BRITISH COLUMBIA

B.C. Gov A.I. Compute Modernization – Recommendation Package

From: Connected Service B.C. – A.I. Technical Team
Date: November 2025
Subject: Recommendation to Extend the Existing Compute Fleet with Intel Xeon 6 P-Core Processors (AMX-Enabled) for A.I. and Data Workloads

---

1. Purpose

This recommendation proposes extending, not replacing, B.C. Gov's existing Intel Xeon–based compute fleet with Intel Xeon 6 P-core (Granite Rapids) processors.

The goal is to modernize compute capability for advanced A.I. and data workloads while maintaining:

- Full compatibility with current x86 infrastructure
  Existing OpenShift / Kubernetes operational patterns

Early benchmarking on 2nd-Gen Xeon (Gold 6244) nodes has shown:

- Encouraging results for document processing and smaller transformer inference/backprop
  Clear architectural limits without AMX and modern memory bandwidth

New systems can be deployed alongside current hardware, with Kubernetes gradually routing A.I.-intensive workloads to the new nodes as adoption increases.

A more detailed integration and validation report can be provided upon request.

---

2. Why Xeon 6 P-Cores

The Xeon 6 P-core family provides a direct, low-risk modernization path from the existing 2nd-Gen Xeon platform and introduces hardware features required for modern transformer workloads.

Performance Cores (P-cores)

Full-width, high-frequency cores optimized for consistent, low-latency compute under sustained load.

AI and Matrix Acceleration (AMX)

Native BF16 / INT8 tiled matrix acceleration for quantized and mixed-precision transformer workloads not efficiently supported on the current fleet.

High Cache Density & Good Cache-per-Core Ratio

Up to 504 MB L3 per socket, with many SKUs maintaining ≥ 5 MB L3 per core.

This enables:

- Key model weights and gradients to remain cache-resident
  Reduced DRAM access during matrix operations (forward/backward)
  Faster gradient accumulation and KV-cache access during inference

Expanded Memory Subsystem

12-channel DDR5 / MRDIMM, delivering:

- ~1.6 TB/s memory bandwidth (dual-socket)
  Multi-terabyte DRAM capacity per node

Supports:

- Large-context inference
  High-throughput embedding generation
  Large activation/KV/optimizer footprints for experimentation

Data Movement Acceleration (DSA)

Offloads memory-copy/set operations so cores spend more cycles on A.I. math and less on data movement.

Next-Generation I/O & Networking

- PCIe 5.0 / CXL 2.0
  200 Gb/s+ Ethernet with RDMA (RoCEv2) for distributed workloads

Continuity and Compatibility

Still x86—compatible with existing:

- Containers and workload images
  CI/CD pipelines
  OpenShift A.I. Operator and platform tooling

Alignment with the A.I. Adoption Roadmap

These capabilities allow experimentation, fine-tuning, and inference to run entirely on government-controlled infrastructure, minimizing reliance on external cloud platforms and ensuring data remains within B.C. Gov networks.

They also enable adaptation of open-source and sovereign A.I. models for diverse ministry needs while maintaining data sovereignty.

3. Recommended Configuration Pattern

This is a **hardware pattern**, not a single SKU, giving Hosting / Platform Services flexibility while ensuring A.I. requirements are met.

Component	Recommendation	Rationale
Processor Class	Intel Xeon 6 P-cores (Granite Rapids only)	Ensures access to P-cores + AMX + AVX-512 + DSA
Sockets	Dual-socket nodes	Maximizes compute & memory density within rack power budgets
TDP Envelope	Select SKUs in two performance tiers: - 300-350W (efficiency-focused) - ~500W (maximum performance)	Allows flexibility based on datacenter power/cooling capacity; both tiers provide AMX acceleration
Core / Cache Profile	Choose SKUs with ≥ 5 MB L3 per core (examples: 32c/336MB, 64c/336MB, 72c/432MB, 96c/480MB)	Keeps weights, KV cache, and gradients hot in cache for AMX + AVX-512 workloads
Memory Configuration	Populate all 12 DDR5/MRDIMM channels (6400–8800 MT/s)	Full channel population maximizes bandwidth and supports memory-bound operations
Total DRAM per Node	1.5–3.0 TB per dual socket	Supports large-context inference, embeddings, activations, optimizer states
Networking Interfaces	1–2 × 200 Gb/s RDMA NICs per socket, expandable based on motherboard	Provides high aggregate bandwidth for distributed A.I. workloads
Switch Fabric	200 Gb/s Ethernet with RDMA (RoCEv2) and lossless fabric	Enables low-latency multi-node training and inference

Accelerator Features	Enable AMX + DSA in BIOS and OS	Required to use matrix and data-movement acceleration
NUMA Behaviour	Keep NUMA enabled (no interleaving)	Supports NUMA-aware scheduling and memory pinning for predictable latency