

The Flexibility Premium

A Strategic Framework for Deciding When Vendor Capacity Should Become an In-House Capability



A practical decision framework for operations leaders, CFOs and sourcing executives.

How to Use This Paper

This paper is designed for operations leaders, CFOs, and sourcing executives evaluating whether vendor-delivered work should move in-house. The frameworks can be applied in 2-4 hours of analysis and are intended to complement—not replace—detailed TCO modeling.

EVIDENCE ARCHITECTURE

The Operative case is a 2018 operating decision. Period sources establish the market context around that decision. 2024-2026 research is used to show how the logic has aged—not to retroactively justify the original decision. Case figures are direct people-cost equivalents, not audited total cost of ownership.

Table of Contents

1.	EXECUTIVE SUMMARY	3
2.	INTRODUCTION: THE SOURCING DECISION	4
3.	THE FLEXIBILITY PREMIUM: DEFINING THE CONCEPT	5
4.	THE OPERATIVE CASE: A 2018 DECISION	6
5.	THE FIVE-LENS COST-TO-DELIVER FRAMEWORK	7
6.	SOURCE BY WORK TYPE, NOT IDEOLOGY	9
7.	THE TRANSITION BLUEPRINT: INTERNALIZE WITHOUT BREAKING DELIVERY	10
8.	TESTING THE FINANCIAL CASE	11
9.	CALCULATING YOUR FLEXIBILITY PREMIUM	12
10.	HOW THE CASE HAS AGED: 2018-2026	13
11.	WHEN OUTSOURCING IS STILL THE BETTER ANSWER	14
12.	GOVERNANCE: MAKE THE MODEL SELF-CORRECTING	15
13.	LEADERSHIP LESSONS	16
14.	APPENDIX A: EXECUTIVE DECISION CHECKLIST	17
15.	APPENDIX B: WORK TYPE MATRIX TOOL	18
16.	REFERENCES & EVIDENCE NOTES	18

1. Executive Summary

When “flexibility” becomes permanently paid capacity, the sourcing model can quietly turn into a margin problem. This retrospective paper revisits a 2018 operating decision at Operative, frames the historical context, and then tests the enduring logic against the 2024-2026 sourcing, GCC and AI landscape.

At Operative, 150 vendor FTE were carried at approximately \$2,100 per FTE per month and the capacity was held at only around 60% utilization. Workflow and demand analysis showed that the same scope could be supported by approximately 60-75 internal FTE at about \$900 per FTE per month. The work was internalized and the transition was completed in roughly four months.



At-a-glance: direct people-cost economics of the 2018 Operative decision.

Using 75 internal FTE, direct monthly people cost fell from approximately \$315K to \$67.5K—a reduction of about 79%, or roughly \$2.97M annualized before infrastructure and shared-service costs. The central lesson is not that internalization is always cheaper. It is that demand, utilization and workflow must be made visible before a sourcing model is renewed.

ACTIONABLE TAKEAWAY

Calculate your flexibility premium before every renewal. If the premium is material and the promised flexibility is rarely used, redesign the work and re-test the sourcing model before extending the contract.

PUBLICATION NOTE

This paper deliberately separates historical operating evidence from later market validation. The Operative case is a 2018 decision; references [1]-[3] establish period context, while references [4]-[6] show how the logic has aged. Case figures are direct people-cost equivalents, not audited total cost of ownership. Suggested governance thresholds and contractual considerations are framework recommendations, not historical case facts.

2. Introduction: The Sourcing Decision

The case in this paper is historical; the operating question is not. In 2018, outsourcing was already shifting from simple labor transfer toward technology-enabled transformation. Deloitte's 2018 Global Outsourcing Survey reported 93% of surveyed executives were considering or adopting cloud solutions, 72% were considering or adopting RPA, and about 84% had initiated discussions, pilots or implementation of disruptive solutions. NASSCOM's 2018 GCC Conclave similarly described in-house centers as moving from cost and capacity plays toward strategic capability.[1][2]

THE 2018 SOURCING CONTEXT: TRANSFORMATION WAS ALREADY RESHAPING DELIVERY

Period context for the 2018 case — later research appears only in the Current Perspective section.



Figure 1: The 2018 sourcing context - cloud, RPA and capability-led GCCs were already changing the operating-model conversation.

Sources: Deloitte Global Outsourcing Survey 2018; NASSCOM GCC Conclave 2018.

Figure 1: Historical context around the 2018 sourcing decision. Sources: Deloitte Global Outsourcing Survey 2018; NASSCOM GCC Conclave 2018.

3. The Flexibility Premium: Defining the Concept

The flexibility premium is the amount an organization pays above the best-fit alternative operating model in exchange for elasticity, speed, scarce skills, resilience, risk transfer or other provider advantages. It is economically justified only to the extent that those advantages are actually used and create measurable value.

THE FLEXIBILITY PREMIUM: WHAT ARE YOU PAYING ABOVE THE BEST-FIT ALTERNATIVE?



1. VENDOR TCO

Contracted capacity
+ tools + governance



2. BEST-FIT ALTERNATIVE

Zero-based capacity
+ internal / hybrid TCO



3. REALIZED FLEXIBILITY VALUE

Ramp-down + speed + skills
+ resilience actually used



**Net unrecovered flexibility premium =
Vendor TCO – Best-fit alternative TCO –
Realized flexibility value**



Use this as a decision screen before renewal;
validate the result with detailed TCO modeling.

Figure 2: A decision-screen definition of the flexibility premium. Detailed TCO modeling should follow.

4. The Operative Case: A 2018 Decision

The strongest part of the decision was the sequence. The organization did not begin by asking, “What would 150 employees cost internally?” That would simply have moved the incumbent staffing pyramid from one employer to another. The first step was to diagnose the work itself.

The utilization trap

At approximately 60% utilization, a 150-FTE vendor roster equated to roughly 90 fully utilized FTE of productive capacity before process redesign. Demand and workflow analysis then indicated that 60-75 internal FTE could absorb the scope. The gap was therefore a combination of unit-rate difference and capacity right-sizing.

THE UTILIZATION TRAP: FLEXIBILITY THAT RARELY FLEXED



Figure 3: Paid capacity, productive capacity and required capacity were materially different.

WHAT THE EVIDENCE DOES—AND DOES NOT—SHOW

The case evidence confirms ~60% utilization and that turn-off flexibility was rarely used. It does not establish whether the root cause was contractual minimums, notice periods, demand volatility, forecast conservatism, internal governance or provider constraints. Leaders should diagnose that cause before renewal rather than assume it.

The zero-based approach

A zero-based sourcing review starts from required outcomes and workload, not the incumbent staffing structure. Remove avoidable work, standardize what remains, automate where possible, then size the residual capacity and choose the sourcing model. This prevents an organization from internalizing—or outsourcing—yesterday’s inefficiency.

5. The Five-Lens Cost-to-Deliver Framework

Rate cards are useful, but insufficient. A robust sourcing decision should evaluate five connected lenses: capacity, process, capability, technology/AI, and risk/elasticity. The objective is to understand what the enterprise is really buying and what it is really paying for.

THE FIVE-LENS OPERATING ECONOMICS MODEL

A structured way to design, operate and continuously optimize high-performance delivery organizations.



When all five lenses are aligned, organizations unlock sustainable performance, scalable growth and measurable business impact.

Figure 4: Five lenses for evaluating the real operating economics of a sourcing model.

How the five lenses applied to Operative

THE FIVE-LENS EVALUATION FRAMEWORK

A structured view of the evidence, what it signaled, and how it shaped the decision.

LENS	CASE SIGNAL	DECISION IMPLICATION
 CAPACITY & UTILIZATION	 150 vendor FTE at ~60% utilization	 Capacity was materially above observed productive demand; redesign supported a 60–75 FTE internal target.
 PROCESS & REWORK	 Workflow and demand analysis preceded the decision	 The staffing requirement was re-sized rather than copied 1:1 from the vendor model.
 CAPABILITY & KNOWLEDGE	 Internalization increased ownership and knowledge retention	 The decision traded vendor elasticity for direct control and retained capability.
 TECHNOLOGY & AI	 No case-specific AI claim is made for the 2018 decision	 Modern reviews should automate before sizing residual labor and ensure productivity gains change commercials.
 RISK & ELASTICITY	 Vendor turn-off flexibility was rarely used	 The four-month transition required continuity controls, overlap, KT and stabilization.



INSIGHT: The decision was not just about cost. It was about right-sizing, capability retention, and sustainable control.

OPERATING EQUATION

True cost-to-deliver = paid capacity + technology + management + rework/quality cost + transition/attrition cost + risk premium – productivity and automation benefits.

6. Source by Work Type, Not Ideology

The correct sourcing answer changes by work type. Stable, strategic work often benefits from ownership and accumulated knowledge. Highly variable, standardized work can favor providers that aggregate demand across clients. Strategic but volatile work often benefits from hybrid models. The matrix is a starting point, not a substitute for TCO and risk review.

SOURCE THE WORK BY ECONOMICS AND STRATEGIC VALUE – NOT IDEOLOGY

Choose the right sourcing model based on strategic criticality and demand volatility.

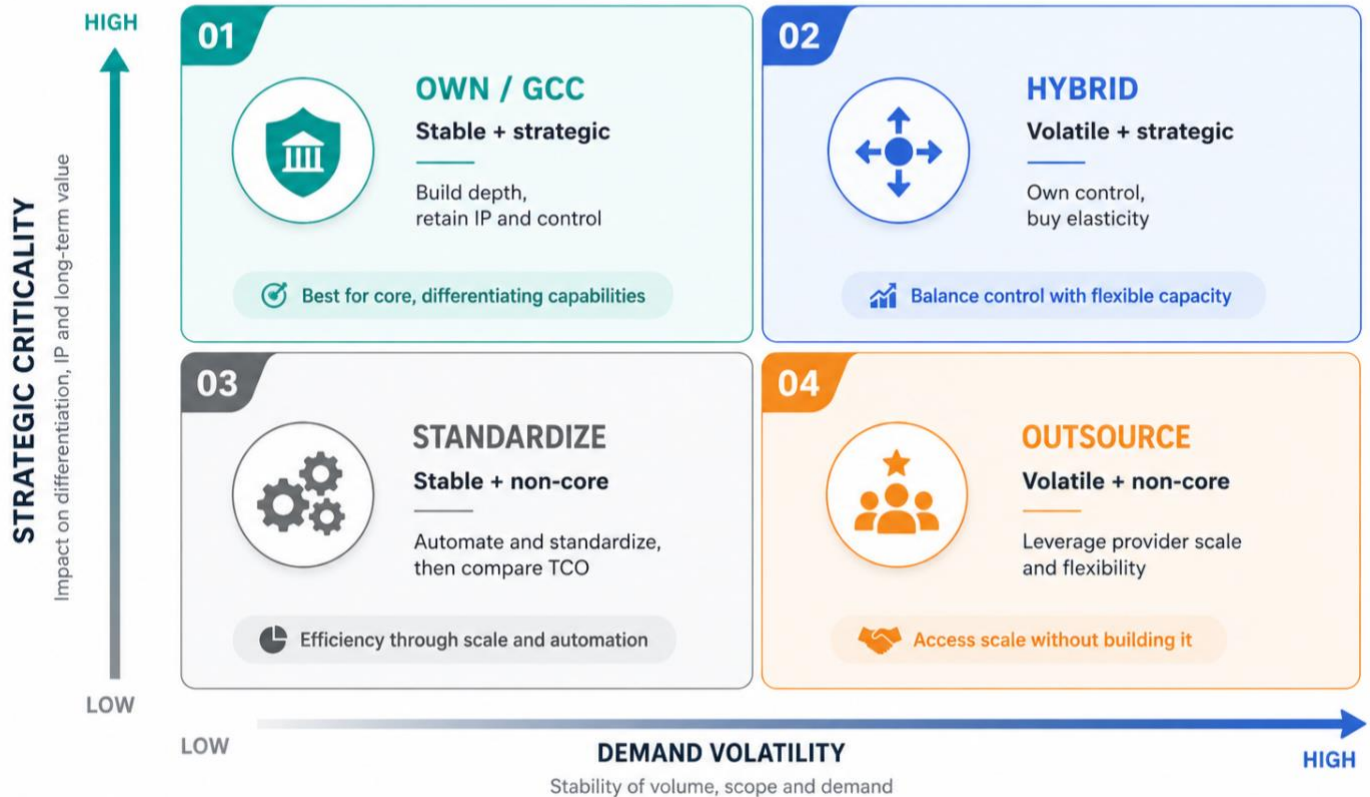


Figure 5: A practical portfolio matrix for deciding where work should live.

Contractual traps that can make “flexibility” expensive



These are general contract-review considerations, not claims about the Operative agreement.

7. The Transition Blueprint: Internalize Without Breaking Delivery

A favorable cost model is meaningless if the transition destabilizes service. The Operative move was completed in roughly four months, so economics and execution had to be managed together. The roadmap below turns that experience into a reusable transition pattern.

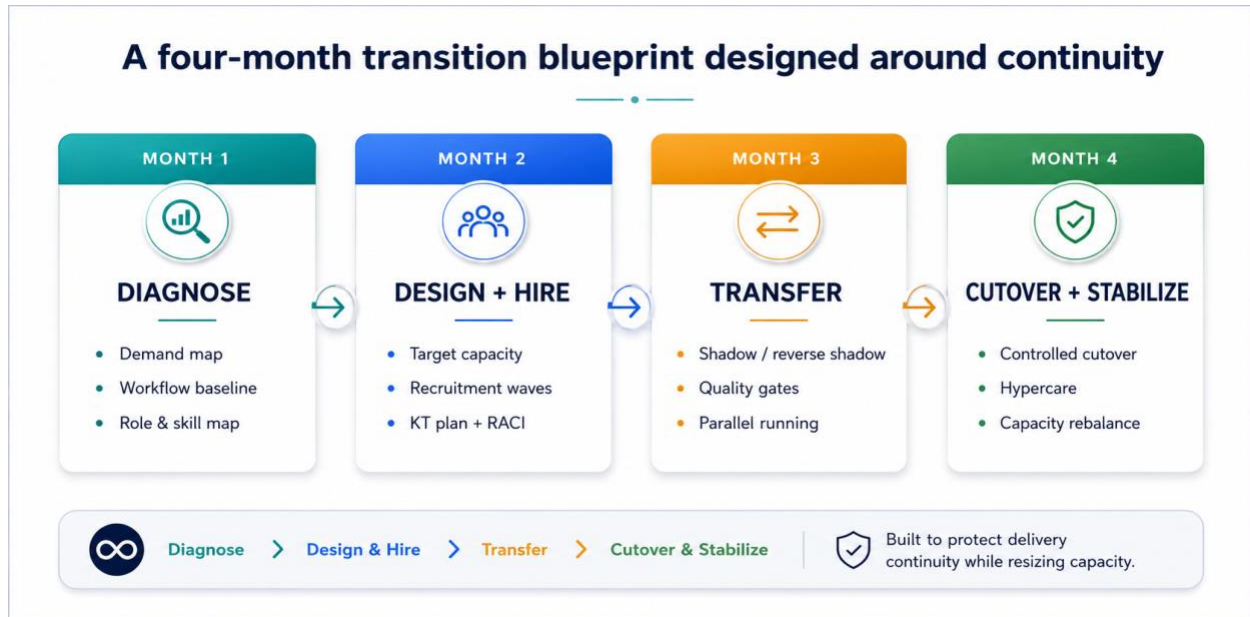


Figure 6: A four-month internalization roadmap centered on capacity design, knowledge transfer and controlled cutover.

Risk mitigation

 RISK	 FAILURE MODE	 MITIGATION
 Service continuity	Backlog, missed SLA, delayed cutover	Parallel running, daily control tower, queue-level cutover gates
 Knowledge loss	Vendor SMEs exit before KT is complete	Named knowledge owners, shadow/reverse-shadow, competency sign-off
 Hiring / skill availability	Internal recruitment lags transition schedule	Wave hiring, critical-role prioritization, contingency capacity
 Access / security	Late systems provisioning blocks readiness	Access checklist, lead-time ownership, pre-cutover validation
 Transition cost	Dual-running and hiring costs erode the business case	Time-box overlap, weekly cost burn, exit criteria for hypercare

The human factor

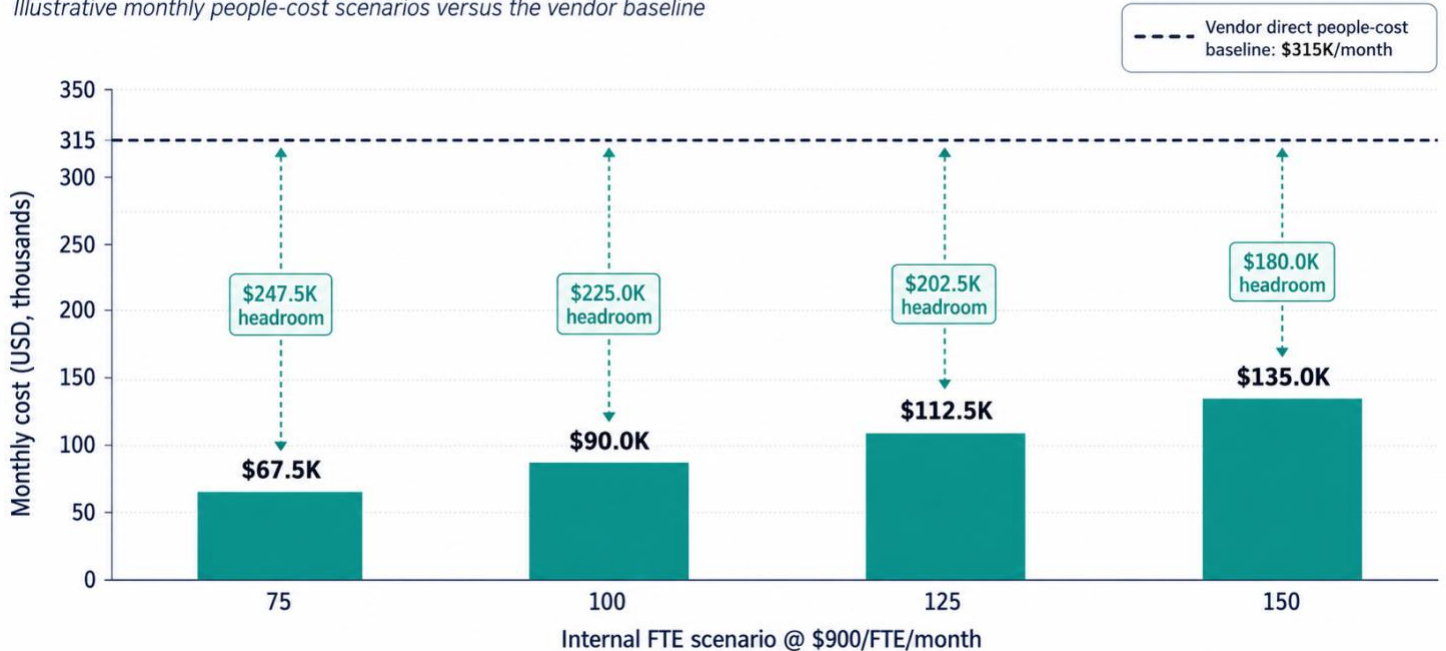
Internalization affects people on both sides of the boundary. Leaders should avoid framing the move as a “winner versus loser” exercise. Communicate only after commercial and legal readiness, protect critical knowledge, create clear internal career paths, onboard managers before volume moves, and use transparent readiness criteria so service teams know what “good” looks like. Where practical, redeployment or fair transition arrangements can reduce disruption and preserve relationships.

8. Testing the Financial Case

The Operative calculation is intentionally narrow: it compares direct people-cost equivalents. A credible business case should then add the costs that do not appear in payroll—recruitment, leadership, facilities, IT, benefits, support functions, tooling, transition overlap, compliance and contingency.

Direct people-cost stress test: how much non-people TCO headroom remains?

Illustrative monthly people-cost scenarios versus the vendor baseline



Headroom = baseline vendor direct people cost minus internal direct people cost.

Figure 7: At 75-150 internal FTE, the direct people-cost gap leaves substantial room for non-people TCO before reaching the \$315K vendor baseline.

DIRECT-LABOR-ONLY BREAK-EVEN

At \$900 per internal FTE per month, direct people cost alone reaches \$315K at 350 internal FTE. This is not the true TCO breakeven because infrastructure, management, benefits, tools and transition costs must also be added.

TCO components to include

- Direct labor and benefits
- Management and governance
- Recruitment, training and attrition
- Facilities, devices, connectivity and IT support
- Enterprise software, tooling and licenses
- Transition overlap and one-time migration costs
- Risk, compliance, BCP and geographic resilience
- Expected productivity and automation benefits

9. Calculating Your Flexibility Premium

Use the below model as a two-to-four-hour decision screen before renewal. It does not replace a finance-approved TCO model; it helps leaders determine whether a deeper re-sourcing analysis is warranted.

Step 1: Calculate the gross premium



THE FLEXIBILITY PREMIUM

A 2-to-4-hour decision screen for fast sourcing decisions



Timebox: 2-4 hours



Goal: quantify the premium



Use case: renew / redesign / re-source



Output: gross and net premium

INPUT	WHAT TO INCLUDE	YOUR VALUE
A A. Vendor monthly TCO	 Contracted capacity + provider fees + tools + governance	\$ _____
B B. Best-fit alternative monthly TCO	 Zero-based required capacity + internal/hybrid overhead	\$ _____
C C. Gross flexibility premium	 A – B	\$ _____
D D. Realized flexibility value	 Quantified value of ramp-down, speed, scarce skills, resilience	\$ _____
E E. Net unrecovered flexibility premium	 C – D	\$ _____

DECISION READOUT



Gross flexibility premium = A – B



Net unrecovered flexibility premium = C – D



Decision question:
Is the premium justified by value actually used?

HOW TO USE

1



1. Estimate vendor TCO



2



2. Model the best-fit alternative



3



3. Quantify the flexibility actually used



4



4. Compare gross vs. net premium



Use this screen as a fast decision aid; validate major decisions with detailed TCO modeling and transition-risk review.

Step 2: Measure whether flexibility was actually used

1. How many FTE, hours or units were contractually available to ramp down during the last 12 months?
2. How much capacity was actually reduced, for how long, and what cost was avoided?
3. How often was extra capacity added faster than an internal team could reasonably have responded?
4. Which provider-owned skills, tools, geographic coverage or risk transfer created measurable business value?
5. Would the same work, after standardization and automation, require materially less capacity today?

OPERATIVE SCREENING EXAMPLE

Using direct people-cost equivalents only: \$315K vendor baseline – \$67.5K internal baseline = \$247.5K/month gross people-cost premium, or 78.6%. Because infrastructure/shared-service costs and a quantified value for elasticity are not available, this is a screening result—not a completed TCO case.

10. How the Case Has Aged: 2018-2026

The historical case has aged well because the operating logic has become more—not less—relevant. Modern enterprises increasingly combine vendor ecosystems, GCCs, shared services, automation and AI. The boundary between “inside” and “outside” is now a portfolio decision that can move over time.



Figure 8: Historical case, modern validation and the evolving operating-model boundary.

Deloitte’s 2024 Global Outsourcing Survey found that 70% of executives had selectively insourced scope previously handled by third parties, 78% were using Global In-house Centers, and 80% planned to maintain or increase outsourcing. EY’s 2026 view of GCC operating models highlights the move toward hybrid models, while KPMG frames GBS as an integrated system spanning shared services, GCCs, outsourcing and AI-enabled support.[4][5][6]

AI-specific guidance for sourcing leaders

- Automate before sizing: estimate residual human effort after AI/automation, not before it.
- Reset commercials: if provider productivity improves, move from pure FTE pricing toward transaction, outcome or capacity-band models where feasible.
- Protect control points: retain human judgment for quality, exceptions, compliance and customer-impacting decisions.
- Measure realized productivity: separate demo-level AI potential from sustained production savings.
- Re-open sourcing boundaries: automation can make previously outsourced work small enough to own—or make internal work standardized enough to outsource.

11. When Outsourcing Is Still the Better Answer

The Operative case supports internalization because economics, utilization and capacity redesign all pointed in the same direction. It should not be read as an argument that vendor delivery is inherently inefficient. Outsourcing remains a strong answer when a provider supplies capability the enterprise would struggle to reproduce economically.

- Demand is genuinely volatile and contractual capacity can ramp up or down without large stranded-cost penalties.
- The provider owns scarce specialist skills, technology, market access or geographic reach that would be slow to build internally.
- Speed to launch matters more than long-run unit economics for the relevant business horizon.
- The work is standardized enough to be governed through measurable outcomes rather than loosely governed effort.
- The provider aggregates demand across clients and can achieve utilization or scale the enterprise cannot.
- Commercial incentives reward productivity and outcomes instead of preserving billable headcount.

WHEN THIS FRAMEWORK SHOULD **NOT** DRIVE A DECISION BY ITSELF

	Regulated / restricted work Legal, data-residency or licensing constraints may dominate economics.		Very small scope The management overhead of building a captive capability can outweigh savings.
	Short-lived demand If the work is temporary, internal fixed capacity can become the wrong risk.		Severe talent scarcity Provider access to skills may be worth a significant premium.
	Transformation in flight Do not freeze a sourcing decision before the future-state process is stable.		No reliable workload data If demand, productivity and rework are not measurable, the model will produce false precision.
 Key takeaway: Use this framework as an input, not the verdict. Pair it with context, judgment and stakeholder priorities to make the right sourcing decision.			

12. Governance: Make the Model Self-Correcting

Sourcing choices age. Demand changes, AI removes work, salaries move, vendors improve, strategic priorities shift, and capabilities become more or less critical. The operating model therefore needs a mechanism to challenge its own assumptions.

Make the sourcing model self-correcting

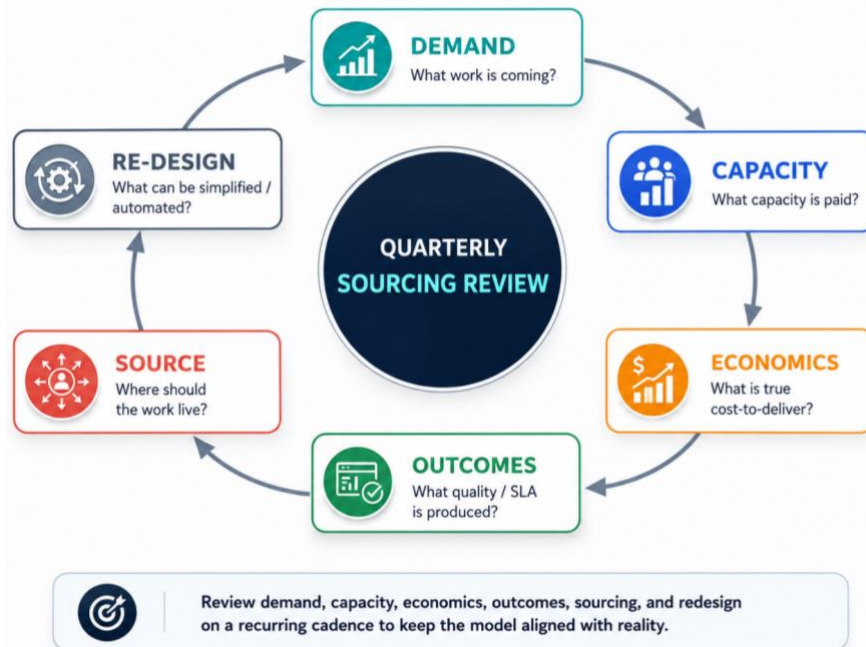


Figure 9: A recurring sourcing review connects demand, capacity, economics, outcomes, sourcing and redesign.

Suggested review cadence and triggers

Use a formal quarterly review for strategic sourcing portfolios, but trigger an out-of-cycle review when operating economics move materially. The thresholds below are practical defaults—not historical Operative evidence—and should be tuned to workload volatility and service criticality.

TRIGGERS FOR AN OUT-OF-CYCLE SOURCING REVIEW

— Early signals. Timely course-correction. —



Figure 10: Example triggers for an out-of-cycle sourcing review.

Decision output: Every review should end with one of five explicit outcomes—retain, renegotiate, automate, hybridize, or internalize. “Continue as is” should be a conscious decision supported by current economics.

13. Leadership Lessons


01

Price flexibility—and measure its use

A premium for elasticity is rational only if ramp-up, ramp-down, speed or risk transfer are actually used and valued.


02

Re-size before you re-source

Zero-base the workload, remove avoidable work, standardize and automate before deciding how many people—and whose people—are required.


03

Compare true cost-to-deliver, not rate cards

Include utilization, rework, governance, tools, transition, risk and productivity alongside labor rates.


04

Treat internalization as a capability decision

The durable case can include knowledge retention, control, data ownership, process velocity and easier automation—not just lower cost.


05

Make sourcing a living portfolio

Use explicit review triggers so work can move as demand, economics, technology and strategic value change.

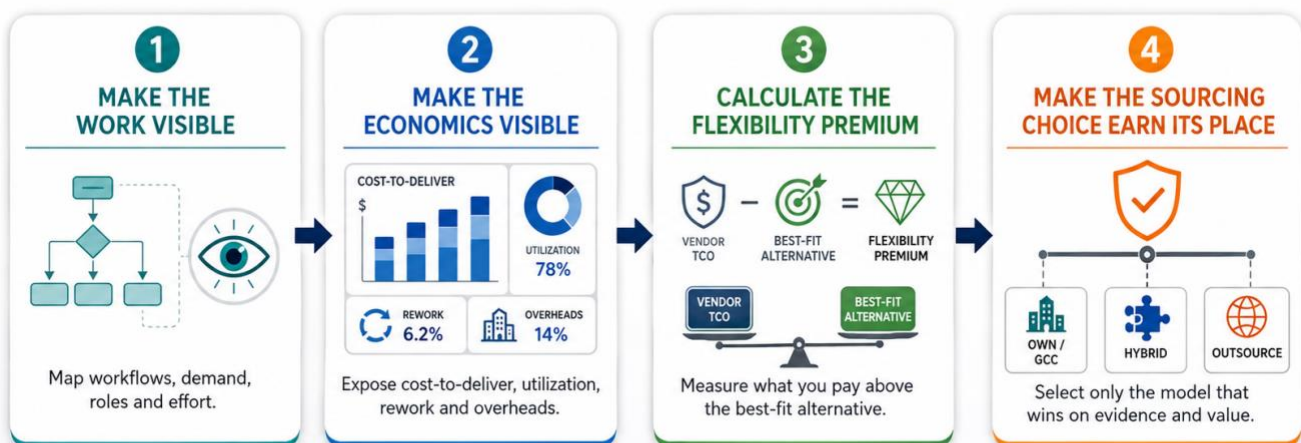
Conclusion: Measure the Premium, Then Choose the Model

Flexibility, control, capability and speed all have economic value, but none should be treated as free. A vendor premium is justified when it buys something the enterprise genuinely uses. An internal model is justified when ownership, utilization and scale produce better total economics and strategic capability. A hybrid model is justified when both needs coexist.

At Operative, paid vendor flexibility had become largely fixed capacity. Re-sizing the work before re-sourcing it created a much smaller internal capacity requirement and a substantial direct people-cost reset. In another operation, the same framework might support continued outsourcing.

THE DURABLE ASSET IS THE METHOD

A repeatable way to make sourcing decisions evidence-led, not assumption-led.



VISIBILITY → ECONOMICS → PREMIUM → CHOICE

The lasting advantage is not the answer once — it is the method used every time.

14. Appendix A: Executive Decision Checklist



Executive checklist: 14 questions before you make a sourcing decision

Use this as a practical decision screen before renewing, redesigning, insourcing, outsourcing, or shifting to a hybrid model.

1 DEMAND & CAPACITY 	
1	What percentage of paid capacity is productively utilized? ☐
2	How frequently did we use the flexibility we pay for? ☐
3	What caused unused capacity—contract terms, notice periods, forecast buffers, demand volatility, governance or provider constraints? ☐

2 PROCESS & REDESIGN 	
4	What proportion of effort is rework, coordination, duplicate control or avoidable handoff? ☐
5	What work can be eliminated, standardized or automated before capacity is sized? ☐
6	How many FTE are required after redesign—not before? ☐

3 SOURCING ECONOMICS & CONTROL 	
7	What is the total cost-to-deliver under vendor, GCC, in-house and hybrid scenarios? ☐
8	What scarce capability would a provider deliver more economically? ☐
9	Which knowledge, data, customer relationships or IP should remain under enterprise control? ☐
10	What is the transition risk, and what overlap is required to protect service? ☐

4 AI, GOVERNANCE & REVIEW 	
11	How should AI productivity change headcount, SLAs and commercial terms? ☐
12	What operating metric would trigger an out-of-cycle sourcing review? ☐
13	What is the quantified gross and net unrecovered flexibility premium? ☐
14	What decision will be re-tested at the next formal review date? ☐



Best practice: a sourcing decision should earn its place on economics, strategic value, control, resilience and measurable outcomes.

USE THE CHECKLIST AS A GATE

If demand, utilization, required capacity and total cost-to-deliver cannot be answered with credible data, do not make a binary outsource/insource decision yet. Improve measurement first.

15. Appendix B: Work Type Matrix Tool

Use one row per workflow, service line, queue or capability. Score each factor from 1 (low) to 5 (high), then use the sourcing matrix as a discussion tool—not as an automatic formula.

Work type	Demand volatility 1-5	Strategic criticality 1-5	Standardization 1-5	Internal capability 1-5	Provider advantage 1-5	Current model	Recommended test

Interpretation guide

Pattern	Starting model	Question to test
Stable + strategic	Own / GCC	Build depth, retain IP and decision rights.
Volatile + strategic	Hybrid	Own control points; buy elasticity or specialist capacity.
Stable + non-core	Standardize first	Automate and simplify, then compare TCO.
Volatile + non-core	Outsource	Exploit provider scale and cross-client utilization.

MATRIX RULE

Do not score the incumbent organization. Score the future-state work after zero-based redesign and automation. Otherwise the matrix will simply rationalize today's structure.

16. References & Evidence Notes

Operative case evidence

The case evidence supplied by Sudiip Ghosh relates to the 2018 operating exercise: 150 vendor FTE; approximately \$2,100 vendor cost per FTE per month; approximately 60% utilization; workflow/demand analysis indicating 60-75 internal FTE; approximately \$900 internal cost per FTE per month; transition completed in roughly four months. The \$315K versus \$67.5K comparison uses 75 internal FTE and represents direct people-cost equivalents only. Infrastructure and shared-service costs are excluded.

References

- [1] PERIOD SOURCE — Deloitte (2018). The Deloitte Global Outsourcing Survey 2018: Traditional outsourcing is dead. Long live disruptive outsourcing. [Source](#)
- [2] PERIOD SOURCE — NASSCOM (2018). 8th Global Capability Centers Conclave 2018: Leading the New Normal — Speed, Smart & Skill. [Source](#)
- [3] PERIOD SOURCE — EY India (2019; benchmarking data for 2018). EY India GCC Cost, Value and Operations Benchmarking Study 2018. [Source](#)
- [4] CURRENT PERSPECTIVE — Deloitte (2024). Global Outsourcing Survey 2024: Multidimensional sourcing — orchestrating the extended workforce ecosystem. [Source](#)
- [5] CURRENT PERSPECTIVE — EY India (2026). Why GCC operating models offer a new competitive advantage. [Source](#)
- [6] CURRENT PERSPECTIVE — KPMG (accessed September 2026). Insights into Global Business Services and Outsourcing. [Source](#)