# Allocating Sovereign AI Compute: A Mechanism-Design Audit, an Impossibility, and a Credit-Market Response

Christian Hansen
Western Academy of Beijing
Beijing, China

## ABSTRACT

National and institutional allocators of scarce AI compute — the US National AI Research Resource (NAIRR), the UK AI Research Resource, the EU's EuroHPC AI Factories, Japan's ABCI — ration a fixed accelerator fleet among research teams whose valuations for compute are private and whose budgets are set by funders independently of that value. We make three contributions. First, we audit the allocation rule each program actually uses, sourced to program documentation: none elicits cardinal preferences through a formal mechanism; they ration by merit review, by review-threshold-plus-lottery, by first-in-first-out queueing, or by administered or subsidized fees. Second, we design and analyze a two-stage credit mechanism in which a planner mints an endowment as a function of a public-value assessment and then runs a Vickrey–Clarke–Groves (or clinching) auction denominated in those credits; we show the planner can always choose the endowment profile so the target allocation is credit-feasible, incentive-compatible, and individually rational at once. Third, we frame both against a budget-constrained impossibility that is *not* ours: specialized to compute, it is a corollary of Che and Gale (1998) and Dobzinski, Lavi and Nisan (2008, 2012). We work a fully divisible numerical instance, verified in code, on which the efficient truthful (VCG) allocation charges one truthful winner \$120,000 against a \$60,000 budget, and we compute VCG, a uniform posted price, a budget-capped clock auction, Dominant Resource Fairness, Probabilistic Serial, and the credit mechanism on that single instance, reporting realized welfare for each. We show the budget-capped clock has no instance-independent efficiency guarantee relative to first-best, and we confront the criterion question directly: because these budgets track institutional wealth rather than research value, a willingness-to-pay-efficient rule shifts compute toward better-resourced entrants, the stated opposite of what public programs exist to do. Under a distributionally weighted objective the recommendation reverses, and we state the reversal rather than smoothing it over.

## CCS CONCEPTS

• **Theory of computation → Algorithmic mechanism design**; *Algorithmic game theory*.

## KEYWORDS

mechanism design; budget constraints; compute governance; VCG; clinching auctions; fair division; AI infrastructure

## 1 INTRODUCTION

Every large allocator of AI compute is running a mechanism-design problem, named or not. A government dividing a cluster of H100s among university labs, a cloud provider pricing spot instances, and a frontier lab rationing internal GPU-hours across teams all face one question: how to allocate a scarce, heterogeneous resource among agents who know their own needs better than the allocator does.

The scarcity is real. On-demand H100 rental sits in a roughly \$4–6 per-GPU-hour band (Lambda quotes \$3.99–\$4.29/GPU-hr for H100 SXM depending on instance size; CoreWeave lists \$6.16/GPU-hr for an 8×H100 HGX on-demand instance, and \$2.46/GPU-hr on its spot tier), and next-generation B200 capacity commands a premium (Lambda \$6.69–\$6.99/GPU-hr; CoreWeave \$8.60/GPU-hr for 8×B200 HGX on-demand). Prices accessed 4 August 2026 from the vendors' own pricing pages; Figure 1 plots the on-demand headline figures.

Governments have responded with explicit allocation programs: the US NAIRR pilot, the UK's AI Research Resource built around Isambard-AI and Dawn, and the EU's AI Factories under EuroHPC. None solves its allocation problem with anything auction theory would recognize as an efficient, truthful mechanism. As Section 2 documents from each program's own materials, sovereign compute is rationed by application and review, by review-threshold-plus-lottery, by queueing, or by administered fee schedules, never by a preference-eliciting auction.

This paper starts from that observed gap and works in three directions. It first characterizes what the deployed programs do, precisely and from primary sources, because that characterization is itself absent from the literature. It then asks what a budget-aware market mechanism would add, and designs one. Finally it asks whether adding it is a good idea, and gives an answer that depends, sharply, on what these programs are for.

*Contributions.*

- A mechanism-design audit of every operating sovereign compute program (Section 2), sourced to program documentation with access dates, establishing that none elicits willingness to pay and characterizing the designed rules they use instead.
- A two-stage credit mechanism (Section 7) with an existence result: the planner can always choose the credit endowment so a target allocation is simultaneously efficient (in credit-value), dominant-strategy truthful, individually rational, and credit-feasible (Proposition 7.1). We place it honestly within the Hylland–Zeckhauser and Quota Marketplace lineage.
- The budget-constrained impossibility, stated as framing and correctly attributed to Che–Gale and Dobzinski–Lavi–Nisan (Proposition 4.1), with the desideratum that pins it down made explicit.
- A verified, fully worked numerical comparison of six mechanisms on one divisible instance (Sections 5–6), including a

proof that the budget-capped clock has no instance-independent efficiency bound (Proposition 6.3).
- A criterion analysis (Section 8) showing the policy recommendation reverses under a distributionally weighted objective, with the reversal stated.

One motivating fact organizes the theory. Research teams bidding for compute have private budgets that need not equal their private valuations. A grant-funded lab may value a large training run at far more than its grant permits it to pay; a well-capitalized startup may have deep pockets and modest need. This is definitionally true of any pool that mixes grant-funded academic labs, whose budgets are fixed by a funding agency, with commercial bidders, exactly the population NAIRR, the UK AIRR, and EuroHPC's AI Factories are built to serve together. Classical mechanism design has a name for what follows: when private budgets bind below private valuations, the efficient dominant-strategy-truthful mechanism (VCG) can demand a payment a truthful winner cannot make. Che and Gale (1998) established this for single-item auctions; Dobzinski, Lavi and Nisan (2008, 2012) extended it to multi-unit settings. We are not the first to observe the impossibility, and Section 4 states it as a corollary of that work rather than a result of ours.

The paper proceeds as follows. Section 2 audits the operating programs. Section 3 sets up the model, with the budget entering preferences directly. Section 4 states the impossibility as framing and fixes the desideratum that pins it down. Section 5 works the verified instance. Section 6 computes six mechanisms and reports welfare. Section 7 develops the credit mechanism. Section 8 confronts the criterion and revises the recommendation. Section 9 lists limitations; Section 10 concludes. Appendix A contains the verification code.

## 2 WHAT SOVEREIGN COMPUTE PROGRAMS ACTUALLY DO

The starting observation is empirical: no operating sovereign compute program elicits cardinal willingness to pay through a formal mechanism. This section establishes that claim program by program, from each program's published documentation, with access dates. We know of no prior mechanism-design treatment that characterizes these rules as mechanisms, and the characterization turns out to be more interesting than "ad hoc rationing" would suggest. Two of the programs use rules with recognizable formal content.

*NAIRR (United States).* The National Science Foundation leads the NAIRR pilot with federal partners. Access is by application and independent merit review across several tracks. Full proposals are peer-reviewed for technical merit and feasibility, then routed to per-resource matching committees under confidentiality agreement; a lighter Start-Up track grants smaller, single-resource, three-month allocations (on the order of a couple thousand GPU-hours), sent directly to resource providers rather than through full peer review, with decisions in about two weeks. There is no price and no lottery. (nairrpilot.org; /opportunities/allocations; /opportunities/startup-project; accessed 4 Aug 2026.)

*UK AIRR (United Kingdom).* Access to Isambard-AI and Dawn is by application through UKRI, assessed by experts against published criteria, with routes at different scales (Gateway ≈10k GPU-hours; Rapid Access ≈20k, aimed at SMEs; Innovator 50k–150k over six months). Two features are worth naming. First, allocation among applications that clear the merit threshold uses explicit partial randomization: "Resources will be allocated to the applications in the top tier as priority, using partial randomisation as required." Second, the Innovator route runs distributed peer review, in which applicants serve as reviewers of other applications. Compute is provided as a resource, with no cash grant. An SME may instead route through a separate Sovereign AI compute offer. (ukri.org Innovator/Rapid/Gateway route pages; gov.uk "AI Research Resource: AI open access call"; accessed 4 Aug 2026.)

The UK rule is, in mechanism terms, a merit-threshold filter followed by a lottery over the qualifying tier. That is closer to a designed random-assignment mechanism than to discretionary judgment, and Section 6 returns to it when discussing Probabilistic Serial.

*EuroHPC AI Factories (European Union).* Access is offered through EuroHPC Joint Undertaking calls, free of charge, in tiered modes distinguished by size rather than price. The load-bearing distinction is the selection rule, not a fee: Fast Lane (up to 50,000 GPU-hours, up to three months) uses first-in-first-out admission on a continuous deadline, while Large Scale (above 50,000 GPU-hours, up to one year) uses technical and peer-review evaluation on fixed cut-off dates. (eurohpc-ju.europa.eu AI-Factories access-modes, Fast Lane, and Large Scale pages; EuroHPC JU Access Policy, April 2025; accessed 4 Aug 2026.)

We flag a correction to a natural but wrong reading: EuroHPC AI Factory access is not "free below a GPU-hour threshold and paid above it." All the calls above are free; the 50,000 GPU-hour line separates a FIFO queue from a peer-review process. FIFO is itself a non-preference-eliciting rule, and it is the only sovereign allocation rule in our set that ignores merit as well as value.

*ABCI (Japan).* ABCI (operated by AIST) prices usage. Under the ABCI 3.0 tariff, users buy prepaid "ABCI Points" (1 point = 220 yen, tax included; minimum purchase 1,000 points) and spend them per node-hour: an 8-GPU rt_HF node costs 16 points/hour on the spot/on-demand rate and 24 points/hour reserved, with a discounted "Development Acceleration" tier at 7.5 points/hour spot. Points expire at fiscal-year end. Free use exists only by special institutional approval. (abci.ai/en/how_to_use/tariffs.html and /faq.html; accessed 4 Aug 2026.)

*IndiaAI Mission (India).* IndiaAI does not ration a government-owned cluster; it subsidizes a market. MeitY, through the IndiaAI IBD, empanelled private cloud and GPU providers via a Request for Empanelment (competitive lowest-bid selection), and pays a percentage subsidy on the per-GPU-hour rate; eligible users (academia, startups, MSMEs, researchers, government) consume compute at the subsidized rate through the IndiaAI compute portal. The subsidy is 40%, quoted against a high-end reference price near INR 150/GPU-hour, bringing effective cost below INR 100/GPU-hour; per-GPU-hour L1 bid prices are published per accelerator. Roughly 34,000 GPUs are empanelled. (pib.gov.in press releases

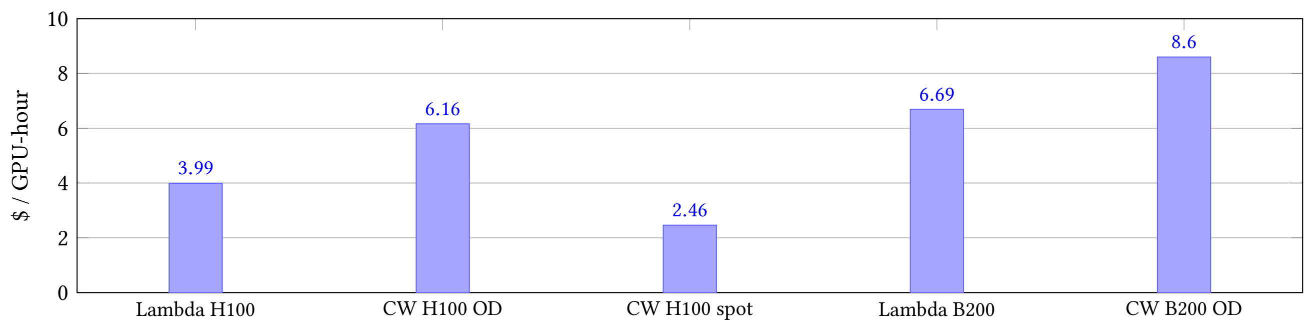


**Figure 1: On-demand GPU rental, per GPU-hour, headline configurations (vendor pricing pages, accessed 4 Aug 2026). Lambda figures are the 8× instance rate (H100 SXM ranges to $4.29 at 1×; B200 SXM6 to $6.99 at 1×). CoreWeave per-GPU figures are the published per-instance-hour rate divided by eight.**

PRID 2097709, 30 Jan 2025, and PRID 2132817, 30 May 2025; compute.indiaai.gov.in; accessed 4 Aug 2026.) This is administered, subsidized posted pricing over third-party capacity: a price mechanism, but with the subsidy, not a bidder's willingness to pay, doing the allocative work.

*A note on scope.* We restrict the audit to programs whose purpose is to allocate AI compute. Canada's annual Resource Allocation Competition, run by the Digital Research Alliance of Canada, is a merit-reviewed peer competition, but it is a general-purpose advanced-research-computing allocation across all scientific disciplines, not a dedicated AI-compute program; AI work competes within it like any other domain. We leave it out of the taxonomy rather than miscast it. (Canada's AI-specific analogue is the separate Sovereign AI Compute Strategy, which we did not verify here.)

Table 1 collects the rules. The column that matters is the last: the mechanism each program uses to choose among qualified claimants when demand exceeds supply.

Three of the five ration without money. The two that charge (ABCI, IndiaAI) post an administered or subsidized tariff rather than clearing a market. No program lets a claimant express, and pay according to, how much a given allocation is worth to it. That is the space the theory in Sections 4–7 occupies, and Section 8 asks whether the space should be filled.

## 3 THE MODEL

*Environment.* A planner controls a fixed stock of compute over one allocation window, described by $K$ resource types indexed $k = 1, \ldots, K$ (H100-hours, H200-hours, interconnect slots), with total quantity $Q_k$ of each. There are $N$ bidders indexed $i = 1, \ldots, N$. An allocation gives bidder $i$ a bundle $x_i = (x_{i1}, \ldots, x_{iK}) \in \mathbb{R}^K_+$, feasible when $\sum_i x_{ik} \le Q_k$ for every $k$.

*Preferences, with the budget as a primitive.* Bidder $i$ has a private valuation $v_i : \mathbb{R}^K_+ \to \mathbb{R}_+$ and a private budget $b_i > 0$. Utility is quasi-linear up to the budget and $-\infty$ beyond it:

$$u_i(x_i, t_i) = \begin{cases} v_i(x_i) - t_i, & t_i \le b_i, \\ -\infty, & t_i > b_i. \end{cases}$$

The infinite penalty is the modeling commitment that distinguishes this setting, and it follows Che and Gale (1998). A payment above budget is not merely disliked; it cannot be made. A grant-funded lab cannot conjure money past its grant, and a sovereign allocator with a fixed appropriation cannot collect what a bidder does not have. With this specification, budget-feasibility and individual rationality are linked properties of one preference structure rather than two independent constraints layered on quasi-linear utility, and the language of payments that "cannot be collected" describes the model instead of gesturing past it.

*First-best.* Absent budget or incentive constraints, the efficient allocation solves $\max \sum_i v_i(x_i)$ subject to $\sum_i x_{ik} \le Q_k$ for all $k$ and $x_i \ge 0$.

*Desiderata.* A mechanism is a pair $(x(\cdot), t(\cdot))$ from reported types to allocations and payments. We use five properties.

- (EFF) Allocative efficiency: the mechanism implements a first-best allocation.
- (DSIC) Dominant-strategy incentive compatibility: truthful reporting $(\hat{v}_i, \hat{b}_i) = (v_i, b_i)$ is dominant for every bidder.
- (IR) Individual rationality: $u_i \ge 0$ for every truthful bidder.
- (BF) Budget feasibility: $t_i \le b_i$ for every bidder, ex-post, with certainty.
- (WBB) Weak budget balance: the planner never subsidizes, $\sum_i t_i \ge 0$.

WBB is not a triviality, and Section 4 shows it is the property that makes the impossibility bite. A planner who could pay bidders to accept compute could escape the tension entirely; a sovereign allocator with a fixed appropriation cannot, so WBB is the right constraint here and we state it as a maintained hypothesis.

## 4 THE IMPOSSIBILITY, CORRECTLY ATTRIBUTED

The tension among these properties is real, and it is not new. Specialized to a multi-resource compute environment, it is a corollary of Che–Gale (1998) and Dobzinski–Lavi–Nisan (2008, 2012). We keep the name *compute allocation trilemma* for the three-way choice a compute allocator faces, while stating plainly that the underlying impossibility belongs to that prior literature. Our contribution here is expository precision: identifying which desideratum drives the result, and confirming the domain on which it applies to

**Table 1: Allocation rule in operating sovereign compute programs (program documentation, accessed 4 Aug 2026).**

| Program | Operator | Price? | Selection among qualified claimants |
|---|---|---|---|
| NAIRR (US) | NSF + partners | No | Merit peer review + matching committees; lighter Start-Up track |
| UK AIRR | UKRI / DSIT | No (resource) | Expert merit threshold, then partial randomization over top tier; distributed peer review |
| EuroHPC AI Factories | EuroHPC JU | No | FIFO (Fast Lane ≤50k GPU-h); peer review (Large Scale >50k) |
| ABCI 3.0 (Japan) | AIST | Yes (points) | Administered per-node-hour tariff; buy points, spend per use |
| IndiaAI Mission | MeitY / IndiaAI | Yes (subsidized) | Empanelled providers, 40% subsidy on posted per-GPU-hour rates |

our own example. Figure 2 places the mechanisms of Sections 6–7 on the three corners.

Proposition 4.1 (Five-property incompatibility, after Che–Gale and Dobzinski–Lavi–Nisan). *In the environment of Section 3 with $N \geq 2$, when budgets $b_i$ are private and can bind strictly below the valuations $v_i$ relevant to the efficient allocation, no mechanism satisfies EFF, DSIC, IR, BF, and WBB simultaneously.*

Proof. On quasi-linear domains without budget constraints, the Groves class is the set of mechanisms satisfying EFF and DSIC (Groves 1973), and on suitably rich domains it is the *only* such class (Green–Laffont 1977; Holmström 1979). A Groves mechanism charges $t_i = h_i(\theta_{-i}) - \sum_{j \neq i} v_j(x^*(\theta))$, where $x^*$ is the efficient allocation and $h_i$ any function of the others' reports. WBB together with IR pins the pivot down to the Clarke rule $h_i(\theta_{-i}) = \max_x \sum_{j \neq i} v_j(x)$: this is the smallest $h_i$ compatible with WBB in general, and any larger $h_i$ violates IR for some type. Under the Clarke rule, $t_i$ equals the externality bidder $i$ imposes on the others, a quantity determined by the others' values and the capacity constraint, not by $b_i$. Che and Gale (1998) construct a single-item instance in which this Clarke payment exceeds a truthful winner's budget; Dobzinski, Lavi and Nisan (2008, 2012) show the failure persists, and is generically larger, in multi-unit settings where the payment depends on the combinatorial structure of competing demands. Hence within the EFF+DSIC+IR+WBB class the Clarke payment can violate BF, and no mechanism attains all five. □

*Why WBB, not "triviality," closes the argument.* A five-property statement is necessary because EFF, DSIC, IR, BF are jointly satisfiable inside this very framework. Take any Groves mechanism with pivot $h_i \equiv 0$, so $t_i = -\sum_{j \neq i} v_j(x^*(\theta)) \leq 0$. It is efficient and dominant-strategy truthful because it is Groves; individually rational because every winner is paid rather than charged; and budget-feasible because every payment is non-positive and therefore below every $b_i > 0$. All four hold at once. What fails is WBB: the planner pays bidders. So the constraint that forces the Clarke rule, and with it the impossibility, is precisely the one an earlier treatment dismissed as automatically satisfied. It is the operative hypothesis, and a sovereign allocator with a fixed appropriation is exactly the setting in which WBB cannot be relaxed.

*A precision about uniqueness.* The Green–Laffont/Holmström results characterize the Groves *class*, not the Clarke mechanism alone; it is WBB and IR that select Clarke within the class. This matters because the characterization's domain conditions (smoothness or richness of the type space) are not automatic. Our numerical instance in Section 5 uses linear-with-cap valuations over divisible goods with a single binding capacity constraint. On that domain the efficient rule is greedy-by-rate, the Groves payments are well defined, and the Clarke payment for the relevant bidder exceeds its budget by an explicit factor, so the impossibility applies to the specific domain we exhibit and not only in the abstract. We verify the payment computation, and the domain's regularity, in code (Appendix A).

We name the practical form of the choice the trilemma: an allocator can keep EFF and DSIC and accept that some truthful winners are billed more than they can pay (undeployable when the planner cannot collect the money); or keep BF and DSIC and lose efficiency, as Section 6 shows constructively; or keep BF and EFF and give bidders a reason to misreport. Sections 6 and 7 place one mechanism at each corner.

## 5 A VERIFIED NUMERICAL INSTANCE

The instance below is a constructed illustration, not empirical data. Every quantity in this section and the next was recomputed by brute force and by closed form in the script of Appendix A; the two agree.

*Setup.* Two resource types, $Q = (1{,}000$ H100-hrs, $500$ H200-hrs$)$, both divisible. Five labs have linear valuations in H100-hours up to a per-lab capacity cap, plus a fixed complementary H200-hour requirement and a private budget (Table 2). Total H200 demand is $0 + 100 + 0 + 200 + 150 = 450 \leq 500$, so the H200 constraint never binds and the problem reduces to a single-resource allocation of H100-hours. Divisibility is deliberate: it makes the gross-substitutes condition of Section 6 hold, lets Dominant Resource Fairness and Probabilistic Serial apply without ad hoc combinatorial surgery, and lets one instance carry all six mechanisms so the comparison is like-for-like.

*First-best.* Greedy by value-rate fills D (150, the cap), then C (400), then B (250), then E (200), exhausting 1,000 H100-hours and excluding A. Surplus is $1{,}300 \cdot 150 + 700 \cdot 400 + 600 \cdot 250 + 450 \cdot 200 =$ \$715,000. A is excluded not because it has the lowest per-unit value

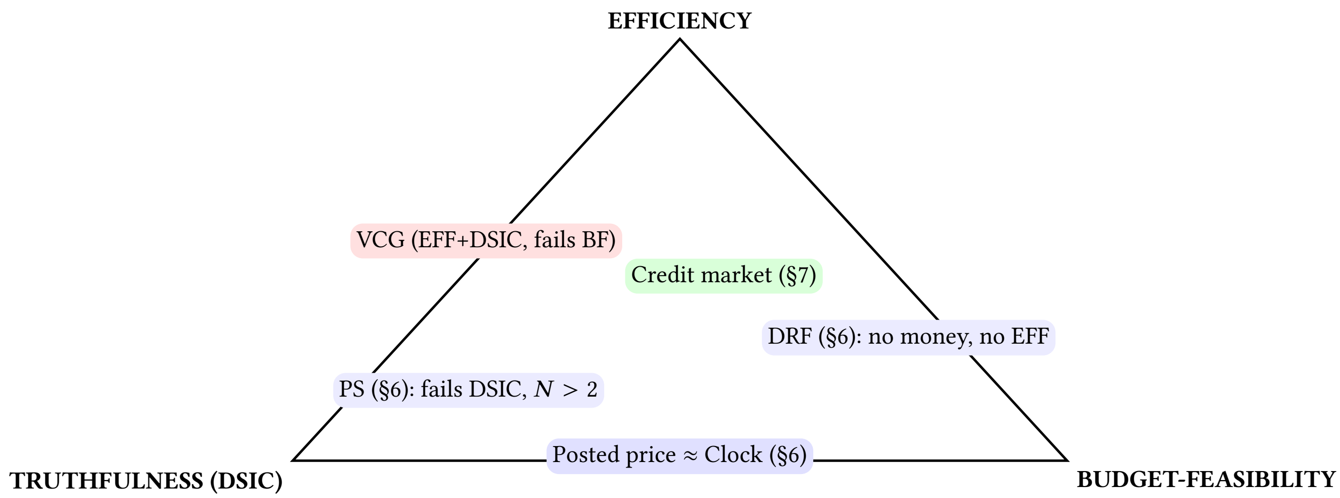


**Figure 2: The compute allocation trilemma. No mechanism reaches all three vertices at once (Proposition 4.1). Positions are qualitative; the credit market of Section 7 escapes the corners by denominating payments in planner-minted credits, at the cost of relocating the value judgment to the endowment stage.**

**Table 2: The instance (constructed illustration).**

| Lab | Value $/H100-hr | Cap (H100-hrs) | H200-hrs required | Budget ($) |
|---|---:|---:|---:|---:|
| A | 400 | 300 | 0 | 90,000 |
| B | 600 | 250 | 100 | 150,000 |
| C | 700 | 400 | 0 | 60,000 |
| D | 1,300 | 150 | 200 | 180,000 |
| E | 450 | 200 | 150 | 90,000 |

(it does not), but because greedy-by-rate exhausts capacity before reaching it. H200 usage among winners is 450, within the 500 available.

*VCG (Clarke) payments.* For each winner, the Clarke payment is the value the others could obtain without that winner, minus the value the others obtain with it present. Table 3 gives all four winners.

VCG revenue is $360,000. Lab C exhibits the impossibility in the concrete: the efficient, dominant-strategy-truthful mechanism asks C for $120,000, twice its $60,000 budget and most of its gross value. The planner cannot quietly collect a partial payment without either reallocating C's capacity (losing efficiency) or letting C shade its report (losing truthfulness, and inviting every other bidder to do the same).

## 6 SIX MECHANISMS ON ONE INSTANCE

We now run every mechanism on the Section 5 instance and report realized welfare. Table 4 is the summary and Figure 3 the picture; the subsections give the derivations and the incentive analysis. "Surplus" is realized true value $\sum_i v_i(x_i)$; "% of first-best" divides by 715,000.

### 6.1 Unconstrained VCG and the uniform posted price

VCG is in the table for contrast: it attains first-best (715,000) and is truthful, but it is not deployable here, because its bill for C cannot be collected. That is the whole motivation.

A uniform posted price $p$ per H100-hour is the simplest budget-feasible, strategyproof benchmark, and it is exactly the corner the clock is engineered for. At price $p$ each lab buys $\min(\text{cap}_i,\ b_i/p)$ if its value-rate exceeds $p$, and nothing otherwise. Sweeping $p$ (Appendix A), the welfare-maximizing uniform price is $p^\star = \$375/\text{H100-hr}$, which clears the market exactly: A buys 240, B 250, C 160, D 150, E 200, totalling 1,000, with realized surplus $643,000 and revenue $375,000. Posted prices satisfy DSIC (a take-it-or-leave-it price is not manipulable by the buyer), IR, BF, and WBB, and fail only EFF. This corrects an earlier table that marked administered pricing "N/A" under DSIC: posted-price selling is the canonical strategyproof, budget-feasible mechanism, and it belongs in the DSIC column with a checkmark.

### 6.2 The budget-capped combinatorial clock

*Design.* We adapt the ascending clock auction (Ausubel–Milgrom 2002; Cramton 2013) with one addition: each bidder's demand is mechanically capped at what its stated budget affords at current prices. The auctioneer posts a price vector, starting low. Each round, bidder $i$ demands the bundle maximizing $v_i(x) - p \cdot x$ subject to $p \cdot x \le b_i$, "straightforward bidding under a budget cap." Prices on over-demanded resources rise by an increment $\delta$; the auction stops when no resource is over-demanded.

Proposition 6.1 (Budget feasibility by construction). *Under the budget-capped clock, $t_i \le b_i$ for every bidder and every realized outcome.*

**Table 3: Clarke payments, entire winner set (verified, Appendix A).**

| Winner | Others' best without $i$ | Others' value with $i$ | Clarke payment | Budget | Feasible? |
|---|---|---|---|---|---|
| B | 665,000 | 565,000 | 100,000 | 150,000 | Yes |
| C | 555,000 | 435,000 | **120,000** | **60,000** | **No — 2× budget** |
| D | 580,000 | 520,000 | 60,000 | 180,000 | Yes |
| E | 705,000 | 625,000 | 80,000 | 90,000 | Yes |

**Table 4: Mechanisms on the Section 5 instance (all figures verified, Appendix A). *VCG revenue is nominal; it is not collectible, because C cannot pay. †Approximate and conditional; see §6.3. ‡Against the planner's chosen objective; if that objective is willingness to pay, the credit mechanism implements the 715,000 allocation (§7).**

| Mechanism | Surplus | % FB | Revenue | Budget viol.? | EFF | DSIC | IR | BF | WBB |
|---|---|---|---|---|---|---|---|---|---|
| Unconstrained VCG | 715,000 | 100.0 | 360,000* | **Yes (C)** | ✓ | ✓ | ✓ | × | ✓ |
| Uniform posted price ($375) | 643,000 | 89.9 | 375,000 | No | × | ✓ | ✓ | ✓ | ✓ |
| Budget-capped clock ($375) | 642,999 | 89.9 | 375,000 | No | × | approx.† | ✓ | ✓ | ✓ |
| Dominant Resource Fairness | 615,464 | 86.1 | 0 | No | × | ✓ | ✓ | ✓ | ✓ |
| Probabilistic Serial | 653,333 | 91.4 | 0 | No | × | × ($N>2$) | ✓ | ✓ | ✓ |
| Credit market (VCG-in-credits) | 715,000‡ | 100.0‡ | 0 (real) | No | ✓‡ | ✓ | ✓ | ✓(cr.) | ×(real) |

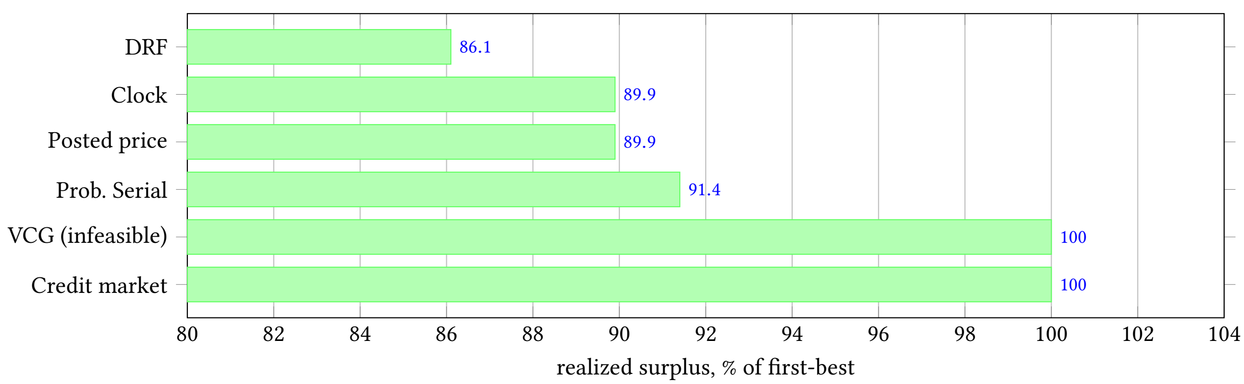


**Figure 3: Realized surplus as a percent of first-best on the Section 5 instance. VCG and the credit market both reach 100%, but VCG is budget-infeasible (uncollectible bill for C) and the credit market attains it only in minted credits, not real revenue. Among deployable money mechanisms the clock ties the optimal posted price at 89.9%.**

No bidder can demand a bundle costing more than its stated budget at current prices, so the final payment (clearing price times quantity) cannot exceed it. This is the mechanism's one unconditional guarantee.

*Realized outcome.* Only H100 binds, so the H100 clock runs alone (H200 stays at reserve). Straightforward budget-capped demand at price $p$ is $\min(\text{cap}_i, b_i/p)$ for labs with rate above $p$. Aggregate demand equals supply at $p^\star = \$375$ (closed form: in the region $p \in (300, 400)$ demand is $150{,}000/p + 600$, set to 1,000). The allocation is A 240, B 250, C 160, D 150, E 200, with payments A $90,000, B $93,750, C $60,000, D $56,250, E $75,000, each within budget (A and C bind exactly). Realized surplus is $642,999, which is 89.9% of first-best, and full capacity is sold.

*What the clock demonstrates.* First-best gives C 400 H100-hours; the clock gives C 160, because at $375 C's $60,000 budget affords only 160. The 240 hours C loses go substantially to A, which receives 240 in the clock versus 0 at first-best. Capacity moves from a $700/hr bidder to a $400/hr bidder because the higher-value bidder's budget binds. That is the efficiency cost of budget-feasibility, made visible and ex-ante predictable, in place of VCG's habit of discovering infeasibility only after printing an unpayable bill.

*The clock does not beat a well-chosen posted price here.* On this instance the clock and the optimal uniform posted price coincide, both at $375, both at $643,000. This is not a coincidence of the numbers: with a single scarce divisible good and straightforward budget-capped bidding, the clock is a tâtonnement that finds the uniform competitive price, so its allocation equals the market-clearing posted-price allocation. The clock's advantage over the posted price is informational, not allocative: it locates the clearing price without the planner knowing the demand curve in advance.

We report this rather than claim a welfare edge the instance does not support.

### 6.3 Incentives in the clock

An earlier treatment asserted that under-demanding "can never lower the eventual clearing price." That is false, and the failure has a standard name.

*Demand reduction.* Reducing one's own demand reduces aggregate demand, which can terminate the clock earlier and at a lower price on the units one still wins (Ausubel–Cramton 2002/2014). Straightforward bidding is therefore not exactly dominant. The correct statement is an $\varepsilon$-dominance claim.

Proposition 6.2 ($\varepsilon$-dominance under gross substitutes). *If valuations satisfy gross substitutes (Kelso–Crawford 1982), then in the budget-capped clock with increment $\delta$, straightforward budget-capped bidding is $\varepsilon$-dominant with $\varepsilon = O(\delta\,\bar{q})$, where $\bar{q}$ bounds the quantity a single bidder can be allocated. As $\delta \to 0$, $\varepsilon \to 0$.*

The gain from any demand-reduction deviation is at most one price increment on the affected units, because a unilateral one-increment-early stop changes the clearing price by at most $\delta$; the surplus swing on a bidder's own allocation is then bounded by $\delta\bar{q}$. Under gross substitutes no deviation does better than this bound, by the convergence machinery for ascending auctions (Gul–Stacchetti 1999; Ausubel 2004). We state the bound rather than assert exact truthfulness. Bundles with strong complementarities violate gross substitutes, and the guarantee weakens accordingly; that case is left to the approximation program noted in Section 9.

*Budget misreporting is the sharper problem.* Because the mechanism caps demand at the *stated* budget $\hat{b}_i$, understating one's budget is a direct instrument for lowering one's own effective demand and, through it, the clearing price. This is the failure mode the Che–Gale/DLN literature is about, and valuation-truthfulness results say nothing about it. On the instance, a bidder whose budget does not bind (D, at \$56,250 paid against \$180,000) has no profitable budget deviation, but a bidder whose budget binds and who has market power over the clearing price can gain: reporting $\hat{b}_i < b_i$ can lower $p^\star$ enough that the saving on inframarginal units exceeds the value of units shed. The honest characterization is that the budget-capped clock is *not* strategyproof in budgets; it is budget-feasible and approximately valuation-truthful under substitutes, and manipulable in the budget report.

*The benchmark the design should be measured against: clinching.* The positive half of Dobzinski–Lavi–Nisan is the adaptive clinching auction, which achieves Pareto efficiency, IR, budget feasibility, and truthfulness when budgets are *public*. Their impossibility is exactly that this cannot survive making budgets private. Goel, Mirrokni and Paes Leme (2015) extend clinching to polyhedral (heterogeneous, multi-resource, budget-constrained) environments, this paper's setting; two of those authors (Paes Leme, Mirrokni) are also authors of the deployed Quota Marketplace system discussed in Section 7. The budget-capped clock is the cruder, more transparent cousin of clinching: it buys budget-feasibility unconditionally and pays for it in exact truthfulness. Where budgets can credibly be treated as public (a funder that sets and observes grant sizes), clinching dominates it, and Section 7 builds on exactly that observation.

*Termination and idle capacity.* An ascending clock that stops when demand no longer exceeds supply can leave capacity unsold when demands are lumpy, and idle GPUs on a sunk-cost sovereign cluster are pure waste. On this divisible instance the clock clears exactly, so idle capacity is zero. With a discrete price grid of increment $\delta$, the last step can overshoot; the residual is at most the demand drop across one increment, bounded by $\sum_i b_i/(p^\star(p^\star - \delta))$ H100-hours, which vanishes as $\delta \to 0$. For the terminal round we specify proportional rationing at the terminal price among bidders still demanding, with free disposal only if no bidder demands the residual at that price. On this instance the rule never triggers.

### 6.4 Dominant Resource Fairness

DRF (Ghodsi et al. 2011) equalizes each bidder's dominant share (its largest single-resource fraction) subject to feasibility, using no money and no cardinal values. Over the divisible demand rays (each lab's cap and H200 requirement as its bundle shape), the equalized dominant share is $s = 0.2474$, with H100 binding. The allocation gives A 247.4, B 247.4, C 247.4, D 92.8, E 165.0 H100-hours, for realized surplus \$615,464, 86.1% of first-best.

DRF's cost is visible in D. D has the highest value-rate (1,300) and the smallest cap (150), so first-best satisfies it fully; DRF, blind to value, gives D only 92.8 hours while handing low-value A a large 247.4-hour slice. DRF cannot tell a bidder that mildly prefers more compute from one desperate for it, by construction. Its offsetting virtue is real: it is strategyproof, envy-free, and gives every claimant a defensible share, close to what the merit-rationed programs of Section 2 are reaching for.

### 6.5 Probabilistic Serial

Where ordinal preferences are more credible than dollar figures, Probabilistic Serial (Bogomolnaia–Moulin 2001) has agents "eat" probability shares of their most-preferred remaining resource at a uniform rate. Over the single scarce divisible good here (H100, with H200 slack), uniform eating reduces to equal division capped at each lab's demand. The resulting shares are A 216.7, B 216.7, C 216.7, D 150, E 200, for expected surplus \$653,333, 91.4% of first-best. (The assignment is deterministic because the good is divisible; the "random" in random assignment collapses to fractional shares.)

PS scores highest among the money-free rules on this instance, and higher than the clock, because equal division happens to satisfy the high-value, low-cap bidder D in full while the clock throttles the high-value, budget-poor bidder C. This is instance-specific and we do not generalize it: PS is not strategyproof once $N > 2$ (Bogomolnaia–Moulin 2001), and on instances where the high-value bidders also have large caps, equal division would strand more value than the clock does. We report the number because it is what the instance produces, not because it recommends PS.

The UK AIRR rule of Section 2 (merit threshold, then partial randomization over the qualifying tier) is the closest deployed relative of a lottery mechanism, and it is more principled than ad hoc peer review: it is a designed randomization, though over a merit-filtered set rather than over PS-style ordinal reports, and it carries none of PS's ordinal-efficiency guarantee.

### 6.6 No instance-independent efficiency bound for the clock

A mechanism paper owes a worst-case statement. For the budget-capped clock, the honest one is negative.

Proposition 6.3 (Unbounded efficiency loss). *For every $\eta > 0$ there is a single-good instance on which the budget-capped clock's realized welfare is below $\eta$ times first-best.*

Proof. Take capacity 1 (divisible), bidder 1 with value $V$ per unit and budget $\beta$, bidder 2 with value $v < V$ per unit and a large budget. First-best gives everything to bidder 1: welfare $V$. In the clock, bidder 2 demands the whole unit at any price below $v$, so the price rises to $v$, at which bidder 2 exits and bidder 1, budget-capped, wins $\beta/v$ units. Realized welfare is $V\beta/v$, and the ratio $\beta/v \to 0$ as $\beta \to 0$. Choosing $\beta < \eta v$ gives the claim; no constant-factor guarantee exists. □

(Verified numerically in Appendix A: with $V$=$10^6$, $v$=$10^3$, budget $\beta$=100 gives ratio 0.1, falling with $\beta$.)

*A conditional bound.* The loss is controlled when budgets are not too small relative to efficient prices. If every bidder's budget covers at least a $\gamma \in (0, 1]$ fraction of the payment it would owe at the efficient allocation's competitive prices, then no budget binds below $\gamma$ of its clearing cost, and the clock's realized welfare is at least $\gamma$ times first-best on single-good instances. The bound degrades gracefully to the unbounded case as $\gamma \to 0$. We state the assumption because without it there is nothing to state.

## 7 THE CREDIT MECHANISM, AND WHAT IT DOES AND DOES NOT ADD

The mechanisms so far treat budgets $b_i$ as exogenous private primitives. For the programs of Section 2 that is the wrong model. NAIRR issues cloud credits; UKRI sets the grant and rations the cluster; the allocator and the funder are the same institution or are tightly coupled. The budget is a policy instrument, not private information the planner must live with.

*Design (two stages). Stage one:* the planner assigns each bidder a credit endowment $b_i^{cr}$ as a function of a public-value or peer-review assessment. *Stage two:* the planner runs VCG (or, better, the adaptive clinching auction) denominated in credits rather than currency. Credits are minted by the planner and are not convertible to money.

Proposition 7.1 (The endowment can always make the target allocation feasible). *Let $x^\dagger$ be any allocation the planner wishes to implement, and let $\pi_i$ be the Clarke payment of bidder i under VCG-in-credits when $x^\dagger$ is the efficient allocation for the reported credit-values. If the planner sets $b_i^{cr} \geq \pi_i$ for every winner, then VCG-in-credits implements $x^\dagger$ and satisfies EFF (with respect to credit-values), DSIC, IR, and credit-BF simultaneously.*

The trilemma of Section 4 dissolves because BF is now slack by construction: the planner mints whatever credits the Clarke payments require. On the Section 5 instance, the efficient allocation's Clarke payments are B 100k, C 120k, D 60k, E 80k credits, so any endowment with $b_C^{cr} \geq 120{,}000$ (against C's real \$60,000 budget) and the others above their Clarke prices makes first-best implementable, truthful, and credit-feasible at once (Appendix A).

*This is not a new mechanism, and saying so is the point.* The core idea — a planner mints an endogenous credit budget and runs a market denominated in it — is the design deployed at scale in Quota Marketplace (Sivan, Paes Leme, Tiuca, McFarlane, Gkatzelis, Mehta, Hassas Yeganeh, Mirrokni & Vahdat, OSDI '26), where pool administrators mint credits at configured income rates and a dynamic-price market clears against them roughly once a minute. It is also the pseudo-market of Hylland and Zeckhauser (1979), who endow agents with artificial money to clear a random-assignment problem in equilibrium; an earlier treatment cited Hylland–Zeckhauser and then set it aside on computational grounds without noticing it answers the policy question. What Section 7 adds to that lineage is narrow and specific, and we claim only it:

(1) *The incentive analysis Quota Marketplace omits.* Quota Marketplace proves Pareto efficiency and approximate max-min fairness for its dynamic-price market, but makes no truthfulness claim and does not engage the budget-constrained impossibility. Denominating VCG or the adaptive clinching auction in the minted credits, rather than a price-adjustment market, is what buys DSIC in Proposition 7.1. The endowment neutralizes the impossibility (BF is slack), and clinching-in-credits then recovers truthfulness that a dynamic-price market does not guarantee.

(2) *The transposition to the sovereign setting, with the value judgment made explicit.* Quota Marketplace allocates inside one firm, where the minting rule is an internal-priority decision. In a sovereign program the minting rule *is* the public-value judgment, moved out of an implicit funding system and into an auditable stage-one endowment.

*Where the difficulty relocates, honestly.* Proposition 7.1 does not create surplus; it relabels the constraint. Three things move rather than vanish.

First, the credit market maximizes *credit-weighted* value, not willingness to pay. If the planner endows credits in proportion to peer-review score, the stage-two efficient allocation maximizes score-weighted value, which equals the 715,000 willingness-to-pay optimum only if scores are proportional to dollar valuations. Choosing the endowment is choosing the objective.

Second, real-money WBB fails, or rather becomes vacuous. The planner collects no dollars; it collects credits it printed. Credit-BF is a real constraint on the *allocation* (a bidder cannot spend credits it was not given) but not a revenue constraint. This is the same non-zero-sum property Quota Marketplace relies on, and it is why the credit design escapes the Section 4 impossibility that binds any real-money mechanism.

Third, stage one is itself a reporting game. Peer review can be gamed by applicants and is noisy across reviewers; a team that learns credits track a rubric will write to the rubric. The credit mechanism converts a truthful stage-two auction into a mechanism whose overall incentive properties are only as good as the stage-one assessment's resistance to manipulation, an open problem in its own literature (Bhattacharya, Conitzer, Munagala & Xia 2010 on incentive-compatible budget elicitation; Lavi & May 2012 on the incompatibility of strategyproofness and Pareto-optimality even with public budgets). We claim the trilemma dissolves at stage two, and we claim nothing about stage one beyond flagging that the hard part now lives there.

## 8 WHICH WELFARE, AND THE RECOMMENDATION THAT FOLLOWS

The mechanisms of Sections 6–7 all maximize some version of $\sum_i v_i(x_i)$. Whether that is the right objective is the question a policy paper cannot skip, and answering it reverses the recommendation an earlier draft reached.

*The circularity.* The paper's own premise is that budgets are set by funders independently of research value, and budgets correlate with institutional wealth. A willingness-to-pay-efficient rule therefore moves compute toward better-resourced bidders. But broadening access beyond the best-resourced labs is the stated reason NAIRR, the UK AIRR, and EuroHPC exist. Recommending "capture the efficiency gains from a priced pool" while asserting "budgets do not track value" is close to self-contradiction, and it is the first objection a referee will raise. Pre-empting it is worth more than the recommendation.

*What the programs say they are for.* NAIRR's charter is to broaden participation across institutions that lack compute; the UK AIRR and EuroHPC AI Factories are framed around access for academics and SMEs, not revenue or throughput. Their objective is closer to capability-broadening and research-output-per-dollar-of-public-money than to willingness-to-pay efficiency. Under that objective, a bidder's dollar budget is not a signal of social value; it is closer to noise correlated with the very inequality the program is chartered to offset.

*Distributionally weighted welfare.* Replace $\sum_i v_i(x_i)$ with $\sum_i \omega_i v_i(x_i)$, where $\omega_i$ down-weights well-resourced bidders. Under such weights the ranking in Table 4 changes. The value-blind rules (DRF, PS), which spread capacity across claimants, do relatively better; the priced rules (posted price, clock), which concentrate capacity on whoever can pay, do relatively worse, because the bidders they favor are the ones $\omega$ penalizes. The credit mechanism spans the range: setting $b_i^{cr}$ to encode $\omega$ makes its stage-two market implement the distributionally weighted optimum directly. That is the strongest argument for the credit design, and it is an argument for using it to *pursue* a public objective, not to override one with willingness to pay.

*Resale.* A secondary market would substantially undo the budget distortion. If a cash-poor high-value lab could sell part of its allocation to a cash-rich low-value firm, the two would trade toward the willingness-to-pay-efficient allocation regardless of the initial rule, and the budget constraint would stop binding on the final allocation. Sovereign programs prohibit resale, and the prohibition is a deliberate choice: allowing resale would convert a capability-broadening subsidy into a cash transfer to whichever eligible lab is best placed to flip its allocation, defeating the distributional purpose. The prohibition is coherent precisely because these programs do not hold willingness-to-pay efficiency as their objective. That is one more reason to doubt that a priced pool belongs in a public program.

*Revised recommendation.* For firm-internal allocation, where the objective is to route scarce compute to its highest-value use inside one budget-setting entity, a budget-aware market is the right tool, and Quota Marketplace is evidence that it works at scale. For sovereign programs whose charter is capability-broadening, merit-based and randomized rationing of the kind already deployed is defensible on its own terms, and the efficiency it "leaves on the table" is measured against an objective those programs do not hold. The credit mechanism of Section 7 is the bridge worth building: it lets a public allocator pursue an explicit, auditable distributional objective through a truthful market, rather than either an implicit wealth-based one (a dollar-priced pool) or an unexamined one (discretionary review). We do not recommend a dollar-priced pool for public compute. We recommend that if a public allocator wants a market, it be a credit market whose endowment encodes the program's actual objective, and that the objective be written down.

## 9 LIMITATIONS

The impossibility of Section 4 is prior work, specialized; the expository and domain-checking contribution is ours, the theorem is not. The numerical instance is a constructed illustration, not data. The model is a single static window; real allocation is repeated, which introduces reputation effects and cross-round strategic behavior we do not model. We assume private, independent valuations; consortium-level coordination can undermine both the clock's and DRF's guarantees. Every value-eliciting mechanism here assumes the planner can verify ex-post that compute was used consistently with the bid, a measurement problem for training workloads that admits a usage-misreporting manipulation distinct from the valuation- and budget-misreporting we do analyze. The gross-substitutes condition behind the clock's incentive guarantee fails for strongly complementary bundles, and the right treatment there is an approximation guarantee rather than an exact incentive result. The credit mechanism's stage-one assessment is a reporting game whose incentive properties we do not resolve. These are the load-bearing gaps; we state them once, here.

## 10 CONCLUSION

Sovereign AI-compute allocators ration a scarce, privately valued, heterogeneous resource, and none of them uses a preference-eliciting market to do it. We characterized what they use instead, from their own documentation, and found designed rules (merit-plus-lottery, FIFO, administered and subsidized tariffs) rather than the void the phrase "ad hoc" suggests. We placed those rules against a budget-constrained impossibility that is not new, and attributed it. On a single verified instance we showed what each candidate mechanism does, including that the efficient truthful rule bills a truthful winner twice its budget, that the budget-capped clock recovers feasibility at a measured 10% welfare cost while merely tying a well-chosen posted price, and that no instance-independent efficiency bound for the clock exists. The one design that dissolves the trilemma, a market in planner-minted credits, is already deployed inside a firm and already known in the pseudo-market literature; its value for public programs is that it can carry an explicit distributional objective through a truthful auction. Whether public compute should be allocated by any market at all turns on a criterion question, and under the objective these programs actually hold, the answer may be no. We think that answer is more useful than a tidier one.

## REFERENCES


Ausubel, L. M. (2004). An efficient ascending-bid auction for multiple objects. *American Economic Review*, 94(5), 1452–1475.

Ausubel, L. M., & Cramton, P. (2002). *Demand reduction and inefficiency in multi-unit auctions*. Working Paper, University of Maryland. [Peer-reviewed successor: Ausubel, Cramton, Pycia, Rostek & Weretka (2014), *Review of Economic Studies*, 81(4), 1366–1400.]

Ausubel, L. M., & Milgrom, P. (2002). Ascending auctions with package bidding. *Frontiers of Theoretical Economics*, 1(1), Article 1.

Bhattacharya, S., Conitzer, V., Munagala, K., & Xia, L. (2010). Incentive compatible budget elicitation in multi-unit auctions. *Proc. 21st ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 554–572.

Bhattacharya, S., Goel, G., Gollapudi, S., & Munagala, K. (2010). Budget constrained auctions with heterogeneous items. *Proc. 42nd ACM Symposium on Theory of Computing (STOC)*, 379–387.

Bogomolnaia, A., & Moulin, H. (2001). A new solution to the random assignment problem. *Journal of Economic Theory*, 100(2), 295–328.

Che, Y.-K., & Gale, I. (1998). Standard auctions with financially constrained bidders. *Review of Economic Studies*, 65(1), 1–21.

Clarke, E. H. (1971). Multipart pricing of public goods. *Public Choice*, 11(1), 17–33.

Cramton, P. (2013). Spectrum auction design. *Review of Industrial Organization*, 42(2), 161–190.

Dobzinski, S., Lavi, R., & Nisan, N. (2012). Multi-unit auctions with budget limits. *Games and Economic Behavior*, 74(2), 486–503. [Conference version: *FOCS 2008*, 260–269.]

Ghodsi, A., Zaharia, M., Hindman, B., Konwinski, A., Shenker, S., & Stoica, I. (2011). Dominant resource fairness: fair allocation of multiple resource types. *Proc. 8th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 323–336.

Goel, G., Mirrokni, V. S., & Paes Leme, R. (2015). Polyhedral clinching auctions and the AdWords polytope. *Journal of the ACM*, 62(3), Article 18. [Conference version: *STOC 2012*, 107–122.]

Green, J., & Laffont, J.-J. (1977). Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica*, 45(2), 427–438.

Green, J., & Laffont, J.-J. (1979). *Incentives in Public Decision-Making*. North-Holland.

Groves, T. (1973). Incentives in teams. *Econometrica*, 41(4), 617–631.

Gul, F., & Stacchetti, E. (1999). Walrasian equilibrium with gross substitutes. *Journal of Economic Theory*, 87(1), 95–124.

Holmström, B. (1979). Groves' scheme on restricted domains. *Econometrica*, 47(5), 1137–1144.

Hylland, A., & Zeckhauser, R. (1979). The efficient allocation of individuals to positions. *Journal of Political Economy*, 87(2), 293–314.

Kelso, A. S., & Crawford, V. P. (1982). Job matching, coalition formation, and gross substitutes. *Econometrica*, 50(6), 1483–1504.

Lavi, R., & May, M. (2012). A note on the incompatibility of strategy-proofness and Pareto-optimality in quasi-linear settings with public budgets. *Economics Letters*, 115(1), 100–103.

Mahajan, K., Balasubramanian, A., Singhvi, A., Venkataraman, S., Akella, A., Phanishayee, A., & Chawla, S. (2020). Themis: fair and efficient GPU cluster scheduling. *Proc. 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 289–304.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, 6(1), 58–73.

Myerson, R. B., & Satterthwaite, M. A. (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2), 265–281.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., et al. (2024). Computing power and the governance of artificial intelligence. arXiv:2402.08797.

Sivan, B., Paes Leme, R., Tiuca, M., McFarlane, I., Gkatzelis, V., Mehta, N., Hassas Yeganeh, S., Mirrokni, V., & Vahdat, A. (2026). Quota marketplace: dynamic pricing for efficient allocation of ML training resources. *Proc. 20th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. arXiv:2607.09802.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1), 8–37.

Vuppalapati, M., Fikioris, G., Agarwal, R., Cidon, A., Khandelwal, A., & Tardos, É. (2023). Karma: resource allocation for dynamic demands. *Proc. 17th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 645–662.

*Program documentation (accessed 4 August 2026).*

- NAIRR: nairrpilot.org; /opportunities/allocations; /opportunities/startup-project.
- UK AIRR: ukri.org Isambard-AI/Dawn Innovator, Rapid Access, and Gateway route pages; gov.uk "AI Research Resource: AI open access call."
- EuroHPC AI Factories: eurohpc-ju.europa.eu AI-Factories access-modes, Fast Lane, Large Scale; EuroHPC JU Access Policy (April 2025).
- ABCI: abci.ai/en/how_to_use/tariffs.html; /faq.html.
- IndiaAI: pib.gov.in PRID 2097709, PRID 2132817; compute.indiaai.gov.in.
- GPU pricing (Fig. 1): lambda.ai/pricing; coreweave.com/pricing.


## A VERIFICATION CODE

The script recomputes every number in Sections 5–7: the first-best allocation and surplus, all four Clarke payments, the budget-capped clock price path and realized welfare, the welfare-optimal uniform posted price, the DRF allocation, the Probabilistic-Serial allocation, the credit-mechanism feasibility thresholds, and the unbounded worst-case ratio. It is pure-stdlib Python and self-checks the load-bearing values with `assert`; running it prints `all checks pass`.

```
LABS = {
    'A': dict(rate=400,  cap=300, h200=0,   budget=90_000),
    'B': dict(rate=600,  cap=250, h200=100, budget=150_000),
    'C': dict(rate=700,  cap=400, h200=0,   budget=60_000),
    'D': dict(rate=1300, cap=150, h200=200, budget=180_000),
    'E': dict(rate=450,  cap=200, h200=150, budget=90_000),
}
Q_H100, Q_H200 = 1000, 500
NAMES = list(LABS)

def val(lab, x):
    return LABS[lab]['rate'] * min(x, LABS[lab]['cap'])
```

```
def first_best(subset, cap_h100=Q_H100):
    order = sorted(subset, key=lambda l: -LABS[l]['rate'])
    alloc = {l: 0 for l in subset}; left = cap_h100
    for l in order:
        take = min(LABS[l]['cap'], left); alloc[l] = take; left -= take
    return alloc, sum(val(l, alloc[l]) for l in subset)

fb_alloc, fb_surplus = first_best(NAMES)
assert fb_surplus == 715_000
assert sum(LABS[l]['h200'] for l in NAMES if fb_alloc[l] > 0) <= Q_H200

# Clarke payments (all winners)
clarke = {}
for i in NAMES:
    _, w_without = first_best([l for l in NAMES if l != i])
    clarke[i] = w_without - (fb_surplus - val(i, fb_alloc[i]))
assert clarke['C'] == 120_000 and clarke['B'] == 100_000
assert clarke['D'] == 60_000 and clarke['E'] == 80_000

# Budget-capped clock (single binding good)
def clock_demand(p):
    return {l: (0.0 if LABS[l]['rate'] <= p
                else min(LABS[l]['cap'], LABS[l]['budget']/p)) for l in NAMES}
def clock_clear(step=0.001):
    p = 0.0
    while True:
        p += step
        if sum(clock_demand(p).values()) <= Q_H100 + 1e-9:
            return p, clock_demand(p)
p_star, cl = clock_clear()
assert abs(p_star - 375) < 0.5
assert abs(sum(val(l, cl[l]) for l in NAMES) - 643_000) < 5

# Welfare-optimal uniform posted price
def welfare_posted(p):
    dem = {l: (min(LABS[l]['cap'], LABS[l]['budget']/p) if LABS[l]['rate'] > p
      else 0.0)
           for l in NAMES}
    tot = sum(dem.values())
    if tot > Q_H100:
        f = Q_H100/tot; dem = {l: dem[l]*f for l in dem}
    return sum(LABS[l]['rate']*dem[l] for l in NAMES)
best_p = max(range(1, 1401), key=welfare_posted)
assert best_p == 375  # ties the clock

# DRF (divisible)
dem = {l: (LABS[l]['cap'], LABS[l]['h200']) for l in NAMES}
dom = {l: max(dem[l][0]/Q_H100, dem[l][1]/Q_H200) for l in NAMES}
cH100 = sum(LABS[l]['cap']/dom[l] for l in NAMES)
cH200 = sum(LABS[l]['h200']/dom[l] for l in NAMES)
s = min(Q_H100/cH100, Q_H200/cH200)
drf = {l: LABS[l]['cap']*s/dom[l] for l in NAMES}
assert abs(sum(val(l, drf[l]) for l in NAMES) - 615_464) < 5

# Probabilistic Serial (single scarce good = capped equal division)
def capped_equal_division(caps, supply):
    alloc = {l: 0.0 for l in caps}; active = set(caps); left = supply
    while active and left > 1e-9:
        share = left/len(active)
        b = min(active, key=lambda l: caps[l]-alloc[l]); need = caps[b]-alloc[b]
        if need <= share + 1e-12:
            for l in list(active): alloc[l] += need
            left -= need*len(active); active = {l for l in active if caps[l]-
       alloc[l] > 1e-9}
        else:
            for l in list(active): alloc[l] += share
            left = 0
    return alloc
ps = capped_equal_division({l: LABS[l]['cap'] for l in NAMES}, Q_H100)
assert abs(sum(val(l, ps[l]) for l in NAMES) - 653_333) < 5

# Credit mechanism: efficient alloc credit-feasible iff endowment >= Clarke
       price
assert clarke['C'] == 120_000  # vs real budget 60,000

# Worst-case clock/first-best is unbounded
def clock_ratio(V, v, beta):
    return min(1.0, beta/v)
assert clock_ratio(10**6, 10**3, 100) == 0.1
print("all checks pass")
```