# Pan-Canadian AI Compute Environment (PAICE) Operating Plan

Year 1 - FY 2022/2023

**Digital Research Alliance** of Canada | **Alliance de recherche** numérique du Canada

# Table of Contents

# Statement of Work

The initial year (2022-23) will predominantly consist of planning and coordinating procurement of the new systems as well as preparing for their coordinated operation as an integrated national platform.

The Alliance, AI Institutes, host sites, and federation national teams will be working on multiple operational components to ensure that the new national AI platform can deliver a coordinated service to the AI researchers once the new hardware is installed.

The activities are listed below by Eligible Activity, as defined in the contribution agreement.

## AI Compute Infrastructure Acquisitions and Operations

| **Activity**:  Coordinated RFI for AI Compute |
| --- |
| The Alliance will lead a Request for Information (RFI) process to gather technical information about possible AI computing solutions and technologies for all three AI institutes to consider in planning systems acquisition. |
| **Target Outcome**:  RFI posted and response in Q3 of FY2022/23 |

| **Activity**:  RFP's for 3 AI Compute Systems |
| --- |
| The three AI Institutes in coordination with the Alliance, regional consortia, and host sites will develop Request for Proposals (RFP) for each of the three proposed AI systems using the information gathered during the coordinated RFI. This involves development, issuance, and evaluation of the proposals. |
| **Target Outcome**:  RFP posted and response in Q4 of FY2022/23 |

| **Activity**:  Selection of systems |
| --- |
| A direct follow on of the RFP evaluations will be selection, contract negotiations and delivery planning for the new systems.  Depending on equipment availability and barring any major supply chain issues it is expected that equipment delivery, installation, and acceptance will be completed in Q1 FY2023/2024 |
| **Target Outcome**: Systems selected in Q4 of FY2022/23 |

| **Activity**: Datacenter Preparations for Installation |
|---|
| In parallel with the procurement process, detailed power/cooling requirements will be finalized.  Host site datacenter modifications, primarily electrical and cooling related, will be performed ahead of the systems arrival, however exact timelines are unknown at this time.  Work will be started in Q4 FY2022/23, however as the full scope of work is not yet known it may only be partially completed and thus funds have been allocated for datacenter renovations for FY2023/24 as well.  Some sites will also require external networking upgrades which will also be procured and installed during this time. |
| **Target Outcome**:  Datacenter preparations started Q4 FY2022/23 |

| **Activity**:  Development of National AI Platform Common Services |
|---|
| Common components such as user environments, AI specific user support mechanisms, and software stacks will be developed by the AI compute delivery teams in coordination with the Alliance and wider federation national teams where appropriate in preparation of the new systems commissioning. |
| **Target Outcome**:  Draft integration & AI support plan by Q4 FY2022/23 |

| **Activity**:  Monitoring & Metrics for AI National Platform |
|---|
| Instrumentation & monitoring of systems, allocation process, support and user statistics to be performed by AI Institutes, regional ARC and host sites in coordination with the Alliance to ensure efficient user and system operations as well as provide performance metrics required for KPI reporting. |
| **Target Outcome**:  Monitoring design by Q3 FY2022/23 |

## AI Compute Resource Allocations:

| **Activity:**  AI Compute Resource Allocations |
|---|
| The AI institutes in coordination with the Alliance will develop an effective AI compute resource allocation process that uses a transparent, merit-based validated research model |

that leverages expert peer review of research with equity, diversity and inclusion as key components. This novel allocation model is inspired by the Compute Canada consultations on the Resource Allocation Competition (RAC) processes and replicates the Rapid Access Service, which provides all eligible researchers with resources upon request while taking advantage of already existing Peer-reviews to provide elevated access. The AI allocation model is specifically designed to address existing pain points in the RAC and reduce the burden on both researchers and administrators, while at the same time allocating resources fairly and transparently.

**Target Outcome:** Process in place by Q4 FY2022/23, Implementation Q1 FY2023/24

## Support for Highly Qualified Personnel:

**Activity:** Hiring AI Support

The AI Institutes and host sites will be determining divisions of labour and hiring staff to fulfil those roles. This includes local clusters' system administrators and AI specific research support staff.

**Target Outcome:** New staff hired started in Q3 FY2022/23

## Cybersecurity:

**Activity:** Cybersecurity Initiatives

AI Platform cybersecurity operations to be coordinated and integrated with the existing Alliance national frameworks.

**Target Outcome:** Q4 FY2022/23

## Management and Administration:

N/A

# Financial Overview

## Annual Cash Flow Requirements

The following is a summary of the annual cash flow requirements in support of the activities planned for each year.

| Fiscal Year | Projected Annual Cash Flow Requirements |
|---|---|
| 2022/23 - Q1 | $0.00 |
| 2022/23 - Q2 | $0.00 |
| 2022/23 - Q3 | $687,364.00 |
| 2022/23 - Q4 | $687,364.00 |
| 2023/24 | $23,662,872.58 |
| 2024/25 | $12,423,633.29 |
| 2025/26 | $14,684,664.94 |
| 2026/27 | $13,016,019.19 |
| Total | $64,998,919.01 |

## Expenditures by Eligible Activity

The following is a summary of the planned costs for each eligible activity for FY2022/23.

| Activity | Expenditure | Incurred Cost (Alliance/Recipient) |
|---|---|---|
| **AI Comp. Infra. and Ops.** | | |
| Datacenter Preparations for Installation | $850,000.00 | Recipient |
| Networking Equipment | $164,729.00 | Recipient |

| Support for Highly Qualified Personnel | | |
|---|---|---|
| Hiring AI Support | $360,000.00 | Recipient |
| Cybersecurity | | |
| | | |
| AI Compute Resource Allocations | | |
| Management and Administration | | |
| Total | $1,374,729.00 | |

# Risk Assessment

### Risk 1: Program Delays

There is a risk that not all procurement competitions complete.  There can also be delays in approval considering the number of stakeholders involved.

Mitigation strategies:

- Use a common RFI to inform and focus follow on RFP process.
- Execute procurement in the AI Compute Delivery teams to reduce delays and variability in procurement times while leveraging efficiency and standardization from shared standards set by the AI Compute Steering Committee.
- The AI Compute Delivery teams and host sites have significant experience in efficient procurement, align with provincial regulatory requirements, and reduce the complexity of aligning requirements across multiple host sites.
- We expect to achieve increased efficiency from previous procurements and to leverage prior process improvements and from maturation of the Alliance standards throughout the program.

### Risk 2: Delays in Hiring

There is a high demand for highly qualified personnel in the IT sectors, including research computing. Hiring technical experts to  help manage and operate the digital research infrastructures, implement the cybersecurity policies, and support the AI community may take longer than expected.

Mitigation strategy:

- Leverage existing expertise (Alliance, AI institutes, regional ARC and host sites), by re-using and adapting existing tools and services at the Alliance and AI institutes to build a common national service.

# Performance Monitoring & Collection

As mentioned in the statement of work, one of the activities that will be performed in FY 2022/23 is developing the "Monitoring & Metrics for AI National Platform". As part of this work the Alliance will be developing the required performance framework.by November 15, 2022.

# Specific Objectives

## Cybersecurity

In this first year the cybersecurity activities primarily involve planning for the coordinated integration of the 3 new AI systems into the Alliance and federation cybersecurity framework. Existing infrastructure will be leveraged where possible with additional infrastructure investment made where required such as for firewall/VPN services and cybersecurity monitoring capabilities.

## Equity, Diversity and Inclusion (EDI)

Advancing equity, diversity and inclusion (EDI) in AI is a strategic priority of the Pan-Canadian AI Strategy, and CIFAR and the AI Institutes are taking action within their respective organization and collectively to advance EDI in the Chairs, affiliated principal investigators, and trainees. EDI will also be a key component of the Allocation of AI computing resources process. EDI metrics will be monitored and reported under the performance monitoring of the PAICE to ensure that underrepresented groups receive equitable access and allocation of resources.

# Technology Profile

The technology profile is not fully known at this point and will be developed as part of the RFI and RFP activities planned for this year. The Alliance will complete this section and submit to ISED 30 days prior to issuing any RFP's.