Beyond Scarcity: A Unified Theory of Input Allocation and Market Concentration

Hanna Halaburda*
NYU Stern School of Business

Extended abstract, August 2026

Summary. Why do some technology-intensive markets support many firms while others tip toward one or two, even when they depend on similarly scarce inputs? Standard explanations emphasize the amount of the input, economies of scale, or demand-side network effects. This paper identifies a different technological determinant of concentration: the rule through which firms compete for the input. Even when physical scarcity, downstream demand, and the set of potential firms are unchanged, a shift from proportional access toward winner-take-all allocation can sharply concentrate production.

I develop a unified model of service markets in which buyers choose providers before performance is realized and service quality depends on firms' subsequent access to a rival input—compute, attention, spectrum, capacity, a procurement award, or transaction priority. Firms compete for the input through effort, investment, or bids. The model nests proportional contests, probabilistic rankings, shared-input markets, and all-pay, first-price, and second-price auctions, as well as commodity market. These institutions, normally treated as separate market structures, are ordered by a single parameter: *decisiveness*, the elasticity of relative input shares with respect to relative effort. The framework therefore turns comparisons across auction formats, algorithms, procurement rules, and capacity contracts into one comparative static.

The results follow a simple logic. Firms that can earn more from the input compete harder and obtain more of it. A more proportional allocation rule preserves modest differences in access; a more decisive rule magnifies the same heterogeneity into large differences in input shares and service quality. For a fixed set of active firms and regular equilibrium responses, greater decisiveness increases concentration and the leading firm's rents. The winner-take-all limit delivers a stronger general result: if several firms remain in buyers' provider set as allocation becomes perfectly decisive, their effective rewards from the input must converge to exact equality. Unless prices and demand can produce that equality, multi-firm equilibria eventually disappear and the largest sustainable provider set converges to one. Multihoming can slow concentration but cannot robustly prevent it when smaller firms' input shares vanish. A demand-side feedback—where reliability attracts demand and demand raises the value of the input—amplifies the result but is not needed to generate it.

While it can be observed in many industries, blockchain execution provides an unusually clean illustration of the theory. The bottleneck input is explicit—the right to construct and order the next block—and Ethereum moved from probabilistic proof-of-work toward more decisive auction-based block construction. Effort or bids, winning shares, transaction values, and private routing can be observed at high frequency, while transaction ordering makes the reward from winning heterogeneous across builders. Existing evidence on builder concentration and exclusive order flow illustrates the model. The same empirical strategy applies more broadly: after estimating or ordering the decisiveness of an allocation rule, and measuring dispersion in firms' value of the input, it's possible to test whether their interaction predicts input concentration, downstream provider concentration, multihoming, and rents. Applications include AI compute reservations, platform rankings, spectrum allocation, and procurement.

Keywords: market concentration; input allocation; technology; contests; auctions; multihoming; blockchain; artificial intelligence. *JEL:* D43, L13, L22, L86, O33.

*The author has no relevant research support, financial interests or other conflicts of interest to disclose.

1 The question: not only how scarce, but how allocated

The mechanism is easiest to see in four familiar settings.

1. An AI developer can promise fast and reliable model access only if it secures accelerator clusters, electrical power, and specialized talent. If compute can be rented in small increments as demand arrives, several firms may scale. If the same compute is tied up through long-term reservations and dedicated infrastructure, a modest financing or forecasting advantage can translate into a large capacity advantage before customers ever observe realized performance.

2. A mobile app or content creator needs exposure. Under a gradual recommendation system, a small improvement in engagement may produce a small increase in impressions. Under a top-chart or featured-slot rule, the same improvement can determine who receives nearly all of the valuable attention. The scarce object is not merely "attention"; it is attention allocated through a particular algorithm.

3. A defense contractor may need to win a large contract to produce at scale. A procurement agency can split a contract among suppliers, award several lots, or choose a single winner. The physical production technology may be unchanged, but the allocation rule turns a small bid advantage into either a somewhat larger share or the entire contract.

4. Blockchain execution makes the contrast especially stark. In every round, only one next block is accepted. Under proof-of-work, an additional unit of mining effort raises the probability of winning roughly in proportion to the miner's share of total effort. Under proposer–builder separation, specialized builders submit bids and the highest bid wins the block. The resource—one block—is equally scarce in both regimes; what changes is how decisively competition maps differences in effort or value into production shares.

These examples are often placed in different literatures: capacity investment, platform rankings, procurement, auctions, or blockchain design. They are not analogies appended to an industry-specific model. They are instances of the same primitive: how elastically the relative effort becomes the relative access. The model below uses one allocation rule and a flexible payment structure to nest the principal institutions through which firms obtain scarce inputs. A steeper ranking algorithm, a change in auction format, an exclusive capacity contract, or a move from split to single-winner procurement can therefore be studied as changes in the same parameter rather than as unrelated market designs. The key object is *decisiveness*: does a 10 percent advantage in effort create approximately a 10 percent advantage in access, or does it determine the winner?

Illustrative cross-market pattern: input-allocation decisiveness and downstream concentration

Market	Bottleneck input	How firms compete for the input	Decisiveness	Typical downstream concentration
Frontier AI models	Advanced compute, power, talent	Reserved clusters, dedicated infrastructure, capital races, and talent bidding	High	High
AI inference and applications	Compute and model access	Short-term rental, cloud purchase, and API access	Low–moderate	Lower; more dispersed
Content and mobile apps	Attention, exposure, rankings	Probabilistic recommendations, rankings, or top slots	Moderate	Moderate
Mobile telecom	Spectrum	Auctions tempered by caps, set-asides, or geographic division	Moderate–high	Moderate–high
Search and digital advertising	Users and attention	Ranking, default placement, and algorithmic attraction	Moderate–high	High
Defense procurement	Contracts and specialized capacity	Split awards, multiple lots, or a single winner	Low to high	High under single awards
Agriculture and routine services	Land, labor, standard capacity	Divisible market purchase or rental	Low	Low
Blockchain execution	Right to construct and order the next block	Probabilistic mining or staking; highest-bid block construction	Low under PoW; high under PBS	Moderate under PoW; high under PBS

Notes: Concentration refers to the downstream provider market, not to concentration among suppliers of the bottleneck input. The labels summarize motivating patterns rather than causal estimates; the model isolates the allocation margin and generates comparative statics and within-market tests.

The final two columns make the motivating pattern visible: markets in which small effort advantages translate into disproportionate input access tend also to support fewer viable downstream providers. The paper does not infer causality from this cross-section. It develops a unified model that holds other primitives fixed, identifies decisiveness as the relevant technological margin, and delivers predictions that can be tested using changes in allocation rules within markets.

> **Contribution.** Holding the physical scarcity of an input fixed, a more decisive technology for allocating that input makes downstream production more concentrated, reduces the set of providers buyers can profitably use, and increases the rents available to the leading firm.

This margin is distinct from standard explanations based on fixed costs, economies of scale, or demand-side network effects. Those forces may be present, but the results below arise even when the number of potential firms is fixed, entry is open, and the resource itself does not become more scarce. The allocation rule creates an additional source of concentration. Market structure is an equilibrium object in the model: firms choose prices and input-seeking effort, buyers choose how many providers to use, and the allocation technology determines which provider sets can survive.

2 A unified model of input competition

2.1 The environment

There are $N \geq 2$ potential firms and a continuum of buyers. The game has three stages.

1. Firms set performance-contingent fees or commissions r_i.
2. Buyers choose one firm or a set of firms S. Using several firms provides redundancy, but incurs a convex multihoming cost $\ell(|S|)$.
3. Firms compete for a rival input. The realized share of that input determines the quality or reliability of the service they deliver.

Performance-contingent pricing covers familiar arrangements such as per-click advertising, transaction fees, platform commissions, and payments made when a contract is successfully delivered.

The timing captures markets in which the buyer commits before final performance is known. A trader submits an order before learning who will build the next block; an advertiser chooses a platform before the impression and conversion are realized; an enterprise customer selects one or more service providers before future capacity is called upon; and a procuring organization chooses suppliers before delivery quality is realized.

Firm i has a baseline capability or monetization parameter b_i. If it is selected by buyers, its effective reward from obtaining the input is

$$R_i = R(b_i, r_i, T_i), \tag{1}$$

where T_i is the demand routed to the firm. The reward is increasing in capability, the fee, and demand. Economically, R_i is the dollar value to firm i of obtaining one unit of the bottleneck input. It may reflect advertising revenue generated from attention, the profit from serving an enterprise contract with scarce compute, the margin on a procurement award, or the transaction-ordering value captured by a block builder.

2.2 One allocation rule, many institutions

Firm i chooses effort, investment, or a bid $e_i \geq 0$. For a finite decisiveness parameter $a \geq 1$, its expected input share is

$$q_i(e; a) = \frac{e_i^a}{\sum_{k=1}^{N} e_k^a}. \tag{2}$$

The parameter a has a direct economic interpretation. For any two active firms,

$$\log \frac{q_i}{q_j} = a \log \frac{e_i}{e_j}. \tag{3}$$

Thus a is the elasticity of relative input shares with respect to relative effort. At $a = 1$, a 10 percent effort advantage creates a 10 percent advantage in the odds of receiving the input. At $a = 5$, the same effort advantage creates roughly a 61 percent odds advantage. As $a \to \infty$, the highest-effort firm receives the entire input.

Payoffs take the flexible form

$$\Pi_i(e; R_i, a) = q_i(e; a)\big[R_i - \beta(e_i) - \gamma(e)\big] - \alpha(e_i), \tag{4}$$

where $\alpha(e_i)$ is paid regardless of whether the firm obtains the input, $\beta(e_i)$ is paid conditional on obtaining it, and $\gamma(e)$ is a label-neutral payment or externality that can depend on rivals' actions. This cost structure is what allows the same allocation rule to encompass institutions usually modeled separately.

Canonical institution	**Decisiveness**	**Mapping into the model**
Proportional Tullock contest	$a = 1$	Effort is paid up front: $\alpha(e_i) = ce_i$, with no winner-contingent payment.
Probabilistic ranking or contest	finite a	Investment raises exposure or success smoothly; larger a makes the ranking steeper.
All-pay auction	$a \to \infty$	Highest effort wins and every firm pays its effort: $\alpha(e_i) = e_i$.
First-price auction	$a \to \infty$	Highest bid wins and the winner pays its own bid: $\beta(e_i) = e_i$.
Second-price auction	$a \to \infty$	Highest bid wins and the winner pays the runner-up's bid through $\gamma(e) = max_{j \neq i} e_j$.
Divisible input purchased or rented	low a	Access changes gradually with purchases or capacity commitments rather than discontinuously at one winning bid.

The theoretical contribution is therefore broader than comparing a few named markets. Equation (4) provides a common environment, while a orders how selective the input-allocation technology is. A change in procurement rules, a steeper ranking algorithm, an exclusive supply contract, or a shift from probabilistic selection to an auction can be studied as a change in the same primitive.

2.3 Buyer choice and multihoming

A firm's input share is an endogenous dimension of quality. Let $\phi(q_i)$ be the value to a buyer of the firm's expected ability to deliver, with ϕ increasing. If buyers select a set S, their utility is

$$B(S) = \sum_{i \in S} (1 - r_i)\phi(q_i) - \ell(|S|), \tag{5}$$

where $\ell(1) = 0$ and ℓ is increasing and convex. A buyer may use two cloud providers, route a transaction to several executors, retain multiple suppliers, or install several apps, but every additional route creates integration, duplication, privacy, monitoring, or switching costs.

A set of m firms can be sustained only if the contribution of every included firm is at least the marginal cost of the mth route:

$$(1 - r_i)\phi(q_i) \geq \Delta\ell(m) \equiv \ell(m) - \ell(m-1) \qquad \text{for all } i \in S. \tag{6}$$

This condition connects the upstream allocation rule to downstream market structure. More decisive input competition changes not only who obtains the resource, but which firms remain valuable enough for buyers to keep in their provider set.

3 Main results

3.1 Result 1: competition sorts firms by the value of the input

Under common cost primitives and a label-neutral allocation environment, a firm with a higher effective reward chooses greater effort and obtains a larger input share. Among strictly profitable firms,

$$R_i > R_j \quad \Longrightarrow \quad e_i^* > e_j^* \quad \Longrightarrow \quad q_i^* > q_j^*. \tag{7}$$

The result is simple but consequential. Open participation does not make the allocation symmetric when firms differ in what they can earn from the resource. A builder with better extraction technology, a platform with better monetization, or an AI firm with more valuable enterprise demand rationally competes harder for the same input. Competition sorts firms according to these differences instead of washing them out.

This is the first step from technology to market structure: heterogeneity in the value of the input becomes heterogeneity in access to the input, even when firms face the same formal rules and the same cost technology.

3.2 Result 2: decisiveness turns small advantages into large share differences

Equation (3) shows the mechanical channel. For a given effort ranking, raising a stretches every pairwise share difference. Under the regularity conditions in the paper—a locally fixed profitable set, a regular equilibrium selection, and effort ratios that do not move against the mechanical effect—the equilibrium share vector becomes more concentrated as a rises. The leading firm's share and any Schur-convex concentration index, including HHI, increase whenever active firms do not choose identical effort.

This comparative static replaces a collection of institution-specific statements with one result. A proportional contest, a probabilistic algorithm, and a winner-take-all auction are not simply different labels. They are different values of the elasticity in (3). The same underlying capability or valuation gap can therefore generate a broad market under one allocation technology and a highly concentrated market under another.

There are two reinforcing channels. First, decisiveness mechanically magnifies any existing effort difference. Second, once a small lead produces a much larger input share, the return to defending or extending that lead rises. The local comparative static formalizes both channels. The decisive-limit result below is stronger in a different sense: it does not require global monotonicity of equilibrium effort ratios.

3.3 Result 3: high decisiveness makes diversified provision a knife-edge

Suppose buyers continue to use exactly $t \geq 2$ firms along a sequence of equilibria in which $a \to \infty$. Supportability requires every selected firm to retain a strictly positive input share. At the perfectly decisive limit, however, only firms tied at the highest effort can receive positive shares. Because equilibrium effort is increasing in the effective reward, firms with unequal rewards cannot remain tied at the top.

It follows that persistent coexistence of t firms requires their effective rewards to converge to exact equality:

$$\overline{R}_i = \overline{R}_j \qquad \text{for every pair } i, j \text{ in the active set.} \tag{8}$$

Prices can in principle offset capability differences: a stronger firm can charge less, or a weaker firm can charge more, so that both value the input equally. But such equalization must also keep every firm attractive to buyers. It is therefore a demanding equilibrium requirement, not an automatic consequence of competition.

If heterogeneous firms cannot use prices and demand to make their rewards converge to an exact common tie, then for every target size $m \geq 2$ there exists a finite threshold $\bar{a}_m$ beyond which no equilibrium supports m or more firms. Under a mild condition ensuring that at least one firm remains active, the maximal equilibrium provider set converges to one as $a \to \infty$.

This result gives multihoming a precise role. Multihoming can preserve several providers when smaller firms still add enough reliability to justify another route. It cannot robustly offset an allocation technology that makes the smaller firms' input shares vanish. In the simpler monotone case, the size of the multihoming set falls as decisiveness rises; the limit theorem shows that, even without assuming monotonicity at every intermediate value, large active sets eventually disappear unless rewards exactly equalize.

3.4 Rents and an endogenous demand feedback

Concentration affects rents as well as shares. In a single-homing market, the leading firm wins demand but its fee is constrained by the best credible offer of the runner-up. A more decisive input rule lowers the runner-up's attainable input share and therefore the quality it can credibly promise. The leader need not give buyers as much surplus to win the market, so equilibrium fees and rents rise while consumer surplus falls.

The core sorting and decisiveness results do not assume an exogenous network effect. A routed-demand extension creates one endogenously. If attracting demand raises the reward from the input, the model contains the feedback

$$q_i \uparrow \Longrightarrow \text{ reliability}_i \uparrow \Longrightarrow T_i \uparrow \Longrightarrow R_i \uparrow \Longrightarrow e_i \uparrow \Longrightarrow q_i \uparrow . \tag{9}$$

This extension is especially important for digital markets in which buyers can route activity toward the firms most likely to deliver. It strengthens the concentration result, but is not its source: the allocation technology already sorts and magnifies heterogeneity before the feedback is added.

4 Blockchain as a laboratory

Blockchain is useful here not because the theory is about one niche industry, but because the elements of the model are unusually visible and the allocation technology change has been observed within the largest ecosystem, Ethereum.

First, the bottleneck input is explicit. In each round, the system must choose who constructs the next block and therefore which transactions are included and in what order. The realized resource is indivisible, but repeated rounds generate an observable production share for every miner, validator, or builder.

Second, firms' rewards from the input are measurably heterogeneous. Suppose a trader submits a large buy order before settlement. The producer of the next block can place a transaction before the order and another after it, or can exploit related arbitrage and liquidation opportunities. Producers differ in the speed, algorithms, private information, and order flow they can use to capture this value. The formal protocol may

offer the same base reward to everyone, yet the effective reward R_i from winning the block differs across producers. This is the blockchain microfoundation for the heterogeneity that drives the general model.

Third, Ethereum provides variation in decisiveness. Under proof-of-work, mining effort translated probabilistically into block-winning shares. After the move to proof-of-stake and proposer–builder separation, specialized builders increasingly competed through auctions for the right to supply the block. The transition maps naturally into a movement from a relatively diffuse contest toward a more decisive allocation rule.

Fourth, downstream routing is visible. Traders can submit transactions to a public mempool or route them privately, sometimes exclusively, to particular builders. A builder with a higher probability of inclusion is a more reliable destination. Private flow then raises the value of the blocks that builder can assemble, which raises its bids and future winning probability. The feedback in (9) is therefore not a metaphor in this setting; its components can be observed.

Existing empirical work documents the patterns that motivate this mapping. Transaction ordering creates substantial extractable value (Daian et al 2019); block building under proposer–builder separation is concentrated (Heimback et al 2023); winning is closely related to access to exclusive order flow (Oz et al 2024; and one recent study estimates that transactions exclusive to a single builder account for 77–84 percent of builder revenue (Pahari and Canidio 2025). These facts do not establish the comparative static by themselves. They make blockchain a disciplined laboratory in which the model's primitives—effort, reward heterogeneity, decisiveness, input shares, and routing—have concrete empirical counterparts.

This is a theory paper. The blockchain evidence motivates the primitives, illustrates the equilibrium mechanism, and shows how the model can be taken to data; the paper does not report a new empirical estimate.

5 How can this theory be tested empirically

The model is designed around objects that can be measured. Its most distinctive prediction is not simply that scarce-input markets are concentrated. It is an interaction:

> A change toward more decisive input allocation should increase concentration most strongly where firms' rewards from obtaining the input are already more dispersed.

That prediction separates the theory from a generic claim that auctions, algorithms, or new technologies always concentrate markets. If firms value the input equally, greater decisiveness has little heterogeneity to magnify. If reward dispersion is large, the same change in the allocation rule has a much larger effect. A natural reduced-form implementation is

$$C_{mt} = \mu_m + \tau_t + \beta a_{mt} + \delta D_{mt} + \theta\, a_{mt} D_{mt} + X'_{mt}\Gamma + \varepsilon_{mt}, \qquad \theta > 0, \tag{10}$$

where C_{mt} is an input-share or downstream concentration measure, a_{mt} captures the decisiveness of the allocation rule, and D_{mt} measures dispersion in firms' rewards from the input. The equation is a guide to measurement rather than a claim that the structural relationship is linear.

A direct empirical implementation has four parts.

1. **Measure decisiveness.** When effort or bids and realized shares are observed, equation (3) provides a direct elasticity. When they are not, institutional rules can order regimes: proportional sharing, probabilistic ranking, split awards, and spot rental are less decisive than top-slot rules, exclusive contracts, or highest-bid-wins auctions.

2. **Measure reward heterogeneity.** Relevant proxies include revenue per unit of input, builder-specific block value, access to exclusive transactions, conversion rates from attention, pre-existing customer value, or margins on awarded capacity.

3. **Use rule changes that move decisiveness while holding scarcity as fixed as possible.** Candidate settings include changes in ranking algorithms, procurement from split to single-winner awards, spectrum caps or reservations, shifts between flexible rental and long-term capacity commitments, and Ethereum's movement from probabilistic mining toward auction-based block building.

4. **Study the predicted outcomes.** The theory speaks to concentration in input shares and downstream demand, the number of providers used by each buyer, the persistence of leaders, and markups or rents. Near a highly decisive regime, multiple surviving firms should display increasingly similar effective rewards; persistent reward gaps predict a shrinking provider set. In the routed-demand extension, lagged input share should predict subsequent demand, which should in turn predict future reward and input share.

Blockchain offers the most complete measurement environment. Block-level execution records identify realized shares; proof-of-work data, staking records, and relay bid logs measure effort or bids; transaction-level values and builder-specific block contents proxy for R_i; public-mempool traces and private or exclusive transaction measures capture routing. The model predicts that a rise in decisiveness has a larger concentration effect during periods or for transaction classes with greater dispersion in order-sensitive value. It also predicts a two-way lead–lag relation between builder reliability and private order flow, and fewer simultaneously used builders when duplication, latency, or privacy costs are higher.

The same empirical design travels to other industries even when the exact data differ. Platform exposure histories reveal input shares; ranking scores, advertising expenditures, or content investment proxy for effort; conversion revenue measures the value of exposure. Procurement records contain bids, contract allocation, split versus sole awards, and supplier concentration. AI capacity contracts can distinguish spot-like access from reserved or dedicated infrastructure, while downstream enterprise adoption provides a measure of demand and multihoming. The theory thus provides both a set of hypotheses and a way to organize otherwise disparate data.

6 Broader implications for firms and technology

The decisiveness of input allocation is not fixed. New technologies and organizational forms can change it even when they leave the physical input unchanged. Open-market purchase can give way to exclusive supply contracts; short-term flexible access can give way to long-term reservations; continuous capacity expansion can give way to lumpy commitments; a recommendation algorithm can become more top-heavy; and a procurement process can move from split awards to a single winner.

This is particularly relevant for AI. At the foundation-model layer, advanced compute, power, and specialized talent are often acquired through large capital commitments, reserved clusters, dedicated infrastructure, and talent bidding. These are relatively decisive ways of securing inputs. Inference capacity can be more divisible and rented through cloud markets. The model predicts different concentration patterns across these layers even though both belong to the same technological stack. It also predicts that a shift toward exclusive or long-term capacity commitments can concentrate the downstream model market without any change in the total quantity of compute.

Platform markets provide a second application. A recommendation or ranking algorithm is an input-allocation technology: it determines how attention is distributed among creators, apps, merchants, or advertisers. Changes in the steepness of the mapping from engagement scores to exposure are changes in a. The model predicts that making the ranking more selective increases concentration most where monetization or conversion ability is heterogeneous, and that the effect appears not only in exposure but eventually in downstream participation and rents.

Procurement and spectrum illustrate how institutions can deliberately temper decisiveness. A sole award or unrestricted winner-take-all auction can turn a small bid or capability advantage into exclusive access. Lotting, caps, reservations, or geographic fragmentation divide the input and keep smaller firms' shares positive. The model provides a common language for comparing these arrangements without treating each as an unrelated special case.

Finally, the framework suggests a different way to interpret technology-driven changes in firm dynamics. Technology affects market structure not only through productivity or fixed costs. It also changes the rules through which firms acquire the inputs needed to deploy the technology. A firm can exit the downstream market not because it is unable to produce in principle, but because a more decisive upstream allocation rule makes its attainable input share too small to support a credible offer to buyers.

7 Conclusion

Physical scarcity alone does not determine market concentration. The same scarce input can support a broad set of firms when access changes smoothly with effort and a narrow set when a small advantage determines the winner. By nesting proportional contests, probabilistic rankings, shared-input markets, and standard auctions in one framework, the paper turns institution-specific market structures into a comparative static with respect to input-allocation decisiveness.

The central result is that more decisive allocation sorts firms more sharply, magnifies reward heterogeneity, reduces robust multihoming, and increases rents. At the winner-take-all limit, coexistence requires exact equalization of active firms' rewards; without it, the provider set collapses. Blockchain shows the mechanism unusually cleanly, but the economic force is broader. In markets for compute, attention, spectrum, contracts, capacity, and transaction priority, the allocation technology is not a neutral plumbing choice. It is itself a determinant of downstream market structure.