# On the Incentive Compatibility of Optimistic Blockchain Mechanisms

Jiasun Li

George Mason University

# "Optimistic" designs

**Optimistic** designs in many new blockchain infrastructures/applications

- System proceeds assuming all participants always well behave
  - As compared to "**correct by construction**" designs...
  - ...where technological constraints prevent misbehavior in the first place
- "Optimistic" assumption is meant to be enforced by incentives
  - (most commonly/almost exclusively) "**stake-and-slash**" mechanisms
  - stake: first put down a certain amount of stakes as collateral
  - slash: whoever caught misbehaving be confiscated some or all stakes

Optimistic systems...
so designed that participants can be assumed to behave well, as otherwise, they will face penalties that go against their rational incentives

# "Optimistic" designs

**Optimistic** designs in many new blockchain infrastructures/applications

- System proceeds assuming all participants always well behave
  - As compared to "**correct by construction**" designs...
  - ...where technological constraints prevent misbehavior in the first place
- "Optimistic" assumption is meant to be enforced by incentives
  - (most commonly/almost exclusively) "**stake-and-slash**" mechanisms
  - stake: first put down a certain amount of stakes as collateral
  - slash: whoever caught misbehaving be confiscated some or all stakes

Optimistic systems...
so designed that participants can be assumed to behave well, as otherwise, they will face penalties that go against their rational incentives

# Wide adoption of optimistic designs

The earliest mention dates back to circa 2017-2018

- the Plasma project for scaling Ethereum

Adopted (explicitly or implicitly) in many new blockchain applications

- Layer-2 scaling solutions, e.g.,
  - Optimistic rollups
    - leading protocols Arbitrum and Optimism
    - Coinbase's recently launched Base
  - Sidechains (e.g., Polygon), etc.
- Proof-of-stake (PoS) layer-1 blockchain, e.g.,
  - Ethereum (post-merge), Solana, etc.
- Optimistic cross-chain bridges, e.g.,
  - Nomad, Connext, etc.

# Wide adoption of optimistic designs

The earliest mention dates back to circa 2017-2018

- the Plasma project for scaling Ethereum

Adopted (explicitly or implicitly) in many new blockchain applications

- Layer-2 scaling solutions, e.g.,
  - Optimistic rollups
    - ★ leading protocols Arbitrum and Optimism
    - ★ Coinbase's recently launched Base
  - Sidechains (e.g., Polygon), etc.
- Proof-of-stake (PoS) layer-1 blockchain, e.g.,
  - Ethereum (post-merge), Solana, etc.
- Optimistic cross-chain bridges, e.g.,
  - Nomad, Connext, etc.

# Wide adoption of optimistic designs

The earliest mention dates back to circa 2017-2018

- the Plasma project for scaling Ethereum

Adopted (explicitly or implicitly) in many new blockchain applications

- Layer-2 scaling solutions, e.g.,
  - Optimistic rollups
    - ★ leading protocols Arbitrum and Optimism
    - ★ Coinbase's recently launched Base
  - Sidechains (e.g., Polygon), etc.
- Proof-of-stake (PoS) layer-1 blockchain, e.g.,
  - Ethereum (post-merge), Solana, etc.
- Optimistic cross-chain bridges, e.g.,
  - Nomad, Connext, etc.

# An open question

Despite wide adoption, it remains a theoretical question

- *Are optimistic designs secure among rational, self-interested parties*?
  - does the current design incentivize all participants to behave?
- many casual arguments in the community
- but lacking rigorous inputs from economists

$\Rightarrow$ i.e., practice is leading theory

  - typical in blockchain and other fast-evolving tech fields

This paper

- an (embarrassingly) simple game-theoretical model
  - to capture the mechanism of existing optimistic systems
  - question whether these mechanisms are fully incentive-compatible

# An open question

Despite wide adoption, it remains a theoretical question

- *Are optimistic designs secure among rational, self-interested parties*?
  - does the current design incentivize all participants to behave?
- many casual arguments in the community
- but lacking rigorous inputs from economists

$\Rightarrow$ i.e., practice is leading theory

  - typical in blockchain and other fast-evolving tech fields

This paper

- an (embarrassingly) simple game-theoretical model
  - to capture the mechanism of existing optimistic systems
  - question whether these mechanisms are fully incentive-compatible

# An open question

Despite wide adoption, it remains a theoretical question

- *Are optimistic designs secure among rational, self-interested parties*?
  - does the current design incentivize all participants to behave?
- many casual arguments in the community
- but lacking rigorous inputs from economists

$\Rightarrow$ i.e., practice is leading theory

  - typical in blockchain and other fast-evolving tech fields

This paper

- an (embarrassingly) simple game-theoretical model
  - to capture the mechanism of existing optimistic systems
  - question whether these mechanisms are fully incentive-compatible

# Related literature

Blockchain scaling and layer-2 solutions

- Whitehat, Gluchowski, HarryR, Fu and Castonguay (2018), Kalodner, Goldfeder, Chen, Weinberg and Felten (2018), Bousfield, Bousfield, Buckland, Burgess, Colvin, Felten, Goldfeder, Goldman, Huddleston, Kalodner, Lacs, Ng, Sanghi, Wilson, Yermakova and Zidenberg (2018), John, Rivera and Saleh (2020), Cong, Hui, Tucker and Zhou (2023), Jermann (2023), etc.

Proof-of-stake layer-1 blockchains ("stake-and-slash" mechanism)

- Halaburda, He and Li (2021), Kogan, Fanti and Viswanath (2021), John, Rivera and Saleh (2021), Saleh (2021), Benhaim, Hemenway Falk and Tsoukalas (2021), Amoussou-Guenou, Biais, Potop-Butucaru and Tucci-Piergiovanni (2023), and He, Li and Wu (2023), etc.

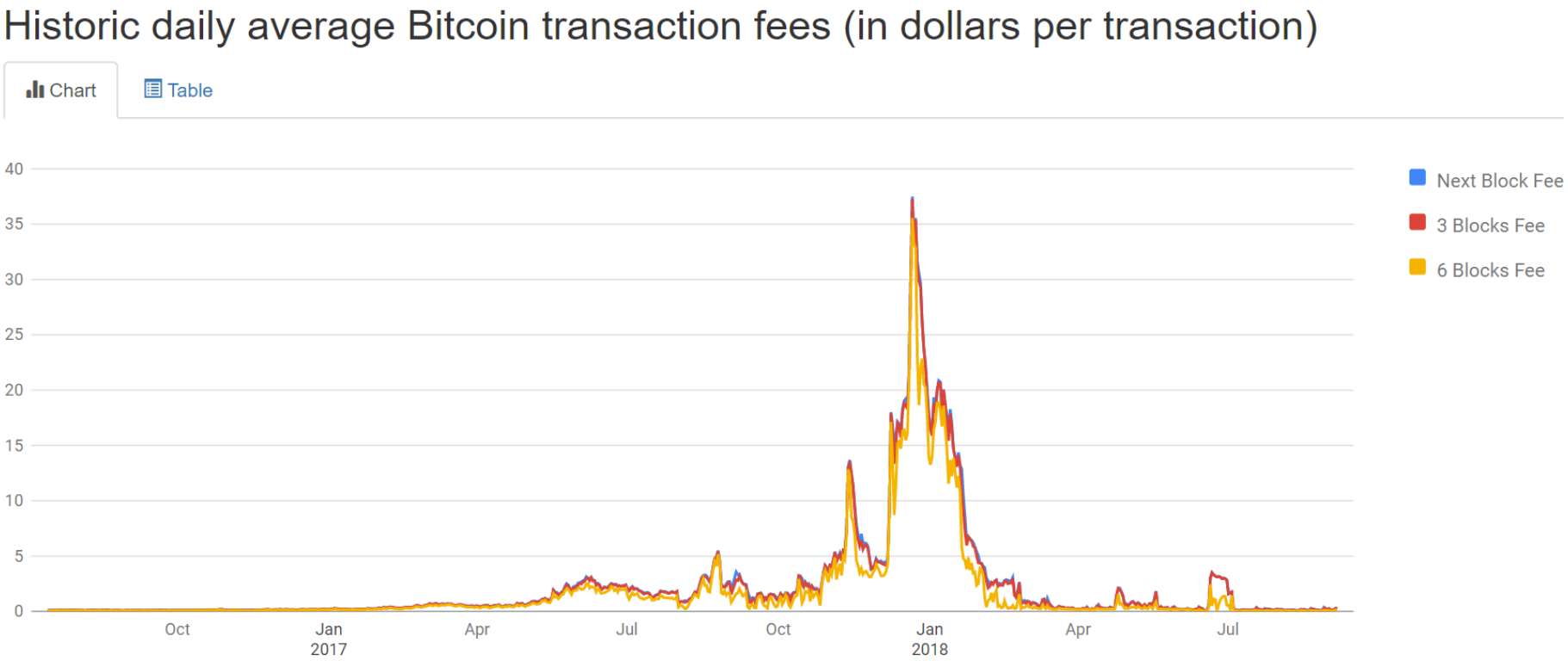
Cross-chain communication bridges

- McCorry, Buckland, Yee and Song (2021) and Li and Wu (2023), etc.

# Roadmap

1. Institutional Details
2. (Simplest) Model
3. Model extensions and robustness of the main argument
   - Challenger's internalization of harms from attacks
   - An extension to multiple attackers and challengers
   - Breaking budget balance?
4. Connections to and implications for practices
5. Conclusion

# The blockchain scaling problem (review)




Source:

https://masterthecrypto.com/blockchain-scalability-bitcoin-scalability-problem-effects/.

# The blockchain scaling problem (review)

BBC Sign in Home News Sport Reel Worklife

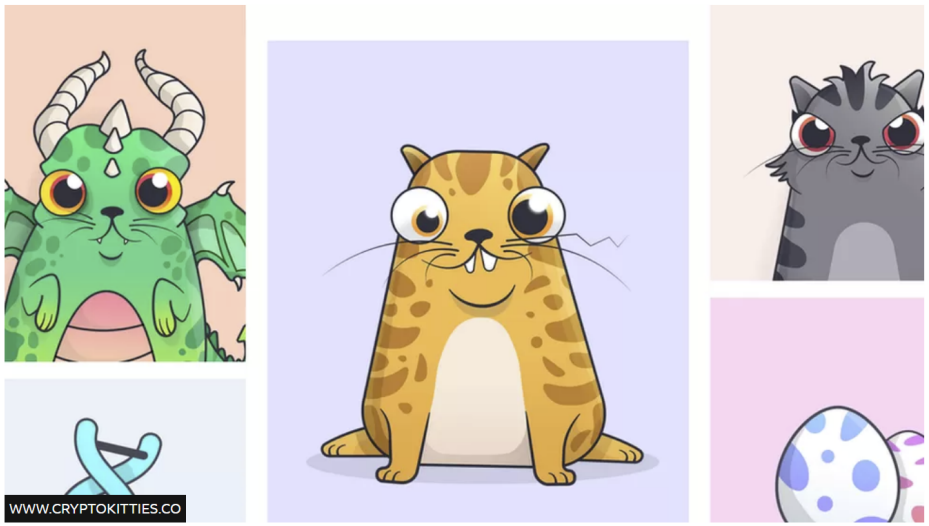
NEWS

Home | War in Ukraine | Climate | Video | World | US & Canada | UK | Business | Tech | Science

Tech

## CryptoKitties craze slows down transactions on Ethereum

5 December 2017




**A new craze for virtual kittens is slowing down trade in one of the largest crypto-currencies.**

Source: https://www.bbc.com/news/technology-42237162.

# Layer-2 scaling solutions

To increase a blockchain's throughput (TPS) by

- processing transactions on (potentially multiple) layer-2 sidechains...
- ...alongside the underlying layer-1 blockchain

How to keep the layer-1 and layer-2 chains in sync?

- layer-2 connects to the underlying layer-1 by "checkpointing"...
- ...periodically summarize layer-2 activities onto the underlying layer-1

How to ensure that layer-2 activities will be correctly sent to layer-1?

- "correct by construction" (make it impossible to misbehave)
  - built upon cryptography and complexity theory
- "**optimistic**" (make it unprofitable to misbehave)
  - built upon economic incentives
  - "**stake-and-slash**" mechanisms

# Layer-2 scaling solutions

To increase a blockchain's throughput (TPS) by

- processing transactions on (potentially multiple) layer-2 sidechains...
- ...alongside the underlying layer-1 blockchain

How to keep the layer-1 and layer-2 chains in sync?

- layer-2 connects to the underlying layer-1 by "checkpointing"...
- ...periodically summarize layer-2 activities onto the underlying layer-1

How to ensure that layer-2 activities will be correctly sent to layer-1?

- "correct by construction" (make it impossible to misbehave)
  - built upon cryptography and complexity theory
- "**optimistic**" (make it unprofitable to misbehave)
  - built upon economic incentives
  - "**stake-and-slash**" mechanisms

# Layer-2 scaling solutions

To increase a blockchain's throughput (TPS) by

- processing transactions on (potentially multiple) layer-2 sidechains...
- ...alongside the underlying layer-1 blockchain

How to keep the layer-1 and layer-2 chains in sync?

- layer-2 connects to the underlying layer-1 by "checkpointing"...
- ...periodically summarize layer-2 activities onto the underlying layer-1

How to ensure that layer-2 activities will be correctly sent to layer-1?

- "correct by construction" (make it impossible to misbehave)
  - built upon cryptography and complexity theory
- "**optimistic**" (make it unprofitable to misbehave)
  - built upon economic incentives
  - "**stake-and-slash**" mechanisms

# Layer-2 scaling solutions (early attempts)

## Plasma: Scalable Autonomous Smart Contracts

Joseph Poon
joseph@lightning.network

Vitalik Buterin
vitalik@ethereum.org

August 11, 2017
WORKING DRAFT
https://plasma.io/

**Abstract**

Plasma is a proposed framework for incentivized and enforced execution of smart contracts which is scalable to a significant amount of state updates per second (potentially billions) enabling the blockchain to be able to represent a significant amount of decentralized financial applications worldwide. These smart contracts are incentivized to continue operation autonomously via network transaction fees, which is ultimately reliant upon the underlying blockchain (e.g. Ethereum) to enforce transactional state transitions.

We propose a method for decentralized autonomous applications to scale to process not only financial activity, but also construct economic incentives for globally persistent data services, which may produce an alternative to centralized server farms.

Source: https://plasma.io/plasma.pdf.

# Layer-2 scaling solutions (Rollup era)


**Arbitrum: Scalable, private smart contracts**

**Harry Kalodner, Steven Goldfeder, Xiaoqi Chen, S. Matthew Weinberg, and Edward W. Felten, *Princeton University***

https://www.usenix.org/conference/usenixsecurity18/presentation/kalodner

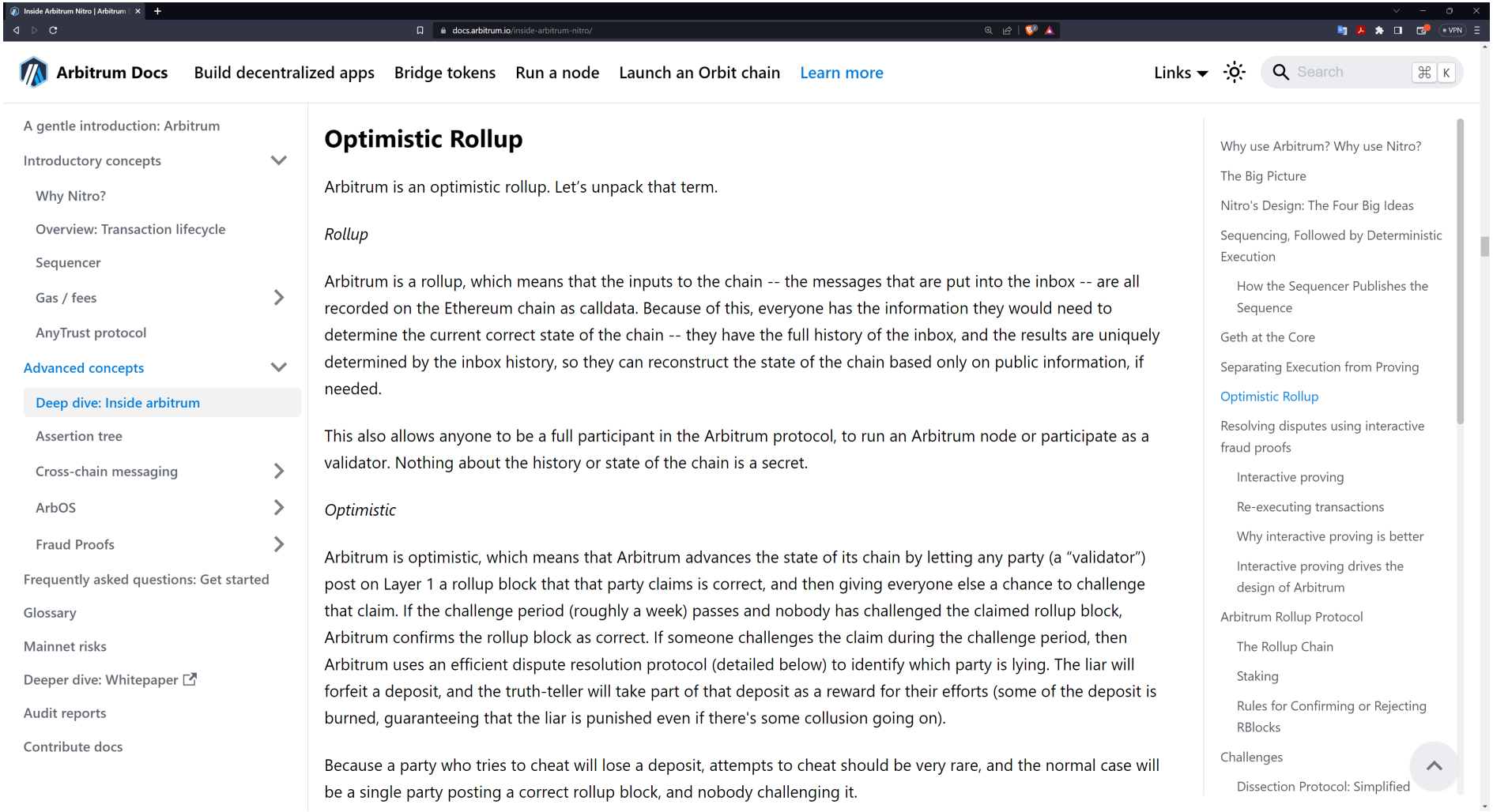
**This paper is included in the Proceedings of the 27th USENIX Security Symposium.**

**August 15–17, 2018 • Baltimore, MD, USA**

978-1-939133-04-5

Source: Kalodner, Goldfeder, Chen, Weinberg and Felten (2018).

# Arbitrum's vision as a (permissionless) optimistic rollup



Inside Arbitrum Nitro | Arbitrum

docs.arbitrum.io/inside-arbitrum-nitro/

Arbitrum Docs | Build decentralized apps | Bridge tokens | Run a node | Launch an Orbit chain | Learn more | Links | Search ⌘ K

A gentle introduction: Arbitrum
Introductory concepts
- Why Nitro?
- Overview: Transaction lifecycle
- Sequencer
- Gas / fees
- AnyTrust protocol

Advanced concepts
- Deep dive: Inside arbitrum
- Assertion tree
- Cross-chain messaging
- ArbOS
- Fraud Proofs

Frequently asked questions: Get started
Glossary
Mainnet risks
Deeper dive: Whitepaper
Audit reports
Contribute docs

## Optimistic Rollup

Arbitrum is an optimistic rollup. Let's unpack that term.

*Rollup*

Arbitrum is a rollup, which means that the inputs to the chain -- the messages that are put into the inbox -- are all recorded on the Ethereum chain as calldata. Because of this, everyone has the information they would need to determine the current correct state of the chain -- they have the full history of the inbox, and the results are uniquely determined by the inbox history, so they can reconstruct the state of the chain based only on public information, if needed.

This also allows anyone to be a full participant in the Arbitrum protocol, to run an Arbitrum node or participate as a validator. Nothing about the history or state of the chain is a secret.

*Optimistic*

Arbitrum is optimistic, which means that Arbitrum advances the state of its chain by letting any party (a "validator") post on Layer 1 a rollup block that that party claims is correct, and then giving everyone else a chance to challenge that claim. If the challenge period (roughly a week) passes and nobody has challenged the claimed rollup block, Arbitrum confirms the rollup block as correct. If someone challenges the claim during the challenge period, then Arbitrum uses an efficient dispute resolution protocol (detailed below) to identify which party is lying. The liar will forfeit a deposit, and the truth-teller will take part of that deposit as a reward for their efforts (some of the deposit is burned, guaranteeing that the liar is punished even if there's some collusion going on).

Because a party who tries to cheat will lose a deposit, attempts to cheat should be very rare, and the normal case will be a single party posting a correct rollup block, and nobody challenging it.

Why use Arbitrum? Why use Nitro?
The Big Picture
Nitro's Design: The Four Big Ideas
Sequencing, Followed by Deterministic Execution
- How the Sequencer Publishes the Sequence

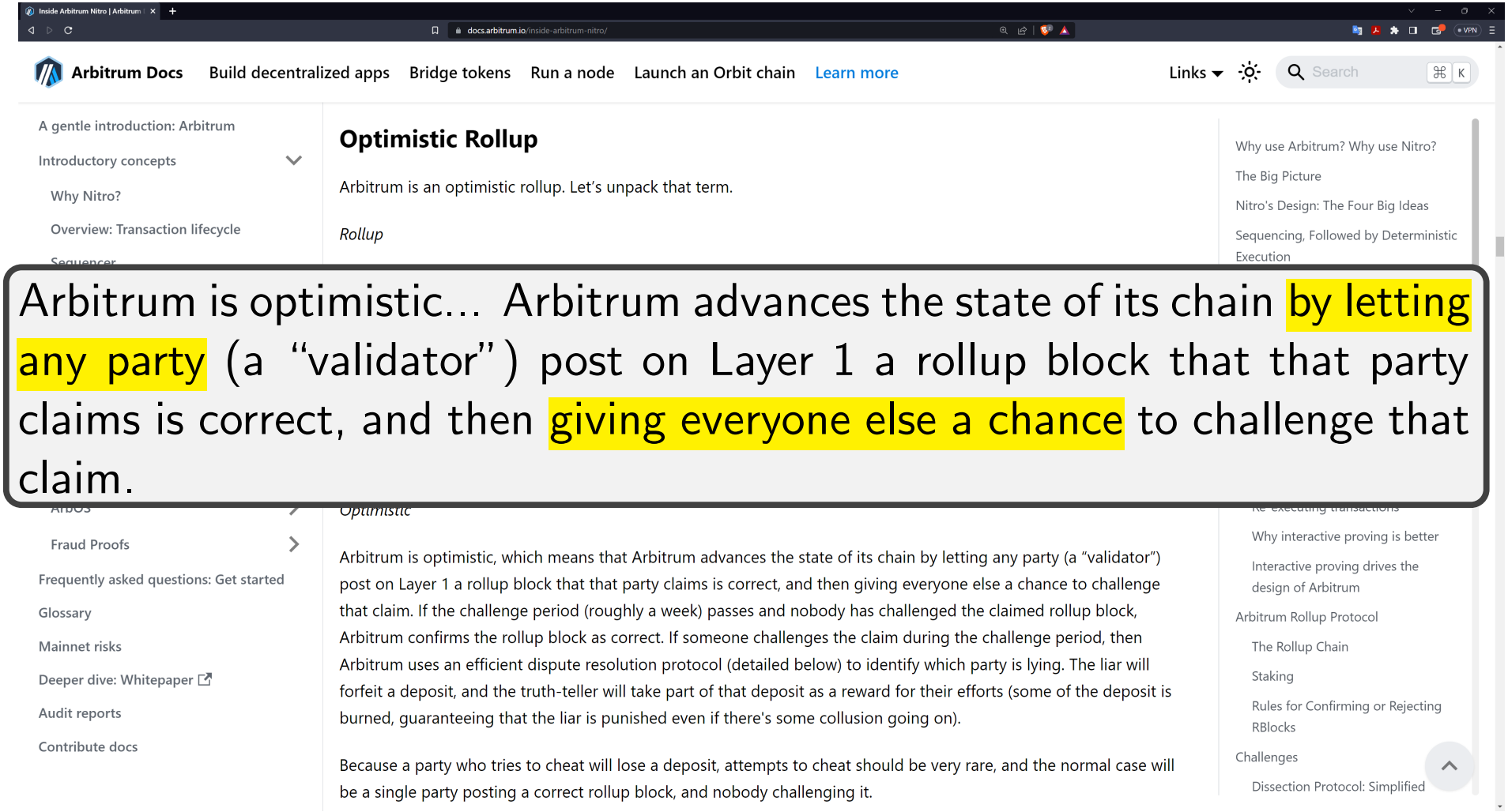
Geth at the Core
Separating Execution from Proving
Optimistic Rollup
Resolving disputes using interactive fraud proofs
- Interactive proving
- Re-executing transactions
- Why interactive proving is better
- Interactive proving drives the design of Arbitrum

Arbitrum Rollup Protocol
- The Rollup Chain
- Staking
- Rules for Confirming or Rejecting RBlocks

Challenges
- Dissection Protocol: Simplified

Source: Arbitrum documentation https://docs.arbitrum.io/inside-arbitrum-nitro/ (retrieved on August 13, 2023)

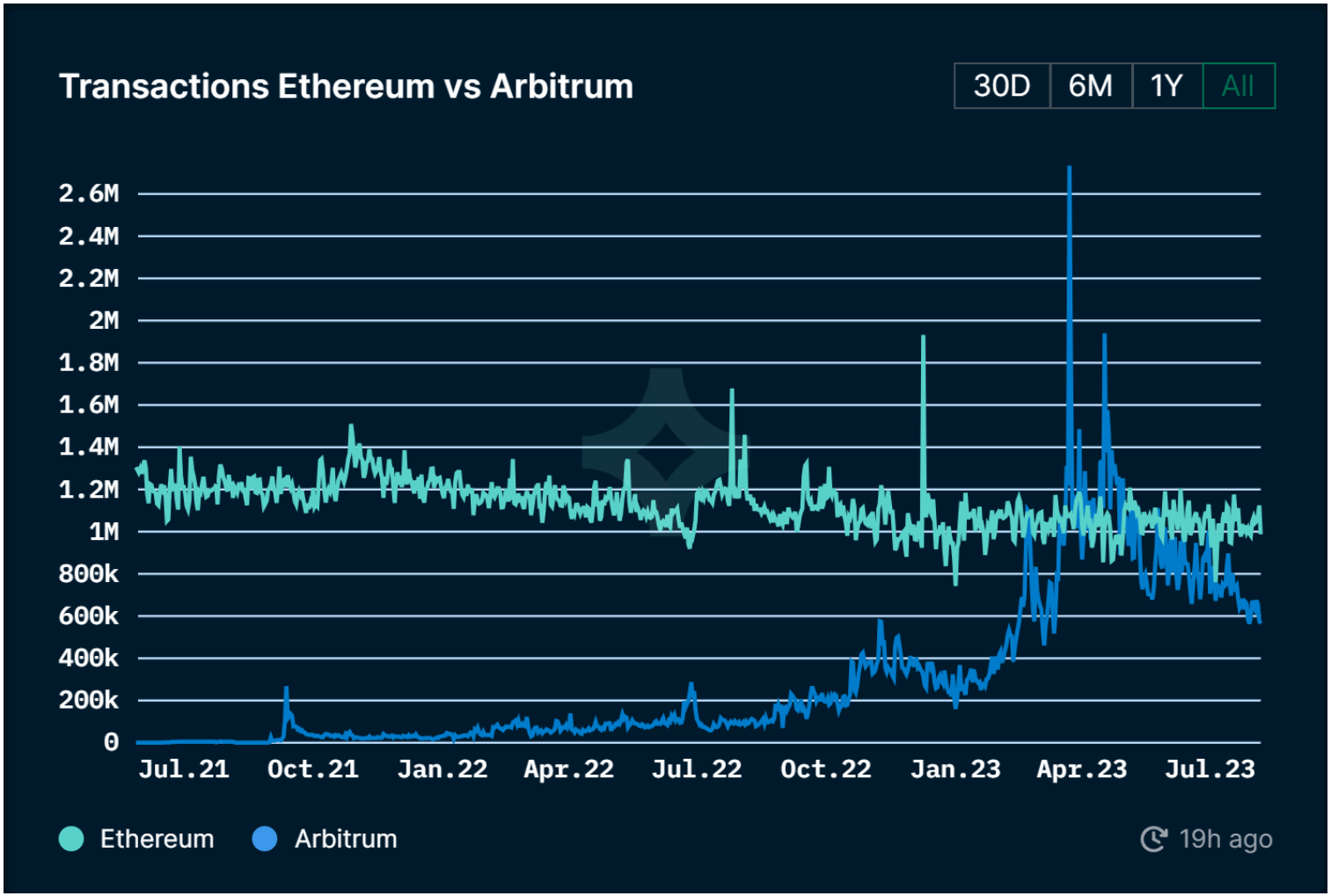
# Arbitrum's vision as a (permissionless) optimistic rollup




Arbitrum is optimistic... Arbitrum advances the state of its chain by letting any party (a "validator") post on Layer 1 a rollup block that that party claims is correct, and then giving everyone else a chance to challenge that claim.

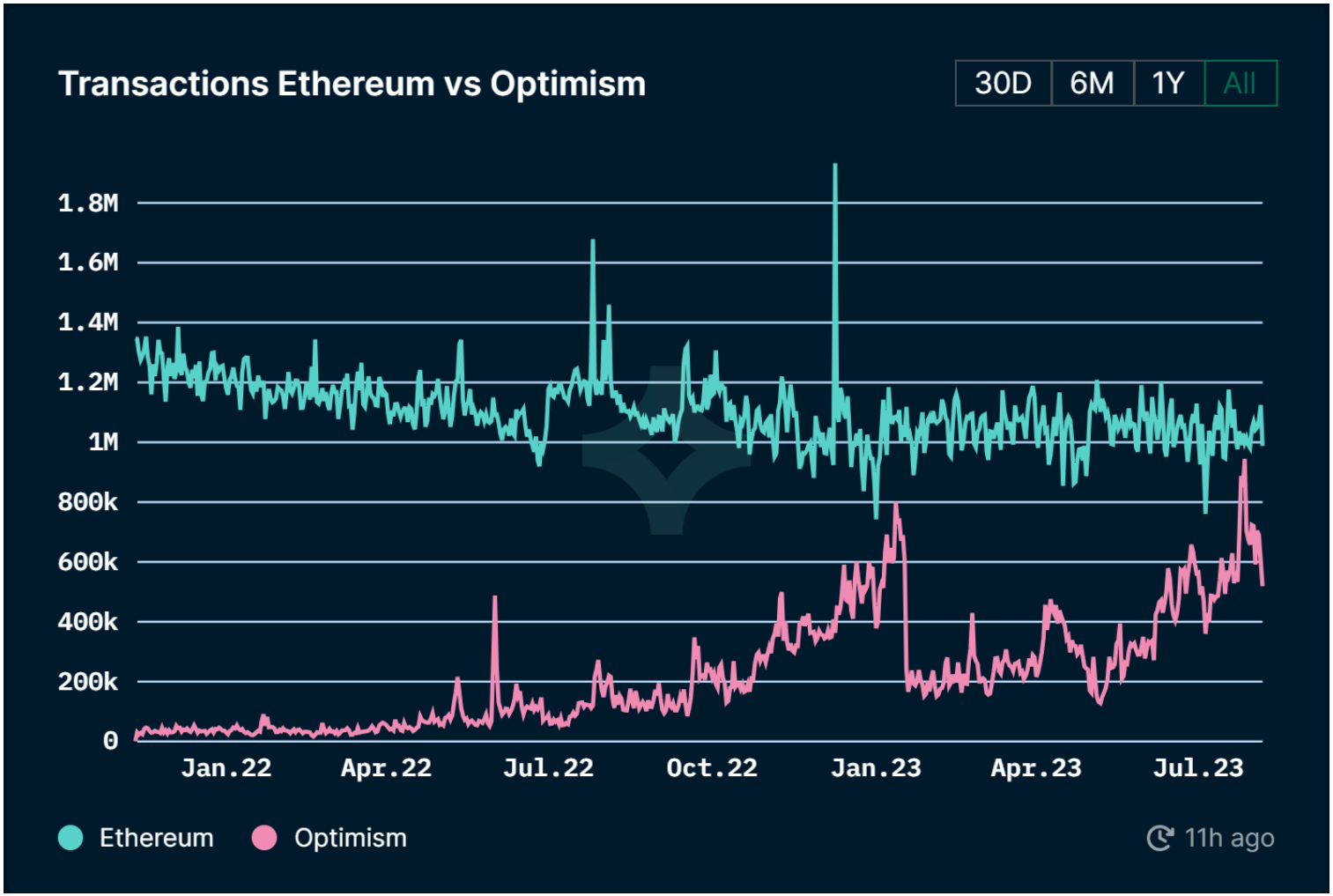
Source: Arbitrum documentation https://docs.arbitrum.io/inside-arbitrum-nitro/ (retrieved on August 13, 2023)

# Layer 1 vs Layer 2 transaction counts comparison over time



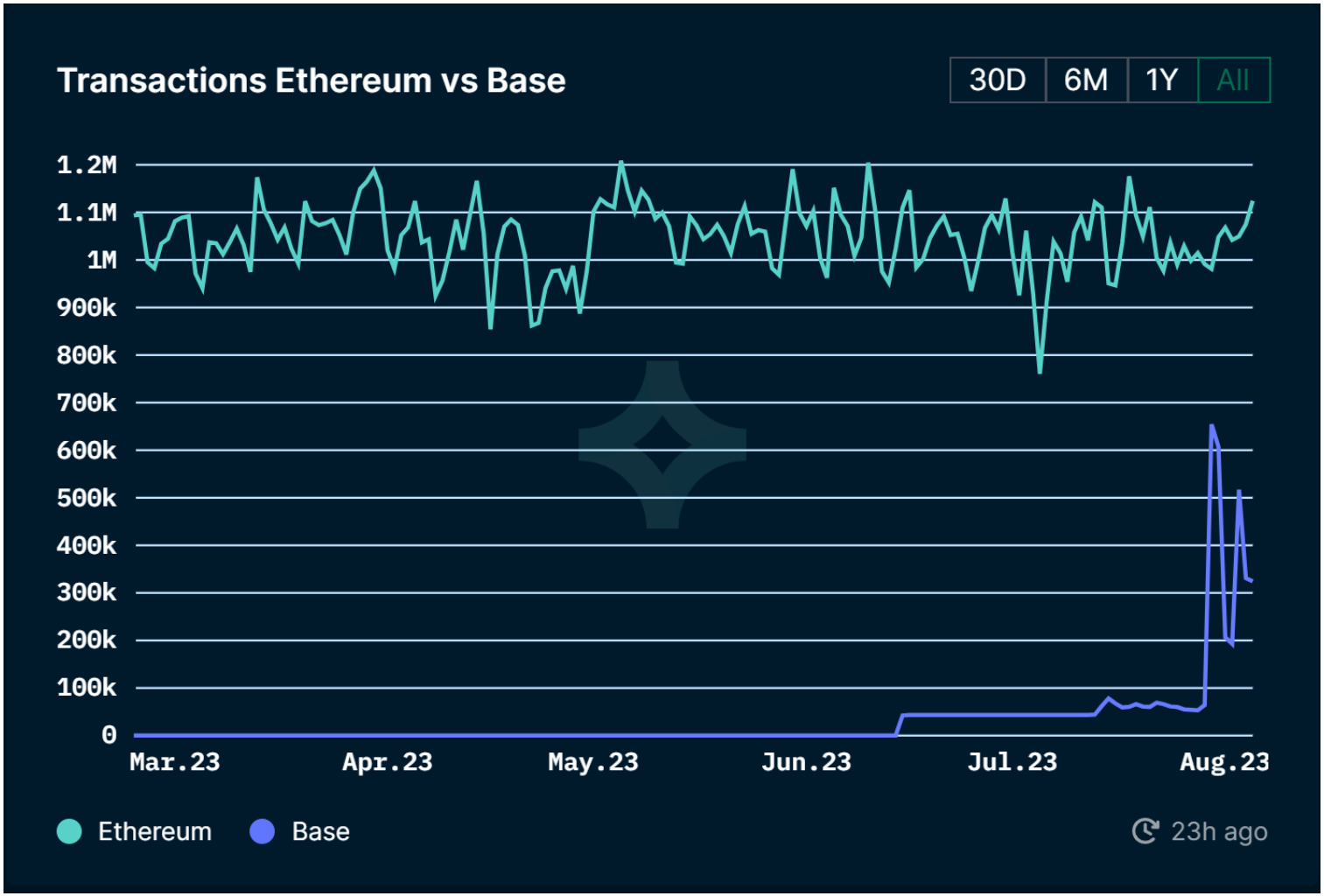

Source: https://pro.nansen.ai/multichain/eth-vs-arbitrum

# Layer 1 vs Layer 2 transaction counts comparison over time



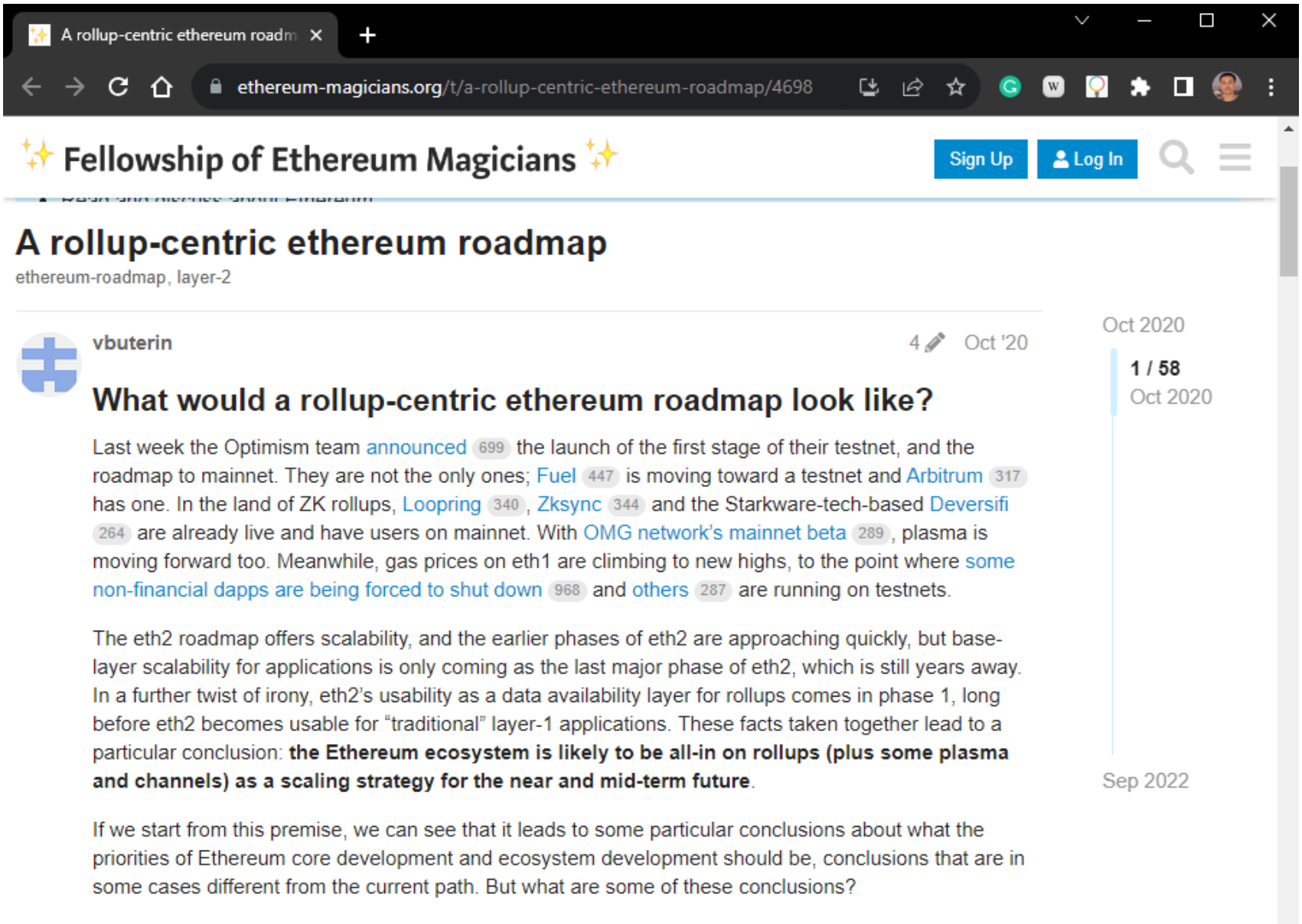

Source: https://pro.nansen.ai/multichain/eth-vs-optimism

# Layer 1 vs Layer 2 transaction counts comparison over time




Source: https://pro.nansen.ai/multichain/eth-vs-base

# A Rollup Centric Ethereum Roadmap




Source: https://ethereum-magicians.org/t/a-rollup-centric-ethereum-roadmap/4698

# Proof-of-stake layer 1 blockchains

"Stake-and-slash" in (many) proof-of-stake (PoS) blockchains

- e.g., Cosmos, Polkadot, Solana, and **Ethereum** post-merge:
- anyone can become an Ethereum validator by staking 32 ethers.
- staked validators repeatedly and pseudo-randomly shuffled over time to **propose** or **attest** new blocks.
- behaving validators rewarded for compliant proposals/attestations.
- validators who are caught misbehaving will be "slashed"
  - deprived of the privilege to validate (and collect rewards) anymore
  - stakes deducted depending on the severity of their violations.
- To trigger slashing:
  - evidence of violation gathered by some validator(s) aka whistleblower(s)
  - then be included by a future block proposer in a new block
- slashable violations may be left unslashed if
  - no whistleblower detection ($\because$ costly detection/inadequate incentives)
  - proposed block containing whistleblow message fails consensus

# Proof-of-stake layer 1 blockchains

"Stake-and-slash" in (many) proof-of-stake (PoS) blockchains

- e.g., Cosmos, Polkadot, Solana, and **Ethereum** post-merge:
- anyone can become an Ethereum validator by staking 32 ethers.
- staked validators repeatedly and pseudo-randomly shuffled over time to **propose** or **attest** new blocks.
- behaving validators rewarded for compliant proposals/attestations.
- validators who are caught misbehaving will be "slashed"
  - deprived of the privilege to validate (and collect rewards) anymore
  - stakes deducted depending on the severity of their violations.
- To trigger slashing:
  - evidence of violation gathered by some validator(s) aka whistleblower(s)
  - then be included by a future block proposer in a new block
- slashable violations may be left unslashed if
  - no whistleblower detection ($\because$ costly detection/inadequate incentives)
  - proposed block containing whistleblow message fails consensus

# Proof-of-stake layer 1 blockchains

"Stake-and-slash" in (many) proof-of-stake (PoS) blockchains

- e.g., Cosmos, Polkadot, Solana, and **Ethereum** post-merge:
- anyone can become an Ethereum validator by staking 32 ethers.
- staked validators repeatedly and pseudo-randomly shuffled over time to **propose** or **attest** new blocks.
- behaving validators rewarded for compliant proposals/attestations.
- validators who are caught misbehaving will be "slashed"
  - deprived of the privilege to validate (and collect rewards) anymore
  - stakes deducted depending on the severity of their violations.
- To trigger slashing:
  - evidence of violation gathered by some validator(s) aka whistleblower(s)
  - then be included by a future block proposer in a new block
- slashable violations may be left unslashed if
  - no whistleblower detection ($\because$ costly detection/inadequate incentives)
  - proposed block containing whistleblow message fails consensus

# Cross-chain bridges

Multi-blockchains interoperability $\Rightarrow$ cross-chain communications

- e.g., asset-transfer (or arbitrary message passing) "bridges"

often built on a "**lock-and-mint/burn-and-unlock**" design.

1. user sends some tokens (e.g. ethers) to a specific smart contract on a source chain (e.g. Ethereum) to "lock" funds within the contract
2. bridge passes a message with proof of locked funds to another smart contract deployed on a destination chain (e.g. Solana) to redeem ("mint") a corresponding number of tokens (e.g. sols).
3. user can then freely use the minted sols on the destination chain, while his ethers are locked on the source chain.

To reverse the process

1. user sends sols to a smart contract to have them burned
2. bridge passes the evidence of burn to a smart contract on the source chain to have the locked funds (ethers in our example) "unlocked."

# Cross-chain bridges

Multi-blockchains interoperability $\Rightarrow$ cross-chain communications

- e.g., asset-transfer (or arbitrary message passing) "bridges"

often built on a "**lock-and-mint/burn-and-unlock**" design.

1. user sends some tokens (e.g. ethers) to a specific smart contract on a source chain (e.g. Ethereum) to "lock" funds within the contract
2. bridge passes a message with proof of locked funds to another smart contract deployed on a destination chain (e.g. Solana) to redeem ("mint") a corresponding number of tokens (e.g. sols).
3. user can then freely use the minted sols on the destination chain, while his ethers are locked on the source chain.

To reverse the process

1. user sends sols to a smart contract to have them burned
2. bridge passes the evidence of burn to a smart contract on the source chain to have the locked funds (ethers in our example) "unlocked."

# Optimistic cross-chain bridges

How to ensure the messages of "lock" or "burn" events on one chain can be accurately transmitted to another chain to trigger "mint" events?

- safety: no mint event is triggered without a corresponding lock or burn event initiated by the correct user with the appropriate amount
- liveness: a correct lock or burn event should expect the corresponding mint event to eventually occur – i.e., no censorship

Various implementation methods, including **optimistic** ones:

- a permissionless set of "relayers" (**staked and slashable**)
- sign/send a message to trigger mint on the destination chain
  - upon hearing a lock/burn transaction on the source chain
- incorrectly relayed messages can be challenged within a window
  - e.g., 30 minutes on Nomad

# Roadmap

1. Institutional Details
2. (Simplest) Model
3. Model extensions and robustness of the main argument
   - Challenger's internalization of harms from attacks
   - An extension to multiple attackers and challengers
   - Breaking budget balance?
4. Connections to and implications for practices
5. Conclusion

# (Simplest) Model

A two-(representative) player, two-action simultaneous-move game.

- **Attacker**:
  - Attack the protocol / Do not attack
  - Interpretations of attacking:
    - ⋆ Erroneously recording layer-2 activities onto layer-1 in scaling solutions.
    - ⋆ Violating consensus protocol rules in PoS blockchains.
    - ⋆ Transmitting wrong messages in optimistic cross-chain bridges.
- **Challenger**:
  - Monitor potential attacks / Do not monitor
  - Interpretations of monitoring:
    - ⋆ Verifying layer-2 transactions on layer-1 blockchain in scaling solutions.
    - ⋆ Running "slasher" algorithm to catch violations in PoS blockchains.
    - ⋆ Monitoring chains to detect wrong messages in optimistic bridges.

# (Simplest) Model

A two-(representative) player, two-action simultaneous-move game.

- **Attacker**:
  - Attack the protocol / Do not attack
  - Interpretations of attacking:
    - ★ Erroneously recording layer-2 activities onto layer-1 in scaling solutions.
    - ★ Violating consensus protocol rules in PoS blockchains.
    - ★ Transmitting wrong messages in optimistic cross-chain bridges.
- **Challenger**:
  - Monitor potential attacks / Do not monitor
  - Interpretations of monitoring:
    - ★ Verifying layer-2 transactions on layer-1 blockchain in scaling solutions.
    - ★ Running "slasher" algorithm to catch violations in PoS blockchains.
    - ★ Monitoring chains to detect wrong messages in optimistic bridges.

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Payoffs

| | | Challenger: Challenge | Challenger: Not Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, 0)$ |
| | Not Attack | $(0, -c)$ | $(0, 0)$ |

Private benefit $\pi > 0$; Penalty $T$:

- set $T > \pi$ to deter attacks

Private cost $c > 0$; Reward $t$

- set $t \leq T$ to balance the budget ($t - T$ typically "burned")
- set $t > c$ to incentivize monitoring

Summarizing the payoff matrix

- Attacker: profitable to attack if and only if not getting caught
- Challenger: profitable to monitor if and only if catching an attack

# Solving the game

## Proposition (No pure strategy Nash equilibrium)

*The strategic interaction between the attacker and challenger features no pure strategy Nash equilibrium. Specifically, the "optimistic" outcome of no attacks cannot be sustained in equilibrium.*

Straightforward proof by contradiction:

- If there were no attacks, then why costly monitor at all?
- But If no one monitors, why not attack in the first place?

# Solving the game: Mixed-strategy Nash equilibrium

## Lemma (Unique mixed-strategy Nash equilibrium)

*In a unique mixed strategy Nash equilibrium, the attacker attacks with prob.* $\frac{c}{t} \in (0,1)$*, and the challenger monitors with prob.* $\frac{\pi}{T} \in (0,1)$*.*

Attacking probability:

- increases in the challenger's monitoring cost $c$
- decreases in the challenger's reward of finding an attack $t$.

$\Rightarrow$ to make the challenger indifferent between monitoring or not.

Monitoring probability

- increases in the attacker's private benefit $\pi$
- decreases in the attacker's penalty from being caught attacking $T$.

$\Rightarrow$ to make the attacker indifferent between attacking or not.

# Crushing the "optimistic" assumption

**Proposition (Strictly positive prob. of attack and uncaught attack)**

*In equilibrium, an attack happens with probability* $\frac{c}{t} > 0$.
*With probability* $\frac{c}{t} \times \left(1 - \frac{\pi}{T}\right) > 0$, *an uncaught attack happens.*

Doubts on the security of optimistic protocols:

- "optimistic" assumption:
  - Participants behave well (if not, face penalties)
- ...cannot hold when all participants are rational

$\Rightarrow$
  - either assume some participants are irrational (e.g., altruistic)
  - or accept that systems cannot be 100% secure as intended

# Roadmap

1. Institutional Details
2. (Simplest) Model
3. Model extensions and robustness of the main argument
   - Challenger's internalization of harms from attacks
   - An extension to multiple attackers and challengers
   - Breaking budget balance?
4. Connections to and implications for practices
5. Conclusion

# Challenger's internalization of harms from attacks

The challenger may internalize the harm of a successful attack

- due to vested interest in the system

| | | Challenger: Challenge | Challenger: No Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, -\tau)$ |
| | No Attack | $(0, -c)$ | $(0, 0)$ |

**Solving for equilibrium**

- no pure strategy Nash equilibrium
- unique mixed strategy NE (attack w.p. $\frac{c}{t+\tau}$ / monitor w.p. $\frac{\pi}{T}$)

Still, attacks happen (and get uncaught) with strictly positive probability!

- If some parties face 0 cost of monitoring (due to side benefit)?
- these parties then become trusted entities $\rightarrow$ centralization

# Challenger's internalization of harms from attacks

The challenger may internalize the harm of a successful attack

- due to vested interest in the system

| | | Challenger: Challenge | Challenger: No Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, -\tau)$ |
| | No Attack | $(0, -c)$ | $(0, 0)$ |

## Solving for equilibrium

- no pure strategy Nash equilibrium
- unique mixed strategy NE (attack w.p. $\frac{c}{t+\tau}$ / monitor w.p. $\frac{\pi}{T}$)

Still, attacks happen (and get uncaught) with strictly positive probability!

- If some parties face 0 cost of monitoring (due to side benefit)?
- these parties then become trusted entities $\rightarrow$ centralization

# Challenger's internalization of harms from attacks

The challenger may internalize the harm of a successful attack

- due to vested interest in the system

| | | Challenger: Challenge | Challenger: No Challenge |
|---|---|---|---|
| Attacker | Attack | $(\pi - T, t - c)$ | $(\pi, -\tau)$ |
| | No Attack | $(0, -c)$ | $(0, 0)$ |

## Solving for equilibrium

- no pure strategy Nash equilibrium
- unique mixed strategy NE (attack w.p. $\frac{c}{t+\tau}$ / monitor w.p. $\frac{\pi}{T}$)

Still, attacks happen (and get uncaught) with strictly positive probability!

- If some parties face 0 cost of monitoring (due to side benefit)?
- these parties then become trusted entities $\rightarrow$ centralization

# Multiple attackers and challengers

A game with an arbitrary number of $M$ attackers and $N$ challengers.

- independent attacking/monitoring decisions
- reflecting decentralization/max-distance from "representative agent"

Each challenger's payoff (normalized to zero when not monitoring):

- monitoring incurs a private cost of $-c < 0$ plus
- $\frac{1}{n}$ of the total reward amount $mt$, where:
  - $t > 0$: reward *per attack*.
  - $n \leq N$: # monitoring challengers.
  - $m \leq M$: # realized attack(s).

Each attacker's payoff (normalized to zero when not attacking):

- attacking gives a private benefit of $\pi < 0$ plus
- penalty $-T$ if at least one challenger monitors

Again, we assume that $t \leq T$ and $t > c$.

# Multiple attackers and challengers

A game with an arbitrary number of $M$ attackers and $N$ challengers.

- independent attacking/monitoring decisions
- reflecting decentralization/max-distance from "representative agent"

Each challenger's payoff (normalized to zero when not monitoring):

- monitoring incurs a private cost of $-c < 0$ plus
- $\frac{1}{n}$ of the total reward amount $mt$, where:
  - $t > 0$: reward *per attack*.
  - $n \leq N$: # monitoring challengers.
  - $m \leq M$: # realized attack(s).

Each attacker's payoff (normalized to zero when not attacking):

- attacking gives a private benefit of $\pi < 0$ plus
- penalty $-T$ if at least one challenger monitors

Again, we assume that $t \leq T$ and $t > c$.

# Multiple attackers and challengers

A game with an arbitrary number of $M$ attackers and $N$ challengers.

- independent attacking/monitoring decisions
- reflecting decentralization/max-distance from "representative agent"

Each challenger's payoff (normalized to zero when not monitoring):

- monitoring incurs a private cost of $-c < 0$ plus
- $\frac{1}{n}$ of the total reward amount $mt$, where:
  - $t > 0$: reward *per attack*.
  - $n \leq N$: # monitoring challengers.
  - $m \leq M$: # realized attack(s).

Each attacker's payoff (normalized to zero when not attacking):

- attacking gives a private benefit of $\pi < 0$ plus
- penalty $-T$ if at least one challenger monitors

Again, we assume that $t \leq T$ and $t > c$.

# Pure strategy Nash equilibria

Assume symmetry WLOG

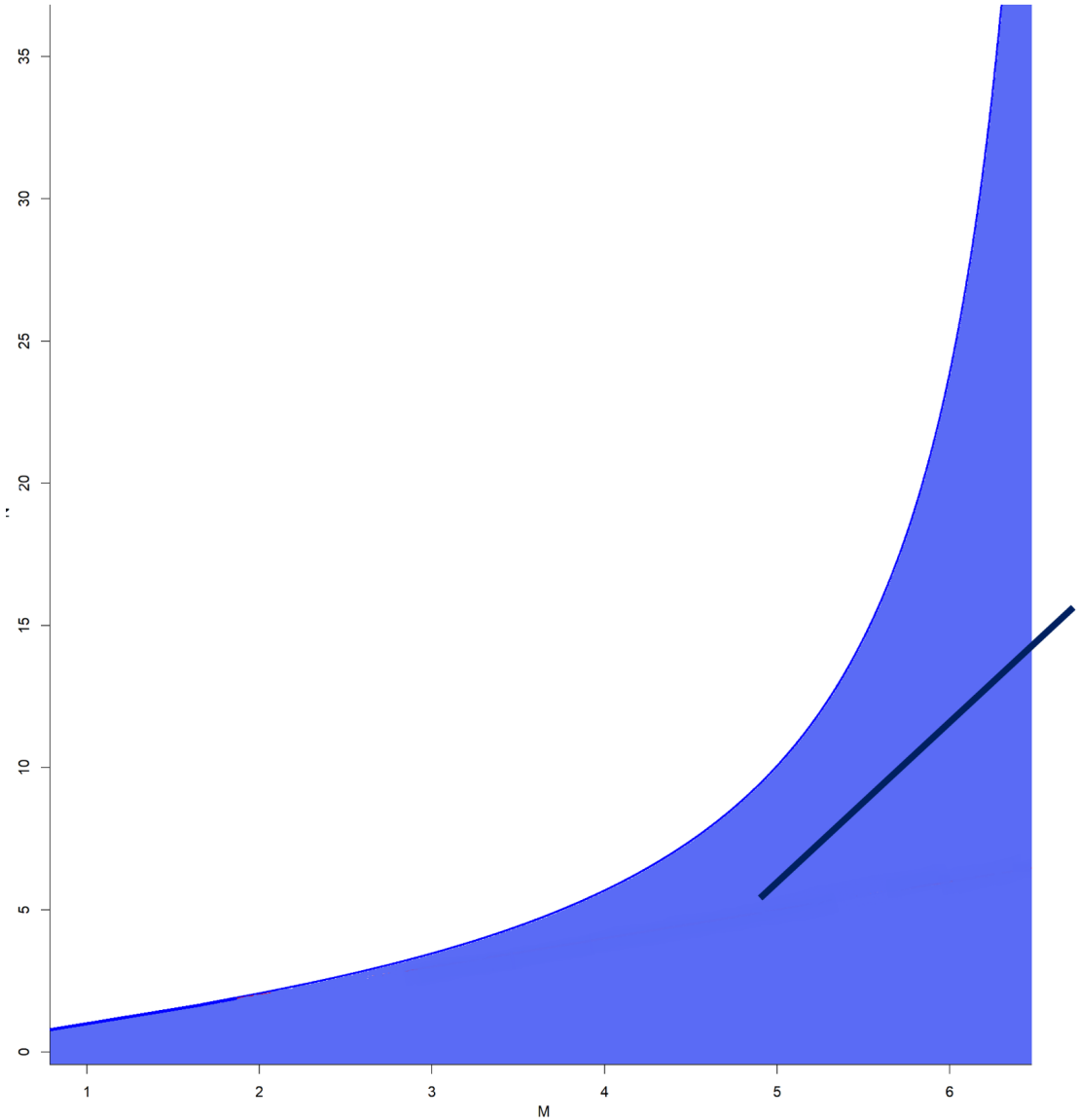
## Proposition

- *With a small number of challengers, the strategic interaction between attackers and challengers features no pure strategy Nash equilibrium*
- *With a large number of challengers, there is a pure strategy Nash equilibrium in which all attackers attack and no challengers monitor*

*Neither sustains the "optimistic" outcome of no attacks in equilibrium*

- when the number of challengers $N < \frac{Mt}{c}$, the incentive analysis is rather similar to that in the two-player game.
- when the number of challengers $N \leq \frac{Mt}{c}$, the challengers would face too fierce competition to cover their private cost of monitoring, so they do not monitor and let attackers freely attack.

# Characterization of all symmetric Nash equilibria




Unique *symmetric* mixed-strategy Nash equilibrium:

Attackers attack with prob. $\frac{1}{M}\frac{N\left(1-\left(\frac{T}{T+\pi}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T+\pi}}\frac{c}{t}$;

Challengers monitor with prob. $1-\left(\frac{T}{T+\pi}\right)^{\frac{1}{N}}$.

Accurate numerical values when $\frac{\pi}{T}=0.999$ and $\frac{c}{t}=0.999$.

▸ Formal Statements

# Crushing the "optimistic" assumption, again

## Proposition ("Optimistic" outcomes are infeasible)

*In the pure-strategy Nash equilibrium:*

- *attack probability* $100\% > 0$
- *expected total number of attacks* $N > 0$
- *all* $N$ *attacks will be left uncaught*

*In the mixed strategy Nash equilibrium:*

- *attack probability* $\frac{1}{M} \frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}} \frac{c}{t} > 0$
- *expected total number of attacks* $\frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}} \frac{c}{t} > 0$
- *in expectation,* $\left(1-\frac{\pi}{T}\right) \frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}} \frac{c}{t} > 0$ *attacks will be left uncaught (and unpenalized)*

# Asymptotics

The number of attackers and challengers converges to infinity

- permissionlessness: potentially larger number of attackers/challengers

## Proposition

*When* $M \to \infty$ *or* $N \to \infty$

- *the total number of attacks is expected to be* $\frac{\ln\left(\frac{1}{1-\frac{\pi}{T}}\right)}{\frac{\pi}{T}} \frac{c}{t} > 0$,
- *in expectation* $\left(1 - \frac{\pi}{T}\right) \frac{\ln\left(\frac{1}{1-\frac{\pi}{T}}\right)}{\frac{\pi}{T}} \frac{c}{t} > 0$ *attacks will be left uncaught*

# Breaking budget balance?

Reward actions (monitoring) rather than outcomes (catching attack)

- blindly reward all self-claimed challengers? How to prevent cheating?
- one solution: introduce system-wide intentional attacks
  - Pure strategy Nash equilibrium exists.
  - Challenger always monitors; attacker never attacks.

Remaining issues: Who gets the reward?

- the first successful challenger only?
  - effectively a "mining" mechanism (PoW in the form of monitoring).
- all potential challengers?
  - costly to fund the reward (without slashing attackers)
  - natural solution: new token issuance
  - Critical considerations:
    - ★ Will a separately minted token accrue value?
    - ★ Even yes, users bear the cost of rewarding monitoring (inflation)
    - ★ likely more expensive than current optimistic systems

# Roadmap

1. Institutional Details
2. (Simplest) Model
3. Model extensions and robustness of the main argument
   - Challenger's internalization of harms from attacks
   - An extension to multiple attackers and challengers
   - Breaking budget balance?
4. Connections to and implications for practices
5. Conclusion

# Theory vs. reality

- **Theory**: optimistic assumptions cannot hold all participants are rational and act on their self-interests.
- **Reality**: various "optimistic" solutions have been widely adopted, gaining attraction, and so far seeming to "work."

Why?

# The ugly truth about the reality

While high-profile optimistic projects all started with the goal of full incentive compatibility, they seem to have all pivoted (given up) over time

- Arbitrum's document now requires additional trust assumptions
  *"assuming that a client believes there to be at least one well behaved active Arbitrum validator..."*
- Ethereum's proof-of-stake update now also requires the presence of altruistic participants, e.g., in Prysm (consensus client) document:
  *"Running a slasher is meant to be an altruistic action... "*
- Ethereum's PoS update also gives up on full incentive-compatibility
  - final spec version removed whistleblower rewards entirely
  - ...instead resort to altruistic behavior.

# Why should we care?

Bring awareness of optimistic systems' incentive incompatibility

1. blockchain community is used to the trustlessness motto
   - e.g., Bitcoin
   - early discussions about optimistic systems, PoS, etc.
2. but optimistic systems require additional trust...
   - even if current industry environment may allow a certain level of trust
   - trust is likely only sustainable within a given financial amount
   - can incentive-incompatible solutions remain secure...
     - ★ ...once gaining further adoption?
   - more money at stake $\Rightarrow$ more temptation to abuse trust

# Cryptographic Alternatives to Optimistic Protocols

Succinct non-interactive arguments of knowledge (SNARK):

- e.g., ZK-rollup vs. optimistic rollup for blockchain scaling
  - Participants must include a succinct proof for messages.
  - Only messages with correct proofs are accepted.
- Examples: Aztec, Starknet, zkSync.
- bottleneck in prover efficiency

Problem mitigated but not fully eliminated (unless verification cost $= 0$)

Insight:

- Our study: Optimistic protocols cannot ensure perfect security.
- Community views: optimistic $=$ temporary; SNARK $=$ long-term.
- maybe...

# Roadmap

1. Institutional Details
2. (Simplest) Model
3. Model extensions and robustness of the main argument
   - Challenger's internalization of harms from attacks
   - An extension to multiple attackers and challengers
   - Breaking budget balance?
4. Connections to and implications for practices
5. Conclusion

# Conclusion

What have we done?

- formalize the strategic interactions in optimistic blockchain protocols
- show that current designs cannot be fully incentive-compatible

Not my intention to criticize optimistic protocols

- they do provide valuable *short-term* solutions for several pressing issues in blockchain (e.g., scaling, proof-of-stake, & interoperability)
- need to tolerate new ideas to meet community needs when alternative solutions are still in development

My hope:

- dissipate community misconceptions
- guide towards more secure systems for crucial applications

# Reference


**Amoussou-Guenou, Yackolley, Bruno Biais, Maria Potop-Butucaru, and Sara Tucci-Piergiovanni.** 2023. "Committee-based blockchains as games between opportunistic players and adversaries." *Review of Financial Studies*, hhad051. tex.eprint: https://academic.oup.com/rfs/advance-article-pdf/doi/10.1093/rfs/hhad051/50684345/hhad051.pdf.

**Benhaim, Alon, Brett Hemenway Falk, and Gerry Tsoukalas.** 2021. "Scaling blockchains: Can elected committees help?" *Available at SSRN 3914471*.

**Bousfield, Lee, Rachel Bousfield, Chris Buckland, Ben Burgess, Joshua Colvin, Edward W. Felten, Steven Goldfeder, Daniel Goldman, Braden Huddleston, Harry Kalodner, Frederico Arnaud Lacs, Harry Ng, Aman Sanghi, Tristan Wilson, Valeria Yermakova, and Tsahi Zidenberg.** 2018. "Arbitrum nitro: A second-generation optimistic rollup."

**Cong, Lin William, Xiang Hui, Catherine Tucker, and Luofeng Zhou.** 2023. "Scaling smart contracts via layer-2 technologies: Some empirical evidence." National Bureau of Economic Research.

**Halaburda, Hanna, Zhiguo He, and Jiasun Li.** 2021. "An economic model of consensus on distributed ledgers." National Bureau of Economic Research.

**He, Zhiguo, Jiasun Li, and Zhengxun Wu.** 2023. "Don't trust, verify: The case of slashing from a popular ethereum explorer." *WWW '23*, 1078–1084. Association for Computing Machinery.

**Jermann, Urban J.** 2023. "A Macro Finance Model for Proof-of-Stake Ethereum." *Available at SSRN 4335835*.

**John, Kose, Thomas J Rivera, and Fahad Saleh.** 2020. "Economic implications of scaling blockchains: Why the consensus protocol matters." *Available at SSRN 3750467*.

**John, Kose, Thomas J Rivera, and Fahad Saleh.** 2021. "Equilibrium staking levels in a proof-of-stake blockchain." *Available at SSRN 3965599*.

**Kalodner, Harry, Steven Goldfeder, Xiaoqi Chen, S Matthew Weinberg, and Edward W Felten.** 2018. "Arbitrum: Scalable, private smart contracts." *USENIX Security '18*, 1353–1370.

**Kogan, Leonid, Giulia Fanti, and Pramod Viswanath.** 2021. "Economics of proof-of-stake payment systems."

**Li, Jiasun, and Zhengxun Wu.** 2023. "Arbitrary message passing across blockchains." *SSRN 4417670*.

**McCorry, Patrick, Chris Buckland, Bennet Yee, and Dawn Song.** 2021. "Sok: Validating bridges as a scaling solution for blockchains." *Cryptology ePrint Archive*.

**Saleh, Fahad.** 2021. "Blockchain without waste: Proof-of-stake." *Review of Financial Studies*, 34(3): 1156–1190. Publisher: Oxford University Press.

**Whitehat, Barry, Alex Gluchowski, HarryR, Yondon Fu, and Philippe Castonguay.** 2018. "Roll_up/roll_back snark side chain˜ 17000 tps."

# Symmetric mixed strategy Nash equilibria

- When $\frac{1}{M}\frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}}\frac{c}{t} > 1$ (so $N$ is large), the game features a unique pure strategy Nash equilibrium in which attackers attack for sure and challengers do not monitor at all;
- When $\frac{Mt}{c} < N$ (so $N$ is small), the game features a unique mixed strategy Nash equilibrium in which attackers attack with probability

$$\frac{1}{M}\frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}}\frac{c}{t}$$

  and challengers monitor with probability $1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}$;
- When $\frac{1}{M}\frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}}\frac{c}{t} \leq 1$ and $\frac{Mt}{c} \leq N$ (so $N$ take intermediate values), the game features both the pure and mixed strategy Nash equilibrium.

Back

# Symmetric mixed strategy Nash equilibria

## Lemma

*The expression* $N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)$ *is strictly increasing in* $N$*, and obtains a limit of* $\ln\left(\frac{1}{1-\frac{\pi}{T}}\right)$ *when* $N$ *approaches infinity.*

As a result, when $\frac{\ln\left(\frac{T}{T-\pi}\right)}{\frac{\pi}{T}}\frac{c}{t} > M$,

$$\frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}}\frac{c}{t} = M$$

as an equation of $N$ has a unique solution, which we denote as $N^*$. Back

# Symmetric mixed strategy Nash equilibria

## Lemma

- *When* $\frac{\ln\left(\frac{T}{T-\pi}\right)}{\frac{\pi}{T}}\frac{c}{t} > M$ *and* $N > N^*$, *the strategic interaction between attackers and challengers features a unique symmetric mixed strategy Nash equilibrium which happens to be a pure-strategy Nash equilibrium. In this equilibrium, attackers attack for sure while challengers do not monitor at all.*
- *Otherwise, there is a unique symmetric mixed strategy Nash equilibrium, in which attackers attack with probability*

$$\frac{1}{M}\frac{N\left(1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}}\right)}{\frac{\pi}{T}}\frac{c}{t} \in (0,1],$$

*and the challengers monitor with probability* $1-\left(1-\frac{\pi}{T}\right)^{\frac{1}{N}} \in (0,1)$.

Back

# Comparative Statics

- Blue region (no all-attack-no-monitoring pure-strategy NE)
  - enlarges when $\frac{t}{c}$ increases.
- Red region (no mixed strategy NE)
  - shrinks when $\frac{t}{c}$ increases or when $\frac{\pi}{T}$ decreases.

Back