HYPERION RESEARCH

Summary Report of Phase 1 Study to Support Compute Ontario ARC Planning

Executive Summary

Compute Ontario engaged Hyperion Research to conduct a study to help inform Compute Ontario's five-year strategic plan for providing advanced research computing (ARC) resources to Ontario researchers. For the study:

- Hyperion conducted a broad survey that yielded 231 completions from lists totaling about 900 names supplied by Compute Ontario. (This exceptionally high response rate argues that the sample closely reflects the larger universe from which it's drawn.)
- In addition, we completed in-depth interviews with 13 key ARC officials at Ontario universities, again from a list supplied by Compute Ontario.
- Finally, Hyperion assembled data to compare ARC/HPC server spending as a percentage of GDP in G8 countries, including Canada.

The **broad survey** revealed that most of the surveyed Ontario researchers are "somewhat satisfied," and a substantial, smaller number are "very satisfied," with the ARC allocations they receive. The in-depth interviews with key ARC officials at Ontario universities confirmed that most are also fairly satisfied with the system for determining allocations.

The broad survey exposed a serious disconnect between Ontario academic researchers and their counterparts in industry. Other global jurisdictions, including the two we were asked to investigate in this study for comparison with Ontario—Germany and Australia—have made major efforts in recent years to extend the benefits of ARC to domestic industry, with IP treatment a necessary component of these initiatives. Efforts by SOSCIP to promote academic-industry collaborations are impressive, but greater financial resources are needed to scale up these efforts. As the Opinion section (above) states, we recommend that Compute Ontario hold or co-sponsor seminars and workshops to promote these collaborations in Ontario.

Another important finding of the broad survey is the desire for more accelerators and other ARC resources designed to support emerging applications involving Big Data and artificial intelligence. Examples of these use cases include precision medicine, automated driving systems, smart cities and the Internet of Things. It is widely acknowledged in the global IT community that ARC/HPC is at the forefront of R&D in these fields. Compute Ontario should expand these efforts in Ontario, perhaps by collaborating with organizations such as the Pan-Canadian Artificial Intelligence Strategy at CIFAR and the Vector Institute.

Perhaps the biggest complaint of all was about the simultaneous upgrading of ARC computers, a practice that gives researchers access to up-to-date computers at the start of the cycle and outdated systems toward the end.

The **in-depth interviews** with key ARC officials at Ontario universities confirmed the main findings of the broad survey and added important commentary on these topics. An additional important need highlighted in the in-depth interviews is for ARC workforce development. (The separate study Compute Ontario engaged Malatest to complete addresses this topic in detail.)

The **comparison of Canada's and other G8 countries' spending on ARC/HPC servers as a percentage of GDP** showed that Canada's ratio is next-to-lowest. As noted in the Option section, above, Hyperion recommends that Canada aim to reach the middle of the pack in the next five years by boosting spending by about one-third.

Summary of Recommendations from Hyperion Study for Compute Ontario

Note – the Technology Study Group which oversaw this project has organized and condensed the recommendations into the form below.

1. On ARC Access and Capacity

Recommendation 1a. The overall number one recommendation by Hyperion is for Ontario to break the pattern of budgeting that often results in near-simultaneous upgrades of ARC systems, and to establish a predictable, sustainable pattern that provides upgrades as needed. This would likely mean making a reasonable budget available every year, ensuring that Compute Ontario could introduce/upgrade systems on a regular and timely basis.

Recommendation 1b. Hyperion recommends that over a five-year period, Ontario and Canada push to boost their ratios of ARC server spending to GDP by about 33%, to place them in the middle of the G8 pack, rather than today's second-to-last position.

Recommendation 1c. Compute Ontario's new five-year ARC plan include an objective to make ARC use more pervasive in Ontario, including among smaller universities and industrial firms of varying size.

2. On Access for Industrial Researchers

Recommendation 2a. The overall number two recommendation by Hyperion is that Compute Ontario should make a concerted effort to broaden industry's access to ARC resources, including human expertise to train industrial researchers in the use of ARC systems in Ontario.

Recommendation 2b. Create a program of seminars and workshops aimed at making the benefits of ARC more pervasive by attracting new scientific and industrial users.

3. On ARC Capability

Recommendation 3a. There is a need for greater access to GPUs and accelerators, especially to support emerging AI and other Big Data problems that respondents expect to become important for Ontario's scientific and industrial competitiveness. The emerging use cases that would benefit from accelerators include precision medicine, automated driving systems, smart power grids, smart cities and the Internet of Things, among others.

Recommendation 3b. Compute Ontario, CIFAR, the Vector Institute and other interested parties discuss the need for a new ARC initiative dedicated to HPDA and AI. This new initiative should be given funding to acquire a supercomputer system designed to support these workloads effectively and could be hosted in an existing ARC facility.

Recommendation 3c. Hyperion's studies of cloud computing for HPC show that public cloud infrastructures are still exploited by HPC users primarily for "embarrassingly parallel," non-mission-critical workloads and still create data security concerns for many HPC users. More ARC/HPC data centers are creating seamless environments that allow users to exploit on-premise or cloud resources. There is a need for Ontario researchers to have appropriate access to "bare metal" cloud infrastructures.

Recommendation 3d. Provide a strategy and resources for long-term data storage

4. On ARC Policy in Canada and Ontario

Recommendation 4a. It is important for Compute Ontario to continue assertively representing Ontario researchers vis-à-vis Compute Canada—and, equally important, to make sure Ontario researchers are aware of this assertive representation on an ongoing basis, through whatever communications channels exist or need to be created. It might be prudent, for example, for Ontario's five-year ARC plan to include a few objectives that would place Ontario in the national ARC limelight, including working to have an Ontario university with competency in these areas host a new national HPDA-AI ARC facility.

5. On Workforce Development

Recommendation 5a. Help development of ARC talent in Ontario.