Online Resource 1

for the article

When the Optimum Matters Less: Certified Near-Optimal Window-Count Selection for a Single ARINC-653 Partition

Submitted to Real-Time Systems. Anonymized for double-anonymous peer review.
Supporting derivations, regression details, verification and preregistration records, point-prediction diagnostics, and secondary results.
This is Online Resource 1 for the article "When the Optimum Matters Less: Certified Near-Optimal Window-Count Selection for a Single ARINC-653 Partition". Sections, tables, and figures carrying an S prefix belong to this document; all other cross-references—section numbers, Tables 1–11, Figs. 1–5, and Appendix A items A.1–A.7, including Lemma A.4b—refer to the main paper.

S1 The Surrogate Optimizer

For the fluid guarantee $G(k) = \alpha_k(t_c - \Delta_k)$ of Theorem 1 (written G to keep it apart from the necessity bound $L(k)$ of Theorem 2) (slack term omitted; it is $O(k)$ and immaterial to the optimizer's location at scale) with evaluation scale $t_c > \delta$:

$$H\,G(k) \;=\; Q(t_c{-}\delta) + \delta(H{-}Q) \;-\; \left[\frac{Q(H-Q)}{k} + \delta(t_c{-}\delta)\,k\right].$$

The δ-dependent cross terms cancel on expansion; the bracketed sum $u/k + vk$ with $u = Q(H{-}Q) > 0$, $v = \delta(t_c{-}\delta) > 0$ is uniquely minimized at $k_0 = \sqrt{u/v}$ by AM–GM, with minimum $2\sqrt{uv}$. Concavity of L over the integers makes the projected rule (evaluate $\lfloor k_0 \rfloor, \lceil k_0 \rceil$, clip to $[1, k_{\max}]$) exact for the surrogate—confirmed against brute force on 5,000 random instances with zero mismatches. Rules **R1** (this projection), **R2** (the same optimization over the actual deadline set), and **R1-FP** (a fixed-priority variant) are the predictors evaluated—and rejected for the exact problem—in Sect. S4. The scaling prediction $k_0 \propto \delta^{-1/2}$ is the origin of hypothesis H4.

S1.1 Why the two fluid surrogates diverge

Both Q_{lin} and R2 descend from the same fluid relaxation, yet their immediate-certification rates differ by a factor of 2.4 at $\delta{=}100\,\mu s$ (98.5% vs. 40.5%). Corollary 2 explains this. Certification requires $Q_{\text{best}} \le (1+\varepsilon)(k_{\min}\delta + L_0)$, while $Q_{\min}(\hat{k})$ carries the initializer's own overhead $\hat{k}\delta$; the test therefore rewards initializers that keep $\hat{k}\delta$ small, independently of how well $\hat{k}$ predicts the optimizer. The measured initializer distributions match: Q_{lin} selects a median $\hat{k}$ of 8 at $\delta{=}100\,\mu s$ and 4 at $\delta{=}1000\,\mu s$, whereas R2—whose surrogate optimum scales as $\delta^{-1/2}$ (above)—selects 24 and 8 respectively (uniform sampling of the 64 candidates would give a mean of 32.5). The gap is thus a consequence of the bound's affine structure rather than of R2's point accuracy, and it is the mechanism behind the workflow's saving: the surrogate initializer's preference for small $\hat{k}$ is a population property, not a property of any individual taskset (the Sect. 7 case study selects $\hat{k}{=}11$, well inside the observed range $[7, 27]$ at $\delta{=}100\,\mu s$).

S1.2 Empirical surrogate count

The surrogate count. The count $\hat{N}(\varepsilon) = |\{k : \lceil Q_{\text{lin}}(k) \rceil \le (1{+}\varepsilon)L^*\}|$ combines the surrogate with the certified bound. Theorem 3 makes every counted candidate certified: $Q_{\min}(k) \le \lceil Q_{\text{lin}}(k) \rceil \le (1+\varepsilon)L^* \le (1+\varepsilon)Q^*$, so the counted set is contained in W_ε and $\hat{N}$ is a certified

lower bound on $|W_\varepsilon|$—the left side of the sandwich of Corollary 3. It is not a two-sided estimate: nothing here bounds how many members it misses, and the measurements below quantify that slack. Measured, it is set-conservative in every evaluated case—the counted set is a subset of W_ε at every evaluated tolerance ($\varepsilon = 1\%, \ldots, 10\%$) over all 720 taskset–overhead combinations at $k \leq 64$—informative where the set is wide (δ=100: median certified count 17 against a measured median of 33.5 at ε=7%, and 32.5 against 46.5 at 10%) and vacuous where it is narrow: the surrogate–necessity gap must be crossed before it counts anything, and the minimal ε at which it becomes non-vacuous has median **4.5%** at $\delta = 100\,\mu s$ and **12.3%** at $\delta = 1000\,\mu s$ (Table S1, Fig. S1). Theorem 3 supplies the sufficient-budget result behind this certified lower count; the necessity bound of Theorem 2 supplies the ingredients for the upper count, and Corollary 3 combines the two into the sandwich; its evaluation on constrained-deadline populations remains open (Sect. 9.4).

Table S1 Median surrogate count $\hat{N}(\varepsilon)$ versus median measured $|W_\varepsilon|$ per overhead level, over all feasible $k \leq 64$; half-integer medians (even sample sizes) are shown exactly. The count is a certified lower bound on $|W_\varepsilon|$ (Corollary 3, left side); its tightness against the measured set is the empirical question measured here, and it is set-conservative in every evaluated case, substantial in the wide regime once ε clears the surrogate–necessity gap and vacuous below it.

ε	3%	5%	7%	10%	20%
δ=100: count / measured	0 / 15.5	7 / 25.0	17 / 33.5	32.5 / 46.5	59 / 63
δ=1000: count / measured	0 / 2	0 / 3	0 / 5	0 / 6	5 / 11

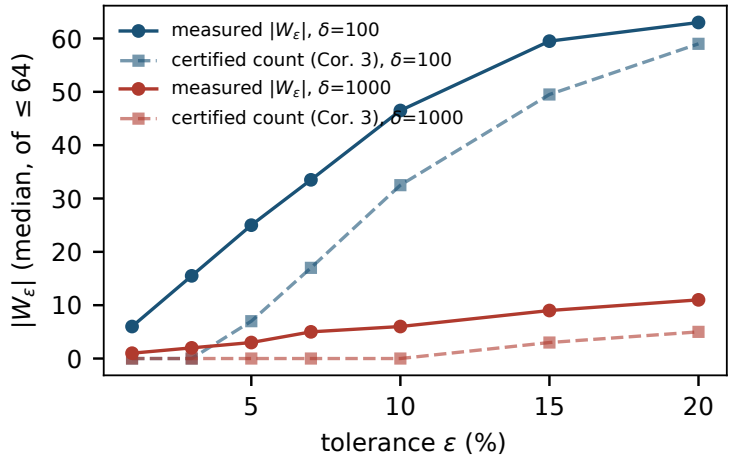

Fig. S1 Surrogate count $\hat{N}(\varepsilon)$ versus the measured $|W_\varepsilon|$: informative where the set is wide; conservative in the narrow regime, with the entry tolerance quantified in the text

S2 Regression and Ladder Detail

What governs the width—and is the scale available ex ante? An exploratory multivariate analysis (G14) tests this. In a standardized regression of $\log |W_{5\%}|$ on $\log \eta_\delta$, utilization, task count, a harmonicity index, and period spread (n=400), the relative-overhead term carries a standardized coefficient of -0.98—close to the exponent -1 that the width law itself posits—while every other variable adds essentially nothing (partial $R^2 \leq 0.044$ each; full coefficients in Sect. S2.1). Three caveats fix the result's epistemic status (a fourth check comes

from Sect. 6.1: grouped cross-validation leaves the predictive R^2 at 0.953). First, the fit largely *confirms the width law's functional form*; the informative content is the measured unimportance of utilization, task count, harmonicity, and spread once η_δ is present (partial $R^2 \leq 0.044$ each). Second, δ enters at two levels a decade apart, so only the covariates' small partial R^2 should be read as a finding. Third, $\eta_\delta = \delta/Q^*$ uses the measured optimum in its denominator, so regressor and response derive from the same curves—an *explanatory* scale, not automatically an ex ante predictor. We therefore re-ran the identical regression (G14b, same folds) with denominators computable *before any exact sizing*: the certified bound L^* and the surrogate minimum $\min_k \lceil Q_{\text{lin}}(k) \rceil$. The explanatory power is essentially unchanged (Sect. S2.1), reflecting how closely the certified bound tracks the optimum (median $L^*/Q^* = 0.99$ at δ=100, 0.93 at δ=1000). The dominant scale is thus not an artifact of the shared-curve construction and *is* available ex ante; what the exact denominator buys is per-taskset ranking, not the scale: the width law's rank fidelity at δ=100 is Spearman 0.95 with Q^* against 0.77 with L^* (0.80 with the surrogate), while its median-scale prediction is unchanged (20.1–21.0 vs. 20.4). Within the sampled factor grid, the synthetic workload families, and the examined feature set, no measured factor competes with relative per-window overhead in explaining plateau width, and an actionable ex ante estimate of it exists.

Provenance of the overhead-ladder study (G16): prospectively registered, frozen before execution and measured with the cap-free procedure of Sect. 4.3. A two-level design cannot identify the width-law exponent, and a regression whose response is algebraically related to its predictor cannot either; the ladder addresses both.

S2.1 Width-dominance regression

The regression of Sect. 4.2 is reported here in full. Standardized OLS of $\log |W_{5\%}|$ on $\log \eta_\delta$, utilization, task count, the harmonicity index and period spread (n=400 taskset–overhead conditions from 200 unique tasksets; study G14, prospectively registered as exploratory). The ex ante re-analysis (G14b, same folds) replaces $\eta_\delta = \delta/Q^*$ by proxies computable before any exact sizing. Uncertainty for the exact-η_δ model is cluster-robust on the 200 base tasksets: the CR1 standard error on the $\log \eta_\delta$ coefficient is 0.0097 against 0.0104 under an independent-rows assumption, an SE ratio of 0.94, so the two-level design does not inflate the nominal precision; the 2,000-draw cluster bootstrap interval is $[-0.995, -0.958]$. Table S2 reports all three models.

Table S2 Width-dominance regression (computed on the original preregistered curves; under the cap-free correction every coefficient moves only at the third decimal—e.g. the $\log \eta_\delta$ partial R^2 0.955 → 0.957—and no competing covariate exceeds 0.044). Every competing covariate (utilization, task count, harmonicity index, period spread) reaches partial $R^2 \leq 0.044$ in all three models. The full model attains $R^2 = 0.959$; removing $\log \eta_\delta$ collapses it to 0.049. The caveats of Sect. 4.2 apply: these figures largely confirm the width law's functional form rather than decomposing the variance independently, and δ enters the design at two levels a decade apart. Median L^*/Q^* is 0.99 at δ=100 μs and 0.93 at δ=1000 μs.

model	standardized coef. on $\log \eta_\delta$	partial R^2	5-fold CV R^2
G14 (exact $\eta_\delta = \delta/Q^*$)	−0.98	0.955	0.955
G14b (certified bound, δ/L^*)	−0.98	0.952	0.952
G14b (surrogate minimum)	−0.97	0.953	0.952

S3 Verification and Preregistration Records

This section reproduces the audit material summarized in Sects. 4.3 and 5: the resolution-status census, the verification and integrity gate counts, the study development history, the registered factor grids, the deviation log, the full cap-free re-measurement diagnostic, and the per-analysis registration ledger.

Resolution status of the measured curves (Sect. 5). Every released curve row carries a `status` field recording whether its budget was resolved exactly or to a certified two-sided bracket. In the cap-free census (24,100 rows) 19.4% are exactly resolved, 72.4% are certified two-sided brackets and 8.2% were never cap-affected, with bracket width a median of 5 and at most 98 time units. The overhead ladder (54,400 rows) is 25.0% exact with median width 3. The deadline and structured-family populations (46,607 rows) are 63.2% exact overall, a figure that splits sharply by study: 99.6% exact for the deadline families against 11.7% for the structured families, the latter being where the cap bound most often. These widths are four orders of magnitude below the budgets themselves, which is why the endpoint-invariance audit of Sect. 5 leaves every reported median unchanged.

Per-family detail (Sect. 8.2). G8-implicit was not re-run: it is bit-identical to G7, with 0 mismatches across 15,360 integrity-gated cells. Among these families, G8-tight resolves exact at all 12,921 rows (following the layout-domain reclassification of Lemma A.2) and G8-moderate at 99.2%, the deadline-aware bound leaving the horizon cap no room to bind on constrained deadlines; the harmonic and semi-harmonic families resolve exact at 11–13% of rows and to certified two-sided brackets elsewhere, of median width 0.006–0.008% of the budget. Under the endpoint-invariance audit, Table 6's G8-moderate row is unchanged at either endpoint (86.2/1/98.1) and G8-tight has no brackets at all. By interval analysis the optimizer identity is ambiguous in 1/80 harmonic and 2/80 semi-harmonic conditions at δ=100 μs and in none elsewhere; 17/80 and 11/80 of those conditions carry at least one borderline set member, yet the median $|W_{5\%}|$ range collapses to a point in every cell (harmonic 16 [16, 16]).

Table S3 lists the verification and integrity gates with their case counts.

Study development and evidence roles. The study developed in two stages. The original registered experiments (E1–E5) evaluated low-cost point-selection rules and the secondary PRM and packing questions. Subsequent separately registered studies (C3/G7, G8, G13, G15) characterized the complete objective geometry, deadline sensitivity, perturbation stability, and workload-structure robustness. The present manuscript organizes these results by analytical role rather than by execution order:

Role	Studies
Full-curve geometry (Q1, Q2)	C3, G7
Deadline sensitivity and local stability	G8, G13
Workload-structure robustness	G15
Certification and workflow evaluation (Q3)	Theorem 2, workflow replays
Point-selection diagnostics	E1, E2, round B
Secondary system results	E3, E4

Factors and scale. Five experiments were frozen. **E1** (original point-selection primary, now the Sect. S4 diagnostic; H1): policy EDF, FP $\times$ regime short-period, long-period $\times$ $U \in \{0.2, 0.3, 0.4, 0.5\} \times \delta \in \{10, 100, 1000, 10000\}\,\mu s$ (plus δ=2 μs spot cells anchoring tightly coupled kernels), 150 replicates per cell (about 10,200 taskset–policy rows): per taskset, k^* by exhaustive exact sizing over $[1, k_{\max}]$ (ties to smaller k), rule choices $k_{R1}, k_{R2}, k_{R1\text{-}FP}$ computed from a single k=1 sizing pass (rule inputs must not require search), and regret = $(Q_{\min}(k_{\text{rule}}) - Q_{\min}(k^*))/Q_{\min}(k^*)$. **E2** (H4): δ swept over seven half-decade levels $\{10, \ldots, 10000\}\,\mu s$ with *paired* tasksets (the same 200 tasksets per cell re-analyzed at every level), so within-taskset trajectories $k^*(\delta)$ are observable; OLS on per-cell medians. **E3** (H2): PRM sizing $\Theta_{\min}$ via

Table S3 Verification and integrity gates. The analyzer-verification tiers were passed before the main experiments; the certificate-implementation and dataset-integrity checks were required before the corresponding results were accepted. Counts under the certified-bound row are implementation deviations from the theorem-guaranteed inequality, not tests of the theorem itself. The 12/800 differential disagreements were all conservative (analysis rejects at a worst phase the synchronous-release simulation passes).

tier	check	cases	violations
closed form	PRM sbf vs Shin–Lee formula	3,000	0
closed form	single-window sbf vs hand formula	3,000	0
property	superadditivity (Lemma A.1)	2,000	0
property	Lemma 1 (candidate phases) vs exhaustive-phase brute force	120	0
property	budget monotonicity / Theorem 1 validity (incl. residue grid, 17,926 configs, slack attained)	300 / 17,926	0
property	Proposition A.3 identity & divisible dominance	1,300	0
property	delay-identity exhaustive grid; surrogate optimizer vs brute force	grid / 5,000	0
differential	independent simulator, EDF+FP	800	0 unsound
certified bound	Theorem 2 necessity $L(k) \leq Q_{\min}(k)$	720 taskset–overhead combos, all $k \leq 64$	0
integrity	seed-stability check; per-group seed uniqueness; G8-implicit ≡ G7 bit-identity; guard consistency	full data (71,680 rows)	0

$\Gamma(H/k, \Theta)$ binary search with effective-overhead semantics matched to A3, versus exact window sizing, over $k \in \{1, 2, 4, 8\} \times \delta \in \{0, 100, 1000\}$ (28,800 measurements); inflation $= (k\Theta_{\min} - Q_{\min}(k))/Q_{\min}(k)$. **E4** (H3): four-partition systems, $U_{\text{tot}} \in \{0.4, \ldots, 0.7\}$ split by UUniFast, three pipelines (Sect. S5), mandatory exact re-verification after packing, 150 systems per cell. **E5** (exploratory): sampled structural checks of Proposition A.3's divisible dominance (results in Sect. 5) and a non-divisible counterexample search (recorded example: H=240, Q=198, k=2 vs 3 mutually non-dominating).

Deviations. Three occurred, logged with timestamps and rationale. **D0** (pre-authorized): E1 replicates 300→150 on a compute-budget overrun. **D1** (not pre-authorized; disclosed): $k_{\max}$ 32→16 for $\delta \geq 100\,\mu s$ cells, retaining 32 where the surrogate can place k^* above 16; the truncation risk was monitored by the boundary-hit rate, observed at $\leq 2.7\%$ in affected cells and 0.9% elsewhere, so at most a 2.7% share of optima could lie beyond the cap. The dependence on the (later rejected) surrogate was removed by the expanded-range study G7, whose $k \leq 64$ curves place $k^* \leq 16$ in 100% of cases. **D2** (not pre-authorized; disclosed) — *analysis prioritization*: the original protocol designated PRM inflation (RQ2/H2) as the primary follow-up if H1 was not supported. The subsequent C3/G7 study was designed to examine the hit–regret separation observed in E1, and it now provides the primary evidence for this manuscript's geometric claims; the preregistered PRM analysis is retained as a secondary result (Sect. S5). This changes the presentation priority, not the specified analyses or their decision criteria. **Scheduler scope.** E1 evaluates EDF and fixed-priority variants; the structural follow-ups G7–G8, Theorem 2, and the certificate workflow of Sects. 4, 6, and 7 are EDF-based, and no fixed-priority claim is made for the near-optimality certificate.

Layout-domain deviation. The analyzer as preregistered searched budgets up to H although the residue-first layout is realizable only for $Q \leq k\lfloor H/k \rfloor$ (Lemma A.2); its δ>0 path validated disjointness of the effective intervals only, so raw overruns of up to H mod k ticks passed whenever δ exceeded them. Raw-layout guards were added to `a653ext.effective_supply` and both budget searches; re-running the two affected cells (G8 tight, δ=1000 μs: U=0.3, rep 22, k=49, stored 99,986 vs ceiling 99,960; rep 57, k=64, stored 99,994 vs 99,968) returns infeasible, and near-ceiling control rows (margins 5–22 ticks) reproduce their stored budgets exactly. Effect audit: both cells sit at roughly twice their condition's minimal budget, so k^* and $|W_{5\%}|$ are unchanged in both conditions (3/3 and 6/6, 3/3 and 4/4); the tightness census drops from 41,791 to 41,789 rows with every ratio median unchanged at the reported precision; all verification gates remain green.

Cap-free re-measurement: full diagnostic record (Sect. 4.3). In the $k \leq 64$ census (24,100 taskset–overhead–count combinations), $\lceil Q_{\text{lin}}(k) \rceil < Q_{\min}(k)$ occurs in 11,932 combinations (49.5%), always by at most 0.393%. All 11,932 share one cause: at $Q = \lceil Q_{\text{lin}} \rceil$ the analyzer's verification horizon $t^* = (\sum_i C_i + \alpha\Delta_{\text{int}})/(\alpha - U)$ exceeds its $400H$ safety cap—because $k\delta$ thins $\alpha - U$ toward zero at high k—and the implementation then rejects conservatively without checking any deadline. In all 11,932 combinations the Theorem 1 sufficiency condition $\alpha_k(t - \Delta_k) - \bar{r}_k \geq \text{dbf}(t)$ holds at every deadline point and at the tail, and in the five worst cases a direct check of every deadline point up to the full analytic horizon (up to $1{,}036H$; $\geq 3{,}484$ points each) confirms dbf $\leq$ sbf throughout. The Q_{lin}-based detector, however, understates the phenomenon: it can only flag combinations where $\lceil Q_{\text{lin}} \rceil$ undercuts the measurement, and Q_{lin} is loose at low k. The exact detector is analytic—the cap can have bound at $(taskset, \delta, k)$ only if the analyzer's internal horizon at $Q^{\text{meas}}_{\min} - 1$ exceeds $400H$—and it flags **22,132 of the 24,100 combinations (91.8%)**, across the entire k range. Every flagged combination is therefore evaluated with a cap-free exact decision procedure: verify all deadline points up to a staged horizon T_0 and accept exactly when the affine tail $\alpha_k(T_0 - \Delta_k) - \bar{r}_k \geq UT_0$ takes over—sound in both directions, since acceptance is the Theorem 3 argument and rejection exhibits a demand violation. The re-measurement confirms inflation in all 22,132 flagged combinations: median 0.36%, ninetieth percentile 0.51%, and maximum 0.70% of the measured budget (each released

row carrying an exact value or a certified schedulable upper value with its certified lower bound; near-critical rows whose exact minimum sits within microseconds of the utilization bound retain a bounded bracket; the residual set-size ambiguity these brackets admit across the tolerance sweep is quantified in Sect. 4.3).

Provenance of the evidence. Table S4 records the registration status of every reported analysis; protocols, decision criteria, and deviation logs are released with the code, and each frozen criterion and verdict is reported verbatim in its analytical section.

Table S4 Registration status of each reported analysis.

Analysis (section)	Registration status
Predictor accuracy and scaling (Sect. S4)	original preregistered (E1/H1, E2/H4)
PRM inflation; system packing (Sect. S5)	original preregistered (E3/H2, E4/H3)
Geometry, width, tolerance sweep (Sect. 6)	prospectively registered (C3, G7)
Deadline sensitivity (Sects. 4 and 6)	prospectively registered (G8)
Perturbation stability (Sect. 6)	prospectively registered (G13)
Structured workload families (Sect. 8)	prospectively registered (G15)
Overhead/utilization sensitivity (Sect. 4)	exploratory (round C)
Alignment / resonance diagnostics (Sect. S4)	exploratory (round B)
Width-dominance regression (Sect. 4)	prospectively registered as exploratory (G14; ex ante re-analysis G14b)
Published-parameter replays (Sect. 8.4)	not preregistered
Certified sufficiency, counting sandwich, deadline-aware bound (Sects. 4.3 and 7)	not preregistered
Overhead-ladder width study (Sect. 6.1)	prospectively registered (G16)

S4 Point-Prediction Diagnostics

Supporting diagnostic. Point-selection experiments show that exact-hit performance and budget regret diverge: the tested rules achieve low typical regret (median 0.8%) but insufficient tail reliability for uncertified deployment (p95 12–42%). These experiments motivate the geometric question and demonstrate why certification rather than prediction is the operational answer; the full-curve studies, not the rules, provide the evidence for the geometry itself (main paper, Sect. 6). *Secondary results.* The PRM supply function equals the static single-window supply delayed by one blackout (Prop. 1), and realizing PRM interfaces as evenly spaced windows over-allocates budget by a median 19.5% in the short-period regime, with an unexplained long-period reversal left open; system-level packing delimits the cost of rule-based over-splitting (Sect. S5).

The preceding sections establish the objective geometry from complete measured curves and provide a certificate independent of any point predictor. This section examines a narrower question about the point-selection experiments: how such rules should be evaluated when exact optimizer identity and practical budget loss are not equivalent. We therefore report both exact-hit performance and regret, with particular attention to tail reliability. These results

diagnose the distinction between point identification and decision quality; they do not establish the plateau, which Sect. 6 measures directly. We evaluate whether two natural low-cost feature families can recover the exact optimizer, and whether exact-hit performance reflects practical regret. Point identification is practically valuable only when optimizer identity is consequential, and effective only when the available features contain stable information about optimizer identity; Sect. 6 showed the first condition often fails, and these experiments test the second. The selection workflow of Sect. 7 does not depend on any predictor evaluated here—its initializer is the first-order surrogate of Sect. 4—so their outcome bears on the point-identification formulation, not on the workflow. Rules R1/R2/R1-FP are as defined in Sect. 4 and Sect. S1; outcomes are reported against their frozen criteria.

S4.1 The overhead–gap surrogate (H1, H4)

H1 (rule accuracy). Criterion: pooled median regret $< 2\%$ *and* p95 $< 10\%$. Outcome (Table S5): **rejected**. The best rule (R2 under EDF) achieves excellent typical accuracy—median regret 0.80% (CI [0.76, 0.89]), zero-regret rate 24.8%—but p95 = 12.8% (CI [11.9, 13.5]); R1 reaches 13.9% and the FP variant 42.0%. The rules are right on average and wrong in the tail.

Table S5 E1 pooled rule performance (150 replicates per cell; bootstrap 95% CIs: R2 median [0.76, 0.89], p95 [11.9, 13.5]).

Policy / rule	median regret	p95 regret	exactly optimal
EDF / R1	1.72%	13.9%	17.0%
EDF / R2	**0.80%**	12.8%	24.8%
FP / R1-FP	3.50%	42.0%	5.8%

H4 (scaling law). The Sect. S1 surrogate predicts $k^* \propto \delta^{-1/2}$. Preregistered criterion: fitted log–log slope of per-cell median k^* against δ within $[-0.6, -0.4]$ with $R^2 \geq 0.9$ in at least 3 of the 4 (regime $\times$ U) cells. Outcome: the preregistered band was met in **0 of 4 cells**, so H4 (the $\delta^{-1/2}$ scaling law) is rejected (Table S6, Fig. S2). The short-period-regime slopes are an order of magnitude too shallow, and in the long-period regime the per-cell median k^* is 1 at every δ level—the aggregate optimum simply does not move.

Table S6 E2 log–log OLS fits of per-cell median k^* against δ (seven half-decade levels, paired tasksets), against the preregistered band $[-0.6, -0.4]$ with $R^2 \geq 0.9$.

cell (regime, U)	slope	R^2	criterion
short-period, 0.3	-0.133	0.880	fail
short-period, 0.5	-0.117	0.872	fail
long-period, 0.3	0.000	—	fail
long-period, 0.5	0.000	—	fail

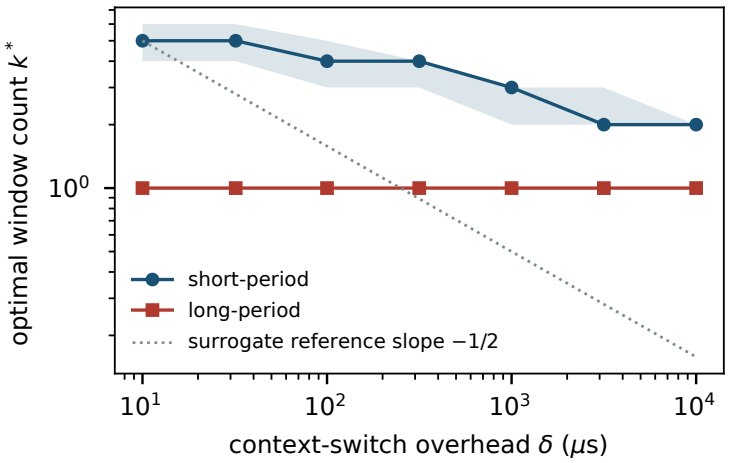

Fig. S2 Median optimal window count k^* versus overhead δ (interquartile bands), against the surrogate reference slope $-1/2$ (Sect. S1). The measured dependence is real but an order of magnitude shallower than the surrogate predicts; the long-period regime is flat at k^*=1

S4.2 The resonance hypothesis (round B)

A second, exploratory identifiability diagnostic (round B) tested whether period *alignment* between deadlines and the window period H/k determines k^*. Three alignment metrics (distance of $D_{\min}$ to the H/k grid; a nearest-rational divisor score; a utilization-weighted all-deadline score) were evaluated as k^* predictors against the failed rule and a random baseline. Outcome: **rejected**—top-3 hit rates 0.22–0.24 (short-period) against R1's 0.234, far below the preregistered criterion of a 2× improvement; no alignment metric separated itself from the rule baseline.

A variance decomposition on the paired E2 data resolves the apparent paradox of H4: within-taskset variance of k^* across δ levels accounts for **66%** of total variance in the short-period regime, so overhead does move each taskset's optimum—but in taskset-specific directions that cancel in aggregates. The aggregate flatness of H4 is thus an aggregation artifact rather than evidence that overhead is irrelevant, while "alignment dominates" fails independently on the prediction test. Individual trajectories are orderly rather than chaotic, yet predicting a taskset's trajectory class from a-priori features reached only AUC 0.72 against a preregistered bar of 0.75 (not met); the class taxonomy and its per-class shares are released with the code.

Open question. Can $k^*(\delta)$ trajectory classes be predicted substantially beyond AUC 0.72?

S5 Secondary Results: PRM Pessimism and System-Level Packing

Two result groups are interesting in their own right but are not steps of the central argument; we separate them so the main line—geometry, certificate, diagnostics, workflow—reads uninterrupted. Both were preregistered (E3/H2, E4/H3) and are reported against their frozen criteria.

S5.1 PRM pessimism (H2)

Sizing a partition through a PRM interface $\Gamma(H/k, \Theta)$ and realizing it as k windows overallocates relative to exact window sizing—consistent with the one-blackout delay identity of Proposition 1. Measured over 28,800 configurations (Table S7, Fig. S3): in the short-period

regime the median inflation is **19.5%** (p90 68.7% pooled) and grows with the blackout-to-deadline ratio (Spearman $\rho = 0.708$), as hypothesized; the effect is nearly invariant to δ, isolating the abstraction itself—not the overhead model—as the cause. In the long-period regime the correlation *reverses* ($\rho = -0.733$); a preregistered re-analysis using the absolute (rather than relative) penalty did not resolve the reversal ($\rho = -0.773$). H2 is therefore **rejected as stated**; the short-period-regime quantification stands as a finding, and the long-period-regime reversal is reported as an open question (Sect. 9.4).

Table S7 PRM budget inflation (%) over exact window sizing, per regime and overhead level (4,800 configurations per cell).

regime	δ=0	δ=100 μs	δ=1000 μs
short-period, median [p90]	19.7 [70.3]	19.6 [69.8]	19.1 [66.1]
long-period, median [p90]	3.9 [14.8]	3.9 [14.5]	3.5 [12.2]

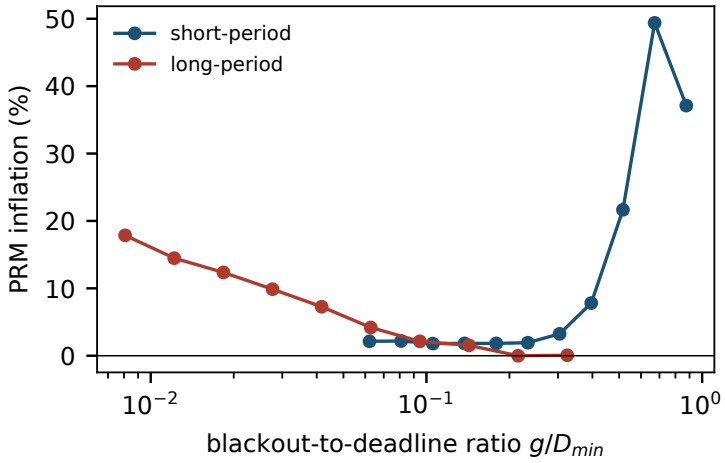

Fig. S3 PRM budget inflation versus blackout-to-deadline ratio: the short-period regime rises (median 19.5%); the long-period regime reverses—an open question

S5.2 System-level packing (H3)

E4 packs four-partition systems ($U_{\text{tot}} \in \{0.4, \ldots, 0.7\}$, $\delta \in \{100, 1000\}\,\mu s$, 150 systems per cell) under three pipelines: **SW** sizes every partition at k=1 (single-window); **PR** picks each partition's k by rule R2 then sizes exactly; **PX** sizes exhaustively over $k \in [1, 16]$. Windows are interleaved round-robin at even anchors with first-fit shift, and every packed system is re-verified exactly (mandatory). Table S8 gives the raw outcomes.

Against the frozen criterion (PR recovers $\geq 90\%$ of the SW−PX bandwidth gap at $\leq 1\%$ of PX's time in ≥ 6 of 8 cells) the target was met in **0 of 8 cells**; H3 is rejected—but the verdict is uninformative on both axes, so we report it for preregistration fidelity and read the unconditional quantities instead: the $\leq 1\%$ time bar is unsatisfiable by construction (one exact sizing per partition, observed 6–11%), and the closure metric conditions on the single-window pipeline's 1–5% of *verified* systems, an unrepresentatively easy subset (the statistic even turns negative in two cells). What the table does show: under the evaluated round-robin/first-fit packer, single-window configurations rarely pack, while the rule pipeline packs nearly all systems at 6–11% of exhaustive cost, carrying a δ-dependent bandwidth premium (5–10% at δ=100, 18–28% at δ=1000) with success dropping to 79% in the hardest cell—the narrow-set

Table S8 E4 packing outcomes per cell: packing success rates, mean allocated bandwidth among verified systems, and analysis-time ratio.

cell (U_{tot}, δ)	pack% SW/PR/PX	bandwidth PR	bandwidth PX	PR time (% of PX)
0.4, 100	5 / 100 / 100	0.459	0.418	11.2
0.4, 1000	5 / 100 / 100	0.650	0.509	6.2
0.5, 100	5 / 100 / 100	0.561	0.519	11.4
0.5, 1000	3 / 100 / 100	0.755	0.611	6.2
0.6, 100	2 / 100 / 100	0.663	0.620	11.4
0.6, 1000	1 / 100 / 100	0.857	0.715	6.3
0.7, 100	1 / 100 / 100	0.764	0.723	11.2
0.7, 1000	1 / 79 / 100	0.952	0.810	6.2

regime of Sect. 6 surfacing at system scale. These are heuristic-specific observations, not an impossibility result for single-window layouts.

Open question. What causes the long-period-regime reversal of PRM inflation ($\rho = -0.73$ against the predicted positive association, robust to the absolute-penalty re-analysis)? Proposition 1 fixes the absolute penalty at one blackout but not the relative behavior when $D_{\min} \gg \Pi$.