[46]	~$735 Million[47]	x		x	
Amazon	~$2,570 Million[48]			x	x

The high costs of AI Compute remain a barrier to early AI firms.

AIC resilience in Canada

Domestic policies for expanding and supporting AI Compute ecosystem

In 2022, the Pan-Canadian AI Strategy (PCAIS) allocated $40 million to expand domestic computing capacity for its research ecosystem.[49] The Government of Canada offers a number of programs for Canadian businesses to adopt AI capabilities through various programs alongside research and innovation-centric funding through CIFAR. These include indirect subsidies such as tax incentives, wage compensation, small grants, and business loans. A few examples of industry support programs are summarized below:

- Scientific Research and Experimental Development (SR&ED) provides a tax incentive for R&D-related expenditures in AI advancement and advisory services on maximizing claim potential.

- The Canadian Digital Adoption Program (CDAP) provides grants for digital adoption and transformation consulting and cybersecurity resources.

- Digital Work Integrated Learning (WIL) programs subsidize wages for recent graduates and current post-secondary students for AI firms and organizations

- Zero-interest loans from Business Development Bank of Canada to invest in and implement AI capabilities.

While the majority of these programs support businesses looking to adopt AI capabilities, they do not provide direct subsidies for AIC cloud computing costs or hardware procurement.

International policy subsidies and investment for AIC infrastructure

Considerations also need to be made for the resilience of domestic AIC capacity in the event of market disruptions such as geopolitical tensions, and cloud computing enterprise insolvency or restrictions. More explicitly, there is a growing need for protections for the AI ecosystem to ensure AI Compute access in the event of AIC market failures or disruptions. In particular, access to US-based cloud computing infrastructure depends on a reliable American policy regime around data protection, and access for non-US-based companies.

A number of national bodies have introduced policy and strategy for these circumstances, listed below.

Both the US and Japan have promised to support domestic AIC chip manufacturing, reducing the reliance on the international supply of AIC infrastructure supply chain with multi-billion-dollar investments into manufacturing, R&D, and its skilled workforce.[50] [51]

Under the European High-Performance Computing Joint Undertaking (EuroHPC JU), the European Union (EU) will invest over $10 billion across its jurisdictions through 2027. The EuroHPC JU has allocated a total of $3 billion to expanding AIC capacity among its member nations. The EU has promised close to $700 million per year to support small to medium AI enterprises in providing access to high-performance computing (HPC) sites and equity financing of cloud computing resources to ensure access and scalability of their AI ecosystem.[52] Its public investing arm, InvestEU, provides over $1.6 billion in financial support for start-up incubation, and over $700 million to established enterprises to adopt and deploy AI.[53]

While the majority of these programs support businesses looking to adopt AI capabilities, they do not provide direct subsidies for AIC cloud computing costs or hardware procurement.

In Finland, 20 percent of the LUMI supercomputer is available to its AI industry, partnering with research institutions. Industry can also pay for high-performance computing resources, and SME firms can apply for grants valued up to $117,000 to use in this capacity.

In Finland, 20 percent of the LUMI supercomputer is available to its AI industry, partnering with research institutions. Industry can also pay for high-performance computing resources, and SME firms can apply for grants valued up to $117,000 to use in this capacity.[54] The program partially offsets the high operational and upfront costs of public AIC infrastructure investments.

France has promised $2.7 billion to support its AI ecosystem. Its strategy looks to support domestic and secure cloud infrastructure in protecting sensitive data and supports French cloud-service providers for trusted computing access.[55]

In 2022, the UK introduced an AI action plan outlining over $1.7 billion of support for the sector, supplementing the $3.8 billion it had already invested.[56] The UK has recently opened a call for AIC infrastructure grants of up to $900 million to domestic organizations that can host and operate a minimum of 2,000 GPUs, with specific mentions to start-up and research firms.[57]

Germany is investing over $2.4 billion in the current fiscal year into AI and plans to deploy AIC-ready HPC sites with at least 100,000 GPUs per site. Its federal Ministry of Education and Research's (BMBF) AI Action Plan looks to improve access to funding and public infrastructure for SMEs in its ecosystem in 2024.[58]

Table 4. National AIC investment policies

Countries	Title	Total Investment (CAD$ Millions)[59]	Start Term	End Term	Policy Type	Criteria
EU	High Performance Computing Joint Undertaking	$10,000	2022	2027	Subsidize and expand AIC access	Start-ups, SMEs, research ecosystem
Finland	LUMI Business Resource Program	N/A	2021	2027	Subsidize and expand AIC access	Industry, research ecosystem
US	CHIPS and Science Act	$70,000	2022		Semiconductor industry support	Industry, research ecosystem
France	National Cloud Strategy	$2,700	2021	2025	Subsidize and expand AIC access	Start-ups, SMEs, research ecosystem
United Kingdom	AI Action Plan	$1,700	2021	2031	Subsidize and expand AIC access	Start-ups, SMEs, research ecosystem
Germany	AI Action Plan	$2,400	2023	2024	Subsidize and expand AIC access	Industry, research ecosystem
Japan	2023/24 Supplemental Budget	$18,000	2023	2024	Semiconductor industry support	Industry

AIC infrastructure as a path to enhancing AI governance

A significant build-out of AI Compute infrastructure supports resilience in another fashion. Although the exact mechanisms are not yet fully developed, AI Compute infrastructure can be a vehicle for AI systems that are more secure, more responsive to public policy objectives, and that can support international collaboration on AI. The authors of "Computing Power and the Governance of Artificial Intelligence" make this clear: *"Relative to other key inputs to AI (data and algorithms), AI-relevant compute is a particularly effective point of intervention: it is detectable, excludable, and quantifiable, and is produced via an extremely concentrated supply chain. These characteristics, alongside the singular importance of computing for cutting-edge AI models, suggest that governing compute can contribute to achieving common policy objectives, such as ensuring the safety and beneficial use of AI."*[60]

However, for Canada to participate more fully in AI governance conversations through AI Compute, it must have a greater share of, or greater access to, AI Compute infrastructure.

In summary, the advances in direct access to public AIC, direct business support, and investments in AI procurement and production from international comparators further widen the gap in AIC capacity between Canada and its peers.

In summary, the advances in direct access to public AIC, direct business support, and investments in AI procurement and production from international comparators further widen the gap in AIC capacity between Canada and its peers.

4

Key Insights and Policy Scenarios

In the previous section, we identified the following key issues:

- **AIC capacity:** Canada lags behind all other G7 countries in AIC infrastructure. The available AIC infrastructure is not sufficient to support frontier-based research. Less than 10 percent of domestically situated AIC infrastructure is owned by the private sector. Investment to build domestic infrastructure is comparatively low, and the supply chain to support such projects is only in its nascent stage.

- **AIC effectiveness:** Canada primarily uses its research AIC capacity in biomedical science, climate science, and aerospace engineering. Private companies have to rely on cloud-based solutions to access AIC outside of Canada, and price-related pressures decrease their use.

- **AIC resilience:** Canadian public policy that shapes the ability to build a resilient AIC relies on broad-based programs that encourage the adoption of technologies in general, and no tailored policies exist that specifically target AIC access. This is in contrast to other jurisdictions with specifically-targeted policies.

Public policy objectives and trade-offs

We presume as a public-policy objective that Canada needs to expand access to AIC to close the AI Compute gap, retain AI firms and talent, and promote greater AI development and adoption. This objective can promote a larger public-policy goal of greater national productivity and competitiveness. Like the Internet before it, AI capacity could power a wide variety of areas of economic growth that require immediate, medium, and long-term solutions. AIC has the potential to be treated like a utility-like infrastructure.

Public policy must be designed so that public investment benefits the broadest array of AI ecosystem firms and stakeholders, that public investment does not simply replace private investment, and that any public involvement is well governed and adaptive to quickly-changing policy and industry circumstances.

We presume as a public-policy objective that Canada needs to expand access to AIC to close the AI Compute gap, retain AI firms and talent, and promote greater AI development and adoption.

The following options summarize the limited range of scenarios that we see available to the Government of Canada. The options are ordered from shorter to longer time horizons, in consideration of the significant present needs of Canada's AI ecosystem.

Scenario 1 | Centralize and subsidize AIC via federal procurement from cloud providers

Proposal: AI Compute resources can be purchased via enterprise cloud computing services. To help close the AI Compute gap, this scenario proposes a national effort to purchase capacity at as large a scale as possible to address immediate AIC access barriers for the domestic AI ecosystem. The scale of immediate AIC access needs could be assessed through a request-for-proposal process and operationalized through cloud-platform-allocation programs.

Providing direct support for cloud-computing costs, similar to the recent programs administered by France, would address the considerable need for computing resources endemic to Canadian innovators and nascent AI firms.

Ideally, the federal government would acquire the cloud AIC resources in bulk to increase the likelihood of accessing the expedient and scalable capacity at the lowest cost, based on demand, growth potential, and equity considerations for AIC.

Implementation: Given the urgency of the matter, the government would need to allocate funding and propose any legislative provisions through the 2024 Budget process, with an intention to have a market in place for Canadian firms to buy from as soon as possible. While the total demand from Canadian firms and their ability in the short term to switch providers is currently unknown, a purchasing program on the scale of hundreds of millions annually would both reduce overall costs and improve access to the domestic AI ecosystem.[61]

The fiscal cost to the government would depend on the extent of the subsidy or discounts provided to buyers, and which types of organizations would be eligible (e.g. research, start-ups, SMEs, large enterprises).

An alternative option is to provide direct subsidies vis-a-vis cloud service provider discount programs with firms already registered in existing technology-related programs. This direct subsidy reduces the public allocation and procurement burden and can be used to gauge the total market demand for cloud AI Compute by start-ups, SMEs, and researchers. The uptake and cost of the direct-subsidy approach would be covered by public resources, and the discount rate could be adjusted accordingly based on firm type and size in future iterations of the programs.

Pros: Centralized procurement and direct cloud-service-provider purchases are the most immediate path to increased affordable Canadian AIC access. This policy would be of substantial benefit, especially to early stage AI companies

Cons: These subsidies will further increase Canada's dependence on the US and cloud computing suppliers for access. Even with federal purchasing power, our market size may not afford sufficient economies of scale. This policy carries a fiscal cost, though this can vary with decisions around the size of the subsidy available.

Scenario 2 | Work with key trade partners to jointly purchase AIC at scale

Proposal: Canada can acquire even larger quantities of AIC and increase its negotiating power by joining existing international collaboratives, and through trade collaboration, by buying with some combination of European, British, Japanese, and other partners. Canada is uniquely positioned to lead such an initiative based on our international trade, diplomacy, and AI expertise. These partnerships would drive down the costs of AIC infrastructure procurement and access to additional cloud computing from existing public and private high-performance computing infrastructure internationally through joint purchasing, procurement and research agreements.

Implementation: The Canadian Government could expand on recent agreements with the UK and chipmaker Nvidia, and older multilateral work with the Global Partnership on Artificial Intelligence, among others, to support the development of national AIC Compute capacity through friend-shoring and expanding our ties to various players across the AI supply chain.[62] [63]

Canada is concluding negotiations in 2024 to accede to Pillar II of the Horizon Europe research and innovation program. As a result, Canada could explore joining the European High-Performance Computing Joint Undertaking (EuroHPC JU), and joining other non-EU states such as Israel and Iceland in this undertaking. Canada should also support and take advantage of massive US investments in the CHIPS and Science Act and our unique relationship with the US to negotiate greater access to long-term AIC hardware procurement.

Despite some loud voices to the contrary, Canada should not be building domestic chip manufacturing capacity (i.e. should not build chip foundries).[64] [65] [66] It would be an inefficient use of federal dollars and would not, from security, economic development, or other potential perspectives, deliver a satisfactory return on investment. By contrast, proactive procurement of future AIC hardware and infrastructure over long time horizons could ensure Canada has sustainable supply chains relative to its international peers. Multi-firm procurement of AI infrastructure and cloud services may require more complex upfront legal and contractual obligations, but will reduce the risk of long-term lock-in to subpar technology and services.[67] While the federal government has recently hinted[68] at the need to work with allies to look at "joint opportunities" in providing more computing power, more details need to be thought out before implementing this idea.

Pros: While trade agreements and multi-national initiatives usually take time to build, they should be faster than building domestic AIC capacity and they play to Canada's unique advantages. Scale and purchasing power are of the greatest importance for AIC due to their impact on price, after the immediacy of access. Robust multi-partner procurement reduces some costs for the domestic AIC ecosystem. It also has the potential to make more overall compute available to Canadians, compared to peers who might have more domestic computing capacity.

Cons: The way technology is governed internationally is changing fast, and the possibility of joining alliances may go down over time. This approach may also result in Canada having multiple partnerships with either interoperability issues, or lower performance per dollar in the long run.

Scenario 3 | Build domestic AI supercomputing capacity

Proposal: Canada procures pre-built high-performance computing hardware from third-party AIC computing vendors, with performance specifications similar to South Korea's Olaf supercomputer, which would provide deployment of additional AIC hardware for the AI ecosystem over a three- to five-year time horizon.[69] In addition, a public-private AIC resource allocation program, similar to Finland's business research program, could be introduced to offset the upfront investment and address domestic AIC access challenges for industry.

Currently, Canada's domestic AIC infrastructure has less than half the capacity it needs to reach parity with international competitors. For example, for Canada to triple its capacity and add 123 PFLOPS of Nvidia H100 AIC-specific compute infrastructure and reach internationally comparable AIC performance capacity, our preliminary estimates suggest it would require upwards of $400 million in AIC-specific hardware procurement, integration, and operation costs.[70] [71] [72] [73] Yoshua Bengio has estimated that the cost of sufficient high-performance computing infrastructure as a whole is about one billion dollars.[74]

Implementation: The uncertain future of AIC cost-to-performance makes it difficult to forecast the clear option for future domestic infrastructure investments. Vendors that handle the increasingly complex landscape of large-scale computing infrastructure in conjunction with a portfolio of AIC chipset design firms could offset the risk of potential AIC supply chain shortages. It would be prudent to secure multi-year, interoperable infrastructure agreements to build out domestic AIC capacity that could be offset by reselling or repurposing aging and incompatible legacy AIC infrastructure for industry needs and educational institutions. While the immediate intention is to ensure specialized AIC infrastructure is being procured over a longer horizon, the need for multi-firm public-private partnerships ought to be considered to ensure that resilient, long-term supply chains are established for unforeseen innovation in Canada's AI ecosystem.

A large domestic build-out of AI Compute could result from a public-private partnership. The federal government could help convene and join a consortium of major Canadian capital providers who could build supercomputing capacity. The capacity could be made available to a variety of private-sector actors as well as the federal government at globally market-competitive rates. The federal government could then use its membership in the consortium to help subsidize or make access available to others, including start-ups, to help achieve objectives to ensure that more AI talent and organizations stay and grow in Canada.

Pros: This scenario gets Canada closer to self-reliance, to greater performance parity with our international peers, and to more domestic sovereignty in AI capacity. Computing allocation can be more cost-effective, closely monitored, and distributed more securely among research and industry stakeholders.

Cons: Potentially the highest fiscal cost to the government, and dependent on many factors including foreign nations and supply shortages. This option has a relatively longer time horizon to procure and operationalize with existing domestic computing infrastructure compared to cloud service providers' resources.

Policy considerations under all scenarios

The three options are not mutually exclusive and can be sequenced. For example, “Centralize and Subsidize” (Scenario 1) and “Work with Key Trade Partners” (Scenario 2) can be pursued in the near term to ramp up AIC capacity more quickly and offer a bridge to the domestic “build” (Scenario 3) of new capacity over a longer horizon.

The aforementioned scenarios are focused in particular on AIC capacity and do not address allocation mechanisms (i.e. AIC effectiveness) of expanded domestic AIC capacity. Since the majority of private enterprises access AI Compute resources internationally, understanding the aggregate demand for such services would require considerable AIC ecosystem consultation. Allocation mechanisms need to consider potential market disruptions of private AIC infrastructure investment. A separate analysis would be necessary to identify the magnitude and dynamic response to public AIC infrastructure and subsidy programs. This analysis would specifically address the scope, timelines, and costs that would dynamically arise from differing levels of public intervention in the domestic market for AIC infrastructure and cloud service provision.

Bridging the AI Compute gap should align with Canada’s positioning as a leader in responsible AI innovation and adoption, including the Pan-Canadian AI Strategy, the *Artificial Intelligence and Data Act* (AIDA) legislative and regulatory process, and other initiatives.[75] In line with the thinking of some of our AI leaders, investment in AIC can be directed towards positive use cases including health and safety.[76]

Bridging the AI Compute gap must factor implications for energy use and climate objectives. Canadian AIC infrastructure is currently half as energy efficient as AI in the United States.[77] Yet, Canada’s electricity is predominantly “green”[78], the aggregate cost of power is cheaper than in many other nations including the US and the UK,[79] and the colder climate[80] offers the potential for AIC to be more efficient in its cooling needs. These factors create the conditions for cost-competitiveness in attracting AIC infrastructure investments to Canada, while also aligning with Canada’s climate goals.

These options remain preliminary and are aimed at informing fast-moving discussions across policy and AI ecosystem actors. Further research, analysis, and industry engagement are required to assess the feasibility and cost-effectiveness of these various options.

Policy design considerations and risks

As the government deliberates investing in private-sector AIC, considerations include:

- Insufficient demand: One major risk of large government purchases of AIC is capacity being underutilized, for the purchase rise far exceeding current real demand for AIC. As the procurement mechanism of Scenario 1 relies on non-market mechanisms, a clear understanding of demand must be established before any policy implementation.

- Private investment crowd-out: Given the very small current domestic AIC capacity accessible to the private sector, any large government procurement/subsidization programs will have second-order pricing impacts that affects private companies accessing AIC resources outside of these programs. These dynamic crowding-out effects should be carefully modelled and understood in all scenarios.

- Innovative market mechanisms: In the ideal, medium-to-long-term future for Scenarios 2 and 3, the organization(s) deploying domestic AIC would have sufficient capacity to employ markets to either recoup the costs of this investment through margins on the savings generated by scale, or even through taking equity in private AIC buyers.

- Leverage past tech market models: While very concentrated, the market for cloud storage has also become very dynamic with the ability to move computing and storage functions across vendors in real time. As the AIC market matures, Canada will need to ensure its AIC can similarly compete on dynamism.

- Implementation inefficiencies: It is of particular concern that if these options are to be undertaken, the way a particular program is delivered impacts program use and uptake. Decisions around eligibility, processing timelines, and direct subsidy thresholds are more critical than the resources being offered by public intervention. This can be particularly important in deciding on the correct vehicle of distribution and eligibility for Scenario 1, as expediency and access are most important for addressing immediate cost and access barriers to AIC in Canada.

What’s next

Further research considerations:

Having begun to engage key players across Canada’s AIC ecosystem, the Dais is committed to continuing to inform this strategy. In particular, we believe we can support development of this strategy over the coming year by:

- convening roundtables and engaging in a thorough and balanced (e.g. across industry, civil society, academia, and government) consultation spanning the AIC ecosystem in understanding total procurement and operational needs for domestic AIC

- helping policy-makers understand the climate and energy considerations of additional AIC infrastructure investment, namely the regional capacity, electrical constraints, and interoperability of existing high-performance computing infrastructure

- creating a framework to estimate the aggregate demand and total cost burden for AI Compute as the magnitude and service delivery (i.e. preference for more cloud-based compute or AI-specific infrastructure) remain uncertain. This will help in understanding the appropriate procurement or subsidy amount in Scenario 1.

By conducting further economic analysis of domestic AIC, the Dais can provide evidence required for decision-makers to better allocate resources towards building out and attracting the capital investment (investment and pension funds for example) for domestic AIC infrastructure expansion and support programs.

5

Conclusion

To quote Canada's Minister of Innovation, Science and Economic Development (ISED) during the recent Canada-UK AI partnership announcement: "We have the brain. Now we need...the mainframe."[81] Canada's AI strategy is talent-focused, but lacks the infrastructure to provide opportunities for early-growth firms, protect public interests, and capture AI's potential for long-term economic prosperity. To scale our AI start-ups and responsibly commercialize our AI research, we need to work with others around the world and invest in computing capacity for Canadians of a quality and scale to match these ambitions. If Canada is to be taken seriously as a leader in the international AI ecosystem, we must act, as our current capacity does not fulfill the requirements for applied research, industry innovation, and governance. At the same time, Canada must support responsible innovation.

Artificial Intelligence Compute capacity is the most important factor missing in the growth of Canada's artificial intelligence ecosystem. If we agree with the justifications for investing further in this ecosystem to the tune of hundreds of millions to billions of dollars—and the productivity gains alone, if distributed appropriately, may justify this—then we need to act in months, and not wait years, to negotiate for and build, a world-class AIC strategy.

If Canada is to be taken seriously as a leader in the international AI ecosystem, we must act, as our current capacity does not fulfill the requirements for applied research, industry innovation, and governance. At the same time, Canada must support responsible innovation.

Endnotes

[1] “Real Talk: How Generative AI Could Close Canada’s Productivity Gap and Reshape the Workplace—Lessons From the Innovation Economy,” *The Conference Board of Canada* (blog), February 20, 2024, https://www.conferenceboard.ca/product/real-talk/.

[2] Amba Kak West and Sarah Myers, “The Problem With Public-Private Partnerships in AI,” *Foreign Policy* (blog), February 22, 2024, https://foreignpolicy.com/2024/02/12/ai-public-private-partnerships-task-force-nairr/.

[3] “2018 ACM A.M. Turing Award Laureates,” Association for Computing Machinery, accessed February 21, 2024, https://awards.acm.org/about/2018-turing.

[4] “AICan: The Impact of the Pan-Canadian AI Strategy,” CIFAR, accessed March 12, 2024, https://cifar.ca/ai/impact/.

[5] “Built to Scale? Microcredentials Use Among Digital Professionals,” The Dais, October 21, 2023, https://dais.ca/reports/built-to-scale-microcredentials-use-among-digital-professionals/.

[6] “U.S. AI Workforce.” Center for Security and Emerging Technology (blog), accessed March 17, 2024, https://cset.georgetown.edu/publication/u-s-ai-workforce/.

[7] Deloitte Canada, “Impact and Opportunities: Canada’s AI Ecosystem – 2023,” accessed February 21, 2024.

[8] “Future of Work Report: AI at Work,” LinkedIn, accessed February 21, 2024, https://economicgraph.linkedin.com/research/future-of-work-report-ai.

[9] Crunchbase. “List of Top Canada Artificial Intelligence (AI) Companies - Crunchbase Hub Profile.” Accessed February 21, 2024, https://www.crunchbase.com/hub/canada-artificial-intelligence-companies.

[10] OECD, “Live Data from OECD.AI,” accessed March 17, 2024, https://oecd.ai/en/data.

[11] Deloitte Canada, “Impact and Opportunities: Canada’s AI Ecosystem – 2023,” accessed February 21, 2024, https://www2.deloitte.com/ca/en/pages/deloitte-analytics/articles/impact-and-opportunities-canadas-ai-ecosystem-2023.html.

[12] Scale AI, “AI at Scale,” accessed February 21 2024, https://www.scaleai.ca/aiatscale-2023/.

[13] “Automation Nation? AI Adoption in Canadian Businesses,” The Dais, August 14, 2023, https://dais.ca/reports/automation-nation-ai-adoption-in-canadian-businesses/.

[14] Scale AI, “AI at Scale,” Accessed February 21, 2024. https://www.scaleai.ca/aiatscale-2023/.

[15] Guido Appenzeller, Matt Bornstein, and Martin Casado, “Navigating the High Cost of AI Compute,” Andreessen Horowitz, April 27, 2023, https://a16z.com/navigating-the-high-cost-of-ai-compute/.

16 Murad Hemmadi, “Cloud Giants Ride Wave of AI Enthusiasm in Canada,” *The Logic*, January 16, 2024, https://thelogic.co/news/cloud-giants-ride-wave-of-ai-enthusiasm-in-canada/.

[17] Murad Hemmadi, “‘To Compete, You Must Compute’: Powering Canada’s AI Surge,” *The Logic*, November 30, 2023, https://thelogic.co/news/special-report/to-compete-you-must-compute-powering-canadas-ai-surge/.

[18] OECD, “A Blueprint for Building National Compute Capacity for Artificial Intelligence,” Paris: OECD, February 28, 2023, https://doi.org/10.1787/876367e3-en.

[19] “AI Principles,” OECD, accessed March 18, 2024, https://oecd.ai/en/ai-principles.

[20] “Stuart Russell, Karine Perset and Marko Grobelnik, “Updates to the OECD’s Definition of an AI System Explained,” OECD. AI (blog), November 29, 2023, https://oecd.ai/en/wonk/ai-system-definition-update.

[21] OECD, “A Blueprint for Building National Compute Capacity for Artificial Intelligence,” OECD Digital Economy Papers, no. 350, 2023, OECD Publishing, https://doi.org/10.1787/876367e3-en.

[22] Graphical Processing Units, or GPUs, are computer processing units that specialize in making calculations in processing and displaying computer graphics, particularly complex ones involved in video game graphics, or other 3D computer graphic applications. Because the calculations that are required to process computer graphics are numerically similar to the calculations required in machine-learning models, they have been favoured by many AI companies over traditional computer processing units, or CPUs. Tensor Processing Units (TPU), a more recent product, are computer processing units that specifically specialize in training machine learning models.

[23] Jai Vipra and Sarah Myers West, ”Computational Power and AI,” AI Now Institute (blog), September 27, 2023, https://ainowinstitute.org/publication/policy/compute-and-ai.

[24] OECD, “A Blueprint for Building National Compute Capacity for Artificial Intelligence,” Paris: OECD, February 28, 2023, https://doi.org/10.1787/876367e3-en.

[25] PFLOPS, or petaflops, is the measure of total computing performance on the LINPACK benchmarking standard for large-scale computing infrastructure. 1 PFLOP is equivalent to 10^15 (1 Quadrillion) FLOPS.

[26] To account for economic market size and population differences, we calculate a per-capita adjustment ratio, (https://data.worldbank.org/indicator/SP.POP.TOTL?locations=CA-US-GB-FR-DE-JP-IT) and GDP (https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=CA-US-GB-FR-DE-JP-IT) adjustment ratio using TOP500’s supercomputer sub-list generator as of November 2023, https://www.top500.org/statistics/list/.

[27] Simon Nakonechny, “AI Pioneer Yoshua Bengio Urges Canada to Build $1B Public Supercomputer, *CBC News*, January 29, 2024, https://www.cbc.ca/news/canada/montreal/bengio-asks-canada-to-build-ai-supercomputer-1.7094858.

[28] Chuan Li, “OpenAI’s GPT-3 Language Model: A Technical Overview,” Lambda (blog), June 3, 2020, https://lambdalabs.com/blog/demystifying-gpt-3.

[29] “Vass Bednar, ”Why Canada Needs a Publicly Owned Cloud,” *Financial Post*, January 17, 2023, https://financialpost.com/telecom/why-canada-needs-publicly-owned-cloud.

[30] Accessed by querying Canada in Top 500’s sub-list generator: https://www.top500.org/statistics/sublist/. RMax is the result of a popular benchmarking test, with higher scores indicating better. Rpeak divides the total PFlop by the number of instructions that can be issued per second, and can be interpreted as an efficiency measure.

[31] Learn about AWS's Long-Term Commitment to Virginia." 2023. US About Amazon. June 7, 2023, https://www.aboutamazon.com/news/aws/aws-commitment-to-virginia.

[32] Anita Balakrishnan, "Here's How Canada's Semiconductor Industry Stacks up," *The Logic*, July 5, 2023, https://thelogic.co/news/special-report/heres-how-canadas-semiconductor-industry-stacks-up/.

[33] Murad Hemmadi, "Tenstorrent Joins Japan's AI Compute Efforts," *The Logic*, February 27, 2024, https://thelogic.co/briefing/tenstorrent-joins-japans-ai-compute-efforts/.

[34] "Fact Sheet : CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China," The White House, August 9, 2022, https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/.

[35] "About SciNet," SciNet: Advanced Research Computing at the University of Toronto," accessed February 21, 2024, https://www.scinethpc.ca/about-scinet/.

[36] "National Systems - Alliance Doc," accessed February 21, 2024, https://docs.alliancecan.ca/wiki/National_systems#Compute_clusters/.

[37] Simon Nakonechny, "AI Pioneer Yoshua Bengio Urges Canada to Build $1B Public Supercomputer," *CBC News,* January 29, 2024, https://www.cbc.ca/news/canada/montreal/bengio-asks-canada-to-build-ai-supercomputer-1.7094858.

[38] Murad Hemmadi, "'To Compete, You Must Compute': Powering Canada's AI Surge", *The Logic*, November 30, 2023, https://thelogic.co/news/special-report/to-compete-you-must-compute-powering-canadas-ai-surge/.

[39] Scale AI. "AI at Scale," accessed February 21, 2024, https://www.scaleai.ca/aiatscale-2023/.

[40] Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over, *WIRED*, April 17, 2023, https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

[41] "FTC Launches Inquiry into Generative AI Investments and Partnerships," Federal Trade Commission, January 25, 2024, https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships.

[42] Tom Krazit, "Do Big Cloud Companies Control AI Startups?" *Runtime*, January 26, 2024, https://www.runtime.news/do-big-cloud-companies-control-ai-startups/.

[43] Murad Hemmadi, "'To Compete, You Must Compute': Powering Canada's AI Surge," *The Logic*, November 30, 2023, https://thelogic.co/news/special-report/to-compete-you-must-compute-powering-canadas-ai-surge/.

[44] "Data Residency in Azure," Microsoft Azure," accessed March 14, 2024, https://azure.microsoft.com/en-ca/explore/global-infrastructure/data-residency/.

[45] "Microsoft Expands Digital Footprint in Quebec with USD $500 million Investment in Infrastructure and Skilling Initiatives," Microsoft News Center Canada," November 22, 2023, https://news.microsoft.com/en-ca/2023/11/22/microsoft-expands-digital-footprint-in-quebec-with-usd500-million-investment-in-infrastructure-and-skilling-initiatives/.

[46] "Cloud locations - Regions & Zones," Google Cloud, accessed March 14, 2024, https://cloud.google.com/about/locations.

[47] Murad Hemmadi,, "Cloud Giants Ride Wave of AI Enthusiasm in Canada," *The Logic*. January 16, 2024, https://thelogic.co/news/cloud-giants-ride-wave-of-ai-enthusiasm-in-canada/.

[48] "Canadian Cloud Hosting Services," Amazon Web Services, accessed March 14, 2024, https://aws.amazon.com/local/canada/.

[49] "Pan-Canadian Artificial Intelligence Strategy," Innovation, Science and Economic Development Canada, July 20, 2022, https://ised-isde.canada.ca/site/ai-strategy/en/pan-canadian-artificial-intelligence-strategy.

[50] Tetsushi Kajimoto and Sam Nussey, "Japan to Spend $13 Bln for Chip Industry Support in Extra Budget, Reuters, November 10, 2023, https://www.reuters.com/markets/asia/japan-allocate-13-bln-chip-industry-support-extra-budget-2023-11-10/.

[51] "Fact Sheet: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China," The White House, August 9, 2022, https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/.

[52] "AI Innovation Package to Support Artificial Intelligence Startups and SMEs Policy," OECD, accessed February 28, 2024, https://oecd.ai/en/dashboards/policy-initiatives/http://aipo.oecd.org/2021-data-policyInitiatives-27588.

[53] European Commission, "Communication on Boosting Startups and Innovation in Trustworthy Artificial Intelligence," European Commission: Shaping Europe's Digital Future, January 24, 2024, https://digital-strategy.ec.europa.eu/en/library/communication-boosting-startups-and-innovation-trustworthy-artificial-intelligence.

[54] "Solutions for companies - Implementation and Pricing," LUMI, accessed February 28, 2024, https://www.csc.fi/ratkaisut-yrityksille-laskentapalveluiden-kaytto.

[55] "National Cloud Strategy: Launch of the Industrial Plan to Support the Sector,"Enterprises.Gouv.Fr., accessed February 8, 2024, https://www.entreprises.gouv.fr/fr/actualites/numerique/strategie-nationale-cloud-lancement-du-plan-industriel-de-soutien-la-filiere.

[56] International Trade Administration, "United Kingdom Artificial Intelligence Market 2023," March 29, 2023, https://www.trade.gov/market-intelligence/united-kingdom-artificial-intelligence-market-2023.

[57] "Host Sites for the next Wave of UK Government AI Infrastructure," UK Research and Innovation, January 24, 2024, https://www.ukri.org/opportunity/host-sites-for-the-next-wave-of-uk-government-ai-infrastructure/.

[58] "BMBF Action Plan ' Artificial Intelligence,'" Bundesministerium Für Bildung Und Forschung, November 7, 2023, https://www.bmbf.de/bmbf/de/forschung/digitale-wirtschaft-und-gesellschaft/kuenstliche-intelligenz/ki-aktionsplan.html.

[59] Converted in March 2024 CAD dollars, original value is based on announcement year.

[60] Girish Sastry et al., "Computing Power and the Governance of Artificial Intelligence," *Arxiv*, February 13, 2024, arXiv:2402.08797.

[61] In 2023, 180 Canadian AI firms received over $2.5 billion in venture capital funding, according to Crunchbase and the OECD; estimates are that over half of funding for genAI companies goes to AIC costs. https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=vc-investments-in-generative-ai-by-country.

[62] Canada, Global Affairs, "Memorandum of Understanding between the Department of Foreign Affairs, Trade and Development of Canada and the Department for Science, Innovation and Technology of the United Kingdom of Great Britain and Northern Ireland Concerning Cooperation over Scientific Research and Innovation," January 30, 2024, https://www.tradecommissioner.gc.ca/tcs-sdc/innovators-innovateurs/mou_science_innovation_protocole_entente.aspx?lang=eng.

[63] Brian Caulfield, "Canada Partners With NVIDIA to Supercharge Computing Power," NVIDIA (blog), February 5, 2024, https://blogs.nvidia.com/blog/canada/.

[64] "OpenAI's Altman in Talks to Raise Funds for Chips, AI Initiative – WSJ," Reuters, February 9, 2024, https://www.reuters.com/technology/openais-altman-talks-raise-funds-chips-ai-initiative-wsj-2024-02-09/.

[65] Anton Shilov, "Legendary Chip Architect Jim Keller Responds to Sam Altman's Plan to Raise $7 Trillion to Make AI Chips — 'I Can Do It for Less than $1 Trillion,'" *Tom's Hardware*, February 17, 2024, https://www.tomshardware.com/tech-industry/artificial-intelligence/jim-keller-responds-to-sam-altmans-plan-to-raise-dollar7-billion-to-make-ai-chips.

[66] Anton Shilov, "Nvidia CEO Jensen Huang Says $7 Trillion Isn't Needed for AI — Cites 1 Million-Fold Improvement in AI Performance in the Last Ten Years," *Tom's Hardware*, February 14, 2024, https://www.tomshardware.com/pc-components/gpus/nvidia-ceo-jensen-huang-says-dollar7-trillion-isnt-needed-for-ai-cites-1-million-fold-improvement-in-ai-performance-in-the-last-ten-years.

[67] While Nvidia holds a dominant share of the AIC hardware market share, AMD, Intel and potentially Apple are catching up in AIC performance. See: https://www.top500.org/statistics/list/. As of November 2023, Nvidia co-processors hold over 25 percent of the Top 500 supercomputers list, compared to less than three percent for AMD and two percent for Intel.

[68] Murad Hemmadi, "G7 Nations Exploring Cooperating on AI Compute: Champagne," *The Logic*, March 15, 2024.

[69] "Olaf - Lenovo ThinkSystem SR675 V3, AMD EPYC 9334 32C 2.7GHz, NVIDIA H100, Infiniband NDR 400" TOP500, accessed February 21, 2024, https://www.top500.org/system/180180/.

[70] In real terms, the cost of NVIDIA H100 graphics processing cards is approximately $45,000, capable of performing 34 TFLOPS on the LINPACK benchmark. NVIDIA. "NVIDIA H100 Tensor Core GPU Datasheet," accessed February 21, 2024, https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet.

[71] "NVIDIA H100 - GPU Computing Processor - NVIDIA H100 Tensor Core - 80 GB - 900-21010-0000-000 - Graphic Cards," CDW.ca, accessed February 21, 2024, https://www.cdw.ca/product/nvidia-h100-gpu-computing-processor-nvidia-h100-tensor-core-80-gb/7354651.

[72] Assuming domestic infrastructure has the PCIe slot and power capacity to add over 3,600 PCIE-based Nvidia H100 accelerator cards.

[73] "NVIDIA H100 - GPU Computing Processor," SHI.ca, accessed February 21, 2024, https://www.shi.ca/Product/45671009/NVIDIA-H100-GPU-computing-processor.

[74] Simon Nakonechny, "AI Pioneer Yoshua Bengio Urges Canada to Build $1B Public Supercomputer," *CBC News*, January 29, 2024, https://www.cbc.ca/news/canada/montreal/bengio-asks-canada-to-build-ai-supercomputer-1.7094858.

[75] "Submission on the Proposed Artificial Intelligence and Data Act," The Dais, November 13, 2023, https://dais.ca/reports/submission-on-the-proposed-artificial-intelligence-and-data-act/.

[76] As noted in a recent article arguing for more Canadian AIC capacity, Yoshua Bengio would like to see that class of machine built in Canada, funded by governments, so public entities have the digital firepower to keep up with the private tech giants they'll be tasked with monitoring or regulating. "I think government will need to understand at some point, hopefully as soon as possible, that it's important for [them] to have that muscle," said Bengio. Simon Nakonechny, "AI Pioneer Yoshua Bengio Urges Canada to Build $1B Public Supercomputer," *CBC News*, January 29, 2024, https://www.cbc.ca/news/canada/montreal/bengio-asks-canada-to-build-ai-supercomputer-1.7094858.

[77] As measured by performance as a function of kilowatts needed to power public-use AIC infrastructure, Canada's PFLOPS per kilowatt is 0.007 compared to the United States, measured at 0.016 PFLOPS per kilowatt. This implies Canada's current public-use AIC infrastructure is less than half as energy efficient as the United States' larger-scale public-use computing infrastructure.

[78] Nour Abdelaal, Adams Aghmien and Andre Cote, "Clean Connection: How Digitization Can Support Canada's Path to Net-Zero," The Dais, June 23, 2023, https://dais.ca/reports/clean-connection-how-digitization-can-support-canadas-path-to-net-zero/.

[79] "Electricity Prices around the World," GlobalPetrolPrices.com, accessed March 12, 2024, https://www.globalpetrolprices.com/electricity_prices/.

[80] "Climate Comparison: Canada / United States," WorldData.info, accessed March 12, 2024, https://www.worlddata.info/climate-comparison.php?r1=canada&r2=usa.

[81] Murad Hemmadi, "Canada and U.K. Are Allies, Not Rivals, on AI: Champagne and Donelan," *The Logic*, January 31, 2024, https://thelogic.co/news/canada-and-u-k-are-allies-not-rivals-on-ai-champagne-and-donelan/.