

Introducing Continuous Cost Governance

Extending FinOps from Cost Optimization to Financial Accountability

Version: 1.5.6

Status: Draft for Industry Review

Date: September 2026

DOI: 10.5281/zenodo.23555398

Executive Summary

Between 2019 and 2026, cloud financial management matured from an improvised practice into a recognized discipline. The FinOps Foundation's framework gave the field a shared vocabulary, a persona model, and a capability map; the FOCUS specification began to normalize billing data across providers; and the major cloud vendors embedded cost optimization into their reference architectures. The discipline has succeeded on its own terms.

It has also moved. The *State of FinOps 2026* survey of 1,192 practitioners managing more than \$83 billion in annual cloud spend reports that 78% of FinOps practices now sit within the CTO or CIO organization, while the share reporting to the CFO has fallen to 8% (FinOps Foundation, 2026a). Over the same period, the Foundation revised its mission from advancing the people who manage the value of *cloud* to the value of *technology*, and practitioners reported that scope expansion, governance, organizational alignment and forecasting now collectively outweigh optimization as a priority. FinOps has shifted **up** into technology strategy and **left** into engineering workflow. It has not, on the evidence, shifted **across** into the finance control environment. (Reporting line is an imperfect proxy for control maturity, as a CTO-reporting practice can maintain a strong finance liaison, but the direction of travel is unambiguous.)

That is the seam this paper examines. A technical recommendation becomes a financial result only when someone is accountable for it, someone authorizes it, someone implements it, and someone independently measures what actually changed. Those steps are the ordinary substance of internal control in every other category of corporate expenditure. In consumption-based technology spend they are, in most organizations, not systematically performed, and the industry's headline metric, "potential savings," measures the absence of exactly that discipline.

This paper proposes **Continuous Cost Governance (CCG)**: a control specification that treats each optimization decision as a governed transaction with an owner, an evidentiary record, an authorization, and an independently verified outcome. It is offered as an extension to FinOps, not a replacement for it, and not as an industry standard.

The paper also states plainly what CCG is not: it is not novel. Nearly every component described here already exists somewhere: in the FinOps Framework's Policy & Governance capability, in

Technology Business Management, in COSO's control activities, in ITIL change enablement, in continuous controls monitoring, and in the thirty-year-old measurement and verification protocols of the energy efficiency sector. Section 3.3 catalogues those overlaps honestly. What appears genuinely additive is narrower than a new discipline: a specification for the object being governed, an evidentiary standard for what may be reported as a financial result, and a name for an organizational gap that recent survey data now makes visible.

Continuous Cost Governance in one sentence

FinOps determines what should be done with technology spend. Continuous Cost Governance establishes what was actually done, on whose authority, and what it verifiably delivered.

Key Contributions

This paper contributes three things, and claims no more:

1. **A decision-centric governance model.** It proposes the individual cost decision (not the period, the resource, or the service) as the unit of governance, with a defined lifecycle and a retained evidentiary record.
2. **A reportability standard for realized savings.** It proposes a rule separating what may be tracked operationally from what may be reported as a financial result, grounded in established measurement and verification practice.
3. **A control specification that extends rather than competes with FinOps.** It locates that specification inside the FinOps Framework's Policy & Governance capability, and names the organizational seam between engineering-owned telemetry and finance-owned control that current practice leaves unattended.

Continuous Cost Governance is abbreviated **CCG** throughout the remainder of this paper.

Who This Paper Is For

This paper is written for those with fiduciary or operational responsibility for technology expenditure:

- **CFOs, controllers and finance business partners** evaluating whether internal control over variable, consumption-based spending is proportionate to its materiality.
- **FinOps practitioners and cloud architects** whose recommendations are not converting into implemented change at the rate their analysis warrants.
- **CIOs, CTOs and infrastructure executives** balancing delivery velocity against budget predictability.
- **Audit committees, internal audit functions and boards** assessing the maturity of governance over cloud, SaaS and AI expenditure now that it rivals labor as a principal cost line in many organizations.

What This Paper Is Not

- **Not an accounting or legal standard.** CCG does not interpret GAAP or IFRS, does not prescribe capitalization treatment for cloud refactoring labor, and is not legal advice. It may generate evidence useful to compliance activity; it does not constitute compliance.
- **Not a replacement for FinOps.** The FinOps Framework describes how to build a practice that generates and prioritizes optimization. CCG addresses what happens to an individual optimization decision after it is generated.
- **Not a claim of novelty.** See Section 3.3.
- **Not a product specification.** CCG originated in observations made during the development of a commercial cost governance tool, which now operates an early, partial implementation of the loop's assignment and verification stages. That origin is disclosed here for transparency; the framework is described in vendor-neutral terms and should be evaluated independently of any implementation, and no results from that tool are cited as evidence in this paper.

1. The Evolution of Cloud Financial Management

The economic shift underlying this discipline is well documented. Capital expenditure on depreciating assets, governed through procurement cycles and approved in advance, was displaced by operating expenditure incurred continuously by engineers through API calls. The buyer moved from finance to engineering. The commitment moved from a signed purchase order to a running instance. Gartner forecast worldwide end-user spending on public cloud services at \$723.4 billion for 2025, growing 21.5% year over year, with cloud infrastructure and platform services alone reaching approximately \$301 billion (Gartner, 2024).

1.1 What FinOps Solved

FinOps emerged as a practitioner response to that shift. Its contribution is substantial and should not be understated:

- **A shared language.** The FinOps Framework's principles, personas, domains and capabilities gave finance and engineering a common vocabulary where none existed (FinOps Foundation, 2025a).
- **Data normalization.** The FinOps Open Cost and Usage Specification (FOCUS) established a provider-neutral schema for billing data, addressing a problem that had previously consumed disproportionate practitioner effort (FinOps Foundation, 2026b).
- **Decentralized accountability.** The principle that "everyone takes ownership for their technology usage" moved cost consciousness into engineering teams rather than concentrating it in a central cost-control function.
- **Rate optimization at scale.** Commitment-based discount instruments (reserved instances, savings plans, and committed use discounts) became a managed portfolio rather than an ad hoc purchase.

- **Vendor reinforcement.** The AWS Well-Architected Framework elevated "Implement Cloud Financial Management" to a design principle of its Cost Optimization pillar (Amazon Web Services, 2024); Microsoft and Google published comparable guidance.

1.2 Where the Discipline Has Moved Since 2025

Two developments are material to the argument of this paper.

First, **scope expanded.** Framework 2025 introduced Scopes as a core element, formalizing the "Cloud+" reality that practitioners were already managing SaaS, licensing, private cloud and data center spend alongside public cloud (FinOps Foundation, 2025a). By 2026, 98% of respondents reported managing AI spend (up from 63% in 2025 and 31% in 2024), alongside 90% for SaaS, 64% for licensing, 57% for private cloud and 48% for data center (FinOps Foundation, 2026a).

Second, **the practice moved up the technology organization rather than into finance.** The 78%/8% reporting split noted above is the clearest available signal. The Foundation's own analysis of Fortune 100 practices describes an emerging "FinOps Enabled Executive": a VP- to EVP-level technology leader, typically with 20–25 years of engineering background, positioned as a "COO for the CTO" (Storment, 2025). This is a credible and probably correct trajectory for the discipline. It is also, unambiguously, a trajectory *within the technology function*.

Meanwhile practitioners report diminishing returns from the optimization work that built the discipline. One characterized the position directly: "We have hit the 'big rocks' of waste and now face a high volume of smaller opportunities that require more effort to capture" (FinOps Foundation, 2026a). Another organization reported reaching 97% of identified optimization in its cost optimization hub, with the remaining 3% deliberately not actioned for business reasons (FinOps Foundation, 2026a), which is a mature outcome that only means something because the organization could distinguish *actioned* from *identified*.

No published dataset establishes how many organizations can make that distinction reliably. Practitioner testimony, including the example above, suggests many cannot; this paper treats that as a plausible and widely reported condition rather than a measured one, and Section 7 identifies closing that gap as the framework's most pressing empirical need. That gap, the unmeasured space between what tooling identifies and what an organization can show it actually did, is the problem this paper addresses.

2. The Remaining Governance Challenge

2.1 Potential Savings Are Not Financial Outcomes

Cloud cost tooling is optimized to produce recommendations. A rightsizing engine surfaces an oversized instance and attaches a dollar figure. That figure is routinely aggregated across an estate, presented as "potential savings" or "cost avoidance," and reported to leadership as an achievement.

It is not an achievement. It is a hypothesis about a counterfactual. A financial outcome exists when a change is implemented, when the reduction survives comparison against a baseline adjusted for business volume and rate changes, and when the resulting figure is reflected in a budget, a forecast, or the ledger. Every step between the hypothesis and the outcome is a step where the value can be, and frequently is, lost.

The FinOps community is candid about this. The measurement problem was raised repeatedly in the 2026 survey: "Once you fix it, it's gone. How do we give developers credit for shift-left activities?" (FinOps Foundation, 2026a). The Foundation's own guidance on usage optimization notes that estimating savings requires accounting for the full cost of change (labor hours, service disruption, and transformation complexity), not merely the delta in the bill. Procurement disciplines have long distinguished realized savings from cost avoidance and tracked them separately. Cloud cost management, for the most part, has not adopted that distinction.

2.2 Why Recommendations Fail to Become Outcomes

Four mechanisms are observable, and none of them is a tooling failure.

Evidence asymmetry. A recommendation asserts a saving but rarely proves safety. The engineer receiving it bears the entire downside of a production incident and captures none of the upside of a reduced bill. Absent evidence (sustained utilization percentiles, headroom against peak, and dependency and blast-radius analysis), declining the recommendation is the individually rational choice. Recommendation volume without evidence produces fatigue, and fatigue produces silent attrition.

Ownership ambiguity. A resource tag identifies a cost center. It does not identify a person with the authority and the standing to approve a change to a production system. Where the routing from finding to decision-maker is unreliable, findings accumulate in a queue that nobody owns.

Absent financial context. Engineers cannot reasonably assess whether a rightsizing action will drop utilization below a committed-use threshold, alter the allocation of shared costs, or affect a unit-economic metric reported externally. That context sits in finance. In most organizations it is not consulted before the change, only after the variance.

No verification loop. Where implemented savings are never independently measured, there is no feedback signal distinguishing accurate recommendations from inaccurate ones, and no basis on which finance can rely on the number. The estimate is never falsified, so it is never improved.

Supporting evidence is directionally consistent but methodologically imperfect, and should be read with care. Flexera's *State of the Cloud* series reports self-estimated wasted cloud spend of 27% in 2025, rising to 29% in 2026, marking the first increase in five years and attributed to AI and new service complexity (Flexera, 2026). These are respondent estimates, not audited measurements. McKinsey research based on a survey of more than 800 executives at manufacturing firms in Europe and North America found that 74% of cloud transformations failed to deliver expected savings or business value, with 40% of respondents exceeding cloud

budgets (McKinsey & Company, 2021; CFO Dive, 2021). The sample limits generalizability. What these sources establish is that a gap is widely perceived; they do not establish its precise magnitude, and this paper does not assert one.

2.3 The CFO Perspective

For finance leadership the issue is not the absolute quantum of cloud waste. It is that a large, volatile, and rapidly growing expense category is being managed through a control environment materially weaker than the one applied to categories of comparable size.

Three consequences follow. **Forecast reliability degrades** when the inputs to the forecast include unverified savings assumptions. **Internal control coverage becomes uneven**: technology spend increasingly rivals labor as a principal cost line in large enterprises (Storment, 2025), yet the delegation of authority, segregation of duties and evidence retention that finance takes for granted in procurement are inconsistently applied to consumption decisions made by engineers. And **fiduciary reporting weakens**: a board asking whether a stated efficiency program delivered has, in many organizations, no answer that survives scrutiny, only a dashboard of opportunities.

CFO reporting relationships in this domain declined to 8% in 2026 (FinOps Foundation, 2026a). Whatever else that number means, it does not describe a discipline moving toward the finance control environment.

2.4 Why AI Changes the Calculus

The argument above would be a modest one if consumption-based spend were stable. It is not, and the direction of change is the strongest practical case for the framework this paper proposes.

Artificial intelligence decomposes technology expenditure into a far larger population of far smaller decisions. Tokens, inference calls, embedding generation, vector database operations, GPU reservations, retrieval pipelines and increasingly autonomous agentic workloads each carry cost, and each is initiated at a level of the organization well below where budgets are set. Where a traditional infrastructure commitment was a discrete, annual, negotiated event with a purchase order attached, an AI workload generates a continuous stream of consumption decisions with no procurement moment at all. The 2026 survey data captures how fast this arrived: 98% of practices now manage AI spend, against 63% a year earlier and 31% the year before that, and practitioners rank visibility into AI cost, allocation of that cost to business units, and determination of AI value as their three leading difficulties (FinOps Foundation, 2026a).

Three properties compound the governance problem. Pricing models are unfamiliar and vary sharply between providers, so cost cannot be inferred from architecture as reliably as it once could. Attribution is harder, because a single model endpoint may serve many business functions. And the value denominator is contested; as one practitioner put it, "Is your AI providing value? No one can answer that question yet" (FinOps Foundation, 2026a).

The consequence is directional rather than dramatic. As AI moves from experiment to material operating expense, the center of gravity in technology financial management shifts from *annual budgeting*, which governs a small number of large commitments, toward *continuous financial control*, which must govern a large number of small ones. Periodic review scales poorly against that shape of spend. Per-decision control, applied proportionately, scales better. This paper does not claim that CCG is the only response to that shift, but the shift is what makes the response worth specifying now rather than later.

What governance looks like for AI workloads. A common objection is that per-decision governance cannot function when the "decision" is an autonomous agent making thousands of inference calls per hour. The objection rests on a category error. CCG does not propose governing individual tokens or API calls, any more than it proposes governing individual packets on a network. It proposes governing the *configuration decisions* that determine how tokens are consumed, and those decisions are human-scale, infrequent, and consequential.

Consider a concrete example. An ML platform team operates a retrieval-augmented generation (RAG) pipeline serving an internal knowledge assistant. The pipeline's model-routing policy directs all queries to a frontier model at \$15 per million input tokens. A FinOps analysis observes that 62% of queries are simple classification or lookup tasks that a smaller model handles at \$0.60 per million tokens with no measurable accuracy degradation on the organization's evaluation suite. The *finding* is a routing-policy change. The *owner* is the ML platform lead. The *evidence* includes the query classification analysis, the evaluation-suite results, and the projected annualized saving. The *authorization* is the platform lead's approval of the policy change, recorded with the evaluation data that supports it. The *verification*, ninety days later, is a comparison of cost-per-query against an adjusted baseline that accounts for changes in query volume and any model price revisions in the period.

This is not a novel control challenge. It is the same loop applied to a different resource type, and it is the level at which AI cost decisions are actually made. The individual inference call is consumption; the routing policy, the caching strategy, the model selection, the budget cap, and the retrieval configuration are decisions. CCG governs the latter.

3. Introducing Continuous Cost Governance

3.1 Definition

Continuous Cost Governance (CCG) is a proposed control specification for consumption-based technology expenditure, in which each material cost decision is treated as a governed transaction (carrying evidence sufficient to support a decision, a named accountable owner, an authorization proportionate to its risk and materiality, and an independently verified outcome), and in which only verified outcomes are reported as financial results.

The emphasis is deliberate. CCG does not propose new ways to find savings; that problem is well served. It proposes a specification for what happens to a finding once found, and a standard for what may be reported once acted upon.

The central term carries the weight of the framework and is defined precisely:

*A **governed transaction** is a single cost decision that carries, as a retained record: the evidence on which it was proposed, the named party accountable for it, the financial review it received, the authority under which it was approved, declined or deferred, and the independently verified outcome it produced.*

A note on scope. This paper uses public cloud examples because that is where the practice, the data and the tooling are most mature. The control specification is not cloud-specific. It applies to any expenditure that is consumption-based, initiated below the level at which budgets are set, and incurred continuously rather than at a procurement moment, which includes SaaS subscriptions, AI and inference services, data cloud platforms, and consumption-metered licensing. It is deliberately **not** extended in this paper to purchase-order-controlled procurement, accounts payable or ERP-mediated spend. Those categories already carry mature control environments (three-way match, purchase-order approval, and established delegation of authority), and the gap this paper identifies does not exist there. Claiming them would broaden the framework's apparent reach while weakening its actual argument.

3.2 The Unit of Governance

The framework's central proposition is a change in the object being governed.

Existing disciplines govern different units. Budget governance governs the *period*. TBM governs the *service or application*. Cloud policy governance governs the *resource* or the *configuration state*. FinOps allocation governs the *cost*. CCG proposes governing the **decision**: a discrete, dated, evidenced, owned, authorized and verified event with a lifecycle and a retained record.

This is a modest-sounding shift with practical consequences. It makes optimization auditable at the level at which decisions are actually made, it creates a denominator for measuring conversion (findings raised versus findings resolved), and it produces an evidentiary trail in the form finance and audit functions already understand.

3.3 An Honest Assessment of Prior Art

This paper was written under an explicit instruction not to manufacture novelty. Applying that test rigorously, most of CCG is not new.

CCG element	Existing discipline that already covers it	Assessment
Continuous visibility	FinOps <i>Understand Usage and Cost</i> domain; FOCUS; native provider cost tooling	Not novel. Fully addressed by existing practice.
Evidence-based findings	FinOps <i>Workload and Usage</i>	Largely existing. The

	<i>Optimization</i> capabilities; provider recommenders; SRE capacity management	extension is the evidentiary <i>standard</i> (evidence sufficient to bear an authorization decision), not the analysis itself.
Clear ownership	FinOps principle "everyone takes ownership"; <i>Allocation</i> capability; Weill & Ross decision rights (2004)	Not novel. Decision rights and accountability frameworks are foundational IT governance theory.
Finance review	IT Financial Management; TBM taxonomy (TBM Council, 2023); COSO control activities (COSO, 2013)	Not novel as a function. Modestly novel in its <i>insertion point</i> : per-decision and pre-implementation rather than periodic and post-variance.
Human approval	COSO delegation of authority and segregation of duties; ITIL change enablement; SOX IT general controls	Not novel, and actively contested. See Section 6.1.
Verified outcomes	IPMVP measurement and verification (EVO; development begun 1994, first published 1996); procurement realized-versus-avoided savings; FinOps Effective Savings Rate	Method not novel. Its transposition to cloud is the least-solved and most additive element of the framework.
Executive accountability	TBM; FinOps <i>Quantify Business Value</i> domain; the FinOps Enabled Executive archetype	Not novel. The discipline is already moving in this direction.
The composite control loop	Continuous controls monitoring: "technologies to reduce business losses through continuous monitoring and reducing the cost of audits through continuous auditing of the controls in financial and other transactional applications" (Gartner, n.d.); continuous auditing (CICA/AICPA, 1999; Vasarhelyi et al., 2004)	Structurally analogous. CCG is, in substance, continuous controls monitoring applied to consumption-based technology spend.

Two further points of honesty are required.

The terminology is not new. "Cost governance" and "cloud cost governance" are already in general commercial use, defined by multiple vendors and analysts as continuous, policy-driven oversight of technology spend. CCG's contribution, if it has one, is not the phrase. It is the specification behind it.

The academic ground is well trodden. Continuous auditing has a formal definition dating to 1999 and a substantial peer-reviewed literature, principally associated with Vasarhelyi and colleagues at Rutgers. Anyone presenting a "continuous" governance loop to an audit committee should expect that committee to recognize the pattern immediately.

What appears genuinely additive is therefore narrow, and the paper claims no more than this:

1. **The decision as the governed object**, with a defined lifecycle and retained evidentiary record, which is a construct not specified in the FinOps Framework, in TBM, or in cloud policy governance.
2. **A reportability standard** that excludes estimated, potential and avoided savings from financial reporting while permitting them in operational tracking. The FinOps community identifies this measurement problem as unsolved; CCG proposes a resolution rule rather than a metric.
3. **The naming of an organizational seam** that 2026 survey data makes visible: continuous controls monitoring lives in finance and GRC over ERP transactions; FinOps lives in engineering over telemetry; and the space between them (consumption decisions with financial consequence, made by engineers, outside the finance control environment) has no established owner.

CCG is a recombination. Recombinations can be useful. They should not be marketed as inventions.

3.4 The Five Principles of CCG

Version 0.2 of this framework proposed seven principles. On review, those seven conflated two different kinds of statement: durable normative commitments and sequential process steps. The FinOps Framework separates Principles from Phases from Capabilities for precisely this reason, as do COSO and COBIT. Version 1.0 therefore separates them, and adds one principle that was missing.

Principle 1: Cost decisions are governed as transactions, not as reports.

The unit of governance is the individual decision, not the monthly bill or the quarterly review. A finding without a lifecycle is a suggestion.

Principle 2: A finding must carry evidence sufficient to bear the decision it requests.

The evidentiary burden sits with the party proposing the change, not the party absorbing its risk. Sufficiency scales with consequence: a sustained-utilization percentile analysis with headroom

against observed peak may suffice for a stateless workload; a tier-one transactional system warrants more.

Principle 3: Every finding has a named accountable owner with authority to act.

Ownership is a person or a defined role, derived where possible from identity systems rather than inferred from resource metadata alone. A cost center is not an owner.

Principle 4: Control is proportionate to materiality and risk.

Governance applied uniformly to all decisions is governance that will be circumvented. Most findings should flow through lightweight, automatable paths; formal review and separated authorization should be reserved for decisions that are financially material, operationally risky, or both. This is not merely a concession to practicality: DORA identifies "treating all changes equally" as a distinct pitfall of change approval processes, on the grounds that undifferentiated review prevents reviewers from devoting attention to the changes that actually warrant it (DORA, n.d.). This principle is the one most likely to determine whether CCG is adopted or resented.

Principle 5: Only verified outcomes are reported as financial results.

Estimated, potential and avoided savings may be tracked, and are useful operationally. They are not reported to finance, to executives, or to the board as achieved results. Verification means comparison against a baseline adjusted for business volume, rate changes and seasonality.

Crosswalk from Version 0.2. No prior content is discarded. *Continuous Visibility* becomes a precondition of the loop rather than a principle, because it is a data capability and is already fully specified by FinOps and FOCUS. *Evidence-Based Findings* becomes Principle 2. *Clear Ownership* becomes Principle 3. *Finance Review* and *Human Approval* become stages of the loop, governed normatively by Principles 1 and 4; "Human Approval" is renamed **Authorized Change**, because the salient property is authority, not the presence of a human. *Verified Outcomes* becomes Principle 5. *Executive Accountability* becomes the reporting and reallocation stage of the loop, governed by Principle 5.

3.5 The Continuous Governance Loop

The loop retains the seven-stage structure of Version 0.2, now expressed as process rather than principle, and with the exception paths made explicit, because a control with no documented rejection path is not an auditable control.

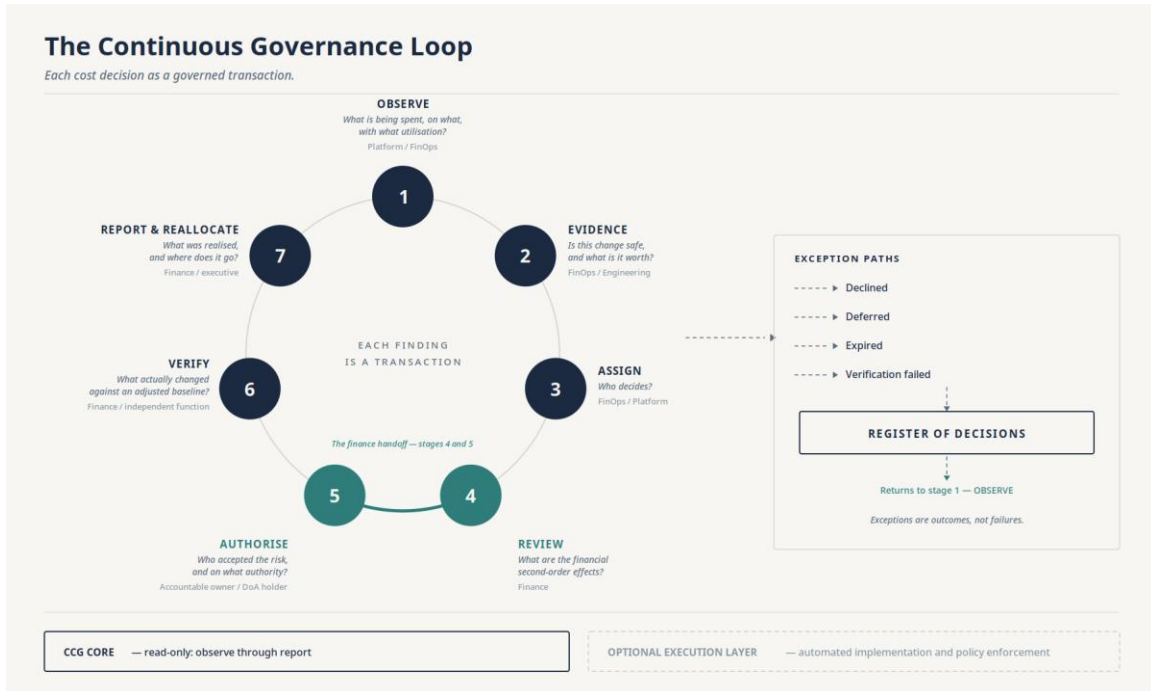


Figure 1: The Continuous Governance Loop

Stage	Question answered	Primary owner
1. Observe	What is being spent, on what, with what utilization?	Platform / FinOps
2. Evidence	Is this change safe, and what is it worth?	FinOps / Engineering
3. Assign	Who decides?	FinOps / Platform
4. Review	What are the financial second-order effects?	Finance
5. Authorize	Who accepted the risk, and on what authority?	Accountable owner / DoA holder
6. Verify	What actually changed against an adjusted baseline?	Finance / independent function
7. Report & Reallocate	What was realized, and where does it go?	Finance / executive

Four exception paths are first-class outcomes, not failures:

- **Declined:** the owner or finance rejects the change, with a recorded reason. A well-run estate will decline a meaningful share of findings; the 97%/3% example cited earlier is a mature outcome precisely because the 3% was a decision rather than a gap.
- **Deferred:** the change is valid but sequenced behind other work, with a review date.
- **Expired:** the finding received no disposition within the organization's defined window. A deferred finding past its review date without disposition does not wait indefinitely; it is

recorded as expired, with its last known state attached, and it counts against conversion. Without this rule the denominator is whatever survived attention, and the conversion metric can be improved by neglect.

- **Verification failed:** the change was implemented but the expected saving did not materialize. This is the most informative outcome in the loop and the one most systems discard. It is the feedback signal that improves estimate quality.

What "independent" means at stage 6. Independence is organizational separation from the decision being verified: the verifying party did not authorize the change, did not implement it, and does not own the budget it affects. In most organizations this will be a finance function or FP&A analyst distinct from the FinOps practice; in regulated environments, internal audit. Where verification is automated (a measurement pipeline comparing metered spend against an adjusted baseline), independence is satisfied by ownership of that pipeline outside the implementing team, together with an immutable record of the baseline and its adjustments. A team verifying its own savings is producing a report, not assurance, and Principle 5 does not accept it as the latter.

The loop is designed to function without infrastructure write access. **CCG Core** (observe, evidence, assign, review, authorize as a recorded decision, verify, and report) requires only read access to cost and telemetry data plus a system of record for decisions. An **optional execution layer** may automate implementation of low-risk actions or enforce policy in CI/CD, but the governance framework is complete without it. This distinction matters to conservative audit and infrastructure functions and was a deliberate design constraint.

3.5.1 Routing and Automation at Scale

CCG does not require human review of every finding, and a reading that assumes it does will conclude, correctly, that the framework cannot scale. Principle 4 governs the loop's throughput: findings should be routed automatically according to materiality, risk classification and policy, with the substantial majority resolving through automated or lightweight paths and never reaching a review queue.

The following table illustrates a representative routing model. The thresholds are organizational choices, not framework prescriptions; the structure is the point.

Tier	Criteria (illustrative)	Stages 3–5 treatment	GTR record weight	Expected share of findings
Auto	Annualized value < \$1,000; non-production; no commitment interaction	Stages 3–5 collapse into a single automated disposition. Policy is cited as the authorizing authority. No	Lightweight: 5–8 auto-populated fields (finding ID, resource, action, policy reference, timestamp).	~70–85%

		human touch.		
Standard	\$1,000–\$25,000 annualized; production but not tier-one; no commitment interaction	Owner notified and assigned automatically. Authorization via recorded peer review or team-lead approval in the existing development platform. No finance review required.	Moderate: adds owner, approval record, evidence summary.	~10–20%
Elevated	>\$25,000 annualized, or tier-one service, or commitment interaction, or cross-team dependency	Full loop: finance review (stage 4), named authorization with recorded delegation of authority (stage 5).	Full: complete evidentiary record as described in §3.2 and Figure 5.	~3–8%
Executive	>\$250,000 annualized, or affects externally reported metric, or requires capital treatment	Full loop with executive-level authorization and, in regulated environments, audit committee visibility.	Full, with executive sign-off and audit trail.	<1%

What the framework requires is not that a human touches every decision, but that every decision carries a record of how it was disposed of and under what authority, including when that authority was a policy rather than a person. Automation is therefore a precondition of CCG at scale, not a departure from it.

Peer review captured in the development platform satisfies stage 5 only where the review record references the governed decision: the evidence, the materiality assessment, and the delegation of authority under which the change is approved. A code review that attests to technical correctness but records nothing of financial consequence is change control, not authorization; the framework requires the latter to be visible in the record.

3.6 The Governance Pyramid

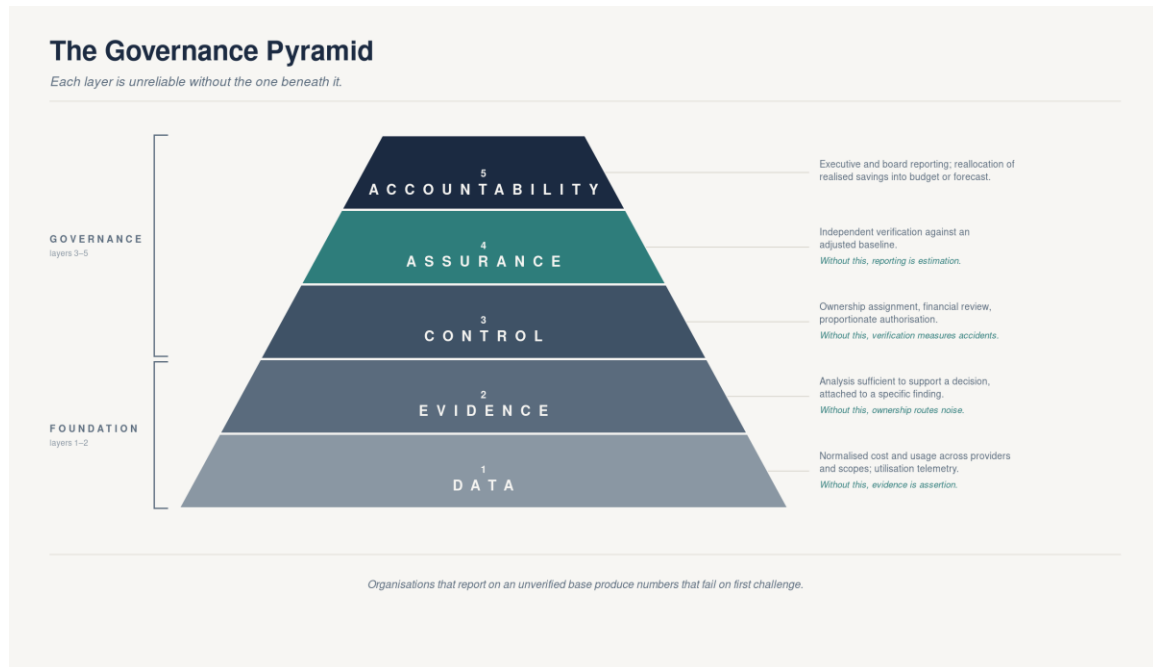


Figure 2: The Governance Pyramid

CCG can be represented as five dependent layers, in which each layer is unreliable without the one beneath it:

1. **Data:** normalized cost and usage across providers and scopes; utilization telemetry. *Without this, evidence is assertion.*
2. **Evidence:** analysis sufficient to support a decision, attached to a specific finding. *Without this, ownership routes noise.*
3. **Control:** ownership assignment, financial review, proportionate authorization. *Without this, verification measures accidents.*
4. **Assurance:** independent verification against an adjusted baseline. *Without this, reporting is estimation.*
5. **Accountability:** executive and board reporting; reallocation of realized savings into budget or forecast.

The layered dependency is the argument for sequencing: organizations that attempt executive reporting on an unverified base produce numbers that fail on first challenge, which is the most common way credibility is lost.

3.7 The CCG Maturity Model

The FinOps Foundation's Crawl/Walk/Run model already assesses maturity across its capability set, and its caution applies here too: the objective is the appropriate level of maturity per capability, not uniform advancement (FinOps Foundation, 2025b). The model below is offered

not as a replacement but as a finer-grained view of two axes, control and verification, that Crawl/Walk/Run treats only lightly.

Level	Visibility	Ownership	Control	Verification	Reporting
1. Reactive	Fragmented billing data	Unassigned	None	None	Ad hoc
2. Visible	Centralized billing	Tag/metadata based	Manual alerts	Estimated avoidance	Periodic dashboards
3. Optimizing	Cost plus utilization	Team-level mapping	Advisory workflow	Simple before/after	Variance narrative
4. Governed	Normalized multi-scope	Identity-derived, named	Risk-proportionate authorization with recorded decisions	Baseline-adjusted M&V	Reviewable evidence
5. Continuous	Real-time unit economics	Federated with automated routing	Integrated with GRC and change systems	Continuous M&V with estimate calibration	Reconciled to budget and forecast

The transition from Level 3 to Level 4 is where most of the difficulty lies, and it is organizational rather than technical: it requires finance to accept a role in a workflow it does not currently own, and engineering to accept a recorded decision point it does not currently have.

3.8 A Worked Example

Illustrative only. The figures below are synthetic. They are included to demonstrate the method and the arithmetic of verification, not to report a result. No claim is made that these values are typical.

Consider a single finding as it passes through the loop.

1. Observe. Continuous visibility flags a production application tier: fourteen general-purpose instances, thirty-day mean CPU utilization 11%, mean memory utilization 24%. The cost tool attaches a naive rightsizing estimate of **\$24,100 annualized**.

2. Evidence. Analysis attaches to the finding a ninety-day utilization distribution (P95 CPU 34%, P99 61%), a dependency map showing two downstream services, and the observation that the P99 peak occurs within a four-hour window on the last business day of each month, correlated with financial close. The proposed target instance size provides roughly twice the headroom against observed P99 for eleven of the fourteen instances. Three instances carry the month-end path and do not clear the threshold.

*Evidence has already changed the proposal: eleven instances, not fourteen. The three excluded are recorded as **Declined**, with the reason attached, rather than disappearing.* Revised estimate: **\$18,900**.

3. Assign. Resource metadata resolves to a cost center, which is not a person. An identity lookup resolves the owning team and its engineering manager as the accountable owner. The finding is routed to a named individual with authority to act.

4. Review. Finance identifies a second-order effect invisible to the engineering analysis: the account is substantially covered by a three-year compute commitment, and withdrawing on-demand consumption from this instance family would push utilization of that commitment below its break-even point, transferring cost to an unused reservation rather than removing it from the P&L. Finance does not decline the change. It directs that the change be either sequenced after the commitment expires, or offset by migrating a currently uncommitted workload into the freed capacity. The team selects the offset.

This is the stage most commonly absent in practice, and in this example it is the stage that prevents a reported saving from being an accounting illusion.

5. Authorize. The change exceeds the organization's automatic-routing threshold on two counts (annualized value and tier-one service classification), so it does not resolve through the automated path. The accountable owner authorizes it. The record captures who authorized, when, on what evidence, and what risk was accepted.

6. Verify. Ninety days after implementation, the outcome is measured against a baseline adjusted for two exogenous changes in the period: a 6% decrease in transaction volume, and a provider list-price revision.

	Annualized
Naive estimate at finding	\$24,100
Estimate after evidence (11 of 14 instances)	\$18,900
Gross observed reduction	\$19,400
Less: routine baseline adjustment (demand and list rate)	(\$2,300)
Less: commitment reallocation effect	(\$1,150)
Verified realized saving	\$15,950

7. Report & Reallocate. **\$15,950** is reported to finance and carried into the next forecast cycle. **\$24,100 is not reported at any point.** The gap between the two figures, roughly a third of the original estimate, is not error or waste; it is the difference between a hypothesis and a measured outcome, and under Principle 5 only the latter is a financial result.

What the example is intended to show. Four of the seven stages changed the answer. Evidence removed three instances from scope. Finance review prevented a change that would have transferred cost rather than eliminated it. Adjustments removed \$3,450 of apparent saving attributable to factors other than the change itself: \$2,300 of routine baseline adjustment (demand normalization and the list-rate reduction) and the \$1,150 non-routine commitment reallocation. And the decision now exists as a record that an auditor, a controller or a successor

can reconstruct. The framework's value is not that it found the saving (the cost tool did that in stage 1), but that it established which part of the saving was real.

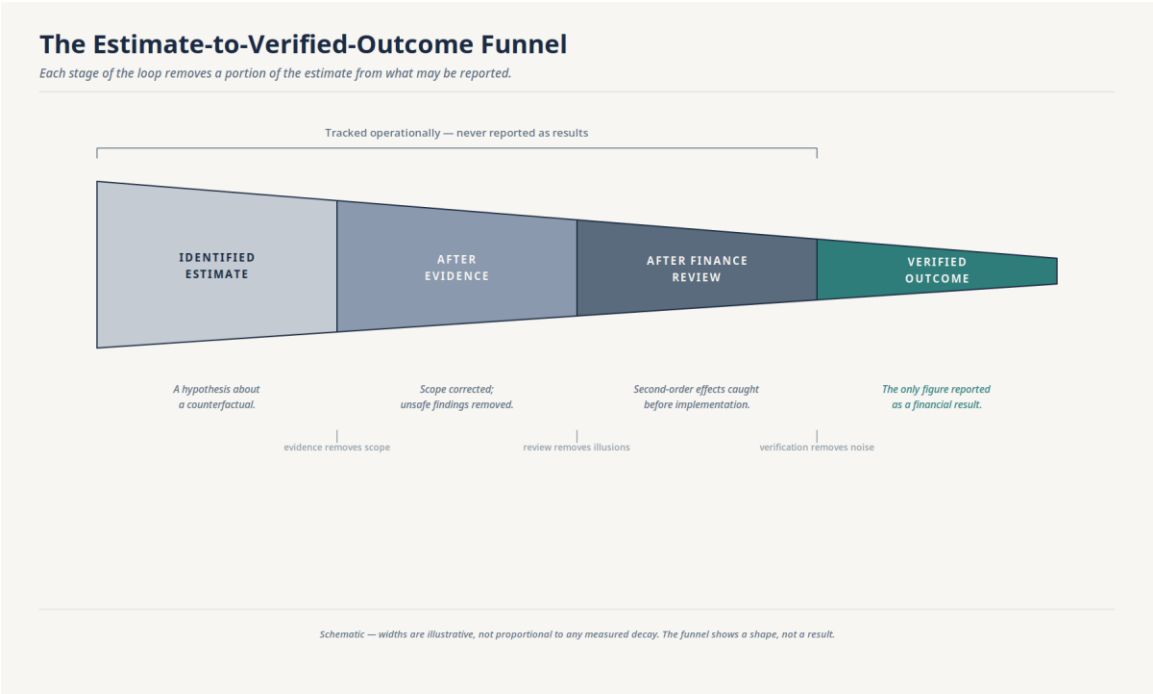


Figure 3: The estimate-to-verified-outcome funnel (schematic; not to scale)

The example above is one instance of a general pattern, shown schematically in Figure 3: each stage of the loop removes a portion of the original estimate from what may be reported, until only the verified outcome remains. The funnel contains no figures of its own (it is a shape, not a result), but it is the visual form of Principle 5: everything to the left of verification is hypothesis, and hypotheses are not financial results.

3.9 The Loop at Enterprise Scale

Illustrative only. The distribution below is synthetic. It is included to demonstrate that the routing model in §3.5.1 produces a manageable workload at enterprise scale, not to assert typical values.

Section 3.8 showed a single finding passing through the loop. The more common objection is not about one finding but about thousands: can the loop function when an enterprise estate generates a large population of recommendations simultaneously?

Consider an organization with \$48 million in annual multi-cloud spend, approximately 38,000 billable resources, and a FinOps practice that surfaces 4,200 active findings per quarter. Applying the routing model from §3.5.1:

Tier	Findings	Treatment	Human effort per finding	Quarterly human effort
------	----------	-----------	--------------------------	------------------------

Auto	3,150 (75%)	Fully automated. Policy-based disposition. GTR auto-created with five fields.	None	0 hours
Standard	714 (17%)	Team-lead or peer-review authorization via existing development platform. Owner auto-assigned from identity resolution.	~15 minutes (review notification, approve/decline with reason)	~178 hours (~1.1 person-months)
Elevated	294 (7%)	Finance review at stage 4. Named authorization with delegation of authority. Full GTR.	~90 minutes (evidence review, commitment check, authorization)	~441 hours (~2.8 person-months)
Executive	42 (1%)	Executive authorization. Audit-grade record.	~3 hours (cross-functional review, formal sign-off)	~126 hours (~0.8 person-months)
Total	4,200			~745 hours (~4.7 person-months; ~1.4 FTE)

Three observations are worth drawing from this distribution.

First, the automation tier absorbs the majority. Three-quarters of findings never enter a human queue. The GTR exists (every finding has a record), but it is five auto-populated fields written by the system at the moment of disposition. The engineer whose non-production dev instance is terminated by a cleanup policy never sees a review request; the record is created for audit completeness, not for human consumption.

Second, the human effort concentrates where value concentrates. The 336 findings in the Elevated and Executive tiers represent 8% of the population by count but, in a typical skewed distribution, 60–80% of the population by dollar value. The quarterly effort is not distributed evenly across 4,200 decisions; about three-quarters of it (567 of ~745 hours) is concentrated on the decisions where evidence, finance review and formal authorization change the answer,

exactly as they did in the §3.8 example where finance review prevented a commitment-transfer illusion.

Third, the loop's own cost is measurable. At approximately 745 hours per quarter governing \$48 million in annual spend, the framework has a unit cost that can be compared against its verified savings output. If the Elevated and Executive tiers produce \$2.4 million in verified annualized savings (a 5% yield on total spend, applied only to the reviewed population), then the governance cost, at typical fully loaded analyst rates, is on the order of 10–15% of the savings it verifies. Where that ratio is unfavorable, Principle 4 directs that the materiality threshold be raised, not that the governance overhead be absorbed as an article of faith. A control whose cost exceeds its yield is a bureaucracy; a control that measures its own cost and adjusts its threshold accordingly is a discipline.

What this example does not show. It does not show finding interactions, namely cases where rightsizing instance A changes the load profile on instance B, invalidating B's recommendation. It does not show the verification pipeline operating across overlapping changes in the same environment. And it does not show the GTR database at steady state, after four quarters of accumulation. These are implementation challenges, not framework challenges, but they are real, and an organization planning a pilot should scope them. The companion document described in Section 7 is the appropriate vehicle for that specification.

4. How CCG Extends FinOps and TBM



Figure 4: TBM, FinOps and CCG: a division of labor

A simple division of labor describes the relationship without diminishing any of the three.

TBM provides the taxonomy: the mapping of technology cost to services, applications and business capabilities. It answers *what does this cost relate to?*

FinOps provides the operating practice: the culture, capabilities and analysis that generate and prioritize optimization across an expanding set of scopes. It answers *what should we do?*

CCG proposes the control specification: the evidentiary, authorization and verification requirements applied to each decision. It answers *did we do it, who said so, and what did it actually deliver?*

Within the FinOps Framework, CCG sits most naturally as a deepening of the **Policy & Governance** capability (renamed from "Cloud Policy & Governance" in Framework 2025; FinOps Foundation, n.d., 2025a; Microsoft, n.d.), with dependencies on Allocation, Workload Optimization, Reporting & Analytics, and Quantify Business Value. If the FinOps Foundation were to absorb these ideas into that capability's guidance rather than as a separate framework, that would be a legitimate and arguably preferable outcome. This paper does not require CCG to exist as a distinct discipline in order for its argument to be useful.

5. Implications

5.1 For Finance Leaders

The practical ask is narrow: accept a defined, bounded role at stage 4 of the loop for decisions above a materiality threshold that finance itself sets. In return, finance gains a defensible realized-savings figure, an evidentiary trail in a familiar form, and an input to forecasting that has been tested rather than assumed. The threshold is the critical design choice. Set too low, finance becomes a bottleneck on trivial decisions and the process collapses. Set sensibly, the reviewed population should be small by count and large by value, following the ordinary distributional logic of any materiality threshold. This paper does not propose specific figures; no published dataset supports one, and organizations should derive their own from their own finding distribution.

Who staffs stage 4. This is the most common practical question the framework raises, and it does not require a new headcount line in most organizations. Stage 4 is a review, not a build: it asks whether a proposed change interacts with a commitment, an allocation rule, or an externally reported metric, which are questions a FinOps analyst, a technology finance business partner, or an existing FP&A function already fields informally today. What CCG adds is not a role but a checkpoint: the review happens before implementation rather than after the variance, and it is recorded rather than remembered. Organizations already running a FinOps practice with a finance-facing capability can typically staff stage 4 from within it; the change is procedural, not organizational. Only where the reviewed population is large enough to require dedicated capacity (itself a function of the materiality threshold discussed above) does this become a staffing decision rather than a reallocation of existing review effort.

The loop's own cost should be measured, not assumed. Governance applied to a finding consumes analyst, finance and engineering time; at scale, the control has a unit cost per decision. Organizations should track cost-per-governed-decision alongside realized savings and set the materiality threshold so that the former is a small fraction of the latter. A control whose cost is unmeasured is, like the estimates it governs, a hypothesis, and Principle 5 applies to the framework's own economics as much as to the savings it reports. Where the reviewed population cannot clear that bar, the correct response is to raise the threshold, not to abandon verification.

5.2 For Engineering

The value proposition is inverted from how governance is usually presented. CCG's Principle 2 places the evidentiary burden on the party proposing the change, not the engineer absorbing its risk. Correctly implemented, it should *reduce* the volume of unevidenced recommendations reaching engineering queues and provide documented cover for changes that are made. Incorrectly implemented as a universal approval gate, it will reproduce exactly the failure mode Section 6.1 describes, and engineering's objection will be correct.

5.3 For CIOs and CTOs

The 78%/8% reporting split in Section 1.2 is not only a measurement fact; it is also the shape of a political question that this paper should not pretend away. FinOps already reports into the technology organization in most enterprises. CCG's stage 4 asks the CIO to route a defined class of decisions through a finance checkpoint before implementation, which can read, at first pass, as ceding a control point that currently sits inside technology.

The more accurate reading is the reverse. An unverified savings number is a liability in a budget conversation, not an asset: it can be, and often is, discounted or disputed by finance precisely because no one outside the technology organization vouches for it. A verified figure, produced through a review the CIO's own organization retains operational ownership of, is a number finance has already agreed to. That is leverage in the next budget cycle, not a loss of authority in this one. What CCG asks the CIO to give up is the unilateral right to assert a savings figure without review; what it offers in exchange is a figure finance cannot later unwind. For a function whose credibility is periodically tested by exactly that kind of dispute, the trade is favorable.

5.4 For Boards and Audit Committees

The relevant question is not "how much did we save?" but "on what basis is that figure stated?" A board-appropriate view of technology cost governance would show: findings raised and their disposition; realized savings measured against an adjusted baseline, stated separately from estimates; the population of decisions subject to formal authorization and the exceptions to it; and the reallocation of realized savings into budget or forecast. Three or four figures, each of which can survive a follow-up question. Estimated savings, aggregate potential, and cost avoidance do not belong in that view. And there is a simple test of whether the underlying record is sufficient: internal audit should be able to re-perform the stage 6 verification from the retained record alone. If it cannot, the record is insufficient, whatever the reported figure.

The Governed Transaction Record

What the register of decisions retains for each cost decision — the schema the loop writes to.

GOVERNED TRANSACTION RECORD

SAMPLE · CCG-2026-0147

1 · IDENTITY

DECISION ID

CCG-2026-0147

TITLE

Rightsize production app tier (14 instances)

STATUS

Stage 6 · Verify — opened 04 Aug, closed 11 Sep 2026

2 · EVIDENCE

EVIDENCE PACK

Telemetry + utilisation refs (CPU 11%, mem 24%)

MATERIALITY

Above threshold — full loop, no collapse

ESTIMATE

\$24,100 annualised — not reportable

3 · OWNERSHIP

ACCOUNTABLE OWNER

Named individual — answers “who decides?”

DECISION RIGHTS

Approve · decline · defer — with recorded reason

A finding with no named owner is not a governed decision.

4 · FINANCIAL REVIEW

SECOND-ORDER EFFECTS CHECKED

Commitments · allocation rules · reported metrics

REVIEWER

Finance (FP&A) — before implementation

Recorded, not remembered.

5 · AUTHORITY

DECISION

Approved — delegation of financial authority stated

GRANTED BY

Person + delegation — or policy, for automated paths

Peer review counts only when it cites the evidence, the materiality assessment, and the authority.

6 · VERIFICATION

VERIFIER

Independent — did not authorise, implement, or own the affected budget

BASELINE

Immutable record + adjustments (-\$3,450)

VERIFIED OUTCOME

\$15,950 — the only figure reported as a result

DISPOSITION — FIRST-CLASS OUTCOMES, NOT FAILURES

REALISED

DECLINED

DEFERRED

EXPIRED

VERIFICATION FAILED

expired counts against conversion

The audit test: internal audit must be able to re-perform the stage 6 verification from this record alone.

If it cannot, the record is insufficient — whatever the reported figure.

Figure 5: The Governed Transaction Record

Figure 5 shows the schema of that retained record: the fields each governed transaction carries from evidence through independently verified outcome, as defined in §3.2.

6. Limitations and Counterarguments

Intellectual honesty requires that the strongest objections be stated rather than answered pre-emptively.

6.1 The change-approval objection is evidence-based and serious

DORA's research program has consistently found that formal change management requiring approval from an external body (a change advisory board or senior manager) is associated with worse software delivery performance. Respondents whose organizations used such processes were 2.6 times more likely to be low performers, and DORA found no evidence that a more formal external review process was associated with lower change failure rates (DORA, 2019; DORA, n.d.). DevOps practice spent fifteen years dismantling approval gates on the strength of that finding.

This paper does not dispute it. Principle 4 exists because of it. But the tension is not fully resolved, and a candid framework should say so. CCG's position is that DORA's finding concerns *external approval of technical change*, while CCG's stage 5 concerns *authorization of financial consequence*, and that where the two coincide, peer review captured in the development platform is an acceptable authorization mechanism, as DORA itself suggests for segregation of duties. Whether that distinction survives contact with practice is an empirical question this paper cannot settle. The practical test it implies is narrow, and worth stating: where DORA's

evidence condemns external approval bodies, CCG requires none; it requires that the approval which already exists, whoever gives it, be recorded against the decision with its financial authority stated.

6.2 The materiality objection

For many organizations, cloud spend is not material to the consolidated P&L, and the governance overhead CCG proposes would not be proportionate. CCG is likely to be relevant only where technology spend is large in absolute terms, volatile, or subject to specific regulatory or margin-reporting pressure. It should not be presented as universally applicable.

6.3 The verification problem is not solved

There is no equivalent of IPMVP for cloud. Baseline adjustment for business volume, rate changes and seasonality is conceptually straightforward and practically difficult. This is the framework's largest open weakness and it warrants more than an acknowledgment.

The specific technical difficulties are catalogued here so that their scope is visible.

Overlapping changes. A verification window rarely contains a single change. If three rightsizing actions, a commitment purchase, and an architectural migration all affect the same account in the same ninety-day period, isolating the financial effect of any one of them requires either sequential implementation with sufficient separation (operationally impractical) or a decomposition methodology that attributes cost movement to individual changes (analytically demanding and inherently approximate). Energy-sector M&V faces the same problem with interactive effects between conservation measures and solves it imperfectly through stipulated values and engineering judgment. Cloud will need to do the same, and the stipulations will be contestable.

Shared and blended infrastructure. Kubernetes clusters, shared databases, multi-tenant platforms and pooled commitment instruments all create attribution problems. Rightsizing a pod within a shared cluster changes the cluster's bin-packing efficiency, which may or may not change the number of nodes, which may or may not change the bill. The saving is real at the pod level and may be invisible at the invoice level until enough pod-level changes accumulate to release a node. Verification must decide whether to measure at the resource level (where the change is visible but the financial effect is not) or at the billing level (where the financial effect eventually appears but attribution to a specific decision is lost).

Rate-change contamination. Enterprise discount programs (EDPs), savings plans, committed-use discounts, private pricing agreements, and provider list-price revisions all change the unit cost of the same resource between the baseline period and the verification period. A rightsizing action that reduces compute hours by 20% may show a 25% bill reduction if a discount program matured in the same period, or a 15% reduction if a list-price increase offset part of the usage reduction. The baseline adjustment must isolate the volume effect from the rate effect, a decomposition familiar in financial analysis (variance analysis separates price, volume and mix) but not yet standardized for cloud cost verification.

Organic demand changes. If a workload's traffic declines due to customer churn, seasonal patterns, or business contraction, the associated cost reduction is not a saving; it is a demand change. Conversely, if traffic grows, a rightsized resource that maintains the same bill has produced a genuine saving measured in cost-per-unit-of-output, even though the absolute bill did not fall. Verification requires a normalization variable (transactions, active users, API calls, or another demand proxy), and the choice of that variable is itself a judgment call that different stakeholders may contest.

Currency and regional pricing. Multi-region deployments introduce exchange-rate effects and region-specific pricing differences that are exogenous to any optimization decision but affect the measured outcome.

Closest existing analogues. IPMVP defines four verification options, ranging from Option A (engineering estimates with partial measurement) through Option D (calibrated simulation). For cloud, the closest practical analogues are:

- **Option A (key parameter measurement):** Measure the key variable that changed (e.g., instance hours eliminated) and stipulate the unit cost. This is the simplest approach and the one most organizations use today when they measure at all. Its weakness is that it ignores interactive effects and rate changes.
- **Option B (retrofit isolation):** Measure all energy (cost) of the affected system before and after, with adjustments. This maps to comparing the metered spend of the specific resources affected, adjusted for volume and rate, which is substantially what the worked example in §3.8 demonstrates.
- **Option C (whole-facility measurement):** Measure total facility (account, project, subscription) cost and attribute the change using statistical methods. This is appropriate when many changes are made simultaneously and individual isolation is impractical, but it requires regression or decomposition techniques that are not yet standardized for cloud.
- **Option D (simulation):** Use a calibrated model to predict what cost would have been without the change. This is the most rigorous and the least practical for cloud at present, though digital-twin approaches to infrastructure modelling may eventually support it.

A realistic assessment of current capability. Most organizations that measure savings at all operate at Option A (key parameter measurement with stipulated unit costs). A smaller number operate at Option B for high-value changes. Options C and D are largely academic in the cloud context today. This means that Principle 5's verification standard can be partially met now, using Option A or B methods with disclosed assumptions and stated error bounds, and more fully met as methods mature. The paper does not claim that perfect verification is achievable today. It claims that *disclosed, methodical, independently reviewed* verification, even with acknowledged imprecision, is materially more useful than the alternative, which is no verification at all and a reported number that is purely an estimate.

Organizations adopting CCG should expect to begin verification at Option A for most finding types, advance to Option B for material changes, and treat Options C and D as research objectives rather than near-term requirements. The maturity model in §3.7 reflects this progression: Level 4 requires "baseline-adjusted M&V," not perfect M&V, and Level 5's "continuous M&V with estimate calibration" is a destination, not a prerequisite.

6.4 Framework proliferation

CIOs face a legitimate objection that the field does not need another framework, another maturity model and another assessment. The response offered here is that CCG is deliberately small (five principles, one loop, and one reportability rule), and that its most likely useful destination is absorption into an existing framework rather than independent survival.

7. Future Research

Five questions would materially strengthen or falsify the propositions in this paper.

1. **Conversion rate measurement.** What proportion of generated recommendations are implemented, declined, deferred or silently expired, and how does that vary by evidence quality, ownership clarity and organization size? No credible public dataset currently answers this. It is the empirical foundation the framework lacks.
2. **AI workload economics.** With 98% of practices now managing AI spend (FinOps Foundation, 2026a) and token- and inference-based pricing lacking established optimization playbooks, does the CCG loop hold, or do decision cycles compress below the point at which per-decision authorization is feasible? Section 2.4 argues that AI economics strengthen the case for continuous control and offers a preliminary answer: that the governed object is the configuration decision (routing policy, model selection, caching strategy), not the individual inference call. The opposite conclusion, that AI decision velocity defeats per-decision governance entirely, is equally testable and would falsify a central premise of this paper. Given that §2.4 now positions AI as the primary urgency driver for the framework, this question cannot be deferred behind the others.
3. **A measurement and verification protocol for cloud.** Can IPMVP's option structure and adjusted-baseline methodology be adapted to consumption-based technology spend with acceptable error bounds? Section 6.3 catalogues the specific technical difficulties (overlapping changes, shared infrastructure, rate-change contamination, and organic demand variation) and maps them to IPMVP options. What normalization variables are required, and how should attribution be handled when multiple changes overlap?
4. **The friction/realization trade-off.** At what materiality thresholds does added authorization control begin to reduce total realized savings by suppressing implementation? This is directly testable and would resolve the tension in Section 6.1. The routing model in §3.5.1 and the enterprise-scale example in §3.9 provide a testable starting structure.
5. **Audit utility.** Do internal and external auditors regard per-decision cost governance evidence as useful, and under what circumstances would it affect testing approach or reliance?

This paper is deliberately conceptual: it addresses *why* the control specification is needed and *what* it must contain. A separate companion document addressing *how* (data structures for the decision record, evidence sufficiency thresholds by workload class, baseline adjustment methods, and reference integrations with change and GRC systems) would be the natural next artefact, and is better written after the questions above have been tested than before.

8. Conclusion

FinOps solved a problem that genuinely needed solving, and continues to solve harder versions of it. Its trajectory (up into technology strategy, left into engineering workflow, and out across an expanding set of technology scopes) is the right trajectory for what it is.

But that trajectory leaves a specific seam unattended. Decisions with financial consequence are made continuously, by engineers, on the basis of estimates, outside the control environment that finance applies to expenditure of comparable size. Continuous Cost Governance proposes a specification for closing that seam: govern the decision, not the report; require evidence proportionate to consequence; name an owner; authorize in proportion to risk; verify against an adjusted baseline; and report only what was verified.

Little of this is new. The components are drawn from internal control, continuous auditing, change management and energy-sector measurement and verification, and the paper has tried to say so plainly. What may be new is the assembly, the object being governed, and the discipline of refusing to report an estimate as a result.

This paper proposes the framework. It does not validate it. Validation requires the empirical work described in Section 7, and the framework should be treated as a hypothesis until that work exists.

References

- Amazon Web Services. (2024). *Cost Optimization Pillar: AWS Well-Architected Framework*. <https://docs.aws.amazon.com/wellarchitected/latest/cost-optimization-pillar/welcome.html>
- ASHRAE. (2023). *ASHRAE Guideline 14-2023: Measurement of Energy, Demand, and Water Savings*. ASHRAE.
- Canadian Institute of Chartered Accountants & American Institute of Certified Public Accountants. (1999). *Continuous Auditing*. Research report. CICA/AICPA.
- CFO Dive. (2021, June 8). *Companies "prone to self-delusion" when using cloud, McKinsey says*. <https://www.cfodive.com/news/companies-prone-to-self-delusion-when-using-cloud-mckinsey-says/596505/>
- Committee of Sponsoring Organizations of the Treadway Commission. (2013). *Internal Control: Integrated Framework*. COSO.

DORA. (2019). *Accelerate: State of DevOps 2019*. Google Cloud / DORA.
<https://dora.dev/research/2019/dora-report/>

DORA. (n.d.). *Capabilities: Streamlining change approval*.
<https://dora.dev/capabilities/streamlining-change-approval/>

Efficiency Valuation Organization. (2018). *IPMVP: Generally Accepted M&V Principles*.
<https://evo-world.org/>

Efficiency Valuation Organization. (2022). *International Performance Measurement and Verification Protocol: Core Concepts 2022*(EVO 10000-1:2022). <https://evo-world.org/>

FinOps Foundation. (2025a). *FinOps Framework 2025*. <https://www.finops.org/insights/2025-finops-framework/>

FinOps Foundation. (2025b). *FinOps Maturity Model*.
<https://www.finops.org/framework/maturity-model/>

FinOps Foundation. (2026a). *State of FinOps 2026*. <https://data.finops.org/>

FinOps Foundation. (2026b). *FinOps Open Cost and Usage Specification (FOCUS)*.
<https://focus.finops.org/>

FinOps Foundation. (n.d.). *Policy & Governance capability*.
<https://www.finops.org/framework/capabilities/>

Flexera. (2026). *State of the Cloud Report 2026*. <https://info.flexera.com/CM-REPORT-State-of-the-Cloud>

Gartner. (2024, November 19). *Gartner forecasts worldwide public cloud end-user spending to total \$723 billion in 2025* [Press release]. <https://www.gartner.com/en/newsroom/press-releases/2024-11-19-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-total-723-billion-dollars-in-2025>

Gartner. (n.d.). *Continuous controls monitoring (CCM)*. Gartner Information Technology Glossary. <https://www.gartner.com/en/information-technology/glossary/continuous-controls-monitoring-ccm>

McKinsey & Company. (2021, February 25). *Clearing the air on cloud: How industrial companies can capture cloud technology's full business value*.
<https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/clearing-the-air-on-cloud-how-industrial-companies-can-capture-cloud-technologys-full-business-value>

Microsoft. (n.d.). *Policy and governance: FinOps Framework*. Microsoft Learn.
<https://learn.microsoft.com/en-us/cloud-computing/finops/framework/manage/governance>

Storment, J. R. (2025, December 8). *FinOps Enabled Executive: Two new archetypes for technology leadership ("COO for the CTO")*. FinOps Foundation.
<https://www.finops.org/insights/finops-executive-technology-leader/>

TBM Council. (2023). *Technology Business Management Taxonomy*. TBM Council.

Vasarhelyi, M. A., Alles, M. G., & Kogan, A. (2004). Principles of analytic monitoring for continuous assurance. *Journal of Emerging Technologies in Accounting*, 1(1), 1–21.

Weill, P., & Ross, J. W. (2004). *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results*. Harvard Business School Press.

Appendix A: A Measurement and Verification Protocol for Consumption-Based Spend

Status. This appendix is a first attempt at the protocol Section 7 identifies as research question 3. It is offered to be attacked rather than adopted. It has not been applied in any organization, it carries no empirical error bounds, and it is not a standard. The figures continue the illustrative, synthetic finding from Section 3.8.

A.1 Why the paper requires this

Principle 5 states that only verified outcomes are reported as financial results, and defines verification as comparison against a baseline adjusted for business volume, rate changes and seasonality. Section 6.3 then concedes that no cloud equivalent of IPMVP exists. Those two statements together leave the framework's most distinctive principle without a method, and a principle without a method is a slogan.

The energy efficiency sector resolved the same problem structurally. IPMVP does not measure savings directly, because savings are the difference between actual consumption and what consumption would have been, which is unobservable. Instead it specifies a baseline, an adjustment procedure, and a set of measurement options whose rigour is matched to the materiality of the measure (Efficiency Valuation Organization, 2018). That structure transposes. What does not transpose is the content: cloud has no stable physical plant, its configuration changes continuously, its rates change unilaterally, and its commitment instruments couple unrelated workloads together. This appendix specifies the minimum method required to execute Principle 5, and is explicit about where the transposition fails.

A.2 Definitions and the savings equation

Governed scope. The set of resources to which the authorized change applies, rather than the account, the service, or the resources considered and declined. In the Section 3.8 example the governed scope is the eleven instances authorized at stage 5, and excludes the three recorded as Declined.

Measurement basis. Amortized cost, FOCUS-normalized via the normative EffectiveCost column (FinOps Foundation, 2026b), rather than invoiced cash (BilledCost). BilledCost moves when commitments are purchased upfront, which would register a commitment purchase as a cost increase and its amortization as a saving; unit rate changes (ρ) are similarly evaluated against FOCUS ListCost or ContractedCost, while commitment portfolio state is tracked via the FOCUS 1.3 Contract Commitment dataset. Any protocol operating on billed cash will produce artefacts.

Periods. A baseline period preceding the change and a reporting period following it. Both must span at least one complete cycle of every driver that materially moves the governed scope's cost; in the worked example, this requires at least one financial close, because the fleet's peak is a month-end phenomenon. Both are annualized by the same stated method.

The equation.

$$S = B \cdot (D_r / D_b)^e \cdot \rho - R - \sum A_{nr}$$

Term	Meaning
S	Verified realized saving, annualized
B	Baseline-period amortized cost (EffectiveCost) of the governed scope, annualized
R	Reporting-period amortized cost (EffectiveCost) of the governed scope, annualized
D_b, D_r	Demand driver in the baseline and reporting periods
e	Cost elasticity of the governed scope to that driver, estimated from the baseline window, not assumed
ρ	Rate factor: reporting-period unit rates divided by baseline unit rates, for the baseline configuration
A_nr	Non-routine adjustments: discrete events in the period that are not the authorized change

The terms for demand and rate are the routine adjustments: predictable, recurring, modelled. Non-routine adjustments capture migrations, re-architecture, tenancy or account restructuring, commitment portfolio changes, model substitution, and one-off provider credits. The distinction is IP MVP's and it carries the protocol's integrity, because non-routine adjustments cannot be modelled from the baseline and must therefore be enumerated by someone, which is why they are the easiest place to flatter a result.

A.2.1 The elasticity term is where most error enters

Setting the elasticity to 1 (assuming cost moves proportionally with demand) is the default in informal before-and-after comparison, and it is wrong in the direction that matters. Statically provisioned capacity has an elasticity near zero: its cost is determined by a provisioning decision, not by demand. In IPMVP terms, the multiplicative factor $(D_r/D_b)^e$ is the local constant-elasticity equivalent of a two-parameter linear baseline model $C(D) = C_{\text{fixed}} + c_{\text{var}} \cdot D$, where $e = (c_{\text{var}} \cdot D_b)/B$ is the variable-cost share at baseline (by first-order Taylor expansion, $B \cdot (D_r/D_b)^e \approx B \cdot [1 + e(D_r - D_b)/D_b] = C_{\text{fixed}} + c_{\text{var}} \cdot D_r$). Practitioners may equivalently use the linear fixed-plus-variable specification where demand crosses zero or shifts sharply; what matters is measuring the variable share rather than assuming proportionality.

Applying a unit-elasticity normalization to a static fleet removes genuine saving and reports a smaller number than was achieved. Section A.7.1 quantifies this on the worked example: the error is 11% of the reported figure, from a single unexamined assumption.

A.2.2 Pre-registration of the adjustment set

The adjustment set (which drivers, which elasticity estimate, which non-routine events, and which windows) must be specified and recorded before implementation, as part of the stage 4 review. Adjustments selected after the result is known are not verification; they are negotiation. This is the protocol's most important control and among the cheapest to implement, because stage 4 already exists and already produces a record.

A.2.3 Rounding

The worked example rounds line items to the nearest \$50 and states so. A protocol that does not fix its rounding convention will produce figures that fail to reconcile across reports, which is a disproportionately common way for an otherwise sound number to lose credibility.

A.3 Four measurement options

Transposed from IPMVP's Options A to D, with rigour matched to materiality per Principle 4.

Option	What is measured	What is stipulated	Cloud application	Reportable under Principle 5?
A: isolation, key parameter	The changed parameter only: resource count or size times amortized unit rate	That the demand envelope is unchanged, within a stated tolerance	High-volume, low-materiality configuration changes resolving on automated paths	Yes, within the tier limits in A.6
B: isolation, all parameters	Full amortized cost of the governed scope in both periods, at resource	Nothing material	A single identifiable finding, which is the default for anything	Yes

	granularity		reaching stages 4 and 5	
C: whole-scope regression	Total amortized cost of an account, subscription or product scope, regressed on drivers	Resource-level attribution	Diffuse or policy-driven change across many resources, where isolation is impossible	Only if the detectability test in A.7.2 is met
D: calibrated counterfactual	Nothing; the baseline is modelled	The entire baseline	Net-new workloads, prevented spend, and architectural avoidance, where no measured baseline exists	No. Operational tracking only

The Option D exclusion is the protocol's sharpest consequence, and it is the mechanism by which Principle 5 excludes cost avoidance. Avoided cost is, by construction, an Option D measurement: it compares actual spend against a counterfactual that was never instrumented. This is not a judgement that avoidance is worthless; preventing a badly provisioned workload from reaching production is often worth more than rightsizing it afterwards. It is a judgement that avoidance cannot be verified, and Principle 5 governs reporting, not value. Section A.9 records the cost of that position.

A.4 Normalization drivers by workload class

The driver must be selected from the cost structure of the workload, not from whichever business metric is nearest to hand. The elasticity column states a priori expectations only, offered as hypotheses to be tested per estate; consistent with this paper's practice elsewhere, no measured values are asserted.

Workload class	Primary driver candidates	Expected elasticity (to be measured, not assumed)	Notes
Statically provisioned compute	Transactions, requests, active users	Low	Cost is set by provisioning, not demand. Demand normalization is small and must be measured.
Autoscaled and serverless compute	Invocations, concurrency-seconds, requests	High	The case where unit elasticity is defensible, and close

			to the only one.
Object and block storage	GB-months retained, objects written, retrieval volume	Low for retained, high for retrieval	Retained volume is strongly autocorrelated; regress on level, not change.
Data platform and warehouse	Queries, bytes scanned, slot- or credit-hours	Moderate to high	Query mix shift is a confound; segment by workload before regressing.
AI inference	Input and output tokens, calls, context length distribution	High, but non-linear in context length	Model substitution is a non-routine event, not a routine adjustment.
SaaS and metered licensing	Active seats, metered units	Step-wise, not continuous	Tier boundaries break the regression; model as steps, not a slope.

A.5 Attribution when changes overlap

Overlapping changes are the problem Section 7 names, and the one most likely to defeat the protocol in a real estate. Five rules, in precedence order:

1. **Sequence where possible.** Verification windows for findings against the same resources should not overlap. Sequencing is cheaper than attribution.
2. **Isolate where separable.** Where overlapping changes touch disjoint resources, Option B isolation resolves the attribution without further machinery.
3. **Aggregate where inseparable.** Where overlapping changes touch the same resources and cannot be separated, the findings are merged into a single verification unit and verified jointly. The joint figure is reported; it is not allocated to the constituent findings for financial reporting. Pro-rata allocation by estimate is permitted for operational tracking and estimate calibration only.
4. **Rate and commitment effects belong to the portfolio.** Savings arising from commitment purchases, rate negotiations or discount instrument changes are attributed to the commitment portfolio owner, never to a usage-side finding that happens to coincide with them. Without this rule the same dollar is credited twice (once to the engineer who reduced usage and once to the analyst who bought the commitment), and the sum of verified savings exceeds the reduction in the bill.
5. **Uniqueness.** The decision register enforces a constraint on resource and period: no resource-period may be credited to more than one verified finding. Where rule 3 merges findings, the merged unit holds the claim.

A.6 Rigour proportionate to materiality

Tier	Trigger	Required option	Minimum windows	Verified by
1	Below the automated-routing threshold	A	30-day reporting against existing telemetry	Automated, with periodic sampled audit
2	Above the routing threshold, non-tier-one service	B	90 / 90	FinOps, with finance sign-off on the pre-registered adjustment set
3	Financially material or tier-one service	B	90 / 90, extended to span a full business cycle	Finance or an independent function
4	Diffuse or program-level	C	180 baseline / 90 reporting	Finance, with the detectability test documented

Tier 1 is deliberately permissive. A protocol that demands Option B of every finding will be applied to none of them, which is the failure mode Principle 4 exists to prevent.

A.7 Worked calculation

Continuing the Section 3.8 finding. Tier 3, Option B. All figures synthetic.

Input

Governed scope	11 instances authorized at stage 5 (3 Declined, out of scope)
Measurement basis	Amortized cost, FOCUS-normalized
Baseline period	90 days pre-change, spanning two financial closes
Reporting period	90 days post-implementation, spanning two financial closes
B: baseline annualized	\$38,400
R: reporting annualized	\$19,000
Driver	Settled transactions
D_b → D_r	42.0m → 39.5m (−6.0%)
e: measured elasticity	0.19 (standard error 0.07)
p: rate factor	0.951 (a 4.9% list-rate reduction on the in-scope instance family, pro-rated over the window)
A_{nr}	\$1,150 commitment reallocation (see below)

The elasticity estimate. The elasticity is estimated by ordinary least squares of daily amortized in-scope cost (EffectiveCost, which includes fixed instance-hour compute alongside usage-metered network egress and storage I/O attached to the eleven instances) on daily settled transactions across the baseline window, in logs. The regression returns 0.19 with an R^2 of 0.21, explaining only a fifth of daily cost variation. That weak fit is not a defect in the estimate; it is the finding. A statically provisioned fleet's cost is largely insensitive to demand (with only its attached transfer and I/O components varying with transaction volume), which is precisely why it was able to sit at 11% mean CPU in the first place.

The non-routine adjustment. Stage 4 directed that freed capacity be backfilled by migrating an uncommitted workload into the three-year commitment. That backfill is a real enterprise benefit, but \$1,150 of the governed scope's observed reduction is a transfer: committed units the in-scope workload no longer consumes are now consumed by the backfilled workload. Under rule 4 of Section A.5 that benefit is attributed to the commitment portfolio, not to this finding. Crediting it here would double-count it against the backfilled workload's own record.

The calculation.

Step		
Demand factor	$(39.5 / 42.0)^{0.19}$	= 0.98841
Demand-adjusted baseline	$38,400 \times 0.98841$	= 37,954.9
Normalized baseline B*	$37,954.9 \times 0.951$	= 36,095.1
Routine savings	$36,095.1 - 19,000$	= 17,095.1
Less non-routine adjustment		- 1,150.0
Verified realized saving		= 15,945.1 → \$15,950

Reconciliation to the Section 3.8 table. Decomposing $B^* - B$ sequentially in equation order (applying the demand factor to B gives -\$445.15, and applying the rate factor to the demand-adjusted baseline gives -\$1,859.79, each rounded to the nearest \$50 per Section A.2.3):

	Annualized
Gross observed reduction (B - R)	\$19,400
Less: demand normalization	(\$450)
Less: provider list-rate reduction	(\$1,850)
Less: commitment reallocation (non-routine)	(\$1,150)
Verified realized saving	\$15,950

A.7.1 What the unit-elasticity assumption would have cost

Had the elasticity been assumed at 1.0 rather than measured at 0.19, the demand normalization alone would have deducted \$2,286 rather than \$445. Carried through the full equation, with the rate factor applied to the smaller demand-adjusted baseline, the verified saving would have

been \$14,195, understating the reported figure by \$1,750, or 11%, from one unexamined modelling assumption. The team would have reported a number that was wrong in the direction of pessimism, and would have had no way to discover it. This term is the strongest practical argument in this appendix for specifying a protocol at all.

Uncertainty. Propagating the standard error on the elasticity through the demand term gives a 95% interval on the verified saving of approximately \$15,600 to \$16,300, or about $\pm 2\%$. The magnitude of that interval matters less than its size relative to the alternative: the gap between the naive estimate at finding (\$24,100) and the verified result (\$15,950) is \$8,150. The protocol's imprecision (a half-width of $\pm \$350$) is roughly 4% of the error it removes. Imprecision in verification is not an argument for reporting the unverified figure.

A.7.2 Option C and the detectability test

Where a change is diffuse (for example, a storage lifecycle policy applied across thousands of buckets), isolation is impossible and the baseline becomes a regression of scope-level cost on drivers. The hazard is that a real but small saving disappears into the scope's own noise, while the regression nonetheless returns a point estimate that looks like a result.

The protocol therefore transposes the compliance threshold of ASHRAE Guideline 14 (ASHRAE, 2023; Efficiency Valuation Organization, 2022), which requires that Fractional Savings Uncertainty ($FSU = \Delta S/S$) not exceed 50% at the 68% confidence level (1σ). Algebraically, requiring $1 \cdot SE(S)/S \leq 0.50$ is identical to a two-sigma detectability test: a saving is reportable under Option C only if its expected magnitude exceeds twice the standard error of the cumulative reporting-period residual ($S \geq 2 \cdot SE(S)$), computed from the baseline regression and adjusted for lag-1 serial correlation (ρ_1) in daily cost via ASHRAE Guideline 14's effective sample size $n' = n(1 - \rho_1)/(1 + \rho_1)$. Illustratively, a scope whose autocorrelation-adjusted baseline regression leaves a residual standard error equivalent to $\pm \$12,000$ annualized (so $2 \cdot SE = \$24,000$) supports reporting a \$96,000 expected effect (= 12.5%) and does not support reporting an \$18,000 one (= 66.7%). The second is tracked operationally and excluded from financial reporting, not because it is untrue, but because the method cannot distinguish it from the scope's ordinary variation.

A.8 Calibration: closing the loop back to stage 2

Verification's second output is a correction to the estimator. For this finding:

Ratio	Value
Verified divided by naive estimate at finding	0.66
Verified divided by post-evidence estimate	0.84

The second ratio is the useful one, because it measures the estimator the organization would actually act on. Accumulated across a population of verified findings and segmented by workload class, it yields a class-level realization factor that stage 2 can apply to future estimates,

converting the "verification failed" outcome of Section 3.5 from an embarrassment into the only available training signal.

One observation is not a calibration. A single ratio of 0.84 establishes nothing; the quantity of interest is the distribution of that ratio across findings within a class, and its stability over time. This is the mechanism by which research question 1 in Section 7 becomes answerable from operational data rather than from a survey.

A.9 What this protocol does not solve

1. The counterfactual remains unobservable. Every figure the protocol produces is an estimate with a stated method. Verification narrows the error and makes it inspectable; it does not convert an estimate into a measurement of fact.
2. Elasticity estimates are unstable where drivers are collinear, where the baseline window is short, or where the workload is itself changing. The protocol offers no minimum window length by workload class, because no evidence supports one.
3. Shared-platform attribution is unresolved. The protocol defers to the organization's existing allocation rules rather than fixing them, and inherits whatever contestation those rules carry.
4. No inter-rater reliability evidence exists. Two competent analysts applying this to the same finding may enumerate different non-routine adjustments and report different figures. Pre-registration constrains this but does not eliminate it, and the magnitude of residual disagreement is unknown.
5. The error bounds in Section A.7 are specific to one synthetic case and should not be read as characteristic. No distribution of protocol error across real findings or organizations exists.
6. The Option D exclusion is costly. It rules prevented and avoided cost, plausibly the largest category of engineering value in a mature practice, out of financial reporting entirely. Some organizations will reasonably judge that trade unacceptable, and the protocol offers them no middle path beyond operational tracking.
7. It is untested. No organization has run this protocol. Its publication here is an invitation to falsify it, and Section 7 states the questions that would do so.

About the Author

Andy Surujnarine (ORCID: 0009-0004-0144-1497) is an IT security and cloud operations leader with more than twenty-two years of experience architecting and governing enterprise cloud, hybrid infrastructure, and security programs, and the founder of Compute LLC. His work focuses on cloud financial governance, security operations, and verifiable control design across cloud and AI estates. More of his writing and research is available at surujnarine.net.

Suggested Citation

Surujnarine, A. (2026). *Introducing Continuous Cost Governance: Extending FinOps from Cost Optimization to Financial Accountability* (Working Paper Version 1.5.6). Compute LLC.
<https://doi.org/10.5281/zenodo.23555398>

Continuous Cost Governance is a proposed framework offered for industry review and critique. It is not an industry standard. The framework originated in observations made during the development of a commercial cost governance tool; that origin is disclosed for transparency and the framework is intended to be evaluated on its own merits.