

BEYOND FREE DELAY

The Business Case for Speed — What Two Years of Crisis Delivery Proved About Cost

On Delay, Queues, and the Unpriced Cost of Portfolio Governance

Giovanni Leonardi

April 2022



Contents

1. Executive Summary	3
2. The Quarter Where the Numbers Came Back	3
3. The Accidental Dataset	4
4. Pricing What Was Always Free	4
5. The Queue No One Is Paying For	5
6. What Stage-Gates Actually Cost	6
7. The Strongest Objection	7
8. Before the Amnesia Sets In	8

The Business Case for Speed — What Two Years of Crisis Delivery Proved About Cost

Portfolio · No. 89

Originally written April 2022 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

“When delay is free, the rational response is to deliberate. When delay is priced, the rational response is to decide.”

1. Executive Summary

For two years, organisations ran an uncontrolled experiment in what happens when portfolio governance is compressed to its minimum. The experiment was not designed — it was imposed by crisis — but it generated a dataset that portfolio management has never had: a sustained comparison of two operating models applied to substantially the same work. This essay examines that dataset through the lens of portfolio economics rather than governance design. Three findings recur: compressed decision cycles reduced total cost far more than they increased risk; the most expensive element in most portfolios was not delivery but the queue; and stage-gate friction, honestly costed, was among the largest line items. The window for capturing this evidence is closing as committees reconvene and pre-crisis processes are restored. The business case for speed is not a cultural argument — it is arithmetic.

2. The Quarter Where the Numbers Came Back

In March, the portfolio governance board met for the first time in its full pre-pandemic format. The standing agenda had been restored — quarterly reviews, gate approvals, resource allocation panels — and the tone was one of quiet satisfaction at the return to order. Buried in the pack, however, was a comparison slide that no one had commissioned. A portfolio analyst, preparing the standard velocity metrics, had run the numbers across two time horizons: the twenty-four months of crisis delivery, and the twenty-four months that preceded them. The comparison was not comfortable.

Cycle times had halved. Cost per initiative had dropped by roughly a fifth. The number of initiatives reaching benefit realisation within the financial year had nearly doubled. And the total risk exposure — the metric that the governance framework existed to minimise — had not materially changed. The slide was noted, not discussed, and the board moved to the next item.

That slide sits at the heart of a conversation the portfolio profession is not yet having in earnest. For two years, organisations ran a large-scale, uncontrolled experiment in what

happens when you strip governance friction from the delivery engine and force the portfolio to move at speed. The experiment was not designed; it was imposed by circumstance. But the data it generated is real, and it tells a story that is uncomfortable for anyone whose operating model depends on the assumption that deliberation is always worth its cost.

This is not an argument for recklessness. It is an argument for arithmetic.

3. The Accidental Dataset

What makes the 2020–2022 period unusual is not that organisations delivered faster — they have always been capable of bursts. It is that the burst lasted long enough to produce a meaningful dataset. Two years is enough to observe patterns, control for outliers, and distinguish structural effects from the noise of individual programme performance. The result is something portfolio management has never had: a large-scale comparison of two operating models — traditional governance and compressed governance — applied to substantially the same portfolios, in substantially the same organisations, over a substantial period.

The conditions were not identical, of course. Crisis delivery operated under emergency authorities, with elevated executive attention and a tolerance for imperfection in areas outside the critical path. These are legitimate caveats. But they do not invalidate the dataset. They mean it must be read carefully rather than dismissed.

When it is read carefully, three findings recur with uncomfortable consistency across sectors and portfolio types.

First, compressed decision cycles reduced total delivery cost far more than they increased rework or failure cost. The net saving was not marginal. In the portfolios I have seen analysed, the reduction in elapsed time translated to a reduction in total cost of between fifteen and twenty-five per cent — a figure driven largely by the elimination of carrying costs during decision delays.

Second, the single largest cost in most portfolios was not delivery. It was the queue. Initiatives waiting for approval, waiting for resources, waiting for a governance slot — these queues carried real financial cost, and under the traditional model they were invisible because no one was pricing them.

Third, stage-gate friction, when honestly costed — including preparation time, scheduling delay, review cycles, and rework arising from misaligned feedback — was among the most expensive line items in the portfolio. Not the most visible, but among the most expensive.

These are not cultural observations. They are financial ones. And the risk of the current moment is that they will be filed alongside the lessons about remote working and crisis leadership — acknowledged, admired, and quietly abandoned as the committees reassemble.

4. Pricing What Was Always Free

The concept of cost of delay is not new. Don Reinertsen formalised much of the mathematics over a decade ago, and lean practitioners have used it as a prioritisation tool for longer still. But in most portfolio contexts, cost of delay remains a theoretical instrument — acknowledged in training materials, absent from business cases.

The crisis period changed this, not through intellectual conversion but through operational necessity. When executives were forced to choose between initiatives on a weekly cadence rather than a quarterly one, they needed a way to compare urgency that went beyond RAG status and strategic alignment scores. Cost of delay — however roughly estimated — became a working tool.

Consider a composite but representative example. A financial services organisation carried a regulatory remediation programme in its portfolio alongside a digital channel enhancement. Under the traditional model, both sat behind the same stage-gate sequence: business case approval, architecture review, security assessment, delivery assurance, benefits review. Each gate added, on average, three and a half weeks of elapsed time — not because the reviews themselves took that long, but because of preparation, scheduling, document circulation, feedback cycles, and the rework that followed misaligned commentary.

The regulatory programme carried a cost of delay of approximately £380,000 per month — the sum of potential enforcement exposure, manual workaround costs, and the operational risk premium the compliance function was carrying. The digital channel carried a cost of delay of roughly £150,000 per month in deferred revenue capture. Neither figure appeared in the portfolio reporting.

Under crisis governance, the regulatory programme was fast-tracked through a single approval and delivered four months ahead of its original timeline. The saving — not the delivery cost saving, but the delay cost saving — was in the region of £1.5 million. The digital channel, reprioritised behind it, was delayed by two months, at a cost of £300,000. Net benefit of the reprioritisation: roughly £1.2 million. This was not a heroic intervention. It was basic arithmetic that the traditional governance model had made structurally impossible to perform, because the information needed to perform it — cost of delay, priced honestly — was not part of the reporting framework.

The lesson is not that every initiative should be fast-tracked. It is that the cost of not fast-tracking is real, measurable, and in most portfolios unmeasured. When delay is free, the rational response is to deliberate. When delay is priced, the rational response is to decide.

5. The Queue No One Is Paying For

If cost of delay is the metric portfolio governance has failed to adopt, the queue is the structure it has failed to see.

A mid-sized enterprise portfolio typically carries between fifty and seventy initiatives at various stages of lifecycle. Of these, perhaps fifteen to twenty are in active delivery at any given time. The remainder sit in some form of queue — awaiting prioritisation, awaiting funding approval, awaiting resource allocation, awaiting a governance slot. In a

well-managed portfolio, these queues are tracked. In most portfolios, they are simply the space between meetings.

The financial cost of a queue is not intuitive, which is why it is so rarely priced. An initiative sitting in a prioritisation queue is not consuming delivery resource. It is not, in the traditional accounting sense, costing anything. But it is carrying cost — the cost of the problem it was raised to solve going unsolved, the cost of the opportunity it was raised to capture going uncaptured, and the overhead cost of maintaining it as a line item in the portfolio: status updates, stakeholder management, re-baselining as assumptions shift.

Little's Law — the relationship between throughput, cycle time, and work in progress — gives us the arithmetic. If a portfolio's average cycle time is fourteen months and its throughput is sixteen initiatives per year, then by Little's Law it is carrying approximately eighteen to nineteen initiatives in progress at any time. Reduce cycle time to nine months and throughput rises to roughly twenty-five — without adding a single resource. The constraint was never capacity. It was flow.

During the crisis, portfolios did not add capacity. Most reduced it, as resources were redeployed to operational priorities. What they did instead was drain the queue. Ruthless prioritisation — forced by the impossibility of running everything simultaneously under emergency conditions — reduced work in progress, which reduced cycle time, which increased throughput. The arithmetic was mechanical, not heroic. But it required something the traditional model does not provide: a reason to say no.

Stage-gate governance provides many things, but a reason to say no is not among them. Gates assess whether an initiative is ready to proceed. They do not assess whether it should proceed given what else is in the queue. The result is a model that is exquisitely equipped to judge readiness and structurally blind to priority — which is why portfolios under traditional governance tend to accumulate work in progress without limit, and why cycle times drift upward in every organisation I have observed over any five-year period.

The most expensive decision in most portfolios is not a bad investment. It is the absence of a decision — an initiative that is neither killed nor accelerated but allowed to occupy a queue, consuming no delivery resource and generating no value, while the problem it was meant to address compounds.

6. What Stage-Gates Actually Cost

It would be unfair to lay the entire cost of portfolio friction at the door of stage-gate governance. Decision delay has many sources — organisational politics, resource contention, strategy ambiguity. But stage-gates are the one source that is entirely within the portfolio function's control, and they are the one source whose cost has never been honestly accounted.

The arithmetic is not complex. Take a typical enterprise initiative subject to five governance gates: concept approval, business case, solution design review, delivery assurance, and benefits review. Each gate requires preparation — the team assembles the documentation, aligns stakeholders, rehearses the narrative. It requires scheduling —

finding a slot in the board's calendar, which in most organisations runs monthly or bi-monthly. It requires the review itself, feedback and rework — gate conditions are rarely zero — and the elapsed time between conditional approval and the next activity starting. Across these components, a conservative estimate is three to four weeks per gate. Five gates multiply this to fifteen to twenty weeks — roughly four months of elapsed time consumed by governance process across the life of a single initiative.

Scale this to a portfolio of thirty-five active initiatives and the aggregate is formidable: between five hundred and seven hundred weeks of elapsed time, across the portfolio, consumed by gate processes in a single year. Not all of this is waste — some gates surface genuine issues that would otherwise reach production. But the question is not whether gates add value. It is whether the value they add exceeds their cost, and until the cost is priced, the question cannot be answered.

During the crisis, most organisations collapsed their gate structure from five stages to two — an initial triage and a delivery checkpoint. The predicted consequence was an increase in failed deliveries, integration defects, and benefit shortfalls. The observed consequence, in the portfolios I have examined, was none of these things. Failure rates did not materially change. What changed was the type of failure: fewer initiatives failed at gate reviews — because there were fewer gates to fail at — and marginally more encountered issues during delivery, issues that were in most cases resolved operationally rather than escalated through governance.

The implication is not that gates are worthless. It is that the relationship between the number of gates and the quality of outcomes is far weaker than the operating model assumes, and that the cost of the gates — in elapsed time, in queue effects, and in cost of delay — is far higher than the operating model acknowledges.

7. The Strongest Objection

The most serious counter-argument to this analysis is not that speed is dangerous — that is a truism, and truisms do not explain anything. The serious objection is that crisis conditions are non-representative because the tolerance for failure was structurally different. Under normal conditions, a failed initiative carries reputational cost, career cost, and audit cost. Under crisis conditions, these costs were suspended. People were forgiven for trying and failing in ways they would not be forgiven under business as usual. The apparent superiority of compressed governance, on this argument, is an artefact of a temporarily relaxed accountability regime.

This deserves a serious answer, because it is partly right. Accountability regimes did change during the crisis. But the change was not what this objection assumes. What was suspended was not the tolerance for failure — organisations did not suddenly become indifferent to whether things worked. What was suspended was the tolerance for delay. The quality bar, in every organisation I have observed, remained exactly where it was. The patience for waiting did not.

This distinction matters because it reframes the comparison. The traditional model does not prevent failure — failure rates under stage-gate governance are well-documented and not low. What the traditional model prevents is fast failure. It ensures that when an

initiative fails, it fails slowly, at high cost, after extensive deliberation, and with comprehensive documentation. Compressed governance allows initiatives to fail faster, at lower cost, with less documentation and more operational recovery. The total cost of failure — including the delay cost of the slow model — favours the compressed approach in every dataset I have seen.

The honest version of the objection is therefore not that speed increases failure, but that speed changes the distribution of failure in ways that organisations are not yet equipped to absorb. Fast failure requires fast learning, and fast learning requires feedback loops that most portfolio functions do not yet possess. This is a genuine constraint, and it is one worth solving. But it is a constraint on implementation, not a refutation of the economics.

8. Before the Amnesia Sets In

The window for this analysis is closing. As governance committees reconvene, as gate processes are reinstated, as the crisis-period data is overwritten by the next two years of traditional delivery metrics, the comparison will become harder to make and easier to dismiss. Already, in the organisations I work with, the narrative is shifting from "we proved we could move faster" to "we moved fast because we had to, and now we can return to doing things properly." The word properly is doing heavy lifting in that sentence. It assumes that the pre-crisis model was proper — that is, that its costs were justified by its benefits — without ever having tested the assumption.

The data from 2020–2022 does not prove that traditional portfolio governance is wrong. It proves that it is unpriced. And an unpriced operating model is one that cannot be rationally evaluated, defended, or improved.

What is needed is not a dismantling of governance but a pricing of it. Every gate, every queue, every decision delay carries a cost, and until that cost is visible in the portfolio reporting, the operating model will continue to optimise for the thing it can see — risk at each gate — at the expense of the thing it cannot: the cumulative cost of the time those gates consume.

The evidence exists, now, to build that pricing model. The crisis period generated the comparison data. The portfolio analytics required to price delay, queue cost, and gate friction are not exotic — they require the same information most portfolios already capture, combined differently. The obstacle is not technical. It is institutional. Pricing governance means subjecting governance to the same cost-benefit discipline that governance applies to everything else. There is a certain circularity to the resistance.

Two years of accidental data will not settle this question permanently. But they have settled it long enough to act. The business case for speed is not an abstraction, a cultural preference, or an agile manifesto. It is arithmetic. The numbers came back in March. The question is whether anyone will read them before the filing is complete.