AI SCALING BLUEPRINT

Module 4: Scaling Systems, Teams & Operations

Northbridge Financial Group

Northbridge Agentic AI Platform: Proposal C Scale-Up

Prepared by Luke Kilpatrick

University of Waterloo AI-CTO Program

August 10, 2026

Planning Premise and Continuity

This Blueprint continues the Northbridge transformation case from Modules 2 and 3, where Proposal C was selected: an owned AI platform using open-weight models, Northbridge-controlled data and infrastructure, and staged capital gates. For Module 4, I assume Proposal C has been approved, the foundation gates have cleared, customer service is entering controlled production, and underwriting support and compliance monitoring remain behind stricter production gates.

Initiative Overview

Initiative or function name	**Northbridge Agentic AI Platform: Proposal C Scale-Up**
What it does	Provides a governed enterprise foundation for customer-service, underwriting-support, and compliance-monitoring workflows using two Northbridge-controlled open-weight models, a replaceable frontier-model slot, and the Three-Computer Agreement validation pattern for designated high-risk decisions.

Assumptions and Provenance
At submission, I am Director of Marketing at Sanctuary Cruises, a small whale-watching operator. It is my current role but not a useful enterprise scaling case for this module. I therefore continue the Northbridge program case from Modules 2 and 3. Post-approval figures are planning assumptions, not claims about a current employer; any planning baseline is validated in the first 30 days of controlled production.

Current scale

Northbridge enters this scale-up with meaningful enterprise AI already in place: approximately 8,000 employees, an in-house fraud model processing 2.1 billion transactions annually, Copilot deployed to 2,200 employees, and the Maple chatbot handling about 4,200 sessions per day. Proposal C is earlier in its lifecycle: the foundation phase is complete, customer service is entering controlled production, and underwriting support and compliance monitoring remain limited deployments.

Target scale: 12 to 18 months

Within 12 to 18 months, Northbridge should operate one enterprise AI platform across customer service, underwriting support, and compliance monitoring in its Canadian and European businesses. Customer service expands into more complex interactions; underwriting uses AI for preparation and analysis while humans retain final credit authority; compliance uses AI for triage and exception detection with accountable human review.

Two Northbridge-controlled open-weight models run primarily on Northbridge datacenter inference, with cloud reserved for burst, recovery, and experimentation. Candidate open-weight families include Llama, Qwen, Mistral, or equivalent successors, but Northbridge does not hard-code a family or version into the architecture. A third frontier-model slot remains interchangeable, preserving the option to change providers as capability, economics, availability, or jurisdictional evidence changes.

Primary scaling constraint

The first constraint to break is organizational decision-making capacity. This is a managerial-capacity and bounded-rationality problem: compute, capital, and headcount can grow faster than leader attention and cognition. The scale-up therefore must redesign where attention goes, where decisions sit, and which pilot-era heuristics no longer generalize.

Centralized expert control is useful during foundation but becomes a bottleneck when routine decisions queue behind the same small group. Northbridge therefore centralizes standards, model admission, security, and risk architecture while Workflow Pods own routine decisions inside those boundaries.

Reliability Charter

Reliability at Northbridge is not synonymous with model uptime. A system can be available and still fail by producing an unsafe hallucinated result, bypassing a required control, taking too long to reach a usable outcome, or losing the audit evidence required to defend the decision. The commitment below therefore measures whether an accepted workflow reaches a safe, auditable terminal state.

Architecture Under Test

Two distinct Northbridge-controlled open-weight models run on separate production systems; candidate families include Llama, Qwen, Mistral, or equivalent successors. The two slots must be filled from different model families and pass a common qualification gate covering capability, family independence, provenance and license, data/privacy fit, jurisdiction, inference cost, latency, hardware fit, and tested recovery. One swappable frontier-model slot comes from a distinct provider lineage and may be filled by a qualified model from OpenAI, Anthropic, Google, or another provider that passes the same operating controls. The approved-model registry records family, version, provenance where available, license, hosting posture, evaluation evidence, and next review date. For designated high-risk workflows, 3-of-3 agreement may advance a recommendation to deterministic policy validation; disagreement routes to recovery, additional retrieval, or authorized human review. We standardize the agreement architecture and qualification process, not the individual models inside it.

Reliability commitment
Northbridge will optimize for controlled workflow completion: the user receives a valid result or a policy-required safe escalation, within the workflow's service window, with all mandatory controls executed and a complete audit record. For high-risk workflows, losing one leg of Three-Computer Agreement never silently changes the control standard.

1. Reliability Commitment

Reliability Element	Northbridge Definition
Service Level Indicator (SLI)	Controlled Workflow Completion Rate (CWCR): the percentage of eligible workflow attempts that reach an allowed final state within the workflow-specific service window. They need to execute every mandatory control and write a complete audit record. Allowed final states are (1) valid completion or (2) safe escalation to an authorized human when policy or model disagreement requires it. Before release, each workflow contract must define its service window. Initial planning windows are <=30 seconds for a customer-service response, <=5 minutes for underwriting case preparation, and <=10 minutes for compliance triage; safe escalation must be initiated within the same window. These planning windows are validated and reset from the first 30 days of controlled production. User cancellations and malformed requests rejected before acceptance are excluded.
Service Level Objective (SLO)	Controlled Expansion: ≥99.5% CWCR measured monthly. Full Scale target: ≥99.9%, activated only after two consecutive quarters with less than 50% of the monthly error budget consumed and successful failover testing. High-risk workflows also carry non-negotiable guardrails: 100% of material 3CA disagreements are routed correctly, 100% of completed regulated workflows have an audit record, and zero unauthorized autonomous regulated actions are tolerated.
Correlation control and sampled false consensus	The two open models must come from distinct model families and distinct production domains; the frontier slot must come from a distinct provider lineage. Diversity is a reliability control, not a procurement preference. Each month, a precommitted sample of unanimous high-risk 3CA outputs is reviewed against authorized human ground truth. Any verified false consensus capable of producing a material regulated or customer-impacting action is P2 even if a downstream human catches it; any unauthorized action is P1. The sampled rate is trended in the weekly reliability review, but the incident response does not wait for a statistical threshold.
Monthly error budget and freeze threshold	At the 99.5% SLO, the monthly error budget is 0.5% of eligible workflow attempts. Example: at 100,000 accepted workflow attempts, the budget is 500 failed controlled completions. At 50% budget consumption, reliability work becomes the first engineering priority. At 80%, non-critical production releases freeze. At 100%, or on any P1 control/security incident, the freeze is immediate regardless of remaining budget.
Freeze authority and external dependencies	The AI Platform Engineering Lead may declare a freeze without executive approval. The same role may lift a normal reliability freeze after the cause is mitigated, required tests pass, no P1/P2 incident remains open, and the trailing 24-hour burn rate is back within target. Security incidents require the AI Security Lead/CISO delegate to co-approve the lift; regulated-workflow control failures also require the accountable Risk/Compliance Owner. User-impacting dependency failures count against the service error budget even when Northbridge did not cause them; attribution is tracked separately so provider reliability can be managed without redefining the customer promise.

The monthly error budget is percentage-based so it scales automatically with volume. This avoids a static failure count becoming meaningless as customer-service usage expands and underwriting/compliance workloads enter production.

2. Failure Modes and Three-Computer Agreement

The primary SLI is intentionally broader than availability. Northbridge treats four outcomes as reliability failures: the workflow is unavailable; it reaches an invalid or unsafe terminal state; it exceeds the service window without a safe escalation; or it cannot produce the audit record required for the action taken. Model disagreement, by itself, is not failure. In a high-risk workflow, a correctly detected 3CA disagreement that routes to a human is the control working as designed.

3CA condition	Required system behavior	Reliability treatment
3 of 3 models agree	Advance only to deterministic policy validation; authorization remains with the applicable rules and/or human decision-maker.	Valid only if controls execute, the audit record is complete, and the output remains subject to monthly false-consensus sampling.
Any material disagreement	Retrieve/recalculate once where appropriate; if disagreement remains, route to the authorized domain reviewer.	Safe escalation counts as a valid terminal state, but disagreement rate is tracked as a model-quality signal.
One open-model service unavailable	Fail over to its tested DR target. Do not substitute a duplicate of the other open model merely to manufacture three votes.	Counts against the budget if the workflow cannot reach a valid terminal state in time.
Frontier provider unavailable	Switch the frontier slot to a pre-qualified alternate model. If no qualified replacement is available, route the high-risk workflow to human review.	User impact counts against the SLO; provider attribution is recorded separately.
False consensus / control bypass	Stop affected autonomy, preserve evidence, and invoke incident response. If a unanimous result would have produced an invalid action, treat as a near miss even if a downstream human catches it.	P1 if an unauthorized or unsafe action occurs. Otherwise P2. A verified false consensus capable of producing a material regulated or customer-impacting action is P2 even if caught downstream; incident response does not wait for the sampled rate to rise.

2.1 Correlation Risk: Agreement Is Not Ground Truth

Agreement does not prove correctness. The three models can be trained on overlapping corpora, share architectural priors, and read the same retrieved context, so their errors can be correlated and unanimity can reflect shared assumptions rather than truth. In Bayesian terms, a 3CA result updates our prior; it does not replace judgment or drive the probability of error to zero. That is why disagreement routes to an authorized human instead of a tiebreak, why the model families and provider lineages are deliberately diverse, and why unanimous high-risk outputs are sampled against human ground truth.

Infrastructure and Correlation Requirement
The two Northbridge open models must come from distinct model families and must not share a single production failure domain. Each has a tested recovery target in a second Northbridge site or approved cloud capacity. The frontier slot comes from a distinct provider lineage and has at least one pre-qualified alternate. A component outage may trigger failover or human fallback; it never converts a 3-of-3 control into a 2-of-3 majority vote.

3. Escalation and Recovery Structure

Incident response prioritizes safe state before service restoration. The first responder classifies severity, assigns an Incident Commander for P1/P2 events, and disables autonomous action when the integrity of a control is uncertain. A business deadline is never grounds to bypass 3CA, deterministic policy checks, or required human authorization.

Severity	What triggers it	Who responds	Target recovery	Postmortem?
P1: Critical	Systemwide outage or imminent monthly SLO breach; customer-data exposure; unauthorized regulated/customer action; loss of audit integrity; 3CA/control bypass; or failure mode requiring immediate shutdown of autonomous action.	On-call Platform Engineer; AI Platform Engineering Lead as Incident Commander; AI Security Lead/CISO delegate when security is implicated; affected Workflow Owner; Risk/Compliance Owner; AI Transformation Lead/CIO notified immediately.	Restore service or safe human fallback within 60 minutes; containment begins immediately.	Yes. Within 3 business days.
P2: High	Material degradation in one workflow; error-budget burn >2x target; loss of an inference component; repeated latency failure; or a verified false consensus with material regulated/customer impact, even if caught downstream.	On-call Platform Engineer; affected Workflow Owner; AI Platform Engineering Lead within 30 minutes; AI Security or Risk/Compliance owner as required by cause.	Mitigate or restore within 4 hours.	Yes. Lightweight if no customer or control impact.
P3: Watch	Degraded performance with no immediate SLO breach; burn >1× but <2× target; rising 3CA disagreement; capacity headroom below plan; intermittent provider errors; or recurring non-critical operational toil.	Workflow engineer/owner and Platform Reliability Engineer; reviewed in the next operations cadence.	Mitigate by next business day; permanent fix planned in sprint.	No, unless recurring.

Incident path:

- Detect through telemetry, automated control checks, provider health, capacity monitoring, or user report.
- Classify P1/P2/P3 within 10 minutes of acknowledgement; assign the AI Platform Engineering Lead or designated delegate as Incident Commander for P1/P2.
- Move to a safe state first: stop autonomous action, fail over a model/inference service, or route the workflow to an authorized human.
- Restore service only after mandatory controls, logging, and rollback/failover behavior are verified.
- Communicate status to workflow owners and executive/risk stakeholders at the cadence appropriate to severity.
- For required postmortems, publish a blameless review with root cause, contributing conditions, action owner, and due date; track actions to closure.

4. Reliability in Engineering and Product Planning

Reliability is a normal planning input, not an incident-only activity. The Platform Reliability Lead maintains the SLO dashboard and error-budget ledger; workflow owners see the same data used by engineering and Finance. Reliability capacity therefore increases automatically as the budget is consumed instead of being negotiated after a failure.

Cadence / gate	Required reliability activity	Decision authority
Every production release	Regression/evaluation suite; audit-log validation; rollback plan; dependency/failover check. High-risk releases must prove 3CA independence and the disagreement-to-human path before production.	Platform Reliability Lead can block the release.
Weekly reliability review	Review error-budget burn, P1/P2 actions, provider health, GPU/capacity headroom, 3CA disagreement trend, sampled false-consensus rate, and recurring operational toil.	AI Platform Engineering Lead prioritizes reliability work and can reduce release scope.
Monthly SLO review	Close the error-budget period, review attributable vs. dependency failures, review the monthly unanimous-output ground-truth sample, confirm whether the 99.5% SLO is met, and decide whether scaling gates remain open.	AI Platform Engineering Lead + Workflow Owners; AI Transformation Lead receives exceptions/escalations.
Scale gate	Do not move a workflow to broader production until it demonstrates its SLO, failover path, complete auditability, and required 3CA/human-control behavior under representative peak load.	AI Transformation Lead may approve expansion only after Engineering and Risk attest that the gate is met.

Release-block rule
The Platform Reliability Lead has an explicit veto when the error-budget freeze is active, a P1/P2 control issue is unresolved, rollback/failover has not been demonstrated, required audit evidence is missing, or a high-risk workflow cannot preserve its 3-of-3 validation/human fallback path. Schedule pressure cannot override these conditions.

This charter deliberately makes safe escalation part of reliability. It prevents the organization from optimizing for automation rate at the expense of control: when the models disagree or infrastructure degrades, the reliable behavior is to preserve the decision boundary, not to force the system to act.

Organizational Resilience Plan

Organizational design principle
Centralize the platform, standards, and irreversible risk decisions; distribute workflow ownership and routine reversible decisions to cross-functional pods. The organization should mirror the modular architecture Northbridge wants the system to have, rather than reproducing a hub-and-spoke approval queue.

1. Organizational Diagnostic

At the start of controlled production, governance and the first senior hires are in place, but coordination still depends on pilot-era relationships. The diagnostic below names the gap between that operating condition and what the 12 to 18 month target requires.

Element	Current State	What Scale Demands	Gap
Strategy	Proposal C has board approval and a common platform direction, but Northbridge's AI estate was built in functional silos: fraud, Copilot, and Maple each have different sponsors, controls, and measures of success. The first new workflow can still be translated through the central transformation team.	One shared scaling objective: a common AI control plane with domain-owned outcomes. Customer service, underwriting, and compliance must use the same reliability, cost, model-risk, and audit standards while retaining domain accountability.	The strategy is shared at the platform level but not yet translated into a common operating scorecard or domain ownership model. If unchanged, Conway's Law predicts vertical AI islands connected by a central approval layer. A biweekly AI Scale Review will bridge those islands: one liaison from each Workflow Pod, Platform & Enablement, Risk, Data, and FinOps will review the same scorecard and decision log, resolve cross-domain dependencies, and push routine decisions back to the accountable pod.
Structure	The proven six-person fraud ML team is fully committed to the fraud platform and should remain protected. The new platform team is still small; Risk, Data, Infrastructure, and business SMEs sit in separate reporting lines. Central AI currently acts as a broker across structural holes between fraud, Copilot, Maple, Risk, Data, and the business lines. That brokerage makes same-day escalation possible today, but it is also the bottleneck we will hit at scale.	Standing workflow pods with named business, technical, data, and risk ownership; a shared Platform & Enablement team; and explicit interfaces between pods and the platform. Pods need dense internal ties for routine work, while the platform team keeps the cross-domain ties that only it can hold. Coordination should occur through defined service contracts and decision rights, not personal access.	No durable pod structure yet. The central AI team risks becoming both integration team and approval queue, while domain knowledge remains outside the teams shipping the workflows.
Processes	Governance, risk tiering, and the Reliability Charter now exist, but the first owned workflow can still be launched through expert attention and bespoke coordination. Model evaluation, 3CA configuration, data onboarding, cost attribution, failover, and human-escalation design are not yet a fully transferable launch system.	A repeatable workflow lifecycle: intake and risk tiering, data contract, approved-model selection, evaluation pack, SLO/error budget, 3CA rule where required, FinOps tags, rollback/failover, human escalation, and production-readiness gate.	The process is partly documented but still dependent on people who know how to assemble the pieces. A successful first workflow could become a handcrafted exception rather than a template.

People and Incentives

Element	Current State	What Scale Demands	Gap
Rewards	Business teams are rewarded for service outcomes and speed; Engineering for delivery; Risk for avoiding control failures; Finance for cost discipline; Infrastructure for availability. Those measures are individually rational but can create pressure to automate faster, centralize more approvals, or cut capacity without a shared trade-off.	Joint workflow scorecards that balance business outcome, CWCR reliability, cost per completed workflow, control adherence, safe escalation, and decision autonomy. Psychological safety is a reliability control, not a culture nicety: an engineer who suspects a control gap must be rewarded for speaking up and stopping the line. Silence is the organizational version of a 3CA leg failing silently.	No shared cross-functional incentive yet. Local optimization can either encourage bypassing controls or make governance the safest place to say no.
People	The fraud team holds deep production ML experience but has no spare delivery capacity. The initial senior AI hires carry disproportionate knowledge of the new open-model stack, 3CA/evaluation architecture, and platform design. GPU/datacenter operations and regulated-domain judgment are also concentrated in specialist roles.	Every critical capability has a primary and trained deputy; senior engineers mentor rather than become permanent gatekeepers; workflow pods contain domain expertise; runbooks, architecture decisions, and model-evaluation evidence are transferable. Fraud specialists advise and help interview/train without being reassigned.	Key-person risk remains high in the new platform. Losing one senior architect, model engineer, infrastructure specialist, or domain approver could materially slow scaling even if the technology continues to run.

2. Scaling Readiness Assessment

Readiness Question	Northbridge Decision
Atomic unit of scaling	A cross-functional AI Workflow Pod operating on the shared Northbridge AI platform. Each pod has a named Workflow Product Owner, AI/Application Technical Lead, one or more AI/application engineers, a domain SME, a data owner/steward, and a Risk/Compliance partner appropriate to the workflow. Platform/SRE, security, model-serving, and FinOps capabilities remain shared services. The pod owns the business outcome, workflow-level SLO, unit economics, data contract, human-escalation path, and releases inside its delegated risk envelope. Scaling means adding another pod on the same platform, not cloning another AI stack.
Competency trap	Trusted-expert centralization. Northbridge's strongest internal AI proof point is the four-year-old fraud platform run by a six-person ML team. Second-order thinking makes the trap visible: the first-order effect of routing every AI decision through that team and the new senior architects is safety; the second-order effect is an approval queue that pulls scarce people off the fraud system and prevents customer-service, underwriting, and compliance teams from building their own operating capability. The fraud team therefore stays protected and advisory; Platform & Enablement sets reusable standards; Workflow Pods own execution and routine decisions within those standards.
Ambidexterity: exploit vs. explore	Northbridge begins from an exploit-heavy position: the fraud system is mission critical, core data remediation has consumed capacity, and a regulated bank naturally prioritizes stability. Scaling Proposal C will intensify that pull as SLOs, incidents, GPU utilization, audits, and cost targets become daily work. Northbridge will reserve 20% of central AI Platform & Enablement engineering capacity each quarter for exploration - model replacement tests, open-model tuning, inference efficiency, 3CA false-consensus analysis, and sandboxed workflow experiments. The commitment can be temporarily suspended only during a Reliability Charter release freeze or active P1 response; it cannot be consumed permanently by the production backlog.
Fastest warning signal and escalation latency	Routine Central Escalation Rate (RCER): the percentage of reversible workflow decisions designated as delegated or policy-governed that still require the central AI team, Governance Council, or executive owner to decide. The AI Transformation Lead owns the decision taxonomy, and the denominator includes every decision class designated 'Delegate' or 'Set a Policy' whether it is resolved locally or escalated, so the metric cannot improve by quietly redefining decisions out of scope. In controlled production, central review is intentionally the default for non-standard decisions and direct access means most escalations can reach an authorized person the same business day. That is a small-network advantage, not a scalable process. By the end of Controlled Expansion, at least 80% of routine reversible decisions should be resolved inside the Workflow Pod or by published policy (RCER <20%). RCER above 25% for two consecutive weeks, or median time-to-authorized-decision above one business day for routine matters, triggers a structural review rather than more headcount in the approval queue.

Why the competency trap is real
The prior Northbridge recommendation explicitly protected the six-person fraud ML team: it should help interview and train the new organization, not be pulled onto new projects. Organizational resilience depends on keeping that boundary. The answer to scarce expertise is to turn expertise into standards, mentoring, runbooks, and delegated capability - not to make the experts a permanent approval service.

3. Target Operating Structure: Platform + Workflow Pods

Conway's Law is useful here as a design test. If Northbridge keeps one central AI group responsible for every integration and every decision, the technical system will become one tightly coupled central service. If the

organization separates stable platform responsibilities from domain workflow ownership, the architecture can remain a common control plane with loosely coupled workflows.

Layer	Owns	Does Not Own
AI Platform & Enablement	Model/provider qualification; open-model serving and lifecycle; datacenter inference and DR; shared 3CA components; security and observability; evaluation tooling; common FinOps standards; enterprise policy implementation.	Customer-service, underwriting, or compliance business outcomes; routine workflow configuration inside an approved envelope; domain-specific human judgments.
Workflow Pod	Workflow outcome; user experience; workflow-level SLO; unit cost; approved data contract; agent flow; model choice from the approved set; human escalation; release decisions inside delegated risk/cost boundaries.	New enterprise providers; changes to risk appetite; new restricted data classes; changes from recommendation to autonomous regulated action; shared platform infrastructure.
Governance / Executive Risk	Risk appetite; high-stakes and hard-to-reverse decisions; new autonomy classes; jurisdictional constraints; material exceptions to enterprise policy.	Routine model switching, ordinary release approvals, or daily workflow operations that already fall inside policy.

4. Resilience Commitments and Scale Gate

- Before a workflow leaves controlled production, it must have a chartered Workflow Pod with named business, technical, data, and risk owners plus a deputy for every role that can block a regulated production decision.
- No critical platform capability may remain dependent on a single individual at Full Scale. Open-model operations, 3CA/evaluation, GPU inference/DR, security response, and each regulated workflow must have at least two trained, authorized operators or decision owners.
- The six-person fraud ML team remains protected from routine Proposal C delivery. Its contribution is standards transfer, interviewing, mentoring, and periodic design review - not backlog execution or standing approval duty.
- Twenty percent of central platform engineering capacity remains reserved for exploration except during an active P1 incident or Reliability Charter freeze.
- RCER is reviewed weekly. If it exceeds 25% for two consecutive weeks, the response is to identify the recurring decision type and delegate it, codify policy, or redesign the interface. Adding another approver is not considered a fix.

Organizational scale gate
Northbridge does not move from Controlled Expansion to Full Scale until all three priority workflows operate through chartered pods, critical capabilities have primary-and-deputy coverage, and RCER remains below 20% for four consecutive weeks. A technically healthy platform with a growing central decision queue has not met the scaling gate.

This plan treats organizational resilience as an operating property, not an org-chart exercise. The objective is to preserve expert standards while removing expert dependence: the platform team makes the safe path easy, Workflow Pods own outcomes inside that path, and senior attention is reserved for decisions whose stakes or irreversibility actually justify it.

FinOps Operating Model

Northbridge chose Proposal C partly to control long-run economics rather than rent every unit of intelligence from one vendor. We are applying the platform through a Lean and Agile operating model: AI should improve flow and support frontline judgment, not become a Taylorist surveillance-and-control layer. The FinOps progression is Inform through visibility and attribution, Optimize through routing and waste removal, then Operate by making cost-to-value trade-offs part of normal planning. The question is not how to make AI cheapest; it is how to align spend with defensible business value. Operator examples show that open-weight routes can be materially cheaper than frontier-model use in some markets. We treat that as an observation, not a universal price ratio; the strategic point is that the difference can be large enough to justify owned base-load inference when utilization and governance support it.

FinOps operating principle
Own the base load; rent the peaks; cache what is deterministic; attribute every dollar to a workflow. Northbridge will run steady-state open-model inference primarily in its own datacenters, use cloud for burst, DR, and experimentation, route repeated deterministic requests to cache where controls permit, and reserve the frontier model and full Three-Computer Agreement for higher-risk work where the cost of control materially exceeds the cost of additional validation.

1. Inform: Where Northbridge Is Today

Northbridge can see spend at the contract and platform level, but the new platform does not yet have a common allocation model that links shared compute, frontier usage, human review, and platform toil to a completed business workflow. That is the FinOps gap we must close before scale.

Pain symptom	Applies?	Specific Northbridge manifestation
Unexplained cloud or AI cost spikes	**Partially**	Vendor invoices can be traced at the contract/provider level, but frontier-model usage, cloud burst, evaluation runs, and retried agent steps are not yet consistently attributable to a specific workflow and outcome.
Weak or inconsistent tagging	**Yes**	Legacy Copilot and Maple were not designed around a shared tagging standard. The new platform does not yet enforce workflow, pod, environment, model, risk tier, jurisdiction, and cost-center metadata on every inference and infrastructure allocation.
Always-on non-production environments	**Partially**	Open-model test clusters and cloud sandboxes can remain provisioned for engineering convenience. There is not yet a universal owner/expiry rule for non-production GPU capacity.
Costs not linked to product metrics	**Yes - primary pain**	The prior brief correctly identified cost per completed workflow as the right metric, but current spend is still viewed as licenses, API usage, people, and infrastructure. Northbridge cannot yet answer what one successful customer-service, underwriting, or compliance workflow costs end to end.
Finance vs. Engineering tension	**Partially**	Finance naturally sees fixed capital plus variable AI spend; Engineering sees capacity, resilience, and model-quality requirements. Without a shared unit metric, both can be locally correct while arguing about different things.
Scaling before proving value	**No**	Proposal C is explicitly staged behind data, governance, talent, reliability, and operating gates. The FinOps model adds economic gates rather than assuming more usage equals more value.

Highest-priority FinOps investment
Build attributable unit economics before scale: every production request and every material shared-infrastructure allocation must carry enough metadata to answer who spent it, on which workflow, in which environment and jurisdiction, using which model, and what controlled business outcome resulted.

2. Optimize: Fully Loaded Cost per Completed Workflow

The unit-cost model annualizes owned hardware through three-year depreciation so operating economics do not double-count capital purchases. It also includes people, data, tooling, cloud burst, frontier usage, and operational toil because those costs do not disappear just because the inference hardware is owned.

Planning assumption	Value used in this model
Current workflow volume	1.0M successfully completed workflows/year, approximately two-thirds of the 4,200 daily Maple sessions in the source case after allowing for sessions that do not become completed workflows. This is an annualized controlled-production planning baseline.
12-month workflow volume	4.0M successfully completed workflows/year across customer service, underwriting support, and compliance monitoring.
Owned inference hardware	Two independent open-model production domains. Current plan assumes 16 high-memory accelerators plus hosts/network/storage (~$1.5M capital base); Full Scale uses ~24 accelerators (~$2.1M capital base) as a planning envelope. Hardware is depreciated over three years. Final accelerator generation and count are gated by controlled-production benchmarks for tokens per completed workflow, peak concurrency, p95 latency, utilization, and N+1 failover headroom; 24 accelerators is not a procurement commitment.
Cloud posture	Cloud is for DR, burst capacity, model experimentation, and temporary capacity. Stable open-model base load is expected to run in Northbridge datacenters.
Risk-tiered routing mix and Three-Computer Agreement	Target routing assumption: 25% of completed workflows use a deterministic or cached response tier for high-frequency repeated requests; 60% use the lowest-cost qualified single open model/control pattern; 15% invoke full 3CA with both open models plus the swappable frontier slot. As a planning sensitivity, removing the cache tier raises modeled Full Scale cost from about $1.34 to about $1.47 per completed workflow by requiring additional inference and capacity for repeated work.
People and toil	Includes incremental AI Platform & Enablement labor plus allocated SRE/ML/FinOps/security effort, and explicit operational toil/human review caused by incidents, model disagreements, evaluation, and on-call work.

Annualized Cost Model

Annualized cost category	Current ($M)	12-mo ($M)	What changes at scale
Hardware depreciation	0.50	0.70	Owned base-load capacity expands; three-year depreciation used.
Datacenter power, cooling & support	0.35	0.45	Higher utilization and support footprint as owned inference grows.
Cloud burst, DR & experimentation	0.30	0.35	Cloud remains a variable safety valve, not the default base load.
Frontier-model usage	0.45	0.65	Grows with high-risk 3CA volume and qualified frontier use.
Data, security, observability & evaluation tooling	0.45	0.55	Shared platform services scale more slowly than workflow volume because Northbridge uses negotiated enterprise tiers for core observability and caps expensive evaluation through risk-based sampling rather than evaluating every workflow at full depth.
AI-specific people allocation	1.90	2.30	Adds platform/pod capacity while avoiding linear headcount growth.
Operational toil & human review	0.30	0.35	Includes incident response, evaluation, 3CA disagreement review, and cost-management toil.
Fully loaded annual run cost	**4.25**	**5.35**	Operating run rate; prior build/acquisition capital remains governed in the Scaling Investment Case.

Unit economics commitment
Current planning baseline: $4.25M / 1.0M successful workflows = $4.25 per completed workflow. 12-month model: $5.35M / 4.0M workflows = $1.34. Full Scale commitment: reach <=$1.50 per successfully completed workflow without breaching the Reliability Charter, 3CA requirements, auditability, or human decision boundaries. The $1.34 model assumes the 25% deterministic/cache routing tier; without it, the sensitivity rises to about $1.47.

3. Operate: Ownership, Tension and Leading Signal

Operating element	Northbridge rule
Cost ownership	Director, AI Platform & Enablement owns the shared AI cost envelope and capacity plan. Each Workflow Pod Owner owns its workflow-level cost per completed workflow. The Platform Director may enforce tags, budgets/quotas, model-routing policy, non-production shutdown schedules, and capacity limits; Workflow Owners may change prompts, retrieval, model choice within the approved registry, caching/batching, and workflow design to improve unit economics. Finance validates allocation and forecast discipline but does not choose the technical architecture.
Attribution standard	100% of production inference and material shared infrastructure must carry workflow, pod/cost-center, environment, model/provider, risk tier, and jurisdiction metadata before Full Scale. Owned GPU cost is allocated by measured or reserved accelerator time; frontier tokens and cloud resources flow directly to the invoking workflow. Non-production resources require an owner and expiry date.
Review cadence	Workflow Pods review cost and value weekly. Finance, the Platform Director, and Workflow Owners hold a monthly FinOps review covering CPW, volume, forecast variance, GPU utilization/headroom, frontier concentration, 3CA share, and low-value toil. Capital capacity decisions remain tied to the Scaling Investment Case gates.
Tension point	The hardest trade-off is cost versus reliability/control: reducing redundancy, using a weaker model, skipping a 3CA leg, or running DR cold may reduce spend but can violate the Reliability Charter or risk controls. The explicit order is: regulatory/control floor first, Reliability Charter second, unit economics third, delivery speed fourth. Cost optimizations inside the first two boundaries are delegated; crossing them requires the accountable Risk/Compliance and Reliability owners, not a budget argument.
Leading indicator	7-day rolling cost per completed workflow (CPW) variance. A >15% unfavorable variance to the workflow budget for seven consecutive days, or a projected monthly spend >10% above budget, triggers a same-week FinOps review. A >25% CPW variance for two consecutive weeks pauses new scale-out until the cause is understood. We review this before the lagging monthly finance close.

4. Lean Optimization Targets

The management failure modes are the test for these targets. We do not start with technology, automate a broken process, add dashboards that create clutter, use AI mainly for surveillance and control, or scale before operational value is visible. The table therefore separates flow improvement from task automation and measures the business process, not the number of AI interactions.

Outcome	Baseline → 12-month target	AI role / process need	Guardrail against low-value automation
Customer-service human handoff rate	38% → <=25%	Replace repetitive lookup/drafting; support complex judgment. Flow improvement.	Do not automate around a broken routing path. Workflow Owner reviews handoff reasons monthly; account actions remain inside pre-approved action boundaries and Reliability Charter escalation rules.
Underwriting analyst preparation time per case	90 min → <=45 min	Supports judgment; gathers and synthesizes evidence. Task automation.	Final credit authority remains human/deterministic. No model approval-rate target; quality sample and override review prevent optimizing speed by weakening evidence.
Compliance triage cycle time	2 business days → <=4 hours	Supports judgment; prioritizes and summarizes regulatory material. Flow improvement.	Compliance owner signs material action. Monthly sample audit checks missed/false escalations; 3CA is applied when the workflow crosses the designated high-risk threshold.

FinOps scale gate
Northbridge does not move to Full Scale until (1) >=95% of production AI spend is attributable to a workflow and owner for two consecutive months, (2) modeled/actual cost per completed workflow is <=$1.50 at representative volume, (3) the 7-day CPW variance is inside the 15% tolerance, and (4) the three Lean targets are moving in the intended direction without a reliability, control, or human-authority regression.

The FinOps model turns Proposal C's ownership thesis into an operating discipline. Owning compute only creates strategic leverage if Northbridge knows what that compute costs, what business outcome consumed it, and who has the authority to change the trade-off. The economic goal is not the cheapest token; it is the lowest defensible cost for a safely completed workflow.

Distributed Decision Charter

Northbridge will scale by moving decisions to the lowest level that has the evidence and authority to make them safely. Reversible choices move to Workflow Pods, repeatable hard-to-reverse choices become policy, high-stakes reversible choices become bounded experiments, and high-stakes irreversible changes remain senior decisions. The AI Transformation Lead owns the decision architecture - not every decision.

Decision principle
Centralize risk appetite, platform standards, jurisdictional boundaries, and irreversible autonomy decisions. Distribute routine model, release, data, and cost choices inside those boundaries. Three-Computer Agreement informs a decision; it never changes who is authorized to make it.

1. Decision Inventory

Decision	Who Currently Makes It	Best Information Lives With	Cost of Wrong Level / Delay
Switch to another model already approved for the same workflow risk tier, jurisdiction, and cost envelope.	Workflow Tech Lead proposes; central Platform/Transformation lead still reviews.	Workflow Pod: evaluation, quality, SLO/error budget, and cost-per-workflow telemetry.	Central review slows a two-way-door choice, delays quality/cost improvements, and builds an approval queue.
Add an approved internal data source to retrieval without changing data class, training use, or jurisdiction.	Data Owner/CDO governance + central AI team are still often consulted.	Domain SME, Data Steward, Workflow Pod, local Risk/Privacy partner.	Delay leaves stale context; over-centralization creates a data queue; unguided local action risks access/sovereignty errors.
Pilot a materially different open-weight or frontier model/provider.	Platform & Enablement + AI Transformation Lead, with Security/Procurement/Risk/ Workflow Owner.	Evaluation harness, 3CA disagreement data, security, FinOps, and workflow outcomes.	Delay creates capability/cost lock-in; a wrong-level decision can create premature vendor, security, or sovereignty commitments.
Expand a workflow from recommendation/support to autonomous regulated or customer-impacting action.	Governance/Executive Risk + accountable business/risk executive.	Workflow outcomes, 3CA/override evidence, SLO history, Legal/Risk/Compliance, human process owner.	Slow approval defers value; approval too low can create persistent regulatory, customer, financial, and reputational exposure.

2. Decision Attention Grid

Northbridge uses reversibility as the operating test. A reversible choice inside an approved envelope should not consume senior attention; a choice that changes risk appetite, data sovereignty, or regulated autonomy should.

	REVERSIBLE	HARD TO UNDO
HIGH STAKES	**EXPERIMENT** Pilot a new model/provider. AI Transformation Lead sets hypothesis, evidence bar, data/traffic boundary, budget, and rollback; Platform + Workflow Pod run the test.	**YOUR FOCUS** Expand regulated/customer-impacting autonomy. Executive Risk + accountable business executive own the call. 3CA, SLO, legal, override, and control evidence inform it; none authorizes it.
LOWER STAKES	**DELEGATE** Switch among pre-approved models and approve ordinary releases inside the same risk, jurisdiction, reliability, and cost envelope. Workflow Pod owns the choice and rollback.	**SET A POLICY** Onboard retrieval sources inside an approved data class, purpose, and jurisdiction. CDO/Legal/InfoSec define the rule; Data Owner + Workflow Pod apply it. Exceptions escalate.

3CA decision boundary
Unanimous 3CA may advance a high-risk recommendation to deterministic policy validation. It never authorizes credit or money movement, waives a jurisdiction rule, or lets a Workflow Pod expand its own autonomy class. The reliability charter treats unanimity as evidence, not truth: correlated errors can survive 3-of-3 agreement, so sampled false consensus is monitored and material cases trigger incident review.

3. What I Will Stop Deciding

Decision I Stop Making	New Owner	What Replaces My Judgment	When It Comes Back Up
Routine model switching among models already approved for the workflow envelope.	Workflow Technical Lead / Product Owner.	Approved-model registry; evaluation pack; SLO/error budget; quality/3CA telemetry; <=$1.50 CPW target; rollback.	New provider/family, risk-tier change, jurisdiction exception, or sustained SLO/cost breach.
Ordinary production release approval inside the delegated autonomy/data envelope.	Workflow Product Owner + Technical Lead; Reliability Lead retains veto.	Automated tests/evals, audit evidence, rollback/failover, Reliability Charter freeze policy, named controls.	Active freeze/P1-P2, autonomy/data-class/shared-control change, or risk-appetite change.
Case-by-case approval of retrieval sources inside an approved data class/purpose/jurisdiction.	Data Owner/Steward + Workflow Pod Risk partner.	Data contract, classification/purpose rules, access/retention controls, jurisdiction profile, logged evidence.	New/restricted class, training use, cross-border transfer, conflicting local rule, or material exception.

4. Anchor to Break: "Safety Requires Central Approval"

The first operating phase set a useful anchor: central approval signaled safety while controls, runbooks, evaluation evidence, and ownership were immature. Early anchors silently become reference points. At scale, this one biases us toward escalation even when a decision is reversible and already inside policy, so the mindset that worked in the pilot stops generalizing.

- Central AI owns the approved-model registry, risk tiers, autonomy classes, enterprise controls, evaluation standards, and jurisdiction profiles - not routine choices made inside them.
- Repeated escalation of the same reversible decision is a design defect: delegate it, codify policy, or redesign the interface rather than add an approver.
- A P1/P2 control failure may temporarily pull a delegated decision back to the center; delegation returns only after the control is corrected.

5. Regulatory, Data-Sovereignty, and Cultural Constraints

Northbridge operates across Canadian and European businesses. Enterprise policy therefore defines a minimum control set, while jurisdiction profiles can be stricter. Open weights do not solve governance by themselves: the approved-model registry records family, lineage/provenance, licensing, hosting posture, and evaluation evidence, and a local profile can refuse a model even when it passes enterprise qualification. Local Legal, Data, and Risk owners decide whether a workload meets the local bar. Because norms for challenging central authority vary across teams and jurisdictions, delegated authority and stop-work rights are written rather than assumed.

Decision Type	Enterprise Standard	Local Authority / Constraint
Open-model hosting / data location	Approved-model registry records model family, lineage/provenance where available, license, hosting posture, auditability, evaluation evidence, and tested recovery. Northbridge-controlled serving must avoid a single production failure domain.	Local Data/Legal/Risk profile sets the approved region/site, whether cloud DR is permitted, and may reject an otherwise qualified model on provenance or sovereignty grounds.
Frontier model in third 3CA slot	Must pass enterprise security, evaluation, contract, reliability, FinOps, and provider-lineage qualification. The slot remains swappable by design.	Global qualification is not enough. The local profile must permit the data class, location, purpose, and provenance. A qualified local-language or local-domain model may be preferred when it produces better local outcomes.
Data source / retention / training use	Purpose, classification, access, retention, auditability, and separation of retrieval from training use.	Local Data/Privacy/Risk decides exceptions/cross-border use; new/restricted classes cannot be delegated by the pod.
Autonomous customer/regulatory action	No workflow expands its own autonomy; 3CA validates but does not authorize.	Local Business/Risk/Legal may impose a stricter autonomy ceiling or human-review requirement.
Challenge and local dissent	Delegated authority and stop-work rights are written and visible. Any Workflow Pod member may raise a suspected control gap and pause work inside the published escalation path without being penalized for slowing delivery.	Local Risk, Data, Legal, and Business owners may challenge or block a global platform choice when local evidence requires it. Dissent is logged as operational evidence, not treated as obstruction; a local profile may always be stricter than the enterprise floor.

6. Governance and Scale Gate

The AI Transformation Lead owns the decision-rights matrix and reviews it monthly with the Organizational Resilience Plan. The role may delegate recurring reversible decisions, require policy codification, and temporarily centralize a delegated class after a material control failure. Governance Council attention is reserved for risk appetite, autonomy classes, jurisdiction exceptions, and other high-stakes hard-to-reverse decisions.

> **Decision scale gate**
> No Full Scale until all four inventory decisions have a named owner and evidence path; RCER is <20% for four consecutive weeks; median routine decision latency is <=1 business day; and there are zero unauthorized autonomy, data-sovereignty, or risk-tier changes in the same period.

The operating test is whether senior leaders see fewer routine AI decisions as volume grows. Their attention should concentrate on choices that change Northbridge's risk posture or are difficult to reverse.

Scaling Investment Case

The Reliability Charter, Organizational Resilience Plan, FinOps Operating Model, and Distributed Decision Charter now become investment commitments. We use one set of measures across operations and capital allocation: reliability, organizational resilience, FinOps, and decision rights. Further capital is released only when those controls survive growth.

Commitments and Targets

These are the outcomes we will use to decide whether Proposal C is scaling safely and economically. Every commitment below comes from the Reliability Charter, Organizational Resilience Plan, FinOps Operating Model, and Distributed Decision Charter, so the investment case uses the same evidence as the operating teams.

Commitment	Current Baseline	12-Month Target	Source
Reliability: Controlled Workflow Completion Rate (CWCR)	Foundation begins under the 99.5% CWCR release gate. The AI Platform Engineering Lead owns instrumentation and must publish the measured CWCR and sampled false-consensus baseline within 60 days of controlled-production start, no later than October 15, 2026.	>=99.9% monthly CWCR after two consecutive quarters with <50% error-budget consumption and successful failover testing; 100% material 3CA disagreements routed correctly; monthly unanimous high-risk sample reviewed against human ground truth; any verified material false consensus is P2 even if caught downstream; any unauthorized autonomous regulated action is P1; zero unauthorized autonomous regulated actions.	Reliability Charter
Operational efficiency: Lean workflow outcomes	Planning baseline: customer-service handoff 38%; underwriting preparation 90 min/case; compliance triage 2 business days.	Customer-service handoff <=25%; underwriting preparation <=45 min/case; compliance triage <=4 hours, with no weakening of human authority or control quality.	FinOps Operating Model
Unit economics: cost per successfully completed workflow	Planning baseline: 1.0M workflows/year at a modeled fully loaded $4.25/workflow ($4.25M annualized).	4.0M workflows/year at <=$1.50/workflow; current model reaches $1.34 on $5.35M annualized cost.	FinOps Operating Model
Organizational resilience: repeatable ownership without expert dependency	One controlled-production workflow still benefits from direct central-expert coordination; three durable Workflow Pods and full primary/deputy coverage are not yet in place.	All three priority workflows operate through chartered pods; every critical platform/regulated-decision role has primary + deputy coverage; RCER <20% for four consecutive weeks.	Organizational Resilience Plan
Leading indicator: Routine Central Escalation Rate (RCER)	n/a - instrumented during Foundation. Current pilot-era coordination relies on same-day access to central experts and is not a scalable baseline.	RCER <20%. Structural review if RCER >25% for two consecutive weeks or median routine-decision latency exceeds one business day.	Organizational Resilience Plan & Distributed Decision Charter

Governance Structure

Ownership is paired with authority. The person accountable for a target can stop expansion, freeze releases, enforce attribution, or pull a delegated decision back to the center when the trigger defined in the relevant operating model is breached.

Blueprint Dimension	Owned By	Review Cadence	Escalation if Targets Slip
Reliability (SLO and error budget)	AI Platform Engineering Lead	Weekly reliability review; monthly SLO close; monthly sampled false-consensus review against human ground truth.	At 50% error-budget use, reliability becomes first engineering priority; at 80%, non-critical releases freeze. Any P1/control-security failure freezes immediately. Any verified material false consensus is P2 even if caught downstream; any unauthorized action is P1. CISO/Risk co-approval is required to lift freezes involving their control domain.
Organizational resilience (knowledge, structure, feedback)	AI Transformation Lead	Weekly RCER/decision-latency review; monthly pod and deputy-coverage review	Stop further expansion when RCER >25% for two weeks, routine decision latency >1 business day, or required pod/deputy coverage is missing. Correct by delegation, policy codification, interface redesign, or capability transfer - not by adding another standing approver.
FinOps (visibility, unit economics, optimization)	Director, AI Platform & Enablement	Pod-level weekly cost view; monthly Technology/Finance/business FinOps review	>15% unfavorable 7-day CPW variance or >10% projected monthly overspend triggers same-week review. >25% CPW variance for two weeks pauses scale-out. Owner may enforce tagging, budgets/quotas, routing, non-production shutdown, and capacity controls.
Decision architecture (distributed effectiveness, regulatory compliance)	AI Transformation Lead	Monthly decision-rights and jurisdiction review with Governance Council	Any unauthorized autonomy, risk-tier, or sovereignty change stops expansion. Repeated reversible escalations are delegated or codified. Governance Council retains risk appetite, autonomy classes, material jurisdiction exceptions, and other high-stakes hard-to-reverse decisions.

Governance rule
An owner without authority cannot act; authority without accountability does not hold. Northbridge therefore gives each owner a defined stop-work or escalation right, while leadership receives the same metrics used by the operating teams.

Scaling Roadmap

A phase is complete only when its gate is answered yes. Dates guide sequencing; they do not override a failed gate. This is Bayesian thinking in operating form: we start with a prior that Proposal C can scale, run the 60-day Foundation test, and update the next capital decision from observed reliability, cost, and organizational evidence rather than treating the pilot as a formality.

Phase	What Happens	Gate: the question that must be answered YES
Foundation - operating proof	The Customer Service Workflow Pod runs the Reliability Charter, Organizational Resilience Plan, FinOps Operating Model, and Distributed Decision Charter as one operating system: 99.5% CWCR/error budget; named incident roles; workflow-pod ownership and deputies; mandatory cost attribution; delegated model/release/data decisions; tested open-model/frontier failover; 3CA and human fallback for designated high-risk paths. The AI Platform Engineering Lead publishes the measured CWCR and sampled false-consensus baseline by October 15, 2026.	Has the first workflow completed 60 days at ≥99.5% CWCR with no unresolved P1; published the operating and sampled false-consensus baseline; demonstrated rollback/failover and safe 3CA disagreement routing; attributed ≥95% of production AI spend; and executed delegated model, release, and permitted-data decisions without an unauthorized autonomy or jurisdiction exception?
Controlled Expansion - add two regulated domains	Add Underwriting Support and Compliance Monitoring pods; expand Customer Service; increase owned inference toward the benchmark-gated ~24-accelerator Full Scale planning envelope; preserve 20% platform exploration capacity; use full 3CA only on designated high-risk paths. Release up to $6M of the previously gated execution envelope for capacity, platform hardening, pod delivery, observability/evaluation, and bounded cloud/frontier use.	Do all three workflows have chartered pods and primary/deputy coverage; has RCER stayed <20% for four weeks with routine-decision latency <=1 business day; has >=95% spend been attributable for two months; is CPW <=$1.50 with 7-day variance inside 15%; are Lean outcomes improving; and have there been zero unauthorized autonomy, sovereignty, or risk-tier changes?
Full Scale - governed enterprise operation	Operate approximately 4.0M safely completed workflows/year across the three priority families and approved jurisdictions. Two open models run primarily on Northbridge-controlled datacenter infrastructure with separate failure domains; the third frontier slot remains swappable; workflow pods own outcomes and unit economics; 3CA remains a high-risk validation pattern, not authorization.	Has Northbridge sustained the Full Scale operating condition without weakening the decision boundary: two consecutive quarters with <50% error-budget consumption to activate the >=99.9% SLO, successful failover testing, CPW <=$1.50, stable distributed ownership, and no regression in human authority, jurisdiction controls, or reliability? Only then may further scope, autonomy, or reserve release be considered.

Investment boundary
Proposal C remains inside the previously approved $15M envelope: $3M foundation + up to $10M execution/acquisition + $2M gated reserve. With Foundation complete, we release up to $6M for Controlled Expansion. The remaining $4M of execution capacity and the $2M reserve stay withheld until the Full Scale gate is met. This is real-options thinking: withholding capital buys the right, not the obligation, to commit later when reliability, unit economics, and organizational evidence are stronger.

Acquisition Option: Governed, Not Planned

Technology acquisition has an unfavorable base rate: estimates commonly cited in the course place failure between 70 and 90 percent, with integration difficulty rising as the target grows. Acquisition therefore remains an option inside the execution envelope, not a dependency in the scaling plan. The value test is simple: standalone value + synergy - price. A good target still destroys value if we overpay.

- Diligence exists to reduce information asymmetry. Before any deal reaches Governance Council, we test talent-retention risk, data rights, licensing, model and training provenance, customer concentration, technical debt, and undisclosed liabilities.
- Anchoring is a named behavioral risk. We set a walkaway price from our own diligence before looking at comparable-deal premiums, so the last industry transaction does not become our reference point.
- We run three valuation approaches together: cost to build in-house, comparable market price, and income or savings generated. They should broadly converge; material divergence is itself a diligence signal. The walkaway price, all three valuations, and fit inside the same $10M execution envelope are prerequisites for bringing a transaction forward.

We are not approving twelve months of activity in advance. We are approving the next bounded operating phase and the authorities required to stop it when evidence says no.

NORTHBRIDGE FINANCIAL GROUP Northbridge Agentic AI Platform - Proposal C Scale-Up	BOARD COMMITMENT REQUEST Controlled Expansion
Prepared by: Luke Kilpatrick, AI Transformation Lead	August 10, 2026

Recommendation

Approve Controlled Expansion of Proposal C to move the Northbridge Agentic AI Platform from one controlled-production workflow to three enterprise workflow families and approximately 4.0M safely completed workflows per year within 12 months. Release up to $6M of the previously gated $10M execution envelope for this phase, while keeping the remaining $4M of execution capacity and $2M reserve behind the Full Scale gate. This converts Proposal C from a successful foundation build into an owned operating capability without committing us to scale before the evidence supports it.

Top Three Reasons

- Strategic necessity - Northbridge cannot scale regulated AI by extending siloed systems or increasing dependence on a single frontier vendor. Proposal C creates one owned control plane that preserves sovereignty, provider optionality, and economic leverage as customer service, underwriting, and compliance scale. Two controlled open-weight models, owned base-load infrastructure, a swappable frontier slot, and 3CA are implementation choices under that strategy.
- Operational readiness - reliability, pod ownership, decision rights, jurisdiction profiles, failover, auditability, and human escalation are explicit operating controls. Expansion is conditional on production evidence, not on a calendar or an adoption target.
- Financial discipline - the plan moves from a modeled $4.25 per completed workflow at 1.0M workflows/year to <=$1.50 at 4.0M/year (current model: $1.34), requires >=95% cost attribution, and uses a clear rule: own the base load, rent the peaks.

Conditions for Success

- Reliability: first workflow sustains >=99.5% CWCR for 60 days with no unresolved P1, tested failover, complete auditability, and safe 3CA disagreement routing; >=99.9% becomes the Full Scale SLO only after two stable quarters.
- Organization and authority: all three workflows have chartered pods and primary/deputy coverage; RCER remains <20% for four weeks; routine decision latency is <=1 business day; pods cannot waive autonomy or jurisdiction controls.
- Economics and Lean outcomes: >=95% of production AI spend is attributable for two months; CPW is <=$1.50 with 7-day variance inside 15%; customer-service, underwriting, and compliance process metrics improve without weakening controls.
- Funding discipline: the remaining $4M execution capacity and $2M reserve are not released until the Full Scale gate is met; any acquisition must fit inside the existing execution envelope and clear separate diligence.

Decision Requested
Approve Controlled Expansion, release up to $6M of the existing execution envelope, endorse the stop-work and delegated-decision authorities in this Blueprint, and keep the remaining $6M ($4M execution + $2M reserve) withheld until the Full Scale gate is met.

AI Use Reflection

I used a ChatGPT project with the worksheet, course transcripts, and prior assignments loaded, then used Claude and Gemini as independent reviewers. The most important work was deciding where not to accept their defaults. AI first built the FinOps cost stack around conventional cloud consumption pricing, platform uptime, and cost per API call. I replaced that with a three-year depreciated datacenter hardware model plus cloud only for burst, DR, and experimentation; added operational toil because incident response, evaluation, and 3CA disagreement review are real costs; and moved the headline measures to controlled workflow completion and fully loaded cost per completed workflow.

The bigger override was 3CA. Every model I prompted initially treated unanimity as confirmation, and several proposed majority voting if one leg was unavailable. I rejected both because three models can share training data, architectural priors, and retrieved context, so their errors can be correlated. That is why disagreement goes to an authorized human, why the models must come from distinct families and provider lineages, and why unanimous high-risk outputs are sampled monthly against ground truth. AI review still improved the submission: Claude and Gemini independently caught that my escalation design named severities but not who could lift a freeze, and that the FinOps Operating Model had a leading indicator while the Organizational Resilience Plan did not. I fixed both, then made the final calls on metrics, authority, capital gates, and risk boundaries.