COMPOUNDING ENERGY LTD

Working Paper Series • CE-WP-2026-05

The Cannibalisation of Flexibility

Data-Centre Demand Response and the Equilibrium Value of Grid Flexibility in Great Britain

Dr Christopher T. M. Clack
Founder, Compounding Energy Ltd
christopher@compoundingenergy.com

September 2026

Abstract. Flexibility from large computing loads is being capitalised as a revenue stream and as a currency for accelerated grid connection. The valuations underpinning this capitalisation are, almost without exception, *price-taking*: each flexible megawatt is valued against today's price shape, today's ancillary prices, and today's connection queue, as though its own deployment at scale leaves those quantities unchanged. This paper extends the CECadence equilibrium framework (Clack, 2026a) — in which storage fleet build-out, wholesale prices, and dispatch are solved jointly to a fixed point — to include flexible demand cohorts representing data-centre load. We quantify three effects for Great Britain to 2030 and 2035: (i) the erosion of the marginal value of flexibility as flexible penetration grows, with the value of the marginal flexible megawatt falling by 62% between 2 GW and 10 GW of flexible data-centre load in the 2030 background (57% in 2035); (ii) the overstatement embedded in price-taking valuations, which exceed equilibrium values by a factor of 1.3–1.7 at the penetrations now being discussed publicly; and (iii) the cross-cannibalisation of battery storage revenues, with GB BESS wholesale-plus-frequency-response revenues falling by 21% at 10 GW of flexible data-centre load (2030 background), an unpriced risk factor in current storage financings. We discuss implications for the underwriting of flexibility revenues, for battery revenue cases that presently assume inflexible demand, and for connection-reform mechanisms that remunerate flexibility with queue position. All results are reproducible from a stated (scenario, code version, timestamp) triple, with *stylised* and *validated-on-settlement-data* flags recorded per run; the case-study bundle is available on request.

Keywords: data-centre demand response, flexibility, cannibalisation, equilibrium value, battery storage, Great Britain

Contents

1 Introduction

The growth of artificial-intelligence computing has made data-centre electricity demand the most discussed source of load growth in advanced power systems, and has produced a correspondingly rapid market in *flexibility*: the proposition that data centres can curtail or shift computation in response to grid conditions, and that this responsiveness can be monetised — as avoided peak cost, as ancillary revenue, and, most consequentially, as a currency exchanged for earlier and larger grid connections. The proposition has moved from demonstration to capital formation quickly. Software that orchestrates computational workloads against grid signals has been demonstrated at utility scale in the United States (Electric Power Research Institute, 2026), deployed in partnership with major power producers (NVIDIA Corporation, 2026; Colangelo et al., 2025), trialled in Great Britain with National Grid (Data Centre Dynamics, 2026), and, in August 2026, financed at a valuation exceeding one billion dollars (Emerald AI, 2026a). Public statements accompanying this capital formation describe flexibility as capable of unlocking on the order of one hundred gigawatts of additional data-centre capacity on the existing United States grid, echoing the 76–126 GW of integrable new load estimated by Norris et al. (2025). Market references throughout are as of September 2026.

This paper does not dispute that flexible operation of large computing loads is technically real, nor that it is valuable. It asks a narrower question with large financial consequences: *what is a megawatt of data-centre flexibility worth when many gigawatts of data-centre flexibility exist?*

The question matters because every value stream that flexibility monetises is a congestion or scarcity rent, and rents of this kind are self-eroding. Flexible load that avoids the highest-priced settlement periods flattens the price shape from which peak-avoidance value derives. Flexible load that stacks into response and reserve products competes in markets whose depth is measured in hundreds of megawatts against a flexible resource measured in gigawatts. Flexible load that trades responsiveness for queue position dilutes the value of that acceleration as flexible connection becomes the norm rather than the exception. The structure of the problem is identical to one already quantified for battery storage: in Clack (2026a) we showed that valuing GB battery revenues against exogenous price forecasts — ignoring the price impact of the storage fleet itself — overstates 2030 revenues by 100–300%+ across a representative 4–8 GW fleet band, rising further at the fleet sizes in national planning scenarios, and that the bias is removed only by solving fleet build-out, prices, and dispatch *jointly* to a fixed point. The declining value of storage with penetration is well established in the academic literature (de Sisternes et al., 2016); the contribution of the equilibrium framework is to make the effect bankable — quantified at half-hourly resolution, graded in public against a measured, open-data index, and reproducible run by run. Demand-side flexibility has been recognised as both valuable and difficult to remunerate for decades (Strbac, 2008); what is new is the arrival of a single load class large enough, and controllable enough, to move the equilibrium by itself.

This paper makes four contributions. First, it formulates data-centre flexibility as a set of demand-side cohorts — constrained, service-level-bounded virtual batteries — inside the CECadence equilibrium framework, so that flexible demand, storage dispatch, and price formation are solved jointly (Section 3). Second, it quantifies the equilibrium value of GB data-centre flexibility by value stream and by penetration, and the ratio by which price-taking valuation overstates it (Section 5). Third, it quantifies the cross-effect on battery storage revenues — we are not aware of a prior published estimate of what flexible data-centre load does to the revenue case of the GB BESS fleet.[1] Fourth, it draws the implications for underwriting, for storage financing, and for connection-reform design (Section 6).

Great Britain is the right first market for this analysis for three reasons: the storage equilibrium machinery is already calibrated and publicly graded there (Clack, 2026a); the connections-

[1] Search of Google Scholar, arXiv, SSRN and the grey literature of NESO, Ofgem, EPRI, LBNL and the Duke Nicholas Institute for flexible data-centre load or data-centre demand response combined with battery storage revenue for Great Britain, conducted August 2026; no prior quantification was found.

reform programme and the siting of AI growth infrastructure make the flexibility-for-connection exchange a live policy design question rather than a hypothetical; and a GB demonstration of computational flexibility is already under way (Data Centre Dynamics, 2026).

2 The economics of flexibility rents

A flexible data centre in Great Britain can monetise responsiveness through five channels. Each is examined here for the property that governs its behaviour at scale: the depth of the rent pool relative to the size of the flexible resource.

Wholesale price shaping. Avoiding consumption in the highest-priced settlement periods and recovering it in cheaper ones. The rent derives from the steepness of the price-duration curve. It is the deepest of the five pools, but it is precisely the pool that flexible operation at scale flattens: the value is a decreasing function of the aggregate flexible response occupying the same periods. This is the demand-side mirror of storage arbitrage, and it competes for the same spread.

Balancing and ancillary services. Frequency response and reserve products procured by the system operator. These markets are shallow relative to the resource: product requirements are measured in hundreds of megawatts (National Energy System Operator, 2026b), against a prospective flexible data-centre resource measured in gigawatts. Saturation therefore arrives first here, as it did for battery storage — whose ancillary revenue share collapsed as fleet capacity overtook product depth, quantified across the seven-stream stack in Clack (2026b).

Capacity market participation. Demand-side response is eligible GB capacity (Ofgem, 2026); flexible load can earn capacity payments for committed turn-down. The rent pool is set by the capacity auction clearing price, which additional reliable flexibility itself suppresses.

Network charging. Post-reform GB network charging leaves limited scope for peak-avoidance of residual charges (Ofgem, 2019); forward-looking and locational elements retain some value. We treat this channel conservatively and report it as a distinct row.

Connection acceleration. The most novel channel, and the one with the least developed price. Under connections reform, and in emerging flexible-connection arrangements, a load that accepts curtailability or offers dispatchable flexibility may connect earlier or larger than an inflexible load. The private value of acceleration is enormous for AI infrastructure — it is the binding constraint on deployment — but the *exchange rate* between flexibility and queue position is a policy instrument, not a market price, and its value dilutes as flexibility becomes a standard feature of connection applications rather than a distinguishing one. Section 6 argues that setting this exchange rate without an equilibrium valuation of the flexibility being purchased risks systematic over-payment in queue position.

The common structure across all five channels is that the rent is a function of system conditions that the flexible fleet, at scale, itself changes. Any valuation that holds those conditions fixed while scaling the fleet is internally inconsistent.

One further property sharpens the timeline. The orchestration mechanism is not proprietary in any lasting sense: the architecture and field results are published (Colangelo et al., 2025), and power flexibility is being integrated into the reference designs of the dominant AI-factory vendor, with a first deployment announced in 2026 (Emerald AI, 2026b). If it becomes a default rather than an option, capability diffuses on the vendor's release schedule rather than an innovator's sales cycle — and the erosion quantified in this paper arrives correspondingly sooner. The remainder of the paper makes that inconsistency, and its timing, quantitative.

3 Model

3.1 The CECadence equilibrium framework

CECadence solves GB storage fleet build-out, half-hourly wholesale prices, and dispatch jointly to a fixed point (Clack, 2026a), with the full seven-stream GB revenue stack co-optimised as in Clack (2026b). Let $\mathcal{T}$ index half-hourly settlement periods over the model horizon, $\mathcal{G}$ the thermal and renewable generation stack, and $\mathcal{B}$ the battery cohorts, characterised by power rating, duration (1 h, 2 h, 4 h), round-trip efficiency, and degradation limits. The dispatch layer commits and dispatches the GB stack with interconnectors against demand D_t, producing prices $\boldsymbol{p} = (p_t)_{t\in\mathcal{T}}$ from the marginal system balance; the framework's calibration is graded in public against a measured, open-data record — the CECadence Quarterly GB BESS Index (Compounding Energy Ltd, 2026) and a committed 2022–26 backtest in which realised annual-mean prices fall within $\pm 5\%$ in three of five graded windows, with the pre-registered held-out year published exactly as it landed. Storage cohorts respond optimally to prices; the fleet–price–dispatch loop is iterated to convergence. Standard verification applies to every converged solution: system balance in every period, state-of-charge bounds, flow limits, and complementary slackness between prices and binding constraints. The frequency-response clearing curves follow the stylised, hand-calibrated treatment of Clack (2026b), collapsed to a single aggregate response product in the reduced-form case study (Appendix A); ancillary values in Table 1 inherit that stylised status and are badged accordingly.

3.2 Flexible demand cohorts

We introduce flexible-demand cohorts $d \in \mathcal{D}$, each representing a class of data-centre load with a baseline profile $\ell_{d,t}$ and two modes of response.

Deferral (shifting). A fraction of computation can be delayed and recovered later, with an energy overhead. Let $u_{d,t} \geq 0$ be load deferred in period t, $r_{d,t} \geq 0$ recovery, and $s_{d,t}$ the backlog of deferred energy:

$$s_{d,t+1} = s_{d,t} + \Delta t \left(u_{d,t} - \eta_d\, r_{d,t}\right), \qquad 0 \leq s_{d,t} \leq \bar{s}_d, \tag{1}$$

$$0 \leq u_{d,t} \leq \delta^u_d\, \ell_{d,t}, \qquad 0 \leq r_{d,t} \leq \min\{\bar{r}_d,\ \bar{\ell}_d - \ell_{d,t} + u_{d,t} + c_{d,t}\}, \tag{2}$$

where $\Delta t = 0.5$ h is the settlement-period length (so $s_{d,t}$ is an energy backlog in MWh), δ^u_d is the deferrable depth, $\bar{\ell}_d$ denotes the cohort's connected capacity, $\bar{s}_d = \tau_d \cdot \delta^u_d \bar{\ell}_d$ caps the backlog at τ_d hours of deferrable load, $\bar{r}_d$ limits recovery ramping, and $\eta_d \leq 1$ encodes the energy overhead of deferred computation ($1/\eta_d \geq 1$ units recovered per unit deferred). The second bound in (2) is the connection-headroom constraint — equivalently $\tilde{\ell}_{d,t} \leq \bar{\ell}_d$ in the notation of (4): recovery cannot lift the site above its connected capacity, and at the load factors typical of data centres it, rather than $\bar{r}_d$, is what binds in practice. A payback horizon w_d requires $s_{d,t} = 0$ at the end of each non-overlapping (tumbling) window, that is $s_{d,t} = 0$ for $t \in \{w_d, 2w_d, 3w_d, \dots\}$, encoding service-level commitments on job completion.

Curtailment. A fraction of computation can be dropped or degraded without recovery, subject to service-level caps:

$$0 \leq c_{d,t} \leq \delta^c_d\, \ell_{d,t}, \qquad \sum_t \Delta t\, c_{d,t} \leq E^c_d, \qquad \sum_{t'=t}^{t+H_d-1} \Delta t\, c_{d,t'} \leq \bar{C}_d \quad \forall t, \tag{3}$$

where the annual energy cap E^c_d and the rolling-sum bound $\bar{C}_d$ over any window of H_d periods limit sustained curtailment consistent with checkpoint-and-drop operation. The rolling-sum form is deliberate: it bounds event energy and duration while keeping the feasible set a polytope, where a literal minimum/maximum-run constraint would require binary indicators and break the convexity on which Section 3.3 relies.

The effective load of cohort d is

$$\tilde{\ell}_{d,t} \;=\; \ell_{d,t} - u_{d,t} - c_{d,t} + r_{d,t}, \tag{4}$$

and system demand becomes $\tilde{D}_t = \big(D_t - \sum_{d\in\mathcal{D}} \ell^{\text{emb}}_{d,t}\big) + \sum_{d\in\mathcal{D}} \tilde{\ell}_{d,t}$, where $\ell^{\text{emb}}_{d,t}$ is the inflexible data-centre demand otherwise embedded in the background scenario, netted off at its mean level before the cohorts are added so that no data-centre megawatt is double-counted and only the load's shape and its flexible response enter residual demand — an accommodation assumption that is load-bearing, with its consequences quantified in the limitations of Section 6. Each cohort responds optimally to prices: cohort d solves

$$\max_{u,c,r\in\mathcal{F}_d} \;\sum_{t\in\mathcal{T}} \Delta t\, p_t\,(u_{d,t} + c_{d,t} - r_{d,t}) \;-\; \phi_d \sum_{t\in\mathcal{T}} \Delta t\, c_{d,t} \;+\; A_d(u,c,r), \tag{5}$$

where $\mathcal{F}_d$ is the feasible set defined by (1)–(3), ϕ_d is the private cost of curtailed computation (lost tokens, deadline penalties), and A_d collects frequency-response and capacity-market participation within the clearing-curve program of Clack (2026b), consistent with the same physical constraints, so that no megawatt is double-counted across stacked services; A_d is concave, being the gross-surplus term of that clearing-curve program. The parameterisation $(\delta^u_d, \delta^c_d, \tau_d, \eta_d, w_d, E^c_d, \phi_d)$ is the service-level boundary of the cohort: it is where the engineering claims of workload orchestration — demonstrated power reductions of order 25% during grid events without violating compute service levels (Colangelo et al., 2025) — enter the economics as constraints rather than as assertions.

3.3 Equilibrium characterisation

The composition of a pricing operator with a response operator offers the right intuition — responses and prices must be mutually consistent — but it is not the right computational object. As established in Clack (2026a), the profit-maximising dispatch map is the solution map of a linear program and is therefore discontinuous; the composition is not non-expansive, a pointwise fixed point need not exist, and relaxed Picard iteration on the raw operators demonstrably fails to converge. The competitive equilibrium is instead characterised as the unique minimiser of a strictly convex welfare program, with equilibrium price equal to the dual of power balance.

The augmented program of this paper extends that characterisation to flexible demand. The objective collects: system cost under the strictly convex supply representation of Clack (2026a); storage throughput and degradation costs; the private costs of flexible-demand response — curtailment disutility $\phi_d \sum_t \Delta t\, c_{d,t}$; minus the gross-surplus integrals of the frequency-response clearing curves of Clack (2026b), through which both storage and flexible demand supply stacked response. Minimisation is over the joint feasible set: generation and interconnector limits, storage schedules with shared power and energy constraints as in Clack (2026b), stacked-service commitments, and the flexible-demand polytopes $\mathcal{F}_d$ of Section 3.2. The cohort constraints (1)–(3) are linear and the disutility terms convex, so the augmented program remains convex; strict convexity of the price-forming aggregate is inherited from the supply representation, so existence follows exactly as in the storage-only case and the equilibrium price vector and aggregate response are unique. The per-cohort *allocation* is not pinned by these terms alone (the curtailment disutility is linear); it is pinned by adding a small strictly convex response disutility (ε-quadratic) to each cohort, stated with every run, and reported cohort values are invariant to that tie-break within reporting precision. Wholesale prices are the duals of half-hourly power balance; frequency-response prices are read from the clearing curves at equilibrium quantities; cohort values are surplus terms evaluated at the optimum. Every solution is certified optimal by the solver and passes the primal physical verification suite of Section 3.1 at solver-consistent tolerance (Appendix A); dual feasibility and complementary slackness are checked explicitly in the framework's full implementation (Clack, 2026a).

3.4 Value definitions

Throughout, G denotes *connected* data-centre capacity in gigawatts, operated flexibly; the zero-flexibility baseline ($G = 0$) keeps the same connected load operated *inflexibly*, so every difference is attributable to flexibility rather than to load growth. $F = \sum_d (\delta_d^u + \delta_d^c)\,\mathrm{LF}_d\,\bar{\ell}_d$ denotes flexible capacity at the baseline load factor LF_d, and all £/kW/yr figures are per kilowatt of *connected* capacity unless stated. Let $\boldsymbol{p}^0$ denote equilibrium prices at zero flexible response — the same connected load operated inflexibly, evaluated at the smallest modelled penetration of the sweep, G=2 GW (the "today" price shape of a price-taking pitch; the 0.25 GW cell is solved only as a perturbation point for the discrete marginals, not as the baseline) — and $\boldsymbol{p}^\star(G)$ the equilibrium price vector when G gigawatts of flexible data-centre load are present. Three quantities organise the results.

Price-taking value. V_d^{PT} is the annual value per kilowatt of connected capacity of cohort d's optimal response valued against the frozen baseline $\boldsymbol{p}^0$ — the number produced by every valuation that scales the resource while holding the market fixed.

Equilibrium value. $V_d^{\mathrm{EQ}}(G)$ is the annual value per kilowatt of connected capacity at the converged equilibrium with penetration G, decomposed by channel (wholesale, ancillary, capacity, and network; connection acceleration carries no market price, so it is quantified only qualitatively in Section 6 and is not a row of Table 1).

Overstatement ratio. $\kappa_d(G) = V_d^{\mathrm{PT}}/V_d^{\mathrm{EQ}}(G)$, the factor by which price-taking valuation overstates the equilibrium value; the demand-side analogue of the 100–300%+ storage result, where $\kappa = 1 + (\text{overstatement }\%)/100$, so that storage band corresponds to $\kappa = 2$–4.

Cross-cannibalisation. For battery cohort b, $\Delta R_b(G) = R_b(G) - R_b^0(G)$ is the change in annual revenue attributable to flexible data-centre penetration G, where $R_b^0(G)$ is the same fleet's revenue when the same connected load is operated inflexibly, holding the storage fleet at the reference build-out. We report $\Delta R_b(G)$ as *levels* by storage duration cohort, with discrete marginals between successive penetrations — an average $\Delta R_b(G)/G$ would be uninterpretable for a relation the paper argues is saturating. The revenue definition covers the wholesale and frequency-response streams; Balancing Mechanism and Capacity Market revenues are excluded, which biases ΔR_b towards zero, since turn-down from flexible load competes directly in both.

3.5 Scenarios

The background system follows the stylised GB representation of Clack (2026a), with 2030 and 2035 backgrounds and reference storage build-outs stated as scenario inputs in Appendix A, set deliberately below national pathway ambition (National Energy System Operator, 2026a). Flexible data-centre penetration is swept over $G \in \{0, 2, 5, 10\}$ GW of connected data-centre load with flexibility parameters drawn from a grid: deferrable depth $\delta^u \in \{0.10, 0.25, 0.40\}$, curtailable depth $\delta^c \in \{0.05, 0.15\}$, backlog duration $\tau \in \{2, 4, 8\}$ h, overhead $\eta \in \{0.95, 1.00\}$, with the payback window $w = 24$ h centrally and $w \in \{12, 48\}$ h and load factor $\in \{0.85, 0.95\}$ as one-at-a-time sensitivities, against multiple weather years. Grid corners whose stated backlog duration cannot be paid back within the window under the connection-headroom constraint (in mean-rate terms, $\tau\,\delta^u/\eta > (1 - \mathrm{LF})(w - \tau)$ at load factor LF) are infeasible as stated; they are flagged, excluded from the reported robustness bands, and recorded in the run manifest with their attainable duration. The central case is stated in Appendix A; the demonstration-anchored depth (order 25% event reduction) sits within the grid rather than at its edge. Co-location sensitivities pair flexible cohorts with on-site storage to test whether a flexible data centre erodes the value of its own battery.

4 Data

Price formation and the storage fleet follow the stylised calibration of Clack (2026a): synthetic demand, renewables and inverse supply curve, seed 20260901, targeted at historical GB 2024–2025 day-ahead conditions (as of August 2026) and scaled to the 2030/2035 backgrounds of Appendix A. The measured storage baseline against which levels are contextualised is the CECadence Quarterly GB BESS Index 2025Q3–2026Q2, computed from Elexon settlement cashflows and NESO auction records (Compounding Energy Ltd, 2026). Levels throughout Section 5 are therefore *stylised*: the load-bearing results are the overstatement ratio κ, the shape of the erosion, and the sign and ordering of the cross-effects, not the £/kW/yr levels themselves. Data-centre baseline load is a scenario input — the connected-capacity sweep of Section 3.5 — rather than a forecast, with national demand-projection context from National Energy System Operator (2026a) (retrieved August 2026). Flexibility parameters are anchored to the published field demonstration — a 25% cluster power reduction sustained for three hours at a 256-GPU commercial deployment with quality-of-service guarantees maintained (Colangelo et al., 2025) — for depth and duration, with availability envelopes informed by the curtailment-enabled-headroom analysis of Norris et al. (2025), which finds 76–126 GW of new US load integrable at annual curtailment rates of 0.25–1.0% of the load's potential consumption with mean event durations of order two hours; as of September 2026 these remain the most recent public demonstrations of their kind. Service-level caps are set conservatively, and all cohort parameters are reported with every run so that a reader who disputes a service-level assumption can re-run the disputed cell (the case-study bundle — stage runner, seed, results — is available on request).

5 Results

[Every value in this section is generated by the run matrix of Appendix A at seed 20260901 and passes the physical verification suite; levels are stylised (Section 4). Values are means over three synthetic weather years unless stated.]

5.1 The equilibrium value of data-centre flexibility

Table 1: Annual value of data-centre flexibility in £ per kW of *connected* data-centre capacity, by channel and penetration; central cohort parameters; 2030 background; stylised levels (Section 4). The marginal column is the discrete marginal over the last half-gigawatt at G=10. Price-taking column valued at the frozen baseline $\boldsymbol{p}^0$ (Section 3.4). The memo row gives flexible capacity F at each G. Totals are computed before rounding.

	Equilibrium V^{EQ} (£/kW-connected/yr)				Price-taking
Channel	G=2	G=5	G=10	marginal at G=10	V^{PT}
Wholesale shaping	9.0	8.0	6.6	4.0	9.9
Balancing/ancillary	1.2	1.0	0.1	−0.1	1.2
Capacity market	6.4	5.6	4.3	1.5	6.9
Network	0.5	0.5	0.4	0.4	1.6
Total	17.1	15.1	11.4	5.9	19.7
Memo: flexible capacity F (GW)	0.54	1.35	2.70	–	–

The headline comparisons are (i) the decline of the total column in G — the erosion curve, plotted as Figure 1 — and (ii) the gap between the equilibrium columns and the price-taking column, summarised as the overstatement ratio $\kappa(G)$ in Table 2. The pre-registered hypothesis that saturation arrives first in the shallow ancillary channel is confirmed outright in the 2035 background and up to a tie in 2030 — a nuance the pre-registration did not anticipate: in the 2030 background the ancillary and capacity channels both fall below half their first-tranche marginal value in the same 5–10 GW tranche, but only the ancillary channel is fully exhausted by G=10 — its marginal value is ≈ 0 — while in the 2035 background the ancillary channel crosses first outright, below $\sim$4 GW. The marginal megawatt at G=10 retains 32% of the value of the first quarter-gigawatt tranche (37% in 2035); equivalently, the marginal flexible megawatt loses 62% of its value between G=2 and G=10 (57% in 2035). Across the feasible cells of the cohort grid (run block R2, solved in a single synthetic weather year at the 2030 background), the equilibrium total at G=5 spans £7.0–18.0/kW/yr; the eight grid corners whose stated backlog duration is unattainable under the connection-headroom constraint are excluded and recorded in the run manifest with their attainable durations. The one-at-a-time payback-window and load-factor sensitivities of Section 3.5 move the central total between £11.2 and £18.9/kW/yr.

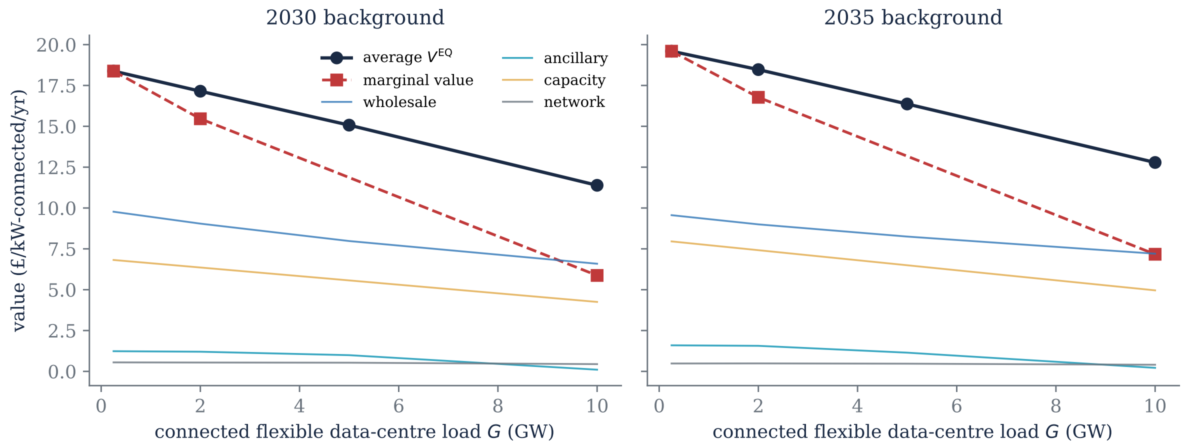

Figure 1: The erosion curve: average and marginal equilibrium value of flexibility (£/kW-connected/yr) versus connected flexible data-centre load G, by channel; 2030 and 2035 backgrounds; stylised levels.

Table 2: Overstatement ratio $\kappa(G) = V^{\mathrm{PT}}/V^{\mathrm{EQ}}(G)$, total across channels.

	G=2	G=5	G=10
2030 background	1.15	1.31	1.73
2035 background	1.14	1.28	1.64

5.2 Cross-cannibalisation of battery storage revenue

Flexible data-centre load is, functionally, a demand-side battery competing for the same spreads and the same shallow ancillary stack as the storage fleet. We are not aware of a prior published quantification of that competition for Great Britain. The pre-registered direction is confirmed: the erosion is real for every cohort and grows with penetration — a fleet-average 21% of baseline revenue at G=10 — an unpriced scenario in current storage financings. One pre-registered expectation is refuted and reported as it landed: in £/MW/yr the erosion is largest for the *four-hour* cohort, not the short-duration cohorts, because the flexible cohorts' defer-and-pay-back cycle competes most directly with longer-duration arbitrage in the stylised setting; as a share of each cohort's own baseline the ordering reverses (−27% for 1 h against −21% for 2 h and −20% for 4 h at G=10), so the short-duration financing case remains the most exposed proportionally

Table 3: Change in GB BESS annual revenue attributable to flexible data-centre load, $\Delta R_b(G)$, in £k per MW of storage capacity per year, by storage duration cohort; reference storage build-out; 2030 background; wholesale-plus-frequency-response streams (BM and CM excluded, Section 3.4); baseline is the same connected load operated inflexibly. Negative values are revenue erosion. The final row is the fleet-average discrete marginal between successive penetrations.

	$\Delta R_b(G)$ (£k/MW/yr)		
Storage cohort	G=2	G=5	G=10
1 h	−1.9	−4.1	−5.7
2 h	−2.0	−5.0	−7.7
4 h	−4.2	−9.2	−14.9
Fleet-average marginal (£k/MW/yr per GW)	−1.2	−1.1	−0.6

on those same stylised baselines (see the limitations of Section 6). The mechanism is already visible in measured data: the CECadence Quarterly GB BESS Index records frequency-response saturation under a 15% fleet build (5.53 to 6.37 GW active), with Dynamic Regulation High clearing negative in 2026Q2 (Compounding Energy Ltd, 2026) — the same shallow stack the flexible fleet enters here. Figure 2 shows the wholesale mechanism: compression of the price-duration curve as flexible load occupies its steep segment.

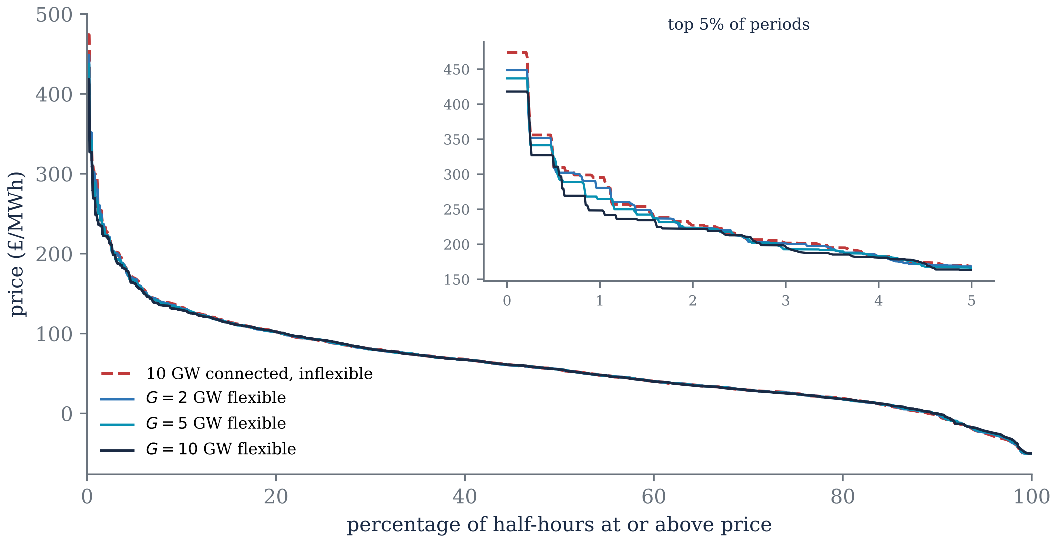

Figure 2: Price-duration-curve compression under flexible data-centre penetration: 2030 background, one weather year, flexible arms at $G \in \{2, 5, 10\}$ GW with the 10 GW-connected inflexible profile as reference; inset — top 5% of periods.

5.3 Co-location

Pairing flexible cohorts with on-site storage tests whether flexibility and co-located batteries are complements or substitutes at equilibrium; the site is evaluated as a marginal entrant against the G-equilibrium price vector (run block R4, Appendix A). The pre-registered hypothesis was substitution: the flexible load and its own battery compete for the same periods. The runs confirm it: the incremental wholesale value of a two-hour battery behind a flexible data centre is 79% of its standalone value at G=5 (78% at G=2, 80% at G=10), and the jointly optimised pair is worth about 8% less than the sum of its separately valued parts at every penetration — the cross-partial is negative — so flexibility and co-located storage are substitutes, not complements.

6 Discussion

For the underwriting of flexibility revenues. Capital is being raised, and connection policy designed, on flexibility values computed at today's system conditions. Table 2 is the correction factor. The direction of the result is not in doubt — rents erode; the contribution is its magnitude, its timing, and its decomposition by channel, which together determine which flexibility business models survive scale and which are artefacts of being early. Investors underwriting flexibility-linked revenues, and lenders to flexible infrastructure, should require equilibrium valuations for the same reason they have begun to require them for storage.

For battery storage financing. Every GB storage revenue case in circulation assumes inflexible demand. Table 3 prices the alternative. A financing that stresses gas prices and build-out but not the arrival of gigawatt-scale flexible load is stressing yesterday's risk factors; we propose flexible-demand penetration as a standard downside scenario in storage credit analysis, alongside the fleet-cannibalisation scenarios of Clack (2026a).

For connections policy. Remunerating flexibility with queue position is an elegant response to the connections constraint, and Great Britain is well placed to design it deliberately rather than inherit it. But an exchange rate between flexibility and acceleration set against price-taking values will over-pay — in the scarce currency of connection capacity — for flexibility whose system value declines with every award. The efficient exchange rate is the equilibrium value, and it falls with cumulative uptake; mechanism design should anticipate that decline rather than lock early-adopter values into standing arrangements.

Limitations. The analysis is conducted at national pricing resolution consistent with the GB market design; locational value within GB — which nodes reward flexibility most — is deferred to subsequent work. Cohort responses are modelled as price-takers in the game-theoretic sense — atomistic, with no strategic withholding. This is distinct from the price-taking *valuation* criticised throughout, in which the price vector itself is held fixed as the resource scales; here prices adjust in equilibrium. Strategic withholding by large flexible portfolios is excluded by assumption and would, if present, alter the division of a smaller rent, not enlarge it. A second scope condition is the accommodation assumption behind the netting of Section 3.2: the background is taken to accommodate the projected data-centre load, so that only the load's shape and its flexible response enter residual demand. That choice is load-bearing, not bookkeeping. In a background that does not accommodate the load, scarcity from unaccommodated load growth inflates the value of flexibility faster than penetration erodes it, and the sign of the price-taking bias reverses: in the retained pre-netting audit run set — which we do not advance as a valuation, since it double-counts load the background already embeds — $\kappa(10)$ is 0.76 rather than 1.73 in the 2030 background (0.81 against 1.64 in 2035). The overstatement quantified in this paper is therefore a statement about systems planned around the load they host; where connection outpaces accommodation, price-taking valuation errs in the opposite direction — and neither error is an argument for holding the price vector fixed. The stylised baseline also inverts one measured ordering: in the model, baseline storage revenue per megawatt rises with duration, whereas the measured Index records the four-hour cohort earning the least per megawatt (Compounding Energy Ltd, 2026); the duration rankings of Section 5.2 — the £/MW/yr ordering within Table 3 and the proportional ordering, whose denominators are those same stylised baselines — should therefore be read as properties of the stylised setting: against the measured per-MW baselines the four-hour cohort's modelled erosion would exceed the whole of its measured Index revenue (itself a lower-bound figure), so neither levels nor proportional shares transfer, while the sign of the erosion and its growth with penetration are the results we stand behind. One computational caveat: the stylised network layer values load reduction over the discrete top-K peak windows of the residual profile, and the candidate windows are near-tied in these runs, so a raw window selection is identified only up to solver noise — worth roughly two points on the marginal-decline and retention figures in either background at the half-gigawatt granularity of the discrete marginals. The printed marginals therefore value the network channel at its expected value under uniform tie-breaking across near-tied windows, a treatment recorded in the run manifest; channel levels move by less than £0.04/kW/yr under the same treatment. Service-level parameters are anchored to public demonstrations and stated per run; readers who dispute them can re-run the disputed cell, which is the review model this series invites (the case-study bundle is available on request).

7 Conclusion

Flexibility is real and valuable, and at the penetrations now under discussion it is systematically over-valued by the arithmetic used to underwrite it. The same fixed-point discipline that corrected storage revenue expectations applies unchanged to flexible demand: value the resource at the equilibrium its own scale creates, or overstate it by a factor that grows with every megawatt deployed. For Great Britain we quantify that factor as rising from 1.15 at 2 GW to 1.73 at 10 GW of connected flexible load, locate the earliest saturation in the shallow ancillary stack, and price — we believe for the first time — the erosion that flexible data-centre load imposes on the battery fleet's revenue case. As everywhere in this series, the claim to be trusted on is not the mechanism but the record: pre-registered, publicly graded, with the fails published alongside the passes (Compounding Energy Ltd, 2026). The grid will be rebuilt around flexible load once. The numbers should be right.

A Run matrix and manifest

Every run is identified by (scenario, code version, timestamp) with *stylised* and *validated-on-settlement-data* flags, and passes the physical verification suite before its values enter the paper. All results in Section 5 were generated at seed 20260901; scenario `CE-WP-2026-05-stylised-GB`, code version `2cedb716f388`, assembled 2026-08-29 23:37 UTC. The full manifest, per-cell results and figures ship with the case-study bundle (available on request). Simulation horizon: 26 representative weeks at half-hourly resolution, annualised; three synthetic weather years.

Run block	Sweep	Populates	Notes
R1	$G \in \{0.25, 2, 2.5, 5, 10, 10.5\}$, central cohort	Tables 1, 2; Figs 1, 2	both backgrounds, both arms[a]
R2	cohort grid $(\delta^u, \delta^c, \tau, \eta)$ at G=5	robustness band in Section 5.1	28 of 36 cells feasible[b]; single weather year
R3	ΔR_b at reference storage fleet	Table 3	assembled from R1 arms
R4	co-location pairs	Section 5.3	standalone vs paired vs joint

[a] Half-gigawatt perturbation points (2.5, 10.5) supply the discrete marginals; each G is solved in a flexible arm and an inflexible-baseline arm (same connected load, zero flexibility). [b] The eight pruned corners are the $(\delta^u, \tau) \in \{(0.25, 8\,\text{h}), (0.40, 8\,\text{h})\}$ cells, whose stated backlog duration is unattainable under the connection-headroom constraint at the central load factor; their attainable durations are recorded in the run manifest.

Central case. The central cohort and background parameters, all scenario inputs:

Parameter	Value	Parameter	Value
Deferrable depth δ^u	0.25	Payback window w	24 h (tumbling)
Curtailable depth δ^c	0.05	Baseline load factor	0.90
Backlog duration τ	4 h	Curtailment cap E^c	0.5% of hours, full depth
Overhead η	0.95	Curtailment disutility ϕ	£250/MWh
Recovery limit $\bar{r}$	connection headroom	Event limit $(\bar{C}, H)$	2 h depth per 12 h window
Storage build-out 2030	3/5/2 GW (1/2/4 h)	Storage build-out 2035	3/7/4 GW (1/2/4 h)

The deferrable depth follows the demonstrated order-25% event reduction of Colangelo et al. (2025); the curtailment envelope follows Norris et al. (2025). The case study implements the framework in reduced form: storage cohorts and the flexible cohort are co-optimised with an aggregate frequency-response product inside the welfare program, while the capacity-market and network layers are applied as annual post-solve schedules. The per-run verification suite here is the primal suite — system balance under the netting convention, state-of-charge and backlog bounds, tumbling-window zeroing, connection headroom, and the curtailment caps — applied at solver-consistent tolerances recorded in the run manifest, together with certified-optimal solver status for every cell; the flow-limit and duality checks of Section 3.1 belong to the full framework. Served prices are the stylised inverse supply curve evaluated at the equilibrium residual, with its price floor and cap applied as scenario inputs; the floor binds in a small share of trough periods, and valuing at the unclipped curve moves channel values by less than £0.2/kW/yr. The frequency-response clearing curves, capacity-market suppression schedule and network-charge layer are stylised scenario inputs recorded in the run manifest; their parameterisations are deliberately not reproduced here.

References

Christopher T. M. Clack. Cannibalisation as a fixed point: revenue forecasting for battery energy storage in liberalised electricity markets. Working Paper CE-WP-2026-01, Compounding Energy Ltd, 2026a. URL https://www.compoundingenergy.com/library/CE-WP-2026-01_Cannibalisation_Fixed_Point.pdf.

Christopher T. M. Clack. Co-optimising the GB BESS revenue stack: wholesale energy and the NESO dynamic-response suite at the cannibalisation equilibrium. Working Paper CE-WP-2026-02, Compounding Energy Ltd, 2026b. URL https://www.compoundingenergy.com/library/CE-WP-2026-02_GB_BESS_Revenue_Stack.pdf.

Philip Colangelo, Ayse K. Coskun, Jack Megrue, Ciaran Roberts, Shayan Sengupta, Varun Sivaram, Ethan Tiao, Aroon Vijaykar, Chris Williams, Daniel C. Wilson, Zack MacFarland, Daniel Dreiling, Nathan Morey, Anuja Ratnayake, and Baskar Vairamohan. Turning AI data centers into grid-interactive assets: Results from a field demonstration in phoenix, arizona, 2025. URL https://arxiv.org/abs/2507.00909. arXiv:2507.00909.

Compounding Energy Ltd. CECadence quarterly GB BESS index, 2025Q3–2026Q2. Technical report, Compounding Energy Ltd, 2026. URL https://www.compoundingenergy.com/gb-bess-index-2026q2.html. Released 22 July 2026; computed from Elexon settlement cashflows and NESO auction records for the full 168-BMU battery registry.

Data Centre Dynamics. National Grid partners with Emerald AI on data center grid flexibility demonstration project in UK, 2026. URL https://www.datacenterdynamics.com/en/news/national-grid-partners-with-emerald-ai-on-data-center-grid-flexibility-demonstration-project-in-uk/. 7 May 2026.

Fernando J. de Sisternes, Jesse D. Jenkins, and Audun Botterud. The value of energy storage in decarbonizing the electricity sector. *Applied Energy*, 175:368–379, 2016. doi: 10.1016/j.apenergy.2016.05.014.

Electric Power Research Institute. DCFlex initiative: Data center flexibility demonstrations, 2026. URL https://www.epri.com/dcflex. Initiative pages, retrieved August 2026.

Emerald AI. Emerald AI raises $150 million series A at $1.05 billion valuation to scale power-flexible AI data centers, 2026a. URL https://www.businesswire.com/news/home/20260825127649/en/. Business Wire, 25 August 2026.

Emerald AI. Emerald AI announces first DSX Flex deployment with NVIDIA and silicon valley power, 2026b. URL https://www.emeraldai.co/blog/nvidia-dsx-pilot-framework.

National Energy System Operator. Future energy scenarios. Technical report, NESO, 2026a. URL https://www.neso.energy/publications/future-energy-scenarios-fes. 2026 edition; retrieved August 2026.

National Energy System Operator. Balancing services: response and reserve market information, 2026b. URL https://www.neso.energy/industry-information/balancing-services. Published requirement volumes for the dynamic response and reserve suite; retrieved August 2026.

Tyler H. Norris, Tim Profeta, Dalia Patiño-Echeverri, and Adam Cowie-Haskell. Rethinking load growth: Assessing the potential for integration of large flexible loads in US power systems. Technical Report NI R 25-01, Nicholas Institute for Energy, Environment & Sustainability, Duke University, 2025. URL https://nicholasinstitute.duke.edu/publications/rethinking-load-growth.

NVIDIA Corporation. NVIDIA and Emerald AI join leading energy companies to pioneer flexible AI factories as grid assets, 2026. URL https://investor.nvidia.com/news/. Press release, 23 March 2026; retrieved from the NVIDIA investor news index, August 2026.

Ofgem. Targeted charging review: decision and impact assessment, 2019. URL https://www.ofgem.gov.uk/publications/targeted-charging-review-decision-and-impact-assessment.

Ofgem. Capacity market (CM) rules, 2026. URL https://www.ofgem.gov.uk/guidance/consolidated-version-capacity-market-rules. Consolidated rules; retrieved August 2026.

Goran Strbac. Demand side management: Benefits and challenges. *Energy Policy*, 36(12):4419–4426, 2008. doi: 10.1016/j.enpol.2008.09.030.