

Who manages the variance? Operational concentration and the regulation of financial infrastructure



Patrick T. Harker | University of Pennsylvania

Keywords: Operational resilience, critical third parties, cloud concentration, core banking, systemic risk, macroprudential policy

JEL codes: G21, G28, D62, D81, L86

Abstract

In July 2026 the United Kingdom placed four cloud providers under direct supervisory oversight as its first designated critical third parties. Comparable regimes are in operation or under construction across the European Union and the United States. I argue that all of them act on the wrong half of the problem. Expected systemic damage from operational failure is governed by a mean and a variance. Markets already manage the mean, because every institution cares about its own uptime. Nobody manages the variance, which is set by who fails together, and concentration and cross-layer alignment are variance rather than mean. Five recommendations follow, each using authority supervisors already hold. I also take on the strongest objection directly: in core processing, no bank can realistically run two vendors. The answer is not to force dual sourcing. It is to move the instrument to market structure.

Disclaimer: The views expressed are those of the author and do not represent those of any institution with which he is, or has been, affiliated.

The event

On 19 July 2024, a faulty update from a single software vendor disabled some 8.5 million computers worldwide within hours, by Microsoft's own count. Hospitals lost patient records. Airlines grounded fleets. A substantial share of banking, payment and trading operations was impaired, and the cloud outage analytics firm Parametrix put direct losses to Fortune 500 companies alone at roughly \$5.4 billion.

The technical event was a kernel-level driver bug. The systemic event was something else. One vendor's update channel was simultaneously the failure mode of machines sitting inside thousands of nominally independent institutions, and no market-share statistic, no institution-level examination and no vendor-level reliability record would have flagged that exposure the day before.

Two years later, the regulation of this shared infrastructure stopped being hypothetical. On 10 July 2026 HM Treasury designated the first critical third parties under the United Kingdom's new regime, namely Amazon Web Services EMEA, Google Cloud EMEA, Microsoft Ireland and Oracle UK, and joint oversight by the Bank of England, the Prudential Regulation Authority and the Financial Conduct Authority began three days later. The European Union's Digital Operational Resilience Act, or DORA, has applied since January 2025. The Financial Stability Board has published its third-party risk toolkit, and in the United States the Treasury has documented the sector's cloud concentration while the banking agencies have issued interagency third-party guidance.

So the oversight architecture now exists. My argument is that the instruments it carries are aimed at the wrong half of the problem, and that the half they miss is the one that produced 19 July.

That argument rests on a formal model, developed and calibrated in a companion paper, in which institutions choose provider portfolios from markets with scale economies, providers fail through both idiosyncratic and common-cause channels, and correlated failure imposes losses on firms that never made the sourcing decision. None of that machinery appears below.

Markets manage the mean. Nobody manages the variance

Let the system-wide failure rate be the share of financial institutions operationally down at the same time in a given period. Damage to financial stability is convex in that share. Why? Because an event taking down 20% of the system at once does more than twice the damage of two events that each take down 10%. Simultaneous impairment forces correlated asset sales, saturates recovery capacity, congests the healthy institutions that would otherwise absorb the volume, and disrupts the markets in which the impaired institutions intermediate.

Take damage to be proportional to the square of the failure rate. That is the simplest convex form and the one calibrated in the companion paper, and it is conservative here, since squaring delivers exactly the factor of two while the saturation and congestion effects deliver the rest. Expected damage then obeys an identity out of elementary statistics:

$$\text{expected damage} \propto (\text{average failure rate})^2 + \text{variance of the failure rate}.$$

The identity sounds innocuous. It is the whole policy problem.

The mean is average downtime: how often a typical institution is impaired. Private incentives already govern that object. Every institution cares about its own uptime, providers compete on reliability and are bound to service-level agreements, and the platforms and processing vendors at the centre of the system are, by any historical standard, extremely reliable. Nothing here disputes that. No recommendation I make is aimed at the mean.

The variance is a different object. It measures whether the failures that do occur arrive scattered, a few institutions this quarter and a few more next year, or bunched, with long stretches of nothing and then an afternoon when a large share of the system is dark at once.

Here is the point on which the rest of this note rests. Holding each provider's reliability fixed, every structural feature I am concerned with moves the variance and leaves the mean alone. Concentration of institutions on few providers raises it, because the same expected downtime arrives in larger simultaneous blocks. Alignment of failure causes across layers of the technology stack, through shared operating systems, shared update channels or shared subcontractors, raises it as well, because outages that would have landed at different times now land together.¹ Diversification and dispersion of dependencies lower it.

How much is at stake in that distinction? Exhibit 1 holds every provider's reliability exactly fixed and changes nothing but the timing.

Exhibit 1. Same reliability, same average downtime, more damage

	Independent timing	One shared root cause
Both layers down together	16% of years	40% of years
Exactly one layer down	48% of years	never
Neither layer down	36% of years	60% of years
Average failure rate	8%	8%
Damage index	0.0112	0.0160

Note: Let each of two layers of the technology stack, say cloud and core software, carry a failure event with annual probability 0.4 that disables 10% of institutions. Nothing below changes either event's probability or its severity. Only the timing changes.

Average downtime is identical in the two columns. Expected damage is 43% higher in the second, from timing alone. The amplification grows with the number of aligned layers and with the rarity of the event: measured against the companion paper's fine-grained benchmark, in which the same failure mass arrives instead as many small independent events, a once-a-decade aligned event carries a factor near ten.

July 2024 was this arithmetic realised across institutions. A single update channel synchronised, in one afternoon, failures at thousands of firms that every market-share statistic treated as independent.

So the vocabulary of the live regimes, operational resilience, is necessary and incomplete. Resilience programmes, provider by provider and firm by firm, manage the mean, and by making each provider safer they scale part of the variance down with it. What they cannot reach is the part of the variance that is a property of the joint structure: who depends on whom, through which shared substrates, with how much overlap.

No individual institution can see that structure. No provider owns it. It is, in the most literal sense, unmanaged.

Why the market will not fix it

If concentrated, aligned infrastructure is expensive in expectation, why does the market choose it? Three parts of the answer matter for policy.

First, the damage is external. When a shared provider fails, the losses land overwhelmingly on parties other than the institution whose sourcing decision created the exposure: on counterparties, customers, markets and the public. In the companion paper's calibration, an individual institution internalises on the order of a fifth of the systemic loss its concentration choices help create, while a planner would internalise most of it. July 2024 showed the same wedge from

¹ With one condition, made exact in the companion paper: alignment amplifies insofar as the aligned event reaches different institutions through the different layers. Where the same institutions are exposed through both, part of the synchronisation lands on institutions already down, and the effect attenuates.

the vendor side, since direct losses ran to billions while standard vendor terms cap liability at roughly the fees paid. The losses stayed where they fell. When most of the cost of a choice lands somewhere else, the choice gets made too often.

Second, scale makes concentration privately rational. The largest providers are cheaper, better integrated and often individually more reliable. Choosing the market leader is the right decision for almost every institution almost all of the time, which is exactly how a system ends up with a variance problem no participant chose. Industry surveys, including the Cloud Security Alliance's State of Financial Services in Cloud reports and the supervisory analyses collected in the US Treasury's 2023 report on cloud adoption, put the top three providers above 80% of regulated US institutions as primary or secondary platforms. On the companion paper's calibration anchors, that is equivalent to fewer than four equally sized providers serving the entire sector. Four is also, and not by coincidence, the number of cloud firms the United Kingdom has just designated.

Third, concentration sustains itself. Scale economies mean that once a market has tipped towards one or two providers, moving back is prohibitively expensive for any individual institution, because the small competitor is expensive precisely by virtue of being small. The companion paper shows that concentrated configurations, once reached, are stable for all parameter values. To be clear about what that implies: the market does not drift back towards diversity on its own, and an instrument that merely re-prices concentration at the margin can leave the system exactly where it stands. Prevention is cheaper than cure, and that is an equilibrium property rather than a slogan.

The instruments, ranked

Suppose a supervisor accepts the diagnosis. What should it do?

The standard toolkit offers two families. *Price instruments* charge for concentration, through a capital surcharge on concentrated exposures or a levy on the use of dominant providers. *Quantity instruments* constrain structure directly, through a cap on any provider's share of the sector or a floor on the number of providers that must be used. With perfect knowledge the two are equivalent. Under uncertainty they are not, and the classic price-versus-quantity analysis of Martin Weitzman (1974) settles which one to prefer: when the marginal damage curve is steep and the marginal cost of compliance is flat, quantity instruments win, because a mispriced charge lets the system slide a long way down a steep damage curve, while a quantity rule pins the outcome and lets the flat compliance cost absorb the error.

That is the configuration here. The damage side is steep, because near the concentrated region the system sits close to the tipping arithmetic of Exhibit 1, where small increases in concentration or alignment produce large increases in expected damage. The cost side is comparatively flat wherever diversification is feasible at all, since the incremental cost of a second provider is integration and operating overhead, roughly constant per provider added.

The calibration sharpens the point. The charge that would decentralise the planner's allocation runs to about a two-thirds surcharge on the price of concentrated provision, and its correct level moves several-fold with the convexity of systemic damage and the breadth of a common-cause event, two parameters no supervisor can estimate today. That is the textbook setting in which one should stop trying to price an externality and constrain the quantity instead. A regime built on supervisory expectations and firm-level risk pricing is reaching for the weaker tool.

I am not alone in the instinct, though the argument for it has been intuitive rather than formal. Hilary Allen has proposed ex ante limits on reliance on shared infrastructure, and Peter Sands, Gordon Liao and Yueran Ma have documented that regulators of other industries carrying large external operational costs rely on standards and quantity restrictions rather than priced capital. What the model supplies is the condition under which that instinct is welfare-improving.

Five recommendations follow. Each names an owner, and each uses authority that supervisors already hold.

1. **Target system-wide concentration, not firm-level diversity alone.** *Owner: the designating authority in each jurisdiction.* The externality is a function of the sector's structure, not of any one firm's portfolio. A cap on a provider's system-wide share within a category, equivalently a floor on the effective number of providers serving the sector, acts on the variance directly. Firm-by-firm multi-sourcing requirements without a system-level constraint close only part of the gap.
2. **Prefer quantity instruments and reserve price instruments for the margin.** *Owner: the same authority.* For the Weitzman reasons above.
3. **Calibrate the instrument to each provider market.** *Owner: the same authority, market by market.* The same diversity floor that lowers concentration where entry is elastic can raise it where fixed costs push a marginal incumbent out and the survivors absorb its share. Cloud, core processing, payments and data markets have different entry structures and need different instruments, which is the subject of the next section.
4. **Mandate dependency disclosure and build the system-level map.** *Owner: the supervisors of designated providers, using existing examination authority.* The first-order driver of the multi-layer externality is alignment: shared operating systems, update channels, subcontractors, data centres, and the overlap of institutional exposures. No market-share statistic reveals any of it and no single institution can measure it. Providers should disclose architectural dependencies to supervisors, and supervisors should aggregate the map and publish enough of it that institutions can diversify intelligently. This is the highest-value and lowest-cost item on the list.
5. **Pair any concentration instrument with reliability standards.** *Owner: the existing designation regimes.* A provider's optimal reliability investment rises with its scale, so an instrument that cuts a dominant provider's demand erodes its reliability spending unless a standard holds the floor. The converse matters more today: reliability standards on their own, which is the current design, manage the mean and leave concentration and alignment untouched.

Is any of this scoreable? It has to be, or it is a slogan. The measurable objects are the sector's concentration within each infrastructure category and the overlap of failure causes across categories, and neither is published anywhere today. Build the map in Recommendation 4 and both become series a supervisor can track quarterly. Here is what would count as failure: if, five years after those series exist, sector concentration has not fallen and the incident record shows no reduction in the share of outages reaching multiple institutions at once, then the argument I am making here is wrong on its own terms.

The hard case: when running two vendors is not realistic

Every experienced banker will meet those recommendations with the same objection, and it deserves to be answered in the text rather than in a footnote. Running vendors in parallel is itself costly and itself risky. It means duplicate integration, duplicate contracts and audits, duplicated staff expertise, reconciliation between systems, and more surfaces on which a change can go wrong.

In some markets that cost is manageable. In at least one, core processing, it is prohibitive. The core is the ledger, the single source of truth for deposits and loans, and an institution running two cores is for practical purposes running two banks. Almost no bank does it. Where institutions do operate multiple cores it is nearly always merger legacy, accidental complexity they are paying to eliminate rather than resilience they chose. Core conversions are so costly and so hazardous that a given institution attempts one perhaps once in a generation, and the market's shorthand, that you marry your core, is only half a joke.

A diversity floor of two core processors per bank would be somewhere between useless and harmful. Any framework that appears to recommend it discredits itself with the audience that matters.

So does the argument collapse in that market? No, and the reason is worth stating carefully, because it converts the objection into guidance. In the model, the cost of running an additional vendor is a parameter, and the analysis delivers a boundary result: where that cost is prohibitive, the planner does not dual-source either. Firm-level redundancy in the core market is not an efficient outcome being blocked by an externality. It is an inefficient outcome for everyone, market and planner alike.

The externality in that market is nonetheless real. The top four vendors are estimated to serve more than 70% of US community and regional banks, and a common-cause event at any one of them is a sector event. What has to change is the instrument, because the portfolio margin is closed. Four instruments remain open.

First, defend the system-level structure that already exists. Exhibit 2 carries the arithmetic that makes “you cannot dual-source” compatible with “concentration still matters.” Suppose every institution single-sources its core, as they do, and consider a sector event catching each vendor independently. Expected damage conditional on that event rises with the concentration of the vendor market, even though no bank holds more than one vendor. Concentration policy in the core market is therefore not a question about bank portfolios at all. It is a question about the number and size distribution of vendors, which puts merger review in the front line. The core market reached its present structure through consolidation, each further merger raises the sector’s damage index mechanically, and that systemic cost belongs in the merger analysis alongside the competition analysis, every time.

Exhibit 2. Concentration matters even when every bank single-sources

Structure of the vendor market	Herfindahl index	Damage index per event
Four equal vendors, 25% each	0.25	0.3125
One vendor at 70%, three at 10%	0.52	0.3800
Two equal vendors, 50% each	0.50	0.3750

Note: Every institution uses exactly one core vendor. A sector event catches each vendor independently with probability one half. Average downtime in such an event is then 50% of institutions no matter how the market is structured. Expected damage is not. Concentration is measured by the Herfindahl index, the sum of squared market shares.

Moving from four equal vendors to a market with one dominant vendor raises expected damage per event by about a fifth, with every bank still single-sourced and every vendor exactly as reliable as before. Consolidation from four vendors towards two moves the index the same way. This is the formal reason merger policy is a systemic-risk instrument in markets where portfolio diversification is closed off.

Second, separate the substrates. Exhibit 2 assumes vendors fail independently. What if they do not? Suppose the four core vendors share a cloud platform, an authentication backbone, a software supply chain or a common subcontractor. Then the effective number of vendors is smaller than the nominal count and the sector’s true Herfindahl is higher than its market shares suggest. This is alignment again, one layer down. Sub-dependency disclosure and, where warranted, required separation, meaning no single substrate sitting underneath all the majors, raise the effective vendor count without asking any bank to run two cores. The dependency map in Recommendation 4 is what makes this instrument operable.

Third, make exit real even where switching is rare. The resilience value of a four-vendor market depends on institutions being able to move after a failure, even if they never move before one. Escrowed data in standardised formats, tested conversion runbooks, contractual portability rights, and vendor living wills, meaning resolution-style plans for the orderly transfer of a failed vendor’s client base to survivors, on the model of bridge-bank arrangements, convert nominal alternatives into real ones. None of this requires dual operation. All of it caps the tail of a vendor failure.

Fourth, and this is the long game, treat the cost of diversification as a policy variable rather than a constant. The integration cost that closes the portfolio margin in core processing is not a law of nature. It is the product of proprietary data formats, closed interfaces and decades of accumulated customisation. Open-API mandates, standardised data

schemas and interoperability requirements lower that cost structurally, slowly but in a directed way, and every reduction converts a little more of the market from the hard case into the tractable one. The United Kingdom's open-banking programme, which began with the Competition and Markets Authority's 2017 order, shows that interface mandates can move entrenched infrastructure inside a decade.

The lesson of the hard case is that these recommendations are feasibility-ranked by market. Cloud sits at the tractable end, where multi-cloud operation is costly but practised today for critical workloads, and where partial redundancy, meaning portable workloads and recovery capacity held on a second platform, captures much of the variance reduction at a fraction of the cost of full duplication. There the diversity floor applies as written. Payments already run multi-rail. Core processing sits at the far end, where the binding instruments are structural: merger discipline, substrate separation, exit machinery, and standards that lower the cost of switching over time. A regime applying one uniform rule across these markets would be wrong in both directions at once.

What the live regimes do, and do not do

So how do the live regimes score against this framework? They are a real start, aimed at the wrong moment of the problem. The UK designation regime, DORA's oversight of critical ICT providers, and the FSB toolkit share a design: identify the providers that matter, subject them to resilience standards, testing and incident-reporting obligations, and strengthen firm-level third-party risk management. In the language of this note those are mean instruments and reliability instruments. They are valuable, Recommendation 5 requires them, and the designation lists are themselves the seed of the system-level map in Recommendation 4.

But no live regime yet carries an instrument acting on the variance. None constrains system-wide concentration. None measures cross-layer alignment, though DORA's registers of information and the FSB toolkit now collect the subcontracting-chain data on which a mapping function could be built. None acts on the sub-provider substrates that data describes. And none engages the merger margin in the markets where structure is the only instrument available. The architecture exists. The quantity instruments it should carry do not. That, rather than another round of resilience standards, is where the next increment of financial-stability value sits.

For the United States, three near-term steps follow, and none of them requires new legislation. The interagency examination of significant service providers under the Bank Service Company Act should be extended into a genuine mapping function, with dependency disclosure from designated providers, aggregation across agencies, and publication of the system-level picture to the institutions that have to diversify against it. Merger review involving core processing and comparable high-switching-cost infrastructure should carry a systemic operational-risk analysis built on the logic of Exhibit 2. And the Financial Stability Oversight Council's third-party risk recommendations, a standing item in its annual reports for over a decade, should graduate from monitoring to a named instrument: a system-wide concentration standard for each infrastructure category, with the diversity margin applied only where it is feasible.

What we know, and what we do not

How much of this should a supervisor believe? Honesty about limits is part of the case. The quantitative statements here, the two-thirds surcharge, the factor-of-two-to-ten alignment amplification, and the planner's two-and-a-half-provider target in cloud, come from a calibrated model rather than an estimated one. The calibration is disciplined by the incident record and by observed market structure, but the deep parameters, the convexity of systemic damage and the share of losses institutions internalise, are chosen for plausibility.

Two conclusions survive that admission, because they turn on signs and shapes rather than levels. The market under-supplies diversity and over-supplies alignment. Quantity instruments dominate price instruments in this setting. The specific numbers should discipline intuition. They should not appear in a rulebook.

The empirical programme that would harden them, estimating the externality from the outage record and from supervisory data, is the natural next step, and the dependency map in Recommendation 4 is, not incidentally, the data set that programme requires. That is the deeper argument for building the map first. Every other instrument, and every future estimate, depends on it.

On the morning of 19 July 2024, every institution running that software was individually well managed. Each had a resilience programme, a vendor policy, and a provider with an excellent reliability record. None of it mattered, because the thing that failed was not inside any of them. It was the structure they shared, and nobody owned it.

The first designations have now been made and the oversight architecture is built. What it lacks is the recognition that its object is a variance rather than a mean, and the small set of structural instruments, ranked by feasibility across markets, that I have set out here. None of them requires a bank to run two cores. All of them require someone, finally, to manage the variance.

References

- Allen HJ (2024) Reinventing operational risk regulation for a world of climate change, cyberattacks, and tech glitches. *Journal of Corporation Law* 49(4):727–785.
- Amromin G et al. (2025) Technology providers and financial stability: Overview of risks and regulatory frameworks. Working Paper 2524, Federal Reserve Bank of Dallas.
- Bank of England (2026) UK financial regulators to begin overseeing critical third parties announced by HM Treasury. News release, 13 July.
- European Union (2022) Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA). *Official Journal of the European Union* L 333/1.
- Financial Stability Board (2023) *Enhancing Third-Party Risk Management and Oversight: A Toolkit for Financial Institutions and Financial Authorities*. Final report, December.
- Harker PT (2026a) Critical infrastructure concentration and systemic financial fragility. Working paper, The Wharton School, University of Pennsylvania. <https://doi.org/10.2139/ssrn.7282881>.
- Harker PT (2026b) Critical infrastructure concentration and systemic financial fragility: General model and proofs. Technical companion note, The Wharton School, University of Pennsylvania. <https://doi.org/10.2139/ssrn.7282120>.
- HM Treasury (2026) The Critical Third Parties (Designation) Regulations 2026. Statutory instrument, 10 July.
- Kotidis A, Schreft SL (2022) Cyberattacks and financial stability: Evidence from a natural experiment. Finance and Economics Discussion Series 2022-025, Board of Governors of the Federal Reserve System.
- Microsoft (2024) Helping our customers through the CrowdStrike outage. Microsoft Official Blog, 20 July.
- Parametrix (2024) The CrowdStrike outage: Impact assessment for Fortune 500 companies. Research report, July.
- Sands P, Liao G, Ma Y (2018) Rethinking operational risk capital requirements. *Journal of Financial Regulation* 4(1):1–34.
- US Department of the Treasury (2023) *The Financial Services Sector's Adoption of Cloud Services*. Report, February.
- Weitzman ML (1974) Prices vs. quantities. *Review of Economic Studies* 41(4):477–491.

About the author

Patrick T. Harker is the Rowan Distinguished Professor at The Wharton School of the University of Pennsylvania, Director of Academic Engagement for Penn Washington, and Co-Academic Director of the Wharton Future of Finance Initiative. From 2015 to 2025 he was the eleventh President and Chief Executive Officer of the Federal Reserve Bank of Philadelphia and a member of the Federal Open Market Committee. He previously served as President of the University of Delaware and as Dean of the Wharton School. He has authored or co-authored more than 100 scholarly articles on financial institutions, risk and operations, is an elected Fellow of INFORMS and of the National Academy of Inventors, and is a former Editor-in-Chief of Operations Research. He holds BS, MS and PhD degrees in engineering and an MA in economics, all from the University of Pennsylvania.

SUERF Policy Notes and Briefs disseminate SUERF Members' economic research, policy-oriented analyses, and views. They analyze relevant developments, address challenges and propose solutions to current monetary, financial and macroeconomic themes. The style is analytical yet non-technical, facilitating interaction and the exchange of ideas between researchers, policy makers and financial practitioners.

SUERF Policy Notes and Briefs are accessible to the public free of charge at <https://www.suerf.org/publications/suerf-policy-notes-and-briefs/>

The views expressed are those of the authors and not necessarily those of the institutions the authors are affiliated with.

© SUERF – The European Money and Finance Forum. Reproduction or translation for educational and non-commercial purposes is permitted provided that the source is acknowledged.

Editorial Board: Ernest Gnan, David T. Llewellyn, Donato Masciandaro

Designed by the Information Management and Services Division of the Oesterreichische Nationalbank (OeNB)

SUERF Secretariat

c/o OeNB, Otto-Wagner-Platz 3A-1090 Vienna, Austria

Phone: +43 1 40 420 7206 E-mail: suerf@oenb.at Website: <https://www.suerf.org/>

ISSN 2791-5395 (Online)