Optimal Trading Speed on Blockchain[1]

Shihao Yu[2] Ruizhe Jia[3]

Current version: March 27, 2026

[1]We thank seminar participants for helpful comments. Yu: Lee Kong Chian School of Business, Singapore Management University; shihaoyu@smu.edu.sg. Jia: Department of Management Science and Engineering, Stanford University; ruizhejia@stanford.edu.

[2]Singapore Management University

[3]Stanford University

Abstract

We study the optimal block speed for blockchain trading infrastructure. Block speed determines how agents compete for execution priority: slow blocks induce within-block fee auctions whose revenue accrues to the block producer, while fast blocks shift competition toward private latency investment. We characterize conditions under which welfare is maximized at an interior block speed. Slow blocks make it unprofitable for sophisticated liquidity providers to enter, leaving the market with only passive capital. Fast blocks reward active position management but displace passive providers, concentrating liquidity supply. A difference-in-differences analysis around a tenfold block-speed increase on Base, using Arbitrum as control, confirms compressed price deviations and concentrated provision.

1 Introduction

Blockchain-based financial infrastructure is entering production deployment for settlement and trading. The Depository Trust & Clearing Corporation (DTCC) and blockchain infrastructure firm Digital Asset announced a partnership in December 2025 to tokenize DTC-custodied U.S. Treasury securities on the Canton Network, which processes approximately $350 billion per day in U.S. Treasury activity.[1] Legacy exchanges are also adopting blockchain infrastructure: in March 2026, the SEC approved Nasdaq's application to list and trade tokenized securities—enabling Russell 1000 stocks and ETFs to settle as blockchain-based tokens—and the NYSE partnered with Securitize to launch a 24/7 tokenized securities platform.[2] In digital asset trading, the blockchain-based exchange Hyperliquid recorded roughly $2.6 trillion in perpetual-futures volume in 2025, surpassing Coinbase, and decentralized exchange (DEX) spot volume held near one-fifth of centralized exchange (CEX) spot volume by late 2025.

Yet this infrastructure differs from traditional exchanges in three respects. First, transactions do not settle continuously but in discrete blocks produced at a fixed cadence—every 200 milliseconds on Base, every 250 milliseconds on Arbitrum, every 12 seconds on Ethereum. The block cadence is therefore the effective frequency at which the market clears, analogous to the trading frequency parameter in periodic-auction models. Second, the entity that produces each block—the sequencer or validator—exercises temporary monopoly control over transaction ordering within that block and can monetize this control by selling priority through fee auctions or auctioned access. The infrastructure operator is not a passive matcher but an active participant in rent allocation. Third, unlike traditional exchanges, where each venue owns and controls its own matching engine, blockchain is shared infrastructure: a single chain's execution clock governs all markets deployed on it simultaneously. On Base alone, over 512,000 liquidity pools, 22 DEX protocols, and dozens of

[1] The Canton Network connected 22 capital-markets applications in its 2024 pilot and completed live 24/7 Treasury financing tests in July 2025. BlackRock's BUIDL tokenized money market fund surpassed $2 billion in assets across seven blockchains by year-end 2025.

[2] Nasdaq separately partnered with the digital asset exchange Kraken to distribute tokenized U.S. equities to international investors, extending blockchain settlement beyond domestic markets.

lending and derivatives applications all operate under the same ordering cadence. A change in the execution clock therefore propagates across the entire on-chain ecosystem—a public infrastructure design problem, not a single-venue decision.

Across the blockchain ecosystem, infrastructure is getting faster. Ethereum's base layer moved from variable block times to fixed 12-second slots at the Merge. Layer-2 networks have pushed further: Arbitrum operates at 250-millisecond blocks, Base's Flashblocks reduced its effective ordering cadence from 2 seconds to 200 milliseconds in July 2025, and Solana targets roughly 400-millisecond slots. The presumption driving this race is that faster execution must improve market quality—if stale-price arbitrage is a timing problem, shorter settlement intervals should mechanically reduce it. Theoretical support comes from loss-versus-rebalancing models showing that higher block frequency reduces arbitrage profit rates when blocks arrive as a Poisson process (Milionis, Moallemi, and Roughgarden 2024), and empirical evidence that LP capital efficiency improves on faster, cheaper chains (Adams 2024). A large literature in traditional market microstructure warns that this reasoning is incomplete. Exchange latency can either improve or worsen liquidity depending on the balance between reduced payoff risk and intensified adverse selection (Menkveld and Zoican 2017). Higher trading frequency can reduce welfare when strategic costs dominate gains from faster response to news (Du and Zhu 2017b). Private speed investment can be socially wasteful when it reallocates a fixed prize rather than creating surplus (Budish, Cramton, and Shim 2015b). All three lines of research point to the same conclusion: speed and market quality are not monotonically related.

Blockchain infrastructure poses the same question, but in a setting where the economic channels differ in ways that require new theory. On a traditional exchange, the venue is a passive matcher: it sets rules but does not directly profit from individual trade timing. On blockchain, the block producer sells priority on an order-by-order basis: each transaction's position within the block is determined by its fee bid, and the block producer captures this fee as direct revenue—the median arbitrageur on Ethereum transfers approximately 96% of profits to validators as infrastructure fees (Capponi and Jia forthcoming). On a traditional limit order book, each market maker

quotes independently and can cancel stale orders on her own. On blockchain, liquidity is pooled: adverse-selection losses are shared pro rata across all liquidity providers (LPs), but withdrawal costs are borne individually, creating a free-rider problem absent from dealer markets. And on a traditional exchange, speed competition operates on one margin at a time—either latency investment or priority fees. On blockchain, agents compete on *two nested levels*: they race to land a transaction before the block boundary, and conditional on landing together, they compete again for within-block priority through fee auctions. Execution is batched on the outside but serial and monetized on the inside. These structural differences mean that the block cadence simultaneously shapes stale-price exposure, the allocation of timing rents between traders and infrastructure, and the composition of liquidity supply—a compound design problem in which speed investment, priority bidding, and LP participation are jointly determined through a single protocol parameter—the block speed.

Understanding how infrastructure speed affects rent allocation, liquidity composition, and welfare on blockchain markets is therefore an open question that requires both theoretical and empirical analysis. The theoretical challenge is to build a unified model that simultaneously captures the infrastructure operator's priority-fee extraction, the pooled structure of liquidity provision and withdrawal, endogenous speed investment by heterogeneous agents, and the resulting feedback between block speed and market composition—moving parts that existing models treat in isolation. The empirical challenge is that observing exogenous variation in infrastructure speed is rare: traditional exchanges change latency gradually and in response to competitive pressure, confounding speed with other design choices. Blockchain infrastructure upgrades, by contrast, generate discrete, dated policy changes—making the setting unusually well-suited for quasi-experimental analysis.

We build a model of trading on blockchain infrastructure in which the block production rate λ_B is the protocol designer's control variable. Two arbitrageurs, a strategic liquidity provider, a continuum of passive liquidity providers, and a sequencer trade a risky asset on an automated market maker. Blocks arrive stochastically, and between blocks agents observe external price

movements and prepare transactions; at each block boundary, the sequencer orders all arrived transactions by fee bid and executes them against the pool.

The central mechanism is that block speed determines whether agents *clash*—arrive in the same block and compete for priority—or *race*—compete across blocks for who lands first. When blocks are slow, the submission window is long enough that both arbitrageurs almost certainly land before the block boundary. They clash, and the resulting priority-fee auction drives bids toward the full arbitrage prize: intense competition among arbitrageurs produces a Bertrand-like outcome in which the infrastructure operator, not the traders, captures the economic surplus from stale prices. When blocks are fast, the window is short: at most one agent lands per block, clashes are rare, and competition shifts toward private speed investment that determines who arrives first. The form of competition—within-block fee auction versus inter-block speed race—is therefore endogenous to the block rate, and with it the allocation of rents between traders and infrastructure.

The existing literature on AMM liquidity provision models LPs as fundamentally passive: they deposit capital, accept the pricing rule's adverse-selection costs, and earn fees in return (Lehar and Parlour 2023; Capponi and Jia forthcoming; Milionis, Moallemi, Roughgarden, and Zhang 2023). Even models of concentrated liquidity treat the LP's range choice as a static allocation decision rather than a dynamic, speed-dependent strategy (Hasbrouck, Rivera, and Saleh 2022). To our knowledge, this paper provides the first equilibrium model of *active* liquidity provision on AMMs—where LPs dynamically withdraw and redeposit in response to block-by-block price movements—and shows why this active behavior emerges endogenously as infrastructure speeds up.

The mechanism is a threshold problem created by block speed. In each state, the LP compares fee income from remaining in the pool against the expected loss from adverse selection by an arbitrageur who lands in the same block. When blocks are slow, the LP's probability of withdrawing safely—landing a withdrawal transaction in a block where no arbitrageur also lands—converges to zero: the LP is trapped. Entry becomes unprofitable, and the market relies on passive capital alone. When blocks are fast, the LP can exit and re-enter rapidly, adopting a threshold policy

that withdraws liquidity whenever the price gap exceeds a critical level. The LP effectively monitors external prices, pulls liquidity when adverse selection is imminent, and reposts when the gap narrows—a strategy that requires speed technology and continuous monitoring, and that only technologically sophisticated providers can execute.

This speed-induced transformation from passive provision to active market making has two consequences. First, active LPs who can rebalance at the block cadence capture fee income while avoiding the worst adverse-selection episodes, displacing passive providers who remain continuously exposed—concentrating liquidity supply among a small number of participants. The displacement operates through fee dilution: when the active LP enters the pool, the passive LP's share of fee income falls, but her per-execution adverse-selection loss is unchanged because losses are shared pro rata across the pool. The pooled structure of AMM liquidity, which in slow markets shares adverse-selection costs symmetrically across all providers, becomes in fast markets a channel through which active participants extract fee income while passive capital absorbs a disproportionate share of losses. Second, the composition of liquidity supply becomes endogenous to infrastructure speed—a dimension absent from models that treat liquidity provision as a passive, one-time allocation decision.

Aggregate welfare decomposes into three components: gains from trade, which increase with pool depth and therefore with LP participation; speed costs, the deadweight expenditure on latency technology; and sequencer leakage, the fraction of timing rents captured by the block producer through priority fees and removed from the welfare population. The sequencer's rent share is monotonically decreasing in block speed—faster infrastructure reduces priority-fee revenue per execution, shifting the rent channel from block-level extraction toward network-level speed competition. We characterize conditions under which welfare is maximized at an interior block rate. At low block rates, the strategic LP's entry condition fails and the market operates at baseline passive depth. At high block rates, private speed investment and the displacement of passive liquidity providers impose costs that can offset the incremental gains from deeper, faster-repricing pools. The arbitrage prize is itself endogenous to the block rate—unlike models in which it is a

fixed parameter (Budish, Cramton, and Shim 2015b)—because LP participation responds to the speed-dependent execution environment. Speed investment can therefore be self-defeating: it improves LP protection, which reduces the pool depth available for arbitrage, which shrinks the very prize that the speed investment was designed to capture. Liquidity and speed competition are complements—a deeper pool creates a larger arbitrage prize, which intensifies speed investment, which in turn changes LP incentives—and the planner cannot sustain deep liquidity without some associated speed competition. Whether the welfare function ultimately declines at high block rates depends on the elasticity of this response; we identify three regimes separated by observable parameter conditions. The source of the interior optimum differs from periodic-auction models (Du and Zhu 2017b), where non-monotonicity arises from thinner demand schedules at higher frequency. In our setting, the non-monotonicity reflects the joint equilibrium response of liquidity participation, speed investment, and infrastructure rent extraction to the protocol's execution clock.

We test the model's predictions using a difference-in-differences design around the rollout of Base Flashblocks on July 7, 2025, which shortened effective execution intervals from 2 seconds to 200 milliseconds while preserving the priority gas auction ordering rule. Arbitrum, operating under first-come-first-served ordering at 250-millisecond blocks, serves as the comparison chain. The analysis focuses on the Uniswap v3 WETH–USDC 5 basis point pool, using on-chain transaction data matched to Binance Level-2 order book snapshots at 100-microsecond precision. Time-weighted CEX–DEX price deviations compress after the speed increase, consistent with the model's prediction that faster blocks reduce stale-price exposure. LP rebalancing activity increases and concentrates sharply: on Base, a single LP accounts for 58% of all position adjustments (Herfindahl-Hirschman index of 0.36), compared to 8% on Arbitrum (0.01). The composition of blockchain fees shifts from within-block priority toward network-level costs, consistent with the monotonically decreasing sequencer rent share.

Related Literature

This paper relates first to the literature on trading speed, trading frequency, and transaction-ordering design. Budish, Cramton, and Shim (2015b) show that continuous serial processing creates mechanical arbitrage rents and a socially wasteful speed race; Menkveld and Zoican (2017) show that faster exchange speed can either improve or worsen liquidity; Du and Zhu (2017b) derive an interior optimal trading frequency in repeated double auctions; and Biais, Foucault, and Moinas (2011) show that investment in fast trading technology can be privately excessive. Budish, Cramton, and Shim (2015c) further show that exchanges earn economic rents from selling speed infrastructure and may therefore resist socially desirable market-design innovation—a force that has a direct analog in blockchain sequencer revenue. We share with this literature the insight that faster markets are not unambiguously better. We add a blockchain setting in which the protocol's execution clock jointly governs exchange latency, trading frequency, the private value of speed, LP participation, and sequencing-rent extraction—not one speed margin at a time, but the joint equilibrium that arises when these margins are bundled into a single protocol lever. Recent blockchain work studies transaction-ordering auctions and hybrid sequencing rules such as TimeBoost (Mamageishvili, Kelkar, Schlegel, and Felten 2023; Fritsch, Silva, Mamageishvili, Livshits, and Felten 2025). We differ by embedding the ordering contest in a block-gated AMM with endogenous active and passive liquidity, so the protocol's clock changes not only rent dissipation and price efficiency, but also the composition of liquidity supply.

Our paper also relates to the AMM literature on adverse selection and LP losses. Lehar and Parlour (2023) characterize equilibrium in Uniswap-style liquidity pools and compare the AMM to a centralized limit order market; Milionis, Moallemi, Roughgarden, and Zhang (2023) formalize loss-versus-rebalancing as the stale-price cost borne by LPs; and Park (2023) shows how transparent invariant-based pricing invites sandwich attacks and how alternative pricing rules change that exposure. We share this literature's focus on stale-price adverse selection. We add a dynamic execution layer: block speed and sequencing determine when arbitrage reaches the pool, how much

of the resulting rent is dissipated or extracted, and whether LPs remain willing to supply liquidity. The key difference is that we make the execution clock itself the central design variable rather than taking chain infrastructure as given.

A related literature studies active liquidity management on AMMs. Fan et al. (2023) analyze strategic liquidity provision when LPs choose concentrated ranges and face gas costs of reallocation; Wan and Adams (2022) document just-in-time liquidity and show that it is rare on Ethereum mainnet; and Capponi, Jia, and Zhu (2024) show that JIT LPs can avoid toxic flow and crowd out passive LPs. Malinova and Park (2023) show that the risk-sharing properties of AMM pooled liquidity work best when infrastructure is slow and all participants are equally trapped—as speed increases, AMM dynamics converge toward traditional market structures where fast intermediaries dominate. We share this literature's distinction between active and passive liquidity and strengthen the Malinova-Park insight by showing the precise condition—the block rate λ_B—under which pooled risk-sharing erodes. We add a general-equilibrium timing layer in which the protocol clock determines when active LP strategies are profitable, when they dilute passive fee income, and how those forces feed back into arbitrage and welfare.

Our paper further connects to the literature on MEV and ordering rents. Daian et al. (2020) document the widespread use of arbitrage bots on DEXs and their reliance on priority gas auctions. The proposer-builder separation (PBS) literature makes explicit how validators, builders, and relays interact in a market for profitable block construction (Schwarz-Schilling et al. 2023; Bahrani, Garimidi, and Roughgarden 2024). We share with this literature the view that ordering is an economically valuable object. We add a tractable equilibrium bridge from the protocol clock to the division of ordering rents among arbitrageurs, liquidity providers, and infrastructure. The MEV literature shows that blockspace is scarce and valuable; our paper shows how the execution clock changes who pays for it and when that rent extraction starts to erode liquidity provision.

We also relate to the emerging literature on L2 infrastructure and AMM market structure. Adams (2024) studies how cheaper gas and faster blocks on L2s affect Uniswap v3 and finds better execution and higher capital efficiency; Caparros, Chaudhary, and Klein (2023) show that

lower gas fees allow LPs to reposition more frequently; and Gogol, Messias, Miori, Tessone, and Livshits (2024) document persistent cross-rollup arbitrage opportunities on rollups. We share this literature's premise that infrastructure is not a neutral wrapper around AMMs. We add the missing theory: block cadence and sequencing jointly determine landing probabilities, contest incidence, speed adoption, passive liquidity participation, and welfare. Our contribution is not another reduced-form comparison across chains but a welfare theory of the execution clock.

Finally, our paper relates to work proposing alternative AMM or blockchain execution designs. Canidio and Fritsch (2023) propose a fair-market-maker AMM that uses batched execution to eliminate arbitrage profits; Adams, Moallemi, Reynolds, and Robinson (2024) auction temporary pool-management rights to reduce losses to informed order flow. We share this literature's mechanism-design motivation. We add a positive benchmark for the distortions these designs are trying to cure: how a block-gated AMM allocates rents across speed, priority, LP participation, and gains from trade as the execution clock varies. Those papers redesign the mechanism; our paper characterizes the equilibrium distortion that makes such redesigns potentially valuable in the first place.

More broadly, blockchain offers an unusually clean environment in which to study this problem. The clock is explicit. Ordering rules are publicly documented. Fee flows and rent transfers are measurable transaction by transaction. And infrastructure upgrades generate discrete, dated policy changes that are difficult to find in traditional exchanges. In that sense, blockchain is not a separate theory of markets—it is a setting in which the normally hidden interaction between speed, liquidity supply, and venue rents becomes observable enough to analyze jointly.

The remainder of the paper is organized as follows. Section 2 describes the institutional setting: how blockchain settlement, AMM pricing, and L2 execution clocks shape the economic environment. Section 3 presents the model. Section 4 characterizes equilibrium speed investment, LP participation, and rent allocation. Section 5 discusses the equilibrium results, characterizes the three welfare regimes, and draws implications for infrastructure design. Section 6 takes the model's predictions to data using the Base Flashblocks quasi-experiment. Section 7 concludes.

2 Institutional Details and the Execution Clock

This section describes the institutional features of blockchain-based trading that motivate and inform the model. Section 2.1 explains how transactions are settled on a blockchain and how block production creates the discrete execution clock. Section 2.2 describes how decentralized exchanges operate through automated market makers and explains how liquidity provision, pricing, and arbitrage work in this setting. Section 2.3 describes how Layer 2 networks have introduced diverse execution clocks and sequencing rules. Section 2.4 explains why block time is a public infrastructure parameter that simultaneously affects all markets on a chain. Section 2.5 describes the natural experiment that provides the empirical setting.

2.1 Blockchain Settlement

On a traditional exchange, a matching engine processes incoming orders continuously: each order is matched against the standing book as soon as it arrives. On a blockchain, settlement is discrete. Transactions submitted by market participants accumulate in a pending pool—the *mempool*—until the network produces a *block*: a batch of transactions that the protocol commits to the shared ledger simultaneously (Nakamoto 2008). The interval between successive blocks therefore determines the finest granularity at which economic exchange can occur. On Ethereum, blocks are produced every 12 seconds; on Layer 2 networks, the interval ranges from 200 milliseconds to 2 seconds.

Within each block, a designated entity—the *validator* on Ethereum's base layer, or the *sequencer* on Layer 2 networks—exercises temporary monopoly control over two decisions: which pending transactions to include, and in what order to arrange them. For every transaction submitted to the blockchain, users bid a transaction fee, commonly referred to as a *gas fee*.[3] Blocks are constrained by size, and transaction inclusion is determined by bid gas fees. Transactions with higher fees are typically prioritized to maximize the block producer's revenue; transactions with lower

[3]On Ethereum, the gas fee is calculated by multiplying the gas price by the amount of gas needed to execute the transaction. The gas price represents the amount of ETH paid per unit of gas, and the gas amount indicates the computational resources required. The unit for gas price is Gwei, where 1 Gwei equals 10^{-9} ETH.

fees may remain pending for a longer time before being confirmed. This ordering discretion—the block producer's ability to choose which transactions execute and in what sequence—is the institutional feature that creates the economic rents central to our analysis.[4]

2.2 Decentralized Exchanges and Automated Market Makers

The majority of trading on blockchain occurs through decentralized exchanges (DEXs)—smart contracts deployed on the blockchain that hold reserves of tokens and execute trades algorithmically, without the need for a centralized operator or traditional order book.[5] The dominant form of DEX is the automated market maker (AMM), a market structure characterized by liquidity pools and algorithmic pricing (Capponi and Jia forthcoming).

Liquidity pools and provision. An AMM holds reserves of two tokens—say, token A and token B—in a *liquidity pool*. Any participant can act as a *liquidity provider* (LP) by depositing an equivalent value of both tokens into the pool. In return, the LP receives a proportional ownership share, representing a claim on the pool's reserves at the time of withdrawal. For example, if the pool initially contains 100 units of each token and an LP deposits 10 units of each, the LP owns 10% of the pool. The value of this stake depends on the pool's composition at withdrawal, which may differ from the composition at deposit if trades have occurred in the interim.

LPs earn revenue from trading fees: each trade against the pool incurs a fee proportional to the trade size, and the fee accrues to all LPs in proportion to their pool shares.[6] Liquidity provision on an AMM is therefore analogous to market making on a traditional exchange, with an important structural difference: on a limit order book, each market maker submits and manages individual quotes, while on an AMM, all liquidity is *pooled*. Gains and losses are shared pro rata across all

[4]The block producer's ordering discretion is the blockchain analog of the latency advantages studied in traditional market microstructure (Budish, Cramton, and Shim 2015b; Budish, Cramton, and Shim 2015c). The key difference is that on blockchain, the ordering power is explicit and per-block, whereas on traditional exchanges, intermediary advantages arise from physical proximity and processing speed.

[5]DEX spot volume reached approximately one-fifth of centralized exchange (CEX) spot volume by late 2025, with individual platforms such as Uniswap processing over $1 trillion in annual volume.

[6]Common fee tiers on Uniswap v3 are 1, 5, 30, and 100 basis points. Lower-fee pools attract more volume but expose LPs to tighter pricing and greater arbitrage intensity.

providers, regardless of whether a given provider actively monitors the market.

Pricing and the constant product rule. The AMM determines exchange rates algorithmically through a *pricing function* that relates the pool's token reserves. The most widely used specification is the constant product rule, $x \cdot y = k$, where x and y denote the quantities of tokens A and B in the pool and k is an invariant set by the total liquidity deposited (Capponi and Jia forthcoming). Any trade that withdraws Δ_B units of token B must supply Δ_A units of token A such that the product of the post-trade reserves equals k. The exchange rate is therefore determined by the pool's current reserves: the marginal rate equals the ratio of reserves, y/x, and each trade moves the pool along the pricing curve, generating price impact proportional to trade size relative to pool depth.

This automated pricing mechanism has two consequences for market quality. First, the AMM always offers a quote—it cannot "go dark" or withdraw liquidity unilaterally—which guarantees continuous availability but also means the pool trades at stale prices when external market conditions change between blocks. Second, unlike a limit order book where market makers actively update their quotes in response to news, the AMM's price adjusts only when a trade executes against the pool. Between trades, the AMM price remains fixed at its last-traded level, even as the asset's fundamental value changes on external venues.

Adverse selection and arbitrage. The staleness of AMM pricing creates a systematic adverse-selection cost for liquidity providers. When the fundamental value of a token changes—for instance, because of a price movement on a centralized exchange such as Binance—the AMM pool continues to offer the old price until a trade corrects it. Arbitrageurs monitor the deviation between the AMM price and external reference prices, and when this deviation exceeds the pool's trading fee, they submit a transaction to trade against the stale quote. The arbitrageur profits from the price discrepancy, while the pool—and therefore all LPs in proportion to their shares—absorbs a loss equal to the difference between the AMM execution price and the current fundamental value.

This adverse-selection cost is the primary friction in liquidity provision on AMMs (Capponi

and Jia forthcoming; Lehar and Parlour 2023; Milionis, Moallemi, Roughgarden, and Zhang 2023). The loss is larger when the time between blocks is longer, because longer intervals allow larger price deviations to accumulate before the pool can be corrected. The loss is also shared pro rata: if one LP owns 10% of the pool, she bears 10% of every arbitrage loss, regardless of whether she could have anticipated or avoided that particular trade. This pooled loss-sharing structure means that a passive LP who deposits capital and does not monitor external prices bears the same per-unit adverse-selection cost as an active LP who watches the market continuously.

Active versus passive provision. A passive liquidity provider deposits capital into the pool and leaves it for an extended period—days, weeks, or longer—earning fees continuously but bearing all adverse-selection losses that occur during that interval. An active provider, by contrast, monitors external prices and adjusts her position frequently: she deposits liquidity to earn fees when conditions are favorable and withdraws it when a price gap signals imminent adverse selection, then re-enters once the gap closes. This operational distinction—the frequency and speed of deposit-and-withdrawal cycling—is orthogonal to the range of prices over which liquidity is supplied.[7] An active LP who repositions every few minutes resembles a traditional high-frequency market maker more than a passive capital supplier. The feasibility and profitability of this active strategy depend on the cost and speed of submitting transactions, both of which are directly governed by the blockchain's execution clock and sequencing rules. We refer readers to Capponi and Jia (forthcoming) for a comprehensive treatment of the institutional details of liquidity provision on blockchain-based exchanges.

[7]Recent AMM designs, notably Uniswap v3, allow LPs to specify a concentrated price range (Hasbrouck, Rivera, and Saleh 2022). Concentrated liquidity increases capital efficiency but does not by itself determine whether a provider is active or passive in the operational sense we describe. An LP can supply concentrated liquidity passively (set a range and leave) or actively (reposition the range every few minutes). Our model's distinction between strategic and passive LPs concerns the latter dimension: the frequency and feasibility of position adjustment.

2.3 Layer 2 Networks and the Design of Execution Speed

Ethereum's base layer faces three constraints that limit its suitability as trading infrastructure. First, blocks are produced at fixed 12-second intervals, so an AMM pool can reprice at most five times per minute—while external reference prices on centralized exchanges update thousands of times per second. The resulting stale-price windows create large adverse-selection costs for liquidity providers. Second, Ethereum's block space is scarce: each block can accommodate a limited number of transactions, and demand for inclusion drives gas fees to levels that make small or frequent trades uneconomical. Third, the 12-second cadence imposes a hard lower bound on the latency of any on-chain application, making real-time responsive trading infeasible.

These constraints have motivated the development of Layer 2 (L2) networks. An L2 network is a separate execution environment that runs on top of Ethereum: it processes transactions on its own chain—with its own block time, throughput capacity, and sequencing rules—and periodically posts a compressed summary of the resulting state changes back to Ethereum for final settlement and security. By operating a faster execution clock than the base layer, an L2 can reduce stale-price windows, lower transaction fees through higher throughput, and offer a more responsive trading environment. The tradeoff is that L2 execution is typically managed by a single entity—the *sequencer*—who controls block production and transaction ordering, rather than by the decentralized validator set of Ethereum itself. L2 networks have compressed execution clocks by one to two orders of magnitude, reducing stale-price windows but simultaneously altering the competitive dynamics among traders and liquidity providers.

Each L2 chain bundles two design choices: an execution clock (the block interval) and a sequencing rule (how transactions are ordered within each block). Under a *priority gas auction* (PGA), the sequencer ranks transactions by fee bid, highest first, so that agents who value execution priority can purchase it directly; the sequencer captures the bid revenue. Under *first-come-first-served* (FCFS) ordering, the sequencer commits to processing transactions in the order they arrive, which shifts competition from fee bidding toward latency investment. The economic im-

plications differ: PGA ordering channels rents to the infrastructure operator, while FCFS ordering dissipates rents through private speed expenditure.

The variation across chains illustrates the active design experimentation in this space. Base operates a centralized sequencer with PGA ordering at an effective 200-millisecond execution clock.[8] Arbitrum operates a centralized sequencer with FCFS ordering at 250-millisecond block times; TimeBoost overlays a second-price sealed-bid auction that sells a 200-millisecond priority advantage for 60-second rounds, hybridizing FCFS with auction-based priority.[9] Each combination of clock and sequencing rule implies a different allocation of rents among traders, liquidity providers, and infrastructure operators—the central tradeoff that the theoretical analysis formalizes.

2.4 Block Time as Shared Infrastructure

On a traditional exchange, each venue controls its own matching engine and latency infrastructure: a change at one exchange affects only the markets traded there. On a blockchain, execution infrastructure is shared. A single L2 chain's block time simultaneously determines the effective trading speed for every DEX pool, lending protocol, derivatives market, and decentralized finance (DeFi) application deployed on that chain.

The scale of this shared infrastructure is substantial. Base alone hosts over 512,000 Uniswap v2 liquidity pools and more than 22 distinct DEX protocols; Arbitrum hosts 39 DEX protocols and 65 DeFi applications. Every one of these markets operates under the same block cadence. When Base reduced its effective execution clock from 2 seconds to 200 milliseconds, the change simultaneously affected execution quality, LP adverse selection exposure, and arbitrage dynamics across all deployed markets. This feature makes the block cadence a *public infrastructure design choice*: the infrastructure operator's decision is not analogous to a single exchange choosing its match-

[8] Base's standard 2-second block interval was reduced to 200 milliseconds via Flashblocks, introduced on mainnet on July 7, 2025. Flashblocks stream sub-blocks within each sealed slot, locking transaction ordering every 200 milliseconds while preserving the PGA rule inside each sub-block. Each sub-block commits ordering irrevocably upon broadcast, so the economically relevant execution clock is the 200-millisecond interval.

[9] TimeBoost launched on Arbitrum on April 17, 2025 and generated approximately $2 million in auction revenue in its first three months of operation.

ing engine speed, but rather to a regulator setting the minimum tick size or trading frequency for an entire class of markets. Table 1 summarizes the key differences between L1 and L2 trading environments.

Table 1. Comparison of L1 and L2 decentralized exchange environments

Dimension	L1 (Ethereum mainnet)	L2 (Base, Arbitrum)
Execution clock	Fixed 12-second blocks	200–250ms blocks
Clock authority	Determined by consensus protocol; exogenous to any application	Sequencer operator chooses interval as design parameter
Transaction ordering	Validator/builder ordering; fee-based priority	Chain-specific: PGA, FCFS, or hybrid
Stale-price window	Long; AMM reprices at most 5×/min	Short; repricing up to 5×/sec
LP adjustment cost	High gas fees; repositioning expensive	Low fees; frequent repositioning feasible
Rent channel	Gas fees and validator/MEV extraction	Sequencer fees, private speed investment, or hybrid auction
Infrastructure redesign	Requires protocol-wide consensus	Operator can unilaterally modify

2.5 The Base Flashblocks Natural Experiment

On July 7, 2025, Base introduced Flashblocks on mainnet, reducing the effective execution clock from 2 seconds to 200 milliseconds while preserving the PGA ordering rule within each sub-block. Arbitrum, operating under FCFS ordering at 250-millisecond blocks, did not implement a comparable change during the sample period. This combination—a discrete, dated change in block speed on one chain with a stable comparison chain—provides the quasi-experimental variation exploited in Section 6.

3 Model

Our model consists of a repeated block-interval game on a blockchain automated market maker with four types of participants: two arbitrageurs who exploit stale pool prices, a strategic liquidity provider who actively manages her pool position, a competitive fringe of passive liquidity providers who supply capital but cannot rebalance within a block, and a sequencer who monetizes transaction ordering. The block production rate λ_B governs how frequently the market clears and determines both the intensity of within-block competition and the LP's ability to manage adverse-selection exposure. Retail trading demand is not modeled strategically; whenever liquidity is posted, retail flow generates fee income at an exogenous rate $\bar{r} > 0$, so the paper's focus is how block speed changes the division of fee income and adverse-selection losses across liquidity suppliers.

A single risky asset trades on an AMM hosted on a blockchain. Time is continuous, but transactions execute only when a block is produced. Blocks arrive as a Poisson process with rate $\lambda_B > 0$, so the expected block interval is $1/\lambda_B$. The block rate is set by the protocol operator and is exogenous from the viewpoint of market participants.

3.1 Trading Environment and Players

There is an external reference price for the asset, which we interpret as the efficient price observed on broader markets. The AMM cannot update its price continuously; it updates only when someone trades against it or removes liquidity. Between blocks, public information may move the efficient price, but the on-chain quote remains stale until the next relevant transaction executes. This is the source of stale-quote risk on blockchain.

Arbitrageurs. We think of arbitrageurs as low-latency traders who monitor public price discrepancies between the AMM and the external market and race to capture stale-quote profits. There are two arbitrageurs, A_1 and A_2. Two is the smallest number that generates contested blocks and sequencer rent extraction; with only one arbitrageur there is no within-block priority auction. Be-

fore trading begins, each arbitrageur chooses a speed technology $s_i \in \{S, F\}$. The corresponding message-arrival intensity is

$$\lambda(s_i) \in \{\lambda_S, \lambda_F\}, \qquad 0 < \lambda_S < \lambda_F.$$

Adopting the fast technology entails a flow cost $K > 0$. Speed serves two purposes: it increases the probability of landing a transaction before the block boundary, and it increases the probability of arriving before one's rival in contested blocks.

Strategic liquidity provider. We think of the strategic LP as a sophisticated liquidity manager—for example, a professional market-making firm that can deposit capital into and withdraw capital from the pool in response to public price movements. There is one strategic LP with inventory $q_L > 0$. When her capital is deposited in the pool, she supplies depth q_L and earns her pro rata share of retail fee flow. When she has withdrawn, she earns no fee income but is protected from arbitrage losses. Like the arbitrageurs, the strategic LP chooses a speed technology $s_L \in \{S, F\}$ at flow cost K, which governs how quickly her withdrawal transaction reaches the sequencer. She also chooses how aggressively to manage her position by setting a withdrawal threshold $d^* \in \{M, H\}$: a threshold of H means she waits until arbitrage is profitable before attempting to withdraw; a threshold of M means she withdraws preemptively at the warning state, before an arbitrage opportunity materializes. Her central tradeoff is between earning fees by remaining in the pool and avoiding adverse selection by withdrawing when prices move. The strategic LP's outside option is normalized to zero.

Passive liquidity providers. A competitive fringe of passive LPs can either supply aggregate depth $Q^P = \bar{Q}$ or stay out entirely, so $Q^P \in \{0, \bar{Q}\}$. We think of this fringe as slower capital that is willing to post liquidity but cannot manage it within a block. Passive LPs do not choose a speed technology and do not submit state-contingent control transactions. Their participation decision isolates the extensive margin that is economically relevant in the data: does slow capital remain

in the pool once faster liquidity managers are present, or is it crowded out? The passive fringe's outside option is also normalized to zero.

Sequencer. A risk-neutral sequencer (equivalently, validator or block builder) collects all transactions that arrive before a block boundary and orders them by priority fee, highest first. The sequencer's role is central because blockchain execution is not merely batched; it is batched and then serialized by a revenue-collecting intermediary. When both arbitrageurs land in the same profitable block, they bid for priority, and the winning bid is transferred to the sequencer. At equal priority fees, an arbitrage transaction has priority over a withdrawal transaction, and blocks have capacity for all relevant transactions—so the LP's strategic variable is *when* to withdraw, not how much to bid. The sequencer has no strategic choice beyond the protocol rule of ordering by fee.

3.2 Price Gap and Information Arrival

There is an external reference price p_t^* for the asset, which we interpret as the efficient price observed on broader markets, and an AMM marginal price p_t^{AMM}. The AMM price updates only when a transaction executes against the pool; between blocks, public information may move the efficient price while the pool quote remains stale. The payoff-relevant state variable is the absolute price gap

$$d_t := |p_t^* - p_t^{\text{AMM}}|.$$

We discretize the gap into three levels:

$$d_t \in \{0, M, H\}, \qquad 0 < M < f < H,$$

where f is the AMM swap fee. The gap $d = 0$ means the pool price is aligned with the external price: no arbitrage opportunity exists. The gap $d = M$ is a warning level: the pool quote is stale, but the deviation is smaller than the fee, so arbitrage is not yet profitable. The existence of this intermediate level is what makes the LP's withdrawal decision nontrivial—an LP who withdraws

at M gives up fee income in a state where no arbitrageur will trade, but avoids the risk of a further jump to H. The gap $d = H$ means the deviation exceeds the fee and stale-quote arbitrage is profitable. Arbitrageurs trade only in this state, and each execution fully corrects the pool price, resetting the gap to $d = 0$. Arbitrage participation depends on the publicly observable gap, not on whether the strategic LP has capital deposited in the pool.

Public information arrives as a Poisson process with rate $\lambda_J > 0$. Because we track the absolute gap, one innovation moves the system one step:

$$0 \xrightarrow{\lambda_J} M, \qquad M \xrightarrow{\lambda_J/2} 0, \qquad M \xrightarrow{\lambda_J/2} H.$$

From 0, any price movement creates a small deviation. From M, a further same-direction jump deepens the gap to H, while a reversal restores alignment. Once arbitrage is profitable ($d = H$), only an arbitrage execution can reset the gap.

All agents observe the external price and therefore the gap state. There is no private information about fundamentals—adverse selection arises entirely from stale prices, not from informed trading. Transactions that do not arrive before the next block boundary expire and can be resubmitted, so there is no pending-message state. The only asymmetry across agents is technological: they differ in how quickly a message reaches the sequencer after a common public trigger.

3.3 Payoffs

Total pool depth is $Q = Q^P + q_L$ when the strategic LP has capital deposited and $Q = Q^P$ when she has withdrawn. The AMM's depth impact function is

$$\kappa(Q) = cQ. \tag{1}$$

The linear form is the minimal reduced form that preserves the key AMM feature: deeper pools scale both the stale-quote prize and the associated LP loss proportionally.

Arbitrageur payoff. At the dangerous state H, an arbitrage execution against depth Q generates net stale-quote surplus

$$V_{\text{arb}}(Q) = \kappa(Q)(H^2 - f^2) = cQ(H^2 - f^2). \tag{2}$$

Each term has a direct economic interpretation: $H^2 - f^2$ is the per-unit-depth profit from trading at a price that deviates by H from fundamentals while paying a fee f, and cQ scales this profit by pool depth. When both arbitrageurs land in the same block, they compete in a pay-on-landing first-price auction. The winner receives $V_{\text{arb}}(Q)$ minus her priority fee bid; the loser receives zero and pays nothing. Any agent who has adopted the fast technology pays the flow cost K while fast.

Strategic LP payoff. The strategic LP's per-period payoff has two components. While her capital is deposited in the pool, she earns fee income at rate

$$\sigma\bar{r}, \qquad \text{where } \sigma = \frac{q_L}{Q} \tag{3}$$

is her pool share. Each time an arbitrageur executes while she has capital deposited at $d = H$, she suffers a loss of

$$L_L(Q) = \sigma\, V_{\text{arb}}(Q) = \sigma\, cQ(H^2 - f^2). \tag{4}$$

The LP's long-run average payoff is therefore the difference between cumulative fee income and cumulative adverse-selection losses, less any speed-technology cost.

Passive LP payoff. When both the strategic LP and the passive fringe have capital deposited, the passive fringe earns a fee share of $\bar{Q}/(\bar{Q} + q_L)$ and bears the residual arbitrage loss proportional to $\bar{Q}$. Because the depth impact is linear, the passive LP's per-execution loss is invariant to whether the strategic LP has deposited or withdrawn—a property of the linear functional form that simplifies the passive fringe's participation condition to a comparison of expected fee income against expected adverse-selection losses.

Sequencer payoff. The sequencer collects the winning priority fee whenever at least one arbitrageur lands in a profitable block. Sequencer revenue per block is therefore the equilibrium bid of the winning arbitrageur, which depends on the within-block competition intensity and hence on the speed profile $\mathbf{s}$ and the block rate λ_B.

3.4 Timing and Stages of Play

Stage 1: Speed and threshold choice. Each arbitrageur and the strategic LP simultaneously choose a speed technology $s_i \in \{S, F\}$. The strategic LP also chooses a withdrawal threshold $d^* \in \{M, H\}$. The passive LP fringe chooses whether to supply $\bar{Q}$ or to stay out.

Stage 2: Repeated trading. After Stage 1, news, message arrivals, and block boundaries evolve continuously and independently. Within each block interval: (i) a public information shock may move the gap from 0 to M or from M to either 0 or H; (ii) whenever the gap reaches a player's trigger, that player immediately submits a transaction—arbitrageurs submit at $d = H$, the strategic LP submits a withdrawal at d^*; (iii) the next block boundary arrives and the sequencer orders all arrived transactions by priority fee, executing them sequentially against the AMM; (iv) transactions that did not arrive before the boundary expire.

Stage 3: Participation. The passive LP's participation decision $Q^P \in \{0, \bar{Q}\}$ is based on the expected long-run payoff conditional on the speed profile and threshold chosen in Stage 1. The strategic LP's deposit decision is similarly evaluated against her outside option of zero.

3.5 Equilibrium Concept

Our solution concept is stationary subgame-perfect Nash equilibrium.

Definition 3.1 (Equilibrium)**.** *A* stationary subgame-perfect Nash equilibrium *is a profile of strategies* $(\mathbf{s}^*, d^{**}, Q^{P*})$*—consisting of speed choices for each arbitrageur and the strategic LP, a with-*

drawal threshold for the strategic LP, and a participation decision for the passive fringe—such that:

1. *Conditional on* $(\mathbf{s}^*, d^{**}, Q^{P*})$*, the within-block bidding strategies constitute a Nash equilibrium of the pay-on-landing first-price auction in every block with* $d = H$*.*
2. *The strategic LP's threshold* d^{**} *and speed* s_L^* *maximize her long-run average payoff given the arbitrageurs' strategies and pool depth.*
3. *Each arbitrageur's speed* s_i^* *maximizes her long-run average payoff given the other players' strategies.*
4. *The passive fringe participates (*$Q^{P*} = \bar{Q}$*) if and only if expected fee income weakly exceeds expected adverse-selection losses.*

4 Equilibrium Analysis

The model describes a game in which arbitrageurs, a strategic LP, passive LPs, and a sequencer interact within discrete block intervals. We now solve this game, tracing how block speed propagates through the market. Section 4.1 shows that block speed governs whether rents are allocated through within-block priority auctions or through inter-block speed races. Section 4.2 shows that block speed transforms LP behavior from passive provision to active market making. Section 4.3 characterizes the joint determination of speed investment and liquidity provision. Section 4.4 establishes conditions under which an interior block rate maximizes aggregate welfare.

4.1 Within-Block Competition

How does block speed determine who captures timing rents? On slow chains, both arbitrageurs almost certainly arrive before the block boundary, triggering intense priority fee bidding that transfers rents to the sequencer. On fast chains, at most one arbitrageur typically arrives per block, so priority competition is rare and rents stay with the arbitrageur. The block rate λ_B continuously interpolates between these two regimes through its effect on the probability that multiple agents

arrive in the same block.

The probability that both arbitrageurs arrive in the same block—and therefore compete for execution priority—depends on block speed. Because blocks arrive as a Poisson process and transaction latencies are independent exponentials, the joint arrival probabilities have closed-form expressions.

Proposition 4.1 (Arrival Probabilities)**.** *For block rate λ_B and arbitrageur speeds $\lambda_1, \lambda_2 > 0$, the joint probabilities that each combination of agents arrives before the block boundary are:*

$$p_{00} = \frac{\lambda_B}{\lambda_B + \lambda_1 + \lambda_2}, \tag{5}$$

$$p_{11} = 1 - \frac{\lambda_B}{\lambda_B + \lambda_1} - \frac{\lambda_B}{\lambda_B + \lambda_2} + \frac{\lambda_B}{\lambda_B + \lambda_1 + \lambda_2}, \tag{6}$$

$$p_{10} = \frac{\lambda_B \lambda_1}{(\lambda_B + \lambda_2)(\lambda_B + \lambda_1 + \lambda_2)}, \tag{7}$$

$$p_{01} = \frac{\lambda_B \lambda_2}{(\lambda_B + \lambda_1)(\lambda_B + \lambda_1 + \lambda_2)}. \tag{8}$$

Let $q := p_{11}$ denote the collision probability (both arbs present), $c_i := p_{i0}/q$ the ratio of uncontested to contested blocks for arb i, and $P_{\geq 1} := 1 - p_{00}$ the probability that at least one arb arrives.

The collision probability q measures how often both arbitrageurs are present in the same block and must compete for priority. The ratio c_i measures how often arb i enjoys an uncontested block relative to a contested block. Together, these objects determine both the intensity of priority bidding and the share of rents captured by the sequencer.

The limiting behavior reveals the asymmetry that drives all subsequent results. On slow chains ($\lambda_B \to 0$), blocks are long enough that both arbs almost certainly submit transactions before the block boundary, so every block is contested ($q \to 1$). On fast chains ($\lambda_B \to \infty$), the probability of any arb arriving before the boundary vanishes, and the probability of a collision vanishes even faster ($q = O(\lambda_B^{-2})$). Under symmetry ($\lambda_1 = \lambda_2 = \lambda$), the probability that at least one arb arrives simplifies to $P_{\geq 1} = 2\lambda/(\lambda_B + 2\lambda)$.

Because block arrivals are memoryless, delaying submission can only reduce the probability

of arriving before the next block, so immediate submission is optimal (Appendix A). In practice, this corresponds to the empirical observation that arbitrageurs and active LPs submit transactions as soon as they observe an opportunity, with transactions expiring if not included in the next block.

When both arbitrageurs arrive in the same block, they compete by bidding priority fees in a first-price auction. The winner pays its bid and captures the arbitrage profit; the loser pays nothing. This mechanism corresponds to the priority gas auction used by Base and other L2 chains.

Proposition 4.2 (Priority-Fee Auction Equilibrium)**.** *Each arbitrageur's expected payoff equals its probability of arriving alone times the full prize, and the sequencer captures all rents from contested blocks. Under symmetric speeds ($\lambda_1 = \lambda_2$),*

$$\Pi_i = p_{i0}\, V, \qquad E[R] = qV. \tag{9}$$

Under asymmetric speeds, expected priority fee revenue is

$$E[R] = qV \cdot \frac{1 + \min(c_1, c_2)}{1 + \max(c_1, c_2)}. \tag{10}$$

The intuition mirrors Bertrand competition. In contested blocks (probability q), intense bidding drives priority fees up to the full prize value, transferring nearly all rents to the sequencer. In uncontested blocks (probability p_{i0}), the arb faces no competition and bids minimally, keeping the entire prize. The arb's expected profit therefore equals the value of arriving alone, which happens precisely when the rival fails to arrive.

On slow chains where $q \approx 1$, nearly every block is contested, so arbitrage profits are almost entirely converted into priority fees—infrastructure rents captured by the sequencer. This is consistent with the empirical finding in Capponi and Jia (forthcoming) that DEX arbitrageurs transfer approximately 96% of their profits to validators as gas fees on Ethereum L1, where 12-second block times ensure that multiple arbs are almost always present.

The LP's optimal priority bid is zero (Appendix A). The economic reason is a *tragedy of the*

commons: arbitrage losses are distributed across all liquidity providers proportional to their pool shares, but the cost of preempting arbitrageurs through priority bidding falls on the individual LP. To preempt, an LP would need to outbid the arbs' expected profits, but this cost exceeds any individual LP's share of the potential loss. No individual LP has sufficient incentive to pay the full cost of preemption, even though collective preemption would be valuable.[10]

Combining the arrival probabilities with the priority-fee auction yields a closed-form expression for the sequencer's share of arbitrage rents as a function of block speed.

Proposition 4.3 (Sequencer Rent Share)**.** *The sequencer's share of arbitrage rents is strictly decreasing in block speed. Under symmetric arb speeds, the rent share is*

$$s(\lambda_B) = \frac{\lambda}{\lambda + \lambda_B},$$

with $s \to 1$ *as* $\lambda_B \to 0$ *and* $s \to 0$ *as* $\lambda_B \to \infty$.

Faster blocks reduce the collision probability q, weakening competition among arbitrageurs and lowering the share of rents captured as priority fees. As $\lambda_B \to 0$, nearly all arbitrage rents are captured by the sequencer ($s \to 1$). As $\lambda_B \to \infty$, priority fees collapse because arbs rarely compete in the same block ($s \to 0$). One might expect that faster blocks would benefit the sequencer by increasing throughput. The countervailing force is that faster blocks dilute per-block competition, reducing the intensity of the priority fee bidding. The net effect is a strict decrease in the *share* of rents captured per execution event, even though the *frequency* of execution events may increase. This prediction is directly testable by comparing priority fee patterns across L2 chains with different block times, as we do in Section 6.3.[11]

[10]This contrasts with centralized exchanges, where fast liquidity providers can cancel limit orders microseconds before an informed trade arrives (Budish, Cramton, and Shim 2015a). On blockchain, the LP's withdrawal transaction is included in the same block as the arb's trade, and the block's ordering is determined by priority fees. The pooled nature of AMM liquidity makes individual LP defense unprofitable in a way that has no direct analog in traditional limit order books.

[11]Despite the decentralized nature of blockchain infrastructure, infrastructure rents in DEXs can systematically exceed those in traditional markets. While centralized exchanges capture approximately 30% of speed rents through co-location fees (Budish, Lee, and Shim 2024), the sequencer captures up to $s(\lambda_B)\cdot 100\%$ of arbitrage rents. This stems from two factors: (i) the block-by-block first-price auction creates more intense competition than continuous-time co-

The results in this subsection establish that slow blocks channel rents into priority fee auctions where the sequencer extracts, while fast blocks shift competition to inter-block speed races where private investment is dissipated. The block rate λ_B therefore does not merely scale the intensity of competition—it selects the *form* of competition.

Within-block competition determines how arbitrage rents are divided between traders and infrastructure. The next question is how block speed affects who supplies the liquidity that these rents are extracted from.

4.2 LP Behavior and Liquidity Composition

How does block speed affect who supplies liquidity? The strategic LP faces a tradeoff between earning trading fees while her capital is deposited and avoiding adverse-selection losses by withdrawing before arbitrageurs execute. Her pool position and the price gap jointly evolve as a six-state dynamic system: at each block boundary, the within-block race determines whether an arbitrage execution, a safe withdrawal, or neither occurs. Three rates summarize this interaction: the rate at which stale prices are corrected through arbitrage (A), the rate at which the LP can withdraw safely at the high gap (h), and the rate at which the LP can withdraw preemptively at the warning level (m). All three are closed-form functions of block speed and the speed profile (Appendix B). The long-run fraction of time the LP spends deposited in the pool, π_I, is therefore a closed-form function of λ_B—the workhorse object for all subsequent results.

Block speed affects the LP through two channels: the arbitrage execution rate, which governs how quickly stale prices are corrected, and the withdrawal rates, which govern how effectively the LP can manage her position. Faster blocks increase both—more frequent price correction but also easier withdrawal—creating the fundamental tension that this subsection resolves.

Proposition 4.4 (Long-Run Participation Rate)**.** *The LP's long-run participation rate π_I—the fraction of time she has capital deposited in the pool—is a closed-form function of block speed. It*

location, and (ii) decentralized liquidity provision prevents LPs from coordinating to contest the sequencer's monopoly, unlike traditional exchanges where market makers can negotiate infrastructure fees by threatening to redirect order flow.

depends on λ_B through the arbitrage execution rate A and the LP's withdrawal rates h and m. Closed-form expressions for all six state probabilities appear in Appendix B.

The long-run participation rate is the workhorse object for all subsequent results. It maps the block rate to the fraction of time the LP spends deposited, and hence to her average payoff, pool depth, and welfare. The block rate enters through two channels: the rate at which arbitrage trades execute (which governs how quickly stale prices are corrected) and the rate at which the LP can withdraw safely (which governs how effectively she can manage adverse-selection exposure). All economic quantities derived from this object are therefore closed-form functions of the block rate.

Slow blocks trap LPs. When blocks are slow, the LP cannot withdraw in time to avoid adverse selection. The LP's safe-exit rate vanishes faster than the arbitrage execution rate, trapping the LP in the pool.

Proposition 4.5 (LP Trapping on Slow Chains)**.** *On sufficiently slow chains, the strategic LP is effectively unable to withdraw before adverse selection. As blocks slow down ($\lambda_B \to 0$), the probability of withdrawing safely vanishes, and the long-run participation rate converges to one: the LP is trapped in the pool.*

The mechanism is an order-of-magnitude imbalance. The LP's safe-withdrawal rate requires two events to coincide within a single block: her transaction must arrive, and no arbitrageur's transaction can arrive first. This rate therefore vanishes quadratically in the block rate. The arbitrage execution rate, by contrast, requires only that at least one arbitrageur arrives, and vanishes linearly. As blocks slow down, the LP's withdrawal attempts are almost always preempted by arbitrage executions. The ratio of safe withdrawals to total arbitrage events converges to zero.

The consequence is paradoxical. The long-run distribution concentrates on deposited states not because the LP *wants* to remain in the pool, but because she *cannot withdraw* before being adversely selected. On chains with sufficiently long block times, the technology for LP self-protection is rendered ineffective, and the pool functions as if all liquidity were passive.

Remark 4.1 (Connection to Menkveld, Yueshen, and Zhu (2017))**.** *Menkveld, Yueshen, and Zhu (2017) study exchange latency in a limit-order-book market and show that faster venues can either improve or worsen liquidity. The mechanism differs from ours: in their model, speed works through the probability of rent-bearing states in the recursive equilibrium. In ours, block speed changes the LP's ability to exit before being adversely selected—the operative channel is the ratio of safe-exit rate to arbitrage execution rate, a block-boundary object absent from continuous-time models.*

Faster blocks enable cream-skimming. When blocks are fast enough, the strategic LP can withdraw before adverse selection materializes, capturing fee income while avoiding the losses that passive providers bear. This selective participation displaces passive capital.

Proposition 4.6 (Cream-Skimming)**.** *Whenever the LP's safe-withdrawal rate is positive, there exists a non-empty range of fee rates on which the strategic LP earns positive expected profit while a passive LP earns negative expected profit. The exact bounds on this cream-skimming interval are derived in Appendix A.*

The strategic LP exploits two advantages. First, she collects fee income while her capital is deposited, just as the passive LP does. Second, she can withdraw before adverse selection hits: when the gap reaches the profitable-arbitrage level, her speed technology gives her a positive probability of withdrawing before arbitrageurs execute. The passive LP, by contrast, remains exposed to every arbitrage execution.

The strategic LP imposes a negative externality on the passive LP through *fee dilution*. When the strategic LP deposits capital, the passive LP's share of fee income falls, but her per-execution adverse-selection loss is unchanged—a consequence of the linear depth-impact function, under which the passive LP's loss per arbitrage execution is invariant to the strategic LP's participation. The strategic LP captures fees while withdrawing before bearing the adverse-selection cost that the passive LP cannot avoid.

On chains with faster block times, the LP exit hazard h is larger, widening the cream-skimming interval. The model therefore predicts that faster chains should exhibit higher concentration of

LP activity among sophisticated, high-frequency rebalancers—precisely the pattern documented in our empirical analysis (Section 6.4).

Remark 4.2 (Connection to Malinova and Park (2023))**.** *Malinova and Park (2023) show that AMM pooled liquidity provides risk-sharing among heterogeneous LPs: adverse-selection losses are diluted across the pool, so even uninformed LPs can participate profitably. Our analysis reveals that this risk-sharing benefit depends on infrastructure speed. When blocks are slow, all LPs are trapped equally (Proposition 4.5), so losses are shared and no LP can strategically exit. As block speed increases, the strategic LP acquires a timing advantage that allows her to avoid the shared losses while still collecting fee income. The distribution of who bears losses shifts toward passive capital, and Malinova and Park's risk-sharing degrades into a subsidy from passive to active LPs.*

Passive LP participation. Whether passive providers remain in the market depends on whether fee income exceeds the combination of adverse-selection losses and fee dilution from strategic LPs.

Proposition 4.7 (Passive LP Participation)**.** *The passive LP participates ($Q^P = \bar{Q} > 0$) if and only if expected fee income exceeds expected adverse-selection losses:*

$$\bar{r}\left[\pi_O + \pi_I \frac{\bar{Q}}{\bar{Q} + q_L}\right] \geq A\pi_H \cdot c\bar{Q}(H^2 - f^2).$$

The left side is the passive LP's expected fee income, which reflects fee dilution: when the strategic LP has capital deposited, the passive LP earns a smaller share of the fee flow. The right side is the passive LP's expected adverse-selection loss: the rate at which arbitrage executions occur while the gap is at the profitable level, times the per-execution loss.

The participation condition tightens as the strategic LP becomes more active, because fee dilution increases while per-execution losses remain unchanged. Faster blocks raise the strategic LP's participation rate, which dilutes passive fee income, but also reduce the long-run probability of being at the profitable-arbitrage gap, which lowers adverse-selection losses. Whether passive

participation expands or contracts depends on which force dominates—a tension that the interior optimum resolves.

Remark 4.3 (Connection to Du and Zhu (2017a))**.** *Du and Zhu (2017a) derive an interior optimal trading frequency in repeated double auctions, where higher frequency trades off faster response to news against thinner demand schedules. In our model, the execution clock trades off LP protection (higher λ_B increases P_{safe}) against fee dilution and cream-skimming (higher λ_B increases π_I, diluting passive income). The welfare object in our setting includes not only gains from trade but the endogenous supply and allocation of liquidity—a dimension absent from double-auction models where market-making is exogenous.*

Block speed therefore transforms liquidity provision. On slow chains, all LPs are trapped equally and the pool functions as passive infrastructure. As blocks become faster, strategic LPs acquire the ability to time their participation, capturing fees while avoiding losses. This transforms passive provision into active market making and changes the composition of who supplies liquidity—a channel that has no analog in continuous-time models of traditional exchanges.

LP behavior and liquidity composition depend on the withdrawal rates, which in turn depend on agents' speed choices. The question then becomes: how much do agents invest in speed?

4.3 Speed Investment

Each agent—both arbitrageurs and the strategic LP—decides whether to pay a flow cost K to upgrade from slow to fast, comparing the benefit of faster arrival against the cost. The benefit depends on block speed through all the preceding equilibrium objects: arrival probabilities, the priority-fee auction, and the LP's participation.

Proposition 4.8 (Joint Speed Equilibrium)**.** *Speed investment and liquidity provision are jointly determined. There exist closed-form thresholds $\underline{K}_A(\lambda_B)$ and $\underline{K}_L(\lambda_B)$—the arbitrageur's and the LP's deviation gains from upgrading to fast technology—such that:*

(i) if $K > \max\{\underline{K}_A, \underline{K}_L\}$, no agent invests in speed;

(ii) if $K < \min\{\underline{K}_A, \underline{K}_L\}$*, all agents invest;*

(iii) for intermediate values, partial adoption arises: only the LP invests (K below the LP's threshold but above the arbitrageur's), or only the arbitrageurs invest (the reverse).

The thresholds depend on λ_B *through the arrival probabilities and the LP's participation rate. The equilibrium set is generically non-empty for every block rate.*

An arbitrageur's incentive to invest in speed depends not only on the block rate but also on the LP's speed choice, which affects the LP's withdrawal rate and therefore the effective pool depth available at each execution. The LP's incentive to invest in speed depends on arbitrageur speeds, which affect the execution rate she faces. This joint determination creates a rich equilibrium structure: different speed profiles can coexist at the same block rate for different ranges of K, and the selected profile changes as block speed varies.

Speed adoption can be non-monotone in block speed—a possibility result, not a guarantee. The non-monotonicity arises from two opposing channels. The *arrival-value channel* is straightforward: faster blocks create more execution opportunities per unit time, so the gross benefit of being fast increases with block speed. The *endogenous liquidity channel* works in the opposite direction: as blocks speed up and more agents invest in speed, the LP becomes better protected, which reduces the effective pool depth available at each execution. The net deviation gain therefore peaks at an intermediate block rate and can decline at high speeds. For a fixed speed cost K, the economy may transition from an all-slow regime to a mixed regime and back as block speed increases.

Whether speed investment eventually becomes unprofitable depends on whether LP protection at high block rates sufficiently depletes the arbitrage prize. A complementarity operates in the opposite direction: deep pools create large arbitrage prizes, which intensify speed investment. Liquidity and speed races are complements—the planner cannot sustain deep liquidity without some associated speed competition.

This complementarity generates two qualitatively distinct equilibrium families. When more agents invest in speed, the LP is better protected, which reduces the arbitrage prize and dimin-

ishes the pool's attractiveness to passive capital. Conversely, when agents remain slow, the pool retains high depth but leaks more value to the sequencer through priority competition. Observationally similar L2 chains may therefore settle into very different market structures—a low-speed, high-liquidity regime with larger pool depth but higher sequencer leakage, or a high-speed, low-liquidity regime with smaller arbitrage prizes and lower passive participation. The arbitrage prize is endogenous to block speed: unlike traditional models where the prize is fixed, the block rate changes both how much rent exists and through which channel it is contested.

Speed investment, LP participation, and sequencer extraction are jointly determined by the block rate. What block speed maximizes welfare?

4.4 Welfare and the Interior Optimum

Fee income, arbitrage prizes, and exchange fees are transfers within the market population and cancel in aggregate (Appendix A). Only three objects survive: the surplus from deeper pools, the real resource cost of speed investment, and the rents that flow to the sequencer—who is outside the welfare population.

Proposition 4.9 (Welfare Decomposition)**.** *Aggregate welfare decomposes into three components:*

$$W(\lambda_B) = \underbrace{\gamma\,\mathbb{E}[Q(\lambda_B)]}_{\text{gains from trade}} - \underbrace{K\,n_{\text{fast}}}_{\text{speed costs}} - \underbrace{S(\lambda_B)}_{\text{sequencer leakage}}, \tag{11}$$

where $\gamma > 0$ is the marginal retail surplus per unit of pool depth, n_{fast} counts fast agents, and $S(\lambda_B)$ is the sequencer's priority-fee revenue—rents that exit the welfare population.

Deeper pools generate more retail surplus per unit time, and average depth depends on the strategic LP's long-run participation rate. Speed costs are deadweight—real resources consumed to gain priority. Sequencer leakage is the share of arbitrage rents captured by the block producer through priority fees. The protocol designer's problem therefore reduces to maximizing pool depth net of speed waste and infrastructure extraction.

Neither the slowest nor the fastest block rate maximizes welfare.

Proposition 4.10 (Interior Welfare Maximizer)**.** *Under three conditions—slow blocks exclude the strategic LP, intermediate blocks generate welfare above the passive-only baseline, and speed costs dominate LP depth contributions at high block rates—welfare is maximized at an interior block rate. The formal conditions appear in Appendix A.*

The result formalizes a three-region picture. When blocks are too slow, the strategic LP does not enter: she is trapped (Proposition 4.5), and welfare is flat at the passive-only baseline. At intermediate block rates, the strategic LP enters and average pool depth jumps discretely, creating a welfare improvement because the depth benefit dominates the associated speed costs and sequencer leakage. When blocks are too fast, the LP retains positive participation but the deadweight cost of speed investment dominates the depth benefit, and welfare falls below the baseline.

The mechanism differs from the interior optimum in Du and Zhu (2017a), where faster trading frequency thins demand schedules. In our model, the welfare-relevant non-monotonicity arises from three channels that interact through blockchain market structure. The *LP participation channel* determines whether the strategic LP enters at all, creating a discrete jump in pool depth—a consequence of pooled AMM liquidity, where the deposit decision is all-or-nothing. The *speed-cost channel* reflects the increasing deadweight cost of speed investment as faster blocks create more execution opportunities. The *rent leakage channel* captures the sequencer's share of arbitrage rents through priority fees, a transfer out of the welfare population that decreases with block speed but remains positive at every finite block rate.

Remark 4.4 (Connection to Budish, Cramton, and Shim (2015b))**.** *Budish, Cramton, and Shim (2015b) show that continuous-time serial processing creates socially wasteful speed rents, and propose frequent batch auctions as a remedy. L2 blockchains appear to implement this prescription—blocks are discrete, periodic clearings. But blockchain execution is batched on the outside and serial on the inside: within each block, transactions are ordered by priority fee, and the block producer sells this ordering as a monopolist. Discrete block production does not eliminate speed*

rents—it relocates them from an inter-transaction arms race to a within-block priority auction. The relevant design question is not whether to batch (blockchain already does), but how to set the execution clock when batching and monetized priority coexist.

5 Discussion

The equilibrium analysis establishes that block speed affects welfare through three channels: gains from trade, which accrue to traders and liquidity providers; speed costs, the aggregate expenditure on latency technology; and sequencer extraction, the share of arbitrage rents captured by the block producer through fee auctions. This section interprets these results, characterizes the conditions under which welfare is maximized at an interior block rate, and discusses implications for infrastructure design.

5.1 Three Regimes of the Welfare Curve

When is welfare maximized at an interior block rate? Welfare is not monotone in block speed. The non-monotonicity arises because gains from trade, speed costs, and sequencer leakage respond to block speed with different signs and different elasticities.

When blocks are slow, the strategic LP is trapped (Proposition 4.5): the probability of withdrawing safely vanishes, passive liquidity bears the full weight of adverse selection, and the sequencer captures nearly all arbitrage rents through the priority auction. In this region, faster blocks improve welfare by releasing the LP from the pool, allowing more liquidity to be deposited, and reducing per-block sequencer extraction. When blocks are fast, stale-price exposure shrinks, but competition migrates into private speed investment and a market structure that favors technologically sophisticated LPs (Proposition 4.6). Further speed no longer generates much new trade; instead it raises speed costs and intensifies cream-skimming.

Whether these two forces produce a clean interior maximum depends on the behavior of the endogenous liquidity channel at high block rates. Three regimes are possible.

In *Regime A* (passive exit), cream-skimming drives passive LPs out at very high block rates, collapsing average pool depth. The arbitrage prize falls enough to exhaust speed incentives, and the interior optimum obtains. This regime is most likely when passive LPs have attractive outside options or fee rates are low.

In *Regime B* (liquid, no speed), the endogenous liquidity channel is strong enough that the arbitrage prize falls below the speed-cost threshold at high block rates. No agent invests in speed, the pool remains liquid, and welfare is high but bounded. The interior optimum exists.

In *Regime C* (liquid, persistent speed), the passive LP base is sufficiently inelastic that the arbitrage prize remains high even at large block rates. Speed investment persists, and the welfare curve may not turn down. The interior optimum holds only if the speed-cost channel eventually dominates.

The too-slow and too-fast forces are both economically primitive. Whether they generate a clean interior maximum depends on which regime the economy enters at high block rates. The model identifies all three forces and the conditions separating the regimes, rather than asserting that the interior optimum holds universally.

5.2 Sequencing Rules and Welfare

How should an L2 operator allocate transaction priority? Each sequencing rule creates a different balance between protecting passive traders, extracting sequencer revenue, and controlling speed-investment waste.

Under first-come-first-served (FCFS) ordering, explicit sequencer leakage is limited, but the missing rent reappears as a hidden race for arrival priority. When blocks are slow, FCFS combines stale-price exposure and latency competition: LPs remain trapped while searchers race to be first in line. A priority gas auction (PGA) monetizes the race rather than hiding it. This can improve on FCFS when the alternative is duplicated submissions, but it unambiguously raises sequencer leakage. Because LPs cannot profitably outbid arbitrageurs (Appendix A), the priority auction

does not solve LP protection—it changes the rent-allocation channel.

Flashblocks shorten the risky window itself by dividing the ordering cadence into sub-intervals and locking earlier sub-blocks, reducing the stale-price opportunity per contest. But if pushed far enough, Flashblocks can re-create the too-fast region by favoring responsive LPs and encouraging private speed investment. Batch or fair ordering compresses both sequencer leakage and speed costs directly, but if the batch interval is too coarse or the mechanism impedes LP updates, the venue re-enters the too-slow region through a different door. TimeBoost operates as a hybrid: it lowers hidden speed cost by replacing repeated racing with an ahead-of-time auction, but raises leakage because queue priority is sold rather than earned through latency.

No sequencing rule is unconditionally best. Each creates a different balance among the three welfare components, and the optimal choice depends on which distortion—trapping, leakage, or speed waste—is most binding for a given chain.

5.3 Implications for Infrastructure Design

What should an L2 operator do? The equilibrium analysis yields three concrete design principles.

The policy lever is the pair (λ_B, ordering rule), not either object alone. A chain that speeds up while retaining pure FCFS may shift the distortion from leakage into hidden speed cost. A chain that slows down while adopting fair ordering may reduce rent dissipation but lose liquidity because LPs remain trapped. An operator choosing block time should jointly consider the execution cadence and the queue-allocation mechanism. Base's design illustrates this: it keeps 2-second sealed blocks but orders with 200-millisecond Flashblocks, separating the commitment interval from the economically meaningful ordering frequency.

The optimal clock depends on the asset mix. A DeFi-heavy chain with substantial external price discovery and a large passive-LP base has more to gain from leaving the too-slow region than a payments or gaming chain with comparatively little stale-price arbitrage. The immediate

implication is that there is no single millisecond target correct for all chains.

Current deployments are consistent with the model's predictions. Base launched Flashblocks at 200ms; Arbitrum's TimeBoost operates at 250ms with express-lane auctions; Optimism adopted Flashblocks at 250ms across the Superchain. Major L2s appear to be converging on sub-second effective ordering clocks in the 200–250 millisecond range rather than remaining at 2 seconds or racing to tens of milliseconds. This convergence is consistent with large welfare gains from leaving the too-slow region while marginal benefits of further speed diminish.

5.4 Connection to Traditional Market-Design Debates

How does the interior optimum in blockchain markets relate to classical arguments about batch auctions and speed bumps? The interior optimum result sharpens the batch-auction debate. Budish, Cramton, and Shim (2015b) argue that continuous serial processing creates mechanical arbitrage rents and propose frequent batch auctions as a remedy. Du and Zhu (2017b) derive an interior optimal trading frequency from the tradeoff between faster reaction to news and thinner auctions. The present analysis belongs to this non-monotonic tradition but changes the mechanism: on L2 AMMs, faster execution changes not only the nature of speculative races but also whether liquidity can remain in the market, and whether rents leak to a sequencer-owned queue. The welfare optimum is therefore not "use batches" or "avoid batches," but "choose the cadence and priority rule that jointly maximize liquidity-supply surplus."

This perspective also reframes the speed-bump debate. IEX imposes a 350-microsecond delay so the venue can update prices before stale quotes trade (Aoyagi 2019). The interior optimum reconciles the opposing positions: a delay can improve welfare if it suppresses costly speed races without materially pushing the market into the too-slow region. TimeBoost is a blockchain example of exactly this logic—it delays ordinary transactions by 200 milliseconds but sells an exemption to the delay.

The theoretical analysis identifies three mechanisms—LP trapping, sequencer extraction, and

cream-skimming—each with observable consequences. Can we detect them in blockchain data? The next section tests these predictions using the Base Flashblocks upgrade as a quasi-natural experiment.

6 Empirical Analysis

The equilibrium analysis generates three testable predictions: faster blocks compress stale-price windows, reallocate rents from the sequencer toward private speed returns, and shift the composition of liquidity supply toward sophisticated rebalancers. This section tests each prediction using the Base Flashblocks upgrade as a quasi-natural experiment. On July 7, 2025, Base reduced its effective block time from 2 seconds to 200 milliseconds—a tenfold increase in λ_B—while Arbitrum continued operating at 250-millisecond blocks. The analysis uses Arbitrum as a control chain and organizes results around the three predictions, in order of decreasing causal clarity.

Two scope limitations apply. First, the design uses a single pool (Uniswap v3 WETH–USDC 5bp) on a single event, so the evidence identifies a movement *away from the too-slow corner* consistent with Proposition 4.5, but does not trace the full welfare curve or identify λ_B^*. Second, predictions EP4 (hump-shaped speed adoption) and EP5 (regime clustering) require cross-sectional variation in block rates across many chains, which a single pre/post design cannot provide.

6.1 Institutional Shock, Data, and Identification

The Flashblocks upgrade. Base is an Ethereum L2 rollup operated by Coinbase with a centralized sequencer. Prior to July 7, 2025, Base produced blocks every 2 seconds, with transactions within each block ordered by a priority gas auction (PGA): users bid gas prices, and the sequencer places higher-bidding transactions earlier in the block. On July 7, 2025, Base introduced Flashblocks, co-designed with Flashbots, Uniswap Labs, and OP Labs. Flashblocks streams sub-blocks within each 2-second slot, reducing the effective execution interval to 200 milliseconds. The ordering rule within each sub-block remains PGA; Flashblocks changes only the frequency at which

new transaction batches are committed. In the language of the model, the upgrade increases λ_B by a factor of ten while preserving the PGA sequencing mechanism.

The control chain. Arbitrum is an Ethereum L2 rollup with a centralized sequencer that orders transactions on a first-come-first-served (FCFS) basis by arrival time at the sequencer. Unlike Base, Arbitrum does not use gas price auctions for transaction ordering, so arbitrageurs cannot reliably secure early block positions by bidding higher gas prices. On April 17, 2025, Arbitrum introduced TimeBoost, a second-price sealed-bid auction selling a 200-millisecond time advantage for 60-second rounds. This mechanism differs structurally from the Base PGA. Arbitrum serves as a control chain in our analysis because it did not implement a comparable block-time reduction during the sample period.

Sample periods. The *pre-upgrade* period spans May 1–July 6, 2025 (67 days), and the *post-upgrade* period spans July 7–September 30, 2025 (86 days), with July 7, 2025 marking the Base Flashblocks launch date.

Data. On-chain DEX data come from Allium. The dataset covers all Swap, Mint, and Burn events emitted by the WETH–USDC pool at the 5 basis point (0.05%) fee tier on Arbitrum and Base, over the full sample period. Uniswap v3 represents liquidity positions as ERC-721 non-fungible tokens (NFTs) managed by the Non-Fungible Position Manager (NFPM) contract. Each NFT carries a unique `tokenId` that serves as a position identifier, analogous to an order ID in a limit order book. We link pool-layer Mint and Burn events to IncreaseLiquidity and DecreaseLiquidity events emitted by the NFPM to track the full lifecycle of each liquidity position.

CEX reference prices come from Tardis. We use Binance ETHUSDC spot market Level-2 order book snapshots at 5-level depth, recorded at 100-microsecond precision throughout the sample period. The Binance mid-quote serves as the efficient-price benchmark.

Identification of arbitrageurs. A transaction is classified as an *arbitrage trade* if two conditions hold: (i) the CEX–DEX log price deviation at the time of the swap exceeds 5 basis points, and (ii) the trade direction is consistent with exploiting the deviation. Arbitrageur addresses are identified using the pre-upgrade period. On Base, we retain addresses from the top 10 by arb trading volume that satisfy median transaction index ≤ 10 and arb volume share $\geq 90\%$. On Arbitrum, we apply less restrictive thresholds (transaction index ≤ 20, arb volume share $\geq 45\%$) to the top 10 addresses by arb trading volume. The asymmetric thresholds reflect the different sequencing mechanisms: on Base, PGA concentrates arb activity in early block positions, while Arbitrum's FCFS ordering disperses it.

Identification of active LPs. Active, high-frequency liquidity providers are identified using the pre-upgrade period. We focus on addresses whose liquidity event history consists of *mint-burn sequences*—positions explicitly opened and subsequently closed with the same `tokenId`. From the top 10 addresses by total mint/burn activity, we retain those satisfying: 25th percentile of rebalancing time < 60 seconds, 25th percentile of price range width < 100 basis points, balanced lifecycle ($|\text{burn-to-mint ratio} - 1| < 0.5$), and order-to-trade ratio > 2. These criteria select addresses that actively manage concentrated positions—the empirical counterpart of the strategic LP in the model.

Identification strategy. We implement a standard difference-in-differences regression at the arbitrageur-day (or chain-day) level:

$$Y_{it} = \alpha + \beta_1 \, Post_t + \beta_2 \, Base_i + \beta_3 \, (Post_t \times Base_i) + \varepsilon_{it}, \tag{12}$$

where β_3 is the DiD coefficient of interest. For trader-level and profit outcomes, standard errors are two-way clustered by arbitrageur and date. For market quality outcomes (chain-day level), standard errors are clustered by date. Regression tables appear in the subsections below; robustness across fixed-effect specifications is in Appendix C.

6.2 Faster Sequencing Compresses Stale-Price Windows

Prediction. Proposition 4.5 implies that on slow chains, the pool spends more time at gap $d = H$ before arbitrage execution corrects the price: the LP is trapped because $P_{\text{safe}} \to 0$, and the stale-price window widens. Increasing λ_B raises the arb execution hazard $A = \lambda_B P_{\geq 1}$, accelerating the rate at which deviations are corrected. Three observable implications follow: the time-weighted average CEX–DEX price deviation should fall on Base after the upgrade, the duration during which the deviation exceeds a fixed threshold should decrease, and both improvements should be visible relative to Arbitrum.

Measurement. The block-end price deviation for each block is

$$\delta_b = |p_b^{\text{DEX}} - p_b^{\text{CEX}}| \times 10^4, \tag{13}$$

where p_b^{DEX} is the post-swap log price of the last swap in block b and p_b^{CEX} is the contemporaneous Binance mid-quote. The time-weighted average price deviation aggregates across blocks within a day:

$$\bar{\delta} = \frac{\sum_b \delta_b \cdot \Delta t_b}{\sum_b \Delta t_b}, \tag{14}$$

where Δt_b is the time elapsed until the next block. The duration with large price deviation counts the total daily minutes during which δ_b exceeds a fixed threshold:

$$T^{\text{large}} = \frac{1}{60} \sum_b \Delta t_b \cdot \mathbf{1}[\delta_b > \bar{\delta}], \tag{15}$$

reported for $\bar{\delta} = 6$ basis points, with robustness checks at 5, 7, and 8 basis points.

Results. Figure 1 shows that the time-weighted average CEX–DEX price deviation on Base declines after the Flashblocks upgrade, while the Arbitrum distribution remains stable. The compression is consistent with faster block production shortening the interval during which stale prices

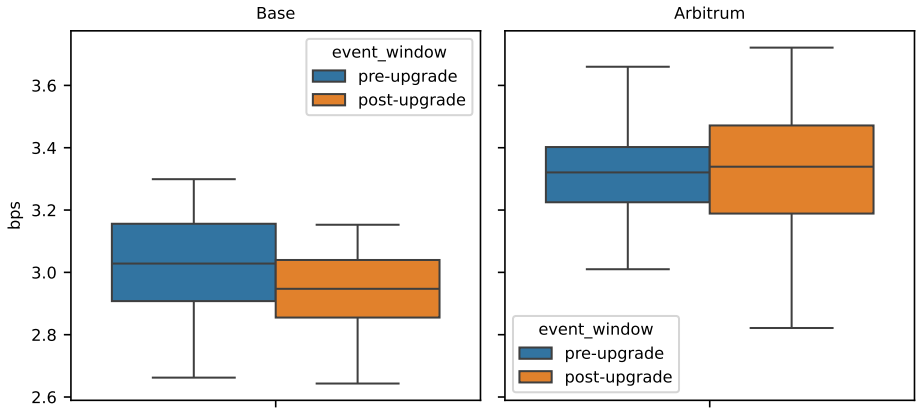


Figure 1. Time-weighted average CEX–DEX price deviation (bps) on Base and Arbitrum, pre- and post-upgrade. Each box plot shows the distribution across trading days. The post-upgrade compression on Base is consistent with faster blocks reducing stale-price exposure (Proposition 4.5).

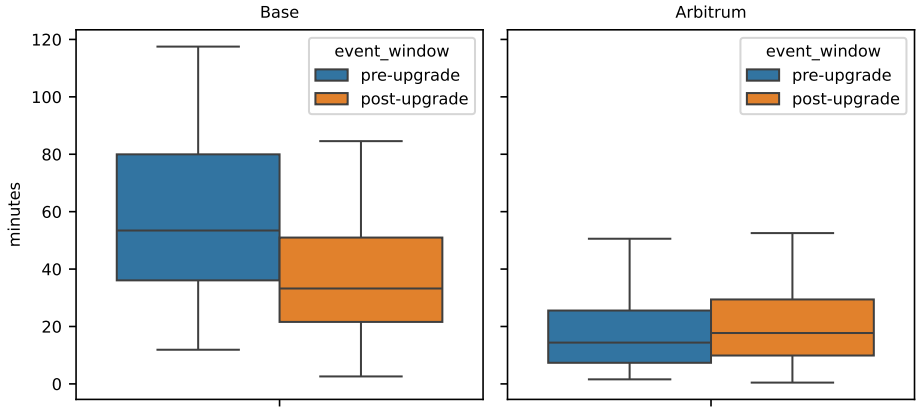


Figure 2. Daily duration with large CEX–DEX price deviation (>6 bps, in minutes) on Base and Arbitrum, pre- and post-upgrade. The reduction on Base indicates that the pool spends less time in the profitable-arbitrage region after Flashblocks.

persist on the AMM. Figure 2 shows the complementary intensive-margin result: the daily duration with large deviations (exceeding 6 basis points) falls on Base post-upgrade, indicating that the price gap not only narrows on average but also spends less time in the region where arbitrage is profitable. Both patterns are consistent with the model's prediction that higher λ_B increases the arb execution hazard $A = \lambda_B P_{\geq 1}$, accelerating the rate at which deviations are corrected.

Table 2 confirms the visual evidence. The Post × Base coefficient is negative for both metrics, indicating that Base experienced a significant reduction in the time-weighted average price deviation and in the duration of large deviations relative to Arbitrum after Flashblocks.

Table 2. DiD Estimates: Market Quality

	(1) TWA price dev. (bps)	(2) Dur. >6bps (min)
Post × Base	-0.0947***	-27.87***
	(0.0213)	(3.30)
Post	0.0159	4.45**
	(0.0322)	(2.25)
Base	-0.2828***	42.37***
	(0.0155)	(2.88)
Constant	3.3072***	16.93***
	(0.0210)	(1.40)
N	306	306
R^2	0.479	0.376

6.3 Faster Sequencing Changes Rent Allocation

Prediction. Proposition 4.3 shows that the sequencer rent share $s(\lambda_B) = \lambda/(\lambda + \lambda_B)$ is strictly decreasing in λ_B: on slow chains, intense within-block competition drives gas fees up to the full prize, transferring nearly all rents to the sequencer; on fast chains, the collision probability q falls and priority bidding weakens. After the Flashblocks upgrade, gas fees paid by arbitrageurs should fall, the decomposition of arb profits should shift away from blockchain fees and toward retained profits, and block-position dynamics should reflect reduced within-block competition.

Profit decomposition. All metrics are expressed in basis points. The price deviation at horizon τ is

$$\Delta p_\tau = (p^{\text{CEX}}_{t+\tau} - p^{\text{DEX}}) \times d \times 10^4, \tag{16}$$

where $d \in \{+1, -1\}$ is the trade direction. The effective spread is

$$\text{Eff. spread} = (p^{\text{exec}} - p^{\text{DEX}}) \times d \times 10^4. \tag{17}$$

Raw and net profits are:

$$\text{Raw profit}_\tau = \Delta p_\tau - \text{Eff. spread} - \text{Swap fee}, \tag{18}$$

$$\text{Net profit}_\tau = \text{Raw profit}_\tau - \text{Blockchain fee}. \tag{19}$$

The blockchain fee includes both L2 execution fees and L1 data posting costs. The gap between raw and net profit measures the total incidence of infrastructure costs on the arbitrageur.

Results. Figure 3 shows that effective gas fees paid by arbitrageurs on Base change after the Flashblocks upgrade, while Arbitrum gas fees remain stable. Figure 4 shows changes in block-position dynamics: arbitrageurs' transaction indices shift post-upgrade, reflecting the changed within-block competition environment under 200-millisecond sub-blocks.

Figures 5–9 present the full profit decomposition. The price deviation at the 10-second horizon (Figure 5) measures the gross arbitrage opportunity available at the time of the trade. The effective spread (Figure 6) captures the execution cost. Raw profit (Figure 7) nets out both the effective spread and the fixed swap fee, measuring the pre-infrastructure-cost profit margin. The blockchain fee (Figure 8) measures the total infrastructure cost. Net profit (Figure 9) is the residual.

The decomposition reveals that faster blocks do not simply compress arb profits uniformly—they change the *composition* of arb costs. This is consistent with the model's prediction that increasing λ_B shifts rents away from the sequencer (lower $s(\lambda_B)$) while potentially increasing returns to private speed investment. The evidence supports the view that faster infrastructure changes the *form* of rent allocation rather than eliminating rents entirely.

Table 3 reports DiD estimates for trader behavior outcomes. Arbitrageur gas fees and block position capture the change in within-block competition intensity; LP gas fees, block position, and rebalancing time capture LP adaptation. Table 4 reports estimates for arbitrage profit components.

Table 3. DiD Estimates: Trader Behavior

	Arbitrageurs		LPs		
	(1)	(2)	(3)	(4)	(5)
	Gas fee (gwei)	Block pos.	Gas fee (gwei)	Block pos.	Reb. time (min)
Post × Base	-2.73***	35.86***	0.01	20.51*	-25.46***
	(0.33)	(4.28)	(0.01)	(10.63)	(9.41)
Post	0.00	0.57	-0.00	4.93***	25.27***
	(0.00)	(0.89)	(0.01)	(1.78)	(9.40)
Base	2.90***	-5.28***	-0.02	64.88***	-27.27*
	(0.33)	(1.55)	(0.02)	(10.57)	(15.11)
Constant	0.01***	7.20***	0.03*	3.37**	27.56*
	(0.00)	(1.55)	(0.02)	(1.33)	(15.10)
N	1,074	1,074	367	367	367
R^2	0.856	0.289	0.010	0.854	0.019

Table 4. DiD Estimates: Arbitrageur Profit Metrics

	(1)	(2)	(3)	(4)	(5)
	Price dev.	Eff. spread	Raw profit	BC fee	Net profit
Post × Base	-1.3776**	-0.5786***	-0.7990**	-1.4645***	0.6655***
	(0.5474)	(0.2003)	(0.3988)	(0.2078)	(0.2496)
Post	-0.0383	-0.0788	0.0405	-0.0126	0.0531
	(0.1467)	(0.1146)	(0.1074)	(0.0091)	(0.1059)
Base	2.9477***	1.6807***	1.2670***	1.7648***	-0.4978***
	(0.5130)	(0.2191)	(0.3436)	(0.2083)	(0.1694)
Constant	6.3431***	0.4637***	0.8793***	0.0414***	0.8379***
	(0.2107)	(0.1560)	(0.1157)	(0.0116)	(0.1196)
N	1,348	1,348	1,348	1,348	1,348
R^2	0.085	0.405	0.016	0.741	0.004

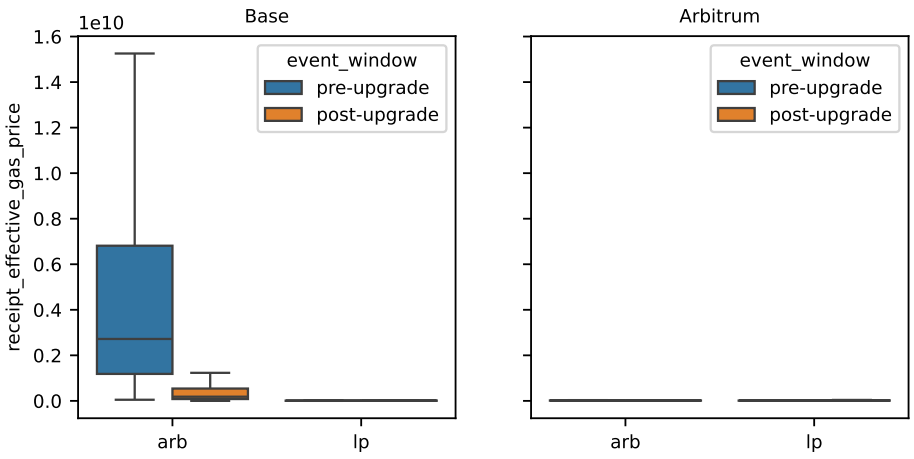


Figure 3. Effective gas fee for arbitrageurs and LPs, pre- and post-upgrade. Changes on Base reflect the altered within-block competition environment under Flashblocks.

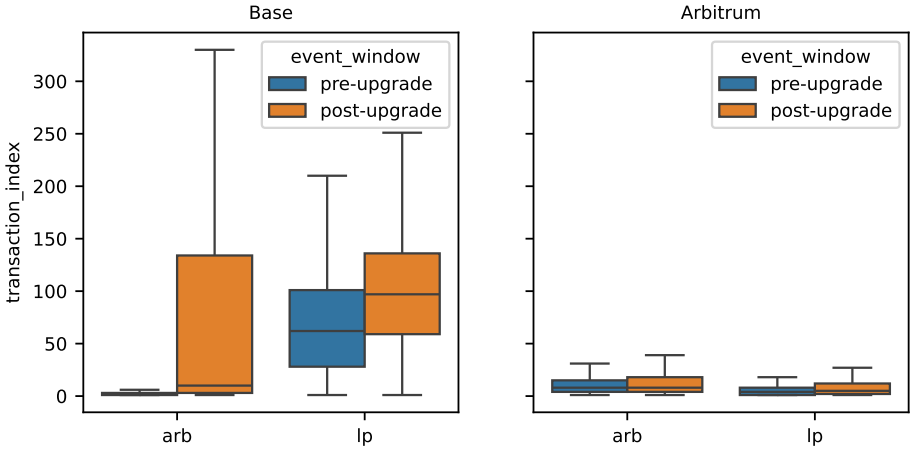


Figure 4. Transaction index (block position) for arbitrageurs and LPs, pre- and post-upgrade. Shifts in block position reflect changes in the PGA competition structure under faster sub-blocks.

6.4 Faster Sequencing Changes Who Supplies Liquidity

Prediction. The cream-skimming result (Proposition 4.6) implies that faster blocks widen the fee interval on which strategic LPs are profitable while passive LPs are not, because the LP exit hazard h increases with λ_B. Chains with faster block times should therefore exhibit higher concentration of LP activity among sophisticated rebalancers, faster rebalancing times as active LPs exploit shorter execution intervals, and potential crowding-out of passive liquidity.

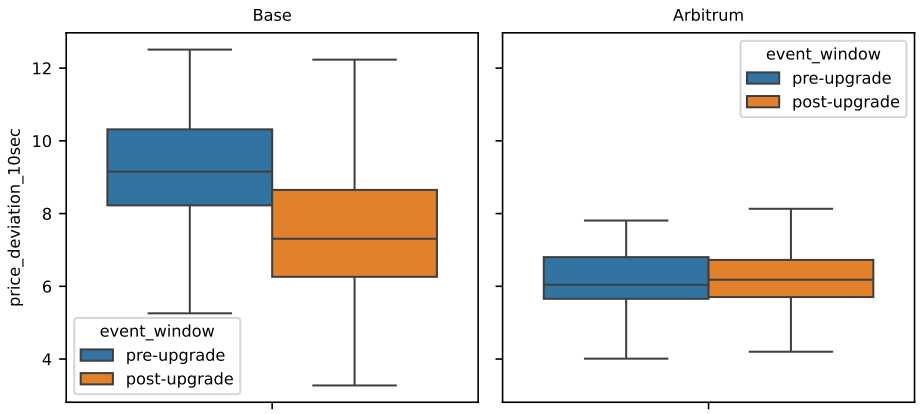


Figure 5. Price deviation at 10-second horizon for identified arbitrageurs on Base and Arbitrum, pre- and post-upgrade. This measures the gross arbitrage opportunity available at the time of the trade.

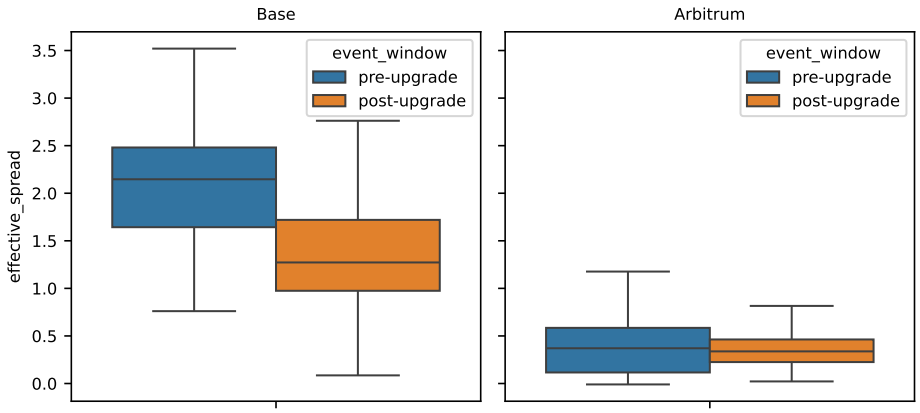


Figure 6. Effective spread for identified arbitrageurs on Base and Arbitrum, pre- and post-upgrade.

Rebalancing behavior. Figure 10 shows the distribution of rebalancing times—the elapsed time between consecutive mint-burn cycles—for identified active LPs on Base and Arbitrum, pre- and post-upgrade. If faster blocks enable the strategic LP to withdraw and re-deposit more quickly (higher withdrawal and re-entry rates), we should observe shorter rebalancing times on Base post-upgrade.

Concentration of LP activity. The model's most distinctive prediction is that faster infrastructure does not simply benefit all LPs—it creates a wedge between technologically sophisticated LPs who can exploit block-boundary timing and passive LPs who cannot. Table 5 presents the Herfindahl-Hirschman Index (HHI) and top-N concentration shares for LP rebalancing activity.

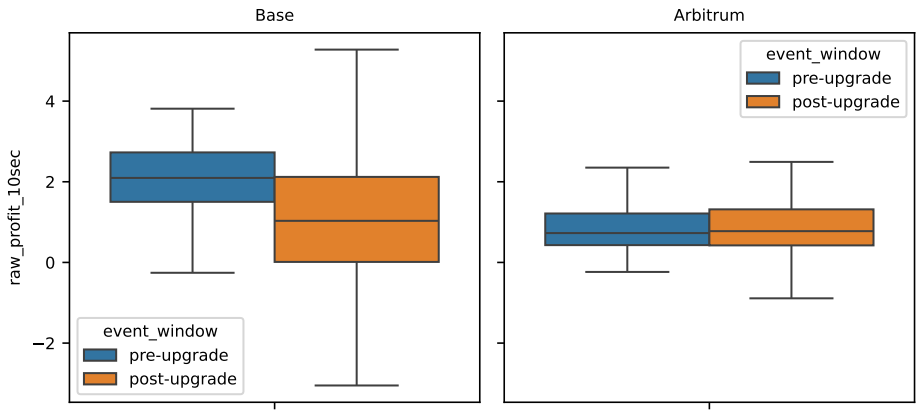


Figure 7. Raw profit (10-second horizon) for identified arbitrageurs on Base and Arbitrum, pre- and post-upgrade.

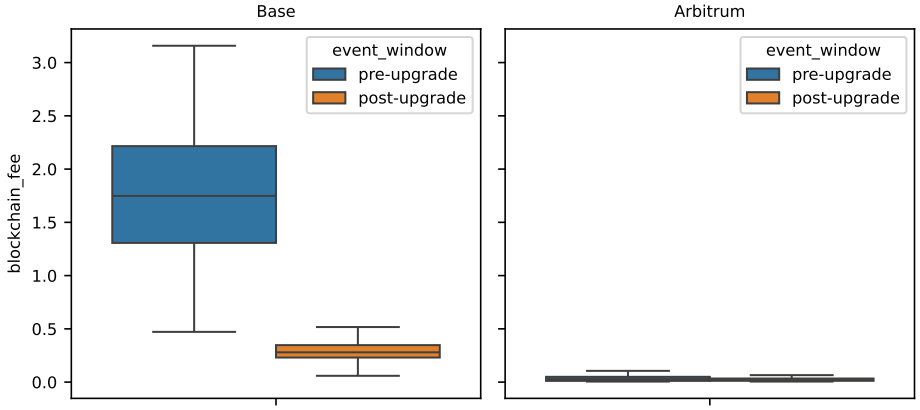


Figure 8. Blockchain fee for identified arbitrageurs on Base and Arbitrum, pre- and post-upgrade.

The cross-chain comparison is striking: Base exhibits an HHI of 0.36, with the top-1 address accounting for 58% of all rebalancing activity and the top-10 for 83%. Arbitrum, by contrast, shows an HHI of 0.01, with top-1 share of 8% and top-10 share of 26%. This pattern is consistent with the cream-skimming mechanism: on Base, where the PGA allows technologically active LPs to secure priority positioning, liquidity provision is dominated by a small number of sophisticated rebalancers who capture fee flow while avoiding adverse selection. On Arbitrum's FCFS ordering, no LP can reliably outpace others, resulting in more dispersed participation.

Interpretation. The LP evidence supports the model's central compositional prediction: faster infrastructure changes *who* viably supplies liquidity. The concentration gap between Base and Arbitrum is large and economically significant, consistent with the cream-skimming interval being

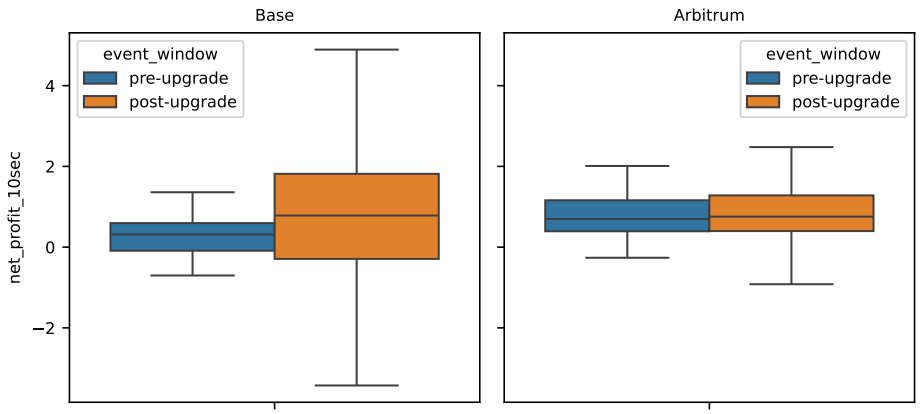


Figure 9. Net profit (10-second horizon) for identified arbitrageurs on Base and Arbitrum, pre- and post-upgrade.

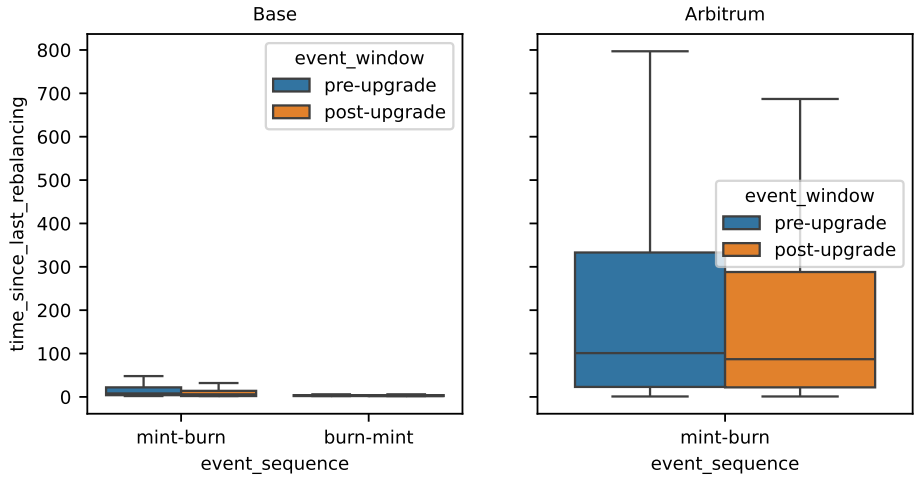


Figure 10. Rebalancing time for active LPs, pre- and post-upgrade. Shorter rebalancing times on Base post-upgrade are consistent with the model's prediction that faster blocks increase the LP exit hazard h, enabling more frequent position adjustment.

non-empty (EP3). The rebalancing-time data suggest that active LPs respond to faster blocks by adjusting positions more rapidly, consistent with the LP exit hazard h increasing in λ_B.

Two confounds require discussion. First, the cross-chain comparison conflates block-time differences with sequencing-rule differences: Base uses PGA ordering, which allows technologically active LPs to purchase early block positions, while Arbitrum's FCFS ordering disperses priority. The concentration gap may therefore partly reflect the sequencing mechanism rather than block speed alone. The within-chain pre/post comparison on Base provides cleaner identification for the effect of λ_B holding the sequencing rule fixed. Second, the Flashblocks upgrade coincided with

Table 5. LP rebalancing concentration. HHI and top-N shares of rebalancing activity (mint-burn events) by chain. Higher concentration on Base is consistent with cream-skimming by technologically active LPs (Proposition 4.6).

	Arbitrum	Optimism	Base
Share (Top 1)	0.08	0.67	0.58
Share (Top 5)	0.19	0.98	0.78
Share (Top 10)	0.26	0.99	0.83
HHI	0.01	0.48	0.36

increased public attention to Base as a DeFi venue, which may have attracted sophisticated LP capital independently of the speed change. The Arbitrum control partially addresses this concern, but the possibility of concurrent demand-side shifts cannot be ruled out with a single-event design.

6.5 Robustness and Heterogeneity

Scope of empirical claims. The empirical analysis uses a single pool, a single chain-level event, and a two-chain comparison. These design features provide clean identification for the directional effects of faster sequencing but impose clear limits on what the evidence can establish. The Flashblocks upgrade moves Base from approximately 0.5 blocks per second ($\lambda_B \approx 0.5$) to approximately 5 blocks per second ($\lambda_B \approx 5$). This is a movement away from the "too slow" corner, consistent with the left side of the interior optimum curve. The evidence does not identify whether Base has reached, overshot, or remains below the welfare-maximizing block rate λ_B^*.

Two predictions of the model are not testable with this design. EP4 (hump-shaped speed adoption) requires observing speed investment across a range of block rates, which a single monotone shock cannot provide. EP5 (regime clustering) requires cross-sectional variation across many chains with different block times and similar pool characteristics, which does not exist in the current L2 landscape with sufficient comparability.

7 Conclusion

Block speed is a first-order determinant of market quality on blockchain trading infrastructure. This paper provides a welfare theory of the execution clock by modeling the joint equilibrium of priority competition, speed investment, and liquidity provision on a block-gated AMM.

The central mechanism is that the block production rate λ_B simultaneously governs three forces. Slow blocks trap liquidity providers: the safe-exit probability vanishes, passive capital bears the full weight of adverse selection, and the sequencer captures nearly all arbitrage rents through priority fees. Fast blocks enable cream-skimming: active LPs exploit block-boundary timing to capture fee income while avoiding adverse selection, displacing passive providers and concentrating liquidity supply. The sequencer's rent share declines monotonically with speed, shifting rents from within-block priority auctions toward private speed investment. Under a tail-dominance condition, these opposing forces produce an interior welfare maximum: blocks can be either too slow or too fast, and the optimal clock depends on the balance between LP protection, speed costs, and infrastructure extraction.

Empirical evidence from the Base Flashblocks rollout—which shortened effective execution intervals from 2 seconds to 200 milliseconds—is consistent with a move away from the too-slow region. CEX–DEX price deviations compress, LP rebalancing concentrates sharply, and the composition of blockchain fees shifts. The evidence identifies a directional effect from a single event; the theory provides the generalizing structure.

The analysis has limitations that point to open questions. The interior optimum is conditional on whether the endogenous liquidity channel eventually exhausts speed incentives at high block rates; primitive conditions separating the three regimes remain to be fully characterized. The model studies a single pool, but blockchain infrastructure serves many markets simultaneously—endogenizing cross-chain competition in speed, where chains race to attract order flow even when the industry-wide result is overinvestment, is a natural extension. Finally, a profit-maximizing sequencer internalizes auction revenue but not the full cost of passive-LP displacement or private

speed spending, creating a wedge between the privately and socially optimal execution clock that deserves formal analysis.

References

Adams, Austin (2024). "Layer 2 Be or Layer Not 2 Be: Scaling on Uniswap v3".

Adams, Austin, Ciamac C. Moallemi, Sara Reynolds, and Dan Robinson (2024). "am-AMM: An Auction-Managed Automated Market Maker".

Aoyagi, Jun (2019). "Strategic Speed Choice by High-Frequency Traders under Speed Bumps". *ISER Discussion Paper No. 1050*.

Bahrani, Maryam, Pranav Garimidi, and Tim Roughgarden (2024). "Centralization in Block Building and Proposer-Builder Separation".

Biais, Bruno, Thierry Foucault, and Sophie Moinas (2011). *Equilibrium high-frequency trading*. Working paper.

Budish, Eric, Peter Cramton, and John Shim (2015a). "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response". *The Quarterly Journal of Economics* 130.4, pp. 1547–1621.

— (2015b). "The high-frequency trading arms race: Frequent batch auctions as a market design response". *The Quarterly Journal of Economics* 130.4, pp. 1547–1621.

— (2015c). "The high-frequency trading arms race: Frequent batch auctions as a market design response". *Quarterly Journal of Economics* 130 (4), pp. 1547–1621.

Budish, Eric, Robin S. Lee, and John J. Shim (2024). "A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?" *Journal of Political Economy* 132.4, pp. 1209–1246.

Canidio, Andrea and Robin Fritsch (2023). "Arbitrageurs' Profits, LVR, and Sandwich Attacks: Batch Trading as an AMM Design Response". *5th Conference on Advances in Financial Technologies (AFT 2023)*. Vol. 282. LIPIcs, 24:1–24:17.

Caparros, Basile, Amit Chaudhary, and Olga Klein (2023). “Blockchain Scaling and Liquidity Concentration on Decentralized Exchanges”. *Working paper*.

Capponi, Agostino and Ruizhe Jia (forthcoming). “Liquidity Provision on Blockchain-based Decentralized Exchanges”. *Review of Financial Studies*.

Capponi, Agostino, Ruizhe Jia, and Brian Zhu (2024). “The Paradox of Just-in-Time Liquidity in Decentralized Exchanges: More Providers Can Sometimes Mean Less Liquidity”.

Daian, Philip et al. (2020). “Flash Boys 2.0: Frontrunning in Decentralized Exchanges, Miner Extractable Value, and Consensus Instability”. *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 910–927.

Du, Songzi and Haoxiang Zhu (2017a). “What Is the Optimal Trading Frequency in Financial Markets?” *The Review of Economic Studies*.

— (2017b). “What is the Optimal Trading Frequency in Financial Markets?” *The Review of Economic Studies* 84.4, pp. 1606–1651.

Fan, Zhou et al. (2023). “Strategic Liquidity Provision in Uniswap v3”. LIPIcs 282, 25:1–25:17.

Fritsch, Robin, Maria Inês Silva, Akaki Mamageishvili, Benjamin Livshits, and Edward W. Felten (2025). *MEV Capture Through Time-Advantaged Arbitrage*. URL: http://arxiv.org/abs/2410.10797. Pre-published.

Gogol, Krzysztof, Johnnatan Messias, Deborah Miori, Claudio Tessone, and Benjamin Livshits (2024). “Cross-Rollup MEV: Non-Atomic Arbitrage Across L2 Blockchains”.

Hasbrouck, Joel, Thomas Rivera, and Fahad Saleh (2022). “The Need for Fees at a DEX: How Increases in Fees Can Increase DEX Trading Volume”. *Working paper*.

Lehar, Alfred and Christine Parlour (2023). “Decentralized exchange: the Uniswap automated market maker”. *Journal of Finance*. (forthcoming).

Malinova, Katya and Andreas Park (2023). “Learning from DeFi: Would Automated Market Makers Improve Equity Trading?” *Working paper*.

Mamageishvili, Akaki, Mahimna Kelkar, Jan Christoph Schlegel, and Edward W. Felten (2023). "Buying Time: Latency Racing vs. Bidding for Transaction Ordering". *5th Conference on Advances in Financial Technologies (AFT 2023)*. Vol. 282. LIPIcs, 23:1–23:22.

Menkveld, Albert J., Bart Zhou Yueshen, and Haoxiang Zhu (2017). "Shades of Darkness: A Pecking Order of Trading Venues". *Journal of Financial Economics* 124.3, pp. 503–534.

Menkveld, Albert J. and Marius Zoican (2017). "Need for speed? Exchange latency and liquidity". *Review of Financial Studies* 30.4, pp. 1188–1228.

Milinois, Jason, Ciamac Moallemi, Tim Roughgarden, and Anthony Lee Zhang (2023). "Automated Market Making and Loss-Versus-Rebalancing". *Working paper*.

Milionis, Jason, Ciamac C. Moallemi, and Tim Roughgarden (2024). "Automated Market Making and Arbitrage Profits in the Presence of Fees". *Financial Cryptography and Data Security (FC 2024)*.

Nakamoto, Satoshi (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*. White paper.

Park, Andreas (2023). "The conceptual flaws of decentralized automated market making". *Management Science*. (forthcoming).

Schwarz-Schilling, Caspar et al. (2023). *Time Is Money: Strategic Timing Games in Proof-of-Stake Protocols*. URL: http://arxiv.org/abs/2305.09032. Pre-published.

Wan, Xin and Austin Adams (2022). *Just-in-time Liquidity on the Uniswap Protocol*. Working paper.

A Proofs

Proof of Proposition 4.1 (Landing Kernel). The block arrival time T_B and arb submission times T_1, T_2 are mutually independent exponential random variables.

Step 1: Derive p_{00}. Since $T_B \sim \text{Exp}(\lambda_B)$, $T_1 \sim \text{Exp}(\lambda_1)$, $T_2 \sim \text{Exp}(\lambda_2)$ are independent, the minimum $\min(T_B, T_1, T_2) \sim \text{Exp}(\lambda_B + \lambda_1 + \lambda_2)$. By the standard Poisson race property,

$$p_{00} = \Pr(T_B < T_1 \text{ and } T_B < T_2) = \Pr(T_B = \min(T_B, T_1, T_2)) = \frac{\lambda_B}{\lambda_B + \lambda_1 + \lambda_2}.$$

Step 2: Derive marginal presence probabilities. Arb i is present iff $T_i < T_B$. By the two-racer exponential formula,

$$\Pr(T_i < T_B) = \frac{\lambda_i}{\lambda_B + \lambda_i}, \qquad i \in \{1, 2\}.$$

Step 3: Derive p_{11} via inclusion–exclusion. Let $A_i = \{T_i < T_B\}$ denote the event "arb i present." Then

$$p_{11} = \Pr(A_1 \cap A_2) = 1 - \Pr(A_1^c) - \Pr(A_2^c) + \Pr(A_1^c \cap A_2^c).$$

We have $\Pr(A_i^c) = \lambda_B/(\lambda_B + \lambda_i)$ and $\Pr(A_1^c \cap A_2^c) = p_{00}$. Substituting:

$$p_{11} = 1 - \frac{\lambda_B}{\lambda_B + \lambda_1} - \frac{\lambda_B}{\lambda_B + \lambda_2} + \frac{\lambda_B}{\lambda_B + \lambda_1 + \lambda_2}.$$

Step 4: Derive p_{10} and p_{01}. By the partition of the marginal of arb 1:

$$p_{10} = \Pr(A_1 \cap A_2^c) = \Pr(A_1) - \Pr(A_1 \cap A_2) = \frac{\lambda_1}{\lambda_B + \lambda_1} - p_{11}.$$

Substituting and simplifying:

$$p_{10} = \frac{\lambda_B}{\lambda_B + \lambda_2} - \frac{\lambda_B}{\lambda_B + \lambda_1 + \lambda_2} = \frac{\lambda_B \lambda_1}{(\lambda_B + \lambda_2)(\lambda_B + \lambda_1 + \lambda_2)}.$$

By identical algebra with indices swapped:

$$p_{01} = \frac{\lambda_B \lambda_2}{(\lambda_B + \lambda_1)(\lambda_B + \lambda_1 + \lambda_2)}.$$

Step 5: Verify the partition identity. One can verify $p_{00} + p_{10} + p_{01} + p_{11} = 1$ algebraically by telescoping after collecting marginals. □

Proof of Proposition 4.2 (P-FP Equilibrium Payoffs). **Part I: Expected payoff given bids.**

Each arb submits a sealed bid $b_i \geq 0$ before the block arrives. Given opponent bid CDF F_j, arb i's expected payoff from bidding b is:

$$U_i(b; F_j) = (p_{i0} + q\, F_j(b))(V - b).$$

Dividing by $q > 0$ and defining $c_i := p_{i0}/q$:

$$\frac{U_i(b; F_j)}{q} = (c_i + F_j(b))(V - b).$$

Part II: Symmetric case ($c_1 = c_2 =: c$).

In a symmetric equilibrium $F_1 = F_2 =: F$. The indifference condition $(p + qF(b))(V - b) = pV$ (where $p := p_{i0}$) gives:

$$F(b) = \frac{c\, b}{V - b}, \qquad \bar{b} = \frac{V}{1 + c}.$$

This is a valid CDF: $F(0) = 0$, $F(\bar{b}) = 1$, $F'(b) = cV/(V - b)^2 > 0$. Each arb's equilibrium payoff

is $\Pi_i = p_{i0}V$, and by the budget identity:

$$\mathbb{E}[R] = P_{\geq 1}V - 2pV = qV.$$

Part III: Asymmetric case ($c_1 \neq c_2$; WLOG $c_1 \leq c_2$).

Arb 2's indifference pins arb 1's CDF:

$$F_1(b) = \frac{c_2 b}{V - b}, \qquad \bar{b}_2 = \frac{V}{1 + c_2}.$$

Arb 1's indifference yields collision surplus:

$$U_1^* = V \cdot \frac{c_2 - c_1}{1 + c_2} \geq 0.$$

Arb 2's strategy has an atom of size $\alpha = (c_2 - c_1)/(1 + c_2)$ at $b = 0$ and is continuous on $(0, \bar{b}_2]$:

$$F_2(b) = \frac{c_2(1 + c_1)V}{(1 + c_2)(V - b)} - c_1, \qquad 0 < b \leq \bar{b}_2.$$

Assembling payoffs:

$$\Pi_2 = p_{01}V, \qquad \Pi_1 = p_{10}V + qV \cdot \frac{c_2 - c_1}{1 + c_2}.$$

By the budget identity:

$$\mathbb{E}[R] = qV - qU_1^* = qV \cdot \frac{1 + c_1}{1 + c_2}.$$

Combining both cases gives the compact revenue form:

$$\mathbb{E}[R] = qV \cdot \frac{1 + \min(c_1, c_2)}{1 + \max(c_1, c_2)}.$$

□

Lemma A.1 (LP Optimally Bids Zero)**.** *The LP's optimal priority bid is zero. Her expected payoff*

is $U_L(b_L) = V_L - p_L\, b_L$*, where* V_L *is independent of* b_L*. Bidding zero is strictly dominant when* $p_L > 0$ *and weakly dominant when* $p_L = 0$*.*

Proof of Lemma A.1 (LP Optimally Bids $b_L = 0$***).*** The LP chooses $b_L \in [0, \infty)$. The priority-fee payment is $F_L(b_L) = I_L \cdot b_L$, where $I_L \in \{0, 1\}$ is the inclusion indicator. Therefore $\mathbb{E}[F_L(b_L)] = p_L(b_L) \cdot b_L$.

By SA6, $p_L(b_L) \equiv p_L$ for all $b_L \geq 0$: LP inclusion probability is independent of the bid. Furthermore, SA6 implies the LP continuation value $V_L(b_L)$ is independent of b_L: payoff-relevant ordering is independent of b_L, LP bids do not affect arb ordering or LP state transitions, and inclusion probability is invariant. Under the α loser rule, losing does not move the LP OUT. Hence $V_L(b_L) =: V_L$ for all b_L.

The LP's total expected payoff is therefore:

$$U_L(b_L) = V_L - p_L \cdot b_L.$$

For any $b'_L \geq 0$: $U_L(0) - U_L(b'_L) = p_L \cdot b'_L \geq 0$.

If $p_L > 0$ and $b'_L > 0$, this difference is strictly positive, so $b_L = 0$ is strictly dominant. If $p_L = 0$, the difference is zero for all b'_L, so $b_L = 0$ is weakly dominant. □

Proof of Proposition 4.4 (Stationary Distribution). We solve $\pi G = 0$ with $\sum_s \pi_s = 1$, where G is the G2b generator matrix with state order $(I_0, I_M, I_H, O_0, O_M, O_H)$.

From the column balance equations:

$$\text{Col } I_M: \quad \pi_{I_0} = \frac{J + m}{J}\, \pi_{I_M},$$
$$\text{Col } I_H: \quad \pi_{I_H} = \frac{J}{2(A + h)}\, \pi_{I_M}.$$

Global balance between I and O classes gives $m\pi_{I_M} + h\pi_{I_H} = \mu\pi_{O_0}$, yielding:

$$\pi_{O_0} = \frac{2m(A+h) + Jh}{2\mu(A+h)}\,\pi_{I_M}.$$

From the remaining balance equations:

$$\pi_{O_M} = \pi_{O_0} + \frac{m}{J}\,\pi_{I_M},$$
$$\pi_{O_H} = \frac{1}{A}\left(\frac{J}{2}\pi_{O_M} + h\pi_{I_H}\right).$$

Every π_s is a known multiple of π_{I_M}. Bringing all terms over the common denominator $D_0 = 4A\mu J(A+h)$ and applying normalization, the sum simplifies to

$$S = \frac{(4A+J)\,\Delta}{4A\mu J(A+h)},$$

where $\Delta := 2AJm+2AJ\mu+2Am\mu+J^2h+2Jhm+2Jh\mu+2hm\mu$. Hence $\pi_{I_M} = 4A\mu J(A+h)/[(4A+J)\Delta]$, and substituting back gives the six closed-form stationary probabilities stated in the proposition. The expressions can be verified by substitution into the balance equations. □

Proof of Proposition 4.5 (Slow Blocks Trap LPs). Since $P_{\text{none}} = p_{00} = \lambda_B/(\lambda_B+\lambda_1+\lambda_2) = \Theta(\lambda_B)$, the LP safe-exit hazard at H satisfies

$$\eta_H = \lambda_B p_L P_{\text{none}} = \Theta(\lambda_B^2),$$

while the arb execution hazard is $A = \lambda_B P_{\geq 1} = \Theta(\lambda_B)$. Thus η_H vanishes faster than A, so

$$P_{\text{safe}} = \frac{\eta_H}{\eta_H + A} = \frac{p_L\lambda_B}{\lambda_1 + \lambda_2 + p_L\lambda_B} = \Theta(\lambda_B) \to 0.$$

The stationary mass on OUT states is $\pi_O = O(\lambda_B) \to 0$ and $\pi_I = 1 - O(\lambda_B) \to 1$. □

Proof of Proposition 4.6 (Cream-Skimming). The strategic LP payoff is $\bar{U}_L = \frac{2A\mu}{(4A+J)\Delta}\,\mathcal{K}(m,h)$,

with $\mathcal{K}(m,h) = \sigma\bar{r}(J^2 + 2(A+h)(2J+m)) - LAJ^2$. Because the prefactor is strictly positive, $\bar{U}_L \geq 0 \iff \mathcal{K}(m,h) \geq 0$, which gives the lower bound:

$$\sigma\bar{r} \geq \frac{LAJ^2}{J^2 + 2(A+h)(2J+m)}.$$

Setting $m = h = 0$ (no-exit limit), $\mathcal{K}(0,0) = \sigma\bar{r}(J^2 + 4AJ) - LAJ^2$. This is negative iff:

$$\sigma\bar{r} < \frac{LAJ}{4A+J}.$$

The interval is non-empty iff the lower bound is strictly below the upper bound:

$$\frac{LAJ^2}{J^2 + 2(A+h)(2J+m)} < \frac{LAJ}{4A+J}.$$

Cross-multiplying (all terms positive) and simplifying:

$$[J^2 + 2(A+h)(2J+m)] - J(4A+J) = 2Am + 4Jh + 2hm > 0$$

whenever $h > 0$. □

Proof of Proposition 4.7 (Passive LP Participation). Define $\sigma_P^{\text{IN}} := \bar{Q}/(\bar{Q} + q_L)$ (passive share when LP IN) and $\sigma_P^{\text{OUT}} := 1$ (when LP OUT). Define $V_{\text{IN}} := c(\bar{Q} + q_L)(H^2 - f^2)$ and $V_{\text{OUT}} := c\bar{Q}(H^2 - f^2)$.

The fee income rate is:

$$\text{FeeIncomeRate} = \bar{r}[\pi_I \sigma_P^{\text{IN}} + \pi_O] = \bar{r}\left[\pi_O + \pi_I \frac{\bar{Q}}{\bar{Q} + q_L}\right].$$

For the loss rate, the key identity is $\sigma_P^{\text{IN}} V_{\text{IN}} = \frac{\bar{Q}}{\bar{Q}+q_L} \cdot c(\bar{Q} + q_L)(H^2 - f^2) = c\bar{Q}(H^2 - f^2) = V_{\text{OUT}}$. This cancellation relies on linearity $\kappa(Q) = cQ$ and implies the passive LP's per-execution loss is

the same whether the strategic LP is IN or OUT. Therefore:

$$\text{LossRate} = A[\pi_{IH} V_{\text{OUT}} + \pi_{OH} V_{\text{OUT}}] = A\pi_H \cdot c\bar{Q}(H^2 - f^2).$$

The passive LP participates ($Q^P = \bar{Q} > 0$) iff $\overline{U}_P = \text{FeeIncomeRate} - \text{LossRate} \geq 0$:

$$\bar{r}\left[\pi_O + \pi_I \frac{\bar{Q}}{\bar{Q} + q_L}\right] \geq A\pi_H \cdot c\bar{Q}(H^2 - f^2). \qquad \square$$

Proof of Proposition 4.8 (Joint Speed Equilibrium). A complete branch is a speed-PSNE if and only if each player's net payoff on that branch weakly exceeds its net payoff on every one-step speed deviation branch. All gross payoff objects are recomputed on the relevant deviation branch, including the profile-specific values of $(p_{ij}, q, P_{\geq 1}, A, p_L, h, m, X, d^*, \pi, A_{\text{ss}}, V_{\text{eff}})$.

1. The all-slow branch (S, S, S). An arb compares staying slow on (S, S, S) with deviating to the high-c position on (F, S, S): the no-deviation condition is $K \geq A_{\text{ss}}^{FSS} \widetilde{\Pi}_H(V_{\text{eff}}^{FSS}) - A_{\text{ss}}^{SSS} \widetilde{\Pi}^{SS}(V_{\text{eff}}^{SSS}) =: \Delta_A^{SSS}$. The LP compares (S, S, S) with (S, S, F): no-deviation requires $K \geq \overline{U}_L^{SSF} - \overline{U}_L^{SSS} =: \Delta_L^{SS}$.

2. The slow-arb / fast-LP branch (S, S, F). Arb no-deviation: $K \geq \Delta_A^{SSF}$. LP no-deviation (staying fast vs. slowing): $K \leq \Delta_L^{SS}$.

3. The all-fast-arb / slow-LP branch (F, F, S). Arb no-deviation (staying fast vs. slowing to low-c position): $K \leq \Gamma_A^{FFS}$. LP no-deviation: $K \geq \Delta_L^{FF}$.

4. The fully fast branch (F, F, F). Arb no-deviation: $K \leq \Gamma_A^{FFF}$. LP no-deviation: $K \leq \Delta_L^{FF}$.

5. The asymmetric slow-LP branches (F, S, S) *and* (S, F, S). The fast arb's no-deviation (vs. slowing to (S, S, S)): $K \leq \Delta_A^{SSS}$. The slow arb's no-deviation (vs. speeding to (F, F, S)): $K \geq \Gamma_A^{FFS}$ (using arb relabeling symmetry). LP no-deviation: $K \geq \Delta_L^{FS}$.

6. The asymmetric fast-LP branches (F, S, F) *and* (S, F, F). The fast arb's no-deviation (vs. slowing to (S, S, F)): $K \leq \Delta_A^{SSF}$. The slow arb's no-deviation (vs. speeding to (F, F, F)): $K \geq \Gamma_A^{FFF}$. LP no-deviation: $K \leq \Delta_L^{FS}$.

All threshold objects are closed-form functions of primitives through the full S3+G2b pipeline.

□

Corollary A.1 (Non-Monotone Speed Adoption)**.** *The equilibrium number of fast agents can be non-monotone in block speed.*

Proof of Corollary A.1 (Speed Adoption Hump-Shaped). By Proposition 4.8, (S, S, S) is a PSNE whenever $K \geq \Delta_A^{SSS}$ and $K \geq \Delta_L^{SS}$. Since both gain differentials vanish as $\lambda_B \downarrow 0$ (because every gross arb payoff has the form $A_{\text{ss}}\widetilde{\Pi}$ with $A_{\text{ss}} = O(\lambda_B)$), the conditions hold for all sufficiently small λ_B, so $0 \in \mathcal{N}_F(\lambda_B)$ in the left tail.

As λ_B rises, the direct arrival-value channel strengthens because $A(\lambda_B) = \lambda_B P_{\geq 1}(\lambda_B)$ moves away from zero, making the first-arrival advantage from choosing F more valuable. This can push arb thresholds above K, creating branches with one or more fast agents.

At higher λ_B, the endogenous-liquidity channel works through $p_L, h, m, \pi, V_{\text{eff}}$. If faster LP protection moves stationary mass away from I_H and toward O_H, then V_{eff} falls. Since every arb threshold is built from terms of the form $A_{\text{ss}}\widetilde{\Pi}(V_{\text{eff}})$, this can push thresholds back down. Therefore the same fixed K can be crossed from below and later from above, producing non-monotonicity.

□

Corollary A.2 (Regime Clustering)**.** *Observationally similar chains can settle into qualitatively different market structures.*

Proof of Corollary A.2 (Regime Clustering). The three arb-speed configurations have execution hazards

$$A^{SS} = \lambda_B \frac{2\lambda^S}{\lambda_B + 2\lambda^S}, \quad A^{FS} = \lambda_B \frac{\lambda^F + \lambda^S}{\lambda_B + \lambda^F + \lambda^S}, \quad A^{FF} = \lambda_B \frac{2\lambda^F}{\lambda_B + 2\lambda^F}.$$

Cross-multiplying and simplifying: $A^{SS} < A^{FS}$ iff $2\lambda^S(\lambda_B + \lambda^F + \lambda^S) < (\lambda^F + \lambda^S)(\lambda_B + 2\lambda^S)$, which reduces to $\lambda_B(\lambda^F - \lambda^S) > 0$. Similarly $A^{FS} < A^{FF}$. Hence $A^{SS} < A^{FS} < A^{FF}$.

For the LP control probability with fixed LP speed $l \in \{S, F\}$:

$$p_L^{l,FF} = \frac{\lambda_L^l}{\lambda_L^l + 2\lambda^F} < p_L^{l,FS} = \frac{\lambda_L^l}{\lambda_L^l + \lambda^F + \lambda^S} < p_L^{l,SS} = \frac{\lambda_L^l}{\lambda_L^l + 2\lambda^S}.$$

Thus moving from SS to FS to FF raises the arb execution hazard A and lowers the LP control probability p_L. The two-fast-arb branches exert the strongest pressure toward early exit and low effective liquidity (high-speed/low-liquidity cluster), while the all-slow branches maintain higher depth (low-speed/high-liquidity cluster). □

Proof of Proposition 4.9 (Ergodic Welfare Accounting). On branch ω, welfare includes three terms: gains from trade, speed costs, and sequencer leakage. When the strategic LP is IN, total pool depth is $Q^P + q_L$; when OUT, it is Q^P. The time-weighted average pool depth is therefore

$$\mathbb{E}[Q_\omega(\lambda_B)] = (Q^P + q_L)\pi_I^\omega(\lambda_B) + Q^P\pi_O^\omega(\lambda_B) = Q^P + q_L\pi_I^\omega(\lambda_B),$$

where the second equality uses $\pi_I + \pi_O = 1$. The closed form for π_I^ω follows from the stationary distribution (Proposition 4.4) evaluated at branch-specific hazards $A^\omega = \lambda_B P_{\geq 1}^\omega$, $h^\omega = \lambda_B p_L^\omega p_{00}^\omega$, $m^\omega = \mathbf{1}\{\tau = M\}\lambda_B p_L^\omega$.

The speed-cost term is $Kn_{\text{fast}}(\omega)$. Sequencer leakage per unit time is $A_{\text{ss}}^\omega(\lambda_B)\widetilde{R}(V_{\text{eff}}^\omega(\lambda_B))$. Combining gives the welfare formula. □

Lemma A.2 (Transfer Cancellation)**.** *All within-population flows—retail fee income, arbitrage trading prizes, and exchange fees—are pure transfers and cancel in aggregate welfare. Only gains from trade, speed costs, and sequencer leakage remain.*

Proof of Lemma A.2 (Transfer Cancellation). The continuous retail fee flow $\bar{r}$ is paid by retail traders and received by LPs; since both are inside the welfare population, this is an internal redistribution. The arb's net trading gain per execution is $\kappa(Q)H^2 - \kappa(Q)f^2 = V_{\text{arb}}(Q)$; the LP side's net payoff is $-V_{\text{arb}}(Q)$. These cancel exactly: $V_{\text{arb}}(Q) + (-V_{\text{arb}}(Q)) = 0$. What does not cancel is

sequencer revenue (the sequencer is excluded from the welfare population) and speed costs (deadweight resource expenditures). Hence aggregate welfare retains $\gamma\mathbb{E}[Q]$, Kn_{fast}, and $A_{\text{ss}}\widetilde{R}(V_{\text{eff}})$. □

Lemma A.3 (Leakage Bounds)**.** *For every active branch ω, the sequencer leakage satisfies $L_\omega(\lambda_B) = O(\lambda_B) \to 0$ as $\lambda_B \downarrow 0$ and $L_\omega(\lambda_B) \to 0$ as $\lambda_B \to \infty$. On the outside-option branch, $W^0(\lambda_B) = \gamma Q^P + o(1)$ as $\lambda_B \downarrow 0$.*

Proof of Proposition 4.10 (Interior Optimum). We establish three regions.

Low block rates. By assumption (i), the outside option is selected on $(0, \lambda_B^{\text{entry}})$. By Lemma A.3 (whose proof uses the bounds $L_\omega(\lambda_B) = O(\lambda_B) \to 0$ as $\lambda_B \downarrow 0$ and $L_\omega(\lambda_B) \to 0$ as $\lambda_B \to \infty$, established from the S3 collision probability $q = \Theta(\lambda_B^{-2})$ at high block rates), $W(\lambda_B) = W^0(\lambda_B) = \gamma Q^P + o(1)$ as $\lambda_B \downarrow 0$.

Intermediate block rates. Assumption (ii) provides an active branch $\hat{\omega}$ and $\hat{\lambda}_B > \lambda_B^{\text{entry}}$ with $W(\hat{\lambda}_B) \geq W_{\hat{\omega}}(\hat{\lambda}_B) > \gamma Q^P$. The sufficient condition follows from Proposition 4.9: $W_{\hat{\omega}}(\hat{\lambda}_B) - \gamma Q^P = \gamma q_L \pi_I^{\hat{\omega}}(\hat{\lambda}_B) - Kn_{\text{fast}}(\hat{\omega}) - A_{\text{ss}}^{\hat{\omega}}(\hat{\lambda}_B)\widetilde{R}(V_{\text{eff}}^{\hat{\omega}}(\hat{\lambda}_B))$.

High block rates. For any persistent active branch ω, by Proposition 4.9 and using $\pi_I^\omega(\lambda_B) \to \pi_I^{\omega,\infty}$ and $L_\omega \to 0$:

$$\lim_{\lambda_B \to \infty} W_\omega(\lambda_B) = \gamma Q^P + \gamma q_L \pi_I^{\omega,\infty} - Kn_{\text{fast}}(\omega).$$

By assumption (iii), $\gamma q_L \pi_I^{\omega,\infty} < Kn_{\text{fast}}(\omega)$, so this limit is strictly below γQ^P. The outside-option welfare satisfies $W^0(\lambda_B) \leq \gamma Q^P$ for all λ_B since $S^0, L^0 \geq 0$. Hence $\limsup_{\lambda_B \to \infty} W(\lambda_B) \leq \gamma Q^P$.

Since W is upper semicontinuous (it is the upper envelope of finitely many continuous regime functions), it attains a maximum on any compact interval. Let $\eta = (W_{\hat{\omega}}(\hat{\lambda}_B) - \gamma Q^P)/2 > 0$. By the tail bounds, there exist $\varepsilon > 0$ and $\Lambda' > \hat{\lambda}_B$ such that $\sup_{0 < \lambda_B \leq \varepsilon} W(\lambda_B) \leq \gamma Q^P + \eta$ and $\sup_{\lambda_B \geq \Lambda'} W(\lambda_B) \leq \gamma Q^P + \eta$. On the compact interval $[\varepsilon, \Lambda']$, W attains a maximum. Since $\hat{\lambda}_B \in (\varepsilon, \Lambda')$ and $W(\hat{\lambda}_B) \geq \gamma Q^P + 2\eta$, this maximum strictly exceeds every value in both tails. Therefore every global maximizer lies in $(0, \infty)$, and an interior maximizer exists. □

B LP Transition Dynamics

Conditional on a speed profile $\mathbf{s} = (s_1, s_2, s_L)$ and a withdrawal threshold d^*, the LP's pool position and the price gap jointly evolve as a six-state continuous-time process. The LP's capital is either deposited (I) or withdrawn (O), and the gap is 0, M, or H. From the arrival probabilities in Proposition 4.1, three reduced-form transition rates summarize the within-block race:

$$\begin{aligned} A(\mathbf{s}) &:= \lambda_B P_{\geq 1}(\mathbf{s}) && \text{(arbitrage execution rate at } d = H\text{)}, \\ h(\mathbf{s}) &:= \lambda_B \, p_L(\mathbf{s}) \, p_{00}(\mathbf{s}) && \text{(safe withdrawal rate at } d = H\text{)}, \\ m(\mathbf{s}, d^*) &:= \mathbf{1}\{d^* = M\} \, \lambda_B \, p_L(\mathbf{s}) && \text{(early withdrawal rate at } d = M\text{)}. \end{aligned}$$

If the LP attempts to withdraw at $d = H$ but an arbitrageur arrives first, she remains deposited and bears the loss. If she has withdrawn, she can re-deposit only from $d = 0$, at rate $\mu_0 > 0$. The transition system is:

$$\begin{aligned} &I0 \xrightarrow{\lambda_J} IM, && IM \xrightarrow{\lambda_J/2} I0, && IM \xrightarrow{\lambda_J/2} IH, && IM \xrightarrow{m} OM, \\ &IH \xrightarrow{A} I0, && IH \xrightarrow{h} OH, && O0 \xrightarrow{\lambda_J} OM, && O0 \xrightarrow{\mu_0} I0, \\ &OM \xrightarrow{\lambda_J/2} O0, && OM \xrightarrow{\lambda_J/2} OH, && OH \xrightarrow{A} O0. \end{aligned}$$

Let π_s denote the long-run fraction of time in state s, and define the aggregate participation and gap probabilities

$$\pi_I := \pi_{I0} + \pi_{IM} + \pi_{IH}, \qquad \pi_O := \pi_{O0} + \pi_{OM} + \pi_{OH}, \qquad \pi_H := \pi_{IH} + \pi_{OH}.$$

All six stationary probabilities share the denominator $D = (4A + \lambda_J)\Delta$, where Δ is a positive polynomial in the transition rates. Closed-form expressions for each π_s are stated in the proof of Proposition 4.4.

C Additional Empirical Results

C.1 Blockchain Fee Computation

On L2 chains, each transaction incurs two components of fees. The **L2 execution fee** compensates the sequencer:

$$\text{L2 fee} = \frac{\text{gas used} \times \text{effective gas price}}{10^{18}} \quad \text{(in ETH)}. \tag{20}$$

The **L1 data fee** compensates for posting calldata to Ethereum mainnet:

$$\text{L1 fee} = \frac{\texttt{receipt_l1_fee}}{10^{18}} \quad \text{(in ETH)}. \tag{21}$$

The total blockchain fee in basis points of trade size is:

$$\text{Blockchain fee} = \frac{\text{L2 fee} + \text{L1 fee}}{|\Delta\text{WETH}|} \times 10^4. \tag{22}$$

C.2 Flashblocks Implementation Timeline

Base. Base implemented Flashblocks following the timeline below:

- 7th July 2025: Enabled Flashblocks on Base Mainnet
- 15th May 2025: Ended testing Flashblocks on Base Mainnet
- 15th May 2025: Started testing Flashblocks on Base Mainnet
- 25th Feb 2025: Enabled Flashblocks on Base Sepolia

Optimism. Optimism implemented Flashblocks on September 30th, 2025.